involve

a journal of mathematics

msp

involve

msp.org/involve

INVOLVE YOUR STUDENTS IN RESEARCH

Involve showcases and encourages high-quality mathematical research involving students from all academic levels. The editorial board consists of mathematical scientists committed to nurturing student participation in research. Bridging the gap between the extremes of purely undergraduate research journals and mainstream research journals, *Involve* provides a venue to mathematicians wishing to encourage the creative involvement of students.

PRODUCTION

Silvio Levy, Scientific Editor

Cover: Alex Scorpan

Occurrence graphs of patterns in permutations

Bjarni Jens Kristinsson and Henning Ulfarsson

(Communicated by Anant Godbole)

We define the occurrence graph $G_p(\pi)$ of a pattern p in a permutation π as the graph whose vertices are the occurrences of p in π, with edges between the vertices if the occurrences differ by exactly one element. We then study properties of these graphs. The main theorem in this paper is that every hereditary property of graphs gives rise to a permutation class.

1. Introduction

The research area of permutation patterns can be traced back to [MacMahon 1915, Section III, Chapter V] where it is shown that permutations without an increasing subsequence of length 3 (avoiding 123 in the language introduced below) are counted by the Catalan numbers. Another famous result is the Erdős–Szekeres theorem [1935] which says that a permutation of length $(n-1)(m-1)+1$ has an increasing subsequence of length n (the pattern $12\cdots n$) or a decreasing subsequence of length m (the pattern $m\cdots 21$). The field came into its own when Knuth [1968] showed that "stack-sortable" permutations are the 231-avoiding permutations and are enumerated by the Catalan numbers. Since then dozens of papers have been written about enumerations of permutations avoiding patterns, their structure, and connections to other objects in mathematics. See [Kitaev 2011] for an overview. The goal of this paper is to connect the study of permutation patterns with properties of graphs.

We define the *occurrence graph* $G_p(\pi)$ of a pattern p in a permutation π as the graph where each vertex represents an occurrence of p in π. Vertices share an edge if the occurrences they represent differ by exactly one element. We study properties of these graphs and show that every *hereditary property* of graphs gives rise to a *permutation class*, which we define below.

The motivation for defining these graphs comes from the algorithm discussed in the proof of the simultaneous shading lemma by Claesson, Tenner and Ulfarsson

[Claesson et al. 2015, Lemma 7.6]. The steps in that algorithm can be thought of as constructing a path in an occurrence graph, terminating at a desirable occurrence of a pattern.

2. Basic definitions

We begin by reviewing some standard definitions.

Definition 2.1. A *graph* is an ordered pair $G = (V, E)$, where V is a set of *vertices* and E is a set of two-element subsets of V. The elements $\{u, v\} \in E$ are called *edges* and connect the vertices. Two vertices u and v are *neighbors* if $\{u, v\} \in E$. The *degree* of a vertex v is the number of neighbors it has. A graph $G' = (V', E')$ is a *subgraph* of G if $V' \subseteq V$ and $E' \subseteq \{\{u, v\} \in E : u, v \in V'\}$.

The reader might have noticed that our definition of a graph excludes those with loops and multiple edges between vertices. We often write uv as shorthand for $\{u, v\}$ and in case of ambiguity we use $V(G)$ and $E(G)$ instead of V and E.

Definition 2.2. Two graphs G and H are *isomorphic* if there exists a bijection from $V(G)$ to $V(H)$ such that two vertices in G are neighbors if and only if the corresponding vertices (according to the bijection) in H are neighbors. We denote this by $G \cong H$.

We let $[\![1, n]\!]$ denote the integer interval $\{1, \dots, n\}$.

Definition 2.3. A *permutation of length n* is a bijective function $\sigma : [\![1, n]\!] \to [\![1, n]\!]$. We denote the permutation by $\sigma = \sigma(1)\sigma(2) \cdots \sigma(n)$. The permutation $\mathrm{id}_n = 12 \cdots n$ is the *identity permutation* of length n.

The set of permutations of length n is denoted by $\mathfrak{S}_n$. The set of all permutations is $\mathfrak{S} = \bigcup_{n=0}^{+\infty} \mathfrak{S}_n$. Note that $\mathfrak{S}_0 = \{\mathcal{E}\}$, where $\mathcal{E}$ is the empty permutation, and $\mathfrak{S}_1 = \{1\}$. There are $n!$ permutations of length n.

Definition 2.4. A *grid plot* or *grid representation* of a permutation $\pi \in \mathfrak{S}_n$ is the subset $\mathrm{Grid}(\pi) = \{(i, \pi(i)) : i \in [\![1, n]\!]\}$ of the Cartesian product $[\![1, n]\!]^2 = [\![1, n]\!] \times [\![1, n]\!]$.

Example 2.5. Let $\pi = 42135$. The grid representation of π is

$$\mathrm{Grid}(42135) = \quad \vcenter{\hbox{}} \quad .$$
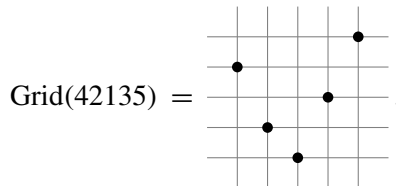
The central definition in the theory of permutation patterns is how permutations lie inside other (larger) permutations. Before we define that precisely we need a preliminary definition:

Definition 2.6. Let $a_1, \ldots, a_k$ be distinct integers. The *standardization* of the string $a_1 \cdots a_k$ is the permutation $\sigma \in \mathfrak{S}_k$ such that $a_1 \cdots a_k$ is order isomorphic to $\sigma(1) \cdots \sigma(k)$. In other words, for every $i \neq j$ we have $a_i < a_j$ if and only if $\sigma(i) < \sigma(j)$. We denote this by $\mathrm{st}(a_1 \cdots a_k) = \sigma$.

For example $\mathrm{st}(253) = 132$ and $\mathrm{st}(132) = 132$.

Definition 2.7. Let p be a permutation of length k. We say that the permutation $\pi \in \mathfrak{S}_n$ *contains* p if there exist indices $1 \leq i_1 < \cdots < i_k \leq n$ such that $\mathrm{st}(\pi(i_1) \cdots \pi(i_k)) = p$. The subsequence $\pi(i_1) \cdots \pi(i_k)$ is an *occurrence* of p in π with the *index set* $\{i_1, \ldots, i_k\}$. The increasing sequence $i_1 \cdots i_k$ will be used to denote the order-preserving injection $i : [\![1, k]\!] \to [\![1, n]\!]$, $j \mapsto i_j$, which we call the *index injection* of p into π for this particular occurrence.

The set of all index sets of p in π is the *occurrence set* of p in π, denoted by $V_p(\pi)$. If π does not contain p, then π *avoids* p. In this context the permutation p is called a (*classical permutation*) *pattern*.

Unless otherwise stated, we write the index set $\{i_1, \ldots, i_n\}$ in ordered form, i.e., such that $i_1 < \cdots < i_n$, in accordance with how we write the index injection.

The set of all permutations that avoid p is $\mathrm{Av}(p)$. More generally for a set of patterns M we define

$$\mathrm{Av}(M) = \bigcap_{p \in M} \mathrm{Av}(p).$$

Example 2.8. The permutation 42135 contains five occurrences of the pattern 213, namely 425, 415, 435, 213 and 215. The occurrence set is

$$V_{213}(42135) = \{\{1, 2, 5\}, \{1, 3, 5\}, \{1, 4, 5\}, \{2, 3, 4\}, \{2, 3, 5\}\}.$$

The permutation 42135 avoids the pattern 132.

Sets of permutations that can be defined by the avoidance of patterns are given a special name:

Definition 2.9. A set of permutations $\mathscr{C}$ that is *closed downwards*, i.e., if $\pi \in \mathscr{C}$ then $p \in \mathscr{C}$ for every pattern p in π, is called a *permutation class*. A permutation class can be written as $\mathrm{Av}(M)$, where M is a set of classical permutation patterns. If M is minimal, then it is called the *basis* of the class.

3. Occurrence graphs

We now formally define occurrence graphs.

Definition 3.1. For a pattern p of length k and a permutation π we define the *occurrence graph* $G_p(\pi)$ of p in π as follows:

- The set of vertices is $V_p(\pi)$, the occurrence set of p in π.
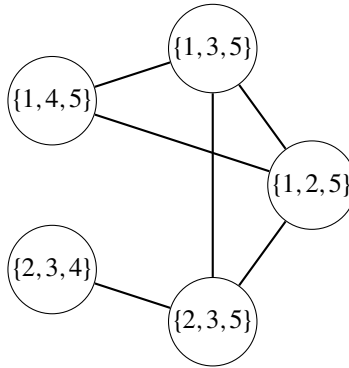
Figure 1. The occurrence graph $G_{213}(42135)$.

- uv is an edge in $G_p(\pi)$ if the vertices $u = \{u_1, \ldots, u_k\}$ and $v = \{v_1, \ldots, v_k\}$ in $V_p(\pi)$ differ by exactly one element, i.e., if

$$|u \setminus v| = 1 = |v \setminus u|.$$

Example 3.2. In Example 2.8 we derived the occurrence set $V_{213}(42135)$. We compute the edges of $G_{213}(42135)$ by comparing the vertices two at a time to see if the sets differ by exactly one element. The graph is shown in Figure 1.

Remark 3.3. For a permutation π of length n the graph $G_{\mathscr{E}}(\pi)$ is a graph with one vertex and no edges and $G_1(\pi)$ is a clique on n vertices.

Following the definition of these graphs there are several natural questions that arise. For example, for a fixed pattern p, which occurrence graphs $G_p(\pi)$ satisfy a given graph property, such as being connected or being a tree? Before we answer questions of this sort we consider a simpler question: what can be said about the graph $G_{12}(\mathrm{id}_n)$?

4. The pattern $p = 12$ and the identity permutation

In this section we only consider the pattern $p = 12$ and let $n \geq 2$. For this choice of p and a fixed n the identity permutation $\pi = 1 \cdots n$ contains the most occurrences of p. Indeed, every set $\{i, j\}$ with $i \neq j$ is an index set of p in π. We can choose this pair in

$$\binom{n}{2} = \frac{n(n-1)}{2}$$

different ways. Therefore, this is the size of the vertex set of $G = G_p(\pi)$.

Every vertex $u = \{i, j\}$ in G is connected to $n-2$ vertices $v = \{i, j'\}$, $j' \neq j$, and $n-2$ vertices $w = \{i', j\}$, $i' \neq i$. Thus, the degree of every vertex in G is $2(n-2)$. By
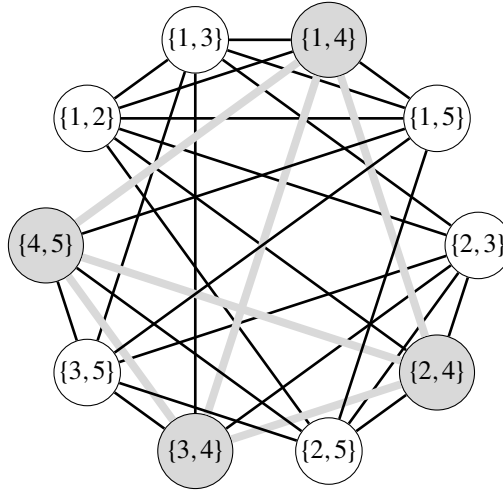
Figure 2. The graph $G_{12}(12345)$.

summing over the set of vertices and dividing by 2 we get the number of edges in G:

$$|E(G)| = \frac{n(n-1)(n-2)}{2} = 3\binom{n}{3}.$$

A triangle in G consists of three vertices u, v, w with edges uv, vw, wu. If $u = \{i, j\}$ (not necessarily in ordered form) then we can assume v is $\{j, k\}$. For this triplet to be a triangle w must connect to both u and v, and therefore w must either be the index set $\{i, k\}$ or $\{j, j'\}$, where $j' \neq i, k$. In the first case, we just need to choose three indices i, j, k. In the second case we start by choosing the common index k and then we choose the remaining indices. Thus the number of triangles in G is

$$\binom{n}{3} + n\binom{n-1}{3} = (n-2)\binom{n}{3}.$$

Example 4.1. The graph $G_{12}(12345)$ is pictured in Figure 2. It has 10 vertices, 30 edges, and 30 triangles. It also has 5 subgraphs isomorphic to K_4, one of them highlighted with thicker gray edges and gray vertices.

The following proposition generalizes the observations above to larger cliques.

Proposition 4.2. *For $n > 0$, the number of cliques of size $k > 3$ in $G_{12}(\mathrm{id}_n)$ is*

$$(k+1)\binom{n}{k+1} = n\binom{n-1}{k}.$$

Proof. The vertices $\{a_1, b_1\}, \{a_2, b_2\}, \ldots, \{a_k, b_k\}$ in a clique of size $k > 3$ must have a common index, say $\ell = a_1 = a_2 = \cdots = a_k$, without loss of generality. The remaining indices $b_1, b_2, \ldots, b_k$ can be chosen as any subset of the other $n-1$ indices. This explains the right-hand side of the equation in the proposition. $\qquad\square$

5. Hereditary properties of graphs

Intuitively one might think that if a pattern p is contained inside a larger pattern q, then one of the occurrence graphs $G_p(\pi)$ and $G_q(\pi)$ (for any permutation π) would be contained inside the other. But this is not the case as the following examples demonstrate.

Example 5.1. (1) Let $p = 12$, $q = 231$ and $\pi = 3421$. The occurrence sets are $V_p(\pi) = \{\{1, 2\}\}$ and $V_q(\pi) = \{\{1, 2, 3\}, \{1, 2, 4\}\}$. The cardinality of the set $V_p(\pi)$ is smaller than the cardinality of $V_q(\pi)$.

(2) If on the other hand $p = 12$, $q = 123$ and $\pi = 123$ then the occurrence sets are $V_p(\pi) = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$ and $V_q(\pi) = \{\{1, 2, 3\}\}$. Thus, in this case, the cardinality of $V_p(\pi)$ is larger than the cardinality of $V_q(\pi)$.

However, for a fixed pattern p, we obtain an inclusion of one occurrence graph in another, in Proposition 5.4. First we need a lemma.

Lemma 5.2. *Let p be a pattern and π, σ be two permutations. For an occurrence of π in σ the index injection induces an injection $\Phi_p : V_p(\pi) \to V_p(\sigma)$.*

Proof. Let p, π, σ be permutations of lengths l, m, n respectively. Every $v = \{i_1, \ldots, i_l\} \in V_p(\pi)$ is an index set of p in π with index injection i. Let j be an index injection for an index set $\{j_1, \ldots, j_m\}$ of π in σ. It is easy to see that $u = \{j_{i_1}, \ldots, j_{i_l}\}$ is an index set of p in σ because $j \circ i$ is an index injection of p into σ. Define $\Phi_p(v) = u$. $\square$

Example 5.3. Let $p = 12$, $\pi = 132$ and $\sigma = 24153$. There are three occurrences of π in σ: 243, 253 and 153 with respective index injections 125, 145 and 345.

For a given index injection, say $i = 345$, we obtain the injection Φ_p by mapping every $\{v_1, v_2\} \in V_p(\pi)$ to $\{i_{v_1}, i_{v_2}\} \in V_p(\sigma)$. We calculate that Φ_p maps $\{1, 2\}$ to $\{i_1, i_2\} = \{3, 4\}$ and $\{1, 3\}$ to $\{i_1, i_3\} = \{3, 5\}$; see Figure 3.
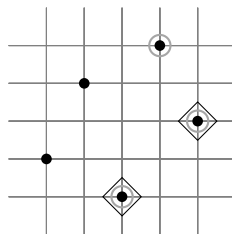
Figure 3. The occurrence of π in σ that is defined by the index injection $i = 345$ is highlighted with gray circles. The occurrence set $\{1, 3\}$ of p in π is mapped with the injection Φ_p, induced by i, to the index set $\{3, 5\}$ of p in σ, highlighted with black diamonds.
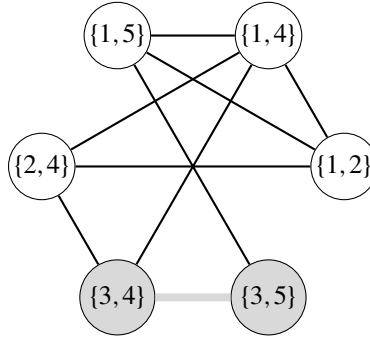
Figure 4. The graph $G_{12}(24153)$ with a highlighted subgraph isomorphic to $G_{12}(132)$.

Proposition 5.4. *Let p be a pattern and π, σ be two permutations. For an occurrence of π in σ the index injection induces an isomorphism of the occurrence graph $G_p(\pi)$ with a subgraph of $G_p(\sigma)$.*

Proof. From Lemma 5.2 we have the injection $\Phi_p : V_p(\pi) \to V_p(\sigma)$. We need to show for every $uv \in E(G_p(\pi))$ that $\Phi_p(u)\Phi_p(v) \in E(G_p(\sigma))$. Let uv be an edge in $G_p(\pi)$, where $u = \{u_1, \ldots, u_l\}$ and $v = \{v_1, \ldots, v_l\}$. For every index injection j of π into σ, the vertices u, v map to $\Phi_p(u) = \{j(u_1), \ldots, j(u_l)\}$, $\Phi_p(v) = \{j(v_1), \ldots, j(v_l)\}$ respectively. Since j is an injection, there exists an edge between these two vertices in $G_p(\sigma)$. $\square$

Example 5.5. We will continue with Example 5.3 and show how the index injection $i = 345$ defines a subgraph of $G_p(\sigma)$ which is isomorphic to $G_p(\pi)$. The occurrence graph of p in π is a graph on two vertices $\{1, 2\}$ and $\{1, 3\}$ with an edge between them. The occurrence graph $G_p(\sigma)$ with the highlighted subgraph induced by i is shown in Figure 4.

The next example shows that different occurrences of π in σ do not necessarily lead to different subgraphs of $G_p(\sigma)$.

Example 5.6. If $p = 12$, $\pi = 312$ and $\sigma = 3412$ there are two occurrences of π in σ. The index injections are $i = 134$ and $i' = 234$. However, as $(i_2, i_3) = (i'_2, i'_3)$ and $\{2, 3\}$ is the only index set of p in π, we obtain the same injection Φ_p and therefore the same subgraph of $G_p(\sigma)$ for both index injections.

We call a property of a graph G *hereditary* if it is invariant under isomorphisms and for every subgraph of G the property also holds. For example the properties of being a forest, bipartite, planar or k-colorable are hereditary properties, while being a tree is not hereditary. A set of graphs defined by a hereditary property is a *hereditary class*.

p	basis	numerical sequence	OEIS
12	123, 1432, 3214	1, 2, 5, 12, 26, 58, 131, 295	A116716
123	1234, 12543, 14325, 32145	1, 2, 6, 23, 100, 462, 2207, 10758	
132	1432, 12354, 13254, 13452, 15234, 21354, 23154, 31254, 32154	1, 2, 6, 23, 95, 394, 1679, 7358	

Table 1. Experimental results for bipartite occurrence graphs, computed with permutations up to length 8.

Given c, a property of graphs, we define a set of permutations:

$$\mathcal{G}_{p,c} = \{\pi \in \mathfrak{S} : G_p(\pi) \text{ has property } c\}.$$

We can now state the main theorem of the paper.

Theorem 5.7. *Let c be a hereditary property of graphs. For any pattern p the set $\mathcal{G}_{p,c}$ is a permutation class; i.e., there exists a set of classical patterns M such that*

$$\mathcal{G}_{p,c} = \text{Av}(M).$$

Proof. Let σ be a permutation such that $G_p(\sigma)$ satisfies the hereditary property c and let π be a pattern in σ. By Proposition 5.4 the graph $G_p(\pi)$ is isomorphic to a subgraph of $G_p(\sigma)$ and thus inherits the property c.   □

In the remainder of this section we consider two hereditary classes of graphs: bipartite graphs and forests. Recall that a nonempty simple graph on n vertices ($n > 0$) is a *tree* if and only if it is connected and has $n - 1$ edges. An equivalent condition is that the graph has at least one vertex and no simple cycles (a sequence of unique vertices $v_1, \ldots, v_k$ with edges $v_1 v_2, \ldots, v_{k-1} v_k, v_k v_1$). A *forest* is a disjoint union of trees. The empty graph is a forest but not a tree. *Bipartite* graphs are graphs that can be colored with two colors in such a way that no edge joins two vertices with the same color. We note that every forest is a bipartite graph.

Table 1 shows experimental results, obtained using software developed by Magnusson and the second author [Magnusson and Ulfarsson 2012], on which occurrence graphs with respect to the patterns $p = 12$, $p = 123$, $p = 132$ are bipartite. Permutations and patterns have the eight symmetries of the square, as can be seen from their grid representation. We only consider one representative from each symmetry class.

In the following theorem we verify the statements in line 1 of Table 1. We leave the remainder of the table as conjectures.

Figure 5. The vertices v_1 and v_2 (shown as line segments inside the permutation π) share the index i_1.

Theorem 5.8. *Let c be the property of being bipartite. Then*

$$\mathscr{G}_{12,c} = \mathrm{Av}(123, 1432, 3214).$$

The OEIS sequence A116716 enumerates a symmetry of this permutation class.

The proof of this theorem relies on a proposition characterizing the cycles in the graphs under consideration.

Proposition 5.9. *If the graph $G_{12}(\pi)$ has a cycle of length $k > 4$ then it also has a cycle of length* 3.

Proof. Let π be a permutation such that $G_{12}(\pi)$ contains a cycle of length $k > 4$. Label the vertices in the cycle $v_1, \ldots, v_k$ with $v_l = \{i_l, j_l\}$, $i_l < j_l$, for $l = 1, \ldots, k$.

The vertices v_1 and v_2 in the cycle have exactly one index in common. If $i_2 = j_1$ then the vertices v_1, v_2, $\{i_1, j_2\}$ form a triangle. So we can assume $i_1 = i_2$. If $j_1 < j_2$ and $\pi(j_1) < \pi(j_2)$ (or $j_1 > j_2$ and $\pi(j_1) > \pi(j_2)$) then $u = \{j_1, j_2\}$ is an occurrence of 12 in π, forming a triangle v_1, v_2, u. So either $j_1 > j_2$ and $\pi(j_1) < \pi(j_2)$ holds, or, without loss of generality (see Figure 5), $j_1 < j_2$ and $\pi(j_1) > \pi(j_2)$.

Next we look at the edge $v_2 v_3$ in the cycle. If the vertices have the index i_1 in common then v_1, v_2, v_3 form a triangle in $G_{12}(\pi)$. So assume that v_2 and v_3 have the index j_2 in common with the conditions $i_3 > i_1$ and $\pi(i_3) < \pi(i_1)$ (because else there are more vertices and edges forming a cycle of length 3 in $G_{12}(\pi)$). Continuing down this road we know that $v_3 v_4$ is an edge with shared index i_3 and conditions $j_3 > i_3$ and $\pi(j_3) < \pi(j_1)$; see Figure 6, where we consider the case $i_3 > j_1$, and $\pi(j_3) < \pi(i_1)$.

Graphically, it is quite obvious that we cannot extend the path in Figure 6 with more southwest-northeast line segments (a sequence of vertices $v_5, \ldots, v_k$) such that the extension closes the path into a cycle without adding more edges (line segments) between vertices that are not adjacent in the cycle and thus forming a cycle of length 3 in the occurrence graph. More precisely, for an edge between v_k and v_1 to exist we must have $v_k = \{i_k, j_k\}$ with a nonempty intersection with v_1. Analyzing each of these cases completes the proof. □
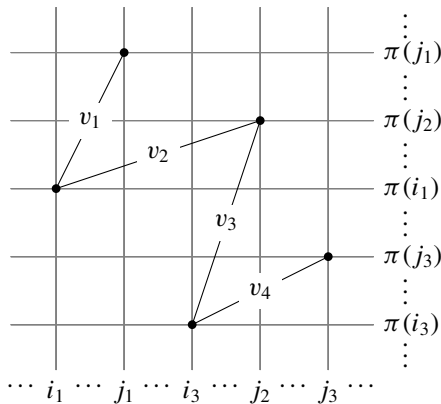
Figure 6. The vertices v_1, v_2, v_3, v_4.

Proof of Theorem 5.8. If π contains any of the patterns 123, 1432, 3214 then $G_p(\pi)$ contains a subgraph that is isomorphic to a triangle. So if $\pi \notin \mathrm{Av}(123, 1432, 3214)$ then $G_{12}(\pi)$ contains an odd cycle and is therefore not bipartite.

On the other hand, let π be a permutation such that $G_{12}(\pi)$ is not bipartite. Then the occurrence graph contains an odd cycle which by Proposition 5.9 implies the graph has a cycle of length 3. The indices corresponding to this cycle form a pattern of length 3 or 4 in π with occurrence graph that is a cycle of length 3. It is easy to see that the only permutations of this length with occurrence graph a cycle of length 3 are 123, 1432 and 3214. Therefore π must contain at least one of the patterns. □

Table 2 considers occurrence graphs that are forests.

p	basis	numerical sequence	OEIS
12	123, 1432, 2143, 3214	1, 2, 5, 11, 24, 53, 117, 258	A052980
123	1234, 12543, 13254, 14325, 21354, 21435 32145	1, 2, 6, 23, 97, 429, 1947, 8959	
132	1432, 12354, 12453, 12534, 13254, 13452, 14523, 15234, 21354, 21453, 21534, 23154, 31254, 32154	1, 2, 6, 23, 90, 359, 1481, 6260	

Table 2. Experimental results for occurrence graphs that are forests, computed with permutations up to length 8.

In the following theorem we verify the statements in line 1 of Table 2. We leave the remainder of the table as conjectures.

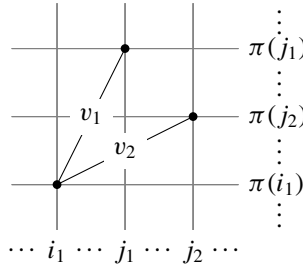**Theorem 5.10.** *Let c be the property of being a forest. Then*

$$\mathscr{G}_{12,c} = \mathrm{Av}(123, 1432, 2143, 3214).$$

Proof. If π contains the pattern 2143 then $G_{12}(\pi)$ contains a subgraph that is isomorphic to a cycle of length 4, according to Proposition 5.4, because $G_{12}(2143)$ is a cycle of length 4. If π contains any of the patterns 123, 1432, 3214 then its occurrence graph is not bipartite by Theorem 5.8, and in particular is not a forest.

On the other hand, let π be a permutation such that $G_{12}(\pi)$ is not a forest. Then the occurrence graph contains a cycle. Proposition 5.9 implies that the cycle must have length either 3 or 4. But it is easy to see that the only permutations with occurrence graphs that are cycles of length 3 or 4 are 123, 1432, 2143, 3214. Therefore π must contain at least one of the patterns. $\square$

6. Nonhereditary properties of graphs

This section is devoted to graph properties that are not hereditary. Thus Theorem 5.7 does not guarantee the permutations whose occurrence graphs satisfy the property form a pattern class. Experimental results in Table 3 seem to suggest that some properties still give rise to permutation classes.

To describe permutations π such that $G_{12}(\pi)$ is connected, we need the language of mesh patterns, which we briefly review here. A *mesh pattern* is a pair (p, s) where p is a classical pattern and the *mesh s* is a subset of $[\![0, |p|]\!] \times [\![0, |p|]\!]$. The

property	basis	numerical sequence	OEIS
connected	see Figure 7	1, 2, 6, 23, 111, 660, 4656, 37745	
chordal	1234, 1243, 1324, 2134, 2143	1, 2, 6, 19, 61, 196, 630, 2025	
clique	1234, 1243, 1324, 1342, 1423, 2134, 2143, 2314, 2413, 3124, 3142, 3412	1, 2, 6, 12, 20, 30, 42, 56	A002378 from $n=2$
tree	very large nonclassical basis	0, 1, 4, 9, 16, 25, 36, 49	A000290

Table 3. Experimental results for the pattern $p = 12$, computed with permutations up to length 8.
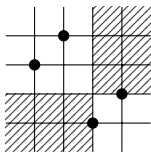
Figure 7. The mesh pattern $m = (p, s)$, where $p = 3412$ and s consists of the boxes $(0, 0)$, $(0, 1)$, $(1, 0)$, $(1, 1)$, $(2, 0)$, $(2, 1)$, $(3, 2)$, $(3, 3)$, $(3, 4)$, $(4, 2)$, $(4, 3)$, $(4, 4)$.
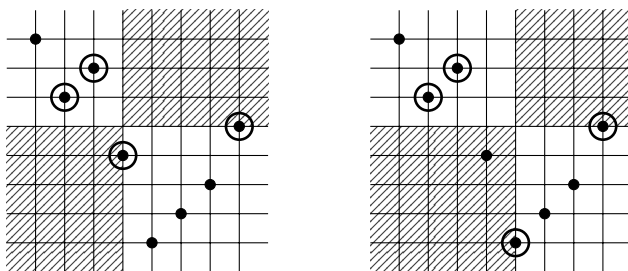
Figure 8. Two occurrences of 3412 in 86741235. Only the left one is an occurrence of the mesh pattern in Figure 7.

elements of s are often called (*shaded*) *boxes*, and informally they denote areas which permutation points are not allowed to occupy in a valid occurrence of (p, s).

For the formal definition of these generalized patterns see [Brändén and Claesson 2011]. We give an example here to illustrate the role the mesh plays.

Example 6.1. There are seven occurrences of the classical pattern $p = 3412$ in the permutation $\pi = 86741235$. In Figure 8 we have highlighted two of them: the subsequences 6745 and 6715. Only the first one is an occurrence of the mesh pattern $m = (p, s)$ in Figure 7, since the mesh can be "stretched" over the grid of the permutation without containing any points. In the second occurrence the point 4 in the permutation occupies a "forbidden" region defined by the mesh.

Theorem 6.2. *Let c be the property of being connected. Then*

$$\mathcal{G}_{12,c} = \mathrm{Av}(m),$$

where m is the mesh pattern in Figure 7. The generating function for the enumeration of these permutations is

$$\frac{F(x) - x}{(1 - x)^2} + \frac{1}{1 - x},$$

where $F(x) = 1 - 1 / \sum k! \, x^k$ is the generating function for the skew-indecomposable permutations (see, e.g., [Comtet 1974, p. 261]).

Before we prove the theorem we recall that the *skew-sum* of two permutations π and σ is the permutation $\pi \ominus \sigma$ obtained by adding the length of σ to every letter of π and then appending σ to it. For example $132 \ominus 2413 = 5762413$. We say that a permutation is *skew-decomposable* if it can be written as a skew-sum of two nonempty permutations. Otherwise it is *skew-indecomposable*. Every permutation can be written as a skew-sum of skew-indecomposable permutations, and we call that the *skew-decomposition* of the permutation.

Proof. If the graph $G_{12}(\pi)$ is disconnected then π has two occurrences of 12 in distinct skew-components, A and B, which we can take to be consecutive in the skew-decomposition of $\pi = \cdots \ominus A \ominus B \ominus \cdots$. Let ab be any occurrence of 12 in A. Let u be the highest point in B and v be the leftmost point in B. Then $abuv$ is an occurrence of the mesh pattern. It is clear that an occurrence $abuv$ of the mesh pattern will correspond to two vertices ab, uv in the occurrence graph, and the shadings ensure that there is no path between them.

The enumeration follows from the fact that these permutations must have no, or exactly one, skew-component of size greater than 1. The first case is counted by $1/(1-x)$, while the second case is counted by $(F(x) - x)/(1-x)^2$. $\qquad\square$

Note that our software suggests a very large nonclassical basis for the permutations with a tree as an occurrence graph. We omit displaying this basis here. However, since a graph is a tree if and only if it is a nonempty connected forest we obtain:

Corollary 6.3. *Let c be the property of being a tree. Then*

$$\mathcal{G}_{12,c} = \mathrm{Av}(123, 1432, 2143, 3214, m) \setminus \mathrm{Av}(12),$$

where m is the mesh pattern in Figure 7.

Proof. This follows from Theorems 5.10 and 6.2. We must remove the decreasing permutations since they have empty occurrence graphs. $\qquad\square$

We end with proving the enumeration for the permutations in the corollary above. The proof is a rather tedious, but simple, induction proof.

Theorem 6.4. *The number of permutations of length n in $\mathcal{G}_{12,\mathrm{tree}}$ is $(n-1)^2$.*

7. Future work

We expect the conjectures in lines 2 and 3 in Tables 1 and 2 to follow from an analysis of the cycle structure of occurrence graphs with respect to the patterns 123 and 132, similar to what we did in Proposition 5.9 for the pattern 12.

Other natural hereditary graph properties to consider would be k-colorable graphs, for $k > 2$, as these are supersets of bipartite graphs. Also planar graphs, which lie between forests and 4-colorable graphs.

It might also be interesting to consider the intersection $\bigcap_{p \in M} \mathcal{G}_{p,c}$ where M is some set of patterns, perhaps all.

We would like to note that Smith (personal communication, 2016) independently defined occurrence graphs and used them to prove a result on the shellability of a large class of intervals of permutations.

Appendix: Proof of Theorem 6.4

We start by introducing a new notation.

Definition A.1. Let $\pi \in \mathfrak{S}_n$ and k be an integer such that $1 \leq k \leq n+1$. The *k-prefix of π* is the permutation $\pi' \in \mathfrak{S}_{n+1}$ defined by $\pi'(1) = k$ and

$$\pi'(i+1) = \begin{cases} \pi(i) & \text{if } \pi(i) < k, \\ \pi(i)+1 & \text{if } \pi(i) \geq k \end{cases}$$

for $i = 1, \ldots, n$. We denote π' by $k \succ \pi$. In a similar way we define the *k-postfix of π* as the permutation $\pi \prec k$ in $\mathfrak{S}_{n+1}$.

Example A.2. Let $\pi = 42135$ and $k = 2$. Visually, if we draw the grid representation of π, we put the new number k to the left on the x-axis and raise all the numbers $\geq k$ on the y-axis by 1. Thus, $2 \succ 42135 = 253146$, as in Figure 9.

We note that for every permutation $\pi' \in \mathfrak{S}_{n+1}$ there is one and only one pair (k, π) such that $\pi' = k \succ \pi$. We let $k = \pi'(1)$ and $\pi = \text{st}(\pi'(2) \cdots \pi'(n+1))$.

Proof of Theorem 6.4. We start by considering three base cases.

For $n = 1$ the occurrence graph is the empty graph. For $n = 2$ we get two occurrence graphs: $G_{12}(12)$ is a graph with a single vertex and $G_{12}(21)$ is the empty graph. For $n = 3$ we have $3! = 6$ different permutations π. Of those we calculate that 132, 213, 231 and 312 result in connected occurrence graphs on one or two vertices but $G_{12}(123)$ is a triangle and $G_{12}(321)$ is the empty graph.

We have thus shown that the claimed enumeration is true for $n = 1, 2, 3$.

For the inductive step we assume $n \geq 4$ and let π be a permutation of length n. We look at four different cases of k to construct $\pi' = k \succ \pi$. We let x, y and z be the indices of $n-1$, n and $n+1$ in π' respectively.
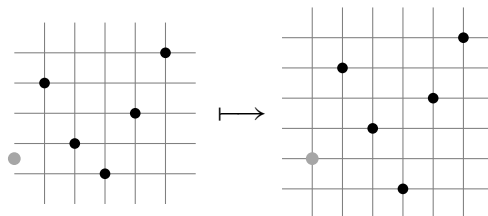
Figure 9. The 2-prefix of 42135 is 253146.
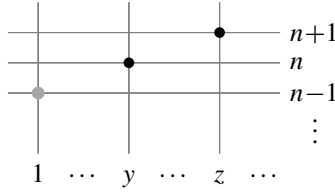
Figure 10. $k = n - 1$ and $y < z$.

(I) $\underline{k \leq n - 2}$: The index sets $\{1, x\}$, $\{1, y\}$ and $\{1, z\}$ of 12 in π' all share exactly one common element and thus form a triangle in $G_{12}(\pi')$. Therefore there are no permutations π such that the occurrence graph $G_{12}(\pi')$ is a tree.

(II) $\underline{k = n - 1}$: Let $T(n + 1)$ denote the number of permutations π' of length $n + 1$ with $\pi'(1) = n - 1$ such that $G_{12}(\pi')$ is a tree. Note that $T(1) = T(2) = 0$, $T(3) = 1$ and $T(4) = 2$. In order to obtain a formula for T we need to look at a few subcases:

(i) If $y < z$ then $\{1, y\}$, $\{1, z\}$ and $\{y, z\}$ form a triangle in $G_{12}(\pi')$; see Figure 10. Independent of the permutation π, the graph $G_{12}(\pi')$ is not a tree.

(ii) Assume $y > z$ and $z \neq 2$, as in Figure 11. Then $\pi'(2) < n - 1$ and $\{1, z\}$, $\{2, z\}$, $\{2, y\}$ and $\{1, y\}$ form a cycle of length 4 in $G_{12}(\pi')$, resulting in it not being a tree.

(iii) Assume $y > z$ and $z = 2$, as in Figure 12. If $y \geq 5$ then the vertices $\{1, y\}$, $\{3, y\}$ and $\{4, y\}$ form a cycle in $G_{12}(\pi')$. If $y = 3$ then $\{1, 2\}$ and $\{1, 3\}$ will be an isolated path component in $G_{12}(\pi')$, making $\pi' = (n - 1)(n + 1)n(n - 2) \cdots 1$ the only permutation such that the occurrence graph $G_{12}(\pi')$ is a tree. If $y = 4$, we need to consider further subcases for the value of $\pi'(3)$.

(a) If $\pi'(3) \leq n - 4$ then $\pi'(3)n$, $\pi'(3)(n - 2)$ and $\pi'(3)(n - 3)$ are all occurrences of 12 in π', with the respective index sets forming a triangle in $G_{12}(\pi')$.

(b) If $\pi'(3) = n - 2$ then $\pi' = (n - 1)(n + 1)(n - 2)n(n - 3) \cdots 1$ is the only permutation resulting in $G_{12}(\pi')$ being a tree.
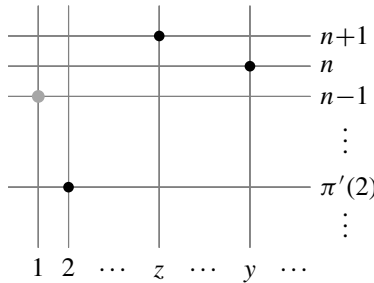
Figure 11. $k = n - 1$, $y > z$ and $z \neq 2$.
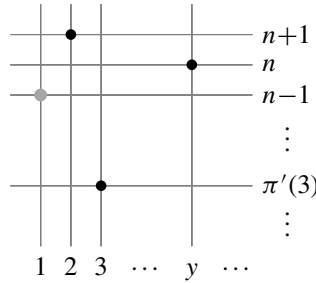
Figure 12. $k = n - 1$, $y > z$ and $z = 2$.

(c) If $\pi'(3) = n - 3$ then we look at Figure 13. The permutation $\sigma = \mathrm{st}(\pi'(3) \cdots \pi'(n+1))$ is just like π' in the case $k = n - 1$ and $z = 2$, only the length of σ is $n - 1$. Because $\{1, 2\}$ is a vertex in $G_{12}(\sigma)$, the occurrence graph of 12 in σ is not the empty graph. Thus it is easy to see that $G_{12}(\pi')$ is a tree if and only if $G_{12}(\sigma)$ is a tree, and according to the aforementioned case there are $T(n-1)$ such permutations σ.

Summing up these possibilities we get a total of $1 + 1 + T(n-1)$ permutations π' making the occurrence graph a tree, i.e., $T(n+1) = 2 + T(n-1)$. Because $T(4) = 2$ and $T(3) = 1$, we deduce that $T(n+1) = n - 1$.

The whole case $k = n - 1$ gives us that there are $n - 1$ permutations π' such that $G_{12}(\pi')$ is a tree.

(III) $\underline{k = n}$: We need to examine three subcases:

(i) If $z \geq 4$ then $\{1, z\}$, $\{2, z\}$, $\{3, z\}$ are all index sets of 12 in π', forming a triangle in $G_{12}(\pi')$.

(ii) If $z = 3$, then $\{1, 3\}$ is an index set of 12 in π making the occurrence graph $G_{12}(\pi)$ nonempty; see Figure 14.

   If $\pi'(2) \leq n - 2$ then $\pi'(2)(n+1)$, $\pi'(2)(n-1)$ and $\pi'(2)(n-2)$ are all occurrences of 12 in π', resulting in $G_{12}(\pi')$ having a triangle. If $\pi'(2) = n - 1$ then $\{1, 3\}$ and $\{2, 3\}$ is an isolated path component in $G_{12}(\pi')$ and $\pi' =$
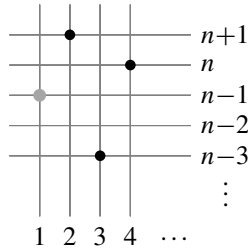
Figure 13. $k = n - 1$, $y = 4$ and $z = 2$.
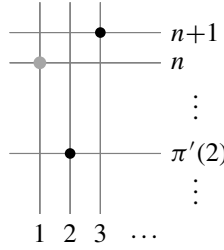
Figure 14. $k = n$ and $z = 3$.

$n(n-1)(n+1)(n-2)\cdots 1$ is the only permutation such that the occurrence graph is a tree. We therefore assume $\pi'(2) = n - 2$; see Figure 15.

Let $\sigma = \mathrm{st}(\pi'(2)\cdots\pi'(n+1))$. Note that the occurrence graphs $G_{12}(\pi')$ and $G_{12}(\sigma)$ are the same except the former has the extra vertex $\{1,2\}$ and an edge connecting it to a graph corresponding to $G_{12}(\sigma)$. Therefore, $G_{12}(\pi')$ is a tree if and only if $G_{12}(\sigma)$ is a tree.

Note that $\sigma(1) = n - 2$ and $\sigma(2) = n$ and therefore σ is like π' in the case $k = n - 1$ and $z = 2$ as in Figure 12, only of length n instead of $n + 1$. By the same reasoning as in that case, the number of permutations σ (and therefore π') such that $G_{12}(\pi')$ is a tree is $T(n) = n - 2$.

(iii) If $z = 2$, then $\{1,2\}$ is an isolated vertex in $G_{12}(\pi')$; see Figure 16. The occurrence graph of 12 in π' is a tree if and only if $G_{12}(\pi)$ is the empty graph, which is true if and only if π is the decreasing permutation. Therefore there is only one permutation $\pi' = n(n+1)(n-1)\cdots 1$ such that $G_{12}(\pi')$ is a tree.

To sum up the case $k = n$ there are $1 + (n-2) + 1 = n$ permutations π' such that $G_{12}(\pi')$ is a tree.

(IV) $\underline{k = n + 1}$: Every occurrence $\pi(i)\pi(j)$ of 12 in π is also an occurrence of 12 in π', but with index set $\{i+1, j+1\}$ instead of $\{i, j\}$. There are no more occurrences of 12 in π' because $\pi'(1) = n + 1 > \pi'(j')$ for every $j' > 1$ so $\pi'(1)\pi'(j')$ is not an occurrence of 12 for any $j' > 1$.
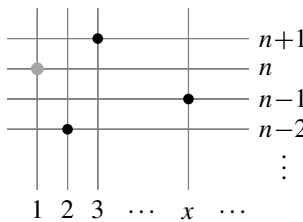
Figure 15. $k = n$, $z = 3$ and $\pi'(2) = n - 2$.
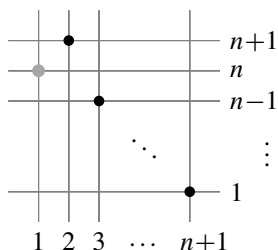
Figure 16. $k = n$ and $z = 2$.

This means that $G_{12}(\pi') \cong G_{12}(\pi)$, so by the induction hypothesis we obtain that there are $(n-1)^2$ permutations π' such that the occurrence graph is a tree for this value of k.

To sum up the four instances there is a total of $0 + (n-1) + n + (n-1)^2 = n^2$ permutations π' such that $G_{12}(\pi')$ is a tree. $\qquad \square$

References

[Brändén and Claesson 2011] P. Brändén and A. Claesson, "Mesh patterns and the expansion of permutation statistics as sums of permutation patterns", *Electron. J. Combin.* **18**:2 (2011), art. id. 5. MR Zbl

[Claesson et al. 2015] A. Claesson, B. E. Tenner, and H. Ulfarsson, "Coincidence among families of mesh patterns", *Australas. J. Combin.* **63** (2015), 88–106. MR Zbl

[Comtet 1974] L. Comtet, *Advanced combinatorics: the art of finite and infinite expansions*, enlarged ed., Reidel, Dordrecht, 1974. MR Zbl

[Erdős and Szekeres 1935] P. Erdős and G. Szekeres, "A combinatorial problem in geometry", *Compositio Math.* **2** (1935), 463–470. MR Zbl

[Kitaev 2011] S. Kitaev, *Patterns in permutations and words*, Springer, 2011. MR Zbl

[Knuth 1968] D. E. Knuth, *The art of computer programming, I: Fundamental algorithms*, Addison-Wesley, Boston, 1968. Zbl

[MacMahon 1915] P. A. MacMahon, *Combinatory analysis, I*, Cambridge Univ. Press, 1915. Zbl

[Magnusson and Ulfarsson 2012] H. Magnusson and H. Ulfarsson, "Algorithms for discovering and proving theorems about permutation patterns", preprint, 2012. arXiv

bjk17@hi.is   *Department of Mathematics, University of Iceland, Reykjavik, Iceland*

henningu@ru.is   *School of Computer Science, Reykjavik University, Reykjavik, Iceland*

Truncated path algebras and
Betti numbers of polynomial growth

Ryan Coopergard and Marju Purin

(Communicated by Kenneth S. Berenhaut)

We investigate a class of truncated path algebras in which the Betti numbers of a simple module satisfy a polynomial of arbitrarily large degree. We produce truncated path algebras where the i-th Betti number of a simple module S is $\beta_i(S) = i^k$ for $2 \leq k \leq 4$ and provide a result of the existence of algebras where $\beta_i(S)$ is a polynomial of degree 4 or less with nonnegative integer coefficients. In particular, we prove that this class of truncated path algebras produces Betti numbers corresponding to any polynomial in a certain family.

1. Introduction

We consider finite-dimensional algebras Λ over an algebraically closed field with $\mathrm{rad}^2 \Lambda = 0$, where $\mathrm{rad}\,\Lambda$ denotes the Jacobson radical of the algebra. We work with these algebras by representing them as quotients of path algebras. The motivation behind investigating these algebras lies in the universality of path algebras. Namely, any finite-dimensional algebra over an algebraically closed field is a quotient of a path algebra. We use quivers (directed graphs) to write down these algebras and provide numerous examples along the way.

We study modules by means of their projective resolutions. Betti numbers are of particular interest in examining the projective resolutions of modules as they provide a method of describing the growth of resolutions. Such growth was examined in the groundbreaking paper [Tate 1957] in the setting of commutative rings and in [Alperin and Evens 1981] for group algebras. Since then the growth of resolutions has been shown to be related to many fundamental properties of an algebra such as, for example, the representation type of an algebra [Diveris and Purin 2014; Erdmann et al. 2004] or codimension of a commutative ring [Avramov 1998; Avramov and Buchweitz 2000; Avramov et al. 1997; Eisenbud 1980].

A fundamental question that is driving our work in this paper is to determine which polynomials are eventually realizable as sequences of Betti numbers. To this

end we introduce a particular class of path algebras, namely those given by quivers
of the form
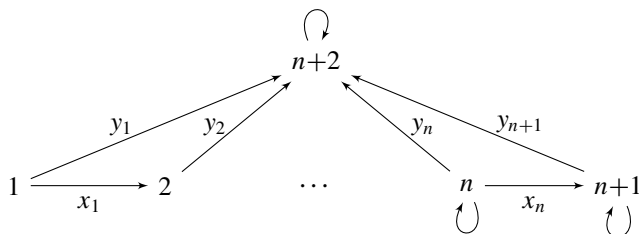
where the x_i and y_{n+1} are positive integers and y_i, $1 \leq i \leq n$, are nonnegative
integers that represent the number of arrows between vertices. To clarify, the general
$n = 0$ case is of the form
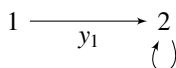
and the general $n = 1$ case is of the form
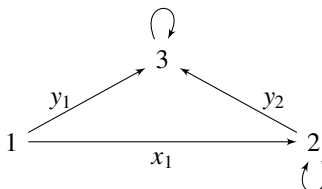
From here on, we refer to these algebras as *pyramidal algebras*. Given a pyramidal
algebra Λ, we refer to the quotient algebra $\Lambda / \operatorname{rad}^m \Lambda$ as an *m-pyramidal algebra*.
For the majority of the paper, we consider only 2-pyramidal algebras and briefly
discuss the more general version at the end.

A key result in the paper is to show that 2-pyramidal algebras have a simple mod-
ule whose Betti numbers have polynomial growth of arbitrarily high degree. More
precisely, the Betti numbers over algebras of pyramidal form satisfy $\beta_i(S_1) = p_n(i)$,
where S_1 is the simple module at vertex 1 and p_n is a polynomial of degree n. In addi-
tion to proving this result, we provide examples of algebras with particularly interest-
ing behavior of Betti numbers. We end with an application of our work to a question
about the existence of algebras in which $\beta_i(S_1)$ is a polynomial of a specific form.

2. Preliminaries

A quiver is a set of vertices and arrows (an oriented graph). In this paper we work
with finite quivers, that is, quivers with finitely many vertices and arrows. Further-
more, we assume that the quiver is connected, which means that the underlying
graph is connected. We concatenate arrows to form paths in the quiver. In addition,
there is a trivial path at each vertex, which we denote by e_i for vertex i.

A path algebra over a field k is the k-vector space that has as its basis the set of all paths. The multiplication of paths is given by concatenation of compatible paths. For incompatible paths the product is zero. With this operation the set of paths has a natural structure as a k-algebra. Furthermore, the Jacobson radical of the algebra is simply the ideal generated by the set of all arrows.

Example 2.1. We illustrate the above notions with the 2-pyramidal algebra where $y_1 = 1$:

$$1 \xrightarrow{\quad \alpha \quad} 2 \circlearrowright \beta$$

The quiver above has two vertices and two arrows. As for paths, there are four nonzero paths: the two trivial paths $\{e_1, e_2\}$ and two arrows $\{\alpha, \beta\}$. Some examples of multiplication are: $e_1 \cdot \alpha = \alpha$, $\alpha \cdot e_2 = \alpha$, $\alpha \cdot \beta = \alpha\beta = 0$ (as the path lies in $\mathrm{rad}^2 \Lambda$), and $\beta \cdot \alpha = 0$ (as the arrows are incompatible).

In this paper we work with finitely generated right modules over finite-dimensional algebras. Every such module has a projective cover and consequently a minimal projective resolution over the algebra. For a path algebra, the number of indecomposable nonisomorphic projective modules corresponds to the number of vertices in the quiver of the algebra. In particular, there are only finitely many such projective modules, while there can be infinitely many indecomposable nonisomorphic modules over such algebras. Therefore projective modules, by means of resolutions, provide a method of studying any module over a finite-dimensional algebra.

We measure the complexity of an algebra by measuring the complexity of the projective resolutions of the modules over the algebra. We do this by examining the growth of the Betti sequence of the resolutions. For $m \geq 0$, the m-th term, the m-th Betti number, is the number of indecomposable projective modules at the m-th step of the resolution. Thus, faster growth of a Betti sequence corresponds to a higher-complexity module.

It suffices to examine the resolutions of the simple modules as the fastest growth rate is always realized by a simple module. The goal in this paper is precisely this — to examine the resolutions of simple modules.

Throughout the paper we use the following notation. We denote by S_n the simple module at vertex n. For $i \geq 0$, the i-th term in a projective resolution of a module M is denoted by $P_i(M)$ and the i-th Betti number is $\beta_i(M)$.

We also make use of dimension vectors of modules. The dimension vector of a module M represents the element $[M]$ in the Grothendieck group $K_0(\Lambda)$ corresponding to M, where $K_0(\Lambda)$ is the free abelian group on a set of isomorphism classes of the simple Λ-modules. As such, dimension vectors record the multiplicity of each composition factor in the composition series of the module. For ease of

notation, k_i copies of S_i in the composition series of a module M are denoted by $1^{k_1} 2^{k_2} \cdots t^{k_t}$. In particular, we are not tracking the radical layers in which the composition factors occur.

Example 2.2. The 2-pyramidal algebra in Example 2.1 has two nonisomorphic simple modules, one at each vertex, denoted by S_1 and S_2. The projective covers of the simple modules can be obtained by recording the maximal path starting at the corresponding vertex, keeping in mind that in a 2-pyramidal algebra the composite of any two arrows vanishes. Thus, the projective cover of the simple module $S_1 = 1$ is $P_0(S_1) = \frac{1}{2} = 1\,2$, the projective cover of $S_2 = 2$ is $P_0(S_2) = \frac{2}{2} = 2^2$.

Note that the zeroth Betti number, corresponding to the zeroth step in the projective resolution, is 1. This will always be the case, and for this reason we will ignore the zeroth Betti number and consider only β_k with $k \geq 1$ for the remainder of this paper. The first syzygy, denoted by $\Omega^1(S_1)$, in the projective resolution of S_1 is the kernel of the epimorphism $P_0(S_1) \to S_1$. It has dimension vector $\Omega^1(S_1) = 2$. A projective resolution of S_1 is obtained by iterating the process and finding a projective cover, denoted by $P_1(S_1)$, for the syzygy $\Omega^1(S_1) = 2$. We obtain the resolution

$$\cdots \frac{2}{2} \to \frac{2}{2} \to \frac{2}{2} \to \frac{1}{2} \to S_1 = 1.$$

In other words, we have $P_i(S_1) = \frac{2}{2}$ and syzygies $\Omega^i(S_1) = 2$ for $i > 0$. The Betti sequence is the constant sequence $\beta_i(S_1) = 1$ for $i \geq 0$.

For more background on modules over path algebras we refer the reader to [Auslander et al. 1995; Assem et al. 2006].

We make frequent use of difference tables of polynomials. Given a polynomial $p(n)$ of degree n, the *difference table* of $p(n)$ is a table of rows and columns, $D = \{d_{i,j}\}$, $i \geq 1$, $j \geq 0$ such that $\{d_{i,0}\} = p(i)$ and the other entries are defined recursively as $d_{i,j} = d_{i+1,j-1} - d_{i,j-1}$. That is, the j-th column in the difference table of $p(n)$ is the difference between the elements in the $(j-1)$-th column. We then refer to the j-th column as the j-th difference of $p(n)$.

Example 2.3. The difference table for the polynomial $p(n) = n^2$ is

1	3	2	0	$\cdots$
4	5	2	0	$\cdots$
9	7	2	0	$\cdots$
16	9	2	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Note that each column produces a sequence that is polynomial of degree one less than the previous column, until we reach a column of zeros.

The difference tables of polynomials eventually reach a column of zeros. Thus, we will refer to the tables only up to, but not including, the first column of zeros.

3. Pyramidal algebras

In this section we examine the behaviour of projective resolutions over pyramidal algebras. We begin with a key observation that describes the syzygies of a projective resolution of a simple module.

Before proceeding with the results, we remark that for 2-pyramidal algebras all syzygies are semisimple. This is because these algebras have radical squared zero. Hence it is sufficient to work with dimension vectors when calculating the syzygies in a resolution.

Lemma 3.1. *In a* 2-*pyramidal algebra of the form,*
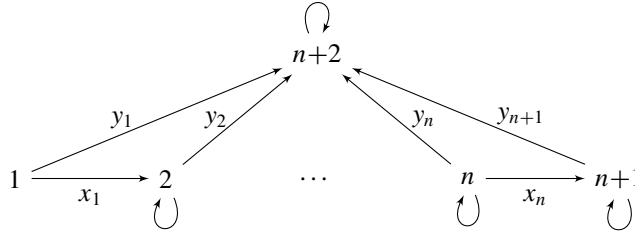
the multiplicity of S_k as a direct summand in the syzygy $\Omega^i(S_1)$, $i \geq k-1$, is

$$\binom{i-1}{k-2} x_1 x_2 \cdots x_{k-1}$$

if $2 \leq k \leq n+1$ and

$$y_1 + \sum_{j=1}^{n} \binom{i-1}{j} x_1 x_2 \cdots x_j y_{j+1}$$

if $k = n+2$ and $i \geq 1$.

In the case where $k \neq n+2$ and $i \leq k-2$, or $k = n+2$ and $i \leq 1$, the multiplicity of S_k in $\Omega^i(S_1)$ is zero.

Proof. Note that S_k appears as a summand of $\Omega^i(S_1)$ if and only if there is a walk of length i from vertex 1 to vertex k in the underlying quiver. The final statement in the lemma is an immediate corollary of this fact.

We will prove the first case where $2 \leq k \leq n+1$ by double induction on the statement "the multiplicity of S_k as a direct summand in the syzygy $\Omega^i(S_1)$, $i \geq k-1$, is

$$\binom{i-1}{k-2} x_1 x_2 \cdots x_{k-1}.\text{"}$$

We will induct on i and k, in that order. When inducting on i, the base case is $k = 2$, $i = 1$, as this is the first syzygy in which S_2 appears. We then proceed by varying i

and fixing $k = 2$ to complete the induction on i. When inducting on k, we must start with the base case $i = k - 1$, as this is the smallest value of i in which S_k appears as a summand of $\Omega^i(S_1)$. Finally, we induct on k given an arbitrary fixed $i \geq k - 1$.

For $k = 2$ and $i \geq 1$ arbitrary, we see that the multiplicity of S_2 in $\Omega^i(S_1)$ is x_1 always. This is equal to $\binom{i-1}{0} x_1$, so this concludes the first part of the induction.

Assume the statement holds for $i = k - 1$ and consider the multiplicity of S_k as a direct summand in $\Omega^{k-1}(S_1)$. Because there is no S_k in $\Omega^{k-2}(S_1)$, only the multiplicity of S_{k-1} in $\Omega^{k-1}(S_1)$ contributes to the multiplicity of S_k in $\Omega^{k-1}(S_1)$. By the induction hypothesis, there are

$$\binom{k-3}{k-3} x_1 x_2 \cdots x_{k-3} x_{k-2} = x_1 x_2 \cdots x_{k-3} x_{k-2}$$

copies of S_{k-1} in the $(k-2)$-th syzygy. Thus the multiplicity of S_k in the $(k-1)$-th syzygy is

$$x_1 x_2 \cdots x_{k-2} x_{k-1} = \binom{k-2}{k-2} x_1 x_2 \cdots x_{k-2} x_{k-1}.$$

Now assume the statement holds up to $k - 1$ and $i - 1$. By induction, the multiplicity of S_{k-1} in the $(i-1)$-th syzygy is given by

$$\binom{i-2}{k-3} x_1 x_2 \cdots x_{k-3} x_{k-2}.$$

Similarly the multiplicity of S_k in the $(i-1)$-th syzygy is

$$\binom{i-2}{k-2} x_1 x_2 \cdots x_{k-1}.$$

Therefore, the multiplicity of S_k in $\Omega^i(S_1)$ is

$$x_{k-1} \binom{i-2}{k-3} x_1 x_2 \cdots x_{k-2} + \binom{i-2}{k-2} x_1 x_2 \cdots x_{k-1} = \left(\binom{i-2}{k-3} + \binom{i-2}{k-2} \right) x_1 x_2 \cdots x_{k-1}$$

$$= \binom{i-1}{k-2} x_1 x_2 \cdots x_{k-1},$$

and the induction is complete. A similar argument can be made for the multiplicity of S_{n+2} as a direct summand of the i-th syzygy. $\square$

Example 3.2. To see an example of this lemma, consider the 2-pyramidal algebra
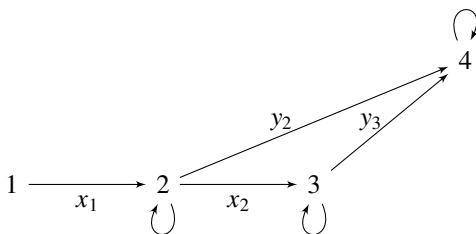
with $y_1 = 0$ and $x_1 = x_2 = y_2 = y_3 = 1$, i.e., there is one arrow from vertex 2 to 3, one from vertex 2 to 4, and one from vertex 3 to 4.

The projective resolution of S_1 is

$$\cdots P_2 \oplus P_3 \oplus P_4 \to P_2 \to P_1 \to S_1,$$

with syzygies

$$\Omega^1(S_1) = 2, \quad \Omega^2(S_1) = 234, \quad \Omega^3(S_1) = 23^2 4^3, \quad \Omega^4(S_1) = 23^3 4^6,$$

etc. The multiplicity of S_3 in the dimension vector of $\Omega^4(S_1)$ is 3, while the multiplicity of S_4 in $\Omega^4(S_1)$ is 6.

Using our formula to calculate the multiplicity of S_3 and S_4 in the dimension vector $\Omega^4(S_1)$ gives the following.

First, for $k = 3$ and $i = 4$ we obtain the multiplicity of S_3 as

$$\binom{3}{1} \cdot 1 \cdot 1 = 3.$$

Similarly for $k = 4$ and $i = 4$, we get the multiplicity of S_4 as

$$\sum_{j=1}^{2} \binom{3}{j} \cdot 1 = 3 + 3 = 6.$$

Note that we interpret $x_j = 0$ for $j \geq 3$ because their corresponding edges in the quiver are not present, so further sums do not appear.

Theorem 3.3. *Every 2-pyramidal algebra with $n + 2$ vertices in the underlying quiver has Betti numbers*

$$\beta_i(S_1) = \begin{cases} 1 & \text{for } i = 0, \\ p_n(i) & \text{for } i \geq 1, \end{cases}$$

where p_n is a polynomial of degree n.

Proof. We proceed by induction on n. If $n = 0$, then we have an algebra of the form

$$1 \xrightarrow{\ y_1\ } 2 \,\circlearrowleft$$

Because $\Omega^i(S_1) = 2^{y_1}$ for all i, it follows that $\beta_i(S_1) = y_1$ for all i, so $\beta_i(S_1)$ is constant, and thus is a polynomial of degree 0.

Suppose the statement holds for all values less than n, and consider an algebra of this form with $n + 2$ vertices. The Betti numbers are calculated by adding the multiplicities of the various S_k together. These multiplicities were calculated in Lemma 3.1, so we see that the i-th Betti number is given by

$$\sum_{j=0}^{n-1} \binom{i-1}{j} x_1 \cdots x_{j+1} + y_1 + \sum_{j=1}^{n} \binom{i-1}{j} x_1 x_2 \cdots x_j y_{j+1}.$$

Now taking the $(i+1)$-th Betti number and subtracting the i-th Betti number yields

$$\sum_{j=0}^{n-1}\binom{i}{j}x_1\cdots x_{j+1}+y_1+\sum_{j=1}^{n}\binom{i}{j}x_1x_2\cdots x_jy_{j+1}$$

$$-\left(\sum_{j=0}^{n-1}\binom{i-1}{j}x_1\cdots x_{j+1}+y_1+\sum_{j=1}^{n}\binom{i-1}{j}x_1x_2\cdots x_jy_{j+1}\right)$$

$$=\sum_{j=0}^{n-1}\left(\binom{i}{j}-\binom{i-1}{j}\right)x_1\cdots x_{j+1}+\sum_{j=1}^{n}\left(\binom{i}{j}-\binom{i-1}{j}\right)x_1x_2\cdots x_jy_{j+1}$$

$$=\sum_{j=1}^{n-1}\binom{i-1}{j-1}x_1\cdots x_{j+1}+\sum_{j=1}^{n}\binom{i-1}{j-1}x_1x_2\cdots x_jy_{j+1}.$$

Observe that this is the i-th Betti number of the following algebra:
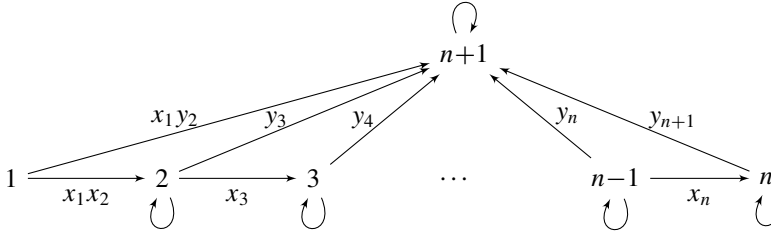
By the induction hypotheses, this 2-pyramidal algebra satisfies $\beta_i(S_1)=p_{n-1}(i)$, where p_{n-1} is a polynomial of degree $n-1$. Thus we see that the difference between the terms of the original algebra's Betti numbers is a polynomial of degree $n-1$, so the Betti numbers follow a polynomial of degree n, as desired. $\square$

It is interesting to mention an alternative approach to the above result, as was suggested by one of the referees. Namely, we may also analyze the Betti sequence by means of the action of the syzygy operator Ω. Because the syzygies of a module over a radical square zero algebra are semisimple, Ω acts as an endomorphism on the Grothendieck group $K_0(\Lambda)$. The action of Ω on S_i is evidently the dimension vector of $\Omega(S_i)$. Considering these vectors over all $n+2$ simple modules, the action of Ω is given by the matrix

$$\Omega=\begin{bmatrix}0 & 0 & 0 & \cdots & 0\\ x_1 & 1 & 0 & \cdots & 0\\ 0 & x_2 & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ y_1 & y_2 & y_3 & \cdots & 1\end{bmatrix}.$$

Thus, $\Omega^m(S_1)$ is the first column of the m-th power of this matrix. Moreover, this matrix is the transpose of the adjacency matrix of the quiver, so the first column of Ω^m gives the number of paths starting at vertex 1 that have length m.

While we have only considered the Betti numbers for the projective resolution of S_1, we may also consider them for projective resolutions of any other simple module, say S_j. In this case, by restricting our quiver to the vertices $j, j+1, \ldots, n+2$, we get another algebra. The Betti numbers of the projective resolution of S_j are evidently the same as that of the 2-pyramidal algebra below:
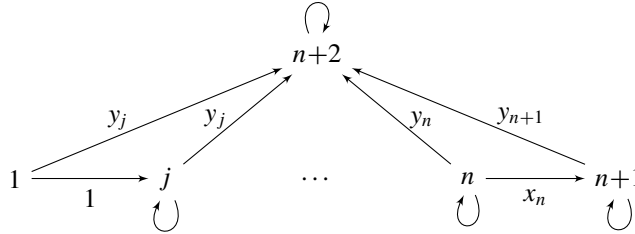
By previous work, we see that the Betti numbers of S_j agree with a polynomial of degree $n + 2 - j$.

Theorem 3.3 is quite useful for the theorems in this paper due to the following corollary.

Corollary 3.4. *Let Λ be a 2-pyramidal algebra as in Lemma 3.1. If the first $n + 1$ Betti numbers are known to fit a polynomial p of degree n, then $\beta_i(S_1) = p(i)$.*

Proof. By Theorem 3.3, we know that $\beta_i(S_1) = p_n(i)$ for some polynomial p_n of degree n. It is well known that given $n + 1$ pairs of points $\{(x_j, y_j)\}_{j=1}^{n+1}$, there is a unique polynomial p of degree n such that $p(x_j) = y_j$ for $1 \le j \le n + 1$. Because $\beta_i(S_1) = p(i)$ for $1 \le i \le n + 1$, it follows that $\beta_i(S_1) = p(i)$ for all $i \ge 1$. $\square$

This theorem and its corollary will help us find algebras with Betti numbers of growth given by $\beta_i(S_1) = i^2$, $\beta_i(S_1) = i^3$ and $\beta_i(S_1) = i^4$. From this, we show that given any polynomial $p(i)$ of degree 4 or less with nonnegative integer coefficients, there exists an algebra such that $\beta_i(S_1) = p(i)$.

Lemma 3.5. *In the following 2-pyramidal algebra, $\beta_i(S_1) = i^2$ for $i \ge 1$:*
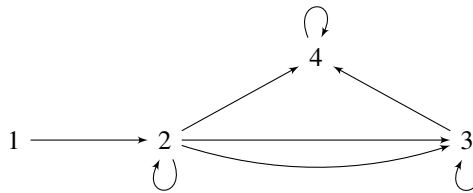
Proof. By Corollary 3.4, we need only show that the first three terms agree with $\beta_i(S_1) = i^2$. Indeed, we can calculate these quite easily:

$$\Omega^1(S_1) = 2, \quad \Omega^2(S_1) = 2\,3^2\,4, \quad \Omega^3(S_1) = 2\,3^4\,4^4.$$

Thus $\beta_i(S_1) = i^2$ for $1 \le i \le 3$. Therefore, $\beta_i(S_1) = i^2$ for all $i \ge 1$. $\square$

Lemma 3.6. *In the following 2-pyramidal algebra, $\beta_i(S_1) = i^3$ for $i \geq 1$:*
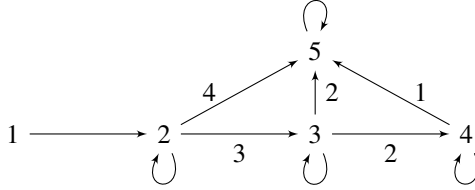
Proof. Again by Corollary 3.4, we need only show that $\beta_i(S_1) = i^3$ for $1 \leq i \leq 4$. We compute the syzygies directly as before. Here we obtain

$$\Omega^1(S_1) = 2, \quad \Omega^2(S_1) = 2\,3^3\,5^4, \quad \Omega^3(S_1) = 2\,3^6\,4^6\,5^{14}, \quad \Omega^4(S_1) = 2\,3^9\,4^{18}\,5^{36}.$$

We see that $\beta_i(S_1) = i^3$ for $1 \leq i \leq 4$. Therefore, $\beta_i(S_1) = i^3$ for all $i \geq 1$.   □

Lemma 3.7. *In the following 2-pyramidal algebra, $\beta_i(S_1) = i^4$ for $i \geq 1$:*
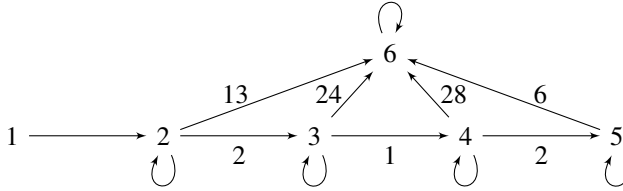
Proof. We find the i-th syzygy of S_1 for $1 \leq i \leq 5$ to show that $\beta_i(S_1) = i^4$ for these values of i. Indeed, the syzygies are as follows:

$$\Omega^1(S_1) = 2, \quad \Omega^2(S_1) = 2\,3^2\,6^{13}, \quad \Omega^3(S_1) = 2\,3^4\,4^2\,6^74,$$
$$\Omega^4(S_1) = 2\,3^6\,4^6\,5^4\,6^{239}, \quad \Omega^5(S_1) = 2\,3^8\,4^{12}\,5^{16}\,6^{588}.$$

By examining the size of these syzygies, we find

$$\beta_1(S_1) = 1, \quad \beta_2(S_1) = 1+2+13 = 16 = 2^4, \quad \beta_3(S_1) = 1+4+2+74 = 81 = 3^4,$$
$$\beta_4(S_1) = 1+6+6+4+239 = 256 = 4^4, \quad \beta_5(S_1) = 1+8+12+16+588 = 625 = 5^4.$$

By Corollary 3.4, it follows that $\beta_i(S_1) = i^4$ for all $i \geq 1$.   □

The next lemma will give us a method of constructing algebras with specific Betti numbers for a simple module.

Lemma 3.8. *Let Λ_1 and Λ_2 be truncated path algebras such that the projective resolution of the simple module at vertex k in Λ_1 follows $\beta_i(S_k) = f(i)$ and the projective resolution of the simple module at vertex m in Λ_2 follows $\beta_i(S_m) = g(i)$ for some functions f and g. Then there exists an algebra with Betti numbers given by $\beta_i(S) = f(i) + g(i)$ for some simple module S and all $i \geq 1$.*

Proof. Begin with the algebras Λ_1 and Λ_2 with underlying quivers Γ_1 and Γ_2. Let R_1 and R_2 be the set of relations in Λ_1 and Λ_2 respectively. Create a new algebra, Λ_3, whose underlying quiver, Γ_3, is obtained by taking the disjoint union of Γ_1 and Γ_2 and adding a new vertex $\tilde{1}$. Additionally, for each arrow $k \xrightarrow{\alpha} n$ in Λ_1, there is an arrow $\tilde{1} \xrightarrow{\tilde{\alpha}} n$ in Γ_3, and for each arrow $m \xrightarrow{\gamma} l$ in Λ_2, there is an arrow $\tilde{1} \xrightarrow{\tilde{\gamma}} l$ in Γ_3. The set of relations of Λ_3, denoted by R_3, is defined as

$$R_3 := R_1 \cup R_2 \cup \{\tilde{\alpha} w_1 \mid \alpha w_1 \in R_1\} \cup \{\tilde{\gamma} w_2 \mid \gamma w_2 \in R_2\},$$

where w_1 and w_2 could be paths of any length. Note that the elements in the last two sets of this union are nonzero because the target of $\tilde{\alpha}$ is the same as that of α, and the target of $\tilde{\gamma}$ is the same as that of γ. The addition of these relations ensures, for example, that if Λ_1 and Λ_2 are radical square zero algebras, then Λ_3 is as well.

By construction, there are bijections

$$\{\text{vertices in } \Gamma_1\} \cup \{\text{vertices in } \Gamma_2\} \Longleftrightarrow \{\text{vertices in } \Gamma_3\} \setminus \{\tilde{1}\},$$

$$\{\text{paths in } \Gamma_1\} \cup \{\text{paths in } \Gamma_2\} \Longleftrightarrow \{\text{paths in } \Gamma_3 \text{ not involving } \tilde{1}\},$$

both induced by inclusion of quivers. Moreover, the bijection of paths is compatible with the bijection of vertices. This, along with the choice of relations in Λ_3, gives a bijection

$$\{\text{projective } \Lambda_1 - \text{modules}\} \cup \{\text{projective } \Lambda_2 - \text{modules}\}$$
$$\Longleftrightarrow \{\text{projective } \Lambda_3 - \text{modules}\} \setminus \{P_{\tilde{1}}\},$$

where $P_{\tilde{1}}$ is the indecomposable projective Λ_3-module at vertex $\tilde{1}$. This correspondence takes radical layers to radical layers bijectively in a manner compatible with the first two bijections. Let

$$\cdots \to Q_1 \to Q_0 \to S_k \to 0,$$
$$\cdots \to R_1 \to R_0 \to S_m \to 0,$$
$$\cdots \to F_1 \to F_0 \to S_{\tilde{1}} \to 0$$

be minimal projective resolutions of S_k, S_m, and $S_{\tilde{1}}$ respectively as Λ_3-modules. We will now show that for $i \geq 1$, we have $F_i \cong Q_i \oplus R_i$ and $\Omega^i(S_{\tilde{1}}) \cong \Omega^i(S_k) \oplus \Omega^i(S_m)$. Note that the bijections above imply that the minimal projective resolutions of S_k and S_m in Λ_3 correspond to those in Λ_1 and Λ_2, so proving this will yield the lemma.

We proceed by induction on i. We compute $\text{rad}(F_0) = \text{rad}(P_{\tilde{1}}) = \Omega^1(S_{\tilde{1}})$. The simple modules in the k-th radical layer of $P_{\tilde{1}}$ correspond to the vertices at the end of paths of length k from $\tilde{1}$ which do not lie in R_3. By the construction of Λ_3, this is precisely the union of the simple modules in the k-th radical layer of P_k and P_m. Also by construction, we in fact get $\text{rad}(P_{\tilde{1}}) \cong \text{rad}(P_k) \oplus \text{rad}(P_m)$, and so

$\Omega^1(S_{\bar{1}}) \cong \Omega^1(S_k) \oplus \Omega^1(S_m)$. Moreover, the projective cover of this syzygy is the direct sum of the covers of its summands, so $F_1 \cong Q_1 \oplus R_1$. Note that F_i does not have $P_{\bar{1}}$ as a summand for any $i > 0$.

Suppose that $F_i \cong Q_i \oplus R_i$ and $\Omega^i(S_{\bar{1}}) \cong \Omega^i(S_k) \oplus \Omega^i(S_m)$ for $i-1$, $i > 1$. The hypothesis implies that at the $(i-1)$-th step of the projective resolution for $S_{\bar{1}}$, we have a projective cover $Q_{i-1} \oplus R_{i-1} \to \Omega^{i-1}(S_k) \oplus \Omega^{i-1}(S_m)$. By the bijection of projective modules and the fact that the radical layers are preserved under this bijection, we get

$$\ker[Q_{i-1} \oplus R_{i-1} \to \Omega^{i-1}(S_k) \oplus \Omega^{i-1}(S_m)] \cong \Omega^i(S_k) \oplus \Omega^i(S_m),$$

so $\Omega^i(S_{\bar{1}}) \cong \Omega^i(S_k) \oplus \Omega^i(S_m)$. From this it also follows that $F_i \cong Q_i \oplus R_i$, and the induction is complete. Thus $\beta_i(S_{\bar{1}}) = f(i) + g(i)$ for all $i \geq 1$. $\qquad\square$

We apply this lemma to 2-pyramidal algebras to construct Betti sequences that realize desired polynomials.

Example 3.9. Let $p(i) = ai^4 + bi^3 + ci^2 + di + e$ for some nonnegative integers a, b, c, d, e. Then there exists an algebra Λ, where $\beta_i(S) = p(i)$ for a simple module S.

Proof. Begin by choosing algebras Λ_4, Λ_3, Λ_2, Λ_1 and Λ_0 and simple modules S_4, S_3, S_2, S_1, and S_0 satisfying

$$\beta_i^{\Lambda_4}(S_4) = i^4, \quad \beta_i^{\Lambda_3}(S_3) = i^3, \quad \beta_i^{\Lambda_2}(S_2) = i^2, \quad \beta_i^{\Lambda_1}(S_1) = i, \quad \beta_i^{\Lambda_0}(S_0) = 1,$$

respectively. Next, take a, b, c, d, and e copies of the algebras Λ_4, Λ_3, Λ_2, Λ_1 and Λ_0, respectively, and apply Lemma 3.8 to these algebras to obtain a new algebra Λ with

$$\beta_i(S) = ai^4 + bi^3 + ci^2 + di + e$$

for a simple Λ-module S. $\qquad\square$

We will see in the following section that these polynomials can be realized as the Betti numbers of some 2-pyramidal algebra.

4. Characterizations

In this section we characterize Betti numbers over 2-pyramidal algebras. We start with some general statements and proceed to provide a characterization of the polynomials that give the growth of Betti sequences over these algebras.

Lemma 4.1. *Let p be a polynomial such that $p(1) \in \mathbb{Z}^+$, and let p' be the polynomial generating the first differences in the difference table of p. Then there exists a 2-pyramidal algebra in which $\beta_i(S_1) = p(i)$ if and only if there exists a 2-pyramidal algebra such that $\beta_i(S_1) = p'(i)$.*

Proof. The forward direction of this proof is made trivial by a fact in the proof of Theorem 3.3. In this proof, we saw that the n-th element of the first difference of the Betti numbers are the Betti numbers of S_1 over the algebra

$$
1 \xrightarrow[x_1 x_2]{x_1 y_2} 2 \quad \cdots \quad n-1 \xrightarrow{x_n} n \quad \xrightarrow[y_{n+1}]{y_n} n+1 \quad \xrightarrow{y_3}
$$
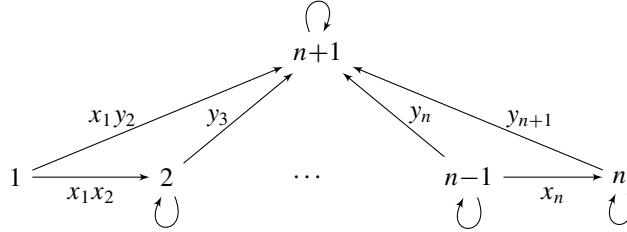
This concludes the first part of the proof.

For the reverse direction, let $p(1) = k \in \mathbb{Z}^+$ and let p' correspond to the Betti numbers of

$$
1 \xrightarrow[x_2]{y_2} 2 \quad \cdots \quad n-1 \xrightarrow{x_n} n \quad \xrightarrow[y_{n+1}]{y_n} n+1 \quad \xrightarrow{y_3} \tag{4-1}
$$
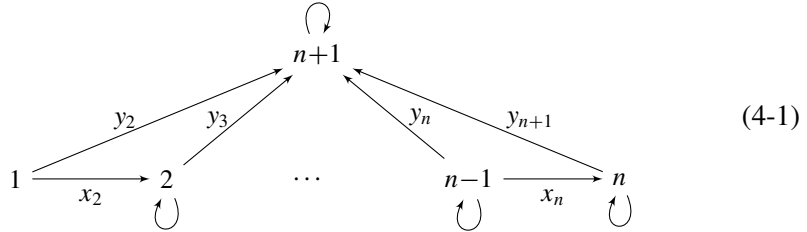
We now consider the algebra

$$
1 \xrightarrow{k-1} 2 \xrightarrow[x_2]{y_2} 3 \quad \cdots \quad n \xrightarrow{x_n} n+1 \quad \xrightarrow[y_{n+1}]{y_n} n+2 \quad \xrightarrow{y_3}
$$
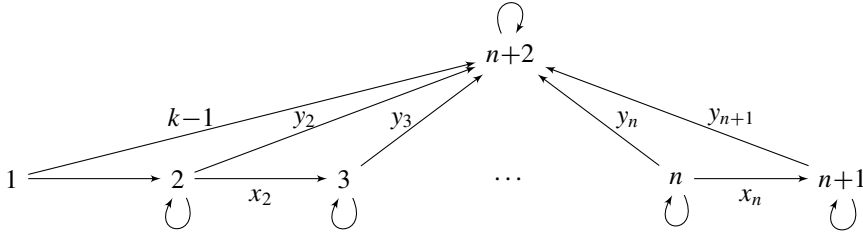
Now this algebra has the property that $\beta_1(S_1) = k$ and the differences are the Betti numbers of (4-1). Because the Betti numbers of (4-1) are given by p', the differences are given by p', as desired. $\qquad\square$

We can now use this result to provide some necessary and sufficient conditions that a polynomial must meet in order to represent the Betti numbers of some 2-pyramidal algebra.

Theorem 4.2. *A polynomial p is such that $\beta_i(S_1) = p(i)$ for some 2-pyramidal algebra if and only if the difference table of p consists of only positive integers.*

Proof. We will prove the forward direction by induction on the columns of the difference table of p. Let p be a polynomial of degree n and let Λ be a 2-pyramidal

algebra such that $\beta_i(S_1) = p(i)$. We first show that the zeroth difference, that is p, has all positive entries. Because x_1 is positive and there is an arrow from 2 to itself, it follows that $p(1) = \beta_1(S_1) \geq x_1$ and in fact, $p(i) \geq x_1$ for all i.

Suppose the statement holds for the k-th difference, and consider the $(k+1)$-th difference. The k-th difference is given by a polynomial of degree $n - k$ and gives the Betti numbers of some algebra. Because the first entry of the k-th column is positive, it follows from the forward direction of Lemma 4.1 that the $(k+1)$-th difference is also a polynomial of this form. By the first step, it follows that all entries for this polynomial are positive, and the induction is complete.

We now prove that every difference gives the Betti numbers over some 2-pyramidal algebra. We proceed by reverse induction on the columns of the difference table of p. Suppose p is a polynomial whose difference table contains only positive integers. In particular, the column of constants is some positive integer m. This polynomial represents $\beta_i(S_1)$ of the 2-pyramidal algebra,

$$1 \xrightarrow{\quad m \quad} 2 \, \circlearrowright$$

so the base case holds.

Assume that the statement holds for the $(n-k)$-th column, and consider the $(n-(k+1))$-th column. Because the first entry of the $(n-(k+1))$-th column is positive, it follows from the reverse direction of Lemma 4.1 that this column gives the Betti numbers of some 2-pyramidal algebra. This completes the induction, and thus p gives the Betti numbers of some 2-pyramidal algebra. $\qquad\square$

Note that in this proof, we only used the fact that the first entry in every column must be a positive integer. Indeed, this leads to a slightly stronger formulation of the theorem.

Corollary 4.3. *A polynomial p is such that $\beta_i(S_1) = p(i)$ for some 2-pyramidal algebra if and only if the first row of the difference table of p contains only positive integers.*

5. Producing pyramidal algebras given a polynomial

So far we have examined the types of polynomial growth possible for the Betti numbers of 2-pyramidal algebras. Another question that arises is: given a polynomial described in Corollary 4.3, can we produce a 2-pyramidal algebra whose Betti numbers follow this polynomial? Moreover, can we produce *all* algebras of this form that correspond to this polynomial?

We answer both of these questions in the affirmative. First, we need to define some notation. Let p be a polynomial. We then define $D_k(p)$ to be the k-th entry

of the first row of the difference table for p. As before, we denote the columns starting at 0 and ending at n.

Theorem 5.1. *Let p be a polynomial of degree n such that $\beta_i(S_1) = p(i)$ for some 2-pyramidal algebra. Then*

$$D_i(p) = \begin{cases} x_1 + y_1 & \text{if } i = 0, \\ x_1 x_2 \cdots x_{i-1} x_i (x_{i+1} + y_{i+1}) & \text{if } 1 \le i \le n-1, \\ x_1 x_2 \cdots x_{i-1} x_i y_{i+1} & \text{if } i = n. \end{cases}$$

Proof. The first case is immediate. We prove the second case by induction on i by looking at the algebras associated with the differences of p. For $i = 1$, we know that the first difference of p gives the Betti numbers for the 2-pyramidal algebra
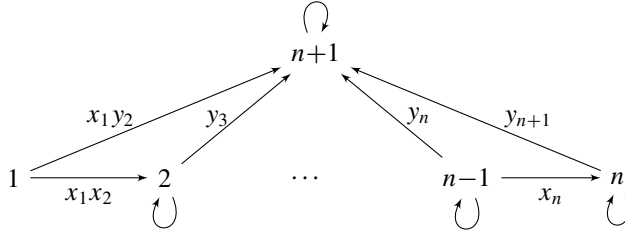
Hence, $D_1 = x_1 x_2 + x_1 y_2 = x_1 (x_2 + y_2)$. For the induction step, we assume that the k-th difference of p produces the Betti numbers over Λ_k, shown below:
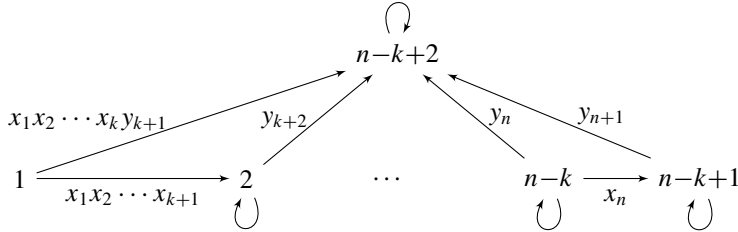
Then $D_k(p) = x_1 x_2 \cdots x_k (x_{k+1} + y_{k+1})$. By previous work, the first difference of the Betti numbers of the simple module S_1 over Λ_k are the Betti numbers of the simple module S_1 over
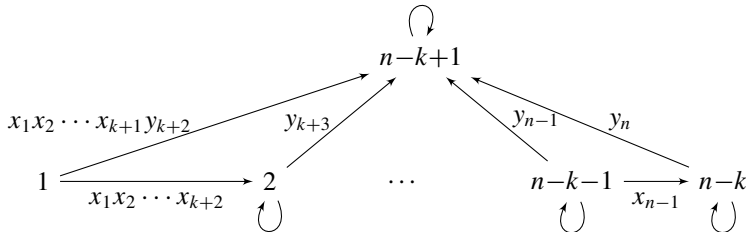
Note that this is also the algebra with the simple module S_1 whose Betti numbers are the $(k+1)$-th difference of p, and thus $D_{k+1} = x_1 x_2 \cdots x_k x_{k+1}(x_{k+2} + y_{k+2})$. This completes the induction for the second case.

For the last case, we know by the previous case that the $(n-1)$-th difference is given by the Betti numbers of the simple module S_1 over the 2-pyramidal algebra
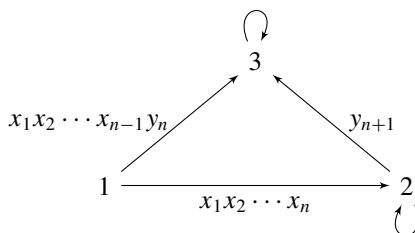
The difference of the Betti numbers of S_1 is given by the Betti numbers of the simple module S_1 over
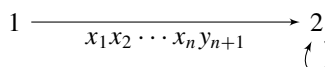
which is clearly the constant $x_1 x_2 \cdots x_n y_{n+1}$. $\qquad \square$

This theorem provides a way to determine restrictions on the x_i in order to produce a pyramidal algebra with a simple module S_1 whose Betti numbers follow a given polynomial. We now reformulate the previous theorem with added emphasis on the values of the x_i.

Corollary 5.2. *Let Λ be a 2-pyramidal algebra such that $\beta_i(S_1) = p(i)$ for some polynomial p. Then $x_1 x_2 \cdots x_k \mid D_k(p)$ and $x_k \leq D_{k-1}(p)/(x_1 x_2 \cdots x_{k-1})$ for all $k \leq n$.*

Theorem 5.3. *Let p be a polynomial of degree n and $x_1, x_2, \ldots, x_n$ be positive integers such that $x_1 x_2 \cdots x_k \mid D_k(p)$ and $x_k \leq D_{k-1}(p)/(x_1 x_2 \cdots x_{k-1})$ for all $k \leq n$. Then there exists a unique 2-pyramidal algebra such that $\beta_i(S_1) = p(i)$, and, for $1 \leq k \leq n$, the number of arrows between vertex k and vertex $k+1$ is x_k.*

Proof. We need only show that given these restrictions, we can choose the appropriate y_k such that $D_k(p)$ is the required value. For $k = 1$, simply choose $y_1 = D_1(p) - x_1$. Because $D_1(p)$ and x_1 are positive integers with $D_1(p) > x_1$, we know y_1 is a nonnegative integer as required.

Suppose that $2 \leq k \leq n - 1$. Then choose $y_k = D_{k-1}(p)/(x_1 x_2 \cdots x_{k-1}) - x_k$. This value is a nonnegative integer by assumption.

Finally, choose $y_{n+1} = D_n/(x_1 x_2 \cdots x_{n-1} x_n)$ to ensure that we have the equality $x_1 x_2 \cdots x_{n-1} x_n y_{n+1} = D_n$.

At each step in this process, there is only one choice for the value of y_k. Thus the 2-pyramidal algebra exists and is unique. $\qquad \square$

Given a polynomial p of degree n with $D_k(p) \in \mathbb{Z}$, this theorem allows us to construct a 2-pyramidal algebra with $\beta_i(S_1) = p(i)$. Simply choose the 2-pyramidal algebra on $n + 2$ vertices with $x_k = 1$ and $y_k = D_{k-1}(p) - 1$ for all k. The existence and uniqueness of these algebras given the appropriate choice of $\{x_i\}_{i=1}^n$ also provides a method of finding the number of algebras of this form whose Betti numbers correspond to a given polynomial.

Corollary 5.4. *Let p be a polynomial of degree n. Then the number of 2-pyramidal algebras such that $\beta_i(S_1) = p(i)$ is equal to the number of n-tuples $\{(x_1, x_2, \ldots, x_n)\}$ such that $x_i \in \mathbb{Z}^+$ for all i and $x_1 x_2 \cdots x_k \mid D_k(p)$ for all k and $x_k \leq D_{k-1}(p)/(x_1 x_2 \cdots x_{k-1})$ for all $k \leq n$.*

6. Generalizing by changing the ideal

Up until now, we have been examining algebras with $\mathrm{rad}^2 \Lambda = 0$. We will now consider algebras with $\mathrm{rad}^m \Lambda = 0$ for $m > 2$ and provide results analogous to the $m = 2$ case.

We use the following notation throughout this section. Given an algebra Λ with $\mathrm{rad}^m \Lambda = 0$ for some $m > 2$, let Λ' be the algebra that has the same underlying quiver as Λ with the relations $\mathrm{rad}^2 \Lambda' = 0$. Denote by S_1' the simple Λ'-module at vertex 1, by $\beta_k(S_1')$ the i-th Betti number and by $\Omega^i(S_1')$ the i-th syzygy of the simple module S_1' over the algebra Λ'.

Lemma 6.1. *Let Λ be an m-pyramidal algebra with $m \geq 2$. Let*

$$Q : \cdots \to Q_2 \to Q_1 \to Q_0 \to S_1 \to 0$$

be a minimal projective resolution of S_1, and let

$$Q' : \cdots \to Q_2' \to Q_1' \to Q_0' \to S_1' \to 0$$

be a minimal projective resolution of S_1' over Λ'. Then the number of indecomposable projective summands of Q_i is equal to the number of projective summands of $Q_{(i/2)m}'$ if i is even, and $Q_{((i-1)/2)m+1}'$ if i is odd. Hence, the Betti numbers of the Λ-module S_1 are given by

$$\beta_i(S_1) = \begin{cases} \beta_{(i/2)m}(S_1') & i \text{ is even,} \\ \beta_{((i-1)/2)m+1}(S_1') & i \text{ is odd.} \end{cases}$$

Note that for $m = 2$, the number of indecomposable projective modules in Q_i and Q_i' are equal, and $\beta_i(S_1) = \beta_i(S_1')$ for all i.

Proof. The $m = 2$ case is trivial. Let $m > 2$ be fixed and let Λ be an m-pyramidal algebra. We construct a list representing simple modules as follows. For each walk of length j starting at vertex 1 in the underlying quiver of Λ, record the vertex at the end of the walk in row j of the list. We use the convention that the trivial walk

from a vertex to itself along no edges is a walk of length 0, and the first written row, which is always a 1, is row 0.

For example, the 2-pyramidal algebra

$$1 \longrightarrow 2$$

generates the list

$$1$$
$$1\,2$$
$$1\,2\,2$$
$$1\,2\,2\,2$$
$$\vdots$$

Observe that the projective module appearing in step j of a minimal projective resolution of S_1' is precisely

$$\bigoplus_{k \in \text{row } j} P_k'.$$

With this in mind, we will prove the first statement by proving the following: for even i

$$Q_i = \bigoplus_{k \in \text{row}(i/2)m} P_k$$

and for odd i

$$Q_i = \bigoplus_{k \in \text{row}((i-1)/2)m+1} P_k.$$

We will prove this by induction on i. For $i = 0$ we have $Q_0 = P_1$. For $i = 1$, note that Q_1 is the projective cover of rad P_1. This is equal to the projective cover of its top radical layer, which is precisely $\bigoplus_{k \in \text{row } 1} S_k$, and this has projective cover $\bigoplus_{k \in \text{row } 1} P_k$.

We examine the syzygies of Q. Note that for any projective Λ-module A,

$$\text{soc } A \cong P(\text{soc } A)/\text{rad } P(\text{soc } A),$$

$$\text{rad } A \cong P(\text{rad } A)/\text{soc } P(\text{rad } A).$$

We will show by induction that for even i

$$\Omega^i(Q) = \text{soc } Q_{i-1}$$

and for odd i

$$\Omega^i(Q) = \text{rad } Q_{i-1}.$$

For $i = 1$, we have

$$\Omega^1(Q) = \ker(Q_0 \to S_1) = \text{rad } Q_0.$$

For $i = 2$

$$\Omega^2(Q) = \ker(Q_1 \to \operatorname{rad} Q_0) = \operatorname{soc} Q_1,$$

because $Q_1 = P(\operatorname{rad} Q_0)$ and

$$\operatorname{rad} Q_0 = P(\operatorname{rad} Q_0)/\operatorname{soc} P(\operatorname{rad} Q_0) = Q_1/\operatorname{soc} Q_1.$$

Assuming i is even and $\Omega^i(Q) = \operatorname{soc} Q_{i-1}$, we have

$$\begin{aligned}
\Omega^{i+1}(Q) &= \ker(Q_i \to \Omega^i(Q)) \\
&= \ker(Q_i \to \operatorname{soc} Q_{i-1}) \\
&= \ker[P(\operatorname{soc} Q_{i-1}) \to P(\operatorname{soc} Q_{i-1})/\operatorname{rad} P(\operatorname{soc} Q_{i-1})] \\
&= \operatorname{rad} P(\operatorname{soc} Q_{i-1}) = \operatorname{rad} Q_i.
\end{aligned}$$

Assuming i is odd and $\Omega^i(Q) = \operatorname{rad} Q_{i-1}$, we have

$$\begin{aligned}
\Omega^{i+1}(Q) &= \ker(Q_i \to \Omega^i(Q)) \\
&= \ker(Q_i \to \operatorname{rad} Q_{i-1}) \\
&= \ker[P(\operatorname{rad} Q_{i-1}) \to P(\operatorname{rad} Q_{i-1})/\operatorname{soc} P(\operatorname{rad} Q_{i-1})] \\
&= \operatorname{soc} P(\operatorname{rad} Q_{i-1}) = \operatorname{soc} Q_i.
\end{aligned}$$

We now return to the proof of the structure of the Q_i. Assume i is even. Then

$$Q_i = P(\Omega^i(Q)) = P(\operatorname{soc} Q_{i-1}).$$

By hypothesis,

$$\operatorname{soc} Q_{i-1} = \operatorname{soc} \bigoplus_{k \in \operatorname{row}((i-2)/2)m+1} P_k = \bigoplus_{k \in \operatorname{row}(i/2)m} S_k.$$

Because Q_i is the projective cover of $\operatorname{soc} Q_{i-1}$, it follows that

$$Q_i \cong \bigoplus_{k \in \operatorname{row}(i/2)m} P_k.$$

Assuming i is odd, we have

$$Q_i = P(\Omega^i(Q)) = P(\operatorname{rad} Q_{i-1}).$$

By hypothesis,

$$\operatorname{rad} Q_{i-1} = \operatorname{rad} \bigoplus_{k \in \operatorname{row}((i-1)/2)m} P_k.$$

Now Q_i is the projective cover of $\operatorname{rad} Q_{i-1}$, so it is the projective cover of $\operatorname{rad} Q_{i-1}/\operatorname{rad}^2 Q_{i-1}$ as well. Because the radical quotient of $\operatorname{rad} \bigoplus_{k \in \operatorname{row}((i-1)/2)m} P_k$

is $\bigoplus_{k \in \text{row}((i-1)/2)m+1} S_k$, it follows that

$$Q_i \cong \bigoplus_{k \in \text{row}((i-1)/2)m+1} P_k. \qquad \square$$

The next theorem gives us asymptotic information about the Betti numbers for m-pyramidal algebras for $m \geq 3$. We will be using the following notation.

Definition 6.2. For a function $f(x)$, we write $f(x) = \Theta(g)$ if there exist positive constants M and N, $M \leq N$ and a real number x_0 such that

$$Mg(x) \leq f(x) \leq Ng(x)$$

for all $x \geq x_0$.

Theorem 6.3. *For all $m \geq 3$ and $n \geq 1$, there exists an m-pyramidal algebra such that $\beta_i(S_1) = \Theta(i^n)$.*

Proof. Let n be a fixed positive integer. Let Λ be the algebra
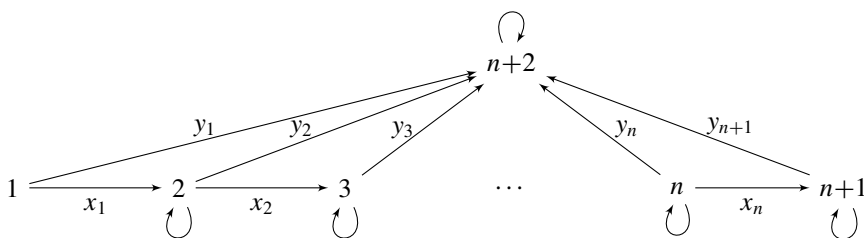
with $\text{rad}^m \Lambda = 0$. It suffices to show that $\beta_i(S_1)$ is bounded above and below by polynomials of degree n. Using Lemma 6.1 and the fact that the Betti numbers are strictly increasing for all $m \geq 2$, we obtain the inequalities

$$\beta_i(S_1') \leq \beta_i(S_1) \leq \beta_{mi}(S_1').$$

By previous work, $\beta_i(S_1') = p(i)$ and $\beta_{mi}(S_1') = p(mi)$, where p is a polynomial of degree n. Because both $p(i)$ and $p(mi)$ are polynomials in i of degree n, we have $\beta_i(S_1) = \Theta(i^n)$. $\qquad \square$

Future work

This work prompts some natural questions. We currently have a class of algebras whose quotients have Betti numbers asymptotic to polynomials of arbitrarily high degree. When does there exist a path algebra Λ such that, for some $m \geq 3$, the quotient $\Lambda / \text{rad}^m \Lambda$ has a simple module whose Betti numbers follow a polynomial exactly, not just asymptotically? Based on the proof of Lemma 6.1, it seems unlikely that there exists an algebra that satisfies this property for multiple m, but perhaps there exists such a path algebra for each m.

We showed that for polynomials of a certain type, we can construct an algebra whose Betti numbers at the simple module at vertex 1 satisfy that polynomial. However, the description of the number of such 2-pyramidal algebras, offered in Corollary 5.4, is complex. Perhaps there is a simpler description of the number of these algebras.

The Betti numbers of simple modules for a 2-pyramidal algebras are different at each vertex. A natural question is whether there exists an algebra where one of its quotients has the same polynomial Betti numbers at all of its simple modules. We can produce an algebra in which two simple modules have the same syzygies: starting with a 2-pyramidal path algebra, add a copy of vertex 1 called $\tilde{1}$, copy all of its arrows, and consider the new algebra modulo its radical squared. Then S_1 and $S_{\tilde{1}}$ have the same syzygies, and by repeating this process we can produce an algebra with arbitrarily many such simple modules. However, this process does not create a path algebra in which *all* simple modules have the same Betti numbers.

Acknowledgements

References

[Alperin and Evens 1981] J. L. Alperin and L. Evens, "Representations, resolutions and Quillen's dimension theorem", *J. Pure Appl. Algebra* **22**:1 (1981), 1–9. MR Zbl

[Assem et al. 2006] I. Assem, D. Simson, and A. Skowroński, *Elements of the representation theory of associative algebras, I: Techniques of representation theory*, London Math. Soc. Student Texts **65**, Cambridge Univ. Press, 2006. MR Zbl

[Auslander et al. 1995] M. Auslander, I. Reiten, and S. O. Smalø, *Representation theory of Artin algebras*, Cambridge Studies in Adv. Math. **36**, Cambridge Univ. Press, 1995. MR Zbl

[Avramov 1998] L. L. Avramov, "Infinite free resolutions", pp. 1–118 in *Six lectures on commutative algebra* (Bellaterra, Spain, 1996), edited by J. M. Giral et al., Progr. Math. **166**, Birkhäuser, Basel, 1998. MR Zbl

[Avramov and Buchweitz 2000] L. L. Avramov and R.-O. Buchweitz, "Support varieties and cohomology over complete intersections", *Invent. Math.* **142**:2 (2000), 285–318. MR Zbl

[Avramov et al. 1997] L. L. Avramov, V. N. Gasharov, and I. V. Peeva, "Complete intersection dimension", *Inst. Hautes Études Sci. Publ. Math.* **86** (1997), 67–114. MR Zbl

[Diveris and Purin 2014] K. Diveris and M. Purin, "Vanishing of self-extensions over symmetric algebras", *J. Pure Appl. Algebra* **218**:5 (2014), 962–971. MR Zbl

[Eisenbud 1980] D. Eisenbud, "Homological algebra on a complete intersection, with an application to group representations", *Trans. Amer. Math. Soc.* **260**:1 (1980), 35–64. MR Zbl

[Erdmann et al. 2004] K. Erdmann, M. Holloway, R. Taillefer, N. Snashall, and Ø. Solberg, "Support varieties for selfinjective algebras", *K-Theory* **33**:1 (2004), 67–87. MR Zbl

[Tate 1957] J. Tate, "Homology of Noetherian rings and local rings", *Illinois J. Math.* **1** (1957), 14–27. MR Zbl

coope786@umn.edu   *Department of Mathematics, University of Minnesota -*
                          Twin Cities, Minneapolis, MN, United States

purin@stolaf.edu   *Department of Mathematics, Statistics, and Computer*
                          Science, St. Olaf College, Northfield, MN, United States

msp

Orbit spaces of linear circle actions

Suzanne Craig, Naiche Downey, Lucas Goad,
Michael J. Mahoney and Jordan Watts

(Communicated by Colin Adams)

We show that nonisomorphic effective linear circle actions yield nondiffeomorphic differential structures on the corresponding orbit spaces.

1. Introduction

Recall that an orbifold is a topological space equipped with an atlas of linear representations of finite groups; in the case that all of these representations are effective, we say that the orbifold is effective (see, for instance, one of [Haefliger 1984; Moerdijk and Pronk 1997] for the precise definition). One can equip an effective orbifold with a "smooth structure" in many different ways [Moerdijk and Pronk 1997; Lerman 2010; Iglesias et al. 2010; Watts 2017]. No matter which notion of smoothness is taken, in [Watts 2017] it is shown that the underlying local semialgebraic set of a smooth (effective) orbifold, equipped with its natural differential structure, holds a complete set of orbifold invariants in its differential structure; that is, an atlas for the orbifold can be recovered from the smooth functions on the orbifold alone. It is natural to ask what happens in the case of a quotient by a smooth circle action on a manifold. The purpose of this paper is to take the first step toward solving this problem by considering the case of linear circle actions: can one recover a linear circle action (up to diffeomorphism) by examining the differential structure of the orbit space alone?

The question and result above can be seen from a broader perspective: there is a functor from Lie groupoids to differential spaces, sending a groupoid to its orbit space [Watts 2013]. Studying this functor, especially when restricted to proper Lie groupoids, leads to a modern connection between two classical subjects: Lie group actions of compact groups, and singular spaces (namely, semialgebraic varieties). The result on orbifolds in [Watts 2017] is that this functor when restricted to proper effective étale Lie groupoids is essentially injective (i.e., injective up to isomorphism). This paper deals with the restriction to linear $\mathbb{S}^1$-actions.

The generalization of smooth structures from manifolds to arbitrary subspaces and quotient spaces has a long history, though the perspective we take was first formally defined by Sikorski [1967; 1971] and is presented in brief here (see [Śniatycki 2013] for more details on these spaces).

Definition 1.1 (differential space). Let X be a nonempty set. A *differential structure* on X is a nonempty family $\mathcal{F}$ of real-valued functions satisfying:

(1) (smooth compatibility) For any positive integer k, functions $f_1, \ldots, f_k \in \mathcal{F}$, and $g \in C^\infty(\mathbb{R}^k)$, the composition $g(f_1, \ldots, f_k)$ is contained in $\mathcal{F}$.

(2) (locality) Equip X with the initial topology induced by $\mathcal{F}$, that is, the weakest topology such that each function in $\mathcal{F}$ is continuous. Let $f : X \to \mathbb{R}$ be a function such that for any $x \in X$ there exist an open neighborhood U of x and a function $h \in \mathcal{F}$ satisfying $f|_U = h|_U$. Then $f \in \mathcal{F}$.

A set X equipped with a differential structure $\mathcal{F}$ is called a *differential space* and is denoted by $(X, \mathcal{F})$. We will drop the notation $\mathcal{F}$ when it is superfluous. In the literature (for example, [Schwarz 1975]), authors use differential structures without naming them, possibly unaware that the structures had been formally named.

Definition 1.2 (smooth maps between differential spaces). Let $(X, \mathcal{F}_X)$ and $(Y, \mathcal{F}_Y)$ be two differential spaces. A map $F : X \to Y$ is *smooth* if $F^*\mathcal{F}_Y \subseteq \mathcal{F}_X$. F is called a *diffeomorphism* if it is smooth and has a smooth inverse. Denote the set of smooth maps between X and Y by $\mathcal{F}(X, Y)$.

Differential spaces with smooth maps between them form a category closed under taking subsets and quotients.

Definition 1.3 (subspace differential structure). Let $(X, \mathcal{F})$ be a differential space, and let $Y \subseteq X$ be a subset. Then Y acquires a differential structure $\mathcal{F}_Y$ as follows: $f \in \mathcal{F}_Y$ if for every $y \in Y$ there exist an open neighborhood $U \subseteq X$ of y and a function $g \in \mathcal{F}$ such that $f|_{U \cap Y} = g|_{U \cap Y}$. We call $(Y, \mathcal{F}_Y)$ a *differential subspace* of $(X, \mathcal{F}_X)$. Note that the subspace topology on Y equals the initial topology induced by $\mathcal{F}_Y$; see [Watts 2012, Lemma 2.28].

Definition 1.4 (quotient differential structure). Let $(X, \mathcal{F})$ be a differential space, let $\sim$ be an equivalence relation on X, and let $\pi : X \to X/\sim$ be the quotient map. Then $X/\sim$ obtains a differential structure $\mathcal{F}_\sim$, called the *quotient differential structure*, consisting of those functions $f : X/\sim \to \mathbb{R}$ whose pullback $\pi^*f : X \to \mathbb{R}$ is in $\mathcal{F}$.

It follows from a famous result of Schwarz [1975] that if G is a compact Lie group acting effectively and orthogonally on $\mathbb{R}^m$, then the orbit space $\mathbb{R}^m/G$ embeds into a Euclidean space $\mathbb{R}^n$ so that the quotient differential structure on $\mathbb{R}^m/G$ equals the subspace differential structure induced by the embedding. Denote this differential structure by $C^\infty(\mathbb{R}^m/G)$. In the case of finite groups, the diffeomorphism class

of $(\mathbb{R}^m/G, C^\infty(\mathbb{R}^m/G))$ as a differential space determines the group G up to isomorphism and the G-representation up to isomorphism (see the proof of the Main Theorem of [Watts 2017]). The invariants obtained from $C^\infty(\mathbb{R}^m/G)$ can be interpreted through two different perspectives: either quotient (linear representation) invariants on $\mathbb{R}^m \to \mathbb{R}^m/G$, or as subset invariants of $(\mathbb{R}^m/G, C^\infty(\mathbb{R}^m/G))$, arising by examining the underlying semialgebraic set of the orbit space (see Section 2A for a definition of semialgebraic set).

Moreover, this recovery of linear actions of finite groups allows one to form an orbifold atlas of a general effective orbifold from the differential structure of its orbit space. Consequently, the diffeomorphism class of the orbit space of a proper, effective, and locally free Lie group action on a manifold determines the corresponding Lie groupoid up to Morita equivalence. What is interesting about this result is that it does not hold for all proper and effective Lie group actions on manifolds, not even in the linear case. Indeed, consider $\mathrm{O}(m)$ acting on $\mathbb{R}^m$ by rotations and reflections. The orbit space is diffeomorphic to the closed ray $[0, \infty)$, which is independent of m, and so the subset invariants do not form a complete set of invariants for the group actions. And so, the original action cannot be recovered. The question remains, however: what information is missing from the subset invariants that would lead to a complete set of invariants so that the action can be recovered from the orbit space?

In this paper, we examine the case of a linear circle action and obtain the following theorem.

Theorem 1.5. *Let $\mathbb{S}^1$ act linearly and effectively on $\mathbb{R}^m$. The diffeomorphism class of the quotient differential space $(\mathbb{R}^m/\mathbb{S}^1, C^\infty(\mathbb{R}^m/\mathbb{S}^1))$ determines the $\mathbb{S}^1$-action up to equivariant linear isomorphism.*

Since linear isomorphisms are examples of diffeomorphisms, this theorem answers the question in the first paragraph above affirmatively. Note that one needs to know that the differential space *came from* a *linear* $\mathbb{S}^1$-action in the theorem above. Even knowing that the space is an orbit space of a smooth effective $\mathbb{S}^1$-action on a manifold is not sufficient to recover the action: consider $\mathbb{S}^1$ acting (effectively) on $\mathbb{S}^2$ by rotation about a fixed axis. The orbit space is diffeomorphic to the manifold-with-boundary $[-1, 1]$. But this action descends through the antipodal action to $\mathbb{RP}^2$, whose orbit space $[0, 1]$ is diffeomorphic to $[-1, 1]$. The missing information that would distinguish between these two orbit spaces is knowledge about the isotropy groups: at the preimage of $-1 \in \mathbb{S}^2/\mathbb{S}^1$, this isotropy group is $\mathbb{S}^1$; whereas at the preimage of $0 \in \mathbb{RP}^2/\mathbb{S}^1$, it is $\mathbb{Z}/2\mathbb{Z}$. Characterizing this information is the subject of future work.

This paper is organized as follows. In Section 2, we review compact group actions, orbit-type stratifications, the result on orbifolds mentioned above, and

linear circle actions. In Section 3, we give a description of $\mathbb{S}^1$-invariant polynomials for a linear circle action, culminating in Corollary 3.2. In Section 4, we take the opportunity to study the structure of the orbit space of a linear $\mathbb{S}^1$-action. While this is not needed in the proof of the main result, this structure is interesting in its own right and important for understanding the singularities that arise. Section 5 contains the proof of Theorem 1.5, and Section 6 contains several examples.

2. Preliminaries

2A. *Linear compact group actions.* Let us begin by reviewing the result of Schwarz and related background material. Fix a compact Lie group G, and assume G is acting linearly and effectively on $\mathbb{R}^m$. Without loss of generality, since G is compact, we can assume that $G \subseteq O(m)$. Denote by $P(\mathbb{R}^m)$ the ring of real-valued polynomials on $\mathbb{R}^m$ and by $P(\mathbb{R}^m)^G$ the subring of polynomials invariant under the G-action, that is, polynomials p satisfying $p(g \cdot x) = p(x)$ for all $x \in \mathbb{R}^m$ and $g \in G$. By Hilbert's basis theorem, $P(\mathbb{R}^m)^G$ is finitely generated. That is, there exist $\sigma_1, \ldots, \sigma_n \in P(\mathbb{R}^m)^G$ such that for any $p \in P(\mathbb{R}^m)^G$, there exists a polynomial $q \in P(\mathbb{R}^n)$ such that $p = q(\sigma_1, \ldots, \sigma_n)$. Moreover, we can choose $\sigma_1, \ldots, \sigma_n$ all to be homogeneous; that is, for each σ_i, its terms are all of the same degree.

Geometrically, what this means is that we can form the *Hilbert map* $\sigma : \mathbb{R}^m \to \mathbb{R}^n$ defined to be the n-tuple $(\sigma_1, \ldots, \sigma_n)$, and for every $p \in P(\mathbb{R}^m)^G$ there exists $q \in P(\mathbb{R}^n)$ such that $p = q \circ \sigma$. That is, $P(\mathbb{R}^m)^G$ is the image of the *pullback map* $\sigma^* : P(\mathbb{R}^n) \to P(\mathbb{R}^m)$ sending q to $\sigma^* q := q \circ \sigma$.

Schwarz [1975] extends this result from polynomials to smooth functions: the image of $\sigma^* : C^\infty(\mathbb{R}^n) \to C^\infty(\mathbb{R}^m)$ is exactly the invariant smooth functions $C^\infty(\mathbb{R}^m)^G$, the set of all smooth functions $f \in C^\infty(\mathbb{R}^m)$ such that $f(g \cdot x) = f(x)$ for all $x \in \mathbb{R}^m$ and $g \in G$.

Let $\pi : \mathbb{R}^m \to \mathbb{R}^m/G$ be the quotient map. Schwarz further shows that σ descends to a topological embedding $i : \mathbb{R}^m/G \to \mathbb{R}^n$; hence we can view $\mathbb{R}^m/G$ as a subset of $\mathbb{R}^n$, and it obtains a subspace differential structure in this way. Since π^* is an isomorphism from $C^\infty(\mathbb{R}^m/G)$ to $C^\infty(\mathbb{R}^m)^G$, which in turn is isomorphic to $\sigma^* C^\infty(\mathbb{R}^n)$, we conclude that $i^* C^\infty(\mathbb{R}^n) = C^\infty(\mathbb{R}^m/G)$; that is, the subspace differential structure equals the quotient differential structure.

We have the following lemma, which we use in the sequel.

Lemma 2.1. *Let G be a compact group acting on a smooth manifold M. The initial topology induced by $C^\infty(M/G)$ equals the quotient topology on the orbit space.*

Proof. Let U be an open set in the initial topology on M/G, and fix $x \in U$. There is a function $f \in C^\infty(M/G)$ such that

$$x \in f^{-1}((0, 1)) \subseteq U.$$

Let $\pi : M \to M/G$ be the quotient map and $\tilde{f} \in C^\infty(M)$ be such that $\pi^* f = \tilde{f}$. Then $\pi^{-1}(f^{-1}((0, 1))) = \tilde{f}^{-1}((0, 1))$ and so is open in M. So, $f^{-1}((0, 1))$ is open in the quotient topology. It follows that U is in the quotient topology.

In the other direction, let U be an open set in the quotient topology. Fix $x \in U$. Then the orbit $\pi^{-1}(x)$ is closed and contained in the G-invariant open set $\pi^{-1}(U)$. Let $\tilde{b} : M \to [0, 1]$ be a smooth bump function equal to 1 on the orbit $\pi^{-1}(x)$ with support in $\pi^{-1}(U)$. After averaging over G, we may assume $\tilde{b}$ is G-invariant. Thus it descends to a smooth function b on M/G such that

$$x \in b^{-1}((0, \infty)) \subseteq U.$$

That is, U is in the initial topology. $\qquad\square$

In particular, there is no ambiguity in the topology on $\mathbb{R}^m/G$ that we use. We can describe how $\mathbb{R}^m/G$ sits in $\mathbb{R}^n$ as a subset. A *semialgebraic set* $S = \bigcup_{i=1}^m S_i$ is a subset of $\mathbb{R}^n$, where the subsets S_i are of the form

$$S_i = \{x \in \mathbb{R}^n \mid r_{i,1}(x), \ldots, r_{i,k_i}(x) > 0 \text{ and } s_{i,1}(x) = \cdots = s_{i,\ell_i}(x) = 0\},$$

where $r_{i,1}, \ldots, r_{i,k_i}, s_{i,1}, \ldots, s_{i,\ell_i} \in P(\mathbb{R}^n)$. For our purposes, we will assume S is equipped with the subspace differential structure induced by $\mathbb{R}^n$. (We call a differential space that is locally diffeomorphic to semialgebraic sets a *local semialgebraic set*.) The Tarski–Seidenberg theorem [Seidenberg 1954; Tarski 1948; 1998] states that the image of a semialgebraic set under a polynomial map (such as the Hilbert map above) is again a semialgebraic set. It follows that since $\mathbb{R}^m/G$ sits inside $\mathbb{R}^n$ as the image of σ, it is a semialgebraic set.

2B. *Compact group actions on manifolds.* We now want to extend these ideas to group actions on manifolds. Again, let G be a compact Lie group acting smoothly on a manifold M with quotient map $\pi : M \to M/G$. Let $x \in M$, and let H be the stabilizer of the action at x. Note that H is compact. Define the *isotropy action* of H on $T_x M$ by $h \cdot v = h_* v$ for any $v \in T_x M$. Here we view elements of G as diffeomorphisms of M, and since elements of H fix x, this action is well-defined. It is also linear, which in the effective case puts us back into the situation described above. Note that for any $v \in T_x(G \cdot x)$, any smooth curve $c : \mathbb{R} \to G \cdot x$ such that $c(0) = x$ and $\dot{c}(0) = v$ and any $h \in H$,

$$h \cdot v = h_* v = \frac{d}{dt}\Big|_{t=0} h \cdot c(t),$$

where $\frac{d}{dt}\big|_{t=0} h \cdot c(t)$ is in $T_x(G \cdot x)$ since it is the derivative of a new smooth curve contained in $G \cdot x$. Thus, $T_x(G \cdot x)$ is an H-invariant linear subspace of $T_x M$, and so the isotropy action descends to a linear H-action (also called the *isotropy action*) on the normal space $V := T_x M / T_x(G \cdot x)$ to the G-orbit at x.

Since H is a subgroup of G, it acts on G by $h \cdot g = gh^{-1}$, and so we have the H-action on the product $G \times V$ defined by

$$h \cdot (g, v) := (gh^{-1}, h \cdot v).$$

Denote the orbit space of this action by $G \times_H V$. Note that G acts on $G \times_H V$ by $g' \cdot [g, v] = [g'g, v]$. We have the following theorem of Koszul [1953]; see also [Duistermaat and Kolk 2000, Section 2.3].

Theorem 2.2 (slice theorem). *Let G be a compact Lie group acting on a manifold M. For any $x \in M$ there exists an open G-invariant neighborhood U of x and a G-equivariant diffeomorphism $F : U \to G \times_H V$, where H is the stabilizer of x and V is the normal space to $G \cdot x$ in M at x equipped with the isotropy action of H.*

Remark 2.3. The slice theorem holds more generally for proper actions [Palais 1961], but we only need the compact case.

It follows from the slice theorem that for any point $\pi(x) \in M/G$ there is an open neighborhood of $\pi(x)$ of the form $V/H \cong (G \times_H V)/G$, where V is the normal space to $G \cdot x$ at x as above. By Lemma 2.1, the quotient topology on M/G is equal to the initial topology induced by the quotient differential structure $C^\infty(M/G)$. Since V/H is semialgebraic, it follows that M/G is a local semialgebraic set.

2C. *Orbit-type stratification.* Given a compact Lie group action of G on M, for any closed subgroup $H \leq G$ we define the subset $M_{(H)}$ as

$$M_{(H)} := \{x \in M \mid \text{there exists } g \in G \text{ such that } \mathrm{Stab}(x) = gHg^{-1}\}.$$

The connected components of these subsets partition M into embedded submanifolds which together form a *stratification*; in particular, the partition is locally finite and if C_1 and C_2 are two such submanifolds such that $C_1 \cap \bar{C}_2 \neq \varnothing$, then either $C_1 = C_2$ or C_1 is contained in the boundary of C_2. We refer to this stratification as the *orbit-type stratification*. (We do not intend to give the full definition of a stratification. This is in fact very involved and will take us away from the point of the paper. The reader who is interested should consult, for example, [Pflaum 2001]. For details on the orbit-type stratification, see [Duistermaat and Kolk 2000, Section 2.7].)

The sets $M_{(H)}$ are G-invariant and so descend to a partition of M/G into subsets $M_{(H)}/G$. The connected components of these again form a stratification, which again we will call the *orbit-type stratification* of M/G (see [Duistermaat and Kolk 2000, Sections 2.7, 2.8]). We are interested in the local form of these stratifications. Fix $x \in M$. By the slice theorem, there is a G-invariant open neighborhood U of x and a G-equivariant diffeomorphism $U \to G \times_H V$, where H is the stabilizer of x and V is the normal space to $G \cdot x$ at x, equipped with the isotropy action. As noted above, $(G \times_H V)/G$ is diffeomorphic to V/H. Since H is compact, there

exists an H-invariant inner product on V, and with respect to this inner product we can write $V \cong E \oplus F$, where E is the linear subspace of H-fixed points (that is, the maximal subspace on which H acts trivially), and F is an H-invariant complement. It follows that $V/H \cong E \times (F/H)$. Denote by k the dimension of F, by $\mathbb{S}^{k-1} \subseteq F$ the unit sphere with respect to an H-invariant norm and by L the quotient $\mathbb{S}^{k-1}/H$. The continuous map $\mathbb{S}^{k-1} \times [0, \infty) \to F$ sending $(x, t) \mapsto xt$ is H-invariant and descends to a homeomorphism between the cone of L, given by $(L \times [0, \infty))/(L \times \{0\})$, and the quotient F/H. The cone itself is a stratified space, with a stratum $S \times (0, \infty)$ for each orbit-type stratum S of L, along with the apex of the cone which we denote by z. The stratification of V/H contains a stratum $E \times S'$ for each stratum S' of F/H. We refer to L as the *link* of this stratification, and the apex of the cone z the *distinguished stratum* of F/H. As a differential space, $F/H \smallsetminus \{z\}$ is diffeomorphic to $L \times (0, \infty)$; however, be aware that the differential structure of F/H in any neighborhood of z does not necessarily equal the quotient differential structure near the apex of $(L \times [0, \infty))/(L \times \{0\})$. We explore the differential structure near z via an example at the end of Section 4.

As a last word on orbit-type stratifications, we have the following theorem [Śniatycki 2013, Theorem 4.3.10]. (While Śniatycki's proof is only for compact connected groups, the proof goes through for any proper action.)

Theorem 2.4 (orbit-type stratification is an invariant). *Let G be a compact Lie group acting smoothly on a manifold M. Then the orbit-type stratification of M/G is an invariant of $C^\infty(M/G)$.*

The proof of the theorem above comes from the fact that the connected components of orbit-type strata are exactly the accessible sets (also called orbits in the literature) of the family of all vector fields on M/G induced by $C^\infty(M/G)$. The details of this would take us too far afield, and so we merely emphasize the fact that the stratification is an invariant of the differential structure.

2D. *Recovering the action: the finite group case.* The purpose of this paper is to address the following question. Given an effective linear $\mathbb{S}^1$-action on $\mathbb{R}^m$, can we recover the action from the differential space $(\mathbb{R}^m/\mathbb{S}^1, C^\infty(\mathbb{R}^m/\mathbb{S}^1))$? As mentioned in the Introduction, in the case of a finite group Γ acting effectively and linearly on $\mathbb{R}^m$, the answer is affirmative: one can obtain invariants of the semialgebraic set $(\mathbb{R}^m/\Gamma, C^\infty(\mathbb{R}^m/\Gamma))$ from which the action of Γ on $\mathbb{R}^m$ can be recovered up to isomorphism. Recall that a compact Lie group action of G on an m-dimensional manifold M is *locally free* if the stabilizer of every point is finite. In this case, the slice theorem implies that for every point x of M/G, there is a finite subgroup Γ, the *isotropy group* of x, a linear Γ-action on $\mathbb{R}^m$, and an open neighborhood of x diffeomorphic to $\mathbb{R}^m/\Gamma$. The orbit space M/G equipped with an open cover by these neighborhoods, and for each of these neighborhoods the

corresponding linear representation, is an (*effective*) *orbifold*. (Again, we do not propose to give a rigorous definition of an orbifold; see the literature cited in the Introduction for details.) As mentioned above, locally, the differential structure encodes the linear representations of these finite groups; piecing together these local pictures into a global one, the following theorem due to one of us [Watts 2017] is obtained.

Theorem 2.5 (orbifold differential structure). *Let G be a compact Lie group acting smoothly*, *effectively*, *and locally freely on a connected manifold M. The diffeomorphism class of the quotient differential space $(M/G, C^\infty(M/G))$ determines the group action of G on M, up to Morita equivalence.*

Remark 2.6. (1) We will not define Morita equivalence in this paper, as this requires the introduction of the language of Lie groupoids. We simply note that the Morita equivalence class of a Lie groupoid representing the orbifold can be recovered from the differential structure.

(2) The "definition" of an effective orbifold we use above is not the typical definition used. In the literature, one usually uses an atlas definition or Lie groupoids. However, it is a theorem that any effective orbifold is the quotient of a manifold by a compact, effective, and locally free Lie group action. See [Satake 1956; 1957, Section 1.5; Haefliger 1984; Moerdijk 2002; Moerdijk and Pronk 1997, Theorem 4.1].

Corollary 2.7. *Let X be an effective orbifold*, *and fix $x \in X$. Then the isotropy group at x is determined up to isomorphism by the differential structure of X.*

While Corollary 2.7 is stated here as a consequence of Theorem 2.5, this fact is actually used in part of the proof of Theorem 2.5; see [Watts 2017, Theorem 5.10]. To prove it, locally about x, one uses an algorithm of [Haefliger and Ngoc Du 1984] that reproduces the orbifold fundamental group (and thus the isotropy group at x) from knowledge of the codimension-0, codimension-1, and codimension-2 strata, and the orders of isotropy groups at these codimension-2 strata. In turn, these orders of isotropy groups at codimension-2 strata can be obtained via the Milnor numbers of the corresponding singularities forming the codimension-2 strata. For details, consult [Watts 2017].

2E. *Linear circle actions.* Let $\mathbb{S}^1$ act linearly on $\mathbb{R}^n$. There is an $\mathbb{S}^1$-equivariant linear change of coordinates $\mathbb{R}^n \cong \mathbb{R}^{n-2m} \times \mathbb{C}^m$, where $\mathbb{S}^1$ acts trivially on $\mathbb{R}^{n-2m}$, and on $\mathbb{C}^m$ we have for all $e^{i\theta} \in \mathbb{S}^1$ and $(z_1, \ldots, z_m) \in \mathbb{C}^m$

$$e^{i\theta} \cdot (z_1, \ldots, z_m) := (e^{i\theta\alpha_1} z_1, \ldots, e^{i\theta\alpha_m} z_m); \tag{1}$$

the numbers α_j are integers called the *weights* of the action. Since complex conjugation is a diffeomorphism, we may assume the weights are nonnegative; in fact, we may assume further that any weight-0 factor is included in the $\mathbb{R}^{n-2m}$ factor,

set	order of stabilizer	codimension	number
$S_{1 \cdots m}$	$1 = \gcd(\alpha_1, \ldots, \alpha_m)$	0	$1 = \binom{m}{0}$
$S_{j_1 \cdots \hat{j}_\ell \cdots j_m}$	$\gcd(\alpha_{j_1}, \ldots, \hat{\alpha}_{j_\ell}, \ldots, \alpha_{j_m})$	2	$\binom{m}{1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
S_j	α_j	$2m - 2$	$\binom{m}{m-1}$
$\{0\}$	∞	$2m$	$1 = \binom{m}{m}$

Table 1. Data for linear circle action.

and so each α_j is positive. Finally, if we also impose the condition that the action be effective, then $\gcd(\alpha_1, \ldots, \alpha_m) = 1$ and all isotropy actions are effective.

Assume $\mathbb{S}^1$ acts on $\mathbb{C}^m$ effectively with positive weights. Denote by $S_{j_1 \cdots j_k}$ the subset

$$\{(0, \ldots, 0, z_{j_1}, 0, \ldots, 0, z_{j_k}, 0, \ldots, 0) \mid z_{j_\ell} \neq 0, \ \ell = 1, \ldots, k\}.$$

The open dense subset $S_{1 \cdots m}$ has trivial stabilizer at all points since

$$\gcd(\alpha_1, \ldots, \alpha_m) = 1.$$

It is a submanifold of dimension $2m$, and there exists exactly one such submanifold. The set $S_{1 \cdots \hat{j} \cdots m}$ is a submanifold of dimension $2m - 2$, and there exist exactly $\binom{m}{1}$ such submanifolds; the hat symbol means that we remove the corresponding index. And so on. If $e^{i\theta}$ fixes a point in $S_{j_1 \cdots j_k}$, then $e^{i\theta \alpha_{j_\ell}} = 1$ for $\ell = 1, \ldots, k$. That is, $e^{i\theta} \in \mathbb{Z}_{\alpha_{j_\ell}}$ for $\ell = 1, \ldots, k$, which is equivalent to $e^{i\theta} \in \mathbb{Z}_{\gcd(\alpha_{j_1}, \ldots, \alpha_{j_k})}$. We tabulate this data in Table 1.

One can also organize this table as integer labels on an $(m-1)$-simplex. Noticing that the m weights appear as the stabilizers of the sets S_j, we place each of these weights at the vertices of the simplex. If an edge connects two vertices labeled α_j and α_k, then we attach the integer label $\gcd(\alpha_j, \alpha_k)$ to the edge. More generally, attach the integer label $\gcd(\alpha_{i_1}, \ldots, \alpha_{i_{\ell+1}})$ to an ℓ-face whose vertices have associated weights $\alpha_{i_1}, \ldots, \alpha_{i_{\ell+1}}$. The interior of the simplex obtains a label of 1 since $\gcd(\alpha_1, \ldots, \alpha_m) = 1$.

The collection of sets in Table 1 partitions $\mathbb{C}^m$ into invariant submanifolds, and an orbit-type stratum is exactly the union of sets above whose points share the same stabilizer.

3. Description of the invariant polynomials

Our first order of business is to obtain a description of the invariant polynomials for an effective linear action of $\mathbb{S}^1$ on $\mathbb{C}^m$ as in (1). A fully satisfactory description

of the algebra of invariant polynomials for a general linear circle action (such as a generating set of invariant polynomials with minimal cardinality along with the relations between these polynomials) remains elusive; at least the authors are not aware of such a description in the literature. Often a minimal generating set can be obtained for specific cases, and on a case-by-case basis the relations can be derived from the invariant polynomials using a Gröbner basis [Sturmfels 1993, Chapter 1] and the techniques of [Procesi and Schwarz 1985]. Also, work has been done on determining the dimension of the subspace of invariant polynomials of a fixed degree; see [Herbig and Seaton 2014]. We take a different approach below, in which we give a simple condition that any invariant polynomial must satisfy. Fix an effective linear action of $\mathbb{S}^1$ on $\mathbb{C}^m$. Considering $\mathbb{C}^m$ as the real vector space $\mathbb{R}^{2m}$, it will be convenient to use coordinates $(z_1, \bar{z}_1, \ldots, z_m, \bar{z}_m)$. Let p be a homogeneous $\mathbb{C}$-valued polynomial on $\mathbb{C}^m$ of degree d. Let $\mathcal{K}$ be the set of all $2n$-tuples $K = (k_1, \bar{k}_1, \ldots, k_n, \bar{k}_n)$ such that $k_1 + \bar{k}_1 + \cdots + k_n + \bar{k}_n = d$. Then, p takes the form

$$p(z_1, \bar{z}_1, \ldots, z_n, \bar{z}_n) = \sum_{K \in \mathcal{K}} P_K z_1^{k_1} \bar{z}_1^{\bar{k}_1} \cdots z_n^{k_n} \bar{z}_n^{\bar{k}_n} \tag{2}$$

for some complex numbers P_K.

Proposition 3.1. *Let $\mathbb{S}^1$ act on $\mathbb{C}^m$ linearly and effectively with positive weights $\alpha_1, \ldots, \alpha_m$. Then a homogeneous $\mathbb{C}$-valued polynomial p as in (2) is invariant if and only if it satisfies the equation*

$$\alpha_1(k_1 - \bar{k}_1) + \cdots + \alpha_m(k_m - \bar{k}_m) = 0 \tag{3}$$

for each $K \in \mathcal{K}$ such that $P_K \neq 0$.

Proof. Fix a homogeneous polynomial as in (2). Then p takes the form

$$p(z_1, \bar{z}_1, \ldots, z_m, \bar{z}_m) = \sum_{K \in \mathcal{K}} P_K |z_1|^{k_1 + \bar{k}_1} \cdots |z_m|^{k_m + \bar{k}_m} e^{i(\psi_1(k_1 - \bar{k}_1) + \cdots + \psi_m(k_m - \bar{k}_m))},$$

where $z_j = |z_j| e^{i\psi_j}$ for each j. Applying $e^{i\theta}$ to p for an arbitrary $e^{i\theta} \in \mathbb{S}^1$, consider the difference $p - (e^{i\theta})^* p$:

$$p(z_1, \ldots, \bar{z}_m) - p(e^{i\theta} \cdot (z_1, \ldots, \bar{z}_m))$$

$$= \sum_{K \in \mathcal{K}} P_K |z_1|^{k_1 + \bar{k}_1} \cdots |z_m|^{k_m + \bar{k}_m} e^{i(\psi_1(k_1 - \bar{k}_1) + \cdots + \psi_m(k_m - \bar{k}_m))}$$
$$\times (1 - e^{i\theta(\alpha_1(k_1 - \bar{k}_1) + \cdots + \alpha_m(k_m - \bar{k}_m))}). \tag{4}$$

If p satisfies (3) for each K such that $P_K \neq 0$, then the right-hand side of (4) is 0, from which it follows that p is invariant.

Conversely, assume p is invariant. Then for any $e^{i\theta} \in \mathbb{S}^1$ the two polynomials p and $(e^{i\theta})^* p$ are equal; in particular, their polynomial coefficients are equal. In

terms of (4), this means that for each $K \in \mathcal{K}$

$$P_K(1 - e^{i\theta(\alpha_1(k_1 - \bar{k}_1) + \cdots + \alpha_m(k_m - \bar{k}_m))}) = 0.$$

Since $e^{i\theta}$ is arbitrary, it follows that for each $K \in \mathcal{K}$, either $P_K = 0$ or (3) holds. $\square$

Corollary 3.2. *Let* $\mathbb{S}^1$ *act on* $\mathbb{C}^m$ *linearly and effectively with positive weights* $\alpha_1, \ldots, \alpha_m$. *Then there exists a generating set of the invariant polynomials consisting solely of real and imaginary parts of monomials* $z_1^{k_1} \bar{z}_1^{\bar{k}_1} \cdots z_n^{k_n} \bar{z}_n^{\bar{k}_n}$ *satisfying* (3).

4. Description of the orbit space

Let $\mathbb{S}^1$ act on $\mathbb{C}^m$ linearly and effectively with positive weights $\alpha_1, \ldots, \alpha_m$. The purpose of this section is to study two main features of the differential structure of the orbit space $\mathbb{C}^m/\mathbb{S}^1$: the link and the distinguished stratum.

Since the stabilizers of the action away from the origin are proper subgroups of $\mathbb{S}^1$, we immediately have the following fact.

Proposition 4.1. *Let* $\mathbb{S}^1$ *act on* $\mathbb{C}^m$ *linearly and effectively. Then* $(\mathbb{C}^m \smallsetminus \{0\})/\mathbb{S}^1$ *is an orbifold diffeomorphic to* $(\mathbb{S}^{2m-1}/\mathbb{S}^1) \times (0, \infty)$.

Remark 4.2. If the action is not effective, one no longer has an effective orbifold. However, there remains a diffeomorphism between the two differential spaces.

In the case of all positive weights, the link $\mathbb{S}^{2m-1}/\mathbb{S}^1$ is a well-known orbifold called a *weighted projective space* $\mathbb{CP}(\alpha_1, \ldots, \alpha_m)$. Although typically a weighted projective space is considered with its complex structure, we discard that here and consider the corresponding differential subspace structure induced by $\mathbb{C}^m/\mathbb{S}^1$. As discussed in Section 2E, for fixed j, the stabilizer at each point $(0, \ldots, 0, z_j, 0, \ldots, 0)$, where $z_j \neq 0$, is $\mathbb{Z}_{\alpha_j}$. Any 1-dimensional orbit-type stratum in $\mathbb{S}^{2m-1}$ is equal to one of these sets intersected with $\mathbb{S}^{2m-1}$. Hence any 0-dimensional stratum of the corresponding weighted projective space has isotropy group $\mathbb{Z}_{\alpha_j}$. Similar statements can be obtained for each higher-dimensional stratum using Section 2E; this will be crucial in the proof of Theorem 1.5.

The distinguished stratum is the image of the origin, the unique fixed point of the action, via the quotient map (again, assuming all weights are positive). For general compact linear actions, the differential structure near such distinguished strata is interesting and important. It detects invariants there that are not topological. For example, consider $\mathbb{Z}_n$ acting on $\mathbb{C}$ by rotations. For each n, the orbit space is homeomorphic to the plane; however, the differential structure detects the so-called Milnor number (also known as a germ codimension) at the distinguished stratum from which the number n can be recovered; see [Watts 2017, Section 5] for more details. The Milnor number makes rigorous what can be interpreted in loose terms as "how fast" the "cone" converges to its apex, without the use of any type of

Riemannian metric. For instance, if we intersect $\mathbb{C}/\mathbb{Z}_n$ as a differential subspace of $\mathbb{R}^3$ with a plane through the distinguished stratum containing the axis of symmetry, we obtain a curve with a singularity diffeomorphic to the graph of $y^2 = x^n$ in $\mathbb{R}^2$ $(x > 0)$. Going back to the $\mathbb{S}^1$-action, we check for similar behavior.

Recall our notation

$$S_j := \{(0, \ldots, 0, z_j, 0, \ldots, 0) \mid z_j \neq 0\}.$$

Proposition 4.3. *Let $\mathbb{S}^1$ act on $\mathbb{C}^m$ linearly and effectively with positive weights $\alpha_1, \ldots, \alpha_m$. There exists a choice of Hilbert map $\sigma = (\sigma_1, \ldots, \sigma_n)$, where the polynomials σ_j are of the form in Corollary 3.2, such that the image of $\bigcup_j S_j$ under σ is closed under scalar multiplication by positive real numbers, forming the nonnegative parts of m coordinate axes of $\mathbb{R}^n$ with the 0-dimensional stratum at the origin. Moreover, the 0- and 1-dimensional orbit-type strata of $\mathbb{C}^m/\mathbb{S}^1$ are contained in these axes.*

Proof. Choose the generating set $\{\sigma_1, \ldots, \sigma_n\}$ to contain the polynomials $|z_i|^2$ for $i = 1, \ldots, m$, but no powers of these polynomials greater than 1. Then for each $k \in \{1, \ldots, n\}$, each polynomial σ_k when restricted to S_j is identically 0 unless it is equal to $|z_j|^2$. The result follows. $\qquad\square$

As an example, consider the case $m = 2$, $\alpha_1 = 1$, and $\alpha_2 = 2$. The orbit space is diffeomorphic to the semialgebraic set in $\mathbb{R}^4$ given by

$$y_1 \geq 0, \quad y_2 \geq 0, \quad y_3^2 + y_4^2 = y_1^2 y_2.$$

(See Example 6.2 for details.) Intersecting this differential subspace with the plane $y_3 = y_4 = 0$, we obtain the nonnegative y_1- and y_2-axes, which together form a curve with differential structure diffeomorphic to the graph of the absolute value function. The intersection with other planes yields different singularities, however. For example, intersecting with $y_3 = y_1 - y_2 = 0$ yields the curve $y_4^2 = y_1^3$, which has a more severe cusp.

5. Linear $\mathbb{S}^1$-actions

A more sophisticated version of Theorem 1.5 is below, along with its proof. We develop an algorithm in the proof for finding the weights of an effective linear circle action. Examples 6.3 and 6.4 illustrate this algorithm.

Let $\mathbb{S}^1$ act linearly on $\mathbb{R}^k$ and $\mathbb{R}^\ell$, and let $\psi : \mathbb{R}^k \to \mathbb{R}^\ell$ be a smooth $\mathbb{S}^1$-equivariant map. Then ψ descends to a map $\hat{\psi} : \mathbb{R}^k/\mathbb{S}^1 \to \mathbb{R}^\ell/\mathbb{S}^1$ making the following diagram commute:

$$
\begin{array}{ccc}
\mathbb{R}^k & \xrightarrow{\ \psi\ } & \mathbb{R}^\ell \\
{\scriptstyle \pi_k} \downarrow & & \downarrow {\scriptstyle \pi_\ell} \\
\mathbb{R}^k/\mathbb{S}^1 & \xrightarrow{\ \hat{\psi}\ } & \mathbb{R}^\ell/\mathbb{S}^1.
\end{array}
$$

Note that $\hat{\psi}$ is smooth. Indeed, let $f \in C^\infty(\mathbb{R}^\ell/\mathbb{S}^1)$. It is sufficient to show that $(\hat{\psi} \circ \pi_k)^* f \in C^\infty(\mathbb{R}^k)$. But $(\hat{\psi} \circ \pi_k)^* f = (\pi_\ell \circ \psi)^* f$. Since π_ℓ and ψ are smooth, we conclude that $(\hat{\psi} \circ \pi_k)^* f \in C^\infty(\mathbb{R}^k)$. This shows that smooth equivariant maps between $\mathbb{S}^1$-representations are sent to smooth maps between orbit spaces. In fact, this is a functor to differential spaces. Theorem 5.1 states that this functor is essentially injective. We say that a functor F is *essentially injective* if given objects c_1 and c_2 in its domain category, $F(c_1) \cong F(c_2)$ implies $c_1 \cong c_2$; that is, it is injective on objects up to isomorphism.

Theorem 5.1 (linear circle actions). *Let $\mathcal{C}$ be the category of all effective linear actions of $\mathbb{S}^1$ on finite-dimensional real vector spaces with smooth $\mathbb{S}^1$-equivariant maps between them. Then the functor from $\mathcal{C}$ to differential spaces sending such an $\mathbb{S}^1$-action on a vector space V to the differential space $(V/\mathbb{S}^1, C^\infty(V/\mathbb{S}^1))$, and sending smooth $\mathbb{S}^1$-equivariant maps to smooth maps between orbit spaces, is essentially injective on objects.*

Remark 5.2. To obtain Theorem 1.5 from Theorem 5.1, we need to show that if there is an $\mathbb{S}^1$-equivariant diffeomorphism φ between two $\mathbb{S}^1$-representations in $\mathcal{C}$, then there is also an $\mathbb{S}^1$-equivariant linear isomorphism between them. This follows from the fact that the actions of $\mathbb{S}^1$ are linear, and so we may identify any such representation with its tangent space at a fixed point with the induced action. Since φ maps the origin to another fixed point, the differential at the origin $d\varphi|_0$ satisfies what is required.

Proof. Let V be an $\mathbb{S}^1$-representation. Let $\dim V = n$, and identify V with $\mathbb{R}^n$. As mentioned previously, the action of $\mathbb{S}^1$ on $\mathbb{R}^n$ will always be isomorphic to an $\mathbb{S}^1$-action on the product $\mathbb{R}^{n-2m} \times \mathbb{C}^m$, where $\mathbb{S}^1$ acts on $\mathbb{R}^{n-2m}$ trivially, and on $\mathbb{C}^m$ by (1) such that the weights α_j are positive and $\gcd(\alpha_1, \ldots, \alpha_m) = 1$. To complete the proof, we need to obtain the integers $\alpha_1, \ldots, \alpha_m$, as well as the dimension $n - 2m$ of the trivial representation, from $C^\infty((\mathbb{R}^{n-2m} \times \mathbb{C}^m)/\mathbb{S}^1)$.

The quotient topology on $(\mathbb{R}^{n-2m} \times \mathbb{C}^m)/\mathbb{S}^1$ is equal to the initial topology induced by $C^\infty((\mathbb{R}^{n-2m} \times \mathbb{C}^m)/\mathbb{S}^1)$ by Lemma 2.1. The dimension of the space $(\mathbb{R}^{n-2m} \times \mathbb{C}^m)/\mathbb{S}^1$, which is $n - 1$, is a topological invariant: it is the topological dimension at generic points of the space. So, the differential structure identifies that the dimension of the $\mathbb{S}^1$-representation is n.

Since $\mathbb{S}^1$ acts trivially on $\mathbb{R}^{n-2m}$, the coordinate functions on this factor can be chosen as generators in a generating set of invariant polynomials on $\mathbb{R}^{n-2m} \times \mathbb{C}^m$. It follows that $(\mathbb{R}^{n-2m} \times \mathbb{C}^m)/\mathbb{S}^1$ is diffeomorphic to $\mathbb{R}^{n-2m} \times (\mathbb{C}^m/\mathbb{S}^1)$, and we shall identify the two spaces.

By Theorem 2.4, the orbit-type stratification on $\mathbb{R}^{n-2m} \times \mathbb{C}^m/\mathbb{S}^1$ can be obtained from $C^\infty(\mathbb{R}^{n-2m} \times \mathbb{C}^m/\mathbb{S}^1)$. In particular, since $\mathbb{C}^m$ contains a unique $\mathbb{S}^1$-fixed point, the orbit space $\mathbb{R}^{n-2m} \times \mathbb{C}^m/\mathbb{S}^1$ has a unique stratum A of minimal dimension

$n-2m$. It follows that the differential structure identifies the dimension of the trivial representation of $\mathbb{S}^1$ on $\mathbb{R}^{n-2m}$. We now only need to find the weights $\alpha_1, \ldots, \alpha_m$ to complete the proof. Note that we know ahead of time that there are m such weights, as we know the orbit space came from a linear circle action on $\mathbb{R}^n$ with a trivial factor $\mathbb{R}^{n-2m}$ of maximal dimension: we can derive m from the two numbers n and $n-2m$.

Removing A from $\mathbb{R}^{n-2m} \times \mathbb{C}^m/\mathbb{S}^1$, we are left with an orbit space of a locally free $\mathbb{S}^1$-action, i.e., an orbifold, by Proposition 4.1. By Corollary 2.7, the orders of the isotropy groups at each stratum of this orbit space can be obtained from its differential structure, and hence from $C^\infty(\mathbb{R}^{n-2m} \times \mathbb{C}^m/\mathbb{S}^1)$. Since these finite orders will correspond to subgroups of $\mathbb{S}^1$, we conclude that those isotropy groups are cyclic groups of the obtained orders.

We claim that there is a natural way to pick out the weights from the orders of these isotropy groups. That the weights appear at all among these orders is clear: the weight α_j is the order of the stabilizer of all points in

$$S_j = \{(0, \ldots, 0, z_j, 0, \ldots, 0) \mid z_j \neq 0\} \subseteq \mathbb{C}^m.$$

The weights α_j completely determine the orbit-type stratification of $\mathbb{C}^m$ (see Section 2E), and hence of $\mathbb{R}^{n-2m} \times \mathbb{C}^m$, and its orbit space. In fact, the orbit-type strata on the orbit space will be unions of images of the sets in Table 1 via the quotient map, and so we can also use the simplex to organize the strata of the orbit space. Since we know ahead of time that the orbit space is the result of an effective linear circle action, we take advantage of this knowledge and now produce an algorithm on the orbit space starting with the differential structure $C^\infty(\mathbb{R}^{n-2m} \times \mathbb{C}^m/\mathbb{S}^1)$ which obtains the weights.

Remove A from $\mathbb{R}^{n-2m} \times \mathbb{C}^m/\mathbb{S}^1$, and denote the collection of the remaining strata by $\mathfrak{S}$. Equip $\mathfrak{S}$ with a partial ordering $\preceq$, defined by $S \preceq T$ if $\overline{T} \cap S \neq \varnothing$, in which case $S \subseteq \overline{T}$. Note that $\mathfrak{S}$ is finite, with open and dense stratum $\mathcal{O}$ as the top stratum, meaning it is maximal with respect to $\preceq$. Also, if $S \preceq T$, then either $S = T$ or $\dim(S) < \dim(T)$. The *depth* of a stratum S is the number of distinct strata in a maximal chain with S as its minimal element:

$$\mathrm{depth}(S) := \sup\{\nu \mid S = S_0 \prec S_1 \prec \cdots \prec S_\nu = \mathcal{O}\}.$$

Start with an element $\mathcal{R}$ of $(\mathfrak{S}, \preceq)$ of maximum depth, and denote its codimension by $2r$; this will be a union of images of sets from Table 1 via the quotient map. $\mathcal{R}$ is represented by an $(m-1-r)$-face in the simplex mentioned above; equivalently, it is the union of the image via the quotient map of a codimension-$2r$ set in Table 1 along with some lower-dimensional such images in its boundary. In particular, we know $m - r$ vertices are contained in this face; as $\mathcal{R}$ is minimal with respect to $\preceq$ we know that the $m - r$ corresponding images of sets S_j via the quotient map are contained in $\mathcal{R}$. Since the isotropy groups at all points of $\mathcal{R}$ share the same

order, and as mentioned above we know what this order is, we conclude that we know what the order is at the $m - r$ vertices. So we have obtained $m - r$ weights. Repeat this step for all strata with the same depth as $\mathcal{R}$, keeping in mind that while a vertex may appear more than once when applying this step to different strata, its associated weight should only be recorded once.

We continue recursively. Fix a depth D. Suppose for each stratum $\mathcal{Q}$ of depth greater than D and for each vertex contained in the face associated to $\mathcal{Q}$, the associated weight is known. Let $\mathcal{P}$ be a stratum of depth D and codimension $2p$, which is represented by an $(m-1-p)$-face in the simplex. Consequently, this face contains $m - p$ vertices. Each stratum in $\overline{\mathcal{P}} \smallsetminus \mathcal{P}$ has depth greater than D, and by assumption, we know the weights associated to the vertices in their corresponding faces. If the total number of these vertices is not $m - p$, then the remaining vertices must be associated to the stratum $\mathcal{P}$ itself, and we obtain the order of the corresponding isotropy groups, which is constant at all points of $\mathcal{P}$. Repeat this for all strata of depth D. The result is that for each stratum of depth at least D, and for each vertex contained in the associated faces, the associated weights are known.

Applying this procedure to all strata of incrementally decreasing depth, we eventually reach the top stratum $\mathcal{O}$. If we have not obtained m weights at this point, we apply the argument above one more time to obtain the remaining weights, all equal to 1.

Since the algorithm above considers every possible vertex in the simplex (equivalently every set S_j in Table 1), it is guaranteed to produce m weights. We now have enough information to reconstruct the $\mathbb{S}^1$-action on $\mathbb{R}^n$. $\square$

6. Examples

Example 6.1 ($\mathbb{S}^1 \circlearrowleft \mathbb{C}$). Consider the action of $\mathbb{S}^1$ on $\mathbb{C}$ given by $e^{i\theta} \cdot z = e^{i\theta} z$. It follows from Proposition 3.1 that $|z|^2$ generates all invariant polynomials. Thus, the orbit space is identified with the closed interval $[0, \infty) \subset \mathbb{R}$. The orbit-type strata in $\mathbb{C}$ are the origin $\{0\}$ and its complement $\mathbb{C} \smallsetminus \{0\}$.

Example 6.2 ($\mathbb{S}^1 \circlearrowleft \mathbb{C}^2$). We compute a generating set of invariant polynomials with their relations, as well as the stabilizer groups, of $\mathbb{S}^1 \circlearrowleft \mathbb{C}^2$ with weights α_1 and α_2. We will assume that the action is effective, and so $\gcd(\alpha_1, \alpha_2) = 1$. The invariant polynomials can be obtained using Corollary 3.2:

$$p_1(z_1, z_2) = |z_1|^2, \quad p_3(z_1, z_2) = \mathfrak{Re}(z_1^{\alpha_2} \bar{z}_2^{\alpha_1}),$$
$$p_2(z_1, z_2) = |z_2|^2, \quad p_4(z_1, z_2) = \mathfrak{Im}(z_1^{\alpha_2} \bar{z}_2^{\alpha_1}).$$

The relations can be verified using a Gröbner basis [Sturmfels 1993, Chapter 1; Procesi and Schwarz 1985]:

$$p_1 \geq 0, \quad p_2 \geq 0, \quad p_3^2 + p_4^2 = p_1^{\alpha_2} p_2^{\alpha_1}.$$

set	order of stabilizer	codimension
S_{123}	1	0
S_{12}	1	2
S_{13}	1	2
S_{23}	1	2
S_1	1	4
S_2	2	4
S_3	3	4

Table 2. Data for Example 6.3.

The stabilizer groups can be computed directly:

$$\text{Stab}(0, 0) = \mathbb{S}^1,$$
$$\text{Stab}(z_1, 0) = \mathbb{Z}_{\alpha_1}, \quad z_1 \neq 0,$$
$$\text{Stab}(0, z_2) = \mathbb{Z}_{\alpha_2}, \quad z_2 \neq 0,$$
$$\text{Stab}(z_1, z_2) = \{1\} \quad \text{elsewhere.}$$

Here, in the case that α_1 or α_2 is 1, we define $\mathbb{Z}_1$ to be the trivial group.

Example 6.3 ($\mathbb{S}^1 \circlearrowright \mathbb{C}^3$). We illustrate the algorithm used in the proof of Theorem 5.1 for a simple example. Consider $\mathbb{S}^1$ acting on $\mathbb{C}^3$ linearly and effectively with weights 1, 2, and 3.

We find the orbit-type strata of the orbit space by constructing a table similar to Table 1; we do so in Table 2. The orbit-type strata are the distinguished stratum, the open dense stratum

$$\mathcal{O} = \pi(S_{123} \cup S_{12} \cup S_{13} \cup S_{23} \cup S_1)$$

and two 1-dimensional strata $\mathcal{P}_1 = \pi(S_2)$ and $\mathcal{P}_2 = \pi(S_3)$, with associated orders of isotropy groups 2 and 3, respectively.

The Hasse diagram for the partial order $\preceq$ introduced in the proof of Theorem 5.1 is as follows:
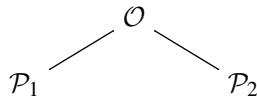
Stratum $\mathcal{P}_1$ has codimension 4, and so corresponds to a vertex of the simplex described in Section 2E (or equivalently a set S_j in Table 2); it has associated order 2, which is one of the weights. Similarly $\mathcal{P}_2$ has associated order 3, another weight. We are expecting three weights in total, and so the remaining weight must be the order associated to $\mathcal{O}$, which is 1.

Example 6.4 ($\mathbb{S}^1 \circlearrowleft \mathbb{C}^5$). We illustrate the algorithm used in the proof of Theorem 5.1 for a more complicated action. Let $\mathbb{S}^1$ act linearly and effectively on $\mathbb{C}^5$ with weights 2, 2, 3, 4, and 6.

To find the orbit-type strata of the orbit space, we could construct a table similar to Table 1; we do not to save space. The resulting strata are, besides the distinguished stratum:

$$\mathcal{O} = \pi(S_{12345} \cup S_{1234} \cup S_{1235} \cup S_{1345} \cup S_{2345} \cup S_{123} \cup S_{134}$$
$$\cup S_{135} \cup S_{234} \cup S_{235} \cup S_{345} \cup S_{13} \cup S_{23} \cup S_{34}),$$

$$\mathcal{P} = \pi(S_{1245} \cup S_{124} \cup S_{125} \cup S_{145} \cup S_{245} \cup S_{12} \cup S_{14} \cup S_{15} \cup S_{24} \cup S_{25} \cup S_{45} \cup S_1 \cup S_2),$$

$$\mathcal{Q} = \pi(S_{35} \cup S_3),$$

$$\mathcal{R}_1 = \pi(S_4),$$

$$\mathcal{R}_2 = \pi(S_5).$$

The associated orders of isotropy groups are

$$\mathcal{O}: 1, \quad \mathcal{P}: 2, \quad \mathcal{Q}: 3, \quad \mathcal{R}_1: 4, \quad \mathcal{R}_2: 6.$$

The Hasse diagram for the partial order $\preceq$ used in the proof of Theorem 5.1 is
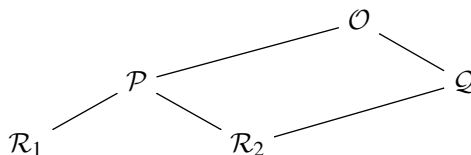
Stratum $\mathcal{R}_1$ has codimension 8, and so corresponds to a vertex of the simplex described in Section 2E (or equivalently a set S_j in Table 1); it has associated order 4, which is one of the weights. Similarly, stratum $\mathcal{R}_2$ also yields a weight, namely, 6. $\mathcal{P}$ has codimension 2, and therefore it corresponds to a 3-face in the simplex (equivalently, a set $S_{j_1 j_2 j_3 j_4}$ in Table 1), and so its closure contains four vertices. Two of the weights have been found, and so the other two must be associated to $\mathcal{P}$ itself, which has order 2. We now have weights 2, 2, 4, and 6. $\mathcal{Q}$ has codimension 6, and so corresponds to an edge in the simplex (equivalently, a set $S_{j_1 j_2}$). There are two vertices in its closure, one of which corresponds to $\mathcal{R}_2$. So the other weight must be the order associated to $\mathcal{Q}$, which is 3. Since we now have five weights, we are done.

Acknowledgements

References

[Duistermaat and Kolk 2000] J. J. Duistermaat and J. A. C. Kolk, *Lie groups*, Springer, 2000. MR Zbl

[Haefliger 1984] A. Haefliger, "Groupoïdes d'holonomie et classifiants", pp. 70–97 in *Transversal structure of foliations* (Toulouse, 1982), edited by J. Pradines, Astérisque **116**, Soc. Math. France, Paris, 1984. MR Zbl

[Haefliger and Ngoc Du 1984] A. Haefliger and Q. Ngoc Du, "Appendice: une présentation du groupe fondamental d'une orbifold", pp. 98–107 in *Transversal structure of foliations* (Toulouse, 1982), edited by J. Pradines, Astérisque **116**, Soc. Math. France, Paris, 1984. MR Zbl

[Herbig and Seaton 2014] H.-C. Herbig and C. Seaton, "The Hilbert series of a linear symplectic circle quotient", *Exp. Math.* **23**:1 (2014), 46–65. MR Zbl

[Iglesias et al. 2010] P. Iglesias, Y. Karshon, and M. Zadka, "Orbifolds as diffeologies", *Trans. Amer. Math. Soc.* **362**:6 (2010), 2811–2831. MR Zbl

[Koszul 1953] J.-L. Koszul, "Sur les représentations linéaires des algèbres de Lie résolubles", *C. R. Acad. Sci. Paris* **236** (1953), 2371–2372. MR Zbl

[Lerman 2010] E. Lerman, "Orbifolds as stacks?", *Enseign. Math.* (2) **56**:3-4 (2010), 315–363. MR Zbl

[Moerdijk 2002] I. Moerdijk, "Orbifolds as groupoids: an introduction", pp. 205–222 in *Orbifolds in mathematics and physics* (Madison, WI, 2001), edited by A. Adem et al., Contemp. Math. **310**, Amer. Math. Soc., Providence, RI, 2002. MR Zbl

[Moerdijk and Pronk 1997] I. Moerdijk and D. A. Pronk, "Orbifolds, sheaves and groupoids", *K-Theory* **12**:1 (1997), 3–21. MR Zbl

[Palais 1961] R. S. Palais, "On the existence of slices for actions of non-compact Lie groups", *Ann. of Math.* (2) **73** (1961), 295–323. MR Zbl

[Pflaum 2001] M. J. Pflaum, *Analytic and geometric study of stratified spaces*, Lecture Notes in Math. **1768**, Springer, 2001. MR Zbl

[Procesi and Schwarz 1985] C. Procesi and G. Schwarz, "Inequalities defining orbit spaces", *Invent. Math.* **81**:3 (1985), 539–554. MR Zbl

[Satake 1956] I. Satake, "On a generalization of the notion of manifold", *Proc. Nat. Acad. Sci. U.S.A.* **42** (1956), 359–363. MR Zbl

[Satake 1957] I. Satake, "The Gauss–Bonnet theorem for *V*-manifolds", *J. Math. Soc. Japan* **9** (1957), 464–492. MR Zbl

[Schwarz 1975] G. W. Schwarz, "Smooth functions invariant under the action of a compact Lie group", *Topology* **14** (1975), 63–68. MR Zbl

[Seidenberg 1954] A. Seidenberg, "A new decision method for elementary algebra", *Ann. of Math.* (2) **60** (1954), 365–374. MR Zbl

[Sikorski 1967] R. Sikorski, "Abstract covariant derivative", *Colloq. Math.* **18** (1967), 251–272. MR Zbl

[Sikorski 1971] R. Sikorski, "Differential modules", *Colloq. Math.* **24** (1971), 45–79. MR Zbl

[Śniatycki 2013] J. Śniatycki, *Differential geometry of singular spaces and reduction of symmetry*, New Math. Monographs **23**, Cambridge Univ. Press, 2013. MR Zbl

[Sturmfels 1993] B. Sturmfels, *Algorithms in invariant theory*, Springer, 1993. MR Zbl

[Tarski 1948] A. Tarski, "A decision method for elementary algebra and geometry", technical report R-109, RAND Corporation, 1948, available at https://www.rand.org/pubs/reports/R109.html. MR Zbl

[Tarski 1998] A. Tarski, "A decision method for elementary algebra and geometry", pp. 24–84 in *Quantifier elimination and cylindrical algebraic decomposition* (Linz, Austria, 1993), edited by B. F. Caviness and J. R. Johnson, Springer, 1998. MR Zbl

[Watts 2012] J. Watts, *Diffeologies, differential spaces, and symplectic geometry*, Ph.D. thesis, University of Toronto, 2012, available at https://search.proquest.com/docview/1346194956.

[Watts 2013] J. Watts, "The orbit space and basic forms of a proper Lie groupoid", preprint, 2013. arXiv

[Watts 2017] J. Watts, "The differential structure of an orbifold", *Rocky Mountain J. Math.* **47**:1 (2017), 289–327. MR Zbl

suzanne.craig@colorado.edu   *Department of Mathematics,*
                                   University of Colorado Boulder, Boulder, CO, United States

naiche.downey@colorado.edu   *Department of Mathematics,*
                                   University of Colorado Boulder, Boulder, CO, United States

lgoad@uccs.edu   *Department of Mathematics, University of Colorado*
                                   Colorado Springs, Colorado Springs, CO, United States

michael.j.mahoney@colorado.edu   *Department of Mathematics,*
                                   University of Colorado Boulder, Boulder, CO, United States

jordan.watts@cmich.edu   *Department of Mathematics, Central Michigan University,*
                                   Mt. Pleasant, MI, United States

■msp

On a theorem of Besicovitch and a problem in ergodic theory

Ethan Gwaltney, Paul Hagelstein, Daniel Herden and Brian King

(Communicated by Kenneth S. Berenhaut)

In 1935, Besicovitch proved a remarkable theorem indicating that an integrable function f on $\mathbb{R}^2$ is strongly differentiable if and only if its associated strong maximal function $M_S f$ is finite a.e. We consider analogues of Besicovitch's result in the context of ergodic theory, in particular discussing the problem of whether or not, given a (not necessarily integrable) measurable function f on a nonatomic probability space and a measure-preserving transformation T on that space, the ergodic averages of f with respect to T converge a.e. if and only if the associated ergodic maximal function $T^* f$ is finite a.e. Of particular relevance to this discussion will be recent results in the field of inhomogeneous diophantine approximation.

Let f be an integrable function on $\mathbb{R}^2$. A classical result in analysis, the *Lebesgue differentiation theorem*, tells us that, for a.e. $x \in \mathbb{R}^2$, the averages of f over disks shrinking to x tend to $f(x)$ itself. More precisely, we have

$$\lim_{r \to 0} \frac{1}{|B(x,r)|} \int_{B(x,r)} f = f(x) \quad \text{a.e.,}$$

where $B(x, r)$ denotes the open disk centered at x of radius r and $|B(x, r)| = \pi r^2$ denotes the area of the disk. For a proof of this result, the reader is encouraged to consult [Stein 1970].

What happens if we average over sets other than disks, say, open rectangles? It turns out that there exist integrable functions f on $\mathbb{R}^2$ such that, for a.e. $x \in \mathbb{R}^2$, there exists a sequence of rectangles $\{R_{x,j}\}$ shrinking toward x for which

$$\lim_{j \to \infty} \frac{1}{|R_{x,j}|} \int_{R_{x,j}} f$$

fails to converge. The news gets even more interesting. In fact, one can construct a function $f = \chi_E$ (that is, *the characteristic function of a set $E \subset \mathbb{R}^2$*) such that, for

---

a.e. $x \in \mathbb{R}^2$, there exists a sequence of rectangles $\{R_{x,j}\}$ shrinking toward x for which

$$\lim_{j \to \infty} \frac{1}{|R_{x,j}|} \int_{R_{x,j}} \chi_E$$

fails to converge. (See [de Guzmán 1975] for a nice exposition of this result. This result is closely related to the well-known *Kakeya needle problem*, and the interested reader is highly encouraged to consult [Falconer 1985].)

If we restrict the class of rectangles that we allow ourselves to average over, we obtain better results. Jessen, Marcinkiewicz, and Zygmund [Jessen et al. 1935] proved that if $\mathcal{B}_2$ consists of all the open rectangles in $\mathbb{R}^2$ *whose sides are parallel to the coordinate axes*, then for any function $f \in L^p(\mathbb{R}^2)$ with $1 < p \leq \infty$ one has

$$\lim_{j \to \infty} \frac{1}{|R_j|} \int_{R_j} f = f(x)$$

for a.e. $x \in \mathbb{R}^2$, where here $\{R_j\}$ is any sequence of rectangles in $\mathcal{B}_2$ shrinking toward x. (In this scenario we would say f is *strongly differentiable*.) Jessen, Marcinkiewicz, and Zygmund proved this by showing that the *strong maximal operator* M_S, defined by

$$M_S f(x) = \sup_{x \in R \in \mathcal{B}_2} \frac{1}{|R|} \int_R |f|,$$

satisfies for every $1 < p < \infty$ the *weak-type* (p, p) estimate

$$|\{x \in \mathbb{R}^2 : M_S f(x) > \alpha\}| \leq C_p \left(\frac{\|f\|_{L^p}}{\alpha} \right)^p.$$

This illustrates a paradigm that has been highly successful in the theory of differentiation of integrals. Namely, suppose one is given a collection of open sets $\mathcal{B} \subset \mathbb{R}^n$ and one wishes to ascertain whether, given a function f on $\mathbb{R}^n$, for a.e. x one must have

$$\lim_{j \to \infty} \frac{1}{|S_j|} \int_{S_j} f = f(x) \tag{0-1}$$

whenever $\{S_j\}$ is a sequence of sets in $\mathcal{B}$ shrinking toward x. (Here we assume that every point x is contained in a set in $\mathcal{B}$ of arbitrarily small diameter.) We may associate to the collection $\mathcal{B}$ a *maximal operator* $M_\mathcal{B}$ defined by

$$M_\mathcal{B} f(x) = \sup_{x \in S \in \mathcal{B}} \frac{1}{|S|} \int_S |f|.$$

It turns out that (0-1) will hold for every f in $L^p(\mathbb{R}^n)$ a.e. provided $M_\mathcal{B}$ satisfies a weak-type (p, p) estimate. A deep theorem of E. M. Stein [1961] tells us that, provided $\mathcal{B}$ is translation invariant in the sense that if $S \in \mathcal{B}$ then every translate of S also lies in $\mathcal{B}$, the limits above will hold for every $f \in L^p(\mathbb{R}^n)$ *only if* $M_\mathcal{B}$ satisfies a weak-

type (p, p) estimate. It is for this reason that maximal operators are an indispensable tool for mathematicians working with the topic of differentiation of integrals.

Having said that, it is interesting to consider the paper in *Fundamenta Mathematicae* immediately preceding the famous paper of Jessen, Marcinkiewicz, and Zygmund. In it, Besicovitch [1935] proved that, given any integrable function f on $\mathbb{R}^2$, if $M_S f$ is finite a.e., then for a.e. x we have

$$\lim_{j\to\infty} \frac{1}{|R_j|} \int_{R_j} f = f(x)$$

whenever $\{R_j\}$ is a sequence of sets in $\mathcal{B}_2$ shrinking to x. Of course, if $f \in L^p(\mathbb{R}^2)$ for $1 < p < \infty$, the quantitative weak-type (p, p) bound satisfied by M_S implies that $M_S f$ will be finite a.e. It is for this reason that this paper of Besicovitch has received comparatively little attention. However, it is of note that Besicovitch provides a mechanism for obtaining a.e. differentiability results bypassing the need for quantitative weak-type bounds on an associated maximal operator.

Let us provide an illustration of the usefulness of this approach. Let $f(x, y) = g(x)\chi_{[0,1]\times[0,1]}(x, y)$ be a function on $\mathbb{R}^2$, where $g \in L^1(\mathbb{R})$. Note f is in $L^1(\mathbb{R}^2)$ but not necessarily in $L^p(\mathbb{R}^2)$ for any $p > 1$. Suppose we wish to show that f is strongly differentiable. We can use the Fubini theorem combined with the weak-type $(1, 1)$ bounds of the Hardy–Littlewood maximal operator to show that $M_S f$ is finite a.e., so by the Besicovitch theorem we know that f is strongly differentiable. However, M_S is *not* of weak-type $(1, 1)$. Note that here we did not show that *every* function in $L^1(\mathbb{R}^2)$ is strongly differentiable, only that *some* of these functions are.

Many results in the study of differentiation of integrals have a "companion" result in ergodic theory; for instance the Lebesgue differentiation theorem is structurally very similar to that of the *Birkhoff ergodic theorem* on integrable functions. This observation may be found at least as far back as [Wiener 1939]. In that regard, it is natural to consider what the companion result of Besicovitch's theorem might be, when replacing the strong maximal operator M_S by an ergodic maximal operator. We are led immediately to the following conjecture.

Conjecture 1. *Let T be a measure-preserving transformation on the nonatomic probability space (X, Σ, μ) and let f be a μ-measurable function on that space. If $T^* f(x)$ is finite μ-a.e., where $T^* f$ is the ergodic maximal function defined by*

$$T^* f(x) = \sup_{N\geq 1} \frac{1}{N}\left|\sum_{j=0}^{N-1} f(T^j x)\right|,$$

then the limit

$$\lim_{N\to\infty} \frac{1}{N}\sum_{j=0}^{N-1} f(T^j x)$$

exists μ-a.e.

We remark that if f is integrable, then by the Birkhoff ergodic theorem the limit above automatically exists. If $f = f^+ - f^-$ is the difference of nonnegative measurable functions f^+ and f^-, then by the proof of the Birkhoff ergodic theorem the limit above still holds provided that *at least one* of the functions f^+ and f^- is integrable. (The reader may consult [Petersen 1983] for a proof of Birkhoff's classical result verifying that Fatou's lemma easily extends the given argument to the more general situation.) Thus, the interesting case is where $f = f^+ - f^-$ with

$$\int_X f^+ \, d\mu = \int_X f^- \, d\mu = \infty. \tag{0-2}$$

The main purpose of this note, aside from advertising the conjecture above, is to consider what happens when T corresponds to an ergodic transformation associated to an irrational rotation on $[0, 1)$ (identified with the unit circle $\mathbb{T}$), and $f(x) = 1/(x - \frac{1}{2})$. This scenario is so natural to consider that the reader might be surprised to find that it has not been treated before. (At least, the authors are unaware of any explicit treatment of this example.) In considering this situation, several issues immediately come to mind. First of all, f clearly satisfies (0-2), so we are not in a situation where we can apply the ergodic theorem. However, f exhibits a natural cancellation, so one might wonder whether the ergodic averages of f tend to 0 a.e. And, moreover, even if the ergodic averages of f did not tend to 0 a.e., it is still possible that the ergodic maximal function $T^* f$ is finite a.e. In that regard, this example seems to be a very worthy candidate for a counterexample of Conjecture 1.

It turns out that neither the ergodic averages of f with respect to T converge a.e. nor is the ergodic maximal function $T^* f$ finite a.e. The proof of the former follows readily from a theorem of Khintchine on the topic of inhomogeneous Diophantine approximation. The proof of the latter is much more subtle, following from relatively recent results of [Kim 2007].

Theorem 1. *Let ξ be an irrational number, and define the measure-preserving transformation T on $[0, 1)$ by*

$$Tx = (x + \xi) \bmod 1.$$

Define the function f on $[0, 1)$ by

$$f(x) = \frac{1}{x - \frac{1}{2}}.$$

If $x \in [0, 1)$, the limit

$$\lim_{N \to \infty} \frac{1}{N} \sum_{j=0}^{N-1} f(T^j x)$$

fails to converge to a finite number. Moreover for a.e. $x \in [0, 1)$ *we have*

$$T^* f(x) = \sup_{N \geq 1} \frac{1}{N} \left| \sum_{j=0}^{N-1} f(T^j x) \right| = \infty.$$

Proof. We first show that at no point $x \in [0, 1)$ does

$$\lim_{N \to \infty} \frac{1}{N} \sum_{j=0}^{N-1} f(T^j x)$$

converge to a finite value. It will be convenient for us to use the notation

$$\|x\| = \min_{n \in \mathbb{Z}} |x - n|.$$

We proceed by contradiction. Suppose for a given $x \in [0, 1)$ that

$$\lim_{N \to \infty} \frac{1}{N} \sum_{j=0}^{N-1} f(T^j x) = L < \infty.$$

Note that since ξ is an irrational number, by a theorem of Khintchine on inhomogeneous Diophantine approximation (see [Hua 1982, p. 267]) we have

$$\left\| q\xi + x - \tfrac{1}{2} \right\| < \frac{1}{q},$$

and thus

$$|f(T^q x)| = |f((q\xi + x) \bmod 1)| > q$$

for infinitely many positive integers q. Observe that

$$\frac{1}{q+1} \sum_{j=0}^{q} f(T^j x) - \frac{1}{q} \sum_{j=0}^{q-1} f(T^j x) = \frac{q}{q+1} \cdot \frac{1}{q} f(T^q x) - \frac{1}{q+1} \cdot \frac{1}{q} \sum_{j=0}^{q-1} f(T^j x).$$

Hence

$$\limsup_{q \to \infty} \left| \frac{1}{q+1} \sum_{j=0}^{q} f(T^j x) - \frac{1}{q} \sum_{j=0}^{q-1} f(T^j x) \right|$$

$$= \limsup_{q \to \infty} \left| \frac{q}{q+1} \cdot \frac{1}{q} f(T^q x) - \frac{1}{q+1} \cdot \frac{1}{q} \sum_{j=0}^{q-1} f(T^j x) \right|$$

$$= \limsup_{q \to \infty} \frac{1}{q} |f(T^q x)| \geq 1.$$

Accordingly, $\lim_{N \to \infty} \frac{1}{N} \sum_{j=0}^{N-1} f(T^j x)$ cannot converge to a finite value L, contradicting the supposition that these ergodic averages indeed did converge to L.

We now show that for a.e. $x \in [0, 1)$ we have

$$T^* f(x) = \sup_{N \geq 1} \frac{1}{N} \left| \sum_{j=0}^{N-1} f(T^j x) \right| = \infty.$$

To show this we use the relatively recent remarkable result of D. H. Kim [2007] that

$$\liminf_{q \to \infty} q \cdot \left\| q\xi + x - \tfrac{1}{2} \right\| = 0$$

for a.e. $x \in [0, 1)$. Thus

$$\limsup_{q \to \infty} \frac{1}{q} |f(T^q x)| = \infty$$

for a.e. $x \in [0, 1)$. Let $x \in [0, 1)$ be such that the limit superior above is infinite. We show that $T^* f(x) = \infty$. Again we proceed by contradiction. Suppose that

$$\sup_{N \geq 1} \frac{1}{N} \left| \sum_{j=0}^{N-1} f(T^j x) \right| = M < \infty.$$

Then, repeating the above calculation, we have the contradiction

$$2M \geq \limsup_{q \to \infty} \left| \frac{1}{q+1} \sum_{j=0}^{q} f(T^j x) - \frac{1}{q} \sum_{j=0}^{q-1} f(T^j x) \right|$$

$$= \limsup_{q \to \infty} \frac{1}{q} |f(T^q x)| = \infty. \qquad \square$$

In addition to Conjecture 1, we wish to indicate another conjecture the reader might find of interest. In Theorem 1 we showed that for a.e. $x \in [0, 1)$ the limit

$$\lim_{N \to \infty} \frac{1}{N} \sum_{j=0}^{N-1} f(T^j x)$$

does not converge to a finite number. What type of divergence is exhibited? By the apparent symmetries involved we would ordinarily expect that the ergodic averages of f do not converge to either positive or negative infinity, noting that the set

$$S = \left\{ x \in [0, 1) : \lim_{N \to \infty} \frac{1}{N} \sum_{j=0}^{N-1} f(T^j x) = \infty \right\}$$

is invariant under the ergodic transformation T and thus either of measure 0 or 1, and it would be strange for these averages to converge to $+\infty$ a.e. but not $-\infty$. Nonetheless, we do not have a proof of this, and the issue appears hard as the techniques involved in Kim's result do not indicate, given $x \in [0, 1)$, on which "side" of $\frac{1}{2}$ the points $(x + q\xi) \bmod 1$ close to $\frac{1}{2}$ lie. We formalize these ideas in the following:

Conjecture 2. *Let ξ be an irrational number, and define the measure-preserving transformation T on $[0, 1)$ by*

$$Tx = (x + \xi) \bmod 1.$$

Define the function f on $[0, 1)$ by

$$f(x) = \frac{1}{x - \frac{1}{2}}.$$

Then, for a.e. $x \in [0, 1)$

$$\limsup_{N \to \infty} \frac{1}{N} \sum_{j=0}^{N-1} f(T^j x) = \infty$$

and

$$\liminf_{N \to \infty} \frac{1}{N} \sum_{j=0}^{N-1} f(T^j x) = -\infty.$$

Conjectures 1 and 2 are topics of ongoing research.

Acknowledgements

References

[Besicovitch 1935] A. S. Besicovitch, "On differentiation of Lebesgue double integrals", *Fund. Math.* **25** (1935), 209–216. Zbl

[Falconer 1985] K. J. Falconer, *The geometry of fractal sets*, Cambridge Tracts in Math. **85**, Cambridge Univ. Press, 1985. MR Zbl

[de Guzmán 1975] M. de Guzmán, *Differentiation of integrals in $\mathbb{R}^n$*, Lecture Notes in Math. **481**, Springer, 1975. MR Zbl

[Hua 1982] L. K. Hua, *Introduction to number theory*, Springer, 1982. MR Zbl

[Jessen et al. 1935] B. Jessen, J. Marcinkiewicz, and A. Zygmund, "Note on the differentiability of multiple integrals", *Fund. Math.* **25** (1935), 217–234. Zbl

[Kim 2007] D. H. Kim, "The shrinking target property of irrational rotations", *Nonlinearity* **20**:7 (2007), 1637–1643. MR Zbl

[Petersen 1983] K. Petersen, *Ergodic theory*, Cambridge Studies in Adv. Math. **2**, Cambridge Univ. Press, 1983. MR Zbl

[Stein 1961] E. M. Stein, "On limits of sequences of operators", *Ann. of Math.* (2) **74** (1961), 140–170. MR Zbl

[Stein 1970] E. M. Stein, *Singular integrals and differentiability properties of functions*, Princeton Math. Series **30**, Princeton Univ. Press, 1970. MR Zbl

[Wiener 1939] N. Wiener, "The ergodic theorem", *Duke Math. J.* **5**:1 (1939), 1–18. MR Zbl

ethan.gwaltney@rice.edu   *Department of Mathematics, Rice University, Houston, TX, United States*

paul_hagelstein@baylor.edu   *Department of Mathematics, Baylor University, Waco, TX, United States*

daniel_herden@baylor.edu   *Department of Mathematics, Baylor University, Waco, TX, United States*

bking@rice.edu   *Department of Statistics, Rice University, Houston, TX, United States*

■msp

Algorithms for classifying points
in a 2-adic Mandelbrot set

Brandon Bate, Kyle Craft and Jonathon Yuly

(Communicated by Kenneth S. Berenhaut)

In her Ph.D. thesis, Jacqueline Anderson identified a nonarchimedean set similar in spirit to the Mandelbrot set which appears to exhibit a fractal-like boundary. We continue this research by presenting algorithms for determining when rational points lie in this set. We then prove that certain infinite families of points lie in (or out) of this set, giving greater resolution to the self-similarity present in this set.

1. Introduction

The Mandelbrot set and its higher-dimensional analogues are well-known sources of continuing research. These sets, which are defined via an archimedean metric, exhibit fascinating fractal-like boundaries. In this paper, we continue the study of a nonarchimedean (2-adic) set which appears to also have an interesting fractal-like boundary [Anderson 2013; Silverman 2013]. The definition of this set, which will be given shortly, is similar to that of the more familiar Mandelbrot set and its higher-dimensional variants. However, since this set is nonarchimedean, determining which elements are in the set is more difficult. To this end, we present two algorithms (Algorithms 4.7 and 5.3) which often determine when points lie in this set. The results of these algorithms reveal a variety of patterns. At various points we take note of patterns which appear to persist indefinitely, and when possible, prove this is indeed the case (Theorems 5.5 and 5.7). We note that Theorem 5.5 in particular expands upon results in [Anderson 2013].

In Section 2 we review some basic facts concerning fields with nonarchimedean absolute values, including a brief introduction to the field of p-adic numbers $\mathbb{Q}_p$, its ring of integers $\mathbb{Z}_p$, and the field $\mathbb{C}_p$, the topological completion of the algebraic closure of $\mathbb{Q}_p$. We review the definition of the Mandelbrot set in Section 3 and discuss generalizations of the Mandelbrot set to $\mathbb{C}_p$ along with stating an important critical radius bound given in [Anderson 2013].

Sections 4 and 5 explore the 2-adic Mandelbrot set $\mathcal{M}_{3,2}$ in more detail, which can be thought of as the set of

$$f(x) = f_{\alpha,\beta}(x) = x^3 - \tfrac{3}{2}(\alpha + \beta)x^2 + 3\alpha\beta x, \qquad (1\text{-}1)$$

with $\alpha, \beta \in \mathbb{C}_p$, for which both $\{f^n(\alpha)\}$ and $\{f^n(\beta)\}$ are bounded. Considering general $\alpha, \beta \in \mathbb{C}_p$ is beyond this scope of this paper; we restrict our attention to determining when $f_{\alpha,\beta}$ is in $\mathcal{M}_{3,2}$ for $\alpha, \beta \in \mathbb{Q}_p$. Section 4 begins with a number of elementary results, the proofs of which rely on little more than basic properties of p-adic numbers. Although fundamentally simple, these results, along with the critical radius bound given in [Anderson 2013], enable us to determine when $f_{\alpha,\beta}$ is in $\mathcal{M}_{3,2}$ for most $\alpha, \beta \in \mathbb{Q}_2$. There are instances where membership of $f_{\alpha,\beta}$ in $\mathcal{M}_{3,2}$ cannot be determined from these results. We break down such instances into two cases, one of which (the case where $\alpha, \beta \in \mathbb{Z}_p$ with $\alpha + \beta$ odd) is the primary focus of the remainder of this paper. For this case, Algorithm 4.7 often determines when $f_{\alpha,\beta}$ is in $\mathcal{M}_{3,2}$. Results of this algorithm are displayed in Figures 1 and 2. We close Section 4 by noting the difficulty of extending this algorithm beyond the case at hand.

Section 5 is focused entirely on understanding the structure of the intersection $\mathcal{M}_{3,2} \cap \{f_{\alpha,0} : \alpha \in \mathbb{Q}_2\}$. Prior work on this case was presented in [Anderson 2013, §6], resulting in the observation that $\mathcal{M}_{3,2}$ appears to have a fractal-like boundary at $\alpha = 1$. We continue this analysis by studying the sequence $\{x_n\}$ (defined in Lemma 5.1) as a proxy for $\{f^n(\alpha)\}$. We adapt the algorithm presented in Section 4 to the analysis of $\{x_n\}$ and present the results of this algorithm in Figures 3 and 4. By working with $\{x_n\}$, certain patterns in $\mathcal{M}_{3,2} \cap \{f_{\alpha,0} : \alpha \in \mathbb{Q}_2\}$ become more readily apparent. Theorems 5.5 and 5.7 classify certain classes of $f_{\alpha,0}$. Section 5 concludes with a discussion on how further improvements in the classification of $f_{\alpha,0}$ might be obtained.

2. Fields with nonarchimedean absolute values

We begin by reviewing some basic facts about nonarchimedean absolute values. We refer the reader to [Gouvêa 1993; Koblitz 1977] for more thorough introductions. Recall that an *absolute value* $|\cdot|$ on a field $\mathbb{K}$ is a function $|\cdot| : \mathbb{K} \to \mathbb{R}$ such that for all $x, y \in \mathbb{K}$,

(a) $|x| = 0$ if and only if $x = 0$,

(b) $|xy| = |x||y|$,

(c) $|x + y| \le |x| + |y|$.

If in addition we have that for all $x, y \in \mathbb{K}$

(d) $|x + y| \le \max\{|x|, |y|\}$,

then $|\cdot|$ is said to be *nonarchimedean*; otherwise $|\cdot|$ is *archimedean*. Note that (d) (the ultrametric inequality) implies (c) (the triangle inequality). For nonarchimedean absolute values, one can show that for all $x, y \in \mathbb{K}$

$$|x| < |y| \quad \Longrightarrow \quad |x + y| = |y|. \tag{2-1}$$

Furthermore, (d) and (2-1) can be extended to any number of elements. For instance,

$$|x + y + z| \le \max\{|x|, |y|, |z|\}, \tag{2-2}$$

$$|x|, |y| < |z| \quad \Longrightarrow \quad |x + y + z| = |z| \tag{2-3}$$

for all $x, y, z \in \mathbb{K}$. A field $\mathbb{K}$ equipped with a nonarchimedean absolute value has associated with it a topology induced by the metric $d(x, y) = |x - y|$. Such topological spaces are called *ultrametric spaces*. Absolute values are *equivalent* if they induce identical topologies on $\mathbb{K}$.

To define a nonarchimedean absolute value on $\mathbb{Q}$, we fix a prime number p. Let $v_p(n) = \max\{k \in \mathbb{Z}_{\ge 0} : p^k \mid n\}$, where $n \in \mathbb{Z}$. We then extend v_p to $\mathbb{Q}$ by defining $v_p(a/b) = v_p(a) - v_p(b)$ for $a, b \in \mathbb{Z}$, $b \ne 0$. The function v_p is called the *p-adic valuation* on $\mathbb{Q}$. The *p-adic absolute value* on $\mathbb{Q}$ is $|\cdot|_p : \mathbb{Q} \to \mathbb{R}$ such that $|x|_p = p^{-v_p(x)}$, where $x \in \mathbb{Q}$. One can show that $|\cdot|_p$ is a nonarchimedean absolute value on $\mathbb{Q}$ and that each nonarchimedean absolute value on $\mathbb{Q}$ is equivalent to a *p*-adic absolute value (Ostrowski's theorem). The set of *p-adic numbers*, denoted by $\mathbb{Q}_p$, is the completion of $\mathbb{Q}$ under the *p*-adic absolute value. This completion is obtained by taking the quotient of the ring of Cauchy sequences in $\mathbb{Q}$ (with respect to the topology induced by the *p*-adic absolute value) over the ideal of sequences in $\mathbb{Q}$ that converge to 0. The real numbers can be constructed similarly, but with the topology on $\mathbb{Q}$ induced from the usual archimedean absolute value. The set of rational numbers $\mathbb{Q}$ is dense in $\mathbb{Q}_p$, allowing the *p*-adic absolute value on $\mathbb{Q}$ to be extended to $\mathbb{Q}_p$, which we also denote by $|\cdot|_p$. One can show that the range of $|\cdot|_p$ on $\mathbb{Q}_p$ is $\{p^k : k \in \mathbb{Z}\}$.

Each $x \in \mathbb{Q}_p$ has a unique representation in the form of a finite-tailed Laurent series in p:

$$x = \sum_{n=n_0}^{\infty} a_n p^n = a_{n_0} p^{n_0} + a_{n_0+1} p^{n_0+1} + \cdots, \tag{2-4}$$

where $n_0 \in \mathbb{Z}$, $a_n \in \{0, 1, \ldots, p-1\}$, and $a_{n_0} \ne 0$. Let $\mathbb{Z}_p = \{x \in \mathbb{Q}_p : |x|_p \le 1\}$, which is called the set of *p-adic integers*. If $x \in \mathbb{Z}_p$ then the Laurent series representation of x has $n_0 \ge 0$. For $x, y \in \mathbb{Z}_p$ and $m \in \mathbb{Z}_{>0}$, we say that $x \equiv y \pmod{p^m}$ if there exists $c \in \mathbb{Z}_p$ such that $x - y = c \cdot p^m$. By (2-4), we see that for given $x \in \mathbb{Z}_p$ and $m \in \mathbb{Z}_{>0}$, there exists a unique $y \in \mathbb{Z}$ such that $x \equiv y \pmod{p^m}$ and $0 \le y < p^m$.

Like $\mathbb{R}$, the set $\mathbb{Q}_p$ is not algebraically closed. Let $\overline{\mathbb{Q}}_p$ denote the algebraic closure of $\mathbb{Q}_p$. We extend $|\cdot|_p$ to $\overline{\mathbb{Q}}_p$ by defining $|\alpha|_p = |N_{\mathbb{Q}_p(\alpha)/\mathbb{Q}_p}(\alpha)|_p^{1/m}$, where

$\alpha \in \overline{\mathbb{Q}}_p$ and $m = [\mathbb{Q}_p(\alpha) : \mathbb{Q}_p]$.[1] Unlike $\mathbb{C}$, the algebraic closure of $\mathbb{R}$, $\overline{\mathbb{Q}}_p$ is not topologically complete. To remedy this, let $\mathbb{C}_p$ denote the (topological) completion of $\overline{\mathbb{Q}}_p$, with $|\cdot|_p$ extending in the natural way. One can show that $\mathbb{C}_p$ is not only topologically complete but also algebraically complete. As one might suspect, $|\cdot|_p$ on $\mathbb{C}_p$ (hence also on $\overline{\mathbb{Q}}_p$ and $\mathbb{Q}_p$) is a nonarchimedean absolute value.

3. Mandelbrot sets over p-adic numbers

We begin by recalling the definition of the Mandelbrot set; further details can be found in [Beardon 1991; Devaney 1989]. Consider $f(z) \in \mathbb{C}[z]$ with $\deg(f) = 2$. For $n \in \mathbb{Z}_{>0}$, let $f^n(z)$ denote the n-th iterate of $f(z)$. For $\alpha \in \mathbb{C}$, let $\{f^n(\alpha)\}$ denote the sequence of n-th iterates of $f(z)$ evaluated at $z = \alpha$. We call $\{f^n(\alpha)\}$ the *orbit* of α under f. We say that $f(z)$ is *critically bounded* if $\{f^n(\alpha)\}$ is a bounded sequence for the critical point $\alpha \in \mathbb{C}$ of $f(z)$.[2] Let $g(z) = h \circ f \circ h^{-1}(z)$, where $h(z) = az + b \in \mathbb{C}[z]$ with $a \neq 0$. We say that $g(z)$ is a *linear conjugate* of $f(z)$. By choosing a and b appropriately, $g(z) = z^2 + c$ for some unique $c \in \mathbb{C}$. One can show that $f(z)$ is critically bounded if and only if $g(z)$ is critically bounded. Notice $g(z)$ has $\alpha = 0$ as its only critical point. The *Mandelbrot set* is

$$\begin{aligned} \mathcal{M} &= \{c \in \mathbb{C} : f_c(z) = z^2 + c \in \mathbb{C}[z] \text{ is critically bounded}\} \\ &= \{c \in \mathbb{C} : \{f_c^n(0)\} \text{ is bounded}\}. \end{aligned} \tag{3-1}$$

Since critical boundedness is well-defined up to linear conjugation, we can also think of $\mathcal{M}$ as the set of classes of linearly conjugate quadratic polynomials in $\mathbb{C}[z]$ that are critically bounded. The Mandelbrot set has a fractal-like boundary, the study of which is an active area of research [Dudko 2017; Lomonaco and Petersen 2017].

The definitions and analysis above translate without issue to $\mathbb{C}_p$. Thus it is natural to wonder whether a similarly defined set in $\mathbb{C}_p$ might also have a fractal-like boundary. Unfortunately, the natural candidate for a p-adic analogue to $\mathcal{M}$,

$$\{c \in \mathbb{C}_p : f_c(z) = z^2 + c \in \mathbb{C}_p[z] \text{ is critically bounded}\}, \tag{3-2}$$

is simply the unit disk $\{c \in \mathbb{C}_p : |c|_p \leq 1\}$ [Anderson 2013, Theorem 4.1]. However, it appears that more interesting sets can be obtained by considering polynomials of higher degree.

Consider $f(z) \in \mathbb{C}_p[z]$ with $d = \deg(f)$. We say that $f(z) \in \mathbb{C}_p[z]$ is *critically bounded* if $\{f^n(\alpha)\}$ is a bounded sequence for all critical points $\alpha \in \mathbb{C}_p$ of $f(z)$. As with quadratics, critical boundedness is well-defined up to linear conjugation,

---

[1] Here, $N_{\mathbb{Q}_p(\alpha)/\mathbb{Q}_p} : \mathbb{Q}_p(\alpha) \to \mathbb{Q}_p$ denotes the *norm* defined in field theory.

[2] A fundamental result in complex dynamics states that the Julia set of $f(z)$ is connected if and only if $f(z)$ is critically bounded. This result helps to explain why the classification of critically bounded $f(z)$ is of interest in its own right.

and in light of this, we can restrict our attention to monic $f(z)$ such that $f(0) = 0$ (every class of linearly conjugate degree-d polynomials has a representative of this form); that is to say, we consider

$$\mathcal{P}_{d,p} = \{x^d + a_{d-1}x^{d-1} + \cdots + a_1 x : (a_{d-1}, \ldots, a_1) \in \mathbb{C}_p^{d-1}\}. \qquad (3\text{-}3)$$

Let

$$\mathcal{M}_{d,p} = \{f \in \mathcal{P}_{d,p} : f \text{ is critically bounded}\}. \qquad (3\text{-}4)$$

If $p \geq d$ then $\mathcal{M}_{d,p} = \{f \in \mathcal{P}_{d,p} : |\alpha|_p \leq 1 \text{ for all critical points } \alpha \in \mathbb{C}_p \text{ of } f\}$ [Anderson 2013, Theorem 4.1, Proposition 4.2], in which case $\mathcal{M}_{d,p}$ lacks a fractal-like boundary. However, if $p < d$ then $\mathcal{M}_{d,p}$ may have a more intricate structure [Anderson 2013, §6]. Such sets are called *p-adic Mandelbrot sets*.

Let

$$r(d, p) = \sup_{f \in \mathcal{M}_{d,p}} \max_{\substack{\alpha \in \mathbb{C}_p \\ f'(\alpha) = 0}} \{-v_p(\alpha)\}, \qquad (3\text{-}5)$$

which is called the *critical radius* of $\mathcal{M}_{d,p}$. As can be easily checked, if α is a critical point of $f \in \mathcal{M}_{d,p}$ then $|\alpha|_p \leq p^{r(d,p)}$. In [Anderson 2013, Theorem 1.2], it was shown that for $d/2 < p < d$, we have $r(d, p) = p/(d-1)$. Therefore, if $d/2 < p < d$ then the critical points of $f \in \mathcal{M}_{d,p}$ are contained in a disk of radius $p^{p/(d-1)}$.

4. Determining when $f_{\alpha,\beta} \in \mathcal{M}_{3,2}$ for $\alpha, \beta \in \mathbb{Q}_2$

We focus our attention on $\mathcal{M}_{3,2}$, which, as was shown in [Anderson 2013, §6], appears to have a fractal-like boundary. Let $f \in \mathcal{P}_{3,2}$. Then

$$\begin{aligned} f(x) = f_{\alpha,\beta}(x) &= x^3 - \tfrac{3}{2}(\alpha + \beta)x^2 + 3\alpha\beta x \\ &= x(x^2 - \tfrac{3}{2}(\alpha + \beta)x + 3\alpha\beta), \end{aligned} \qquad (4\text{-}1)$$

where $\alpha, \beta \in \mathbb{C}_2$ are the critical points of $f(x)$. Since $d/2 < p < d$ for $d = 3$ and $p = 2$, we have $r(3, 2) = 2/(3-1) = 1$. Therefore all critical points of $f \in \mathcal{M}_{3,2}$ are contained in a disk of radius $p^{p/(d-1)} = 2$. Consequently, we only consider $f \in \mathcal{P}_{3,2}$ with $|\alpha| \leq 2$ and $|\beta| \leq 2$, where we write $|\cdot|$ for $|\cdot|_2$. Because of the complexity involved in dealing with elements of $\mathbb{C}_2$, we restrict our attention to $\alpha, \beta \in \mathbb{Q}_2$.

Lemma 4.1. *Let $f = f_{\alpha,\beta}$ with $\alpha, \beta \in \mathbb{Q}_2$ such that $|\alpha|, |\beta| \leq 2$. If $|f^m(\alpha)| > 4$ for some $m \in \mathbb{Z}_{>0}$ then $\lim_{n\to\infty} |f^n(\alpha)| = \infty$, and so $f \notin \mathcal{M}_{3,2}$.*

Proof. Since $|f^m(\alpha)| > 4$, we know $|f^m(\alpha)| = 2^k$ for some $k > 2$. Observe

$$|f^m(\alpha)^2| = 2^{2k}, \quad \left|-\tfrac{3}{2}(\alpha + \beta)f^m(\alpha)\right| \leq 2^{k+2}, \quad |3\alpha\beta| \leq 4.$$

Since $|f^m(\alpha)^2|$ is the largest of the quantities above, by (2-3)

$$|f^{m+1}(\alpha)| = |f^m(\alpha)| \, |f^m(\alpha)^2 - \tfrac{3}{2}(\alpha + \beta)f^m(\alpha) + 3\alpha\beta| = 2^k 2^{2k} = 2^{3k}.$$

By induction, $|f^{m+n}(\alpha)| = 2^{3^n k}$ for all $n \geq 1$. Thus $\lim_{n\to\infty} |f^n(\alpha)| = \infty$. □

Since we are considering $|\alpha|, |\beta| \leq 2$, we have $\alpha = a/2$ and $\beta = b/2$ for some $a, b \in \mathbb{Z}_2$. We say that $c \in \mathbb{Z}_2$ is *odd* if $|c| = 1$ and *even* if $|c| < 1$.

Proposition 4.2. *Let $f = f_{\alpha,\beta}$ with $\alpha = a/2$, $\beta = b/2$, $a, b \in \mathbb{Z}_2$:*

(a) *If $a + b$ is odd then $f \notin \mathcal{M}_{3,2}$.*

(b) *If a and b are odd and $a + b \equiv 2$ (mod 4) then $f \notin \mathcal{M}_{3,2}$.*

(c) *If a and b are even and $a + b \equiv 0$ (mod 4) (i.e., $\alpha, \beta \in \mathbb{Z}_2$ and $\alpha + \beta$ even) then $f \in \mathcal{M}_{3,2}$.*

Proof. For part (a), assume without loss of generality that a is odd and b is even. Thus $|\alpha| = 2$ and $|\beta| \leq 1$. Observe

$$|\alpha^2| = 4, \quad \left|-\tfrac{3}{2}(\alpha + \beta)\alpha\right| = 8, \quad |3\alpha\beta| \leq 2.$$

Since $\left|-\tfrac{3}{2}(\alpha + \beta)\alpha\right|$ is the largest of the quantities above, by (2-3)

$$|f(\alpha)| = |\alpha| \, |\alpha^2 - \tfrac{3}{2}(\alpha + \beta)\alpha + 3\alpha\beta| = 16.$$

Therefore by Lemma 4.1, $f \notin \mathcal{M}_{3,2}$.

For part (b), since $a + b \equiv 2$ (mod 4), we know $a + b = 2k$ for some odd $k \in \mathbb{Z}_2$. Observe

$$|f(\alpha)| = \left|\frac{a}{2}\right| \left|\frac{a^2}{4} - \frac{3(a+b)a}{8} + \frac{3ab}{4}\right| = 2 \left|\frac{a^2 - 3ka + 3ab}{4}\right|.$$

Since $a^2 - 3ka + 3ab$ is odd, $|f(\alpha)| = 8$. Thus by Lemma 4.1, $f \notin \mathcal{M}_{3,2}$.

For part (c), since $\alpha + \beta$ is even, $|\alpha + \beta| \leq \tfrac{1}{2}$. So if $|x| \leq 1$ then by (2-2)

$$|f(x)| = |x| \, |x^2 - \tfrac{3}{2}(\alpha + \beta)x + 3\alpha\beta|$$
$$\leq \max\left\{|x^2|, \left|-\tfrac{3}{2}(\alpha + \beta)\right| |x|, |3\alpha\beta|\right\} \leq 1.$$

Since $|\alpha|, |\beta| \leq 1$, we have $|f(\alpha)|, |f(\beta)| \leq 1$. By induction, $|f^n(\alpha)|, |f^n(\beta)| \leq 1$ for all $n \geq 1$. Thus $\{f^n(\alpha)\}$ and $\{f^n(\beta)\}$ are bounded, and hence $f \in \mathcal{M}_{3,2}$. □

The following lemma gives an improvement on Lemma 4.1 when $a + b$ is even. Although this improvement is slight, having it will make the classification of $f_{\alpha,\beta}$, given by Algorithm 4.7 simpler than it would be otherwise as well as yield simpler proofs for other results given in the remainder of this paper.

Lemma 4.3. *Let $f = f_{\alpha,\beta}$, with $\alpha = a/2$, $\beta = b/2$, $a, b \in \mathbb{Z}_2$ such that $a + b$ is even. If $|f^m(\alpha)| \geq 4$ for some $m \in \mathbb{Z}_{>0}$ then $f \notin \mathcal{M}_{3,2}$.*

Proof. If $|f^m(\alpha)| > 4$ then, by Lemma 4.1, $f \notin \mathcal{M}_{3,2}$. Thus it remains to consider the case when $|f^m(\alpha)| = 4$. Observe

$$|f^m(\alpha)^2| = 16, \quad \left| -\frac{3}{2}\left(\frac{a+b}{2}\right)f^m(\alpha) \right| \leq 8, \quad |3\alpha\beta| \leq 4.$$

Since $|f^m(\alpha)^2|$ is the largest of the quantities above, by (2-3)

$$|f^{m+1}(\alpha)| = |f^m(\alpha)|\left| f^m(\alpha)^2 - \frac{3}{2}\left(\frac{a+b}{2}\right)f^m(\alpha) + 3\alpha\beta \right| = 64.$$

Therefore by Lemma 4.1, $f \notin \mathcal{M}_{3,2}$. □

Notice that Proposition 4.2 considers all possibilities for $a, b \in \mathbb{Z}_2$ except for

(i) a and b even and $a + b \equiv 2 \pmod 4$ (i.e., $\alpha, \beta \in \mathbb{Z}_2$ with $\alpha + \beta$ odd),

(ii) a and b odd and $a + b \equiv 0 \pmod 4$.

We consider each of these cases separately, focusing primarily on (i) and briefly addressing (ii) at the end of this section.

Lemma 4.4. *Let $f = f_{\alpha,\beta}$ with $\alpha, \beta \in \mathbb{Z}_2$ and $\alpha + \beta$ odd. If $|f^m(\alpha)| \leq \frac{1}{2}$ for some $m \in \mathbb{Z}_{>0}$ then $\{f^n(\alpha)\}$ is bounded. Furthermore:*

(a) *If $|f^m(\alpha)| \leq \frac{1}{4}$ for some $m \in \mathbb{Z}_{>0}$, then $\lim_{n\to\infty} f^n(\alpha) = 0$ (i.e., α is in the basin of attraction of zero).*

(b) *If $|f^m(\alpha)| = \frac{1}{2}$ for some $m \in \mathbb{Z}_{>0}$, then $|f^{m+n}(\alpha)| = \frac{1}{2}$ for all $n \in \mathbb{Z}_{>0}$.*

Proof. Suppose $|f^m(\alpha)| = 2^{-k}$ for some $k \in \mathbb{Z}_{\geq 1}$. Since $\alpha + \beta$ is odd, α and β have opposite parity. Thus

$$|f^m(\alpha)^2| = 2^{-2k}, \quad \left| -\tfrac{3}{2}(\alpha+\beta)f^m(\alpha) \right| = 2^{1-k}, \quad |3\alpha\beta| \leq \tfrac{1}{2}. \qquad (4\text{-}2)$$

If $k \geq 2$ then by (2-2)

$$|f^{m+1}(\alpha)| = |f^m(\alpha)|\left| f^m(\alpha)^2 - \tfrac{3}{2}(\alpha+\beta)f^m(\alpha) + 3\alpha\beta \right| \leq 2^{-k-1}.$$

By induction, $|f^{m+n}(\alpha)| \leq 2^{-k-n}$ for all $n \geq 1$. Thus $\lim_{n\to\infty} f^n(\alpha) = 0$.

If instead $k = 1$ then $\left| -\tfrac{3}{2}(\alpha+\beta)f^m(\alpha) \right| = 1$ is the largest of the terms in (4-2). Thus by (2-3),

$$|f^{m+1}(\alpha)| = |f^m(\alpha)|\left| f^m(\alpha)^2 - \tfrac{3}{2}(\alpha+\beta)f^m(\alpha) + 3\alpha\beta \right| = \tfrac{1}{2}.$$

By induction, $|f^{m+n}(\alpha)| = \frac{1}{2}$ for all $n \geq 1$. □

Suppose $\alpha, \beta \in \mathbb{Z}_2$ with $\alpha + \beta$ odd. Without loss of generality, suppose throughout that α is odd and β is even. Notice that if $\beta \equiv 2 \pmod 4$ then $|\beta| = \frac{1}{2}$, and if $\beta \equiv 0 \pmod 4$ then $|\beta| \le \frac{1}{4}$. With this in mind, the following proposition follows from the proof of Lemma 4.4 if we replace $f^m(\alpha)$ with β.

Proposition 4.5. *Let $f = f_{\alpha,\beta}$, with $\alpha, \beta \in \mathbb{Z}_2$, with α odd and β even. Then $\{f^n(\beta)\}$ is bounded. Furthermore*:

(a) *If $\beta \equiv 0 \pmod 4$ then $\lim_{n\to\infty} f^n(\beta) = 0$.*

(b) *If $\beta \equiv 2 \pmod 4$ then $|f^n(\beta)| = \frac{1}{2}$ for all $n \in \mathbb{Z}_{>0}$.*

The following result follows from Lemmas 4.3 and 4.4 and Proposition 4.5.

Proposition 4.6. *Let $f = f_{\alpha,\beta}$, with $\alpha, \beta \in \mathbb{Z}_2$, with α odd and β even*:

(a) *There exists $n \in \mathbb{Z}_{>0}$ such that $|f^n(\alpha)| \ge 4$ if and only if $f \notin \mathcal{M}_{3,2}$.*

(b) *If there exists $n \in \mathbb{Z}_{>0}$ such that $|f^n(\alpha)| \le \frac{1}{2}$ then $f \in \mathcal{M}_{3,2}$.*

For given $\alpha, \beta \in \mathbb{Z}_2$ with α odd and β even, we can sometimes use Proposition 4.6 to determine whether $f = f_{\alpha,\beta} \in \mathcal{M}_{3,2}$. We do so by selecting an upper bound N and computing $f^n(\alpha)$ for $1 \le n \le N$. Proposition 4.6 can then be applied, except when $|f^n(\alpha)| \in \{1, 2\}$ for $1 \le n \le N$. Taking larger N may bring resolution in cases such as these, but as a practical matter, computing $f^n(\alpha)$ can slow substantially for large n. Indeed, even if $\alpha, \beta \in \mathbb{Z}$, we often find that $f^n(\alpha)$ consists of rational numbers whose numerators (and sometimes denominators) are rapidly increasing in size with respect to the archimedean absolute value, even if $|f^n(\alpha)| \in \{1, 2\}$ for $1 \le n \le N$.

Instead of considering fixed $\alpha, \beta \in \mathbb{Z}_2$ with α odd and β even, a more computationally efficient and general method is to first fix $j, k \in \mathbb{Z}_{>0}$ and fix $a_0, b_0 \in \mathbb{Z}$ such that a_0 is odd, b_0 is even, $0 \le a_0 < 2^j$, and $0 \le b_0 < 2^k$. Then consider all $\alpha \in D(a_0, 2^{-j})$ and $\beta \in D(b_0, 2^{-k})$, where

$$D(d, 2^{-m}) = \{x \in \mathbb{Q}_p : |x - d| \le 2^{-m}\} = \{x \in \mathbb{Q}_p : x \equiv d \pmod{2^m}\} \quad (4\text{-}3)$$

for $d \in \mathbb{Q}_p$ and $m \in \mathbb{Z}$. Equivalently, we can also think of $\alpha = a_0 + 2^j p$ and $\beta = b_0 + 2^k q$ for indeterminates $p, q \in \mathbb{Z}_2$. Let $f = f_{\alpha,\beta}$, which we now think of as being dependent upon p and q.

Consider $z = z(p, q) \in \mathbb{Q}_2[p, q]$. We abuse notation by saying that $z \in \mathbb{Z}_2$ if $z(p_0, q_0) \in \mathbb{Z}_2$ for any values $p_0, q_0 \in \mathbb{Z}_2$. We say that $z \notin \mathbb{Z}_2$ otherwise. It is easy to check that under this convention, $4f(\alpha) \in \mathbb{Z}_2$. Thus

$$j_1 = \max\{j' \in \mathbb{Z}_{\ge 0} : 4f(\alpha) \equiv c \pmod{2^{j'}} \text{ for some fixed } c \in \mathbb{Z}\} \quad (4\text{-}4)$$

is well-defined with the understanding that the statement $4f(\alpha) \equiv c \pmod{2^{j'}}$ for some fixed $c \in \mathbb{Z}$ means that regardless of whatever values in $\mathbb{Z}_2$ we may assign to p

in $\alpha = a_0 + 2^j p$ or to q in $\beta = b_0 + 2^k q$, we always have that $4f(\alpha) \equiv c \pmod{2^{j'}}$ for the same fixed $c \in \mathbb{Z}$. Alternatively, we can understand (4-4) as asserting that $4f(\alpha) \in D(c, 2^{-j_1})$.

Therefore

$$4f(\alpha) = c_1 + 2^{j_1} p_1$$

for indeterminate $p_1 \in \mathbb{Z}_2$, which itself is dependent upon p and q, and unique $c_1 \in \mathbb{Z}$ such that $0 \leq c_1 < 2^{j_1}$. To illustrate this, consider $a_0 = 3$, $b_0 = 4$, $j = 2$, and $k = 3$. Then, with some algebraic simplifications, we find that

$$4f(\alpha) = 4(a_0 + 2^j p)\big((a_0 + 2^j p)^2 - \tfrac{3}{2}(a_0 + 2^j p + b_0 + 2^k q)(a_0 + 2^j p)$$
$$+ 3(a_0 + 2^j p)(b_0 + 2^k q)\big)$$
$$= 162 + 2^3(3^2 \cdot 5p + 2^2 \cdot 3p^2 - 2^4 p^3 + 2 \cdot 3^3 q + 2^4 \cdot 3^2 pq + 2^5 \cdot 3p^2 q)$$
$$= 2 + 2^3(2^2 \cdot 5 + 3^2 \cdot 5p + 2^2 \cdot 3p^2 - 2^4 p^3 + 2 \cdot 3^3 q + 2^4 \cdot 3^2 pq + 2^5 \cdot 3p^2 q).$$

This shows that $c_1 = 2$ and $j_1 = 3$, with indeterminate

$$p_1 = 2^2 \cdot 5 + 3^2 \cdot 5p + 2^2 \cdot 3p^2 - 2^4 p^3 + 2 \cdot 3^3 q + 2^4 \cdot 3^2 pq + 2^5 \cdot 3p^2 q$$

(i.e., p_1 is a polynomial in p and q).

Continuing in this manner, we recursively define (possibly finite) sequences $\{c_n\}$ and $\{j_n\}$ as follows. Suppose $4f((c_n + 2^{j_n} p_n)/4) \in \mathbb{Z}_2$; we have established this for $n = 0$ if we take $c_0 = 4a_0$, $j_0 = j + 2$, and $p_0 = p$. Then

$$j_{n+1} = \max\left\{ j' \in \mathbb{Z}_{\geq 0} : 4f\left(\frac{c_n + 2^{j_n} p_n}{4}\right) \equiv c \pmod{2^{j'}} \text{ for some fixed } c \in \mathbb{Z} \right\} \quad (4\text{-}5)$$

is well-defined (with an important disclaimer in the following paragraph). Thus

$$4f\left(\frac{c_n + 2^{j_n} p_n}{4}\right) = c_{n+1} + 2^{j_{n+1}} p_{n+1}$$

for indeterminate $p_{n+1} \in \mathbb{Z}_2$ (ultimately expressible in terms of p and q) and unique $c_{n+1} \in \mathbb{Z}$ such that $0 \leq c_{n+1} < 2^{j_{n+1}}$. If $4f((c_n + 2^{j_n} p_n)/4) \notin \mathbb{Z}_2$ then we terminate $\{c_n\}$ and $\{j_n\}$ at index n.

We have already noted that $4f(\alpha) = c_1 + 2^{j_1} p_1$, and, as one can easily check, we also have

$$4f^n(\alpha) = 4f\left(\frac{c_{n-1} + 2^{j_{n-1}} p_{n-1}}{4}\right) = c_n + 2^{j_n} p_n \quad (4\text{-}6)$$

for $n \in \mathbb{Z}_{>0}$ whenever c_n and j_n are defined, provided that we think of p_n as being expressed in terms of p and q (i.e., as an element of $\mathbb{Z}[p, q]$). Thinking of p_n as a polynomial on two variables, it may be the case that $p_n : \mathbb{Z}_2^2 \to \mathbb{Z}_2$ is not surjective. However, determining the range of p_n is complicated. Rather than keeping track of this in our algorithm, we view p_n as an independent indeterminate

when computing j_{n+1}; that is, we think of p_n as being able to take on any value in $\mathbb{Z}_2$. By divorcing the relationship of p_n to p and q, the statement that $4f^n(\alpha) = c_n + 2^{j_n}p_n$ is, strictly speaking, false. Nevertheless, we will continue to write $4f^n(\alpha) = c_n + 2^{j_n}p_n$ with the understanding that for any choice of values in $\mathbb{Z}_2$ substituted in for p and q there exists a corresponding value in $\mathbb{Z}_2$ that when substituted in for p_n produces $4f^n(\alpha) = c_n + 2^{j_n}p_n$.

There is a close connection between $|f^n(\alpha)|$ and c_n. If $j_n > 0$ and $c_n \neq 0$, then there exists $e \in \mathbb{Z}_{\geq 0}$ such that $c_n \equiv 2^e \pmod{2^{e+1}}$ and $e+1 \leq j_n$. Thus by (4-6),

$$
\begin{aligned}
c_n \equiv 2^e \pmod{2^{e+1}} &\iff c_n + 2^{j_n}p_n \equiv 2^e \pmod{2^{e+1}} \\
&\iff 4f^n(\alpha) \equiv 2^e \pmod{2^{e+1}} \\
&\iff |4f^n(\alpha)| = 2^{-e} \\
&\iff |f^n(\alpha)| = 2^{2-e}. \qquad \text{(4-7)}
\end{aligned}
$$

Therefore by Proposition 4.6, $f \notin \mathcal{M}_{3,2}$ if $e = 0$ and $f \in \mathcal{M}_{3,2}$ if $e \geq 3$. Because of this, we terminate $\{c_n\}$ and $\{j_n\}$ at index n if $e = 0$ or $e \geq 3$.

If instead $c_n = 0$ then by (4-6), $|4f^n(\alpha)|$ is dependent upon p_n, and so we cannot know the exact value of $|f^n(\alpha)|$ from c_n and j_n. However by (4-6),

$$
\begin{aligned}
c_n \equiv 0 \pmod{2^{j_n}} &\implies 4f^n(\alpha) \equiv 0 \pmod{2^{j_n}} \\
&\implies |4f^n(\alpha)| \leq 2^{-j_n} \\
&\implies |f^n(\alpha)| \leq 2^{2-j_n}. \qquad \text{(4-8)}
\end{aligned}
$$

This shows that if $j_n \geq 3$ and $c_n = 0$ then $f \in \mathcal{M}_{3,2}$ by Proposition 4.6(b). If $j_n \leq 2$ and $c_n = 0$ then we cannot determine the behavior of $\{f^n(\alpha)\}$; indeed, one can check that if $j_n = 0$ then $4f((c_n + 2^{j_n}p_n)/4) \notin \mathbb{Z}_2$, if $c_n = 0$ and $j_n = 1$ then $j_{n+1} = 0$, and if $c_n = 0$ and $j_n = 2$ then $c_{n+1} = 0$ and $j_{n+1} = 1$. Given that the behavior of $\{f^n(\alpha)\}$ when $c_n = 0$ is either fully classified (as when $j_n \geq 3$) or impossible to determine (as when $j_n \leq 2$), we terminate $\{c_n\}$ and $\{j_n\}$ at index n whenever $c_n = 0$.

As stated earlier, we terminate $\{c_n\}$ and $\{j_n\}$ at index n if $4f((c_n + 2^{j_n}p_n)/4) \notin \mathbb{Z}_2$. However, the additional termination criteria we just gave (i.e., terminate if $c_n \neq 0$ and $e = 0$ or $e \geq 3$, or if $c_n = 0$) makes it so that we never have to test for this. To see why this is the case, notice that we only compute c_{n+1} and j_{n+1} if $j_n \geq 2$ and $c_n \equiv 2 \pmod{4}$ or if $j_n \geq 3$ and $c_n \equiv 4 \pmod{8}$ (i.e., when $e = 1, 2$ in (4-7)). One can show that in either of these cases $4f((c_n + 2^{j_n}p_n)/4) \in \mathbb{Z}_2$.

In addition to the previously mentioned criteria for determining when $f \in \mathcal{M}_{3,2}$, we also have

$$
\text{if } c_n = c_{n+\ell} \text{ and } j_n = j_{n+\ell} \text{ for some } n, \ell \in \mathbb{Z}_{>0} \text{ then } f \in \mathcal{M}_{3,2}. \qquad \text{(4-9)}
$$

Indeed, if $c_n = c_{n+\ell}$ and $j_n = j_{n+\ell}$ for some n, $\ell \in \mathbb{Z}_{>0}$ then $\{(c_n, j_n)\}$ is preperiodic, which by (4-7) makes $\{|f^n(\alpha)|\}$ also preperiodic.

We summarize the discussion above in the following result.

Algorithm 4.7. Fix $j, k \in \mathbb{Z}_{>0}$ and fix $a_0, b_0 \in \mathbb{Z}$ such that a_0 is odd, b_0 is even, $0 \le a_0 < 2^j$, and $0 \le b_0 < 2^k$. Consider $\alpha, \beta \in \mathbb{Z}_2$ such that $\alpha \equiv a_0 \pmod{2^j}$ and $\beta \equiv b_0 \pmod{2^k}$. Let $f = f_{\alpha,\beta}$, $c_0 = 4a_0$, and $j_0 = j + 2$. We recursively define sequences $\{c_n\}$ and $\{j_n\}$ as follows: if $j_n \ge 2$ and $c_n \equiv 2 \pmod 4$ or if $j_n \ge 3$ and $c_n \equiv 4 \pmod 8$, define

$$j_{n+1} = \max \left\{ j' \in \mathbb{Z}_{\ge 0} : 4f\left(\frac{c_n + 2^{j_n} p_n}{4}\right) \equiv c \pmod{2^{j'}} \right.$$
$$\left. \text{for all } p_n \in \mathbb{Z}_2 \text{ for some fixed } c \in \mathbb{Z} \right\}$$

and define c_{n+1} to be the unique integer such that

$$c_{n+1} \equiv 4f\left(\frac{c_n + 2^{j_n} p_n}{4}\right) \pmod{2^{j_{n+1}}}$$

and $0 \le c_{n+1} < 2^{j_{n+1}}$; otherwise terminate $\{c_n\}$ and $\{j_n\}$ at index n. Then for $n \in \mathbb{Z}_{>0}$ for which c_n and j_n are defined, we have:

(a) If $j_n > 0$ and $c_n \ne 0$ then $|f^n(\alpha)| = 2^{2-e}$, where $e \in \mathbb{Z}_{\ge 0}$ such that $e + 1 \le j_n$ and $c_n \equiv 2^e \pmod{2^{e+1}}$.

(b) If $j_n > 0$ and c_n is odd then $f \notin \mathcal{M}_{3,2}$.

(c) If $j_n \ge 3$ and $c_n \equiv 0 \pmod 8$ then $f \in \mathcal{M}_{3,2}$.

(d) If $c_n = c_{n+\ell}$ and $j_n = j_{n+\ell}$ for some $\ell \in \mathbb{Z}_{>0}$ then $f \in \mathcal{M}_{3,2}$.

We implemented Algorithm 4.7 in Mathematica [Bate et al. 2018]. To do so, we fixed an upper bound $N \in \mathbb{Z}_{>0}$ and computed c_n and j_n (when possible) for $1 \le n \le N$. For the computations in this section, we took $N = 50$. Of course, it may happen that we fail to classify f. This happens when $c_n = 0$ and $j_n \le 2$ or when $|f^n(\alpha)| \in \{1, 2\}$ for $1 \le n \le N$. Failure to classify in the former case is often due to the fact that $f_{\alpha_1,\beta_1} \notin \mathcal{M}_{3,2}$ and $f_{\alpha_2,\beta_2} \in \mathcal{M}_{3,2}$ for some $\alpha_1, \alpha_2, \beta_1, \beta_2 \in \mathbb{Z}_2$ such that $\alpha_1 \equiv \alpha_2 \equiv a_0$ $\pmod{2^j}$ and $\beta_1 \equiv \beta_2 \equiv b_0 \pmod{2^k}$ (i.e., membership of $f_{\alpha,\beta}$ in $\mathcal{M}_{3,2}$ is ill-defined for $\alpha \equiv a_0 \pmod{2^j}$ and $\beta \equiv b_0 \pmod{2^k}$). In the latter case, we can sometimes gain resolution by choosing larger N, but doing so may be futile since there may be $f \in \mathcal{M}_{3,2}$ with $|f^n(\alpha)| \in \{1, 2\}$ for all $n \in \mathbb{Z}_{>0}$ which do not have preperiodicity in $\{|f^n(\alpha)|\}$ detectable by (4-9) or simply lack any preperiodicity at all.

We performed this computation for $\alpha, \beta \pmod{2^7}$; that is, we performed this computation for $j = k = 7$ and all $a_0, b_0 \in \mathbb{Z}$ with a_0 is odd, b_0 is even, and $0 \le a_0, b_0 < 2^7$. Figure 1 depicts these results for $0 \le a_0, b_0 < 30$. In Figure 1, the first column lists values for a_0, and the first row lists values for b_0. For given a_0 and b_0, the corresponding entry in the table is colored red if $f_{\alpha,\beta} \notin \mathcal{M}_{3,2}$, green if $f_{\alpha,\beta} \in \mathcal{M}_{3,2}$, and white if we cannot determine whether $f_{\alpha,\beta}$ is in $\mathcal{M}_{3,2}$.

a_k/b_k	0	2	4	6	8	10	12	14	16	18	20	22	24	26	28
1															
3															
5															
7															
9															
11															
13															
15															
17															
19															
21															
23															
25															
27															
29															

Figure 1. Partial analysis of $\mathcal{M}_{3,2}$ for α, β (mod 2^7).

Figure 1 suggests that the classification of $f_{\alpha,\beta}$ is identical to the classification of $f_{\alpha,0}$ for any α, β (mod 2^7). For $N = 50$ this is indeed the case, and appears to hold for larger N as well. Similar analysis for α, β (mod 2^j) for $j \leq 9$ reveals the same type of conformity, but for $j \geq 10$ this pattern ceases. As an example of this, consider Figure 2, where we give a partial depiction of the results of this computation for α, β (mod 2^{15}) (we restrict to $0 \leq a_0, b_0 < 110$). As before, an entry is colored red if $f_{\alpha,\beta} \notin \mathcal{M}_{3,2}$ and white if we cannot determine whether $f_{\alpha,\beta}$ is in $\mathcal{M}_{3,2}$. But instead of simply coloring an entry green if $f_{\alpha,\beta} \in \mathcal{M}_{3,2}$, we color it cyan (light blue) if $\lim_{n\to\infty} f_{\alpha,\beta}^n(\alpha) = 0$, yellow if $|f_{\alpha,\beta}^n(\alpha)| = \frac{1}{2}$ for all $n \geq M$ for some $M \in \mathbb{Z}_{>0}$, and blue if $\{|f_{\alpha,\beta}^n(\alpha)|\}$ exhibits the preperiodicity detected by (4-9). Delineation between these first two cases can be achieved by applying Algorithm 4.7(a) and Lemma 4.4.

Figure 2 suggests that if $\lim_{n\to\infty} f_{\alpha,0}^n(\alpha) = 0$ then $\lim_{n\to\infty} f_{\alpha,\beta}^n(\alpha) = 0$ for all β (mod 2^{15}). It also suggests that if $|f_{\alpha,0}^n| = \frac{1}{2}$ for all $n \geq M$, for some $M \in \mathbb{Z}_{>0}$, then the same is true for $f_{\alpha,\beta}$ for all β (mod 2^{15}). We know of no evidence that shows
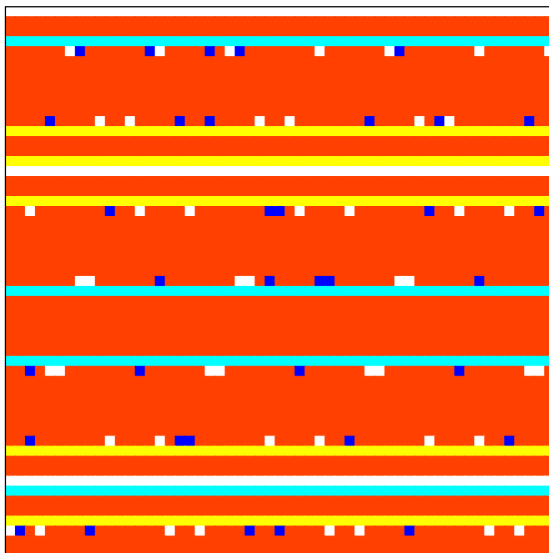
Figure 2. Partial analysis of $\mathcal{M}_{3,2}$ for α, β (mod 2^{15}).

these observations do not persist in general. The differences which arise between $f_{\alpha,\beta}$ and $f_{\alpha,0}$ in Figure 2 occur in part because in some cases $f_{\alpha,\beta} \in \mathcal{M}_{3,2}$ due to (4-9) when $f_{\alpha,0} \notin \mathcal{M}_{3,2}$ (and vice versa). There exist other discrepancies between $f_{\alpha,\beta}$ and $f_{\alpha,0}$ due to the failure of our algorithm to find a classification. As mentioned earlier, we suspect these failures are in part due to a form of preperiodicity in $\{|f^n(\alpha)|\}$ not detected by (4-9). We will discuss other means of detecting preperiodicity, at least for $\{|f_{\alpha,0}^n(\alpha)|\}$, in Theorem 5.7 and the subsequent discussion.

Having discussed the case where α, $\beta \in \mathbb{Z}_2$ with $\alpha + \beta$ odd, we now consider, as promised, the case where $\alpha = a/2$, $\beta = b/2$ for odd $a, b \in \mathbb{Z}_2$ such that $a + b \equiv 0 \pmod{4}$.

Lemma 4.8. *Let $f = f_{\alpha,\beta}$ with $\alpha = a/2$, $\beta = b/2$, and odd $a, b \in \mathbb{Z}_2$ such that $a + b \equiv 0 \pmod{4}$. If $x \in \mathbb{Z}_2$ then $|f(x)| = 4|x|$.*

Proof. Let $x \in \mathbb{Z}_2$. Thus $|x| = 2^{-k}$ for some $k \in \mathbb{Z}_{\geq 0}$. Since

$$\left| x^2 - \frac{3(a+b)}{4}x \right| \leq 1 \quad \text{and} \quad |3\alpha\beta| = \left| \frac{3ab}{4} \right| = 4,$$

by (2-1)

$$|f(x)| = |x|\left| x^2 - \frac{3(a+b)}{4}x + \frac{3ab}{4} \right| = 4|x|. \qquad \square$$

The following proposition follows from Lemmas 4.8 and 4.3.

Proposition 4.9. *Let $f = f_{\alpha,\beta}$ with $\alpha = a/2$, $\beta = b/2$, and odd $a, b \in \mathbb{Z}_2$ such that $a + b \equiv 0 \pmod 4$. If $|f^m(\alpha)| = 4^k$ for some $m \in \mathbb{Z}_{>0}$ and $k \in \mathbb{Z}$ then $\{f^n(\alpha)\}$ is unbounded; hence $f \notin \mathcal{M}_{3,2}$.*

One might hope that we can obtain results similar to Algorithm 4.7 for $\alpha = a/2$, $\beta = b/2$ for odd $a, b \in \mathbb{Z}_2$ such that $a + b \equiv 0 \pmod 4$. However, the definition and resulting analysis of $\{c_n\}$ and $\{j_n\}$ given in Algorithm 4.7 fundamentally depend upon Proposition 4.6. By Lemma 4.8 we cannot have a nice analogue of Proposition 4.6(b), although Proposition 4.9 is an analogue of Proposition 4.6(a). Hence any algorithm that classifies such $f_{\alpha,\beta}$ would likely deviate from Algorithm 4.7 in some significant ways. We will not pursue this matter in this paper.

5. Analysis of $\mathcal{M}_{3,2} \cap \{f_{\alpha,0} : \alpha \in \mathbb{Q}_2\}$

In this section we consider $\mathcal{M}_{3,2} \cap \{f_{\alpha,0} : \alpha \in \mathbb{Q}_2\}$, where

$$f(x) = f_{\alpha,0}(x) = x^2\left(x - \frac{3\alpha}{2}\right) \in \mathcal{P}_{3,2}. \tag{5-1}$$

Restricting our attention to this set is worthwhile since, as was pointed out in Section 4, the dynamics of $f = f_{\alpha,0}$ appear representative of $f_{\alpha,\beta}$ for $\alpha \pmod{2^j}$ and $\beta \pmod{2^k}$. Although we could adapt Algorithm 4.7 slightly to analyze $\{f^n(\alpha)\}$ for fixed $\alpha \pmod{2^j}$, we will instead use Algorithm 5.3 which is more computationally efficient and produces results that bring greater clarity to the structure of $\mathcal{M}_{3,2} \cap \{f_{\alpha,0} : \alpha \in \mathbb{Q}_2\}$.

Anderson's critical radius bound, mentioned at the start of Section 4, together with Proposition 4.2(a,c) show that $f_{\alpha,0} \notin \mathcal{M}_{3,2}$ for $|\alpha| > 1$ and $f_{\alpha,0} \in \mathcal{M}_{3,2}$ for $|\alpha| < 1$. Therefore we restrict our attention to odd $\alpha \in \mathbb{Z}_2$. Rather than focusing on $\{|f^n(\alpha)|\}$ directly, we will instead analyze the sequence $\{x_n\}$ defined in the following lemma, which will serve as a proxy for our study of $\{|f^n(\alpha)|\}$.

Lemma 5.1. *Let*

$$x_n = -\frac{f^{n+1}(\alpha)}{2\alpha f^n(\alpha)^2} \tag{5-2}$$

for $n \in \mathbb{Z}_{>0}$. Then $x_1 = (3 + \alpha^2)/4$ and

$$4x_n = \alpha^2 x_{n-1}(4x_{n-1} - 3)^2 + 3 \tag{5-3}$$

for $n \in \mathbb{Z}_{>1}$.

Proof. Since $f^{n+1}(\alpha) = f^n(\alpha)^2(f^n(\alpha) - 3\alpha/2)$,

$$x_n = -\frac{1}{2\alpha}\left(f^n(\alpha) - \frac{3\alpha}{2}\right) = \frac{3}{4} - \frac{f^n(\alpha)}{2\alpha}, \tag{5-4}$$

and so in particular,

$$x_1 = \frac{3}{4} - \frac{f(\alpha)}{2\alpha} = \frac{3}{4} - \frac{\alpha^2 \cdot (-\alpha/2)}{2\alpha} = \frac{3+\alpha^2}{4}.$$

For $n \in \mathbb{Z}_{>1}$, we have by (5-4) and (5-1) that

$$4x_n = 3 - \frac{2}{\alpha} f^n(\alpha) = 3 - \frac{2}{\alpha}\left(f^{n-1}(\alpha)^2\left(f^{n-1}(\alpha) - \frac{3\alpha}{2}\right)\right)$$

$$= 3 - \frac{2}{\alpha} f^{n-1}(\alpha)^3 + 3 f^{n-1}(\alpha)^2 = 3 - \alpha^2\left(\frac{2 f^{n-1}(\alpha)}{\alpha}\right)^2\left(\frac{f^{n-1}(\alpha)}{2\alpha} - \frac{3}{4}\right)$$

$$= \alpha^2\left(\frac{3}{4} - \frac{f^{n-1}(\alpha)}{2\alpha}\right)\left(\frac{2 f^{n-1}(\alpha)}{\alpha}\right)^2 + 3 = \alpha^2 x_{n-1}(4x_{n-1} - 3)^2 + 3. \qquad \square$$

Suppose

$$|f^k(\alpha)| \in \{1, 2\} \quad \text{for } 1 \le k \le n, \tag{5-5}$$

for some $n \in \mathbb{Z}_{>0}$. Such n must exist for the simple reason that by (5-1), $|f(\alpha)| = 2$. By (5-1),

$$|f^k(\alpha)| = 1 \quad \Longrightarrow \quad |f^{k+1}(\alpha)| = 2; \tag{5-6}$$

hence $|f^k(\alpha)| = |f^{k+1}(\alpha)| = 1$ is impossible. Thus by (5-2),

$$|x_k| = 2\left|\frac{f^{k+1}(\alpha)}{f^k(\alpha)^2}\right| = \begin{cases} 4 & \text{if } |f^k(\alpha)| = 1, \ |f^{k+1}(\alpha)| = 2, \\ \frac{1}{2} & \text{if } |f^k(\alpha)| = 2, \ |f^{k+1}(\alpha)| = 1, \\ 1 & \text{if } |f^k(\alpha)| = 2, \ |f^{k+1}(\alpha)| = 2 \end{cases} \tag{5-7}$$

for $k < n$. We wish to describe the possible values for $|x_n|$. If $|f^{n+1}(\alpha)| \in \{1, 2\}$ then by the same reasoning as in (5-7), we know that $|x_n| \in \{\frac{1}{2}, 1, 4\}$. If $|f^{n+1}(\alpha)| \ge 4$ then in actuality $|f^{n+1}(\alpha)| = 4$. To see why this is the case, notice that if $|f^n(\alpha)| \le 2$ then by (5-1) $|f^{n+1}(\alpha)| \le 4$. If $|f^{n+1}(\alpha)| = 4$ then by (5-6) $|f^n(\alpha)| = 2$, and so

$$|x_n| = 2\left|\frac{f^{n+1}(\alpha)}{f^n(\alpha)^2}\right| = 2. \tag{5-8}$$

Likewise, if $|f^{n+1}(\alpha)| \le \frac{1}{2}$ then by (5-6), $|f^n(\alpha)| = 2$, and so

$$|x_n| = 2\left|\frac{f^{n+1}(\alpha)}{f^n(\alpha)^2}\right| \le \frac{1}{4}. \tag{5-9}$$

These observations, along with Proposition 4.6, prove the following:

Proposition 5.2. *Let* $f = f_{\alpha,0}$ *with odd* $\alpha \in \mathbb{Z}_2$. *Let* $\{x_n\}$ *be as defined in* (5-2). *Let* $n \in \mathbb{Z}_{>0}$ *such that* (5-5) *holds, or equivalently, such that* $|x_k| \in \{\frac{1}{2}, 1, 4\}$ *for* $1 \le k < n$:

(a) *If* $|x_n| = 2$ *then* $f \notin \mathcal{M}_{3,2}$.

(b) *If* $|x_n| \le \frac{1}{4}$ *then* $f \in \mathcal{M}_{3,2}$.

We also have the following converse to Proposition 5.2(a): if $f = f_{\alpha,0} \notin \mathcal{M}_{3,2}$ then there exists $n \in \mathbb{Z}_{>0}$ such that $|x_k| \in \{\frac{1}{2}, 1, 4\}$ for $k < n$ and $|x_n| = 2$; this follows from (5-7) and (5-8) since if $f_{\alpha,0} \notin \mathcal{M}_{3,2}$ then there exists $n \in \mathbb{Z}_{>0}$ such that $|f^k(\alpha)| \in \{1, 2\}$ for $k \le n$ and $|f^{n+1}(\alpha)| = 4$. Therefore since x_n is only dependent upon α^2 by Lemma 5.1, we have then that either both $f_{\alpha,0}, f_{-\alpha,0} \notin \mathcal{M}_{3,2}$ or both $f_{\alpha,0}, f_{-\alpha,0} \in \mathcal{M}_{3,2}$. Because of this, Algorithm 5.3 (which we are about to describe) will consider $\alpha^2 \pmod{2^m}$ for fixed $m \in \mathbb{Z}_{>0}$ rather than $\alpha \pmod{2^j}$ for fixed $j \in \mathbb{Z}_{>0}$

Fix $m \in \mathbb{Z}_{>0}$ and fix odd $d \in \mathbb{Z}$ such that d is a quadratic residue modulo 2^m and $0 \le d < 2^m$; it is well known that d is a quadratic residue modulo 2^m if and only if $d \equiv 1 \pmod 8$. Consider odd $\alpha \in \mathbb{Z}_2$ such that $\alpha^2 \equiv d \pmod{2^m}$, or equivalently, $\alpha^2 = d + 2^m p$, where $p \in \mathbb{Z}_2$ is indeterminate. Let

$$g(x) = \tfrac{1}{4}(\alpha^2 x(4x - 3)^2 + 3). \tag{5-10}$$

By Lemma 5.1, $x_n = g(x_{n-1}) = g^{n-1}(x_1)$, where $x_1 = (3 + \alpha^2)/4$. Recall that in Section 4, we defined sequences $\{c_n\}$ and $\{j_n\}$ from which we could often determine the value of $|f^n(\alpha)|$. We employ the same approach here, defining (possibly finite) sequences $\{d_n\}$ and $\{m_n\}$ in a completely analogous way so as to determine the value of $|x_n| = |g^{n-1}(x_1)|$. Since we covered the aforementioned case in detail, we will give a more abbreviated treatment here, trusting that the reader has a solid grasp of our conventions regarding indeterminates.

We recursively define $\{d_n\}$ and $\{m_n\}$ as follows. Suppose $4g((d_n + 2^{m_n} p_n)/4) \in \mathbb{Z}_2$; this is true for $n = 1$ if we let $d_1 = 3 + d$, $m_1 = m$, and $p_1 = p$. Let

$$m_{n+1} = \max\left\{ m' \in \mathbb{Z}_{\ge 0} : 4g\left(\frac{d_n + 2^{m_n} p_n}{4}\right) \equiv d' \pmod{2^{m'}} \atop \text{for some fixed } d' \in \mathbb{Z} \right\}. \tag{5-11}$$

Thus $4g((d_n + 2^{m_n} p_n)/4) = d_{n+1} + 2^{m_{n+1}} p_{n+1}$ for indeterminate $p_{n+1} \in \mathbb{Z}_2$ and unique $d_{n+1} \in \mathbb{Z}$ such that $0 \le d_{n+1} < 2^{m_{n+1}}$. One can then show that

$$4x_{n+1} = 4g\left(\frac{d_n + 2^{m_n} p_n}{4}\right) = d_{n+1} + 2^{m_{n+1}} p_{n+1}. \tag{5-12}$$

If $4g((d_n + 2^{m_n} p_n)/4) \notin \mathbb{Z}_2$ then we terminate $\{d_n\}$ and $\{m_n\}$ at index n.

If $m_n > 0$ and $d_n \ne 0$, then there exists $e \in \mathbb{Z}_{\ge 0}$ such that $d_n \equiv 2^e \pmod{2^{e+1}}$ and $e + 1 \le m_n$. Just as in (4-7),

$$d_n \equiv 2^e \pmod{2^{e+1}} \quad \Longleftrightarrow \quad |x_n| = 2^{2-e}. \tag{5-13}$$

Thus by Proposition 5.2, $f \notin \mathcal{M}_{3,2}$ if $e = 1$ and $f \in \mathcal{M}_{3,2}$ if $e \ge 4$. We terminate $\{d_n\}$ and $\{m_n\}$ at index n in both of these cases. If $d_n = 0$ then just as in (4-8),

$$d_n \equiv 0 \pmod{2^{m_n}} \quad \Longrightarrow \quad |x_n| \le 2^{2-m_n}. \tag{5-14}$$

Therefore by Proposition 5.2(b), if $m_n \geq 4$ and $d_n = 0$ then $f \in \mathcal{M}_{3,2}$. If $m_n \leq 3$ and $d_n = 0$ then we cannot determine the value of $|x_n|$. Since we use Proposition 5.2 to determine whether f is in $\mathcal{M}_{3,2}$ and since the hypothesis of that proposition requires us to know $|x_k| \in \left\{\frac{1}{2}, 1, 4\right\}$ for all $k < n$ we can never afford to be ignorant of the value of $|x_n|$ when it comes to classifying f via $|x_{n+1}|$. In light of this, we terminate $\{d_n\}$ and $\{m_n\}$ at index n whenever $d_n = 0$.

The additional termination criteria that we just gave shows that we only compute d_{n+1} and m_{n+1} if

(i) $m_n \geq 1$ and $d_n \equiv 1 \pmod 2$ (i.e., $e = 0$),

(ii) $m_n \geq 3$ and $d_n \equiv 4 \pmod 8$ (i.e., $e = 2$), or

(iii) $m_n \geq 4$ and $d_n \equiv 8 \pmod{16}$ (i.e., $e = 3$).

One can show that in all of these cases, $4g((d_n + 2^{m_n} p_n)/4) \in \mathbb{Z}_2$. Thus the additional termination criteria makes it so we never need to check if $4g((d_n + 2^{m_n} p_n)/4) \in \mathbb{Z}_2$.

The preperiodicity condition (4-8) has the following analogue for d_n and m_n:

$$\text{if } d_n = d_{n+\ell} \text{ and } m_n = m_{n+\ell} \text{ for some } n, \ell \in \mathbb{Z}_{>0} \text{ then } f \in \mathcal{M}_{3,2}. \qquad (5\text{-}15)$$

To see why this is true, first note that if $d_n = d_{n+\ell}$ and $m_n = m_{n+\ell}$ for some n, $\ell \in \mathbb{Z}_{>0}$ then $\{(d_n, m_n)\}$ is preperiodic. The termination criteria given above guarantees that each d_n and m_n satisfy either (i), (ii), or (iii). Thus by (5-13), if $\{(d_n, m_n)\}$ is preperiodic then $\{|x_n|\}$ is also preperiodic, with $|x_n| \in \left\{\frac{1}{2}, 1, 4\right\}$. Therefore by (5-7), $\{|f^n(\alpha)|\}$ is preperiodic.

We summarize the discussion above in the following result.

Algorithm 5.3. Fix $m \in \mathbb{Z}_{>0}$ and fix $d \in \mathbb{Z}$ such that $d \equiv 1 \pmod 8$ and $0 \leq d < 2^m$. Consider $\alpha \in \mathbb{Z}_2$ such that $\alpha^2 \equiv d \pmod{2^m}$. Let $f = f_{\alpha,0}$, $d_1 = 3 + d$ and $m_1 = m$. We recursively define sequences $\{d_n\}$ and $\{m_n\}$ as follows. If

(i) $m_n \geq 1$ and $d_n \equiv 1 \pmod 2$ (i.e., $e = 0$),

(ii) $m_n \geq 3$ and $d_n \equiv 4 \pmod 8$ (i.e., $e = 2$), or

(iii) $m_n \geq 4$ and $d_n \equiv 8 \pmod{16}$ (i.e., $e = 3$),

define

$$m_{n+1} = \max\left\{ m' \in \mathbb{Z}_{\geq 0} : 4g\left(\frac{d_n + 2^{m_n} p_n}{4}\right) \equiv d' \pmod{2^{m'}} \right.$$
$$\left. \text{for all } p_n \in \mathbb{Z}_2 \text{ for some fixed } d' \in \mathbb{Z} \right\}$$

and define d_{n+1} to be the unique integer such that

$$d_{n+1} \equiv 4g\left(\frac{d_n + 2^{m_n} p_n}{4}\right) \pmod{2^{m_{n+1}}}$$

and $0 \leq d_{n+1} < 2^{m_{n+1}}$; otherwise terminate $\{d_n\}$ and $\{m_n\}$ at index n. Then for $n \in \mathbb{Z}_{>0}$ for which d_n and m_n are defined, we have:

(a) If $m_n > 0$ and $d_n \neq 0$ then $|x_n| = 2^{2-e}$, where $e \in \mathbb{Z}_{\geq 0}$ such that $e + 1 \leq m_n$ and $d_n \equiv 2^e \pmod{2^{e+1}}$.

(b) If $m_n \geq 2$ and $d_n \equiv 2 \pmod 4$ then $f \notin \mathcal{M}_{3,2}$.

(c) If $m_n \geq 4$ and $d_n \equiv 0 \pmod{16}$ then $f \in \mathcal{M}_{3,2}$.

(d) If $d_n = d_{n+\ell}$ and $m_n = m_{n+\ell}$ for some $\ell \in \mathbb{Z}_{>0}$ then $f \in \mathcal{M}_{3,2}$.

As we did with Algorithm 4.7, we implemented Algorithm 5.3 in Mathematica [Bate et al. 2018]. In the following computations, we computed d_n and m_n (when possible) for $1 \leq n \leq N$ with $N = 50$. As before, it may happen that our algorithm fails to classify f for the same reasons mentioned in Section 4.

It is well known that

$$\{\alpha^2 : \text{odd } \alpha \in \mathbb{Z}_2\} = \{x \in \mathbb{Z}_2 : x \equiv 1 \pmod 8\}. \tag{5-16}$$

If $x \equiv 1 \pmod 8$, then either $x \equiv 1 \pmod{16}$ or $x \equiv 1 + 8 \equiv 9 \pmod{16}$. More generally, if $x \equiv d \pmod{2^m}$ then either $x \equiv d \pmod{2^{m+1}}$ or $x \equiv d + 2^m \pmod{2^{m+1}}$. Topologically speaking, we are simply asserting that the disk

$$D(d, 2^{-m}) = \{x \in \mathbb{Q}_p : |x - d| \leq 2^{-m}\} = \{x \in \mathbb{Q}_p : x \equiv d \pmod{2^m}\} \tag{5-17}$$

decomposes into the disjoint union of $D(d, 2^{-m-1})$ and $D(d + 2^m, 2^{-m-1})$. These disks form a partial order under subset inclusion which can be visualized as a binary tree with $\{\alpha^2 : \text{odd } \alpha \in \mathbb{Z}_2\} = D(1, 2^{-3})$ as the root vertex, $D(1, 2^{-4})$ and $D(9, 2^{-4})$ as its child vertices, and so forth. We can think of Algorithm 5.3 as providing an algorithm for (possibly) determining if a given disk $D(d, 2^{-m})$ is contained entirely within or is entirely disjoint from $\mathcal{M}_{3,2}$.

In Figure 3 we display the results of this algorithm, along with supplemental classification afforded by Lemma 4.4 and Theorem 5.7, out to disks of radius 2^{-11}. A vertex with corresponding disk D is colored

(i) white if $D \cap \mathcal{M}_{3,2} \neq \varnothing$ and $D \not\subset \mathcal{M}_{3,2}$,

(ii) red if $D \cap \mathcal{M}_{3,2} = \varnothing$,

(iii) cyan if for all $\alpha \in D$ we have $\lim_{n \to \infty} f_{\alpha,0}^n(\alpha) = 0$,

(iv) yellow if for each $\alpha \in D$ there exists $M \in \mathbb{Z}_{>0}$ such that $|f_{\alpha,0}^n(\alpha)| = \frac{1}{2}$ for all $n \geq M$,

(v) green if D decomposes into disks of types (iii) and (iv),

(vi) blue if the preperiodicity condition (5-15) is satisfied by $f_{\alpha,0}$ for all $\alpha \in D$,
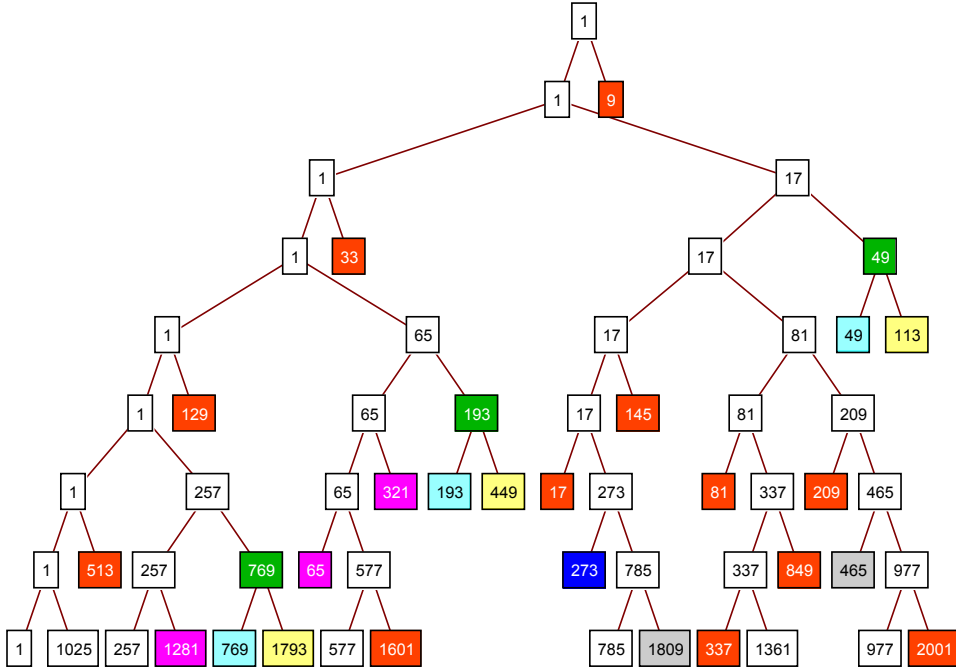
Figure 3. Analysis of $f_{\alpha,0}$ for $\alpha^2 \pmod{2^m}$ for $3 \le m \le 11$.

(vii) magenta if $D \subset D(1, 2^{-5})$ and if for each $\alpha \in D$ there exists $n \in \mathbb{Z}_{>0}$ such that $|f^n(\alpha)| = |f^{n+2}(\alpha)| = 2$ and $|f^{n+1}(\alpha)| = |f^{n+3}(\alpha)| = 1$ for $f = f_{\alpha,0}$ (in such a case, Theorem 5.7 proves $D \subset \mathcal{M}_{3,2}$), and

(viii) gray if we are unable to determine if $D \subset \mathcal{M}_{3,2}$, if $D \cap \mathcal{M}_{3,2} = \varnothing$, or if $D \cap \mathcal{M}_{3,2} \neq \varnothing$ and $D \not\subset \mathcal{M}_{3,2}$ (for the given N).

Disks of types (iii)–(vii) are all contained in $\mathcal{M}_{3,2}$, while disks of type (ii) are disjoint from $\mathcal{M}_{3,2}$. Disks of type (viii) occur when our algorithm is unable to classify a disk or any of the disks that it decomposes into. We can distinguish between disks of types (iii) and (iv) using Lemma 4.4 since we can determine $|f^n(\alpha)|$ from $|x_n|$. Indeed, by (5-2),

$$|f^{n+1}(\alpha)| = \tfrac{1}{2} |x_n||f^n(\alpha)^2|. \tag{5-18}$$

We also recognize disks of type (vii) by examining $\{|f^n(\alpha)|\}$. As an aside, we mention that to our knowledge, there are no disks that are of both type (vi) and (vii). This indicates that the preperiodicity detected by Theorem 5.7 is of a subtler form than that of (5-15); we will discuss this distinction in more detail after Theorem 5.7.

Figure 4 shows the results of the computation above out to disks of radius 2^{-19}. There exists an obvious pattern of disks of types (ii), (iii), (iv), and (vii) along the left
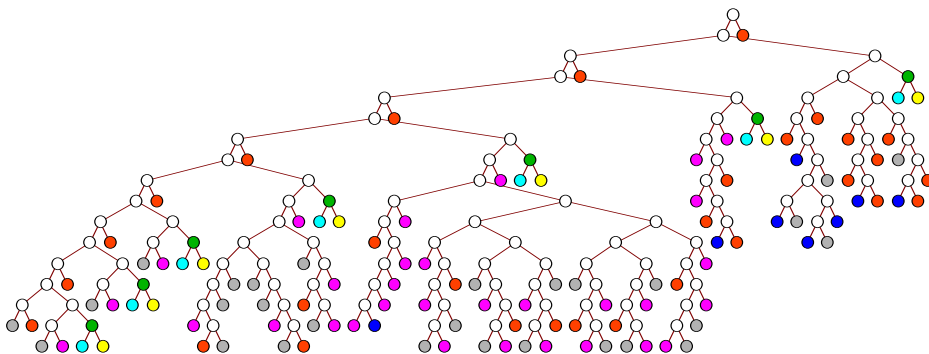
Figure 4. Analysis of $f_{\alpha,0}$ for $\alpha^2 \pmod{2^m}$ for $3 \leq m \leq 20$.

side of the tree. We prove that this pattern persists indefinitely in Theorem 5.5. As was pointed out in [Anderson 2013, §6], this pattern shows that the boundary of $\mathcal{M}_{3,2}$ at $\alpha = 1$ (and hence also $\alpha = -1$) has a degree of self-similarity. The disks of types (ii), (vi), and (vii) in the remainder of Figure 4 seem to indicate that there are other forms of self-similarity, although their exact characteristics seem difficult to quantify.

In Table 1 we list the values of $\{|f^n(\alpha)|\}$ obtained via Algorithm 5.3(a) and (5-18) for $\alpha^2 \pmod{2^m}$ for $3 \leq m \leq 11$. Table 1 has the following conventions:

- A listing terminates at index n if $|f^n(\alpha)| \geq 4$ or $|f^n(\alpha)| \leq \frac{1}{2}$.

- If we know that $|f^n(\alpha)| \leq \frac{1}{2}$ and nothing more, a listing terminates with $\frac{1}{2}^*$ at index n. A similar convention is used for $|f^n(\alpha)| \leq \frac{1}{4}$.

- A listing ending with $[q_1 q_2 \cdots q_m]$ indicates $\{|f^n(\alpha)|\}$ has $q_1 q_2 \cdots q_m$ repeating indefinitely.

- A listing containing $(q_1 q_2 \cdots q_m)_k$ indicates $q_1 q_2 \cdots q_m$ repeats itself k times in $\{|f^n(\alpha)|\}$.

- If upon computing $\{|f_{\alpha_1,0}^n(\alpha_1)|\}$ and $\{|f_{\alpha_2,0}^n(\alpha_2)|\}$ for $\alpha_1^2 \equiv \alpha^2 \pmod{2^{m+1}}$ and $\alpha_2^2 \equiv \alpha^2 + 2^m \pmod{2^{m+1}}$ we find that $\{|f_{\alpha_1,0}^n(\alpha_1)|\}$ and $\{|f_{\alpha_2,0}^n(\alpha_2)|\}$ have in common a sequence which is longer than that obtained by computing $\{|f_{\alpha,0}^n(\alpha)|\}$, we give instead the sequence shared in common by $\{|f_{\alpha_1,0}^n(\alpha_1)|\}$ and $\{|f_{\alpha_2,0}^n(\alpha_2)|\}$ for $\{|f_{\alpha,0}^n(\alpha)|\}$. This means of determining $\{|f_{\alpha,0}^n(\alpha)|\}$ is applied recursively with the disks of radius 2^{-19} serving as the terminating cases.

Notice that the entries in Table 1 correspond to the vertices in Figure 3.[3] Of particular interest are the disks which our algorithm is unable to classify. In Figure 3 these occur for $D(465, 2^{-10})$ and $D(1809, 2^{-11})$. From Table 1, it seems very likely that $\{|f^n(\alpha)|\}$ is preperiodic for $\alpha^2 \in D(465, 2^{-10})$, although we do not present a

---

[3]The *Mathematica* notebook in [Bate et al. 2018] contains values of $|f^n(\alpha)|$ for vertices in Figure 4.

$\alpha^2 \pmod{2^m}$	$\lvert f^n(\alpha)\rvert$	$\alpha^2 \pmod{2^m}$	$\lvert f^n(\alpha)\rvert$
$1 \pmod{2^3}$	22	$1 \pmod{2^{10}}$	22222
		$513 \pmod{2^{10}}$	222224
$1 \pmod{2^4}$	22	$257 \pmod{2^{10}}$	222212
$9 \pmod{2^4}$	224	$769 \pmod{2^{10}}$	$2222\frac{1}{2}{}^{*}$
		$65 \pmod{2^{10}}$	22212[21]
$1 \pmod{2^5}$	222	$577 \pmod{2^{10}}$	2221222
$17 \pmod{2^5}$	22	$273 \pmod{2^{10}}$	[221]
		$785 \pmod{2^{10}}$	$(221)_3 212212$
$1 \pmod{2^6}$	222	$337 \pmod{2^{10}}$	221212212122
$33 \pmod{2^6}$	2224	$849 \pmod{2^{10}}$	221212212224
$17 \pmod{2^6}$	2212	$465 \pmod{2^{10}}$	$22121(221)_{11}2$
$49 \pmod{2^6}$	$22\frac{1}{2}{}^{*}$	$977 \pmod{2^{10}}$	221212212122
$1 \pmod{2^7}$	2222	$1 \pmod{2^{11}}$	222222
$65 \pmod{2^7}$	222	$1025 \pmod{2^{11}}$	22222
$17 \pmod{2^7}$	22122	$257 \pmod{2^{11}}$	2222122
$81 \pmod{2^7}$	2212122	$1281 \pmod{2^{11}}$	222[21]
$49 \pmod{2^7}$	$22\frac{1}{4}{}^{*}$	$769 \pmod{2^{11}}$	$2222\frac{1}{4}{}^{*}$
$113 \pmod{2^7}$	$22\left[\frac{1}{2}\right]$	$1793 \pmod{2^{11}}$	$2222\left[\frac{1}{2}\right]$
		$577 \pmod{2^{11}}$	222122212
$1 \pmod{2^8}$	2222	$1601 \pmod{2^{11}}$	222122224
$129 \pmod{2^8}$	22224	$785 \pmod{2^{11}}$	$(22122122121)_4 22122122$
$65 \pmod{2^8}$	22212	$1809 \pmod{2^{11}}$	22122122121221212212
$193 \pmod{2^8}$	$222\frac{1}{2}{}^{*}$	$337 \pmod{2^{11}}$	22121221212224
$17 \pmod{2^8}$	2212212	$1361 \pmod{2^{11}}$	$(22121)_3 22$
$145 \pmod{2^8}$	2212224	$977 \pmod{2^{11}}$	$(22121)_3 22$
$81 \pmod{2^8}$	2212122	$2001 \pmod{2^{11}}$	22121221212224
$209 \pmod{2^8}$	2212122		
$1 \pmod{2^9}$	22222		
$257 \pmod{2^9}$	2222		
$65 \pmod{2^9}$	222122		
$321 \pmod{2^9}$	22[21]		
$193 \pmod{2^9}$	$222\frac{1}{4}{}^{*}$		
$449 \pmod{2^9}$	$222\left[\frac{1}{2}\right]$		
$17 \pmod{2^9}$	221221212224		
$273 \pmod{2^9}$	2212212212		
$81 \pmod{2^9}$	221212224		
$337 \pmod{2^9}$	221212212		
$209 \pmod{2^9}$	221212224		
$465 \pmod{2^9}$	221212212		

Table 1. Values of $\lvert f^n(\alpha)\rvert$ for $\alpha^2 \pmod{2^m}$.

proof for this. In contrast, there appears to be no such preperiodicity in $\{|f^n(\alpha)|\}$ for $\alpha^2 \in D(1809, 2^{-11})$. In fact, closer inspection of α^2 in smaller disks within $D(1809, 2^{-11})$ results in $\{|f^n(\alpha)|\}$ which are composed of blocks of 21 and 221, but with no apparent preperiodicity emerging. We are uncertain whether such preperiodicity detection is simply beyond our computational limits or whether there is no preperiodicity to be found at all.

In addition to these observations, there are two interesting patterns displayed in Table 1 that are worth pointing out. The first is that if $\{|f^n(\alpha)|\}$ begins with a block of k 2's followed by a 1, then if a block of k 2's reappears later in the sequence, $f \notin \mathcal{M}_{3,2}$. The second is that if at any point $|f^n(\alpha)| = 1$ then $\lim_{n\to\infty} f^n(\alpha) \neq 0$. All numerical evidence we've seen confirms these two observations, but we've been unable to prove either.

The following lemma is used in the proof of Theorem 5.5.

Lemma 5.4. *Let $\alpha \in \mathbb{Z}_2$ odd and $f = f_{\alpha,0}$. If $\alpha^2 \equiv 1 + c \cdot 2^{2n} \pmod{2^{2n+\ell}}$ for $n \in \mathbb{Z}_{\geq 2}$, $c \in \mathbb{Z}$, and $\ell = 1, 2, 3$, then*

$$x_k \equiv 1 + c \cdot 5 \cdot 2^{2(n-k)} \pmod{2^{2(n-k)+\ell}} \qquad (5\text{-}19)$$

for all k such that $2 \leq k \leq n$. Recall x_k is defined in (5-2).

Proof. Since $\alpha^2 \equiv 1 + c \cdot 2^{2n} \pmod{2^{2n+\ell}}$, we have $\alpha^2 \equiv 1 + c \cdot 2^{2n} + d \cdot 2^{2n+\ell}$ for some $d \in \mathbb{Z}_2$. Thus by Lemma 5.1,

$$x_1 = \frac{3 + \alpha^2}{4} = \frac{3 + 1 + c \cdot 2^{2n} + d \cdot 2^{2n+\ell}}{4} = 1 + c \cdot 2^{2(n-1)} + d \cdot 2^{2(n-1)+\ell}$$
$$\equiv 1 + c \cdot 2^{2(n-1)} \pmod{2^{2(n-1)+\ell}}. \qquad (5\text{-}20)$$

Observe

$$(4x_1 - 3)^2 \equiv (4 + c \cdot 2^{2(n-1)+2} - 3)^2 \equiv 1 + c \cdot 2^{2(n-1)+3} + c^2 \cdot 2^{4(n-1)+4}$$
$$\equiv 1 \pmod{2^{2(n-1)+\ell}}. \qquad (5\text{-}21)$$

Since $\alpha^2 \equiv 1 + c \cdot 2^{2n} \pmod{2^{2(n-1)+\ell}}$, by Lemma 5.1, (5-20), and (5-21)

$$4x_2 = \alpha^2 x_1 (4x_1 - 3)^2 + 3 \equiv (1 + c \cdot 2^{2n})(1 + c \cdot 2^{2(n-1)}) + 3$$
$$\equiv 4 + c \cdot 2^{2(n-1)} + c \cdot 2^{2n} + c^2 \cdot 2^{2n+2(n-1)} \equiv 4 + c \cdot (1+4)2^{2(n-1)}$$
$$\equiv 4 + c \cdot 5 \cdot 2^{2(n-1)} \pmod{2^{2(n-1)+\ell}};$$

here we used the fact that $2n + 2(n-1) \geq 2(n-1) + \ell$ since $n \geq 2$. Therefore

$$x_2 \equiv 1 + c \cdot 5 \cdot 2^{2(n-2)} \pmod{2^{2(n-2)+\ell}}.$$

Thus (5-19) is satisfied for $k = 2$. Suppose (5-19) holds for some k such that $2 \leq k < n$. Then

$$
\begin{aligned}
(4x_k - 3)^2 &\equiv (1 + c \cdot 5 \cdot 2^{2(n-k)+2})^2 \\
&\equiv 1 + c \cdot 5 \cdot 2^{2(n-k)+3} + c^2 \cdot 5^2 \cdot 2^{4(n-k)+4} \\
&\equiv 1 \pmod{2^{2(n-k)+\ell}}.
\end{aligned}
\tag{5-22}
$$

Since $\alpha^2 \equiv 1 + c \cdot 2^{2n} \equiv 1 \pmod{2^{2(n-k)+\ell}}$ for $k \geq 2$, by Lemma 5.1 and (5-22)

$$
4x_{k+1} = \alpha^2 x_k (4x_k - 3)^2 + 3 \equiv x_k + 3 \equiv 4 + c \cdot 5 \cdot 2^{2(n-k)} \pmod{2^{2(n-k)+\ell}}.
$$

Thus $x_{k+1} \equiv 1 + c \cdot 5 \cdot 2^{2(n-(k+1))} \pmod{2^{2(n-(k+1))+\ell}}$. By induction, (5-19) holds for all k such that $2 \leq k \leq n$. $\square$

Parts (a) and (b) of the following theorem were proved in a somewhat different form in [Anderson 2013, §6]. All available numerical evidence indicates the converses of parts (b) and (c) are true; however a proof of this has remained elusive.

Theorem 5.5. *Let $\alpha \in \mathbb{Z}_2$ odd and $f = f_{\alpha,0}$. Suppose $\alpha^2 \equiv 1 + c \cdot 2^{2n} \pmod{2^{2n+\ell}}$ for $n \in \mathbb{Z}_{\geq 2}$, $c \in \mathbb{Z}$, and $\ell = 1, 2, 3$. Then for all $k \leq n$, we have $|f^k(\alpha)| = 2$. Furthermore*:

(a) *If $\alpha^2 \equiv 1 + 2^{2n+1} \pmod{2^{2n+2}}$ then $f \notin \mathcal{M}_{3,2}$.*

(b) *If $\alpha^2 \equiv 1 + 3 \cdot 2^{2n} \pmod{2^{2n+3}}$ then $\lim_{n \to \infty} f^n(\alpha) = 0$.*

(c) *If $\alpha^2 \equiv 1 + 7 \cdot 2^{2n} \pmod{2^{2n+3}}$ then $|f^{n+j}(\alpha)| = \frac{1}{2}$ for all $j \in \mathbb{Z}_{>0}$.*

(d) *If $\alpha^2 \equiv 1 + 5 \cdot 2^{2n} \pmod{2^{2n+3}}$ for $n \geq 3$ then*

$$
|f^{n+2j}(\alpha)| = 2 \quad and \quad |f^{n+2j+1}(\alpha)| = 1
$$

   for all $j \in \mathbb{Z}_{\geq 0}$.

Proof of Theorem 5.5(a)–(c). By Lemma 5.4, $|x_k| = 1$ for all $k < n$. Since $|f(\alpha)| = 2$, by (5-7) $|f^k(\alpha)| = 2$ for all $k \leq n$.

For part (a), $\alpha^2 \equiv 1 + c \cdot 2^{2n} \pmod{2^{2n+\ell}}$ for $c = 2$ and $\ell = 2$. By Lemma 5.4, $x_n \equiv 3 \pmod 4$; hence $|x_n| = 1$. Furthermore, by Lemma 5.1,

$$
4x_{n+1} = \alpha^2 x_n (4x_n - 3)^2 + 3 \equiv 1 \cdot 3 \cdot 9^2 + 3 \equiv 2 \pmod 4.
$$

Therefore $|x_{n+1}| = 2$. By Proposition 5.2(a), $f \notin \mathcal{M}_{3,2}$.

For part (b), $\alpha^2 \equiv 1 + c \cdot 2^{2n} \pmod{2^{2n+\ell}}$ for $c = 3$ and $\ell = 3$. By Lemma 5.4, $x_n \equiv 0 \pmod 8$. Thus $|x_n| \leq \frac{1}{8}$. Since $|f^n(\alpha)| = 2$, by (5-18) $|f^{n+1}(\alpha)| \leq \frac{1}{4}$. By Lemma 4.4(a), $\lim_{n \to \infty} f^n(\alpha) = 0$.

For part (c), $\alpha^2 \equiv 1 + c \cdot 2^{2n} \pmod{2^{2n+\ell}}$ for $c = 7$ and $\ell = 3$. By Lemma 5.4, $x_n \equiv 4 \pmod 8$. Thus $|x_n| = \frac{1}{4}$. Since $|f^n(\alpha)| = 2$, by (5-18) $|f^{n+1}(\alpha)| = \frac{1}{2}$. By Lemma 4.4(b), $|f^{n+j}(\alpha)| = \frac{1}{2}$ for $j \in \mathbb{Z}_{>0}$. $\square$

We need the following lemma to prove part (d) of Theorem 5.5.

Lemma 5.6. *Let $\alpha \in \mathbb{Z}_2$ such that $\alpha^2 \equiv 1 \pmod{32}$. Let $f = f_{\alpha,0}$. If $x_n \equiv 2 \pmod 8$ then $x_{n+2} \equiv 2 \pmod 8$. Recall x_n is defined in (5-2).*

Proof. Suppose $x_n \equiv 2 \pmod 8$. Then $x_n = 2 + 8c_1$ for some $c_1 \in \mathbb{Z}_2$. By Lemma 5.1,

$$4x_{n+1} = \alpha^2 x_n (4x_n - 3)^2 + 3 \equiv (2 + 8c_1)(8 + 32c_1 - 3)^2 + 3$$
$$\equiv (2 + 8c_1) \cdot 25 + 3 \equiv 21 + 8c_1 \pmod{32}.$$

Thus there exists $c_2 \in \mathbb{Z}_2$ such that $x_{n+1} = (21 + 8c_1 + 32c_2)/4$. By Lemma 5.1,

$$4x_{n+2} = \alpha^2 x_{n+1}(4x_{n+1} - 3)^2 + 3$$
$$= \alpha^2 \tfrac{1}{4}(21 + 8c_1 + 32c_2)(18 + 8c_1 + 32c_2)^2 + 3$$
$$= \alpha^2 (21 + 8c_1 + 32c_2)(9 + 4c_1 + 16c_2)^2 + 3.$$

By expanding and then reducing coefficients modulo 32, we find that

$$(21 + 8c_1 + 32c_2)(9 + 4c_1 + 16c_2)^2 \equiv 5 + 16(1 + c_1)c_1 \equiv 5 \pmod{32};$$

here we used that $(1 + c_1)c_1$ is even. Therefore

$$4x_{n+2} \equiv (21 + 8c_1)(9 + 4c_1 + 16c_2)^2 + 3 \equiv 8 \pmod{32}.$$

Thus $x_{n+2} \equiv 2 \pmod 8$. $\qquad\square$

Proof of Theorem 5.5(d). Suppose $\alpha^2 \equiv 1 + 5 \cdot 2^{2n} \pmod{2^{2n+3}}$ for $n \geq 3$. Then $\alpha^2 \equiv 1 + c \cdot 2^{2n} \pmod{2^{2n+\ell}}$ for $c = 5$ and $\ell = 3$. By Lemma 5.4, $x_n \equiv 1 + 5^2 \equiv 2 \pmod 8$. Since $\alpha^2 \equiv 1 + 5 \cdot 2^{2n} \equiv 1 \pmod{32}$, by Lemma 5.6 and induction $x_{n+2j} \equiv 2 \pmod 8$ for all $j \in \mathbb{Z}_{\geq 0}$, and hence $|x_{n+2j}| = \frac{1}{2}$ for all $j \in \mathbb{Z}_{\geq 0}$. Therefore by (5-6) and (5-7), $|x_{n+2j+1}| = 4$ for all $j \in \mathbb{Z}_{\geq 0}$. Thus by (5-7), $|f^{n+2j}(\alpha)| = 2$ and $|f^{n+2j+1}(\alpha)| = 1$ for all $j \in \mathbb{Z}_{\geq 0}$. $\qquad\square$

Theorem 5.7. *Let $\alpha \in \mathbb{Z}_2$ such that $\alpha^2 \equiv 1 \pmod{32}$. Let $f = f_{\alpha,0}$. If for some $n \in \mathbb{Z}_{>0}$ $|f^n(\alpha)| = |f^{n+2}(\alpha)| = 2$ and $|f^{n+1}(\alpha)| = |f^{n+3}(\alpha)| = 1$ then*

$$|f^{n+2j}(\alpha)| = 2 \quad and \quad |f^{n+2j+1}(\alpha)| = 1$$

for all $j \in \mathbb{Z}_{\geq 0}$.

Proof. Since $|f^n(\alpha)| = |f^{n+2}(\alpha)| = 2$ and $|f^{n+1}(\alpha)| = |f^{n+3}(\alpha)| = 1$, by (5-7) $|x_n| = \frac{1}{2}$, $|x_{n+1}| = 4$, and $|x_{n+2}| = \frac{1}{2}$. Since $|x_n| = \frac{1}{2}$, we have $x_n = 2c$ for some odd $c \in \mathbb{Z}_2$. Therefore by Lemma 5.1,

$$4x_{n+2} = \alpha^2 x_{n+1}(4x_{n+1} - 3)^2 + 3$$
$$= \alpha^2 \tfrac{1}{4}(\alpha^2 x_n(4x_n - 3)^2 + 3)((\alpha^2 x_n(4x_n - 3)^2 + 3) - 3)^2 + 3$$
$$= \alpha^2 \tfrac{1}{4}(\alpha^2(2c)(8c - 3)^2 + 3)(\alpha^2(2c)(8c - 3)^2)^2 + 3$$

$$= \alpha^2(\alpha^2(2c)(8c-3)^2+3)(\alpha^2 c(8c-3)^2)^2+3$$
$$\equiv (2c(8c-3)^2+3)(c(8c-3)^2)^2+3 \equiv 3+3c^2+2c^3 \pmod{16};$$

in this last congruence we expanded the polynomial and reduced coefficients modulo 16. If $c \equiv 3 \pmod 4$ then $4x_{n+2} \equiv 84 \equiv 4 \pmod{16}$, and so $|x_{n+2}| = 1$, a contradiction. Therefore $c \equiv 1 \pmod 4$. Since $x_n = 2c$, we have $x_n \equiv 2 \pmod 8$. By Lemma 5.6 and induction, we find that $x_{n+2j} \equiv 2 \pmod 8$ for all $j \in \mathbb{Z}_{\geq 0}$. As the proof of Theorem 5.5(d) shows, from this we can then conclude that $|f^{n+2j}(\alpha)| = 2$ and $|f^{n+2j+1}(\alpha)| = 1$ for all $j \in \mathbb{Z}_{\geq 0}$. □

We reiterate that Theorem 5.7 detects preperiodicity in $\{|f^n(\alpha)|\}$ that (5-15) does not detect. Indeed, for $\alpha^2 \equiv 321 \pmod{2^9}$ we find that

$$\{d_n\} = \{324, 20, 8, 5, 8, 1, 0\} \quad \text{and} \quad \{m_n\} = \{9, 7, 5, 3, 4, 2, 2\},$$

which clearly does not satisfy (5-15). Yet from these sequences we find that $\{|f^n(\alpha)|\} = \{2, 2, 2, 1, 2, 1, 2, \ldots\}$, and so the hypothesis of Theorem 5.7 is fulfilled.

Lemma 5.6 is the key to the proofs of Theorem 5.5(d) and Theorem 5.7. We suspect there may be other results such as Lemma 5.6 where a congruence condition on α^2 forces a form of preperiodicity in x_n. If this is the case, then there may exist other results similar to Theorem 5.7 which allow for further detection of preperiodicity in $\{|f^n(\alpha)|\}$.

Acknowledgments

References

[Anderson 2013] J. Anderson, "Bounds on the radius of the p-adic Mandelbrot set", *Acta Arith.* **158**:3 (2013), 253–269. MR Zbl

[Bate et al. 2018] B. Bate, K. Craft, and J. Yuly, "2-adic Mandelbrot set calculations", Mathematica source code, 2018, https://github.com/brandonbate/2-adic-Mandelbrot-Set-Calculations.

[Beardon 1991] A. F. Beardon, *Iteration of rational functions*, Graduate Texts in Math. **132**, Springer, 1991. MR Zbl

[Devaney 1989] R. L. Devaney, *An introduction to chaotic dynamical systems*, 2nd ed., Addison-Wesley, Redwood City, CA, 1989. MR Zbl

[Dudko 2017] D. Dudko, "The decoration theorem for Mandelbrot and multibrot sets", *Int. Math. Res. Not.* **2017**:13 (2017), 3985–4028. MR Zbl

[Gouvêa 1993] F. Q. Gouvêa, *p-adic numbers: an introduction*, Springer, 1993. MR Zbl

[Koblitz 1977] N. Koblitz, *p-adic numbers, p-adic analysis, and zeta-functions*, Graduate Texts in Math. **58**, Springer, 1977. MR Zbl

[Lomonaco and Petersen 2017] L. Lomonaco and C. L. Petersen, "On quasi-conformal (in-)compatibility of satellite copies of the Mandelbrot set, I", *Invent. Math.* **210**:2 (2017), 615–644. MR Zbl

[Silverman 2013] J. H. Silverman, "What is. . . the p-adic Mandelbrot set?", *Notices Amer. Math. Soc.* **60**:8 (2013), 1048–1050. MR Zbl

brandon.bate@houghton.edu   *Department of Mathematics, Houghton College, Houghton, NY, United States*

kyle.craft16@houghton.edu   *Houghton College, Houghton, NY, United States*

jonathon.yuly16@houghton.edu   *Houghton College, Houghton, NY, United States*

Sidon sets and 2-caps in $\mathbb{F}_3^n$

Yixuan Huang, Michael Tait and Robert Won

(Communicated by Joshua Cooper)

For each natural number d, we introduce the concept of a d-cap in $\mathbb{F}_3^n$. A set of points in $\mathbb{F}_3^n$ is called a d-cap if, for each $k = 1, 2, \ldots, d$, no $k + 2$ of the points lie on a k-dimensional flat. This generalizes the notion of a cap in $\mathbb{F}_3^n$. We prove that the 2-caps in $\mathbb{F}_3^n$ are exactly the Sidon sets in $\mathbb{F}_3^n$ and study the problem of determining the size of the largest 2-cap in $\mathbb{F}_3^n$.

1. Introduction

Throughout, let $\mathbb{F}_q$ denote the field with q elements and let $\mathbb{F}_q^n$ denote n-dimensional affine space over $\mathbb{F}_q$. A *cap* in $\mathbb{F}_3^n$ is a collection of points such that no three are collinear. Although this definition is geometric, there is an equivalent definition that is arithmetic: a set of points C is a cap in $\mathbb{F}_3^n$ if and only if C contains no three-term arithmetic progressions.

Here, we consider natural generalizations of caps in $\mathbb{F}_3^n$. For $d \in \mathbb{N}$, we call a set of points a *d-cap* if, for each $k = 1, 2, \ldots, d$, no $k + 2$ of the points lie on a k-dimensional flat. With this definition, a 1-cap corresponds to the usual definition of a cap. We also remark that if C is a set of points in $\mathbb{F}_3^n$, then the points of C are in general linear position if and only if C is an $(n-1)$-cap.

Let $r(1, \mathbb{F}_3^n)$ denote the maximal size of a 1-cap in $\mathbb{F}_3^n$. In general, it is a difficult problem to determine $r(1, \mathbb{F}_3^n)$ — in fact, the exact answer is known only when $n \leq 6$. Table 1 lists the best known upper and lower bounds on $r(1, \mathbb{F}_3^n)$ for $n \leq 10$ [Versluis 2017]. It is also known that in dimension $n \leq 6$, maximal 1-caps are equivalent up to affine transformation [Edel et al. 2002; Pellegrino 1970; Potechin 2008].

The asymptotic bounds on $r(1, \mathbb{F}_3^n)$ are well-studied. Edel [2004] showed that

$$\limsup_{n \to \infty} \frac{\log_3(r(1, \mathbb{F}_3^n))}{n} \geq 0.724851$$

---

dimension	1	2	3	4	5	6	7	8	9	10
lower bound	2	4	9	20	45	112	236	496	1064	2240
upper bound	2	4	9	20	45	112	291	771	2070	5619

Table 1. The best known bounds for the size of a maximal 1-cap in $\mathbb{F}_3^n$.

dimension	1	2	3	4	5	6	7	8	n even	n odd
lower bound	2	3	5	9	13	27	33	81	$3^{n/2}$	$3^{(n-1)/2}+1$
upper bound	2	3	5	9	13	27	47	81	$3^{n/2}$	$\lceil 3^{n/2} \rceil$

Table 2. Bounds for the size of a maximal 2-cap in $\mathbb{F}_3^n$.

and consequently that $r(1, \mathbb{F}_3^n)$ is $\Omega(2.2174^n)$ (using Hardy and Littlewood's Ω notation). In more recent breakthrough work Ellenberg and Gijswijt [2017] (adapting a method of Croot, Lev, and Pach in [Croot et al. 2017]) proved that $r(1, \mathbb{F}_3^n)$ is $o(2.756^n)$.

   In this paper, we focus on the study of 2-caps in $\mathbb{F}_3^n$. We show that there is an equivalent arithmetic formulation of the definition of a 2-cap. In particular, the 2-caps in $\mathbb{F}_3^n$ are exactly the Sidon sets in $\mathbb{F}_3^n$, which are important objects in combinatorial number theory (we refer the interested reader to the survey [O'Bryant 2004]). Using this definition, we are able to compute the exact maximal size of a 2-cap in $\mathbb{F}_3^n$ when n is even. We also examine 2-caps in low dimension when n is odd, in particular considering dimensions $n = 3, 5$, and 7.

   Table 2 lists the bounds we obtain for the size of a maximal 2-cap in $\mathbb{F}_3^n$. The values in dimension 3, 5, and 7 are given by Theorems 3.9 and 3.10, and Proposition 3.12, respectively. The bounds for even dimension follow from Theorem 3.4. The upper bound in odd dimension n follows from Proposition 3.3 and the lower bound is given by adding one affinely independent point to the construction in dimension $n - 1$. Knowing the exact value in even dimension also allows us to conclude that asymptotically, the maximal size of a 2-cap in $\mathbb{F}_3^n$ is $\Theta(3^{n/2})$.

2. Preliminaries

In this section, we establish basic notation, definitions, and background. The set of natural numbers is denoted by $\mathbb{N} = \{1, 2, 3, \dots\}$. Throughout, d and n will always denote natural numbers. An element $\boldsymbol{a} \in \mathbb{F}_3^n$ will be written as a row vector $\boldsymbol{a} = (a_1, a_2, \dots, a_n)$, with each $a_i \in \{0, 1, 2\}$. We will sometimes order the vectors of $\mathbb{F}_3^n$ lexicographically — i.e., by regarding them as ternary strings. We use the notation $\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_n$ to denote the n standard basis vectors in an n-dimensional vector space.

A k-dimensional affine subspace of a vector space is called a k-*dimensional flat*. In particular, a 1-dimensional flat is also called a *line*. In the affine space $\mathbb{F}_3^n$, every line consists of the points $\{\boldsymbol{a}, \boldsymbol{a}+\boldsymbol{b}, \boldsymbol{a}+2\boldsymbol{b}\}$ for some $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{F}_3^n$, where $\boldsymbol{b} \neq \boldsymbol{0}$. Hence, the lines in $\mathbb{F}_3^n$ correspond to three-term arithmetic progressions. It is easy to see that three distinct points in $\mathbb{F}_3^n$ are collinear if and only if they sum to $\boldsymbol{0}$. Likewise, a 2-dimensional flat is called a *plane*. Any three noncollinear points determine a unique plane. For $\boldsymbol{a} = (a_1, a_2, \ldots, a_k) \in \mathbb{F}_3^k$ with $k < n$, the subset of $\mathbb{F}_3^n$ whose first k entries are $a_1, a_2, \ldots, a_k$ is an $(n-k)$-dimensional flat which we call *the $\boldsymbol{a}$-affine subspace* of $\mathbb{F}_3^n$.

Two subsets C and D of a vector space are called *affinely equivalent* if there exists an invertible affine transformation T such that $T(C) = D$. It is clear that affine equivalence determines an equivalence relation on the power set of a vector space. Given a set of points X in a vector space, its affine span is given by the set of all affine combinations of points of X. A set X is called *affinely independent* if no proper subset of X has the same affine span as X. Equivalently, $\{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is affinely independent if and only if $\{\boldsymbol{x}_1 - \boldsymbol{x}_0, \boldsymbol{x}_2 - \boldsymbol{x}_0, \ldots, \boldsymbol{x}_n - \boldsymbol{x}_0\}$ is linearly independent.

Definition 2.1. A subset C of $\mathbb{F}_3^n$ is called a *d-cap* if, for each $k = 1, 2, \ldots, d$, no $k+2$ points of C lie on a k-dimensional flat. Equivalently, C is a d-cap if and only if any subset of C of size at most $d+2$ is affinely independent. A d-cap is called *complete* if it is not a proper subset of another d-cap and is called *maximal* if it is of the largest possible cardinality.

As mentioned in the Introduction, a 1-cap is a classical cap. We will denote the size of a maximal d-cap in $\mathbb{F}_3^n$ by $r(d, \mathbb{F}_3^n)$. We remark that since invertible affine transformations preserve affine independence, the image of a d-cap under an invertible affine transformation is again a d-cap. As a warm-up, we prove some basic facts about maximal d-caps in $\mathbb{F}_3^n$.

Lemma 2.2. *We have that $r(d, \mathbb{F}_3^n) \geq n+1$ with equality if $n \leq d$.*

Proof. The set $\{\boldsymbol{0}, \boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ is an affinely independent subset of $\mathbb{F}_3^n$ of size $n+1$ and hence is a d-cap for any $d \in \mathbb{N}$. Therefore, $r(d, \mathbb{F}_3^n) \geq n+1$.

Now suppose $n \leq d$. Since, by definition, a d-cap must be an n-cap, we have that $r(d, \mathbb{F}_3^n) \leq r(n, \mathbb{F}_3^n)$. A maximal affinely independent set in $\mathbb{F}_3^n$ has size $n+1$ so $r(n, \mathbb{F}_3^n) \leq n+1$, and so $r(d, \mathbb{F}_3^n) = n+1$. $\qquad\square$

Corollary 2.3. *When $n \leq d$, all maximal d-caps in $\mathbb{F}_3^n$ are affinely equivalent.*

Proof. By Lemma 2.2, when $n \leq d$, a maximal d-cap in $\mathbb{F}_3^n$ is a maximal affinely independent set, i.e., an affine basis of $\mathbb{F}_3^n$. All affine bases in an affine space are equivalent up to affine transformation. $\qquad\square$

Lemma 2.4. *For fixed d, $r(d, \mathbb{F}_3^n)$ is a nondecreasing function of n and for fixed n, $r(d, \mathbb{F}_3^n)$ is a nonincreasing function of d.*

Proof. Since $\mathbb{F}_3^{n-1}$ is an affine subspace of $\mathbb{F}_3^n$, a d-cap in $\mathbb{F}_3^{n-1}$ naturally embeds as a d-cap in $\mathbb{F}_3^n$. Hence $r(d, \mathbb{F}_3^{n-1}) \leq r(d, \mathbb{F}_3^n)$ so the first statement follows. The second statement follows since, by definition, a d-cap in $\mathbb{F}_3^n$ must be a $(d-1)$-cap. Hence, $r(d-1, \mathbb{F}_3^n) \geq r(d, \mathbb{F}_3^n)$. $\qquad\square$

3. 2-caps in $\mathbb{F}_3^n$

We now restrict our attention to the study of 2-caps in $\mathbb{F}_3^n$. Our first observation is that in $\mathbb{F}_3^n$, the definition of a 2-cap is equivalent to the definition of a Sidon set.

Definition 3.1. Let G be an abelian group. A subset $A \subseteq G$ is called a *Sidon set* if, whenever $a + b = c + d$ with $a, b, c, d \in A$, the pair (a, b) is a permutation of the pair (c, d).

Theorem 3.2. *A subset C of $\mathbb{F}_3^n$ is a 2-cap if and only if it is a Sidon set.*

Proof. First suppose that C is not a 2-cap. Then C contains three points which are collinear or C contains four points which are coplanar. If C contains three distinct collinear points $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ then $\boldsymbol{a} + \boldsymbol{b} + \boldsymbol{c} = \boldsymbol{0}$ and hence $\boldsymbol{a} + \boldsymbol{b} = \boldsymbol{c} + \boldsymbol{c}$ so C is not a Sidon set.

Suppose therefore that no three points in C are collinear. Then C contains four coplanar points, say $\{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d}\}$. Every set of three distinct noncollinear points in $\mathbb{F}_3^n$ lies on a unique 2-dimensional flat. In particular, the 2-dimensional flat F containing $\boldsymbol{a}, \boldsymbol{b}$, and $\boldsymbol{c}$ is given by

$$F = \begin{array}{|c|c|c|} \hline \boldsymbol{a} & \boldsymbol{b} & -\boldsymbol{a}-\boldsymbol{b} \\ \hline \boldsymbol{c} & -\boldsymbol{a}+\boldsymbol{b}+\boldsymbol{c} & \boldsymbol{a}-\boldsymbol{b}+\boldsymbol{c} \\ \hline -\boldsymbol{a}-\boldsymbol{c} & \boldsymbol{a}+\boldsymbol{b}-\boldsymbol{c} & -\boldsymbol{b}-\boldsymbol{c} \\ \hline \end{array}$$

and since we assumed that no three points in C are collinear, we must have that $\boldsymbol{d} = -\boldsymbol{a}+\boldsymbol{b}+\boldsymbol{c}$, $\boldsymbol{d} = \boldsymbol{a}-\boldsymbol{b}+\boldsymbol{c}$ or $\boldsymbol{d} = \boldsymbol{a}+\boldsymbol{b}-\boldsymbol{c}$. In the first case, $\boldsymbol{a}+\boldsymbol{d} = \boldsymbol{b}+\boldsymbol{c}$, in the second case, $\boldsymbol{b}+\boldsymbol{d} = \boldsymbol{a}+\boldsymbol{c}$, and in the third case $\boldsymbol{c}+\boldsymbol{d}+\boldsymbol{a}+\boldsymbol{b}$. In any case, C is not a Sidon set.

Conversely, suppose that C is not a Sidon set. Then either C contains three distinct points $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ such that $\boldsymbol{a}+\boldsymbol{a} = \boldsymbol{b}+\boldsymbol{c}$, or C contains four distinct points $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d}$ such that $\boldsymbol{a}+\boldsymbol{b} = \boldsymbol{c}+\boldsymbol{d}$. In the first case, $\boldsymbol{a}+\boldsymbol{b}+\boldsymbol{c} = \boldsymbol{0}$ so C contains a line. In the second case, $\boldsymbol{d} = \boldsymbol{a}+\boldsymbol{b}-\boldsymbol{c}$, so $\boldsymbol{d}$ lies in the plane determined by $\boldsymbol{a}, \boldsymbol{b}$, and $\boldsymbol{c}$, and hence the four points are coplanar. In either case, C is not a 2-cap. $\quad\square$

Since, in $\mathbb{F}_3^n$, 2-caps correspond to Sidon sets, we will use the terms interchangeably throughout. We obtain an upper bound on $r(2, \mathbb{F}_3^n)$ by an easy counting argument; see [Cilleruelo et al. 2010, Corollary 2.2].

Proposition 3.3. *For any $n \in \mathbb{N}$,*

$$r(2, \mathbb{F}_3^n) \cdot (r(2, \mathbb{F}_3^n) - 1) \leq 3^n - 1.$$

Proof. Suppose $C \subset \mathbb{F}_3^n$ is a 2-cap and hence, by Theorem 3.2, a Sidon set. For $a, b, c, d \in C$, if $a - b = c - d$ then $\{a, d\} = \{c, b\}$ and so we have either $a = b$, or else $a = c$ and $b = d$. Therefore, the set $\{a - b : a, b \in C, \ a \neq b\}$ has size $|C|(|C| - 1)$. Since these differences are nonzero, we have

$$|C|(|C| - 1) \leq 3^n - 1. \qquad \square$$

Even dimension.

Theorem 3.4. *If n is even, then $r(2, \mathbb{F}_3^n) = 3^{n/2}$.*

Proof. First we will show the lower bound, $r(2, \mathbb{F}_3^n) \geq 3^{n/2}$. Since $\mathbb{F}_3^n$ is additively isomorphic to $\mathbb{F}_3^{n/2} \times \mathbb{F}_3^{n/2}$, it suffices to construct a Sidon set of size $3^{n/2}$ in $\mathbb{F}_3^{n/2} \times \mathbb{F}_3^{n/2}$. As vector spaces over $\mathbb{F}_3$, $\mathbb{F}_3^{n/2}$ is isomorphic to $\mathbb{F}_{3^{n/2}}$, the finite field with $3^{n/2}$ elements. Hence, it suffices to construct a Sidon set of size $3^{n/2}$ in $\mathbb{F}_{3^{n/2}} \times \mathbb{F}_{3^{n/2}}$. This follows easily from the following claim; for a proof, see [Cilleruelo 2012, Example 1].

Claim. *Let q be an odd prime power and $\mathbb{F}_q$ be the finite field of order q. Then the set $\{(x, x^2) : x \in \mathbb{F}_q\}$ is a Sidon set in $\mathbb{F}_q \times \mathbb{F}_q$.*

It is clear that the set $\{(x, x^2) : x \in \mathbb{F}_{3^{n/2}}\}$ has size $3^{n/2}$, so we have $r(2, \mathbb{F}_3^n) \geq 3^{n/2}$. For the upper bound, let $C \subset \mathbb{F}_3^n$ be a 2-cap. Since n is even, $3^{n/2}$ is an integer, and if $|C| \geq 3^{n/2} + 1$, this contradicts Proposition 3.3. Therefore, $r(2, \mathbb{F}_3^n) \leq 3^{n/2}$. $\square$

Corollary 3.5. *As $n \to \infty$, $r(2, \mathbb{F}_3^n)$ is $\Theta(3^{n/2})$.*

The construction above can be leveraged into the following partitioning theorem.

Theorem 3.6. *When n is even, there is a partition of $\mathbb{F}_3^n$ into maximal 2-caps.*

This serves as an analogue to similar results for 1-caps in $\mathbb{F}_3^n$. It is well known that $\mathbb{F}_3^3$ can be partitioned into three maximal 1-caps of size 9. It is possible to partition $\mathbb{F}_3^2$ into a single point and two disjoint maximal 1-caps of size 4. Finally, [Follett et al. 2014, Theorem 3.3] shows that $\mathbb{F}_3^4$ can be partitioned into a single point and four disjoint maximal 1-caps of size 20.

Proof of Theorem 3.6. Since translations of Sidon sets are also Sidon sets, for each $a \in \mathbb{F}_{3^{n/2}}$ the set $S_a := \{(x, x^2 + a) : x \in \mathbb{F}_{3^{n/2}}\}$ is a maximal 2-cap. Since $(x, x^2 + a) = (y, y^2 + b)$ implies $x = y$ and hence $a = b$, we have that S_a and S_b

are disjoint for $a \neq b$. Therefore, as a ranges over $\mathbb{F}_{3^{n/2}}$ the sets S_a cover 3^n points and thus there is the claimed partition. $\qquad\square$

Question 3.7. By Corollary 2.3, all maximal 2-caps in $\mathbb{F}_3^2$ are affinely equivalent. Is this true in $\mathbb{F}_3^n$ when n is even?

We remark that when $n = 4$, a computer program verified that all maximal 2-caps sum to $\mathbf{0}$. If a set of nine points sums to $\mathbf{0}$ in $\mathbb{F}_3^4$, then its image under any affine transformation will likewise sum to $\mathbf{0}$, so this is a necessary condition for all maximal 2-caps in $\mathbb{F}_3^4$ to be affinely equivalent.

Odd dimension.

Lemma 3.8. _If_ $C = \{a, b, c, d\}$ _is a 2-cap of size 4 in_ $\mathbb{F}_3^n$ _then_ $D = \{a, b, c, d, a+b+c+d\}$ _is a 2-cap of size 5._

Proof. First we note that the points of D are distinct since if, without loss of generality, $a+b+c+d = a$, this implies that b, c, and d are collinear, which is impossible since C is a 2-cap.

Now, suppose for contradiction that D is not a 2-cap, so there exist some $x, y, z, w \in D$ with $x + y = z + w$. Since C is a 2-cap, we may assume that $x = a+b+c+d$. Without loss of generality, we then have that one of the following occurs:

(1) $(a+b+c+d)+a = b+c$. Then $a = d$, which is impossible since C has size 4.

(2) $(a+b+c+d)+a = 2b$. Then $a+b = c+d$, which is impossible since C is a 2-cap.

(3) $2(a+b+c+d) = b+c$. Then $a+d = b+c$, which is impossible since C is a 2-cap.

(4) $2(a+b+c+d) = 2a$. Then b, c, and d are collinear, which is impossible since C is a 2-cap.

Hence, D is a 2-cap. $\qquad\square$

Theorem 3.9. _In_ $\mathbb{F}_3^3$, _a maximal 2-cap has size 5; that is,_ $r(2, \mathbb{F}_3^3) = 5$. _Further, all complete 2-caps are maximal and all maximal 2-caps are affinely equivalent._

Proof. Since $\{0, e_1, e_2, e_3\}$ is an affinely independent set in $\mathbb{F}_3^3$, by Lemma 3.8 $\{0, e_1, e_2, e_3, e_1+e_2+e_3\}$ is a 2-cap in $\mathbb{F}_3^3$. Hence, $r(2, \mathbb{F}_3^3) \geq 5$. But by Proposition 3.3, $r(2, \mathbb{F}_3^3) < 6$ and hence $r(2, \mathbb{F}_3^3) = 5$.

Let C be any complete 2-cap in $\mathbb{F}_3^3$. Since $\mathbb{F}_3^3$ is a 3-dimensional affine space, if $|C| \leq 3$, then $\mathbb{F}_3^3$ contains a point which is affinely independent from the points of C, so C cannot be complete. Hence, $|C| \geq 4$. But if $|C| = 4$ then by Lemma 3.8, C is not complete. Hence, $|C| = 5$, and any complete 2-cap in $\mathbb{F}_3^3$ is already maximal.

For the final claim, suppose C is a maximal 2-cap in $\mathbb{F}_3^3$. Pick any four points in C. Since these points are affinely independent, there exists an invertible affine transformation mapping these points to the set $\{\mathbf{0}, e_1, e_2, e_3\}$. Hence, we need only show that all maximal 2-caps containing $\{\mathbf{0}, e_1, e_2, e_3\}$ are affinely equivalent.

It is easy to verify that there are exactly five such maximal 2-caps, namely

$$C_1 = \{\mathbf{0}, e_1, e_2, e_3, (1, 1, 1)\}, \quad C_4 = \{\mathbf{0}, e_1, e_2, e_3, (2, 2, 1)\},$$

$$C_2 = \{\mathbf{0}, e_1, e_2, e_3, (1, 2, 2)\}, \quad C_5 = \{\mathbf{0}, e_1, e_2, e_3, (2, 2, 2)\}.$$

$$C_3 = \{\mathbf{0}, e_1, e_2, e_3, (2, 1, 2)\},$$

It suffices to exhibit an invertible affine transformation T_i mapping C_1 to C_i for $i = 2, 3, 4, 5$. We provide these T_i explicitly, writing $T_i(x) = A_i x + b_i$ for an invertible matrix A_i and $b_i \in \mathbb{F}_3^3$:

$$A_2 = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 2 & 0 & 1 \end{bmatrix} \quad \text{and} \quad b_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \qquad A_4 = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad b_4 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

$$A_3 = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 0 & 1 \\ 0 & 1 & 2 \end{bmatrix} \quad \text{and} \quad b_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \qquad A_5 = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix} \quad \text{and} \quad b_5 = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix}. \qquad \square$$

Theorem 3.10. *A maximal 2-cap in $\mathbb{F}_3^5$ has size 13; that is, $r(2, \mathbb{F}_3^5) = 13$.*

Proof. Let C be a maximal 2-cap in $\mathbb{F}_3^5$. By Theorem 3.4, $r(2, \mathbb{F}_3^4) = 9$ so by Lemma 2.4 we may assume that $|C| \geq 9$. We will apply a sequence of affine transformations to C to conclude that lexicographically, the first points in C are $\{\mathbf{0}, e_5, e_4, e_3, e_3 + e_4 + e_5, e_2\}$ or $\{\mathbf{0}, e_5, e_4, e_3, e_2\}$.

Given any four affinely independent points, there exists an invertible affine transformation mapping them to $\mathbf{0}, e_5, e_4$, and e_3, so without loss of generality we may assume that C contains the subset $\{\mathbf{0}, e_5, e_4, e_3\}$. These points all lie in the $(0, 0)$-affine subspace of $\mathbb{F}_3^5$. Since $r(2, \mathbb{F}_3^3) = 5$, the $(0, 0)$-affine subspace contains four points or five points of C. If it contains five points, then by Theorem 3.9, we may apply an affine transformation (using a block matrix) and assume that the fifth point is $e_3 + e_4 + e_5$.

Consider any other point $a \in C$. Since a is not in the $(0, 0)$-affine subspace of $\mathbb{F}_3^5$, $\{\mathbf{0}, e_5, e_4, e_3, a\}$ is an affinely independent set so there exists an affine transformation T fixing $\mathbf{0}, e_5, e_4$, and e_3 and mapping a to e_2. Notice that if T is given by multiplication by the invertible matrix A followed by addition by $b \in \mathbb{F}_3^5$, we have

$$T(e_3 + e_4 + e_5) = A(e_3 + e_4 + e_5) + b = T(\mathbf{0}) + T(e_3) + T(e_4) + T(e_5) = e_3 + e_4 + e_5,$$

so T fixes $e_3 + e_4 + e_5$.

Hence, up to affine equivalence, we may assume that the lexicographically earliest points in C are $\{\mathbf{0}, \mathbf{e}_5, \mathbf{e}_4, \mathbf{e}_3, \mathbf{e}_3 + \mathbf{e}_4 + \mathbf{e}_5, \mathbf{e}_2\}$ or $\{\mathbf{0}, \mathbf{e}_5, \mathbf{e}_4, \mathbf{e}_3, \mathbf{e}_2\}$. A computer program was used to enumerate all possible complete 2-caps beginning with these sets of points. This verified that $r(2, \mathbb{F}_3^5) = 13$. The C++ code for the program is available on Won's professional website.   □

Remark 3.11. The maximal 2-cap in $\mathbb{F}_3^5$ that is lexicographically earliest is explicitly given by the points

$(0, 0, 0, 0, 0)$, $(0, 0, 0, 0, 1)$, $(0, 0, 0, 1, 0)$, $(0, 0, 1, 0, 0)$, $(0, 0, 1, 1, 1)$,

$(0, 1, 0, 0, 0)$, $(0, 1, 1, 1, 2)$, $(0, 2, 1, 2, 0)$, $(0, 2, 2, 1, 2)$, $(1, 0, 0, 0, 0)$,

$(1, 0, 1, 2, 1)$, $(2, 0, 1, 0, 2)$, $(2, 2, 0, 2, 2)$.

We conclude by giving bounds on $r(2, \mathbb{F}_3^7)$.

Proposition 3.12. *One has that $33 \leq r(2, \mathbb{F}_3^7) \leq 47$.*

Proof. The upper bound on $r(2, \mathbb{F}_3^7)$ is a consequence of Proposition 3.3. For the lower bound, we constructed a 2-cap of size 33 by first embedding a maximal 2-cap in $\mathbb{F}_3^6$ as a 2-cap C of size 27 in $\mathbb{F}_3^7$. We then used a computer program to enumerate all complete 2-caps containing C as a subset. The largest of these complete 2-caps has size 33. The lexicographically earliest one is given by the points

$(0, 0, 0, 0, 0, 0, 0)$, $(0, 0, 0, 1, 0, 0, 1)$, $(0, 0, 0, 2, 0, 0, 1)$,

$(0, 0, 1, 0, 1, 0, 0)$, $(0, 0, 1, 1, 1, 2, 1)$, $(0, 0, 1, 2, 1, 1, 1)$,

$(0, 0, 2, 0, 1, 0, 0)$, $(0, 0, 2, 1, 1, 1, 1)$, $(0, 0, 2, 2, 1, 2, 1)$,

$(0, 1, 0, 0, 1, 2, 0)$, $(0, 1, 0, 1, 0, 2, 1)$, $(0, 1, 0, 2, 2, 2, 1)$,

$(0, 1, 1, 0, 2, 1, 1)$, $(0, 1, 1, 1, 1, 0, 2)$, $(0, 1, 1, 2, 0, 2, 2)$,

$(0, 1, 2, 0, 2, 0, 2)$, $(0, 1, 2, 1, 1, 1, 0)$, $(0, 1, 2, 2, 0, 2, 0)$,

$(0, 2, 0, 0, 1, 2, 0)$, $(0, 2, 0, 1, 2, 2, 1)$, $(0, 2, 0, 2, 0, 2, 1)$,

$(0, 2, 1, 0, 2, 0, 2)$, $(0, 2, 1, 1, 0, 2, 0)$, $(0, 2, 1, 2, 1, 1, 0)$,

$(0, 2, 2, 0, 2, 1, 1)$, $(0, 2, 2, 1, 0, 2, 2)$, $(0, 2, 2, 2, 1, 0, 2)$,

$(1, 0, 0, 0, 0, 0, 0)$, $(1, 0, 0, 0, 0, 0, 1)$, $(2, 0, 0, 1, 0, 2, 0)$,

$(2, 0, 0, 1, 1, 0, 1)$, $(2, 0, 0, 1, 1, 1, 2)$, $(2, 0, 0, 1, 1, 2, 2)$.   □

Acknowledgments

References

[Cilleruelo 2012] J. Cilleruelo, "Combinatorial problems in finite fields and Sidon sets", *Combinatorica* **32**:5 (2012), 497–511. MR Zbl

[Cilleruelo et al. 2010] J. Cilleruelo, I. Ruzsa, and C. Vinuesa, "Generalized Sidon sets", *Adv. Math.* **225**:5 (2010), 2786–2807. MR Zbl

[Croot et al. 2017] E. Croot, V. F. Lev, and P. P. Pach, "Progression-free sets in $\mathbb{Z}_4^n$ are exponentially small", *Ann. of Math.* (2) **185**:1 (2017), 331–337. MR Zbl

[Edel 2004] Y. Edel, "Extensions of generalized product caps", *Des. Codes Cryptogr.* **31**:1 (2004), 5–14. MR Zbl

[Edel et al. 2002] Y. Edel, S. Ferret, I. Landjev, and L. Storme, "The classification of the largest caps in AG(5, 3)", *J. Combin. Theory Ser. A* **99**:1 (2002), 95–110. MR Zbl

[Ellenberg and Gijswijt 2017] J. S. Ellenberg and D. Gijswijt, "On large subsets of $\mathbb{F}_q^n$ with no three-term arithmetic progression", *Ann. of Math.* (2) **185**:1 (2017), 339–343. MR Zbl

[Follett et al. 2014] M. Follett, K. Kalail, E. McMahon, C. Pelland, and R. Won, "Partitions of AG(4, 3) into maximal caps", *Discrete Math.* **337** (2014), 1–8. MR Zbl

[O'Bryant 2004] K. O'Bryant, "A complete annotated bibliography of work related to Sidon sequences", dynamic survey DS-11, *Electron. J. Combin.*, 2004, available at https://tinyurl.com/osurveyds. Zbl

[Pellegrino 1970] G. Pellegrino, "Sul massimo ordine delle calotte in $S_{4,3}$", *Matematiche* (*Catania*) **25** (1970), 149–157. MR

[Potechin 2008] A. Potechin, "Maximal caps in AG(6, 3)", *Des. Codes Cryptogr.* **46**:3 (2008), 243–259. MR Zbl

[Versluis 2017] N. Versluis, *On the cap set problem*, bachelor thesis, Delft University of Technology, 2017, available at https://tinyurl.com/delftvers.

huany16@wfu.edu   *Department of Mathematics and Statistics,*
                                 Wake Forest University, Winston-Salem, NC, United States

mtait@cmu.edu   *Department of Mathematical Sciences,*
                                 Carnegie Mellon University, Pittsburgh, PA, United States

robwon@uw.edu   *Department of Mathematics, University of Washington,*
                                 Seattle, WA, United States

Covering numbers of upper triangular matrix rings over finite fields

Merrick Cai and Nicholas J. Werner

(Communicated by Scott T. Chapman)

A cover of a finite ring R is a collection of proper subrings $\{S_1, \ldots, S_m\}$ of R such that $R = \bigcup_{i=1}^{m} S_i$. If such a collection exists, then R is called coverable, and the covering number of R is the cardinality of the smallest possible cover. We investigate covering numbers for rings of upper triangular matrices with entries from a finite field. Let $\mathbb{F}_q$ be the field with q elements and let $T_n(\mathbb{F}_q)$ be the ring of $n \times n$ upper triangular matrices with entries from $\mathbb{F}_q$. We prove that if $q \neq 4$, then $T_2(\mathbb{F}_q)$ has covering number $q + 1$, that $T_2(\mathbb{F}_4)$ has covering number 4, and that when p is prime, $T_n(\mathbb{F}_p)$ has covering number $p + 1$ for all $n \geq 2$.

1. Introduction

It is well known that no group is equal to the union of two proper subgroups. However, it is possible to achieve such a union if we allow the use of more than two proper subgroups. For example, the Klein 4-group is equal to the union of its three proper, nontrivial subgroups. More generally, any noncyclic group is equal to the union of its proper cyclic subgroups. Given a finite group G, we say that a collection of proper subgroups $\{H_1, \ldots, H_m\}$ forms a *cover* of G if $G = \bigcup_{i=1}^{m} H_i$. If such a cover exists, then the *covering number* of G is the cardinality of the smallest possible cover.

Natural problems to study regarding covers and covering numbers include finding formulas for the covering number of groups or families of groups, and determining which groups have a specified covering number (e.g., "Which groups have covering number 3?"). Consideration of these types of questions dates back at least nine decades [Scorza 1926; Haber and Rosenfeld 1959; Bruckheimer et al. 1970]. The covering number problem for groups began to become more popular following papers by Cohn [1994] and Tomkinson [1997]. Over the past several years, researchers have begun to study the covering numbers for other algebraic structures [Kappe 2014], including rings [Crestani 2012; Lucchini and Maróti 2012; Werner 2015]. All rings with covering number 3 were characterized in [Lucchini and Maróti 2012],

and a formula for the covering number of a matrix ring over a finite field was given in [Lucchini and Maróti 2010] (see also the related article [Crestani 2012]). Covering numbers for some other families of finite rings, including direct products of finite fields, were found in [Werner 2015]. The purpose of this paper is to examine the covering numbers for rings of upper triangular matrices with entries from a finite field.

In this paper, rings are assumed to be associative and have a unit element $1 \neq 0$. Subrings, however, need not contain a unit element. Given a ring R, we say that $S \subseteq R$ is a subring of R if S is a subgroup of R under addition and is closed under multiplication.

Definition 1.1. A *cover* of a ring R is a set $\mathcal{S}$ of proper subrings of R such that $R = \bigcup_{S \in \mathcal{S}} S$. If a cover of R exists, then R is said to be *coverable*. In this case, the *covering number* of R is the cardinality of the smallest possible cover. When R is coverable, $\sigma(R)$ denotes the covering number of R.

Not every ring, or even every finite ring, is coverable. For example, for any $n \geq 2$, $\mathbb{Z}/n\mathbb{Z}$ is not coverable because the unit element of $\mathbb{Z}/n\mathbb{Z}$ cannot lie in any proper subring. There is a similar obstruction with finite fields. For a prime power q, we let $\mathbb{F}_q$ be the finite field with q elements. Then, $\mathbb{F}_q$ is never coverable, because the generator of the unit group of $\mathbb{F}_q$ cannot lie in a proper subring.

By contrast, a noncommutative ring is always coverable (a short proof of this is given in Lemma 2.2). Most of this paper will focus on a particular class of noncommutative rings: those consisting of upper triangular matrices. For any ring R and any integer $n \geq 2$, we let $T_n(R)$ be the ring of $n \times n$ upper triangular matrices with entries from R. The main theorem of the paper is the calculation of the covering number for $T_2(\mathbb{F}_q)$.

Theorem 1.2. *Let q be a prime power. Then $\sigma(T_2(\mathbb{F}_q)) = q + 1$ when $q \neq 4$, and $\sigma(T_2(\mathbb{F}_4)) = 4$.*

When $n \geq 3$ and q itself is not prime, we are not able to determine the exact covering number for $T_n(\mathbb{F}_q)$. However, we are able to provide an upper bound for $\sigma(T_n(\mathbb{F}_q))$, and this bound equals the covering number when q is prime. In fact, we obtain a more general result about finite rings having a residue field of prime order.

Corollary 1.3. (1) *Let q be a prime power. If $n \geq 3$, then $\sigma(T_n(\mathbb{F}_q)) \leq q + 1$.*

(2) *Let R be a finite ring and let p be the smallest prime dividing the order of R. If R has $\mathbb{F}_p$ as a residue field, then $\sigma(T_n(R)) = p + 1$ for all $n \geq 2$. In particular, $\sigma(T_n(\mathbb{F}_p)) = p + 1$ for all $n \geq 2$.*

We prove both Theorem 1.2 and Corollary 1.3 in Section 3 after stating some basic facts about coverings and covering numbers of rings in Section 2. The paper closes with some remarks on the difficulty of establishing equality in part (1) of Corollary 1.3.

2. Basic definitions and properties

For a group G, it is easy to see that G is coverable if and only if G is not cyclic. This is because if G is cyclic with generator g, then g cannot lie in any proper subgroup of G. On the other hand, if G is noncyclic, then a cover is formed by the collection of all cyclic subgroups of G. Furthermore, if $a \in G$ and $\langle a \rangle$ is a maximal subgroup of G, then $\langle a \rangle$ must be part of any cover of G, since a must lie in some subgroup in the cover, and by maximality that subgroup must equal $\langle a \rangle$. Similar statements are true for rings if we use subrings comparable to cyclic subgroups.

Definition 2.1. Let R be a ring. For any $r \in R$, we let $\langle\!\langle r \rangle\!\rangle$ be the subring of R generated by r. The subring $\langle\!\langle r \rangle\!\rangle$ is equal to the intersection of all subrings of R containing r, and $\langle\!\langle r \rangle\!\rangle$ consists of the elements of the form $c_n r^n + \cdots + c_1 r$, where $n \geq 1$ and $c_1, \ldots, c_n \in \mathbb{Z}$.

The relationships between a ring R and a subring $\langle\!\langle r \rangle\!\rangle$ are much the same as those between a group G and a cyclic subgroup $\langle a \rangle$.

Lemma 2.2. *Let R be a ring*:

(1) *R is coverable if and only if for all $r \in R$ we have $R \neq \langle\!\langle r \rangle\!\rangle$.*

(2) *For all $r \in R$, if $\langle\!\langle r \rangle\!\rangle$ is a maximal subring of R, then $\langle\!\langle r \rangle\!\rangle$ is part of any cover of R.*

(3) *If R is noncommutative, then R is coverable.*

(4) *Let I be a two-sided ideal of R. If R/I is coverable, then so is R, and $\sigma(R) \leq \sigma(R/I)$.*

Proof. Items (1) and (2) are proved just as for groups. For (3), notice that $\langle\!\langle r \rangle\!\rangle$ is a commutative ring for all $r \in R$. So, in a noncommutative ring R, $\langle\!\langle r \rangle\!\rangle$ must be a proper subring for all $r \in R$. Finally, for (4), let $\phi : R \to R/I$ be the quotient map. For any proper subring S of R/I, we have $\phi^{-1}(S)$ is a proper subring of R. So, any cover of R/I lifts to a cover of R. $\qquad\square$

Part (4) of the lemma can be used to find upper bounds for $\sigma(R)$. Finding lower bounds is, in general, harder to do. However, a simple counting argument gives a basic lower bound.

Lemma 2.3. *Let R be a finite coverable ring of order m. Let p be the smallest prime dividing m. Then, $p + 1 \leq \sigma(R)$.*

Proof. Every proper subring of R has order at most m/p. Let $S_1, \ldots, S_p$ be proper subrings of R. Since $0 \in S_i$ for each i, the total number of elements in the union of all p subrings is at most $1 + p(m/p - 1) < m$. Thus, R cannot be covered by fewer than $p + 1$ proper subrings. $\qquad\square$

3. Main results

In this section, we prove Theorem 1.2 and Corollary 1.3. Our strategy is to show that (when $q \neq 2$ or 4) we may form a cover of $T_2(\mathbb{F}_q)$ by using $q + 1$ maximal subrings, each of which is generated by a single matrix. By Lemma 2.2, each of these subrings is part of any cover of $T_2(\mathbb{F}_q)$, so we must have $\sigma(T_2(\mathbb{F}_q)) = q + 1$.

Throughout this section, q is a power of the prime p, and $g \in \mathbb{F}_q$ denotes an element of multiplicative order $q - 1$ (so, in particular, $\mathbb{F}_q = \mathbb{F}_p(g)$). We let I denote the 2×2 identity matrix, and let M be the matrix $M = \left(\begin{smallmatrix} g & 1 \\ 0 & g \end{smallmatrix}\right) \in T_2(\mathbb{F}_q)$. The matrix M is useful because — as we will prove — it generates a maximal subring of $T_2(\mathbb{F}_q)$.

Lemma 3.1. *Let S be a subring of $T_2(\mathbb{F}_q)$ that contains all of the scalar matrices. Then, $|S|$ is either q, q^2, or q^3. Consequently, if $|S| = q^2$, then S is maximal.*

Proof. Since S contains all of the scalar matrices, it is closed under $\mathbb{F}_q$-scalar multiplication, and hence is an $\mathbb{F}_q$-vector space. This means that $|S| = q^d$, where $d = \dim_{\mathbb{F}_q}(S)$. Since $\dim_{\mathbb{F}_q}(T_2(\mathbb{F}_q)) = 3$, we see that any subring of $T_2(\mathbb{F}_q)$ of order q^2 that contains all of the scalar matrices must be maximal. $\qquad\square$

Proposition 3.2. *The subring $\langle\!\langle M \rangle\!\rangle$ is maximal, and $\langle\!\langle M \rangle\!\rangle = \left\{ \left(\begin{smallmatrix} a & b \\ 0 & a \end{smallmatrix}\right) \mid a, b \in \mathbb{F}_q \right\}$.*

Proof. Let $S = \left\{ \left(\begin{smallmatrix} a & b \\ 0 & a \end{smallmatrix}\right) \mid a, b \in \mathbb{F}_q \right\}$. Then, S is a subring of order q^2 and contains all of the scalar matrices, so S is maximal by Lemma 3.1. We will prove that $S \subseteq \langle\!\langle M \rangle\!\rangle$.

Note that $M^q = \left(\begin{smallmatrix} g^q & 0 \\ 0 & g^q \end{smallmatrix}\right) = \left(\begin{smallmatrix} g & 0 \\ 0 & g \end{smallmatrix}\right)$. Since $\langle\!\langle g \rangle\!\rangle = \mathbb{F}_q$, $\langle\!\langle M \rangle\!\rangle$ contains all of the scalar matrices as well as the matrix $M - M^q = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$. This implies that $S \subseteq \langle\!\langle M \rangle\!\rangle$, and by maximality $S = \langle\!\langle M \rangle\!\rangle$. $\qquad\square$

By Lemma 2.2, $\langle\!\langle M \rangle\!\rangle$ will be part of any cover of $T_2(\mathbb{F}_q)$. When q is odd, we can form other maximal subrings by using matrices of the form $\left(\begin{smallmatrix} g & b \\ 0 & -g \end{smallmatrix}\right)$, where $b \in \mathbb{F}_q$. This will not work if q is even, because $\left(\begin{smallmatrix} g & b \\ 0 & -g \end{smallmatrix}\right) \in \langle\!\langle M \rangle\!\rangle$ in this case. So, in characteristic 2, we must use different matrices.

Lemma 3.3. *Let $q = 2^k$, where $k \geq 3$. Let $G = \{a \in \mathbb{F}_q \mid \mathbb{F}_2(a) = \mathbb{F}_q\}$ be the set of elements of $\mathbb{F}_q$ that generate $\mathbb{F}_q$ over $\mathbb{F}_2$. Then, there exists $\alpha \in \mathbb{F}_q$ such that $\alpha^2 + \alpha \in G$.*

Proof. Let $H = \{b^2 + b \mid b \in \mathbb{F}_q\}$. We show that $|G| > q/2$ and $|H| = q/2$, which means that $G \cap H$ is nonempty.

For G, note that $c \in \mathbb{F}_q \setminus G$ if and only if c lies in some maximal subfield of $\mathbb{F}_q$. For $q = 2^k$, the proper subfields of $\mathbb{F}_q$ are precisely $\mathbb{F}_{2^d}$ for $d \mid k$ and $d < k$. In particular, $|\mathbb{F}_{2^d}| \leq 2^{k/2}$. When $k = 3$, the only subfield of $\mathbb{F}_8$ is $\mathbb{F}_2$ so $|G| = 6$ and the result holds. For $k = 4$, the only maximal subfield of $\mathbb{F}_{16}$ is $\mathbb{F}_4$, so $|G| = 12$ and the result holds again. Now assume $k \geq 5$. Then consider $\omega(k)$, the number of distinct prime factors of k. There are thus $\omega(k)$ maximal subfields. We easily find

that $\omega(k) < k/2$ for $k \geq 5$. Letting $d_1, d_2, \ldots, d_{\omega(k)}$ be the maximal divisors of k (i.e., k/d_i is a prime), we get

$$\left| \bigcup_{i=1}^{\omega(k)} \mathbb{F}_{2^{d_i}} \right| < \sum_{i=1}^{\omega(k)} 2^{d_i} \leq \frac{k}{2} \cdot 2^{k/2} < 2^{k-1} = \frac{q}{2}.$$

Thus, $|G| > q/2$.

Now, for H, observe that for all $x, y \in \mathbb{F}_q$, we have $x^2 + x = y^2 + y$ if and only if $x + y = x^2 + y^2 = (x+y)^2$, which holds if and only if $x + y = 0$ or $x + y = 1$. Hence, $x^2 + x = y^2 + y$ if and only if $x = y$ or $x = y + 1$. As a result, we get $q/2$ distinct values of $x^2 + x$ as x runs through $\mathbb{F}_q$. So, $|H| = q/2$ and $G \cap H \neq \varnothing$. $\square$

Remark 3.4. Lemma 3.3 fails when $q = 4$. In this case, $\mathbb{F}_4 = \{0, 1, a, a+1\}$, where $a^2 + a + 1 = 0$. So, $\alpha^2 + \alpha$ equals 0 or 1 for all $\alpha \in \mathbb{F}_4$, and neither 0 nor 1 generates $\mathbb{F}_4$ over $\mathbb{F}_2$.

Definition 3.5. When q is odd, for each $b \in \mathbb{F}_q$ let $X_b = \left(\begin{smallmatrix} g & b \\ 0 & -g \end{smallmatrix} \right)$. When q is even and $q \geq 8$, let $\alpha \in \mathbb{F}_q$ be such that $\alpha^2 + \alpha$ generates $\mathbb{F}_q$ over $\mathbb{F}_2$, and for each $b \in \mathbb{F}_q$ let $Y_b = \left(\begin{smallmatrix} \alpha & b \\ 0 & \alpha+1 \end{smallmatrix} \right)$.

The subrings $\langle\!\langle X_b \rangle\!\rangle$ and $\langle\!\langle Y_b \rangle\!\rangle$ are the subrings we require to complete the covers of $T_2(\mathbb{F}_q)$. Proving that this is the case involves several lemmas and propositions.

Lemma 3.6. *When q is odd, g^2 generates $\mathbb{F}_q$ over $\mathbb{F}_p$.*

Proof. Note that $|\langle g^2 \rangle| = (q-1)/2$ as a multiplicative group. Combined with the element 0, we have $|\mathbb{F}_p(g^2)| \geq (q+1)/2$. Since $\mathbb{F}_p(g^2)$ is a subfield of $\mathbb{F}_q$, its order divides q. Hence, $\mathbb{F}_p(g^2)$ must be all of $\mathbb{F}_q$. $\square$

Lemma 3.7. *When q is odd, X_b, X_c, and I are linearly independent over $\mathbb{F}_q$ for all distinct $b, c \in \mathbb{F}_q$. When q is even and $q \geq 8$, Y_b, Y_c, and I are linearly independent over $\mathbb{F}_q$ for all distinct $b, c \in \mathbb{F}_q$.*

Proof. Assume that q is odd. Let $x_1, x_2, x_3 \in \mathbb{F}_q$ be such that $x_1 \cdot X_b + x_2 \cdot X_c + x_3 \cdot I = \left(\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix} \right)$. This corresponds to the matrix equation

$$\begin{pmatrix} g & g & 1 \\ b & c & 0 \\ -g & -g & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

The 3×3 matrix for this system has determinant $2g(c - b)$, which is nonzero because $\mathbb{F}_q$ has odd characteristic and $b \neq c$. Hence, $x_1 = x_2 = x_3 = 0$ and X_b, X_c, and I are linearly independent. The proof for characteristic 2 is the same, except that the determinant of the relevant matrix is $b + c$, which is still nonzero. So, Y_b, Y_c, and I are also linearly independent. $\square$

Proposition 3.8. *When q is odd, $\langle\!\langle X_b \rangle\!\rangle$ is maximal for each $b \in \mathbb{F}_q$ and has order q^2. When q is even and $q \geq 8$, $\langle\!\langle Y_b \rangle\!\rangle$ is maximal for each $b \in \mathbb{F}_q$ and has order q^2.*

Proof. Assume that q is odd. Observe that $X_b^2 = \left(\begin{smallmatrix} g^2 & 0 \\ 0 & g^2 \end{smallmatrix}\right)$. By Lemma 3.6, $\langle\!\langle X_b \rangle\!\rangle$ contains all of the scalar matrices in $T_2(\mathbb{F}_q)$. Since X_b itself is not scalar, Lemma 3.1 shows that $|\langle\!\langle X_b \rangle\!\rangle|$ is either q^2 or q^3. But, $\langle\!\langle X_b \rangle\!\rangle \neq T_2(\mathbb{F}_q)$ because $\langle\!\langle X_b \rangle\!\rangle$ is commutative, so $|\langle\!\langle X_b \rangle\!\rangle| = q^2$ and $\langle\!\langle X_b \rangle\!\rangle$ is maximal by Lemma 3.1. The proof for $\langle\!\langle Y_b \rangle\!\rangle$ is identical after noting that $Y_b^2 + Y_b = \left(\begin{smallmatrix} \alpha^2+\alpha & 0 \\ 0 & \alpha^2+\alpha \end{smallmatrix}\right)$. $\qquad\square$

Proposition 3.9. *Let $F = \{a \cdot I \mid a \in \mathbb{F}_q\}$ be the subring of scalar matrices in $T_2(\mathbb{F}_q)$:*

(1) *If q is odd, then $\langle\!\langle X_b \rangle\!\rangle \cap \langle\!\langle M \rangle\!\rangle = F$ for all $b \in \mathbb{F}_q$, and $\langle\!\langle X_b \rangle\!\rangle \cap \langle\!\langle X_c \rangle\!\rangle = F$ for all distinct $b, c \in \mathbb{F}_q$.*

(2) *If q is even and $q \geq 8$, then $\langle\!\langle Y_b \rangle\!\rangle \cap \langle\!\langle M \rangle\!\rangle = F$ for all $b \in \mathbb{F}_q$, and $\langle\!\langle Y_b \rangle\!\rangle \cap \langle\!\langle Y_c \rangle\!\rangle = F$ for all distinct $b, c \in \mathbb{F}_q$.*

Proof. We will prove part (1); the proof of part (2) is the same. Assume that q is odd. Certainly, $\langle\!\langle X_b \rangle\!\rangle \cap \langle\!\langle M \rangle\!\rangle$ contains F. Suppose that there exists $A \in \langle\!\langle X_b \rangle\!\rangle \cap \langle\!\langle M \rangle\!\rangle$ that is not scalar. Now, $\{X_b, I\}$ forms an $\mathbb{F}_q$-basis for $\langle\!\langle X_b \rangle\!\rangle$, so $A = a_1 \cdot X_b + a_2 \cdot I$ for some $a_1, a_2 \in \mathbb{F}_q$ with $a_1 \neq 0$. But then, $X_b = a_1^{-1}(A - a_2 \cdot I) \in \langle\!\langle M \rangle\!\rangle$, which is impossible by Proposition 3.2. Hence, $\langle\!\langle X_b \rangle\!\rangle \cap \langle\!\langle M \rangle\!\rangle = F$.

Similarly, we know that $\langle\!\langle X_b \rangle\!\rangle \cap \langle\!\langle X_c \rangle\!\rangle \supseteq F$ for distinct $b, c \in \mathbb{F}_q$. As above, if there exists a nonscalar matrix $A \in \langle\!\langle X_b \rangle\!\rangle \cap \langle\!\langle X_c \rangle\!\rangle$, then $X_b \in \langle\!\langle X_c \rangle\!\rangle$. But then, X_b, X_c, and I are all in $\langle\!\langle X_c \rangle\!\rangle$. By Lemma 3.7, these three matrices are linearly independent over $\mathbb{F}_q$, so $|\langle\!\langle X_c \rangle\!\rangle| = q^3$, which contradicts Proposition 3.8. Thus, $\langle\!\langle X_b \rangle\!\rangle \cap \langle\!\langle X_c \rangle\!\rangle = F$. $\qquad\square$

The results above are sufficient to compute $\sigma(T_2(\mathbb{F}_q))$ when q is odd, or q is even and $q \geq 8$. The cases $q = 2$ and $q = 4$ must be dealt with separately. When $q = 4$, we must rule out the possibility that $\sigma(T_2(\mathbb{F}_4))$ equals 3.

Lemma 3.10. $\sigma(T_2(\mathbb{F}_4)) \neq 3$.

Proof. A classification of all rings (with or without unity) with covering number 3 is given in [Lucchini and Maróti 2012, Theorem 1.2]. When applied to a ring R with unity, this classification says that $\sigma(R) = 3$ if and only if R has a residue ring isomorphic to either $\mathbb{F}_2 \times \mathbb{F}_2$ or the ring

$$\left\{ \begin{pmatrix} a & 0 & 0 \\ b & a & 0 \\ c & 0 & a \end{pmatrix} \,\middle|\, a, b, c \in \mathbb{F}_2 \right\}, \tag{3.11}$$

which has order 8. Now, $T_2(\mathbb{F}_4)$ has order 64, and any ideal of $T_2(\mathbb{F}_4)$ is an $\mathbb{F}_4$-vector space, and hence has order equal to a power of 4. So, $T_2(\mathbb{F}_4)$ contains no ideal of order 8, and hence has no residue ring isomorphic to the ring in (3.11). The only ideals of $T_2(\mathbb{F}_4)$ of order 16 are the maximal ideals

$$\left\{ \begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix} \,\middle|\, a, b \in \mathbb{F}_4 \right\} \quad \text{and} \quad \left\{ \begin{pmatrix} 0 & b \\ 0 & c \end{pmatrix} \,\middle|\, b, c \in \mathbb{F}_4 \right\}.$$

For each of these, the associated residue ring is isomorphic to $\mathbb{F}_4$. Hence, $T_2(\mathbb{F}_4)$ does not satisfy the conditions of [Lucchini and Maróti 2012, Theorem 1.2], and so $\sigma(T_2(\mathbb{F}_4)) \neq 3$. □

We can now prove Theorem 1.2 and Corollary 1.3, which are restated for convenience.

Theorem 1.2. *Let q be a prime power. Then $\sigma(T_2(\mathbb{F}_q)) = q + 1$ when $q \neq 4$, and $\sigma(T_2(\mathbb{F}_4)) = 4$.*

Proof. One may check that a cover of $T_2(\mathbb{F}_2)$ is formed by the three subrings

$$\left\{ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \right\},$$

$$\left\{ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \right\},$$

$$\left\{ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \right\}.$$

Since no ring can be covered by only two subrings, we get $\sigma(T_2(\mathbb{F}_2)) = 3$.

For $q = 4$, note that $\mathcal{I} = \left\{ \begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix} \mid b \in \mathbb{F}_4 \right\}$ is an ideal of $T_2(\mathbb{F}_4)$ and $T_2(\mathbb{F}_4)/\mathcal{I} \cong \mathbb{F}_4 \times \mathbb{F}_4$. By [Werner 2015, Theorem 5.3], $\mathbb{F}_4 \times \mathbb{F}_4$ is coverable and $\sigma(\mathbb{F}_4 \times \mathbb{F}_4) = 4$, so $\sigma(T_2(\mathbb{F}_4)) \leq 4$ by Lemma 2.2. But, $\sigma(T_2(\mathbb{F}_4)) \neq 3$ by Lemma 3.10, so we conclude that $\sigma(T_2(\mathbb{F}_4)) = 4$.

Now, assume that either q is odd, or that q is even and $q \geq 8$. By Propositions 3.2 and 3.8, $T_2(\mathbb{F}_q)$ contains $q + 1$ maximal subrings, each generated by a single matrix, and each of order q^2. These subrings are all distinct by Proposition 3.9, so each one must be part of any cover of $T_2(\mathbb{F}_q)$ by Lemma 2.2. Hence, $\sigma(T_2(\mathbb{F}_q)) \geq q + 1$. On the other hand, by Proposition 3.9 any pairwise intersection of these subrings is equal to the set of scalar matrices, so the union of all these subrings has cardinality

$$(q + 1)(q^2 - q) + q = q^3 = |T_2(\mathbb{F}_q)|.$$

Thus, this collection of $q+1$ subrings forms a cover, and hence $\sigma(T_2(\mathbb{F}_q)) = q+1$. □

Corollary 1.3. (1) *Let q be a prime power. If $n \geq 3$, then $\sigma(T_n(\mathbb{F}_q)) \leq q + 1$.*

(2) *Let R be a finite ring and let p be the smallest prime dividing the order of R. If R has $\mathbb{F}_p$ as a residue field, then $\sigma(T_n(R)) = p + 1$ for all $n \geq 2$. In particular, $\sigma(T_n(\mathbb{F}_p)) = p + 1$ for all $n \geq 2$.*

Proof. For (1), let $\mathcal{I}$ be the set of matrices in $T_n(\mathbb{F}_q)$ whose $(1, 1)$, $(1, 2)$, and $(2, 2)$ entries are all 0, and there are no restrictions on the other entries. Then, $\mathcal{I}$ is an ideal of $T_n(\mathbb{F}_q)$ and $T_n(\mathbb{F}_q)/\mathcal{I} \cong T_2(\mathbb{F}_q)$. So, $\sigma(T_n(\mathbb{F}_q)) \leq \sigma(T_2(\mathbb{F}_q)) \leq q + 1$ by Lemma 2.2 and Theorem 1.2.

For (2), note that when $n \geq 2$, $T_n(R)$ is coverable because it is noncommutative. So, $\sigma(T_n(R)) \geq p + 1$ by Lemma 2.3. Let J be an ideal of R such that $R/J \cong \mathbb{F}_p$. Let $\mathcal{J}$ be the set of matrices in $T_n(R)$ whose $(1, 1)$, $(1, 2)$, and $(2, 2)$ entries come from J and there are no restriction on the other entries. Then, $\mathcal{J}$ is an ideal of $T_n(R)$ and $T_n(R)/\mathcal{J} \cong T_2(\mathbb{F}_p)$. By Lemma 2.2, $\sigma(T_n(R)) \leq \sigma(T_2(\mathbb{F}_p))$, and so $\sigma(T_n(R)) = p + 1$ by Theorem 1.2.   □

We suspect that equality holds in part (1) of Corollary 1.3 (except when $q = 4$), and a few words are in order about our inability to prove this. The main obstruction to generalizing Theorem 1.2 for $n \geq 3$ is that the maximal subrings we desire to use are not generated by a single matrix. For instance, in the 3×3 case,

$$\left\{ \begin{pmatrix} a & b & c \\ 0 & a & d \\ 0 & 0 & e \end{pmatrix} \,\middle|\, a, b, c, d, e \in \mathbb{F}_q \right\}$$

(which generalizes $\langle\!\langle M \rangle\!\rangle$) is a maximal subring of $T_3(\mathbb{F}_q)$, but it is not commutative, and hence is not generated by a single matrix. Consequently, we cannot conclude that such a subring must be part of every cover of $T_3(\mathbb{F}_q)$.

It should be possible to compute $\sigma(T_n(\mathbb{F}_q))$ given the complete classification of maximal subrings of $T_n(\mathbb{F}_q)$. Unfortunately, such a classification is not known. Even in the case of 2×2 matrices, identifying all maximal subrings appears to be nontrivial. In this paper, we made use of the maximal subrings $\langle\!\langle M \rangle\!\rangle$, $\langle\!\langle X_b \rangle\!\rangle$, and $\langle\!\langle Y_b \rangle\!\rangle$, but other maximal subrings of $T_2(\mathbb{F}_q)$ exist. For instance, let R be a maximal subring of $\mathbb{F}_q$. Then, the subrings

$$\left\{ \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \,\middle|\, a \in R \text{ and } b, c \in \mathbb{F}_q \right\} \quad \text{and} \quad \left\{ \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \,\middle|\, a, b \in \mathbb{F}_q \text{ and } c \in R \right\}$$

are both maximal in $T_2(\mathbb{F}_q)$. Similar examples exist in $T_n(\mathbb{F}_q)$ when $n \geq 3$. Given these considerations, we propose the classification of maximal subrings of $T_n(\mathbb{F}_q)$ and the associated calculation of covering numbers as problems for further research.

References

[Bruckheimer et al. 1970] M. Bruckheimer, A. C. Bryan, and A. Muir, "Groups which are the union of three subgroups", *Amer. Math. Monthly* **77** (1970), 52–57. MR Zbl

[Cohn 1994] J. H. E. Cohn, "On *n*-sum groups", *Math. Scand.* **75**:1 (1994), 44–58. MR Zbl

[Crestani 2012] E. Crestani, "Sets of elements that pairwise generate a matrix ring", *Comm. Algebra* **40**:4 (2012), 1570–1575. MR Zbl

[Haber and Rosenfeld 1959] S. Haber and A. Rosenfeld, "Groups as unions of proper subgroups", *Amer. Math. Monthly* **66** (1959), 491–494. MR Zbl

[Kappe 2014] L.-C. Kappe, "Finite coverings: a journey through groups, loops, rings and semigroups", pp. 79–88 in *Group theory, combinatorics, and computing* (Boca Raton, FL, 2012), edited by R. F. Morse et al., Contemp. Math. **611**, Amer. Math. Soc., Providence, RI, 2014. MR Zbl

[Lucchini and Maróti 2010] A. Lucchini and A. Maróti, "Rings as the unions of proper subrings", preprint, 2010. arXiv

[Lucchini and Maróti 2012] A. Lucchini and A. Maróti, "Rings as the unions of proper subrings", *Algebr. Represent. Theory* **15**:6 (2012), 1035–1047. MR Zbl

[Scorza 1926] G. Scorza, "I gruppi che possono pensarsi come somma di tre loro sottogruppi", *Boll. Un. Mat. Ital.* **5** (1926), 216–218. Zbl

[Tomkinson 1997] M. J. Tomkinson, "Groups as the union of proper subgroups", *Math. Scand.* **81**:2 (1997), 191–198. MR Zbl

[Werner 2015] N. J. Werner, "Covering numbers of finite rings", *Amer. Math. Monthly* **122**:6 (2015), 552–566. MR Zbl

merrickcai@gmail.com   *Kings Park High School, Kings Park, NY, United States*

wernern@oldwestbury.edu   *Department of Mathematics, Computer and Information Science, SUNY College at Old Westbury, Old Westbury, NY, United States*

msp

Nonstandard existence proofs
for reaction diffusion equations

Connor Olson, Marshall Mueller and Sigurd B. Angenent

(Communicated by Suzanne Lenhart)

——— *To the memory of Terry Millar* ———

We give an existence proof for distribution solutions to a scalar reaction diffusion equation, with the aim of illustrating both the differences and the common ingredients of the nonstandard and standard approaches. In particular, our proof shows how the operation of taking the standard part of a nonstandard real number can replace several different compactness theorems, such as Ascoli's theorem and the Banach–Alaoglu theorem on weak*-compactness of the unit ball in the dual of a Banach space.

1. Introduction

1.1. *Reaction diffusion equations.* We consider the Cauchy problem for scalar reaction diffusion equations of the form

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2} + f(u(x,t)), \quad x \in \mathbb{R},\ t \geq 0, \tag{1a}$$

with prescribed initial condition

$$u(x, 0) = u_0(x). \tag{1b}$$

In the setting of reaction diffusion equations the function $u(x, t)$ represents the density at location $x \in \mathbb{R}$ and time $t \geq 0$ of some substance which diffuses, and simultaneously grows or decays due to chemical reaction, biological mutation, or some other process. The term $D\partial^2 u/\partial x^2$ in the PDE (1a) accounts for the change in u due to diffusion, while the nonlinear term $f(u)$ accounts for the reaction rates. The prototypical example of such a reaction diffusion equation is the Fisher–KPP equation, see [Kolmogorov et al. 1937; Fisher 1937], in which the reaction term is given by $f(u) = u - u^2$.

The Cauchy problem for the reaction diffusion equation (1a) is to find a function $u : \mathbb{R} \times [0, \infty) \to \mathbb{R}$ that satisfies the partial differential equation (1a) as well as the initial condition (1b). This is a classical problem, and the existence of such solutions is well known; see for example [Henry 1981; Pazy 1983]. As various techniques for constructing solutions are known, including the use of finite difference approximations to construct solutions (see [John 1982, Chapter 7.2]), our main goal is not to give another existence proof. Instead, we were inspired by several introductory texts on nonstandard analysis (notably, the undergraduate calculus text [Keisler 1976a], the more advanced introduction to the hyperreals [Goldblatt 1998], the "radically elementary approach" to probability in [Nelson 1987], as well as the blog post [Tao 2007]) and wanted to see what some standard existence proofs would look like in the language of nonstandard analysis.

Keisler [1976a] presented a proof of Peano's existence theorem for solutions to ordinary differential equations

$$\frac{dx}{dt} = f(t, x(t)), \quad x(0) = x_0, \tag{2}$$

using nonstandard analysis. One possible standard proof of Peano's theorem proceeds by constructing the numerical approximation to the solution by solving Euler's method for any small step size $\Delta t > 0$; i.e., one defines numbers $x_{i,\Delta t}$ by setting $x_{0,\Delta t} = x_0$ and then inductively solving

$$\frac{x_{i+1,\Delta t} - x_{i,\Delta t}}{\Delta t} = f(i \Delta t, x_{i,\Delta t}), \quad i = 0, 1, 2, \ldots. \tag{3}$$

The function $x_{\Delta t} : [0, \infty) \to \mathbb{R}$ obtained by linearly interpolating between the values $x_{\Delta t}(i \Delta t) = x_{i,\Delta t}$ is Euler's numerical approximation to the solution of the differential equation (2). The standard analysis proof of Peano's existence theorem then uses Ascoli's compactness theorem to extract a sequence of step sizes $\Delta t_n \to 0$ such that the approximate solutions $x_{\Delta t_n}(t)$ converge uniformly to some function $\tilde{x} : [0, \infty) \to \mathbb{R}$ and concludes by showing that the limit $\tilde{x}$ is a solution to the differential equation (2).

The nonstandard proof in [Keisler 1976a] follows the same outline, but one notable feature of this proof is that instead of using Ascoli's theorem, one "simply" chooses the step size Δt to be a positive infinitesimal number. The approximate solution then takes values in the hyperreals, and instead of applying a compactness theorem (Ascoli's in this case), one "takes the standard part" of the nonstandard approximate solution. The proof is then completed by showing that the function that is obtained actually satisfies the differential equation.

The standard and nonstandard proofs have some common ingredients. In both proofs one must find suitable estimates for the approximate solutions $x_{i,\Delta t}$, where the estimates should not depend on the step size Δt. Namely, the approximate

solutions $x_{\Delta t}$ should be uniformly bounded, and they should be uniformly Lipschitz continuous, i.e., $|x_{\Delta t}(t) - x_{\Delta t}(s)| \leq L|t - s|$ for all $t, s \in [0, \infty)$, $\Delta t > 0$. In the standard proof these estimates allow one to use Ascoli's theorem; in the nonstandard proof they guarantee that the standard part of the approximating solution with infinitesimally small Δt still defines a continuous function on the standard reals.

There appear to be two main differences between the standard and nonstandard proofs. The first, very obviously, is that the nonstandard setting allows one to speak rigorously of infinitely small numbers, and thereby avoid the need to consider limits of sequences. The second difference is that the process of "taking the standard part" of a hyperreal number acts as a replacement for one compactness theorem or another: in the nonstandard proof of Peano's theorem one avoids Ascoli's theorem by taking standard parts. This too is probably well known in some circles (Terry Tao [2007] made the point in a blog post), but is not as obviously stated in the nonstandard analysis texts we have seen.

In this paper we intend to further illustrate this point by proving an existence theorem for weak or distributional solutions of certain partial differential equations that is analogous to the proof of Peano's theorem sketched above (see Section 2.1 below for a very short summary of the theory of distributions). Thus, to "solve" the reaction diffusion equation (1a) we choose space and time steps $\Delta x > 0$ and $\Delta t > 0$, and discretize the PDE by replacing the second derivative with a second difference quotient and the time derivative with a forward difference quotient, resulting in a finite difference equation,

$$\frac{u(x, t+\Delta t) - u(x, t)}{\Delta t} = D \frac{u(x+\Delta x, t) - 2u(x, t) + u(x-\Delta x, t)}{(\Delta x)^2} + f(u(x, t)). \quad (4)$$

This kind of discretization is very common in numerical analysis;[1] see [LeVeque 2007; Press et al. 2007]. For given initial data (but no boundary data) one can use this difference equation to inductively compute the values of $u(x, t)$ for all (x, t) in a triangular grid (see Figure 1).

Given a solution $U(x, t)$ of the difference equation (4), one can define a generalized function, or distribution,

$$\langle U, \varphi \rangle \overset{\text{def}}{=} \sum_t \sum_x U(x, t) \varphi(x, t) \Delta x \Delta t. \quad (5)$$

---

[1]As one of the reviewers pointed out, we have chosen the simplest among the many other difference schemes that approximate the reaction diffusion equation (1a). For many other schemes the arguments in this paper could probably be adapted, although for implicit schemes one would have to arbitrarily select boundary values and observe that these will only have an infinitesimal effect on the solution in bounded regions of the form $|x|, |t| \leq R$ for any standard $R \in \mathbb{R}$. We leave it to the interested reader to pursue these questions.

In a standard existence proof of weak solutions to the equation one would now use a compactness theorem to extract a sequence $(\Delta x_i, \Delta t_i) \to (0, 0)$ for which the corresponding distributions U_i converge in the sense of distributions, and then show that the limiting distribution satisfies the PDE (1a). The compactness theorem that is required in this proof is the Banach-Alaoglu theorem about weak*-compactness in duals of Banach spaces (in our case, $L^\infty(\mathbb{R}^2)$, which is the Banach space dual of $L^1(\mathbb{R}^2)$).

The nonstandard proof, which we give in this paper, avoids the compactness theorem (or notions of Lebesgue integration required to define L^∞) by letting Δx and Δt be infinitesimal positive hyperreals, and by taking the standard part of the expression on the right in (5). In both the standard and nonstandard settings this approach works for the linear heat equation, that is, in the case where the reaction term $f(u)$ is absent (i.e., $f(u) \equiv 0$). The nonlinear case is a bit more complicated because there is no adequate definition of $f(u)$ when u is a distribution rather than a pointwise-defined function. In both the standard and nonstandard proofs we overcome this by proving that the approximating functions are Hölder continuous, so that $f(u(x, t))$ can be defined. In the standard proof this again allows one to use Arzelà–Ascoli and extract a convergent subsequence. However, since the domain $\mathbb{R}^2$ is not compact, Arzelà–Ascoli cannot be applied directly, and the standard proof therefore requires one to apply the compactness theorem on an increasing sequence of compact subsets $K_i \subset \mathbb{R}^2$, after which Cantor's diagonal trick must be invoked to get a sequence of functions that converges uniformly on every compact subset of $\mathbb{R}^2$. As we show below, these issues do not come up in the nonstandard proof.

1.2. *Comments on nonstandard analysis.*

In nonstandard analysis one exploits the existence of an ordered field $*\mathbb{R}$ called the hyperreal numbers, which contains the standard real numbers $\mathbb{R}$, but also contains infinitesimally small numbers, i.e., numbers $x \in *\mathbb{R}$ with $x \neq 0$ that violate the Archimedean axiom by satisfying $n|x| < 1$ for all standard integers $n \in \mathbb{N}$. When two hyperreals $x, y \in *\mathbb{R}$ differ by an infinitesimal, one writes $x \approx y$. For each hyperreal number $x \in *\mathbb{R}$ there is a unique standard real number $\text{St}(x)$, called the standard part of x, such that $x \approx \text{St}(x)$. Beyond this simple description of the hyperreals we will not even try to give an exposition of nonstandard analysis in this paper and instead refer the reader to the many texts that have been written on the subject; see, e.g., a very incomplete list: [Keisler 1976a; 1976b; Goldblatt 1998; Nelson 1987; Albeverio et al. 1986; Tao 2007].

There are a few different approaches to using the hyperreals. Keisler [1976a] gave an axiomatic description of the hyperreals and their relation with the standard reals. In this approach, functions that are defined for standard reals automatically extend to the hyperreals, according to the *transfer principle.* A different approach

that also begins with an axiomatic description of the hyperreals can be found in Nelson's "radically elementary" treatment of probability theory [1987].

Our point of view in this paper is that of internal set theory as explained in [Goldblatt 1998] (see also the "instructor's guide" [Keisler 1976b] to his calculus text). Goldblatt explains the construction of the hyperreals using nonprincipal ultrafilters (which can be thought of as analogous to the construction of real numbers as equivalence classes of Cauchy sequences of rational numbers). He then extends this construction and defines internal sets, internal functions, etc.

2. Distribution solutions

2.1. *The definition of distributions.* We recall the definition of a "generalized function," i.e., of a distribution, which can be found in many textbooks on real analysis, such as [Folland 1999].

A real-valued function f on $\mathbb{R}^2$ is traditionally defined by specifying its values $f(x, y)$ at each point $(x, y) \in \mathbb{R}^2$. In the theory of distributions a generalized function f is defined by specifying its weighted averages

$$\langle f, \varphi \rangle = \int_{\mathbb{R}^2} f(x, y) \, \varphi(x, y) \, dx \, dy \tag{6}$$

for all so-called "test functions" φ. A test function is any function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ that is infinitely often differentiable, and which vanishes outside a sufficiently large ball $B_R = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < R\}$ whose radius R is allowed to depend on the particular test function. The set of all test functions, which is denoted by $C_c^\infty(\mathbb{R}^2)$, or sometimes by $\mathcal{D}(\mathbb{R}^2)$, is an infinite-dimensional vector space. *By definition,* a distribution is any linear functional $T : C_c^\infty(\mathbb{R}^2) \to \mathbb{R}$. The most common notation for the value of a distribution T applied to a test function φ is $\langle T, \varphi \rangle$. For instance, if $f : \mathbb{R}^2 \to \mathbb{R}$ is a continuous function, then (6) defines f as a distribution. The canonical example of a distribution that does not correspond to a function f is the Dirac delta function, which is defined by

$$\langle \delta, \varphi \rangle \stackrel{\text{def}}{=} \varphi(0, 0).$$

The full definition of a distribution T includes the requirement that the value $\langle T, \varphi \rangle$ depend continuously on the test function φ. To state this continuity condition precisely one must introduce a notion of convergence in the space of test functions $C_c^\infty(\mathbb{R}^2)$. We refer the reader to [Folland 1999] for the details, and merely observe that a sufficient condition for a linear functional $\varphi \mapsto \langle T, \varphi \rangle$ to be a distribution is that there exist a constant C such that

$$|\langle T, \varphi \rangle| \le C \iint_{\mathbb{R}^2} |\varphi(x, y)| \, dx \, dy \tag{7}$$

holds for all test functions φ. Alternatively, if a constant C exists such that

$$|\langle T, \varphi \rangle| \leq C \sup_{(x,y) \in \mathbb{R}^2} |\varphi(x, y)| \tag{8}$$

holds for all $\varphi \in C_c^\infty(\mathbb{R}^2)$, then T also satisfies the definition of a distribution. The conditions (7) and (8) are not equivalent: either one of these implies that T is a distribution.

2.2. Distributions defined by nonstandard functions on a grid.

Let dx, dy be two positive infinitesimal hyperreal numbers, and let N, M be two positive hyperintegers such that $N \, dx$ and $M \, dy$ are unlimited. Consider the rectangular grid

$$G = \{(k \, dx, l \, dy) \in {}^*\mathbb{R}^2 \mid k, l \in {}^*\mathbb{N}, \ |k| \leq N, \ |l| \leq M\}. \tag{9}$$

From the point of view of nonstandard analysis and internal set theory, G is a hyperfinite set, and for any internal function $f : G \to {}^*\mathbb{R}$ there is an $(x, y) \in G$ for which $f(x, y)$ is maximal.

Lemma 2.2.1. *If $g : \mathbb{R}^2 \to \mathbb{R}$ is a continuous function with compact support, then*

$$\int_{\mathbb{R}^2} g(x, y) \, dx \, dy \approx \sum_{(x,y) \in G} g(x, y) \, dx \, dy.$$

Recall that $x \approx y$ means that $x - y$ is infinitesimal.

Proof. The statement of the lemma is very close to the nonstandard definition of the Riemann integral of a continuous function, the only difference being that we are integrating over the unbounded domain $\mathbb{R}^2$ rather than a compact rectangle $[-\ell, \ell] \times [-\ell, \ell] \subset \mathbb{R}^2$. Since the function g has compact support, there is a real $\ell > 0$ such that $g(x, y) = 0$ outside the square $[-\ell, \ell] \times [-\ell, \ell]$. By definition we then have

$$\int_{\mathbb{R}^2} g(x, y) \, dx \, dy = \int_{-\ell}^{\ell} \int_{-\ell}^{\ell} g(x, y) \, dx \, dy.$$

Choose hyperintegers $L, L' \in {}^*\mathbb{N}$ for which

$$L \, dx \leq \ell < (L+1) \, dx \quad \text{and} \quad L' \, dy \leq \ell < (L'+1) \, dy.$$

Then the nonstandard definition of the Riemann integral implies

$$\int_{-\ell}^{\ell} \int_{-\ell}^{\ell} g(x, y) \, dx \, dy \approx \sum_{k=-L}^{L} \sum_{l=-L'}^{L'} g(k \, dx, l \, dy) \, dx \, dy.$$

Finally, if $(x, y) \in G$ then $g(x, y) = 0$ unless $|x| \leq \ell$ and $|y| \leq \ell$, so that

$$\sum_{k=-L}^{L} \sum_{l=-L'}^{L'} g(k \, dx, l \, dy) \, dx \, dy = \sum_{(x,y) \in G} g(x, y) \, dx \, dy. \qquad \square$$

Lemma 2.2.2. *Suppose that* $f : G \to {}^*\mathbb{R}$ *is a hyperreal-valued function which is bounded, in the sense that there exists a limited $C > 0$ such that $|f(x, y)| \leq C$ for all $(x, y) \in G$. Then the expression*

$$\langle T_f, \varphi \rangle \overset{\text{def}}{=} \text{St}\left(\sum_{(x,y) \in G} f(x, y)\, \varphi(x, y)\, dx\, dy \right) \tag{10}$$

defines a distribution on $\mathbb{R}^2$.

   If the function f is the nonstandard extension of a (standard) continuous function $f : \mathbb{R}^2 \to \mathbb{R}$, then the distribution T_f coincides with the distribution defined by (6).

Proof. We first verify that the distribution is well-defined. Since $|f(x, y)| \leq C$ for all (x, y) we have

$$\left| \sum_{(x,y) \in G} f(x, y)\, \varphi(x, y)\, dx\, dy \right| \leq C \sum_{(x,y) \in G} |\varphi(x, y)|\, dx\, dy \approx C \int_{\mathbb{R}^2} |\varphi(x, y)|\, dx\, dy.$$

Hence the sum in the definition (10) of $\langle T_f, \varphi \rangle$ is a limited hyperreal, whose standard part is a well-defined real number which satisfies

$$|\langle T_f, \varphi \rangle| \leq C \int_{\mathbb{R}^2} |\varphi(x, y)|\, dx\, dy.$$

Therefore T_f is a well-defined distribution.

   Let $\langle f, \varphi \rangle$ be defined as in (6). Fix φ. We then have

$$\int_{\mathbb{R}^2} f(x, y)\varphi(x, y)\, dx\, dy \approx \sum_{(x,y) \in G} f(x, y)\varphi(x, y)\, dx\, dy,$$

which implies that the distribution defined in (6) coincides with T_f.   □

3. The Cauchy problem for the heat equation

In this section we recall the definition of distribution solutions to the Cauchy problem for the heat equation and show how, by solving the finite difference approximation to the heat equation on a hyperfinite grid, one can construct a distribution solution to the Cauchy problem.

3.1. *Formulation in terms of distributions.* We consider the Cauchy problem for the linear heat equation $u_t = u_{xx}$ with bounded and continuous initial data $u(x, 0) = u_0(x)$. Without losing generality we may assume that the diffusion coefficient D is 1, e.g., by nondimensionalizing space and time and introducing $\tau = Dt$ as the new time variable.

Definition 3.1.1. A distribution u on $\mathbb{R}^2$ is a solution to the heat equation $u_t = u_{xx}$ with initial data u_0 if u satisfies

$$u_t - u_{xx} = u_0(x)\,\delta(t), \quad x \in \mathbb{R},\ t \in \mathbb{R}, \tag{11}$$

in the sense of distributions, and if $u = 0$ for $t \le 0$.

Equality in the sense of distributions in (11) means that both sides of the equation are to be interpreted as distributions, and that they should yield the same result when evaluated on any test function $\varphi \in C_c^\infty(\Omega)$. To explain this in more detail, recall that δ is Dirac's delta function, so that the action of the right-hand side in (11) on a test function is

$$\langle u_0(x)\,\delta(t), \varphi \rangle \overset{\mathrm{def}}{=} \int_{\mathbb{R}} u_0(x)\,\varphi(x, 0)\,dx.$$

The definition of distributional derivative [Folland 1999, Chapter 9] says that the action of the left-hand side in (11) is given by

$$\langle u_t - u_{xx}, \varphi \rangle = \langle u, -\varphi_t - \varphi_{xx} \rangle.$$

If the distribution u is given by a function $u : \mathbb{R}^2 \to \mathbb{R}$ which vanishes for $t < 0$ and is continuous for $t \ge 0$ (so that it has a simple jump discontinuity at $t = 0$) then we get

$$\langle u_t - u_{xx}, \varphi \rangle = \int_{\mathbb{R}} \int_0^\infty u(x, t)\{-\varphi_t - \varphi_{xx}\}\,dt\,dx.$$

A piecewise continuous function u therefore satisfies (11) in the sense of distributions if

$$\int_{\mathbb{R}} u_0(x)\varphi(x, 0)\,dt + \int_{\mathbb{R}} \int_0^\infty u(x, t)\{\varphi_t + \varphi_{xx}\}\,dt\,dx = 0 \tag{12}$$

for all test functions $\varphi \in C_c^\infty(\mathbb{R}^2)$. This is one form of the classical definition of a weak solution to the Cauchy problem.

3.2. _The finite difference equation._ To construct a distribution solution to (11) we introduce a grid with spacing dx and dt, and replace the differential equation by the simplest finite difference scheme that appears in numerical analysis. If u is the solution to the differential equation, then we write U for the approximating solution to the finite difference equation, using the following common notation for finite differences:

$$D_x^+ U(x, t) = \frac{U(x + dx, t) - U(x, t)}{dx},$$

$$D_x^- U(x, t) = \frac{U(x, t) - U(x - dx, t)}{dx},$$

$$D_t^+ U(x, t) = \frac{U(x, t + dt) - U(x, t)}{dt}.$$

See for example, [LeVeque 2007, Chapter 1]. With this notation

$$D_x^2 U(x,t) \stackrel{\text{def}}{=} D_x^+ D_x^- U(x,t) = \frac{U(x+dx,t) - 2U(x,t) + U(x-dx,t)}{(dx)^2}.$$

The operators D_x^+, D_x^-, and D_t^+ all commute. A finite difference equation corresponding to the heat equation $u_t = u_{xx}$ is then $D_t^+ U = D_x^2 U$, i.e.,

$$\frac{U(x,t+dt) - U(x,t)}{dt} = \frac{U(x+dx,t) - 2U(x,t) + U(x-dx,t)}{(dx)^2}. \tag{13}$$

We can solve this algebraic equation for $U(x, t+dt)$, resulting in

$$U(x,t+dt) = \alpha U(x-dx,t) + (1-2\alpha)U(x,t) + \alpha U(x+dx,t), \tag{14}$$

where

$$\alpha \stackrel{\text{def}}{=} \frac{dt}{(dx)^2}.$$

3.3. *The approximate solution.* Let $N \in {}^*\mathbb{N}$ be an unlimited hyperfinite integer, and assume that dx and dt are positive infinitesimals. Assume moreover that N is so large that both $N\,dt$ and $N\,dx$ are unlimited hyperreals. We then consider the hyperfinite grid

$$G_C = \{(m\,dx, n\,dt) \mid m, n \in {}^*\mathbb{N}, |m| + n \le N\}.$$

See Figure 1. The initial function $u_0 : \mathbb{R} \to \mathbb{R}$ extends to an internal function $u_0 : {}^*\mathbb{R} \to {}^*\mathbb{R}$. By assumption there is a $C \in \mathbb{R}$ such that $|u_0(x)| \le C$ for all $x \in \mathbb{R}$, so this also holds for all $x \in {}^*\mathbb{R}$.

We define $U : G_C \to {}^*\mathbb{R}$ by requiring:

- $U(x, 0) = u_0(x)$ for all x with $(x, 0) \in G_C$, i.e., for all $x = k\,dx$ with $k \in \{-N, \ldots, +N\}$.

- U satisfies (4) or, equivalently, (14) at all $(x, t) = (m\,dx, n\,dt) \in G_C$ with $|m| + n < N$.

Theorem 3.3.1. *Let $U : G_C \to {}^*\mathbb{R}$ be the hyperreal solution of the finite difference scheme* (13) *with initial values $U(x, 0) = u_0(x)$, and suppose that $\alpha \le \frac{1}{2}$. Then the expression*

$$\langle u, \varphi \rangle \stackrel{\text{def}}{=} \text{St}\left(\sum U(x,t)\varphi(x,t)\,dx\,dt \right), \quad \varphi \in C_c^\infty(\mathbb{R}^2), \tag{15}$$

defines a distribution on $\mathbb{R}^2$ that satisfies (11).

To show that this expression does indeed define a distribution we must show that the $U(x, t)$ are bounded by a standard real number. This follows from a discrete version of the maximum principle, which we will again use in Section 4, so we state
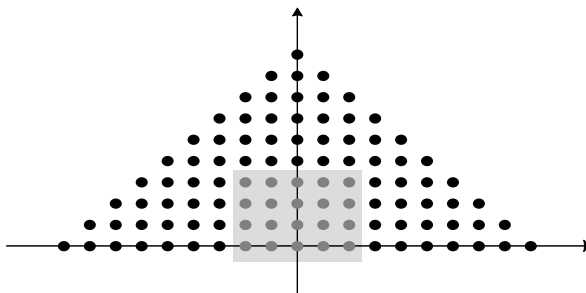
Figure 1. The triangular grid G_C; if $U(x, t)$ is known at all grid points at the bottom of the triangle, then the finite difference equation (13) uniquely determines $U(x, t)$ at all other grid points.

it in slightly greater generality than needed in this section. The lemma explains why we need the condition $\alpha = dt/(dx)^2 \leq \frac{1}{2}$ and is well known in numerical analysis as a necessary condition for stability of the finite difference scheme.

Lemma 3.3.2 (Gronwall-type estimate). *Let $W : G_C \to {}^*\mathbb{R}$ satisfy $W(x, t) \geq 0$ for all $(x, t) \in G_C$, and suppose that for some nonnegative $m \in \mathbb{R}$ one has*

$$W(x, t+dt) \leq \alpha W(x+dx, t) + (1 - 2\alpha) W(x, t) + \alpha W(x - dx, t) + dt\, m W(x, t)$$

at all $(x, t) \in G_C$. For each $t = n\, dt$ with $0 \leq n \leq N$ consider[2]

$$w(t) \overset{\text{def}}{=} \max_x W(x, t).$$

If $0 \leq \alpha \leq \frac{1}{2}$ then

$$w(t) \leq e^{mt} w(0).$$

Proof. (Compare [John 1982, §7.2, Lemma I].) The assumption on α implies that $\alpha \geq 0$ and $1 - 2\alpha \geq 0$. Hence for all x with $(x, t + dt) \in G_C$ we have

$$W(x, t+dt) = \alpha W(x+dx, t) + (1-2\alpha) W(x, t) + \alpha W(x-dx, t) + dt\, m W(x, t)$$
$$\leq (\alpha + (1-2\alpha) + \alpha + m\, dt) w(t)$$
$$= (1 + m\, dt) w(t).$$

Taking the maximum over x we see that $w(t + dt) \leq (1 + m\, dt) w(t) \leq e^{m\, dt} w(t)$. By induction we then have for $t = n\, dt$ that $w(t) \leq (e^{m\, dt})^n w(0) = e^{mt} w(0)$. $\qquad\square$

---

[2]For any given $t = n\Delta t$ there are infinitely many hyperreal numbers $W(x, t)$, so the standard analyst may be surprised to see "max" instead of "sup" in the definition of $w(t)$. However, in the internal-set-theory interpretation, the set of numbers $\{W(x, t) : (x, t) \in G_C\} = \{W(m\Delta x, n\Delta t) : |m| \leq N - n\}$ is a hyperfinite internal set of real numbers, indexed by $m \in \{0, \pm 1, \pm 2, \ldots, \pm(N - n)\}$. Therefore one of the numbers $W(m\Delta x, n\Delta t)$ is the largest, so that the maximum is a well-defined hyperreal number.

3.4. *Proof of Theorem 3.3.1.* The relation (14) which defines $U(x, t)$ implies that $W(x, t) \stackrel{\text{def}}{=} |U(x, t)|$ satisfies

$$W(x, t+dt) = |\alpha U(x-dx, t) + (1-2\alpha)U(x, t) + \alpha U(x+dx, t)|$$
$$\leq \alpha W(x-dx, t) + (1-2\alpha)W(x, t) + \alpha W(x+dx, t),$$

where we have used $\alpha \geq 0$ and $1 - 2\alpha \geq 0$.

Since the initial condition is bounded by $|U(x, 0)| = |u_0(x)| \leq M$, the Gronwall-type Lemma 3.3.2 implies that $|U(x, t)| \leq M$ for all $(x, t) \in G_C$. According to Lemma 2.2.2 this implies that the expression (15) does define a distribution u on $\mathbb{R}^2$.

We want to prove that u satisfies the heat equation in the sense of distributions; i.e., we want to show for any test function φ that

$$\langle u_t - u_{xx}, \varphi \rangle = \langle u_0(x)\,\delta(t), \varphi \rangle = \int_{\mathbb{R}} u_0(x)\,\varphi(x, 0)\,dx.$$

First, we see from the definition of distributional derivative that

$$\langle u_t - u_{xx}, \varphi \rangle = -\langle u, \varphi_t + \varphi_{xx} \rangle.$$

We then have from the definition of u that

$$-\langle u, \varphi_t + \varphi_{xx} \rangle \approx \sum_{(x,t)\in G_C} -U(x, t)(\varphi_t(x, t) + \varphi_{xx}(x, t))\,dx\,dt \stackrel{\text{def}}{=} T.$$

Using Taylor's formula we replace the partial derivatives of the test function with its corresponding finite differences; i.e., we write $\varphi_t(x, t) = D_t^+ \varphi(x, t) + \varepsilon_t(x, t)$ and $\varphi_{xx}(x, t) = D_x^2 \varphi(x, t) + \varepsilon_{xx}(x, t)$, where $\varepsilon_t, \varepsilon_{xx} : G_C \to {}^*\mathbb{R}$ are the infinitesimal error terms in the Taylor expansion. Substituting these gives us

$$T = \sum_{G_C} -U(x, t)(D_t^+ \varphi(x, t) + D_x^2 \varphi(x, t) + \varepsilon_t + \varepsilon_{xx})\,dx\,dt.$$

We can split this sum into three parts, $T = T_1 + T_2 + T_3$, with

$$T_1 = \sum_{G_C} -U(x, t)D_t^+ \varphi(x, t)\,dx\,dt,$$

$$T_2 = \sum_{G_C} -U(x, t)D_x^2 \varphi(x, t)\,dx\,dt,$$

$$T_3 = \sum_{G_C} -U(x, t)(\varepsilon_t + \varepsilon_{xx})\,dx\,dt.$$

We will first handle the error term, T_3. Since the test function φ has compact support, there exists a real $\ell > 0$ such that $\varphi = 0$ outside the rectangle $\Omega = [-\ell, \ell] \times [-\ell, \ell]$. The errors in the Taylor expansion therefore also vanish outside of Ω so that we

can write T_3 as

$$T_3 = \sum_{\Omega \cap G_C} -U(x, t)(\varepsilon_t + \varepsilon_{xx})\, dx\, dt.$$

The key to estimating this sum is that we can estimate all the errors $\varepsilon_t(x, t)$ and $\varepsilon_{xx}(x, t)$ by one fixed infinitesimal $\varepsilon > 0$ that does not depend on (x, t). Indeed, the grid G_C is a hyperfinite internal set, and therefore any internal function such as $|\varepsilon_t| : G_C \to {}^*\mathbb{R}$ attains its largest value at one of the $(x, t) \in G_C$, say at (x_1, t_1). Then $|\varepsilon_t(x, t)| \le |\varepsilon_t(x_1, t_1)|$ for all $(x, t) \in G_C$. Similarly, there is an $(x_2, t_2) \in G_C$ that maximizes $|\varepsilon_{xx}(x, t)|$. Now define

$$\varepsilon_1 = |\varepsilon_t(x_1, t_1)|, \quad \varepsilon_2 = |\varepsilon_{xx}(x_2, t_2)|.$$

Then both ε_1 and ε_2 are positive infinitesimals for which

$$|\varepsilon_t(x, t)| \le \varepsilon_1, \quad |\varepsilon_{xx}(x, t)| \le \varepsilon_2$$

hold at all grid points $(x, t) \in G_C$.

If we let $\varepsilon = \max\{\varepsilon_1, \varepsilon_2\}$, then $|T_3| \le \sum_{G_C \cap \Omega} 2U(x, t)\, \varepsilon\, dx\, dt$. By the construction of U, we have $|U(x, t)| \le M$ for all $(x, t) \in G_C$, so we get

$$|T_3| \le \sum_{G_C \cap \Omega} 2M\varepsilon\, dx\, dt \le 2M\varepsilon\, dx\, dt \frac{\ell}{2\, dx} \frac{\ell}{dt} = M\ell^2\varepsilon,$$

which is infinitesimal, so T_3 is infinitesimal.

From the definition we have $T_1 = \sum_{G_C} -U(x, t)(\varphi(x, t + dt) - \varphi(x, t))\, dx$. Using the compact support of φ, we can then rewrite this sum as

$$T_1 = -\sum_{k=-K}^{K} \sum_{l=0}^{L+1} U(k\, dx, l\, dt)\{\varphi(k\, dx, (l+1)\, dt) - \varphi(k\, dx, l\, dt)\}\, dx,$$

where $K\, dx \approx \ell$ and $L\, dt \approx \ell$.

Applying summation by parts to this sum we then get

$$T_1 = \sum_{k=-K}^{K} \left\{ U(k\, dx, 0)\, \varphi(k\, dx, 0) + \sum_{l=0}^{L} \varphi(k\, dx, (l+1)\, dt)\, D_t^+ U(k\, dx, l\, dt)\, dt \right\} dx.$$

Next, for T_2, we can split the sum into two parts.

$$T_2 = \sum_{l=0}^{L} \sum_{k=-K-1}^{K+1} -U(k\, dx, l\, dt) D_x^+ \varphi(k\, dx, l\, dt)\, dt\, dx$$

$$+ \sum_{l=0}^{L} \sum_{k=-K-1}^{K+1} U(k\, dx, l\, dt) D_x^- \varphi(k\, dx, l\, dt)\, dt\, dx. \quad (16)$$

Applying summation by parts again to both sums we get

$$T_2 = -\sum_{l=0}^{L}\sum_{k=-K}^{K} \varphi(k\,dx, l\,dt)\, D_x^2 U(k\,dx, l\,dt)\, dx\,dt.$$

Putting the terms T_1, T_2, T_3 all together, we have, because $T_3 \approx 0$,

$T_1 + T_2 + T_3 \approx T_1 + T_2$

$$= \sum_{k=-K}^{K} U(k\,dx, 0)\varphi(k\,dx, 0)\,dx$$

$$+ \sum_{k=-K}^{K}\sum_{l=0}^{L} \varphi(k\,dx, l\,dt)$$
$$\times (D_t^+ U(k\,dx, l\,dt) - D_x^2 U(k\,dx, l\,dt))\,dx\,dt. \quad (17)$$

Since U satisfies the difference equation $D_t^+ U = D_x^2 U$ at all grid points this reduces to

$$T = T_1 + T_2 + T_3 \approx \sum_{k=-K}^{K} U(k\,dx, 0)\,\varphi(k\,dx, 0)\,dx \approx \sum_{k=-K}^{K} u_0(k\,dx)\,\varphi(k\,dx, 0)\,dx.$$

Taking the standard part we get the distribution

$$\mathrm{St}(T) = \mathrm{St}\!\left(\sum_{k=-K}^{K} u_0(k\,dx)\,\varphi(k\,dx, 0)\,dx\right) = \int_{-\ell}^{\ell} u_0(x)\,\varphi(x, 0)\,dx = \langle u_0(x)\,\delta(t), \varphi\rangle.$$

This completes the proof that $\langle u_t - u_{xx}, \varphi\rangle = \langle u_0(x)\delta(t),\,\varphi\rangle$ for all test functions φ, and thus that u is a distributional solution of (11).

3.5. *Comments on the proof.* In our construction of solutions to the linear heat equation we completely avoided estimating derivatives of the approximate solution U. The only estimate we used was that the approximate solution $U(x, t)$ has the same upper bound as the given initial function u_0.

We assumed that the initial function u_0 is continuous. The one place in the proof where we needed this assumption was at the end, when we used the fact that for continuous functions $f : \mathbb{R} \to \mathbb{R}$ one has

$$\sum_{|k|\le K} f(k\,dx)\,dx \approx \int_{-\ell}^{\ell} f(x)\,dx$$

and applied this to the function $f(x) = u_0(x)\,\varphi(x, 0)$.

4. The Cauchy problem for a reaction diffusion equation

We consider the reaction diffusion equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(u(x, t)), \quad x \in \mathbb{R},\ t > 0, \quad (1a)$$

with initial condition (1b). It is known that one cannot expect solutions to exist for all times $t > 0$ without imposing some growth conditions on the nonlinearity $f(u)$. We will assume that f is a Lipschitz continuous function;[3] i.e., for some positive real K_1 one has

$$\text{for all } u, v \in \mathbb{R}, \quad |f(u) - f(v)| \leq K_1 |u - v|. \tag{18}$$

This implies that $f(u)$ grows at most linearly in u:

$$\text{for all } u \in \mathbb{R}, \quad |f(u)| \leq K_0 + K_1|u|, \tag{19}$$

where $K_0 \overset{\text{def}}{=} |f(0)|$.

In contrast to the linear heat equation, (1a) contains the nonlinear term $f(u)$, which is meaningless if u is an arbitrary distribution. One can follow the same procedure as in the previous section; i.e., one can replace the differential equation by a finite difference scheme on the hyperfinite grid G_C and construct an approximating solution $U : G_C \to {}^*\mathbb{R}$. After establishing suitable bounds one can then show that by taking standard parts as in Lemma 2.2.2, both $U(x, t)$ and $f(U(x, t))$ define distributions u and F on $\mathbb{R}^2$. The problem is to give a meaning to the claim that "$F = f(u)$", because u is merely a distribution and can therefore not be substituted in a nonlinear function. In this section we show how to overcome this problem by adding the assumption that the initial function is Lipschitz continuous, i.e.,

$$\text{for all } x, y \in \mathbb{R}, \quad |u_0(x) - u_0(y)| \leq L|x - y| \tag{20}$$

for some real $L > 0$, and showing that the standard part of the approximating solution U is a continuous function on $\mathbb{R} \times [0, \infty)$. The substitution $f(U(x, t))$ is then well-defined and we can verify that the continuous standard function corresponding to U is a distributional solution of the reaction diffusion equation (1a).

4.1. *Weak solutions to the reaction diffusion equation.* Rather than writing the initial value problem in the distributional form $u_t - u_{xx} - f(u) = u_0(x)\, \delta(t)$, we use the integral version (12) of the definition of weak solution. Thus we define a *weak solution* to (1a), (1b) to be a continuous function $u : \mathbb{R} \times [0, \infty) \to \mathbb{R}$ that satisfies

$$\iint_{\mathbb{R} \times [0,\infty)} \{u(x, t)(-\varphi_{xx} - \varphi_t) - f(u(x, t))\, \varphi\}\, dx\, dt = \int_{\mathbb{R}} u_0(x)\, \varphi(x, 0)\, dx \tag{21}$$

for all test functions $\varphi \in C_c^\infty(\mathbb{R}^2)$.

---

[3]The assumption that f be globally Lipschitz continuous rules out the Fisher–Kolmogorov nonlinearity $f(u) = u - u^2$. However, for that particular nonlinearity the only solutions that are relevant to the interpretation of u as an allele ratio are those with $0 \leq u \leq 1$. If $f(u) = u - u^2$, then a quick look at our finite difference scheme (22) shows that for initial data that satisfy $0 \leq u(x, 0) \leq 1$ the approximate solution to the difference equation also satisfies $0 \leq U(x, t) \leq 1$, provided $\alpha < \frac{1}{2}$, so that the subsequent arguments still apply.

Theorem 4.1.1. *If f is Lipschitz continuous as in* (18), *and if the initial function u_0 is bounded by*

$$\text{for all } x \in \mathbb{R}, \quad |u_0(x)| \le M$$

for some positive real M, and if u_0 also is Lipschitz continuous, as in (20), *then the reaction diffusion equation* (1a), (1b) *has a weak solution.*

4.2. *Definition of the approximate solution.* To construct the solution we consider the grid G_C as defined in Section 3.3 with infinitesimal mesh sizes $dx, dt > 0$, and consider the finite difference scheme

$$D_t^+ U(x, t) = D_x^2 U(x, t) + f(U(x, t)). \tag{22}$$

Solving for $U(x, t + dt)$ we get

$$U(x, t+dt) = \alpha U(x+dx, t) + (1-2\alpha)U(x, t) + \alpha U(x-dx, t) + dt\, f(U(x, t)), \tag{23}$$

where, as before, $\alpha = dt/(dx)^2$. We extend the continuous function u_0 to an internal function $u_0 : {}^*\mathbb{R} \to {}^*\mathbb{R}$, and specify the initial conditions $U(x, 0) = u_0(x)$ for $x = m\, dx$, $m = -N, \ldots, +N$. The finite difference equation (23) then determines $U(x, t)$ for all $(x, t) \in G_C$.

We now establish a number of a priori estimates for the approximate solution U that will let us verify that its standard part is well-defined and that it is a weak solution of the initial value problem.

4.3. *Boundedness of the approximate solution.* First we bound $|U(x, t)|$.

Lemma 4.3.1. *For all $(x, t) \in G_C$ we have*

$$|U(x, t)| \le e^{K_1 t} M + \frac{K_0}{K_1}(e^{K_1 t} - 1). \tag{24}$$

Proof. Using (19), i.e., $|f(u)| \le K_0 + K_1|u|$, we get

$$|U(x, t+dt)| \le \alpha|U(x+dx, t)| + (1-2\alpha)|U(x, t)|$$
$$+ \alpha|U(x-dx, t)| + dt(K_0 + K_1|U(x, t)|). \tag{25}$$

In terms of $M(t) = \max_x |U(x, t)|$ this implies

$$M(t+dt) \le M(t) + dt(K_0 + K_1 M(t)) = (1 + K_1\, dt)M(t) + K_0\, dt.$$

Setting $t = n\, dt$ we see that this is an inequality of the form $M_n \le a M_{n-1} + b$, with $M_n = M(n\, dt)$. By induction this implies

$$M(t) = M(n\, dt) \le (1 + K_1\, dt)^n M(0) + \frac{(1 + K_1\, dt)^n - 1}{1 + K_1\, dt - 1} K_0\, dt$$

$$\le e^{K_1 t} M(0) + \frac{K_0}{K_1}(e^{K_1 t} - 1).$$

Since $M(0) = M$, this proves (24). $\qquad \square$

4.4. *Lipschitz continuity in space of the approximate solution.* We now show that $U(x,t)$ is Lipschitz continuous in the space variable.

Lemma 4.4.1. *For any two points* (x,t), $(x',t) \in G_C$ *we have*

$$|U(x,t) - U(x',t)| \leq Le^{K_1 t}|x - x'|, \tag{26}$$

where L is the Lipschitz constant for the initial function u_0, as in (20).

Proof. Let

$$V(x,t) \overset{\text{def}}{=} \frac{U(x+dx,t) - U(x,t)}{dx} = D_x^+ U(x,t).$$

Applying D_x^+ to both sides of (4) for U, and using the definition of V and commutativity of the difference quotient operators, we find

$$D_t^+ V = D_x^2 V + D_x^+ f(U).$$

Solving for $V(x,t+dt)$ we find

$$V(x,t+dt) = \alpha V(x+dx,t) + (1-2\alpha)V(x,t) + \alpha V(x-dx,t) + D_x^+ f(U).$$

Examining $D_x^+ f(U)$, we have

$$\begin{aligned}|D_x^+ f(U)| &= \frac{|f(U(x+dx,t)) - f(U(x,t))|}{dx} \\ &\leq \frac{K_1\,|U(x+dx,t) - U(x,t)|}{dx} = K_1|V(x,t)|,\end{aligned}$$

so that

$$|V(x,t+dt)| \leq \alpha|V(x+dx,t)| + (1-2\alpha)|V(x,t)| + \alpha|V(x-dx,t)| + K_1\,dt\,|V(x,t)|.$$

Using Gronwall's inequality on $\max_x V$, we get the inequality

$$\max_x |V(x,t)| \leq e^{K_1 t} \max_x |V(x,0)|.$$

The initial condition u_0 satisfies $|u_0(x) - u_0(x')| \leq L|x-x'|$ for all $x, x' \in \mathbb{R}$, and therefore the extension of u_0 to the hyperreals satisfies this same inequality. Hence $|V(x,0)| \leq L$ for all grid points $(x,0)$, and thus we have $|V(x,t)| \leq Le^{K_1 t}$. This implies (26). $\qquad\square$

4.5. *Hölder continuity in time of the approximate solution.*

Lemma 4.5.1. *Given any real $\bar{t} > 0$, for any two grid points (x_0, t_0), $(x_0, t_1) \in G_C$ with $0 \leq t_0 \leq t_1 \leq \bar{t}$, we have*

$$|U(x_0, t_1) - U(x_0, t_0)| \leq C\sqrt{t_1 - t_0}, \tag{27}$$

where C is a constant that only depends on $\bar{t}$, K_0, K_1, L, and M.

Proof. We begin by observing that $f(U(x,t))$ is bounded on the time interval we are considering. Indeed, for $t \leq \bar{t}$ we have shown for all $(x,t) \in G_C$ that

$$|U(x,t)| \leq A_0 \overset{\text{def}}{=} e^{K_1 \bar{t}} M + \frac{K_1}{K_0}(e^{K_1 \bar{t}} - 1),$$

while the Lipschitz condition for f implies $|f(U(x,t))| \leq K_0 + K_1|U| \leq K_0 + K_1 A_0$. So if we set $A = K_0 + K_1 A_0$, then we have

$$\text{for all } (x,t) \in G_C \text{ with } t \leq \bar{t}, \quad |f(U(x,t))| \leq A. \tag{28}$$

Next, we construct a family of upper barriers for U using parabolas. In particular, for any real $a, b, c > 0$ we consider

$$\bar{U}(x,t) \overset{\text{def}}{=} U(x_0, t_0) + a(t - t_0) + \frac{b}{2}(x - x_0)^2 + c.$$

For any $b > 0$ we will find $a, c > 0$ so that $\bar{U}$ is an upper barrier, in the sense that

$$D_t^+ \bar{U} - D_x^2 \bar{U} \geq A + 1, \tag{29}$$

$$\bar{U}(x,0) > U(x,0) \quad \text{for all } (x,0) \in G_C. \tag{30}$$

A direct computation shows that $D_t^+ \bar{U} - D_x^2 \bar{U} = a - b$, so for a given b we choose $a = b + A + 1$ and (29) will hold.

To satisfy (30) we use (26), i.e., that $U(x,t)$ is Lipschitz continuous with Lipschitz constant $\bar{L} \overset{\text{def}}{=} e^{K_1 \bar{t}} L$:

$$U(x,t) \leq U(x_0, t_0) + \bar{L}|x - x_0| \leq U(x_0, t_0) + \frac{2}{b}(x - x_0)^2 + \frac{\bar{L}^2}{2b}.$$

If we choose $c > \bar{L}^2/(2b)$, e.g., $c = \bar{L}^2/b$, then our upper barrier $\bar{U}$ also satisfies (30).

Next, we apply a maximum principle argument to compare U and $\bar{U}$. Consider $W(x,t) = U(x,t) - \bar{U}(x,t)$. Then we have shown that $W(x,0) < 0$ for all x, and $D_t^+ W - D_x^2 W < 0$, which implies

$$W(x, t + dt) < \alpha W(x - dx, t) + (1 - 2\alpha)W(x,t) + \alpha W(x + dx, t)$$

for all (x,t) for which $(x \pm dx, t) \in G_C$. By induction we get $W(x,t) < 0$ for all $(x,t) \in G_C$. In particular $U(x_0, t) < \bar{U}(x_0, t)$ for all $t > t_0$; i.e., we have shown

$$U(x_0, t_1) < U(x_0, t_0) + (b + A + 1)(t_1 - t_0) + \frac{\bar{L}^2}{b}.$$

This upper bound holds for any choice of $b > 0$. To get the best upper bound we minimize the right-hand side over all $b > 0$. The best bound appears when $b = \bar{L}/\sqrt{t_1 - t_0}$. After some algebra one then finds

$$U(x_0, t_1) - U(x_0, t_0) < (A + 1)(t_1 - t_0) + 2\bar{L}\sqrt{t_1 - t_0}.$$

Finally, using $t_1 - t_0 = \sqrt{t_1 - t_0}\sqrt{t_1 - t_0} \le \sqrt{\bar{t}}\sqrt{t_1 - t_0}$ we get

$$U(x_0, t_1) - U(x_0, t_0) < ((A+1)\sqrt{\bar{t}} + 2\bar{L})\sqrt{t_1 - t_0}.$$

This proves the upper bound in (27). To get the analogous lower bound one changes the signs of the coefficients a, b, c, which will turn $\bar{U}$ into a lower barrier. After working through the details, one finds the appropriate lower bound. ☐

Now that we have Lipschitz in space and Hölder in time, we also have for any pair of points (x, t), $(y, s) \in G_C$ that

$$|U(x, t) - U(y, s)| \le |U(x, t) - U(y, t)| + |U(y, t) - U(y, s)|$$

$$\le L|x - y| + C\sqrt{|t - s|}. \tag{31}$$

4.6. *Definition of the weak solution.*

So far we have been establishing estimates for the solution U to the finite difference scheme. It is worth pointing out that a standard existence proof would have required exactly the same estimates. At this point, however, the standard and nonstandard proofs diverge.

For $(x, t) \in \mathbb{R} \times [0, \infty)$ we choose $(\tilde{x}, \tilde{t}) \in G_C$ with $x \approx \tilde{x}$ and $t \approx \tilde{t}$, and then define $u(x, t) = \mathrm{St}(U(\tilde{x}, \tilde{t}))$. The continuity property (31) of the approximate solution U implies that the value of $\mathrm{St}(U(\tilde{x}, \tilde{t}))$ does not depend on how we chose the grid point $(\tilde{x}, \tilde{t})$, for if $(\hat{x}, \hat{t}) \in G_C$ also satisfied $\hat{x} \approx x$, $\hat{t} \approx t$, then $\tilde{x} \approx \hat{x}$ and $\tilde{t} \approx \hat{t}$, so that $U(\tilde{x}, \tilde{t}) \approx U(\hat{x}, \hat{t})$. It follows directly that the function $u : \mathbb{R} \times [0, \infty) \to \mathbb{R}$ is well-defined, that it satisfies the continuity condition (31), and that it satisfies the same bounds as in (24).

By the transfer principle, the standard function u extends in a unique way to an internal function $^*\mathbb{R} \times {}^*[0, \infty) \to {}^*\mathbb{R}$. It is common practice to abuse notation and use the same symbol u for the extension. The extended function satisfies the same continuity condition (31).

Lemma 4.6.1. *If $(x, t) \in G_C$ is limited, then $u(x, t) \approx U(x, t)$.*

Proof. If (x, t) is limited, then $x' = \mathrm{St}(x)$ and $t' = \mathrm{St}(t)$ are well-defined real numbers. By continuity of both u we have $u(x, t) \approx u(x', t')$. By the definition of u it follows from $x' \approx x$ and $t' \approx t$ that $u(x', t') \approx U(x, t)$. Combined we get $u(x, t) \approx U(x, t)$. ☐

4.7. *Proof that u is a weak solution.*

We will now show that u is a weak solution whose existence is claimed in Theorem 4.1.1; i.e., we verify that u satisfies (21) for any test function $\varphi \in C_c^\infty(\mathbb{R}^2)$.

Since φ has compact support, there is a positive real ℓ such that $\varphi(x, t) = 0$ outside the square $[-\ell, \ell] \times [-\ell, \ell]$. We therefore have to verify

$$\int_0^\ell \int_{-\ell}^\ell \{u(x, t)(-\varphi_{xx} - \varphi_t) - f(u(x, t))\varphi\} \, dx \, dt = \int_{-\ell}^\ell u_0(x)\varphi(x, 0) \, dx.$$

Since the integrands are continuous functions we only make an infinitesimal error when we replace the two Riemann integrals by Riemann sums over the part of the grid G_C that lies within the square $[-\ell, \ell] \times [-\ell, \ell]$. Thus we must prove

$$\sum_{(x,t) \in G_C} \{u(x,t)(-\varphi_{xx} - \varphi_t) - f(u(x,t))\varphi\} \, dx \, dt \approx \sum_{(x,0) \in G_C} u_0(x)\varphi(x,0) \, dx. \quad (32)$$

We now intend to replace u by U, and the derivatives of φ by the corresponding finite differences. In doing so we make errors that we must estimate. Let $G_{C\ell} = G_C \cap [-\ell, \ell]^2$, so that the only nonzero terms in the two sums come from terms evaluated at points in $G_{C\ell}$. The intersection of internal sets is again internal, so the set $G_{C\ell}$ is internal and hyperfinite.

For each $(x,t) \in G_{C\ell}$ the quantities

$$|u(x,t) - U(x,t)|, \quad |\varphi_t(x,t) - D_t^+ \varphi(x,t)|, \quad \text{and} \quad |\varphi_{xx}(x,t) - D_x^2 \varphi(x,t)|$$

are infinitesimal. Since they are defined by internal functions, one of the numbers in the hyperfinite set

$$\{|u(x,t) - U(x,t)|, |\varphi_t(x,t) - D_t^+ \varphi(x,t)|, |\varphi_{xx}(x,t) - D_x^2 \varphi(x,t)| : (x,t) \in G_{C\ell}\}$$

is the largest. This number, which we call ε, is again infinitesimal. Therefore we have

$$\max_{G_{C\ell}}\{|u(x,t) - U(x,t)|, |\varphi_t(x,t) - D_t^+ \varphi(x,t)|, |\varphi_{xx}(x,t) - D_x^2 \varphi(x,t)|\} \le \varepsilon \quad (33)$$

for some infinitesimal $\varepsilon > 0$.

The remainder of the argument is very similar to our proof in Section 3.4 that the distribution u defined was a distribution solution to the linear heat equation. Namely, if we replace u by U and derivatives of φ by finite differences of φ in (32), then (33) implies that we only make an infinitesimal error on both sides. We therefore only have to prove

$$\sum_{(x,t) \in G_C} \{U(x,t)(-D_x^2 \varphi - D_t^+ \varphi_t) - f(u(x,t))\varphi\} \, dx \, dt \approx \sum_{(x,0) \in G_C} u_0(x)\varphi(x,0) \, dx.$$

This follows after applying summation by parts, and using the finite difference equation (22) satisfied by U. This completes the existence proof.

Acknowledgement

References

[Albeverio et al. 1986] S. Albeverio, R. Høegh-Krohn, J. E. Fenstad, and T. Lindstrøm, *Nonstandard methods in stochastic analysis and mathematical physics*, Pure and Appl. Math. **122**, Academic Press, Orlando, FL, 1986. MR Zbl

[Fisher 1937] R. A. Fisher, "The wave of advance of advantageous genes", *Ann. Eugenics* **7**:4 (1937), 355–369. Zbl

[Folland 1999] G. B. Folland, *Real analysis: modern techniques and their applications*, 2nd ed., Wiley, New York, 1999. MR Zbl

[Goldblatt 1998] R. Goldblatt, *Lectures on the hyperreals*, Graduate Texts in Math. **188**, Springer, 1998. MR Zbl

[Henry 1981] D. Henry, *Geometric theory of semilinear parabolic equations*, Lecture Notes in Math. **840**, Springer, 1981. MR Zbl

[John 1982] F. John, *Partial differential equations*, 4th ed., Appl. Math. Sciences **1**, Springer, 1982. MR Zbl

[Keisler 1976a] H. J. Keisler, *Elementary calculus: an infinitesimal approach*, Prindle, Weber & Schmidt, Boston, 1976. Zbl

[Keisler 1976b] H. J. Keisler, *Foundations of infinitesimal calculus*, Prindle, Weber & Schmidt, Boston, 1976.

[Kolmogorov et al. 1937] A. N. Kolmogorov, I. G. Petrovsky, and N. S. Piskunov, "A study of the diffusion equation with increase in the amount of substance, and its application to a biological problem", *Vestnik Moskov. Univ. Mat. Meh.* **1**:6 (1937), 1–26. In Russian; translated in *Selected works of A. N. Kolmogorov, Volume I* (1991), Springer, 242–270.

[LeVeque 2007] R. J. LeVeque, *Finite difference methods for ordinary and partial differential equations*, Soc. Indust. Appl. Math., Philadelphia, 2007. MR Zbl

[Nelson 1987] E. Nelson, *Radically elementary probability theory*, Ann. Math. Studies **117**, Princeton Univ. Press, 1987. MR Zbl

[Pazy 1983] A. Pazy, *Semigroups of linear operators and applications to partial differential equations*, Appl. Math. Sciences **44**, Springer, 1983. MR Zbl

[Press et al. 2007] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes: the art of scientific computing*, 3rd ed., Cambridge Univ. Press, 2007. MR Zbl

[Tao 2007] T. Tao, "Ultrafilters, nonstandard analysis, and epsilon management", blog post, 2007, available at https://tinyurl.com/ultrafilt.

cjo5325@psu.edu   *Department of Mathematics, University of Wisconsin, Madison, WI, United States*

marshall.mueller@tufts.edu   *Department of Mathematics, Tufts University, Medford, MA, United States*

angenent@wisc.edu   *Department of Mathematics, University of Wisconsin, Madison, WI, United States*

Improving multilabel classification via heterogeneous ensemble methods

Yujue Wu and Qing Wang

(Communicated by Sat N. Gupta)

We consider the task of multilabel classification, where each instance may belong to multiple labels simultaneously. We propose a new method, called multilabel super learner (MLSL), that is built upon the problem transformation approach using the one-vs-all binary relevance method. MLSL is an ensemble model that predicts multilabel responses by integrating the strength of multiple base classifiers, and therefore it is likely to outperform each base learner. The weights in the ensemble classifier are determined by optimization of a loss function via V-fold cross-validation. Several loss functions are considered and evaluated numerically. The performance of various realizations of MLSL is compared to existing problem transformation algorithms using three real data sets, spanning applications in biology, music, and image labeling. The numerical results suggest that MLSL outperforms existing methods most of the time evaluated by the commonly used performance metrics in multilabel classification.

1. Introduction

Classification is a task of predicting labels of future instances by learning from the patterns of observed instances with known labels [Herrera et al. 2016]. The traditional classification problem, known as single-label classification, considers data sets with only one output attribute. When the single output attribute has two categories, it is referred to as binary classification; when the output attribute has more than two categories, it is called multiclass classification. In this paper we focus on the problem of multilabel classification, where each instance may be associated with more than one label.

The first literature on multilabel classification dates back to [McCallum 1999]; it focuses on the task of text categorization. In recent decades, multilabel classification has become an emerging research area and has been applied to many different disciplines, including image labeling [Duygulu et al. 2002; Boutell et al. 2004],

sentiment analysis [Turnbull et al. 2008; Sobol-Shikler and Robinson 2010] and bioinformatics [Elisseeff and Weston 2001; Diplaris et al. 2005]. More recent work on multilabel text categorization can also be found in [Klimt and Yang 2004; Lewis et al. 2004; Crammer et al. 2007; Katakis et al. 2008; Sriram et al. 2010; Charte et al. 2015]. A good overview of multilabel classification and its methods is provided in [Tsoumakas and Katakis 2007; Zhang and Zhou 2014; Gibaja and Ventura 2015; Herrera et al. 2016].

We can formally formulate the problem of multilabel classification as follows [Herrera et al. 2016]: Consider a dataset $\mathcal{D}$ with f input attributes $V_1, \ldots, V_f$. Let $\mathcal{V} = \{V_1, \ldots, V_f\}$ be the set of all input attributes in the dataset and $|\mathcal{V}| = f \geq 1$. Let $\mathcal{X} = V_1 \times V_2 \times V_3 \times \cdots \times V_f$. That is, $\mathcal{X}$ is the input space of the dataset, and $\mathcal{D} \subseteq \mathcal{X}$. Let $\mathcal{L} = \{y_1, \ldots, y_k\}$ be a set of distinct labels for $\mathcal{D}$, where each y_j represents a label. Here $|\mathcal{L}| = k \geq 2$. In single-label classification, including both binary and multiclass classification, each instance $\boldsymbol{x} \in \mathcal{X}$ is associated with one and only one label $y_j \in \mathcal{L}$. However, in multilabel classification, each instance $\boldsymbol{x} \in \mathcal{X}$ is associated with a subset of labels $L \subseteq \mathcal{L}$, where $1 \leq |L| \leq k$. The output space in multilabel classification, denoted by $\mathcal{Y}_{\text{multilabel}}$, is defined as the Cartesian product of k sets of binary values 0 and 1; i.e.,

$$\mathcal{Y}_{\text{multilabel}} = \{0, 1\}_1 \times \{0, 1\}_2 \times \cdots \times \{0, 1\}_k.$$

A multilabel classifier, denoted by $\mathcal{C} : \mathcal{X} \rightarrow \mathcal{Y}_{\text{multilabel}}$, learns from the input space $\mathcal{X}$ and predicts outcomes in the output space $\mathcal{Y}_{\text{multilabel}}$.

Generally speaking, there are two fundamental approaches to realize multilabel classification: *problem transformation* and *algorithm adaption* [Herrera et al. 2016]. The problem transformation methodology, at its core, converts a multilabel data set into several single-label data sets, thereby allowing the transformed data sets to be modeled using existing binary or multiclass classification methods. For example, one of the ways to realize problem transformation is through the *one-vs-all binary relevance* method, where a multilabel data set with k labels is converted into k binary data sets, one for each label. On the other hand, the algorithm adaption methodology transforms a single-label classification algorithm so that it can be applied to the original multilabel data set.

In this paper we propose a new method, called multilabel super learner (MLSL), which is an improved multilabel classification algorithm following the problem transformation approach, and is built upon the one-vs-all binary relevance method. MLSL is an ensemble model that makes predictions based on an integration of multiple base classifiers. The weights in the ensemble classifier are determined by optimizing a loss function. Several widely used loss functions are considered and evaluated numerically in this paper. The performance of the proposal is compared to existing problem transformation algorithms using real data sets in Section 4.

The numerical results suggest that MLSL outperforms existing binary relevance algorithms evaluated by almost all of the commonly used performance metrics in multilabel classification. To the best of our knowledge, none of the previous research considers implementing ensemble methods of this kind in multilabel classification.

The rest of the paper is structured as follows. In Section 2 we introduce the two general approaches to realize multilabel classification, and focus on the binary relevance method that the proposed MLSL is built upon. In Section 3 we detail the proposed MLSL algorithm, followed by numerical studies in Section 4. Finally, we conclude the paper with discussions of some future work in Section 5.

2. Existing methods for multilabel classification

We now introduce commonly used methods in multilabel classification. The two main approaches for multilabel classification are problem transformation and algorithm adaption. Problem transformation can be realized in two possible ways: (1) by converting the multilabel dataset into multiple binary data sets, (2) by converting the multilabel data set into one multiclass data set. These two approaches are often referred to as *binary relevance* and *label powerset* respectively. After the conversion, the altered data sets are suitable for single-label classification. Individually predicted labels are obtained from each of these single-label data sets, and then combined to produce the desired multilabeled outputs as the final predictions.

In algorithm adaption, existing single-label classification methods are altered so that they can be applied to multilabel data sets. Common methods under this framework include instance-based and logistic regression (IBLR-ML) [Cheng and Hüllermeier 2009], which is adapted from k-nearest neighbor (kNN) [Cover and Hart 1967] and logistic regression [Cox 1958], MODEL-x [Boutell et al. 2004], which is derived from support vector machines (SVM) [Cortes and Vapnik 1995], and the multilabel k-nearest neighbor lazy learning algorithm (ML-kNN) [Zhang and Zhou 2007]. Figure 1 displays an overview of the methods mentioned above. A detailed introduction of multilabel classification methods, including problem transformation and algorithm adaption, can be found in [Herrera et al. 2016; Tsoumakas and Katakis 2007].

Since our proposal is built upon the binary relevance method, we provide more details of this method in the following subsections. In binary relevance, a multilabel data set is converted into multiple single-label data sets. Such a data conversion process can be realized in two different ways: one-vs-all or one-vs-one.

2.1. *One-vs-all binary relevance.* The one-vs-all binary relevance approach [Herrera et al. 2016], showcased in Figure 2, transforms a multilabel data set $\mathcal{D}$, associated with k labels, into k unique binary-response data sets—one for each label. One then applies k single-label classifiers to the k binary data sets. The
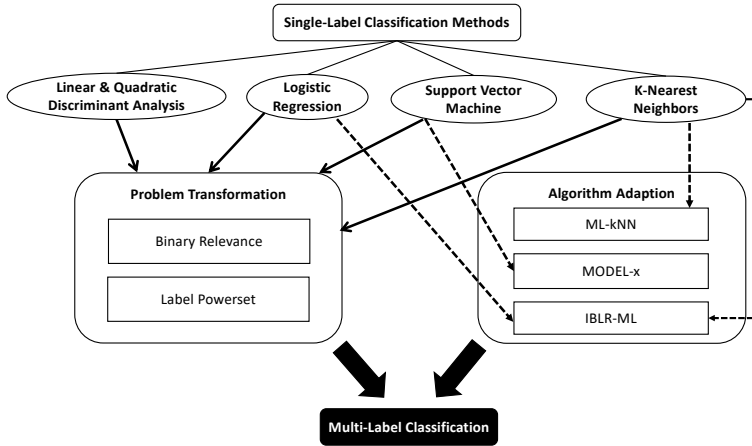
Figure 1. Illustration of some multilabel classification methods and their relationships with single-label classification methods.

k single-label classifiers are often set to be the same classification method, such as support vector machines, although this is not a strict requirement. In the prediction process, each test sample with unknown labels is considered as input for each of the binary classifiers, and based on the inputs, the i-th ($1 \leq i \leq k$) binary classifier produces a binary output, 0 or 1, indicating whether the test sample is associated with label $y_i \in \mathcal{L}$. All outputs generated by the trained binary classifiers will then be combined to form a final multilabel prediction.

The one-vs-all binary relevance approach is easy to implement. In addition, it offers a flexible family of methods in the sense that any binary classifier can be considered and used in the process. However, it suffers from two main disadvantages [Herrera et al. 2016]: First, since the single-label classifiers are independently trained, any potential correlations between labels are not taken into account in producing multilabel predictions. Intuitively, label correlations are valuable information that could help improve the accuracy of multilabel prediction. Second, it is possible that the transformed binary training data sets are more imbalanced than the original multilabel data set. As a result, some challenges may arise in the training stage due to the data conversion.

2.2. *One-vs-one binary relevance.* In the one-vs-one approach [Herrera et al. 2016], a multilabel data set is transformed to binary data sets, each of which is associated with a pair of labels in the label space $\mathcal{L}$. That is, given a data set with k unique labels, one considers $\binom{k}{2}$ binary data sets where each data set is associated with labels y_i and y_j ($y_i, y_j \in \mathcal{L}$ and $i \neq j$). Additionally, any instance that is not categorized by either of the two labels under consideration, or is categorized by both labels, is discarded from the corresponding binary data set.

Legend:
- ....▸ Step 1. Transforming Data
- ▭▸ Step 2. Training Binary Classifiers
- ——▸ Step 3. Predicting Binary Outputs
- ----▸ Step 4. Combining Binary Outputs

Training Sample

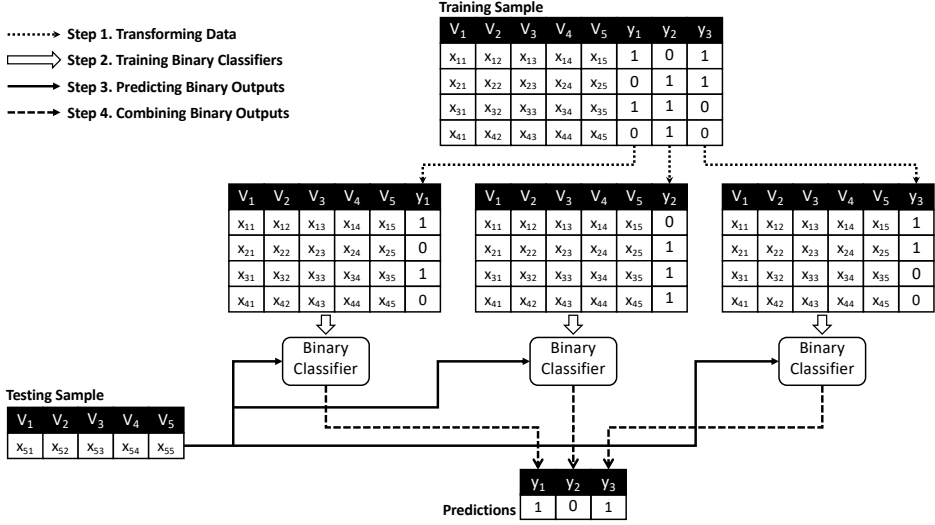| V_1	V_2	V_3	V_4	V_5	y_1	y_2	y_3
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	1	0	1
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	0	1	1
x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	1	1	0
x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	0	1	0

V_1	V_2	V_3	V_4	V_5	y_1
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	1
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	0
x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	1
x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	0

V_1	V_2	V_3	V_4	V_5	y_2
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	0
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	1
x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	1
x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	1

V_1	V_2	V_3	V_4	V_5	y_3
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	1
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	1
x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	0
x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	0

Binary Classifier Binary Classifier Binary Classifier

Testing Sample

V_1	V_2	V_3	V_4	V_5
x_{51}	x_{52}	x_{53}	x_{54}	x_{55}

Predictions

y_1	y_2	y_3
1	0	1

Figure 2. Illustration of one-vs-all binary relevance transformation, assuming five input variables and three possible labels.

In the prediction process, as illustrated in Figure 3, the test sample is considered as input. The output of each binary classifier is then used as "votes", and subsequently a ranking of labels produced by the votes will be generated to decide which labels are to be included in the final multilabel prediction. Examples of ranking algorithms include ranking by pairwise comparison [Hüllermeier et al. 2008] and calibrated label ranking [Fürnkranz et al. 2008].

The one-vs-one binary relevance has the same drawbacks as the one-vs-all binary relevance approach: lack of considerations of label correlations and imbalance in training datasets. In addition, the one-vs-one binary relevance method is likely to be less efficient than the one-vs-all binary relevance method due to the following two reasons: First, any given multilabel dataset with k labels, $k > 2$ and $\binom{k}{2} \geq k$. Thus, the one-vs-one method fits a larger number of binary classifiers than the one-vs-all method. Second, in the prediction process, since the one-vs-one approach incorporates ranking algorithms, it requires additional computation and is therefore likely to introduce errors to the final predictions. As a result, when considering the binary relevance approach, one often prefers the one-vs-all method.

3. Our proposal: multilabel super learner

We propose a stacking-based heterogeneous ensemble method, multilabel super learner (MLSL). MLSL is a multilabel classification algorithm that combines the prediction power of several one-vs-all binary relevance multilabel classification algorithms through an ensemble algorithm, *super learner* [van der Laan et al. 2007].
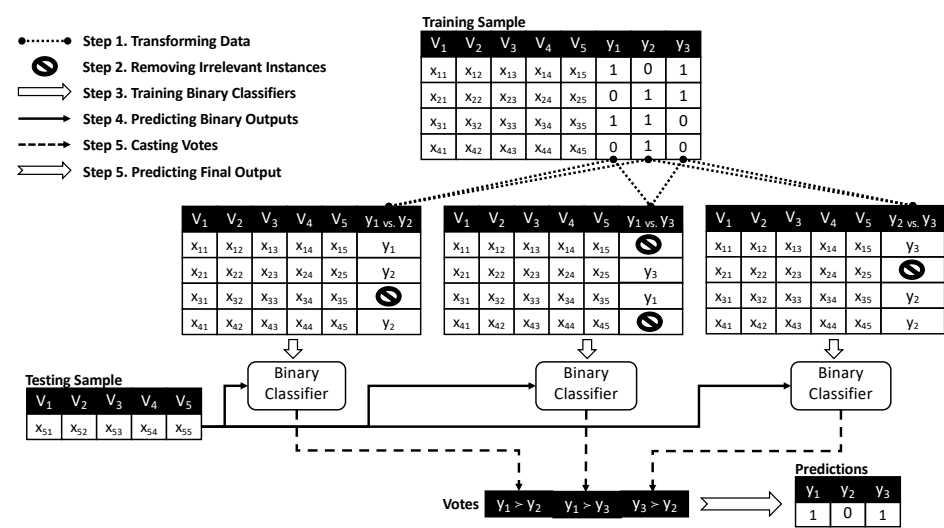
- • · · · · · • Step 1. Transforming Data
- ⊘ Step 2. Removing Irrelevant Instances
- ⇒ Step 3. Training Binary Classifiers
- → Step 4. Predicting Binary Outputs
- - - → Step 5. Casting Votes
- ⇒ Step 5. Predicting Final Output

Training Sample

V_1	V_2	V_3	V_4	V_5	y_1	y_2	y_3
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	1	0	1
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	0	1	1
x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	1	1	0
x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	0	1	0

V_1	V_2	V_3	V_4	V_5	y_1 vs. y_2
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	y_1
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	y_2
x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	⊘
x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	y_2

V_1	V_2	V_3	V_4	V_5	y_1 vs. y_3
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	⊘
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	y_3
x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	y_1
x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	⊘

V_1	V_2	V_3	V_4	V_5	y_2 vs. y_3
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	y_3
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	⊘
x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	y_2
x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	y_2

Testing Sample

V_1	V_2	V_3	V_4	V_5
x_{51}	x_{52}	x_{53}	x_{54}	x_{55}

Binary Classifier   →   Binary Classifier   →   Binary Classifier

Votes $y_1 > y_2$ $y_1 > y_3$ $y_3 > y_2$

Predictions

y_1	y_2	y_3
1	0	1

Figure 3. Illustration of one-vs-one binary relevance transformation, assuming five input variables and three possible labels.

In the following we first provide some background for the development of MLSL, followed by a step-by-step description of the MLSL algorithm. In the end we discuss its properties based on the theorems of super learner [van der Laan et al. 2007].

3.1. *Background.* MLSL is rooted in stacking, which dates back to the discussion of stacked generalization in [Wolpert 1992]. Stacked generalization combines information from multiple generalizers and minimizes the generalization error rate or biases of the generalizers. This model was later studied by Breiman [1996] in the context of regression and is referred to as stacked regression. Later, Freund et al. [1997] and Hansen [1998] adopted the same idea and proposed combining base learners from different methods to form a single learner. Following in the footsteps of previous work, van der Laan and Dudoit [2003] provided a unified framework to select the optimal combination of the set of base learners through cross-validation; they refer to the optimal solution as a "super learner". More recently, van der Laan et al. [2007] improved the previously proposed super learner by (1) extending it to include more flexible base learning algorithms, and (2) controlling over-fitting of the algorithm using cross-validation. Both [van der Laan and Dudoit 2003] and [van der Laan et al. 2006] show that under some regularity conditions the super learner in regression and single-label classification perform asymptotically as well as or even better than any of the base learning algorithms.

However, none of the previous literature considers applying ensemble methods of this kind to multilabel classification problems. Hence, the main contribution of our paper is to propose a multilabel super learner that integrates the strength and power

of several base multilabel classifiers through an optimal linear combination of them that minimizes some cross-validated risk. More specifically, we adapt the one-vs-all binary relevance method following the problem transformation approach, and implement super learner to realize binary classification based on each transformed binary data set. The weights in the linear combination of the binary base learners are optimized by cross-validated risk, which guards against over-fitting. In the end, predictions from binary super learners are combined to form multilabel output. The detailed algorithm of our proposed MLSL method is presented in the next subsection.

3.2. *Methodology.* Suppose we have an input space $\mathcal{X}$ and its associated label space $\mathcal{L}$. In multilabel prediction, one takes any instance $\boldsymbol{x} \in \mathcal{X}$ as an input and predicts an array of outputs

$$\boldsymbol{Y} = [Y_1 \ Y_2 \ \cdots \ Y_k],$$

where

$$Y_j = \begin{cases} 1 & \text{if } \boldsymbol{x} \text{ is labeled by } y_j \in \mathcal{L}, \\ 0 & \text{otherwise} \end{cases} \quad (1 \le j \le k).$$

Under this setting, the MLSL algorithm can be realized by the following five steps:

<u>Step 1</u>: selecting base learners. Define a library of m ($m \ge 2$) *base learners* $\{\phi_1, \dots, \phi_m\}$. Candidates for the base learners include any binary classifier, ranging from simple models, such as support vector machine (SVM) and k-nearest neighbors (kNN), to multistep algorithms that may involve covariate screening, parameter optimization, or model selection.

<u>Step 2</u>: transforming multilabel dataset. Given a training dataset $\mathcal{D} \subseteq \mathcal{X}$, we transform the multilabel datasets into $|\mathcal{L}| = k$ binary datasets, following the one-vs-all binary relevance method. Denote these k transformed binary datasets by $\mathcal{D}_1, \dots, \mathcal{D}_k$.

<u>Step 3</u>: training single-label super learners. For each binary data set $\mathcal{D}_j$ ($1 \le j \le k$), we realize the single-label super learner as follows:

(1) We first randomly split the j-th binary dataset $\mathcal{D}_j$ into V equally sized subsets, denoted by $\mathcal{D}_j^1, \mathcal{D}_j^2, \dots, \mathcal{D}_j^V$. Without loss of generality, assume $|\mathcal{D}_j|$ is divisible by V. Denote the number of observations in each data subset D_j^v by $\tilde{n} = |\mathcal{D}_j| / V$. For $v \in \{1, \dots, V\}$, let $\mathcal{D}_j^v$ be the validation sample and the remaining data be the training sample. Denote the v-th training set by $\mathcal{D}_j^{-v}$ so that $\mathcal{D}_j^{-v} = \mathcal{D}_j \setminus \mathcal{D}_j^v$.

(2) For each v ($1 \le v \le V$), we fit base learners $\phi_h \in \{\phi_1, \dots, \phi_m\}$ on $\mathcal{D}_j^{-v}$. Denote the fitted classifiers, trained on $\mathcal{D}_j^{-v}$, as $\hat{\phi}_{h, \mathcal{D}_j^{-v}}$ for $1 \le h \le m$. Write the prediction for label j based on the s-th instance $\boldsymbol{x}_s \in \mathcal{D}_j^v$ as $\hat{\phi}_h^j(\boldsymbol{x}_s) := \hat{\phi}_{h, \mathcal{D}_j^{-v}}(\boldsymbol{x}_s)$ ($1 \le h \le m$).

(3) Create a $|\mathcal{D}_j| \times m$ prediction matrix by combining the predictions from the m base learners $\phi_1, \ldots, \phi_m$ over the V validation sets. Denote the prediction matrix for label j by Z_j $(1 \leq j \leq k)$:

$$Z_j = \begin{bmatrix} \hat{\phi}_1^j(x_1) & \hat{\phi}_2^j(x_1) & \cdots & \hat{\phi}_m^j(x_1) \\ \hat{\phi}_1^j(x_2) & \hat{\phi}_2^j(x_2) & \cdots & \hat{\phi}_m^j(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\phi}_1^j(x_{|\mathcal{D}_j|}) & \hat{\phi}_2^j(x_{|\mathcal{D}_j|}) & \cdots & \hat{\phi}_m^j(x_{|\mathcal{D}_j|}) \end{bmatrix}.$$

Note that each prediction $\hat{\phi}_k^j(x_s)$ in Z_j is obtained by training on a data subset $\mathcal{D}_j^{-v}$ for $v \in \{1, \ldots, V\}$, where $x_s \in \mathcal{D}_j^v$.

(4) For each label j, let $\{\alpha_{j1}, \ldots, \alpha_{jm}\}$ be a set of weights $(\alpha_{jh} \in \mathbb{R}, \ 1 \leq h \leq m)$. Additional constraints on the weights, such as nonnegativity, may be applied but are not required. For any instance $x_s \in \mathcal{D}_j$, define the predicted single-label output as

$$\widehat{Y}_{js}(x_s) := \sum_{h=1}^m \alpha_{jh} \, \hat{\phi}_h^j(x_s), \tag{3-1}$$

or $\widehat{Y}_{js}$ for short. The coefficients, $\alpha_{j1}, \ldots, \alpha_{jm}$, are obtained under some optimization criterion via cross-validation, such as V-fold cross-validation. For instance, if we denote by $L(Y_{js}, \widehat{Y}_{js})$ a loss function that evaluates the closeness between Y_{js} and $\widehat{Y}_{js}$, then

$$(\hat{\alpha}_{j1}, \ldots, \hat{\alpha}_{jm}) = \arg \min_{\alpha_1, \ldots, \alpha_m} \sum_{v=1}^V \sum_{x_s \in \mathcal{D}_j^v} L(Y_{js}, \widehat{Y}_{js}(x_s)). \tag{3-2}$$

(5) The predicted probability of label j for instance x_s is thus given by

$$\widehat{Y}_j^{\text{sup}}(x_s) = \sum_{h=1}^m \hat{\alpha}_h \, \hat{\phi}_h^j(x_s).$$

Given a discriminating threshold c $(0 < c < 1)$, such as 0.5, one determines the classification output. Instances with predicted probabilities greater than the threshold would be classified as 1 (i.e., associated with label j) and as 0 (i.e., not associated with label j) otherwise. Denote the final predicted outcome of label j for instance x_s by $\mathcal{C}_j^{\text{sup}}$, given by

$$\mathcal{C}_j^{\text{sup}}(x_s) = \begin{cases} 1 & \text{if } \widehat{Y}_j^{\text{sup}}(x_s) \geq c, \\ 0 & \text{otherwise.} \end{cases}$$

<u>Step 4</u>: predicting future instances. In the prediction process, given an unknown instance x_t, the multilabel output is given by combining all k binary outputs

predicted by the k binary super learners:

$$[\mathcal{C}_1^{\text{sup}}(\boldsymbol{x}_t) \ \mathcal{C}_2^{\text{sup}}(\boldsymbol{x}_t) \ \cdots \ \mathcal{C}_k^{\text{sup}}(\boldsymbol{x}_t)].$$

3.3. *Properties.* As noted in [van der Laan and Dudoit 2003; van der Laan et al. 2006], the binary super learner is shown to perform asymptotically at least as well as any of the base binary classifiers. As a result, the binary super learner, i.e., $\mathcal{C}_j^{\text{sup}}$ ($1 \le j \le k$) in Step 4, is asymptotically at least as good as any of the base binary classifiers in $\{\phi_1, \ldots, \phi_m\}$. Therefore, the multilabel prediction, i.e., $[\mathcal{C}_1^{\text{sup}}(\boldsymbol{x}_t) \ \cdots \ \mathcal{C}_k^{\text{sup}}(\boldsymbol{x}_t)]$, should perform at least as well as, or even better than, the one-vs-all binary relevance multilabel classifier based on ϕ_h for all $h \in \{1, \ldots, m\}$.

4. Numerical comparison

We now empirically examine the performance of the MLSL algorithm introduced in Section 3. We consider four different criteria in determining the optimal weights in (3-2). In the context of V-fold cross-validation, the optimal weights under each criterion are selected as follows:

(1) Nonnegative least squares criterion (MLSL-NNLS): For $1 \le j \le k$,

$$(\alpha_{j1}, \ldots, \alpha_{jm}) = \arg\min \sum_{v=1}^{V} \sum_{\boldsymbol{x}_s \in D_j^v} (Y_{js} - \hat{Y}_{js}(\boldsymbol{x}_s))^2$$

subject to $\alpha_{j\ell} \ge 0$ ($1 \le \ell \le m$) and $\sum_{\ell=1}^{h} \alpha_{j\ell} = 1$. Here $\hat{Y}_{js}(\boldsymbol{x}_s)$ represents the predicted response for label j given an instance $\boldsymbol{x}_s \in \mathcal{D}_j^v$, where the base learners used to define $\hat{Y}_{js}$ (3-1) are trained on data D_j^{-v}.

(2) Nonnegative binomial likelihood maximization (MLSL-NNloglik): For $1 \le j \le k$,

$$(\alpha_{j1}, \ldots, \alpha_{jm}) = \arg\max \sum_{v=1}^{V} \sum_{\boldsymbol{x}_s \in \mathcal{D}_j^v} [Y_{js} \log \hat{Y}_{js}(\boldsymbol{x}_s) + (1 - Y_{js}) \log(1 - \hat{Y}_{js}(\boldsymbol{x}_s))],$$

subject to $\alpha_{j\ell} \ge 0$ ($1 \le \ell \le h$).

(3) Negative binomial log-likelihood minimization on the logistic scale using convex combination of weights (MLSL-CC_nloglik): For $1 \le j \le k$,

$$(\alpha_{j1}, \ldots, \alpha_{jm}) = \arg\min \sum_{v=1}^{V} \sum_{\boldsymbol{x}_s \in \mathcal{D}_j^v} [-Y_{js} \log \hat{Y}_{js}(\boldsymbol{x}_s) + (Y_{js} - 1) \log(1 - \hat{Y}_{js}(\boldsymbol{x}_s))],$$

subject to $\alpha_{j\ell} \ge 0$ ($1 \le \ell \le h$) and $\sum_{\ell=1}^{h} \alpha_{j\ell} = 1$.

(4) Area under the ROC (receiver operating characteristic) curve maximization (MLSL-AUC): For $1 \le j \le k$,

$$(\alpha_{j1}, \ldots, \alpha_{jm}) = \arg \max_{\alpha_{j1}, \ldots, \alpha_{jm}} \text{AUC}(\mathcal{C}_j^{\text{sup}}),$$

where AUC stands for the area under the ROC curve computed based on predictions $\mathcal{C}_j^{\text{sup}} = \{\mathcal{C}_j^{\text{sup}}(x_s) : x_s \in \mathcal{D}_j\}$. For more on ROC and AUC, see [Metz 1978; Swets 1973; Fawcett 2006].

In [van der Laan et al. 2007] it is shown, both theoretically and numerically, that the super learner [van der Laan et al. 2007] for single-label classification yields a result that is at least as good as that obtained from any of the base learners. As a result, the ensemble binary classifier $\mathcal{C}_j^{\text{sup}}$ ($1 \leq j \leq k$) should produce predictions that are at least as good as the binary predictions from the one-vs-all binary relevance method. Thus, following the proposed MLSL algorithm and combining the ensemble binary outputs to form multilabel predictions, we expect to see an improvement in the performance of the proposed MLSL method.

We assess the performances of the proposed MLSL method and the benchmarks based on the following commonly used multilabel performance metrics: Hamming loss, accuracy, precision, recall, F-measure, and subset accuracy. Definitions and more details of these performance measures can be found in [Herrera et al. 2016]. Ten 10-fold cross-validation was used when computing the performance metrics, in addition to the 10-fold cross-validation algorithm applied to choosing the optimal weights in (3-2).

4.1. *Data.* We selected three open-source data sets for our real data analysis, namely *emotions* [Trohidis et al. 2011], *birds* [Briggs et al. 2013], and *scene* [Boutell et al. 2004]. Our choice of these data sets is a result of three considerations. First, we focused our attention on data sets that are accessible online and well-known to the field of multilabel classification, so that researchers and practitioners in this area can easily reference our results in comparison to existing literature as well as future research. Second, we chose data sets from diverse real-world applications, with each data set initially collected to answer a different research question. Third, to the best of our efforts, we included data sets that have distinct multilabel characteristics.

Some details of the three data sets are as follows: the *emotions* data set models the relationship between 593 song clips and six kinds of emotions each song clip may evoke; the *birds* data set focuses on identifying which bird species (out of 19) are present in each of the 645 audio clips recorded in forests; the *scene* data set associates each of the 2407 photographs by one or more of the six scenery labels that the photo may capture.

We present some characteristic metrics of the three data sets in Table 1. Among these statistics, cardinality, density, and highest label frequency represent label distribution; diversity, maximum imbalance ratio (MaxIR), mean imbalance ratio (MeanIR), and score of concurrence among imbalanced labels (SCUMBLE)

	emotions	*birds*	*scene*
#instances	593	645	2407
#labels	6	19	6
#attributes	78	279	300
cardinality	1.87	1.01	1.07
density	0.31	0.05	0.18
highest label frequency	81	194	405
diversity	4	73	3
MaxIR	1.78	17.17	1.46
MeanIR	1.58	5.41	1.25
SCUMBLE	0.01	0.03	0.00

Table 1. Characteristic metrics for datasets *emotions*, *birds* and *scene*.

reveal the degree of imbalance in the data, and a larger value in these measures indicates a more imbalanced label structure and higher difficulty level for the task of classification.

From Table 1 we can see that *scene* is the largest data set with 2407 instances and 300 attributes, but it is the least imbalanced data set. In contrast, *birds*, a smaller data set than *scene*, is much more imbalanced than either *scene* or *emotions*. Compared to *scene* and *birds*, the *emotions* data set is the smallest data set and is partially balanced, with all of its imbalance measures, including MaxIR, MeanIR and SCUMBLE, falling between those of *scene* and *birds*.

4.2. Results. In multilabel classification, one often considers performance metrics such as Hamming loss, accuracy, precision, F-measure, recall, and subset accuracy [Herrera et al. 2016]. In particular, F-measure is a trade-off between precision and recall. In Tables 2–7 we summarize the results of these measures after fitting our proposed MLSL model and the benchmark models based on each of the three real data sets. We considered two ways of selecting the base learners in the proposed algorithm. In the first case, we only chose simple binary classifiers for the binary relevance (BR) method. There are four such benchmark models under consideration, i.e., logistic regression (BR-GLM), linear discriminant analysis (BR-LDA), k-nearest neighbor (BR-kNN), and support vector machines (BR-SVM). In the second scenario, in addition to the previously listed simple base learners we also included two machine learning binary classifiers when fitting the binary relevance model, i.e., random forest (BR-RF) and gradient decent (BR-GD). These more powerful benchmark methods are anticipated to yield more accurate multilabel classification results at the expense of higher computational cost. We are interested in investigating how our proposed MLSL method works with or without more complex base learners from different aspects. The R package "SuperLearner" [Polley et al.

	binary relevance (BR)				MLSL			
	GLM	LDA	*k*NN	SVM	NNLS	NN	CC	AUC
Hamming loss	0.215	0.208	0.273	0.181	0.180	0.179	**0.178**	0.181
accuracy	0.785	0.792	0.727	0.819	0.820	0.821	**0.822**	0.819
F-measure	0.674	0.686	0.546	0.725	0.725	0.726	**0.727**	0.722
precision	0.678	0.699	0.573	0.763	0.766	0.766	**0.767**	0.761
recall	0.673	0.675	0.523	0.691	0.690	0.691	**0.692**	0.689
subset accuracy	0.236	0.259	0.183	**0.323**	0.315	0.317	0.320	0.310
computation (min)	0.002	0.001	0.000	0.012	0.158	0.158	0.158	0.421

Table 2. Results for the *emotions* data set using four base learners. Here and in the following tables the column headers "NN" and "CC" stand for "NNlogik" and "CC_nloglik".

	binary relevance (BR)						MLSL			
	GLM	LDA	*k*NN	SVM	RF	GD	NNLS	NN	CC	AUC
Hamming loss	0.215	0.208	0.273	0.181	0.178	0.200	0.178	**0.176**	0.177	0.178
accuracy	0.785	0.792	0.727	0.819	0.822	0.800	0.822	**0.824**	0.823	0.822
F-measure	0.674	0.686	0.546	0.725	0.731	0.688	0.732	**0.738**	0.734	0.729
precision	0.678	0.699	0.573	0.763	0.761	0.704	0.769	**0.774**	0.772	0.769
recall	0.673	0.675	0.523	0.691	0.704	0.674	0.699	**0.706**	0.701	0.694
subset accuracy	0.236	0.259	0.183	0.323	**0.331**	0.280	0.317	0.321	0.317	0.315
computation (min)	0.002	0.001	0.001	0.013	0.165	0.191	3.666	3.667	3.667	4.145

Table 3. Results for the *emotions* data set using six base learners.

	binary relevance (BR)				MLSL			
	GLM	LDA	*k*NN	SVM	NNLS	NN	CC	AUC
Hamming loss	0.128	0.082	0.057	0.043	0.042	0.042	**0.041**	0.056
accuracy	0.872	0.918	0.943	0.957	0.958	0.958	**0.959**	0.944
F-measure	0.327	0.463	0.259	0.698	0.682	0.691	**0.701**	0.516
precision	0.234	0.386	0.354	0.747	0.736	0.752	**0.763**	0.505
recall	0.552	0.588	0.206	**0.659**	0.639	0.643	0.652	0.534
subset accuracy	0.319	0.427	0.468	0.523	0.532	0.536	**0.538**	0.485
computation (min)	0.127	0.030	0.006	0.107	2.765	2.768	2.767	3.139

Table 4. Results for the *birds* data set using four base learners.

2018] was used in the implementation process of MLSL to realize Step 4 of the proposed algorithm. In Tables 2–7 we highlighted the value of performance metric, up to three decimal places, corresponding to the best performance given each criterion. In addition, we also reported the computation time in minutes for fitting each model. Recall that 10-fold cross-validation (CV) was used when determining the coefficients of MLSL in (3-2). So, MLSL essentially fit each base learner ten times, which is reflected in the total running time of MLSL.

The numerical results suggest that the proposed MLSL algorithm is quite competitive compared to the benchmarks based on almost all performance metrics.

	binary relevance (BR)						MLSL			
	GLM	LDA	kNN	SVM	RF	GD	NNLS	NN	CC	AUC
Hamming loss	0.128	0.082	0.057	0.043	0.040	0.042	0.039	**0.038**	**0.038**	0.039
accuracy	0.872	0.918	0.943	0.957	0.960	0.958	0.961	**0.962**	**0.962**	0.961
F-measure	0.327	0.463	0.259	0.698	0.711	0.627	0.713	**0.718**	**0.718**	0.715
precision	0.234	0.386	0.354	0.747	**0.893**	0.763	0.864	0.854	0.853	0.851
recall	0.552	0.588	0.206	**0.659**	0.594	0.535	0.609	0.622	0.622	0.620
subset accuracy	0.319	0.427	0.468	0.523	0.530	0.525	0.547	0.554	**0.555**	0.548
computation (min)	0.118	0.027	0.005	0.097	1.138	1.256	26.03	26.04	26.03	26.94

Table 5. Results for the *birds* data set using six base learners.

	binary relevance (BR)				MLSL			
	GLM	LDA	kNN	SVM	NNLS	NN	CC	AUC
Hamming loss	0.135	0.112	0.093	0.074	0.073	0.073	0.073	**0.074**
accuracy	0.865	0.888	0.907	0.926	**0.927**	**0.927**	**0.927**	0.926
F-measure	0.720	0.780	0.773	0.862	**0.865**	0.862	0.862	0.861
precision	0.677	0.760	0.778	0.861	**0.868**	0.864	0.864	0.865
recall	0.769	0.802	0.769	**0.864**	0.862	0.860	0.860	0.857
subset accuracy	0.443	0.507	0.645	0.664	0.671	**0.672**	**0.672**	0.668
computation (min)	0.104	0.029	0.028	0.225	3.751	3.754	3.753	4.130

Table 6. Results for the *scene* data set using four base learners.

	binary relevance (BR)						MLSL			
	GLM	LDA	kNN	SVM	RF	GD	NNLS	NN	CC	AUC
Hamming loss	0.135	0.112	0.093	0.074	0.083	0.074	**0.069**	**0.069**	**0.069**	0.070
accuracy	0.865	0.888	0.907	0.926	0.917	0.926	**0.931**	**0.931**	**0.931**	0.930
F-measure	0.720	0.780	0.773	0.862	**0.903**	0.874	0.885	0.884	0.884	0.889
precision	0.677	0.760	0.778	0.861	**0.914**	0.871	0.890	0.890	0.890	0.897
recall	0.769	0.802	0.769	0.864	**0.892**	0.877	0.880	0.879	0.879	0.882
subset accuracy	0.443	0.507	0.645	0.664	0.578	0.656	0.676	**0.678**	0.677	0.670
computation (min)	0.098	0.027	0.025	0.198	2.843	2.490	55.35	55.35	55.35	55.95

Table 7. Results for the *scene* data set using six base learners.

Through the binary relevance approach, a given multilabel data set is converted to multiple binary data sets, revealing differences among labels. As a result, each individual classification method, regardless of its complexity, is unlikely to perform well on all of the transformed single-label data sets. By introducing an ensemble classifier that incorporates a diverse group of base learners, the MLSL method increases the chance for each instance to be predicted correctly via the ensemble of different base learners [van der Laan et al. 2007]. When including two more complex base learners, i.e., BR-RF and BR-GD, the performance of MLSL (see Tables 3, 5, and 7) showed further improvement compared to the results based on four simple learners. In theory, one would always expect better performance of MLSL if a larger number of base learners are considered.

Our numerical investigations also revealed that the computation time for finding the optimal weights in (3-2) is negligible compared to the cost of fitting each base learner. As a result, the running time of MLSL is driven by the complexity of each base learner, the total number of base learners, and the number of folds in the cross-validation algorithm. Since 10-fold CV was used in finding the optimal coefficients in (3-2), the computation time of MLSL is approximately equal to ten times the total time for fitting all base learners. (The running time of MLSL would be roughly cut by half, if one uses 5-fold CV, instead of 10-fold CV.) As shown in the tables, the total running time increased significantly when we introduced complex base learners (BR-RF and BR-GD) to the algorithm. In practice, there is a trade-off between performance and computational cost. It is common to consider a few to a dozen base learners for real data implementation.

5. Future work

As noted in Section 2, one of the drawbacks of the binary-relevance multilabel classification method is its lack of consideration of label correlations. Since our proposed multilabel super learner is built upon the one-vs-all binary relevance algorithm, it suffers from the same shortcoming. In practice, there are a few existing methods to account for label correlations in multilabel classification. For our future work, we are interested in exploring the possibility of implementing classifier chain [Read et al. 2011], an algorithm that takes into consideration label correlations, to MLSL. We anticipate that combining classifier chain with MLSL would further improve the performance of multilabel classification.

Acknowledgments

References

[Boutell et al. 2004] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification", *Pattern Recog.* **37**:9 (2004), 1757–1771.

[Breiman 1996] L. Breiman, "Stacked regressions", *Mach. Learn.* **24**:1 (1996), 49–64. Zbl

[Briggs et al. 2013] F. Briggs, Y. Huang, R. Raich, K. Eftaxias, Z. Lei, W. Cukierski, S. F. Hadley, A. Hadley, M. Betts, X. Z. Fern, J. Irvine, L. Neal, A. Thomas, G. Fodor, G. Tsoumakas, H. W. Ng, T. N. T. Nguyen, H. Huttunen, P. Ruusuvuori, T. Manninen, A. Diment, T. Virtanen, J. Marzat, J. Defretin, D. Callender, C. Hurlburt, K. Larrey, and M. Milakov, "The 9th annual MLSP competition: new methods for acoustic classification of multiple simultaneous bird species in a noisy environment", art. id. 34 in *IEEE International Workshop on Machine Learning for Signal Processing* (Southampton, UK, 2013), IEEE, Piscataway, NJ, 2013.

[Charte et al. 2015] F. Charte, A. J. Rivera, M. del Jesus, and F. Herrera, "QUINTA: a question tagging assistant to improve the answering ratio in electronic forums", pp. 255–260 in *IEEE International*

Conference on Computer as a Tool (*EUROCON*) (Salamanca, Spain, 2015), edited by J. Haase et al., IEEE, Piscataway, NJ, 2015.

[Cheng and Hüllermeier 2009] W. Cheng and E. Hüllermeier, "Combining instance-based learning and logistic regression for multilabel classification", *Mach. Learn.* **76**:2-3 (2009), 211–225.

[Cortes and Vapnik 1995] C. Cortes and V. Vapnik, "Support-vector networks", *Mach. Learn.* **20**:3 (1995), 273–297. Zbl

[Cover and Hart 1967] T. Cover and P. Hart, "Nearest neighbor pattern classification", *IEEE Trans. Info. Theory* **13**:1 (1967), 21–27. Zbl

[Cox 1958] D. R. Cox, "The regression analysis of binary sequences", *J. Roy. Stat. Soc. Ser. B* **20**:2 (1958), 215–242. Zbl

[Crammer et al. 2007] K. Crammer, M. Dredze, K. Ganchev, P. P. Talukdar, and S. Carroll, "Automatic code assignment to medical text", pp. 129–136 in *Proceedings of the workshop on BioNLP 2007: biological, translational, and clinical language processing* (Prague, 2007), Assoc. Comput. Linguistics, Stroudsburg, PA, 2007.

[Diplaris et al. 2005] S. Diplaris, G. Tsoumakas, P. A. Mitkas, and I. Vlahavas, "Protein classification with multiple algorithms", pp. 448–456 in *Advances in informatics* (Volas, Greece, 2005), edited by P. Bozanis and E. N. Houstis, Lecture Notes in Comput. Sci. **3746**, Springer, 2005.

[Duygulu et al. 2002] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth, "Object recognition as machine translation: learning a lexicon for a fixed image vocabulary", pp. 97–112 in *Computer vision: ECCV 2002* (Copenhagen, 2002), edited by A. Heyden et al., Lecture Notes in Comput. Sci. **2353**, Springer, 2002. Zbl

[Elisseeff and Weston 2001] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification", pp. 681–687 in *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic* (Vancouver, 2001), edited by T. G. Dietterich et al., MIT Press, Cambridge, MA, 2001.

[Fawcett 2006] T. Fawcett, "An introduction to ROC analysis", *Pattern Recog. Lett.* **27**:8 (2006), 861–874.

[Freund et al. 1997] Y. Freund, R. E. Schapire, Y. Singer, and M. K. Warmuth, "Using and combining predictors that specialize", pp. 334–343 in *Proceedings of the 29th annual ACM Symposium on Theory of Computing* (El Paso, TX, 1997), ACM, New York, 1997. Zbl

[Fürnkranz et al. 2008] J. Fürnkranz, E. Hüllermeier, E. Loza Mencía, and K. Brinker, "Multilabel classification via calibrated label ranking", *Mach. Learn.* **73**:2 (2008), 133–153.

[Gibaja and Ventura 2015] E. Gibaja and S. Ventura, "A tutorial on multilabel learning", *ACM Comput. Surv.* **47**:3 (2015), art. id. 52.

[Hansen 1998] J. V. Hansen, "Combining predictors: some old methods and a new method", pp. 14–16 in *Proceedings of the Joint Conference on Information Sciences (JCIS '98), II* (Research Triangle Park, NC, 1998), edited by P. P. Wang, Assoc. Intelligent Machinery, Durham, NC, 1998.

[Herrera et al. 2016] F. Herrera, F. Charte, A. J. Rivera, and M. J. del Jesus, *Multilabel classification: problem analysis, metrics and techniques*, Springer, 2016. MR

[Hüllermeier et al. 2008] E. Hüllermeier, J. Fürnkranz, W. Cheng, and K. Brinker, "Label ranking by learning pairwise preferences", *Artificial Intelligence* **172**:16-17 (2008), 1897–1916. MR Zbl

[Katakis et al. 2008] I. Katakis, G. Tsoumakas, and I. Vlahavas, "Multilabel text classification for automated tag suggestion", pp. 75–84 in *Proceedings of the ECML/PKDD Discovery Challenge* (Antwerp, 2008), ECML/PKDD, 2008.

[Klimt and Yang 2004] B. Klimt and Y. Yang, "The Enron corpus: a new dataset for email classification research", pp. 217–226 in *Machine learning: ECML 2004* (Pisa, 2004), edited by J.-F. Boulicaut et al., Lecture Notes in Comput. Sci. **3201**, Springer, 2004. Zbl

[van der Laan and Dudoit 2003] M. J. van der Laan and S. Dudoit, "Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and examples", working paper 130, Univ. California, Berkeley, Division of Biostatistics, 2003, available at https://biostats.bepress.com/ucbbiostat/paper130/.

[van der Laan et al. 2006] M. J. van der Laan, S. Dudoit, and A. W. van der Vaart, "The cross-validated adaptive epsilon-net estimator", *Statist. Decisions* **24**:3 (2006), 373–395. MR Zbl

[van der Laan et al. 2007] M. J. van der Laan, E. C. Polley, and A. E. Hubbard, "Super Learner", *Stat. Appl. Genet. Mol. Biol.* **6**:1 (2007), art. id. 25. MR Zbl

[Lewis et al. 2004] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, "RCV1: a new benchmark collection for text categorization research", *J. Mach. Learn. Res.* **5** (2004), 361–397.

[McCallum 1999] A. K. McCallum, "Multi-label text classification with a mixture model trained by EM", preprint, 1999, available at https://tinyurl.com/akmccall.

[Metz 1978] C. E. Metz, "Basic principles of ROC analysis", *Semin. Nuclear Medicine* **8**:4 (1978), 283–298.

[Polley et al. 2018] E. Polley, E. LeDell, C. Kennedy, S. Lendle, and M. van der Laan, "SuperLearner", 2018, available at https://github.com/ecpolley/SuperLearner. R package.

[Read et al. 2011] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification", *Mach. Learn.* **85**:3 (2011), 333–359. MR

[Sobol-Shikler and Robinson 2010] T. Sobol-Shikler and P. Robinson, "Classification of complex information: inference of co-occurring affective states from their expressions in speech", *IEEE Trans. Pattern Anal. Mach. Intell.* **32**:7 (2010), 1284–1297.

[Sriram et al. 2010] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in Twitter to improve information filtering", pp. 841–842 in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Geneva, 2010), edited by H.-H. Chen et al., ACM, New York, 2010.

[Swets 1973] J. A. Swets, "The relative operating characteristic in psychology", *Science* **182**:4116 (1973), 990–1000.

[Trohidis et al. 2011] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas, "Multi-label classification of music by emotion", *EURASIP J. Audio Speech Music Proc.* **2011** (2011), art. id. 4.

[Tsoumakas and Katakis 2007] G. Tsoumakas and I. Katakis, "Multi-label classification: an overview", *Int. J. Data Warehousing Mining* **3**:3 (2007), 1–13.

[Turnbull et al. 2008] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects", *IEEE Trans. Audio Speech Lang. Process.* **16**:2 (2008), 467–476.

[Wolpert 1992] D. H. Wolpert, "Stacked generalization", *Neural Netw.* **5**:2 (1992), 241–259.

[Zhang and Zhou 2007] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: a lazy learning approach to multi-label learning", *Pattern Recog.* **40**:7 (2007), 2038–2048. Zbl

[Zhang and Zhou 2014] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms", *IEEE Trans. Knowledge Data Eng.* **26**:8 (2014), 1819–1837.

ywu3@wellesley.edu   *Department of Mathematics, Wellesley College, Wellesley, MA, United States*

qwang@wellesley.edu   *Department of Mathematics, Wellesley College, Wellesley, MA, United States*

■msp

The number of fixed points
of AND-OR networks
with chain topology

Alan Veliz-Cuba and Lauren Geiser

(Communicated by Kenneth S. Berenhaut)

AND-OR networks are Boolean networks where each coordinate function is either the AND or OR logical operator. We study the number of fixed points of these Boolean networks in the case that they have a wiring diagram with chain topology. We find closed formulas for subclasses of these networks and recursive formulas in the general case. Our results allow for an effective computation of the number of fixed points in the case that the topology of the Boolean network is an open chain (finite or infinite) or a closed chain. We further explore how our approach could be used in "fractal" chains.

1. Introduction

Boolean networks, $f : \{0, 1\}^n \to \{0, 1\}^n$, have been used to study problems arising from areas such as mathematics, computer science, and biology [Akutsu et al. 1998; Albert and Othmer 2003; Mendoza and Xenarios 2006; Jarrah et al. 2010; Wang et al. 2017]. A particular problem of interest is counting the number of fixed points (x such that $f(x) = x$). To simplify this problem one can restrict the class of Boolean functions or the topology of the network [Agur et al. 1988; Aracena et al. 2004; Jarrah et al. 2007; 2010; Aracena 2008; Murrugarra and Laubenbacher 2011; Bollman et al. 2010; Veliz-Cuba et al. 2014a; 2014b; Dimitrova et al. 2015; Weiss and Margaliot 2017], which in some cases allows one to find effective algorithms or formulas in closed form.

In this manuscript we focus on the number of fixed points of AND-OR networks (each Boolean function is either the AND or the OR operator) that have open or closed chain topology. The networks we study in this manuscript also arise by restricting min-max networks to a Boolean set of values $\{0, 1\}$ [Goles et al. 2000].

Figure 1. Wiring diagram with open chain topology.

Although one typically specifies the update order to analyze the dynamics, this is not necessary here as the fixed points would not change [Hansson et al. 2005]. We first consider the case of finite open chain topology and find a recursive formula (Theorem 2.4) and sharp lower and upper bounds. We then consider the case of infinite and closed chain topology and show how they can be reduced to the case of finite open chain topology (Theorems 3.1 and 3.2).

2. Open chain

Let $f = (f_1, \ldots, f_n) : \{0, 1\}^n \to \{0, 1\}^n$ with $n \geq 2$ be an AND-OR network such that its wiring diagram is a chain; see Figure 1. That is, we consider Boolean networks of the form

$$f_1 = x_2, \quad f_2 = x_1 \Diamond_2 x_3, \quad f_3 = x_2 \Diamond_3 x_4, \quad \ldots, \quad f_{n-1} = x_{n-2} \Diamond_{n-1} x_n, \quad f_n = x_{n-1},$$

where $\Diamond_i$ is the AND ($\wedge$) or the OR ($\vee$) operator.

Because this family of Boolean networks is completely determined by the sequence of logical operators $\Diamond_2, \Diamond_3, \ldots, \Diamond_{n-1}$, we can use this sequence to represent the network. Furthermore, consecutive occurrences of the same logical operator can be denoted as $\wedge^k$ or $\vee^k$.

We are interested in the number of fixed points of such Boolean networks. For simplicity we denote the elements of $\{0, 1\}^n$ as binary strings (omitting parentheses). Also, we will use the notation $\mathbf{0} = 00 \cdots 0$ and $\mathbf{1} = 11 \cdots 1$, where the length of the strings will be clear from the context. Note that $\mathbf{0}$ and $\mathbf{1}$ are fixed points of all AND-OR networks with chain topology.

Example 2.1. Our running example will be the AND-OR network

$$\begin{aligned}
&f_1 = x_2, &&f_4 = x_3 \vee x_5, &&f_7 = x_6 \vee x_8, &&f_{10} = x_9 \wedge x_{11}, \\
&f_2 = x_1 \wedge x_3, &&f_5 = x_4 \wedge x_6, &&f_8 = x_7 \vee x_9, &&f_{11} = x_{10} \vee x_{12}, \\
&f_3 = x_2 \wedge x_4, &&f_6 = x_5 \vee x_7, &&f_9 = x_8 \wedge x_{10}, &&f_{12} = x_{11}.
\end{aligned}$$

This network can be represented by the sequence of operators $\wedge\wedge\vee\wedge\vee\vee\vee\wedge\wedge\vee$. We can further simplify this representation to $\wedge^2\vee\wedge\vee^3\wedge^2\vee$. This AND-OR network has 13 fixed points listed in Table 1 (first column).

The next lemma states that the number of fixed points depends only on the powers of the operators. Since we do not know which operator is last ($\wedge$ or $\vee$), we will simply use ellipses without explicitly writing the last operator.

Lemma 2.2. *The AND-OR networks $f = \wedge^{k_1} \vee^{k_2} \wedge^{k_3} \cdots$ and $g = \vee^{k_1} \wedge^{k_2} \vee^{k_3} \cdots$ have the same number of fixed points.*

Proof. Consider $\phi : \{0, 1\}^n \to \{0, 1\}^n$ given by $\phi(x_1, \ldots, x_n) = (\neg x_1, \ldots, \neg x_n)$, where $\neg$ is the logical operator NOT. Using the fact that $\neg(p \wedge q) = \neg p \vee \neg q$ and $\neg(p \vee q) = \neg p \wedge \neg q$, it follows that $f(\phi(x)) = \phi(g(x))$. Then, x will be a fixed point of g if and only if $\phi(x)$ is a fixed point of f. So, ϕ is a bijection between the fixed points of g and f. $\square$

Because we are interested in the number of fixed points, we will simply use $(k_1, k_2, \ldots, k_m)$ to refer to a network. For instance, the AND-OR network seen in Example 2.1 can be represented simply by $(2, 1, 1, 3, 2, 1)$. We denote the number of fixed points by $\mathcal{F}(k_1, k_2, \ldots, k_m)$. A similar approach was used by [Alcolei et al. 2016] to study nonmonotonic Boolean networks.

The following lemma states that consecutive variables that have the same logical operator must be equal.

Lemma 2.3. *Consider an AND-OR network f represented by $(k_1, k_2, \ldots, k_m)$. Denote an element of the domain of f by $x = (x^1, x^2, \ldots, x^m)$, where $x^1 \in \{0, 1\}^{k_1+1}$, $x^m \in \{0, 1\}^{k_m+1}$, and $x^i \in \{0, 1\}^{k_i}$ for $i = 2, \ldots, m - 1$. If x is a fixed point of f, then $x^i = \mathbf{0}$ or $x^i = \mathbf{1}$ for $i = 1, \ldots, m$.*

Proof. Let x be a fixed point of f. We use $(x^i)_j$ to denote the j-th coordinate of x^i. Note that $(x^1)_1 = (x^1)_2$ and $(x^m)_{k_m} = (x^m)_{k_m+1}$ by the definition of f (the first and last coordinate functions of f depend on single variables).

Now, the rest of the proof follows from the fact that if $q = p \wedge r$ and $r = q \wedge s$ or if $q = p \vee r$ and $r = q \vee s$, then $q = r$. This implies that consecutive variables, $(x^i)_j$ and $(x^i)_{j+1}$, that have the same logical operators must be the same. $\square$

The next proposition states that the numbers k_i in $\mathcal{F}(k_1, \ldots, k_m)$ can be assumed to be at most 2 for $2 \le i \le m-1$, and 1 for k_1 and k_m. For example, this will imply that $\mathcal{F}(2, 1, 1, 3, 2, 1) = \mathcal{F}(1, 1, 1, 2, 2, 1)$ and $\mathcal{F}(2, 5, 3, 1, 4, 3) = \mathcal{F}(1, 2, 2, 1, 2, 1)$.

Example 2.1 (continued). The second column of Table 1 highlights the structure of the fixed points of $\wedge^2 \vee \wedge \vee^3 \wedge^2 \vee$.

Proposition 2.3.1. *We have*

$$\mathcal{F}(k_1, k_2, \ldots, k_{m-1}, k_m) = \mathcal{F}(1, \min\{k_2, 2\}, \ldots, \min\{k_{m-1}, 2\}, 1)$$

for all positive integers k_i.

Proof. We will use the notation of Lemma 2.3.

We first show that $f = \wedge^{k_1} \vee^{k_2} \wedge^{k_3} \cdots$ and $g = \wedge \vee^{k_2} \wedge^{k_3} \cdots$ have the same number of fixed points. Let $x = (x^1, \ldots, x^m)$ be a fixed point of f. Then, by Lemma 2.3 we have $x^1 = \mathbf{0}$ or $x^1 = \mathbf{1}$. Consider $y = (z, x^2, \ldots, x^m)$, where

fixed points	structure from Lemma 2.3	"reduced" system (Proposition 2.3.1)
000000000000	000 0 0 000 00 00	00 0 0 00 00 00
000000000011	000 0 0 000 00 11	00 0 0 00 00 11
000001110000	000 0 0 111 00 00	00 0 0 11 00 00
000001111111	000 0 0 111 11 11	00 0 0 11 11 11
000001110011	000 0 0 111 00 11	00 0 0 11 00 11
000111110000	000 1 1 111 00 00	00 1 1 11 00 00
000111110011	000 1 1 111 00 11	00 1 1 11 00 11
000111111111	000 1 1 111 11 11	00 1 1 11 11 11
111100000000	111 1 0 000 00 00	11 1 0 00 00 00
111100000011	111 1 0 000 00 11	11 1 0 00 00 11
111111110000	111 1 1 111 00 00	11 1 1 11 00 00
111111110011	111 1 1 111 00 11	11 1 1 11 00 11
111111111111	111 1 1 111 11 11	11 1 1 11 11 11

Table 1. Fixed points of the AND-OR network $\wedge^2 \vee \wedge \vee^3 \wedge^2 \vee$. First column: fixed points. Second column: fixed points with the structure given by Lemma 2.3 highlighted. Third column: fixed points of reduced network, $\wedge^2 \vee \wedge \vee^2 \wedge^2 \vee$, with the structure given by Lemma 2.3 highlighted. For this example, the fixed points can be found using software [Elmeligy Abdelhamid et al. 2015]. We performed computations using resources from the Ohio Supercomputer Center [OSCC 1987].

$z = ((x^1)_1, (x^1)_2)$. It can be checked that y is a fixed point of g. Now, if $y = (z, x^2, \ldots, x^m)$ is a fixed point of g, Lemma 2.3 implies that $z = 0$ or $z = 1$. We define $x = (x^1, \ldots, x^m)$ in the domain of f, where $x^1 = 0$ if $z = 0$ and $x^1 = 1$ if $z = 1$. Then, it can be checked that x is a fixed point of f. This shows that $\mathcal{F}(k_1, k_2, \ldots, k_{m-1}, k_m) = \mathcal{F}(1, k_2, \ldots, k_{m-1}, k_m)$, and similarly it can be shown that $\mathcal{F}(1, k_2, \ldots, k_{m-1}, k_m) = \mathcal{F}(1, k_2, \ldots, k_{m-1}, 1)$.

We now show that for $k_2 \geq 2$, $f = \wedge^{k_1} \vee^{k_2} \wedge^{k_3} \cdots$ and $g = \wedge^{k_1} \vee^2 \wedge^{k_3} \cdots$ have the same number of fixed points. The general case is analogous. Let $x = (x^1, x^2, \ldots, x^m)$ be a fixed point of f. Then, by Lemma 2.3 we have $x^2 = 0$ or $x^2 = 1$. Consider $y = (x^1, z, x^3, \ldots, x^m)$, where $z = ((x^2)_1, (x^2)_2)$. It can be checked that y is a fixed point of g. Now, if $y = (x^1, z, x^3, \ldots, x^m)$ is a fixed point of g, Lemma 2.3 implies that $z = 0$ or $z = 1$. We define $x = (x^1, x^2, \ldots, x^m)$ in the domain of f, where $x^1 = 0$ if $z = 0$ and $x^1 = 1$ if $z = 1$. Then, it can be checked that x is a fixed point of f. This shows that

$$\mathcal{F}(k_1, k_2, \ldots, k_{m-1}, k_m) = \mathcal{F}(k_1, 2, k_3, \ldots, k_{m-1}, k_m) \quad \text{for } k_2 \geq 2. \qquad \square$$
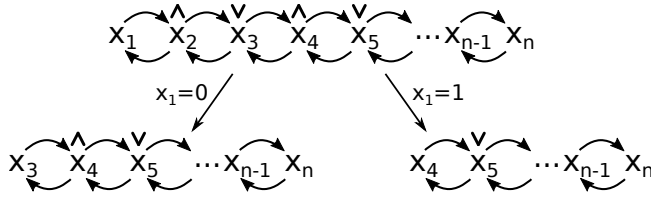
Figure 2. Idea behind the proof of Proposition 2.3.2 (logical operators are included for clarity). Considering the cases $x_1 = 0$ and $x_1 = 1$ yields systems of equations that correspond to smaller AND-OR networks.

Example 2.1 (continued). Proposition 2.3.1 guarantees that $\wedge^2 \vee \wedge \vee^3 \wedge^2 \vee$ and $\wedge \vee \wedge \vee^2 \wedge^2 \vee$ have the same number of fixed points. We can consider the second AND-OR network as a "reduced" version of the original AND-OR network [Veliz-Cuba 2011; Matache and Matache 2016]. This is illustrated in Table 1 (third column).

Proposition 2.3.2. *Let* $r_1, \ldots, r_m \in \{1, 2\}$, *and* $m \geq 2$. *Then, we have*

$$\mathcal{F}(1, r_1, \ldots, r_m, 1)$$
$$= \begin{cases} \mathcal{F}(1, r_3, \ldots, r_m, 1) + \mathcal{F}(r_3, \ldots, r_m, 1) & \textit{for } r_1 = 1, \ r_2 = 1, \\ \mathcal{F}(2, r_3, \ldots, r_m, 1) + \mathcal{F}(1, r_3, \ldots, r_m, 1) & \textit{for } r_1 = 1, \ r_2 = 2, \\ \mathcal{F}(1, 1, r_3, \ldots, r_m, 1) + \mathcal{F}(r_3, \ldots, r_m, 1) & \textit{for } r_1 = 2, \ r_2 = 1, \\ \mathcal{F}(1, 2, r_3, \ldots, r_m, 1) + \mathcal{F}(1, r_3, \ldots, r_m, 1) & \textit{for } r_1 = 2, \ r_2 = 2. \end{cases} \quad (1)$$

Proof. We will use the notation of Lemma 2.3.

If $r_1 = 1, r_2 = 1$, then we claim that any fixed point of $f = \wedge \vee \wedge \vee^{r_3} \wedge^{r_4} \cdots$ is of the form $x = (x^0, x^1, x^2, \ldots, x^m, x^{m+1})$, where either $x^0 = \mathbf{0}$ and $z = (x^1, x^2, \ldots, x^m, x^{m+1})$ is a fixed point of $g = \wedge \vee^{r_3} \wedge^{r_4} \cdots$ or $x^0 = x^1 = \mathbf{1}$ and $z = (x^2, \ldots, x^m, x^{m+1})$ is a fixed point of $h = \vee^{r_3} \wedge^{r_4} \cdots$. Indeed, the system of Boolean equations for fixed points is

$$x_1 = x_2, \quad x_2 = x_1 \wedge x_3, \quad x_3 = x_2 \vee x_4, \quad x_4 = x_3 \wedge x_5, \quad x_5 = x_4 \vee x_6, \quad \ldots, \quad x_n = x_{n-1}.$$

We divide this system of equations into the cases $x_1 = 0$ and $x_1 = 1$. Then, using the fact that $1 = m \wedge n$ implies $m = n = 1$ and that $0 = m \vee n$ implies $m = n = 0$, it follows that we obtain the two systems

$$x_3 = x_4, \quad x_4 = x_3 \wedge x_5, \quad x_5 = x_4 \vee x_6, \quad \ldots, \quad x_n = x_{n-1}$$

and

$$x_4 = x_5, \quad x_5 = x_4 \vee x_6, \quad \ldots, \quad x_n = x_{n-1},$$

corresponding to the cases $x_1 = 0$ and $x_1 = 1$, respectively (see Figure 2). This means that the number of fixed points of f is equal to the number of solutions of these two systems. Since the solutions of the first system are the fixed points of $g = \wedge \vee^{r_3} \wedge^{r_4} \cdots$ and the solutions of the second system are the fixed points of $h = \vee^{r_3} \wedge^{r_4} \cdots$, we obtain

$$\mathcal{F}(1, 1, 1, r_3, \ldots, r_m, 1) = \mathcal{F}(1, r_3, \ldots, r_m, 1) + \mathcal{F}(r_3, \ldots, r_m, 1).$$

The proof for the other three cases is similar. $\square$

We use the convention

$$\mathcal{F}(0, k_1, \ldots, k_m, 0) = \mathcal{F}(k_1, \ldots, k_m, 0) = \mathcal{F}(0, k_1, \ldots, k_m) = \mathcal{F}(k_1, \ldots, k_m),$$

which will simplify the formulation of upcoming results.

Theorem 2.4. *With the convention above, we have that for $m \geq 3$ and $k_i \geq 1$*

$$\mathcal{F}(k_1, \ldots, k_m) = \mathcal{F}(k_2 - 1, k_3, \ldots, k_m) + \mathcal{F}(k_3 - 1, k_4, \ldots, k_m)$$

and

$$\mathcal{F}(k_1, \ldots, k_m) = \mathcal{F}(k_1, \ldots, k_{m-2}, k_{m-1} - 1) + \mathcal{F}(k_1, \ldots, k_{m-3}, k_{m-2} - 1).$$

Also,

$$\mathcal{F}(k_1, k_2) = 3, \qquad \mathcal{F}(k) = 2 \quad \text{for } k \geq 0.$$

Proof. For $m \geq 4$ the result follows directly from Propositions 2.3.1 and 2.3.2. For $m = 3$ the result follows from

$$\mathcal{F}(1, 2, 1) = 5, \quad \mathcal{F}(1, 1, 1) = 4, \quad \mathcal{F}(1, 1) = 3, \quad \mathcal{F}(1) = 2, \quad \text{and} \quad \mathcal{F}(0) = 2,$$

which can be easily checked by complete enumeration. $\square$

Example 2.1 (continued). We now use Theorem 2.4 to find the number of fixed points of $\wedge^2 \vee \wedge \vee^3 \wedge^2 \vee$:

$$\begin{aligned}
\mathcal{F}(2, 1, 1, 3, 2, 1) &= \mathcal{F}(1, 1, 1, 2, 2, 1) \\
&= \mathcal{F}(1 - 1, 1, 2, 2, 1) + \mathcal{F}(1 - 1, 2, 2, 1) \\
&= \mathcal{F}(1, 2, 2, 1) + \mathcal{F}(2, 2, 1) \\
&= \mathcal{F}(2 - 1, 2, 1) + \mathcal{F}(2 - 1, 1) + \mathcal{F}(2 - 1, 1) + \mathcal{F}(1 - 1) \\
&= \mathcal{F}(1, 2, 1) + \mathcal{F}(1, 1) + \mathcal{F}(1, 1) + \mathcal{F}(0) \\
&= \mathcal{F}(2 - 1, 1) + \mathcal{F}(1 - 1) + \mathcal{F}(1, 1) + \mathcal{F}(1, 1) + \mathcal{F}(0) \\
&= \mathcal{F}(1, 1) + \mathcal{F}(0) + \mathcal{F}(1, 1) + \mathcal{F}(1, 1) + \mathcal{F}(0) \\
&= 3 + 2 + 3 + 3 + 2 \\
&= 13
\end{aligned}$$

or

$$\mathcal{F}(2, 1, 1, 3, 2, 1) = \mathcal{F}(1, 1, 1, 2, 2, 1)$$
$$= \mathcal{F}(1, 1, 1, 2, 2-1) + \mathcal{F}(1, 1, 1, 2-1)$$
$$= \mathcal{F}(1, 1, 1, 2, 1) + \mathcal{F}(1, 1, 1, 1)$$
$$= \mathcal{F}(1, 1, 1, 2-1) + \mathcal{F}(1, 1, 1-1) + \mathcal{F}(1, 1, 1-1) + \mathcal{F}(1, 1-1)$$
$$= \mathcal{F}(1, 1, 1, 1) + \mathcal{F}(1, 1) + \mathcal{F}(1, 1) + \mathcal{F}(1)$$
$$= \mathcal{F}(1, 1, 1-1) + \mathcal{F}(1, 1-1) + \mathcal{F}(1, 1) + \mathcal{F}(1, 1) + \mathcal{F}(1)$$
$$= \mathcal{F}(1, 1) + \mathcal{F}(1) + \mathcal{F}(1, 1) + \mathcal{F}(1, 1) + \mathcal{F}(1)$$
$$= 3 + 2 + 3 + 3 + 2$$
$$= 13.$$

In this way, Theorem 2.4 provides a recursive formula to compute the number of fixed points of AND-OR networks with chain topology without the need of exhaustive enumeration. We now study the two special cases of $\mathcal{F}(1, 1, \ldots, 1, 1)$ and $\mathcal{F}(2, 2, \ldots, 2, 2)$.

Define

$$A_n = (1, \underbrace{1, 1, \ldots, 1, 1}_{n \text{ times}}, 1) \quad \text{and} \quad B_n = (2, \underbrace{2, 2, \ldots, 2, 2}_{n \text{ times}}, 2).$$

Also define the sequences $a_0 = 1$, $a_1 = 1$, $a_2 = 1$, and $a_n = a_{n-2} + a_{n-3}$ for $n \geq 3$ and $b_0 = 1$, $b_1 = 1$, and $b_n = b_{n-1} + b_{n-2}$ for $n \geq 2$. Note that (a_n) is the Padovan sequence and (b_n) is the Fibonacci sequence.

Corollary 2.4.1. *With the definitions above we have $\mathcal{F}(A_n) = a_{n+5}$ and $\mathcal{F}(B_n) = b_{n+3}$ for $n \geq 0$, and the sharp bounds $\mathcal{F}(A_n) \leq \mathcal{F}(1, r_1, r_2, \ldots, r_n, 1) \leq \mathcal{F}(B_n)$ for all $r_i \geq 1$.*

Proof. It follows from Theorem 2.4 or Proposition 2.3.2 using induction. □

3. Infinite and closed chain

In this section we study the cases of AND-OR networks with infinitely many variables and when the topology is a closed chain.

When the AND-OR network has infinitely many variables we have a infinite collection of Boolean functions $f = (\ldots, f_{-2}, f_{-1}, f_0, f_1, f_2, \ldots)$ such that $f_i = x_{i-1} \wedge x_{i+1}$ or $f_i = x_{i-1} \vee x_{i+1}$. We can use the notation of Section 2 and denote consecutive logical operators as $\wedge^k$ or $\vee^k$, where k could also be ∞. Also, we can simply use the exponents to represent the AND-OR network. For example, $(\infty, 1, 2, \infty)$ and $\wedge^\infty \vee \wedge^2 \vee^\infty$ represent the AND-OR network $\cdots \wedge\wedge\wedge\vee\wedge\wedge\vee\vee\vee \cdots$.

Similarly, $(\ldots, 1, 1, 2, 1, 1, 2, 1, 1, 2, \ldots)$ and $\cdots \wedge \vee \wedge^2 \vee \wedge \vee^2 \wedge \vee \wedge^2 \cdots$ represent the AND-OR network $\cdots \wedge \vee \wedge \wedge \vee \wedge \vee \vee \wedge \vee \wedge \wedge \cdots$.

The following theorem allows us to use the results from Section 2 to study AND-OR networks with infinitely many variables.

Theorem 3.1. *With the notation above and $k_i \geq 1$ we have*

$$\mathcal{F}(\infty) = 2,$$

$$\mathcal{F}(\infty, k_1, k_2, \ldots, k_{m-1}, k_m, \infty) = \mathcal{F}(1, k_1, k_2, \ldots, k_{m-1}, k_m, 1),$$

$$\mathcal{F}(\infty, k_1, k_2, k_3, \ldots) = \infty,$$

$$\mathcal{F}(\ldots, k_{-3}, k_{-2}, k_{-1}, \infty) = \infty,$$

$$\mathcal{F}(\ldots, k_{-3}, k_{-2}, k_{-1}, k_0, k_1, k_2, k_3, \ldots) = \infty.$$

Proof. To prove the first equality we consider the AND-OR network where all logical operators are $\wedge$. If one of the variables is 0, it follows that all the other variables are also 0. Similarly, if one of the variables is 1, all the other variables are also 1. Thus, the only fixed points of this AND-OR network are **0** and **1**.

The second equality follows the same approach seen in Proposition 2.3.1.

To prove the third equality we first observe that $\mathcal{F}(\infty, k_1, k_2, k_3, \ldots) = \mathcal{F}(1, k_1, k_2, k_3, \ldots)$. Now, we will show that any fixed point of the AND-OR network $\mathcal{F}(1, k_1, k_2, k_3, \ldots, k_r)$ defines a fixed point of $\mathcal{F}(1, k_1, k_2, k_3, \ldots)$. Indeed, using the notation of Lemma 2.3, a fixed point of the AND-OR network $\mathcal{F}(1, k_1, \ldots, k_r)$ has the form $\boldsymbol{x} = (\boldsymbol{x}^0, \boldsymbol{x}^1, \ldots, \boldsymbol{x}^r)$. Then, denoting $\boldsymbol{z} = (1, 1, \ldots)$ if $\boldsymbol{x}^r = \boldsymbol{1}$ and $\boldsymbol{z} = (0, 0, \ldots)$ if $\boldsymbol{x}^r = \boldsymbol{0}$, it follows that $(\boldsymbol{x}^0, \boldsymbol{x}^1, \ldots, \boldsymbol{x}^r, \boldsymbol{z})$ is a fixed point of $\mathcal{F}(1, k_1, k_2, k_3, \ldots)$. Since r is arbitrary, $\mathcal{F}(1, k_1, \ldots, k_r)$ is not bounded (see Corollary 2.4.1) and the number of fixed points of $\mathcal{F}(1, k_1, \ldots)$ is ∞. The last two equalities are similar. $\square$

When the topology of the network is a closed chain, we have the network

$$
\begin{aligned}
f_1 &= x_n \Diamond_1 x_2, & f_{n-2} &= x_{n-3} \Diamond_{n-2} x_{n-1}, \\
f_2 &= x_1 \Diamond_2 x_3, & f_{n-1} &= x_{n-2} \Diamond_{n-1} x_n, \\
&\;\;\vdots & f_n &= x_{n-1} \Diamond_n x_1.
\end{aligned}
$$

We denote this network as $[k_1, k_2, \ldots, k_r]$ or any cyclic permutation that groups consecutive logical operators. Thus, the AND-OR network

$$
\begin{aligned}
f_1 &= x_n \wedge x_2, & f_3 &= x_2 \wedge x_4, & f_5 &= x_4 \vee x_6, \\
f_2 &= x_1 \vee x_3, & f_4 &= x_3 \vee x_5, & f_6 &= x_5 \wedge x_1,
\end{aligned}
$$

will not be denoted by $[1, 1, 1, 2, 1]$ ("splitting" the first and last $\wedge$'s), but by $[1, 1, 2, 2]$, $[1, 2, 2, 1]$, $[2, 2, 1, 1]$, or $[2, 1, 1, 2]$ (combining the first and last $\wedge$'s).

This means that r in $[k_1, k_2, \ldots, k_r]$ will always be an even number or equal to 1. The number of fixed points will be denoted by $\mathcal{F}[k_1, k_2, \ldots, k_r]$. The following propositions and theorem allow us to use the results from Section 2 to study AND-OR networks with closed chain topology.

Proposition 3.1.1. *With the notation above, we have that for $k_i \geq 1$*

$$\mathcal{F}[k_1, k_2, \ldots, k_r] = \mathcal{F}[\min\{2, k_1\}, \min\{2, k_2\}, \ldots, \min\{2, k_r\}].$$

Proof. It is analogous to the proof of Proposition 2.3.1. $\qquad\square$

Proposition 3.1.2. *Consider $k_i \geq 1$, $m \geq 6$, and $l \geq 8$. Then,*

$$\mathcal{F}[2, k_2, \ldots, k_m] = \mathcal{F}(k_2 - 1, k_3, \ldots, k_{m-1}, k_m - 1) + \mathcal{F}(k_3 - 1, k_4, \ldots, k_{m-2}, k_{m-1} - 1),$$

$$\mathcal{F}[1, k_2, \ldots, k_l] = \mathcal{F}(k_3 - 1, k_4, \ldots, k_{l-1} - 1) + \mathcal{F}(k_4 - 1, k_5, \ldots, k_{l-1}, k_l - 1)$$
$$+ \mathcal{F}(k_2 - 1, k_3, \ldots, k_{l-3}, k_{l-2} - 1) - \mathcal{F}(k_4 - 1, k_5, \ldots, k_{l-3}, k_{l-2} - 1).$$

Proof. The first equality is analogous to Proposition 2.3.2. To prove the second equality we use the notation of Lemma 2.3.

We have several cases to consider for $k_{l-2}, k_{l-1}, k_l, k_2, k_3$, and k_4. We focus on the case $k_{l-2} = k_{l-1} = k_l = k_2 = k_3 = k_4 = 1$ since the other cases are analogous. Note that we want to prove

$$\mathcal{F}[1, 1, 1, 1, k_5, \ldots, k_{l-3}, 1, 1, 1] = \mathcal{F}(1, k_5, \ldots, k_{l-3}, 1) + \mathcal{F}(k_5, \ldots, k_{l-3}, 1, 1)$$
$$+ \mathcal{F}(1, 1, k_5, \ldots, k_{l-3}) - \mathcal{F}(k_5, \ldots, k_{l-3}).$$

The fixed points of the AND-OR network are the solutions of

$$\begin{aligned}
x_1 &= x_n \wedge x_2, & x_{n-3} &= x_{n-4} \wedge x_{n-2}, \\
x_2 &= x_1 \vee x_3, & x_{n-2} &= x_{n-3} \vee x_{n-1}, \\
x_3 &= x_2 \wedge x_4, & x_{n-1} &= x_{n-2} \wedge x_n, \\
&\vdots & x_n &= x_{n-1} \vee x_1.
\end{aligned}$$

We now consider the cases $x_1 = 1$ and $x_1 = 0$ (see Figure 3). The case $x_1 = 1$ yields the system of equations

$$\begin{aligned}
x_3 &= x_4, & x_{n-4} &= x_{n-5} \vee x_{n-3}, \\
x_4 &= x_3 \vee x_5, & x_{n-3} &= x_{n-4} \wedge x_{n-2}, \\
x_5 &= x_4 \wedge x_6, & x_{n-2} &= x_{n-3} \vee x_{n-1}, \\
&\vdots & x_{n-1} &= x_{n-2},
\end{aligned}$$

which has $\mathcal{F}(1, k_5, \ldots, k_{l-3}, 1)$ solutions. On the other hand, when we consider $x_1 = 0$ the first equation becomes $x_n \wedge x_2 = 0$. We now have two subcases: $x_n = 0$
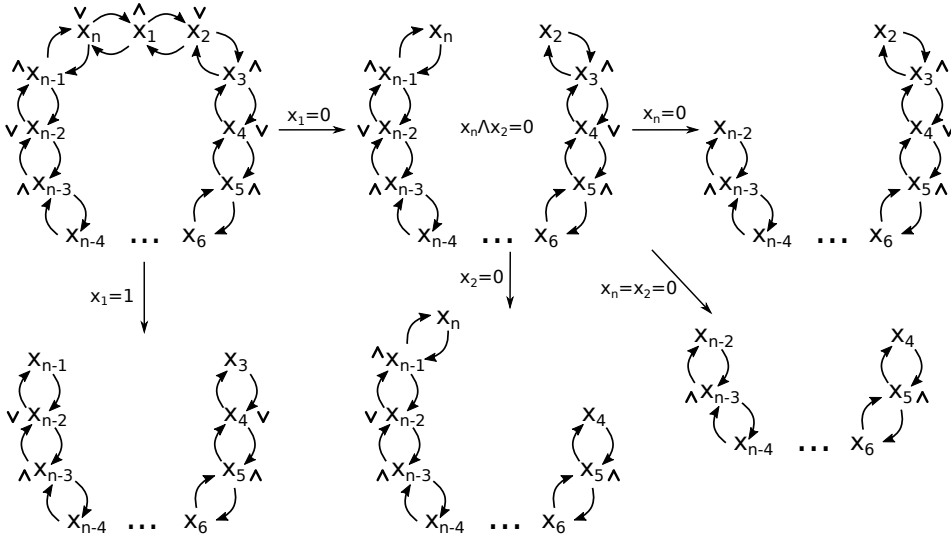
Figure 3. Idea behind the proof of Proposition 3.1.2 (logical operators are included for clarity). Considering the case $x_1 = 1$ yields a system of equations that corresponds to a smaller AND-OR network. Considering the case $x_1 = 0$ yields a system of equations that does not correspond to an AND-OR network (due to the equation $x_n \wedge x_2 = 0$). However, the subcases $x_n = 0$ and $x_2 = 0$ yield systems of equations that do correspond to smaller AND-OR networks. These two systems have overlapping solutions, so we must also take into consideration the common case $x_n = x_2 = 0$ when counting the number of fixed points.

and $x_2 = 0$. The subcase $x_n = 0$ yields

$$x_2 = x_3, \qquad x_{n-4} = x_{n-5} \vee x_{n-3},$$
$$x_3 = x_2 \wedge x_4, \qquad x_{n-3} = x_{n-4} \wedge x_{n-2},$$
$$\vdots \qquad x_{n-2} = x_{n-3},$$

which has $\mathcal{F}(1, 1, k_5, \ldots, k_{l-3})$ solutions. The subcase $x_2 = 0$ yields

$$x_4 = x_5, \qquad x_{n-2} = x_{n-3} \vee x_{n-1},$$
$$x_5 = x_4 \wedge x_6, \qquad x_{n-1} = x_{n-2} \wedge x_n,$$
$$\vdots \qquad x_n = x_{n-1},$$

which has $\mathcal{F}(k_5, \ldots, k_{l-3}, 1, 1)$ solutions. Thus, adding up these three numbers we obtain $\mathcal{F}(1, k_5, \ldots, k_{l-3}, 1) + \mathcal{F}(k_5, \ldots, k_{l-3}, 1, 1) + \mathcal{F}(1, 1, k_5, \ldots, k_{l-3})$. However, this is not $\mathcal{F}[1, 1, 1, 1, k_5, \ldots, k_{l-3}, 1, 1, 1]$, since the subcases $x_n = 0$ and

$x_2 = 0$ overlap. We need to subtract the number of solutions of the system

$$x_4 = x_5, \qquad\qquad x_{n-4} = x_{n-5} \vee x_{n-3},$$
$$x_5 = x_4 \wedge x_6, \qquad\qquad x_{n-3} = x_{n-4} \wedge x_{n-2},$$
$$\vdots \qquad\qquad\qquad x_{n-2} = x_{n-3},$$

which has $\mathcal{F}(k_5, \ldots, k_{l-3})$ solutions. Then, the result follows. $\qquad\square$

We now declare some conventions to write Proposition 3.1.2 more compactly. We define $\mathcal{F}(-1) = 1$, $(k_s - 1, \ldots, k_s - 1) = (k_s - 2)$, and $(k_s - 1, \ldots, k_t - 1) = (-1)$ for $s > t$.

Theorem 3.2. *With the conventions above, we have that for $m \geq 4$ and $k_i \geq 1$*

$$\mathcal{F}[2, k_2, \ldots, k_r] = \mathcal{F}(k_2 - 1, k_3, \ldots, k_{r-1}, k_r - 1) + \mathcal{F}(k_3 - 1, k_4, \ldots, k_{r-2}, k_{r-1} - 1),$$

$$\mathcal{F}[1, k_2, \ldots, k_r] = \mathcal{F}(k_3 - 1, k_4, \ldots, k_{r-1}, k_r - 1) + \mathcal{F}(k_4 - 1, k_5, \ldots, k_{r-1}, k_r - 1)$$
$$+ \mathcal{F}(k_2 - 1, k_3, \ldots, k_{r-3}, k_{r-2} - 1) - \mathcal{F}(k_4 - 1, k_5, \ldots, k_{r-3}, k_{r-2} - 1).$$

Also,

$$\mathcal{F}[k] = 2 \quad \text{for } k \geq 3,$$
$$\mathcal{F}[k, 1] = 2 \quad \text{for } k \geq 2,$$
$$\mathcal{F}[k_1, k_2] = 3 \quad \text{for } k_1, k_2 \geq 2,$$

Proof. The first two equalities follow directly from Propositions 3.1.1 and 3.1.2 using the convention declared above. The last three equalities follow from Proposition 3.1.1 and $\mathcal{F}[3] = \mathcal{F}[2, 1] = 2$ and $\mathcal{F}[2, 2] = 3$, which can be verified by complete enumeration. $\qquad\square$

As in Section 2, we now consider the cases

$$A_n = (1, \underbrace{1, 1, \ldots, 1, 1}_{n \text{ times}}, 1) \quad \text{and} \quad B_n = (2, \underbrace{2, 2, \ldots, 2, 2}_{n \text{ times}}, 2).$$

We denote the number of fixed points of the corresponding AND-OR networks with closed chain topology by $\mathcal{F}[A_n]$ and $\mathcal{F}[B_n]$, respectively.

Corollary 3.2.1. *With the notation above we have* $\mathcal{F}[A_n] = 3a_n - a_{n-2}$ *and* $\mathcal{F}[B_n] = b_{n+2} + b_n$ *for $n \geq 2$, and the sharp bounds*

$$\mathcal{F}[A_n] \leq \mathcal{F}[k_0, k_1, \ldots, k_n, k_{n+1}] \leq \mathcal{F}[B_n]$$

for all $r_i \geq 1$

Proof. The proof follows from first using Theorem 3.2 and then Corollary 2.4.1. $\square$

Example 3.3. We consider

$$f_1 = x_{12} \wedge x_2, \quad f_4 = x_3 \vee x_5, \quad f_7 = x_6 \vee x_8, \quad f_{10} = x_9 \wedge x_{11},$$
$$f_2 = x_1 \wedge x_3, \quad f_5 = x_4 \wedge x_6, \quad f_8 = x_7 \vee x_9, \quad f_{11} = x_{10} \vee x_{12},$$
$$f_3 = x_2 \wedge x_4, \quad f_6 = x_5 \vee x_7, \quad f_9 = x_8 \wedge x_{10}, \quad f_{12} = x_{11} \vee x_1.$$

We will use Theorems 2.4 and 3.2 for the representations $[3, 1, 1, 3, 2, 2]$ and $[1, 3, 2, 2, 3, 1]$ of f.

$$
\begin{aligned}
\mathcal{F}[3, 1, 1, 3, 2, 2] &= \mathcal{F}[2, 1, 1, 2, 2, 2] \\
&= \mathcal{F}(1-1, 1, 2, 2, 2-1) + \mathcal{F}(1-1, 2, 2-1) \\
&= \mathcal{F}(1, 2, 2, 1) + \mathcal{F}(2, 1) \\
&= \mathcal{F}(2-1, 2, 1) + \mathcal{F}(2-1, 1) + \mathcal{F}(2, 1) \\
&= \mathcal{F}(1, 2, 1) + \mathcal{F}(1, 1) + \mathcal{F}(2, 1) \\
&= \mathcal{F}(2-1, 1) + \mathcal{F}(1-1) + \mathcal{F}(1, 1) + \mathcal{F}(2, 1) \\
&= \mathcal{F}(1, 1) + \mathcal{F}(0) + \mathcal{F}(1, 1) + \mathcal{F}(2, 1) \\
&= 3 + 2 + 3 + 3 = 11,
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{F}[1, 3, 2, 2, 3, 1] &= \mathcal{F}[1, 2, 2, 2, 2, 1] \\
&= \mathcal{F}(2-1, 2, 2-1) + \mathcal{F}(2-1, 2, 1-1) + \mathcal{F}(2-1, 2, 2-1) - \mathcal{F}(2-2) \\
&= \mathcal{F}(1, 2, 1) + \mathcal{F}(1, 2) + \mathcal{F}(1, 2, 1) - \mathcal{F}(0) \\
&= \mathcal{F}(1, 1) + \mathcal{F}(0) + \mathcal{F}(1, 2) + \mathcal{F}(1, 1) + \mathcal{F}(0) - \mathcal{F}(0) \\
&= 3 + 2 + 3 + 3 + 2 - 2 = 11.
\end{aligned}
$$

4. Final remarks: coupled chains

Although the results in this manuscript are for chain topology, we now show how our techniques could also be used for coupled chains. These couplings could be considered as "fractal" versions of the 1-dimensional chains that we covered in previous sections. However, due to the complex couplings that could be attained and the different cases that appear (e.g., the proof of Proposition 3.1.1 has 2^6 subcases per case), a single proposition that covers all cases would be unfeasible. Thus, we will consider two examples featuring different couplings of chains: a coupling of three open chains, and a coupling of an open and a closed chain.

First, we will prove two lemmas that will allow us to handle intersections of chains. To make the notation simpler, a pair of edges between two vertices will be simply denoted by a single undirected edge (Figure 4). When it is not required to label vertices, we will use an even simpler representation of the wiring diagram (top
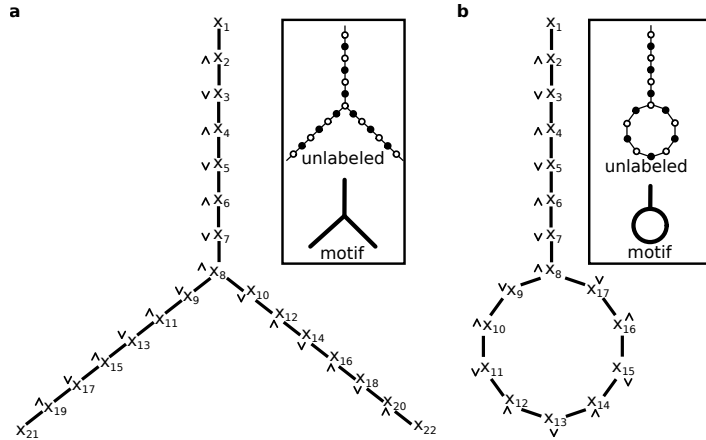
Figure 4. Coupling chains. (a) Wiring diagram of an AND-OR network consisting of the coupling of three open chains. Each undirected edge represents two edges as shown in Figure 1. For example, $f_1 = x_2$, $f_2 = x_1 \wedge x_3$, and $f_8 = x_7 \wedge x_9 \wedge x_{10}$. Vertex x_1 could also be assigned the AND or OR operator. (b) Wiring diagram of an AND-OR network consisting of the coupling of a closed and open chain. In both panels, the insets show the simplified representation of the wiring diagram where the labels of variables are omitted. Open circles indicate the AND operator, whereas filled circles indicate the OR operator. The vertex corresponding to x_1 is left blank, but could also be assigned a filled or open circle. The insets also show an undirected graph highlighting the coupling motif of the AND-OR network. In previous sections the coupling motif would simply be a (finite or infinite) line or a circle.

insets in Figure 4). Also, we will use $\mathcal{F}[f]$ to denote the number of fixed points of an AND-OR network f.

Consider an AND-OR network, $f : \{0, 1\}^n \to \{0, 1\}^n$, and $S \subseteq \{1, 2, \ldots, n\}$. We define a new network $g : \{0, 1\}^{n-|S|} \to \{0, 1\}^{n-|S|}$ in the variables $\{x_i : i \notin S\}$, denoted by $g = f \backslash S$, as follows:

(1) Remove the vertices in $\{x_i : i \in S\}$ from the wiring diagram of f. Note that this also means that we remove the edges of the forms $x_i \to x_j$ and $x_k \to x_i$, where $i \in S$.

(2) For each variable x_k in the new wiring diagram, g_k will be the same logical operator as in the wiring diagram of f, but may possibly depend on less variables. Note that the operators $\vee$ and $\wedge$ on a single variable are simply the identity function.
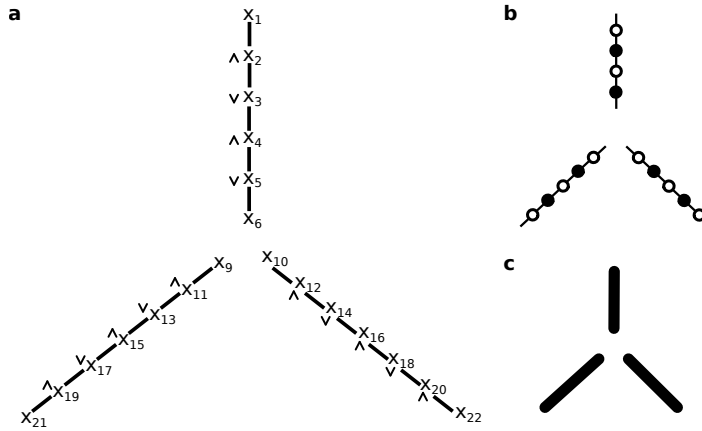
Figure 5. Example of $f \setminus S$. (a) Wiring diagram of the AND-OR network $f \setminus S$, where f is given in Figure 4(a) and $S = \{7, 8\}$. (b) Simplified representation of the wiring diagram with labels omitted. (c) Undirected graph highlighting the coupling motif of $f \setminus S$. Note that this graph now has three connected components, all of them being open chains.

Example 4.1. Consider the AND-OR network with wiring diagram given by Figure 4(a) and let $S = \{7, 8\}$. The new network $g = f \setminus S$ has wiring diagram shown in Figure 5(a). Note that g depends on 20 variables and variables x_6, x_9, and x_{10} each depend on a single variable only (e.g., the Boolean function corresponding to x_6 is $g_6 = x_5$).

We now state and prove the lemmas.

Lemma 4.2. *Consider an AND-OR network f consisting of coupled chains such that x_n depends on two or more variables as shown in Figure 6(a). Denote with $R = \{1, 2, \ldots, r\}$. Then, the number of steady states of f is equal to*

$$\mathcal{F}[f] = \mathcal{F}[f \setminus (\{n\} \cup \{i_s : s \in R\})]$$
$$+ \sum_{\varnothing \neq S \subseteq R} (-1)^{|S|+1} \mathcal{F}[f \setminus (\{n\} \cup \{i_s : s \in S\} \cup \{j_s : s \in S\})].$$

Proof. We proceed by cases as in the proof of Proposition 3.1.2.

If $x_n = 1$, then any fixed point of f will satisfy $x_{i_s} = x_{j_s} \vee 1 = 1$ and $x_{j_s} = x_{k_s} \wedge x_{i_s} = x_{k_s} \wedge 1 = x_{k_s}$. Thus, x is a fixed point of f if and only if $y = (x_s)_{s \in \{1,\ldots,n\} \setminus \{n, i_1,\ldots,i_r\}}$ is a fixed point of the AND-OR network with wiring diagram given in Figure 6(b). This smaller network is precisely $f \setminus (\{n\} \cup \{i_s : s \in R\})$.

If $x_n = 0$, then any fixed point of f will satisfy $x_{i_1} \wedge \cdots \wedge x_{i_r} = 0$ (Figure 6(c)). This does not correspond to a system of equations of an AND-OR network, so we consider the subcases $x_{i_s} = 0$ for each $s \in R = \{1, 2, \ldots, r\}$.
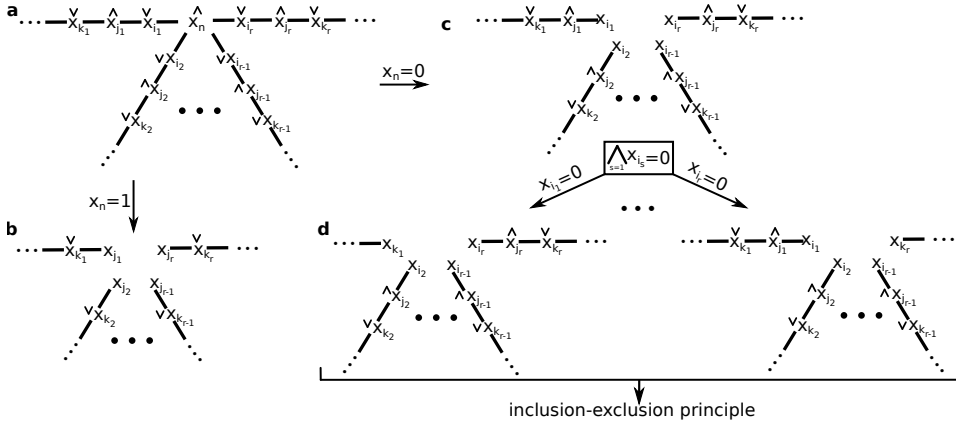
Figure 6. Idea of the proof in Lemma 4.2. (a) Wiring diagram of AND-OR network f that has a vertex that depends on two or more variables. The three dots represent other variables in the r chains that could potentially intersect as in Figure 4(b). We consider two cases, $x_n = 1$ and $x_n = 0$ in the system of equations $f(x) = x$. (b) In the case $x_n = 1$, we obtain a smaller system of equations that corresponds to a smaller wiring diagram of an AND-OR network. (c) In the case $x_n = 0$, the system of equations does not correspond to an AND-OR network due to the condition $\bigwedge_{s=1}^{r} x_{i_s} = 0$. (d) To obtain AND-OR networks we consider the subcases $x_{i_1} = 0$, $x_{i_2} = 0, \dots$, and $x_{i_r} = 0$. However, there is overlap of fixed points between the different subcases, so we use the inclusion-exclusion principle.

If $x_{i_s} = 0$, then $x_{i_1} \wedge \cdots \wedge x_{i_r} = 0$ is satisfied. Also, $x_{j_s} = x_{k_s} \wedge 0 = 0$ and then x_{k_s} will depend on a single variable only (see Figure 6(d) for the cases $s = 1$ and $s = r$). The resulting AND-OR network is $f \backslash (\{n, i_s, j_s\})$. Note that these subcases overlap, so we use the inclusion-exclusion principle to properly account for this. For the case $x_n = 0$, the inclusion exclusion principle implies that the number of fixed points is

$|\{$fixed points of the form $x_n = 0\}|$

$$= \sum_{\varnothing \neq S \subseteq R} (-1)^{|S|+1} |\{\text{fixed points of the form } x_{i_s} = 0 \text{ for } s \in S\}|.$$

We now claim that

$$|\{\text{fixed points of the form } x_{i_s} = 0 \text{ for } s \in S\}| = \mathcal{F}[f \backslash (\{n\} \cup \{i_s : s \in S\} \cup \{j_s : s \in S\})].$$

Indeed, if $x_{i_s} = 0$ for $s \in S$, it follows that $x_{j_s} = 0$ and that x_{k_s} depends on a single variable only for $s \in S$. The AND-OR network corresponding to this is precisely $f \backslash (\{n\} \cup \{i_s : s \in S\} \cup \{j_s : s \in S\})$.

$$\mathcal{F}[\,\lambda\,] = \mathcal{F}[\,\lambda\,] + \mathcal{F}[\,\lambda\,] + \mathcal{F}[\,\lambda\,] + \mathcal{F}[\,\lambda\,]$$

$$\qquad\qquad - \mathcal{F}[\,\lambda\,] - \mathcal{F}[\,\lambda\,] - \mathcal{F}[\,\lambda\,] + \mathcal{F}[\,\lambda\,]$$

$$\mathcal{F}[\,\lambda\,] = \mathcal{F}[\,\vdots\,] \times \mathcal{F}[\,\diagup\,] \times \mathcal{F}[\,\diagdown\,] = (\mathcal{F}[\,\vdots\,])^3$$

$$\mathcal{F}[\,\lambda\,] = \mathcal{F}[\,\vdots\,] \times (\mathcal{F}[\,\vdots\,])^2$$

$$\mathcal{F}[\,\lambda\,] = \mathcal{F}[\,\vdots\,] \times (\mathcal{F}[\,\vdots\,])^2$$

$$\mathcal{F}[\,\lambda\,] = (\mathcal{F}[\,\vdots\,])^3$$

Figure 7. Using our results to find the number of fixed points of a coupling of three open chains.

The proof then follows by adding the total number of fixed points from the cases $x_n = 1$ and $x_n = 0$. □

Lemma 4.3. *Suppose a Boolean network f is the Cartesian product of h and g; that is, up to a relabeling of variables, $f(x, y) = (g(x), h(y))$ (also denoted by $f = g \times h$). Then,*

$$\mathcal{F}[f] = \mathcal{F}[g]\,\mathcal{F}[h].$$

Proof. This follows from the fact that (x, y) is a steady state of f if and only if x is a steady state of g and y is a steady state of h. □

With these lemmas we can now find the number of fixed points of the AND-OR networks given in Figure 4. For notational purposes, we apply the lemmas using the unlabeled representation of the wiring diagrams.

Example 4.4. Consider the AND-OR network with wiring diagram given by Figure 4(a). We use Lemma 4.2 to split the wiring diagram at x_8. The process is shown in Figure 7. Using this lemma, we find that the number of fixed points can be written as a sum/difference of the number of fixed points of disjoint chains. Then, we use Lemma 4.3 to express the number of fixed points as an algebraic combination of the number of fixed points of single chains. Once we have single chains, we can use the results from previous sections. Thus, the number of fixed points is

$$\mathcal{F}[f] = (\mathcal{F}(1, 1, 1, 1))^3 + 3\mathcal{F}(1, 1, 1)(\mathcal{F}(1, 1, 1, 1, 1))^2$$
$$\qquad\qquad - 3\mathcal{F}(1, 1, 1, 1, 1)(\mathcal{F}(1, 1, 1))^2 + (\mathcal{F}(1, 1, 1))^3$$
$$= (5)^3 + 3(4)(7)^2 - 3(7)(4)^2 + (4)^3 = 441.$$

$$\mathcal{F}[\;\text{⬡}\;] = \mathcal{F}[\;]+\mathcal{F}[\;]+\mathcal{F}[\;]+\mathcal{F}[\;]$$

$$-\mathcal{F}[\;]-\mathcal{F}[\;]-\mathcal{F}[\;]+\mathcal{F}[\;]$$

Figure 8. Using our results to find the number of fixed points of a coupling of an open chain and a closed chain.

Example 4.5. Consider the AND-OR network with wiring diagram given by Figure 4(b). We use Lemma 4.2 to split the wiring diagram at x_8 and then we use Lemma 4.3. The process is shown in Figure 8. Analogous to the previous example, we obtain

$$\mathcal{F}[f] = \mathcal{F}(1, 1, 1, 1)\mathcal{F}(1, 1, 1, 1, 1) + \mathcal{F}(1, 1, 1)\mathcal{F}(1, 1, 1, 1, 1, 1, 1)$$
$$+ 2(\mathcal{F}(1, 1, 1, 1))^2 - 3\mathcal{F}(1, 1, 1)\mathcal{F}(1, 1, 1, 1, 1) + (\mathcal{F}(1, 1, 1))^2$$
$$= (5)(7) + (4)(12) + 2(7)^2 - 3(4)(7) + (4)^2 = 113.$$

5. Conclusion

Our results provide recursive formulas and sharp bounds for the number of fixed points of AND-OR networks with chain topology. Other work regarding the number of fixed points has focused on bounds with respect to the number of nodes [Aracena et al. 2004]. Our results, on the other hand, focus on formulas and bounds with respect to the pattern of logical operators. Thus, our findings complement previous results. Our approach can potentially be extended to cases where an AND-OR network has a topology that can be seen as the "combination" of open chains. Then, the number of fixed points of the original AND-OR network will be given by the inclusion-exclusion principle in terms of the number of fixed points of the AND-OR networks with open chain topology. Indeed, Section 4 shows how our approach can be used in such cases.

References

[Agur et al. 1988] Z. Agur, A. S. Fraenkel, and S. T. Klein, "The number of fixed points of the majority rule", *Discrete Math.* **70**:3 (1988), 295–302. MR Zbl

[Akutsu et al. 1998] T. Akutsu, S. Kuhara, O. Maruyama, and S. Miyano, "A system for identifying genetic networks from gene expression patterns produced by gene disruptions and overexpressions", *Genome Inform.* **9** (1998), 151–160.

[Albert and Othmer 2003] R. Albert and H. G. Othmer, "The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster*", *J. Theoret. Biol.* **223**:1 (2003), 1–18. MR

[Alcolei et al. 2016] A. Alcolei, K. Perrot, and S. Sené, "On the flora of asynchronous locally non-monotonic Boolean automata networks", *Electr. Notes Theor. Comp. Sci.* **326** (2016), 3–25. Zbl

[Aracena 2008] J. Aracena, "Maximum number of fixed points in regulatory Boolean networks", *Bull. Math. Biol.* **70**:5 (2008), 1398–1409. MR Zbl

[Aracena et al. 2004] J. Aracena, J. Demongeot, and E. Goles, "Fixed points and maximal independent sets in AND-OR networks", *Discrete Appl. Math.* **138**:3 (2004), 277–288. MR Zbl

[Bollman et al. 2010] D. Bollman, O. Colón-Reyes, V. A. Ocasio, and E. Orozco, "A control theory for Boolean monomial dynamical systems", *Discrete Event Dyn. Syst.* **20**:1 (2010), 19–35. MR Zbl

[Dimitrova et al. 2015] E. S. Dimitrova, O. I. Yordanov, and M. T. Matache, "Difference equation for tracking perturbations in systems of Boolean nested canalyzing functions", *Phys. Rev. E* (3) **91**:6 (2015), art. id. 062812. MR

[Elmeligy Abdelhamid et al. 2015] S. H. Elmeligy Abdelhamid, C. J. Kuhlman, M. V. Marathe, H. S. Mortveit, and S. S. Ravi, "GDSCalc: a web-based application for evaluating discrete graph dynamical systems", *PLOS One* **10**:8 (2015), art. id. e0133660.

[Goles et al. 2000] E. Goles, M. Matamala, and P. A. Estévez, "Dynamical properties of min-max networks", *Int. J. Neural Syst.* **10**:6 (2000), 467–473.

[Hansson et al. 2005] A. Å. Hansson, H. S. Mortveit, and C. M. Reidys, "On asynchronous cellular automata", *Adv. Complex Syst.* **8**:4 (2005), 521–538. MR Zbl

[Jarrah et al. 2007] A. S. Jarrah, B. Raposa, and R. Laubenbacher, "Nested canalyzing, unate cascade, and polynomial functions", *Phys. D* **233**:2 (2007), 167–174. MR Zbl

[Jarrah et al. 2010] A. S. Jarrah, R. Laubenbacher, and A. Veliz-Cuba, "The dynamics of conjunctive and disjunctive Boolean network models", *Bull. Math. Biol.* **72**:6 (2010), 1425–1447. MR Zbl

[Matache and Matache 2016] M. T. Matache and V. Matache, "Logical reduction of biological networks to their most determinative components", *Bull. Math. Biol.* **78**:7 (2016), 1520–1545. MR Zbl

[Mendoza and Xenarios 2006] L. Mendoza and I. Xenarios, "A method for the generation of standardized qualitative dynamical systems of regulatory networks", *Theor. Bio. Medical Model.* **3**:1 (2006), art. id. 13.

[Murrugarra and Laubenbacher 2011] D. Murrugarra and R. Laubenbacher, "Regulatory patterns in molecular interaction networks", *J. Theoret. Biol.* **288** (2011), 66–72. MR Zbl

[OSCC 1987] Ohio Supercomputer Center, "Ohio Supercomputer Center", homepage, 1987, available at https://tinyurl.com/osuosc.

[Veliz-Cuba 2011] A. Veliz-Cuba, "Reduction of Boolean network models", *J. Theoret. Biol.* **289** (2011), 167–172. MR Zbl

[Veliz-Cuba et al. 2014a] A. Veliz-Cuba, A. Kumar, and K. Josić, "Piecewise linear and Boolean models of chemical reaction networks", *Bull. Math. Biol.* **76**:12 (2014), 2945–2984. MR Zbl

[Veliz-Cuba et al. 2014b] A. Veliz-Cuba, D. Murrugarra, and R. Laubenbacher, "Structure and dynamics of acyclic networks", *Discrete Event Dyn. Syst.* **24**:4 (2014), 647–658. MR Zbl

[Wang et al. 2017] Y. Wang, B. Omidiran, F. Kigwe, and K. Chilakamarri, "Relations between the conditions of admitting cycles in Boolean and ODE network systems", *Involve* **10**:5 (2017), 813–831. MR Zbl

[Weiss and Margaliot 2017] E. Weiss and M. Margaliot, "A polynomial-time algorithm for solving the minimal observability problem in conjunctive Boolean networks", preprint, 2017. arXiv

avelizcuba1@udayton.edu   *Department of Mathematics, University of Dayton, Dayton, OH, United States*

geiserl1@udayton.edu   *University of Dayton, Dayton, OH, United States*

Positive solutions to singular second-order boundary value problems for dynamic equations

Curtis Kunkel and Alex Lancaster

(Communicated by Johnny Henderson)

We study singular second-order boundary value problems with mixed boundary conditions on an infinitely discrete time scale. We prove the existence of a positive solution by means of a lower and upper solutions method and the Brouwer fixed-point theorem, in conjunction with perturbation methods used to approximate regular problems.

1. Introduction

This paper continues the work done previously by Kunkel [2008], where he studied a singular second-order boundary value problem in purely discrete time scales of nonuniform step size. Although similar throughout most of the time scale, this result is different in the fact that the time scale itself has a limit point at the right-side boundary condition, forcing a nearly continuous behavior at that end. If this limiting condition were not present, the result would be trivial using [Kunkel 2008], but as it stands, this result continues to expand the work of that paper to another type of time scale.

More specifically, [Kunkel 2008] dealt with the discrete boundary value problem

$$u^{\Delta\Delta}(t_{i-1}) + f(t_i, u(t_i), u^{\Delta}(t_{i-1})) = 0, \quad t \in \mathbb{T}^{\circ},$$
$$u^{\Delta}(t_0) = u(t_{n+1}) = 0,$$

where $\mathbb{T}^{\circ}$ is the discrete interval of nonuniform step size $[t_1, t_n] := \{t_1, t_2, \ldots, t_n\}$ and $f(t, x, y)$ is singular in x. This work was an extension of a previous result by Rachůnková and Rachůnek [2006], where they studied a singular second-order boundary value problem for the discrete p-Laplacian, $\phi_p(x) = |x|^{p-2}x$, $p > 1$.

In particular, Rachůnková and Rachůnek dealt with the discrete boundary value problem

$$\Delta(\phi_p(\Delta u(t-1))) + f(t, u(t), \Delta u(t-1)) = 0, \quad t \in [1, T+1],$$
$$\Delta u(0) = u(T+2) = 0,$$

in which $f(t, x, y)$ was singular in x.

Combine these works with [Kunkel 2006], which deals with the continuous boundary value problem

$$u''(t) + f(t, u(t)) = 0, \quad t \in (0, 1),$$
$$u'(0) = u(1) = 0,$$

where $f(t, x)$ is singular in x, and you have a similar boundary value problem scenario across time scales ranging from being entirely continuous to varying degrees of discrete. This result fits between these two ends of the time scale continuum being a discrete interval with a continuous point, the ultimate goal of which would be to create a unifying theorem for this type of problem across all types of time scales (forthcoming).

The methods in this paper rely heavily on lower and upper solution methods in conjunction with an application of the Brouwer fixed-point theorem [Zeidler 1986]. We consider only the singular second-order boundary value problem, while letting our function range over an infinitely discrete interval of nonuniform step size, included in which is the limit point. We will provide definitions of appropriate lower and upper solutions. The lower and upper solutions will be applied to nonsingular perturbations of our nonlinear problem, ultimately giving rise to our boundary value problem by passing to the limit.

Lower and upper solutions have been used extensively in establishing solutions of boundary value problems for finite difference equations. Representative works include [Bao et al. 2012; Henderson and Kunkel 2006; Precup 2016].

Singular boundary value problems have also received a good deal of attention. Representative works include [Agarwal and O'Regan 1999; Precup 2016; Rachůnková and Rachůnek 2009].

2. Preliminaries

We now state some definitions used throughout the remainder of the paper, many of which can be found in [Bohner and Peterson 2001; Kelley and Peterson 1991]. Some definitions are required prior to the introduction of the problem we intend to solve.

Definition 2.1. For $i = 1, 2, 3, \ldots$, let $t_i = 1 - 1/i$. Define the time scale

$$\mathbb{T} = \{t_i\}_{i=1}^{\infty} \cup \{1\}.$$

We conveniently make note of the standard notation for both forward and backward jump operators on time scales of this nature.

Definition 2.2. The forward step operator $\sigma : \mathbb{T} \to \mathbb{T}$ is defined by

$$\sigma(t) := \inf\{s \in \mathbb{T} : s > t\}.$$

The backward step operator $\rho : \mathbb{T} \to \mathbb{T}$ is defined by

$$\rho(t) := \sup\{s \in \mathbb{T} : s < t\}.$$

Definition 2.3. For the function $u : \mathbb{T} \to \mathbb{R}$, define the delta derivative $u^{\Delta} : \mathbb{T} \to \mathbb{R}$ by

$$u^{\Delta}(t) := \frac{u(\sigma(t)) - u(t)}{\mu(t)},$$

where $\mu(t) := \sigma(t) - t$. Note that μ is the graininess function.

Having introduced these definitions, we can now consider the following second-order dynamic equation, which will be our focus throughout this paper:

$$u^{\Delta\Delta}(t_{i-1}) + f(t_i, u(t_i), u^{\Delta}(t_{i-1})) = 0, \quad t_{i-1} \in \mathbb{T}, \tag{1}$$

satisfying the mixed boundary conditions,

$$u^{\Delta}(0) = u(1) = 0. \tag{2}$$

Our goal is to prove the existence of a positive solution to this problem (1), (2), where f has a specific type of singularity as explained below.

Definition 2.4. Define a solution to problem (1), (2) to mean a function $u : \mathbb{T} \to \mathbb{R}$ such that u satisfies (1) on $\mathbb{T}$ and also satisfies the boundary conditions (2). If $u(t) > 0$ for $t \in \mathbb{T}$, except possibly at the boundary conditions, we call u a positive solution to problem (1), (2).

Definition 2.5. Let $D \subseteq \mathbb{R}^2$. We say f is continuous on $\mathbb{T} \times D$ if $f(\cdot, x, y)$ is defined on $\mathbb{T}$ for each $(x, y) \in D$ and if $f(t, \cdot, \cdot)$ is continuous on D for each $t \in \mathbb{T}$.

Definition 2.6. Let $D \subseteq \mathbb{R}^2$. Let $f : \mathbb{T} \times D \to \mathbb{R}$. If $D = \mathbb{R}^2$, then we call (1), (2) a regular problem. If $D \subsetneq \mathbb{R}^2$ and f has singularities on the boundary of D, then we call (1), (2) a singular problem.

We assume the following throughout this paper:

(A) $D = [0, \infty) \times \mathbb{R}$.

(B) f is continuous on $\mathbb{T} \times D$.

(C) $f(t, x, y)$ has a singularity at $x = 0$; i.e., $\limsup_{x \to 0^+} |f(t, x, y)| = \infty$ for $t \in \mathbb{T}$ and $y \in \mathbb{R}$.

3. Lower and upper solutions method

For the purpose of establishing a lower and upper solutions method to be used in solving our pre-existing singular problem, we first consider the following regular problem:

$$u^{\Delta\Delta}(t_{i-1}) + h(t_i, u(t_i), u^{\Delta}(t_{i-1})) = 0, \quad t_{i-1} \in \mathbb{T}, \tag{3}$$

where h is continuous on $\mathbb{T} \times \mathbb{R}^2$ and the same boundary conditions (2) are satisfied. Now, (3), (2) is clearly a regular problem and it is our current goal to establish a lower and upper solutions method as a means to establish an existence result. To this end, we first must define what is meant by a lower and an upper solution.

Definition 3.1. Let $\alpha : \mathbb{T} \to \mathbb{R}$. We call α a lower solution of problem (3), (2) if

$$\alpha^{\Delta\Delta}(t_{i-1}) + h(t_i, \alpha(t_i), \alpha^{\Delta}(t_{i-1})) \geq 0, \quad t_{i-1} \in \mathbb{T}, \tag{4}$$

satisfying

$$\alpha^{\Delta}(0) \geq 0, \quad \alpha(1) \leq 0. \tag{5}$$

Definition 3.2. Let $\beta : \mathbb{T} \to \mathbb{R}$. We call β an upper solution of problem (3), (2) if

$$\beta^{\Delta\Delta}(t_{i-1}) + h(t_i, \beta(t_i), \beta^{\Delta}(t_{i-1})) \leq 0, \quad t_{i-1} \in \mathbb{T}, \tag{6}$$

satisfying

$$\beta^{\Delta}(0) \leq 0, \quad \beta(1) \geq 0. \tag{7}$$

Theorem 3.3 (lower and upper solutions method). *Let α and β be lower and upper solutions of the regular problem* (3), (2), *respectively, where $\alpha \leq \beta$ on $\mathbb{T}$. Let $h(t, x, y)$ be continuous on $\mathbb{T} \times \mathbb{R}^2$ and nonincreasing in its y-variable. Then* (3), (2) *has a solution u satisfying*

$$\alpha(t) \leq u(t) \leq \beta(t), \quad t \in \mathbb{T}.$$

Proof. We proceed with this proof through a sequence of steps involving modifications of the function h.

<u>Step 1</u>: For $t_{i-1} \in \mathbb{T}$ and $(x, y) \in \mathbb{R}^2$, define

$$\tilde{h}\left(t_i, x, \frac{x-y}{\mu(t_{i-1})}\right)$$

$$= \begin{cases} h\left(t_i, \beta(t_i), \dfrac{\beta(t_i) - S(t_{i-1}, y)}{\mu(t_{i-1})}\right) - \dfrac{x - \beta(t_i)}{x - \beta(t_i) + 1}, & x > \beta(t_i), \\[2ex] h\left(t_i, x, \dfrac{x - S(t_{i-1}, y)}{\mu(t_{i-1})}\right), & \alpha(t_i) \leq x \leq \beta(t_i), \\[2ex] h\left(t_i, \alpha(t_i), \dfrac{\alpha(t_i) - S(t_{i-1}, y)}{\mu(t_{i-1})}\right) + \dfrac{\alpha(t_i) - x}{\alpha(t_i) - x + 1}, & x < \alpha(t_i), \end{cases} \tag{8}$$

where,

$$S(t_{i-1}, y) = \begin{cases} \beta(t_{i-1}), & y > \beta(t_{i-1}), \\ y, & \alpha(t_{i-1}) \le y \le \beta(t_{i-1}), \\ \alpha(t_{i-1}), & y < \alpha(t_{i-1}). \end{cases}$$

Given this construction, $\tilde{h}$ is continuous on $\mathbb{T} \times \mathbb{R}^2$ and there exists $M > 0$ so that

$$|\tilde{h}(t, x, y)| \le M$$

for all $t \in \mathbb{T}$ and $(x, y) \in \mathbb{R}^2$.

We now study the auxiliary equation

$$u^{\Delta\Delta}(t_{i-1}) + \tilde{h}(t_i, u(t_i), u^{\Delta}(t_{i-1})) = 0, \quad t_{i-1} \in \mathbb{T}, \tag{9}$$

satisfying boundary conditions (2). Our immediate goal is to prove the existence of a solution to problem (9), (2).

Step 2: For this existence result, we lay the foundation to use the Brouwer fixed-point theorem. To this end, define

$$E = \{u : \mathbb{T} \to \mathbb{R} : u^{\Delta}(0) = u(1) = 0\}.$$

Also, define

$$\|u\| = \max\{|u(t)| : t \in \mathbb{T}\}.$$

Given E and $\| \cdot \|$, we say E is a Banach space. Further, we define an operator $\mathcal{T} : E \to E$ by

$$(\mathcal{T}u)(t_k) = \sum_{j=k}^{\infty} \mu(t_j) \sum_{i=2}^{j} \mu(t_{i-1})\tilde{h}(t_i, u(t_i), u^{\Delta}(t_{i-1})). \tag{10}$$

$\mathcal{T}$ is a continuous operator. Moreover, from the bounds placed on $\tilde{h}$ in Step 1 and from (10), if $r > M$, then $T(\overline{B(r)}) \subseteq \overline{B(r)}$, where $B(r) := \{u \in E : \|u\| < r\}$. Hence, by the Brouwer fixed-point theorem [Zeidler 1986], there exists $u \in \overline{B(r)}$ such that $u = \mathcal{T}u$.

Step 3: We now show that u is a fixed point of $\mathcal{T}$ if and only if u is a solution to the problem (9), (2).

To this end, let us first assume that u solves the problem (9), (2). Then, since the boundary conditions (2) are satisfied, $u \in E$.

It is convenient for the first part of this subproof to consider a relabeling of the points in $\mathbb{T}$ as follows: let $\tau_{\infty} = \lim_{i \to \infty} \tau_i := t_1 = 0$, let $\tau_0 := 1$, and, for each $i > 0$, let there exist some $j > 0$ so that $t_i = \tau_j$, $t_{i+1} = \tau_{j-1}$, etc. Using this notation, we then consider

$$u^{\Delta}(\tau_1) = \frac{u(\tau_0) - u(\tau_1)}{\mu(\tau_1)} = \frac{-u(\tau_1)}{\mu(\tau_1)},$$

and we have

$$u(\tau_1) = -\mu(\tau_1)u^\Delta(\tau_1).$$

Also,

$$u^\Delta(\tau_2) = \frac{u(\tau_1) - u(\tau_2)}{\mu(\tau_2)} = \frac{-\mu(\tau_1)u^\Delta(\tau_1) - u(\tau_2)}{\mu(\tau_2)},$$

and we have

$$u(\tau_2) = -\mu(\tau_1)u^\Delta(\tau_1) - \mu(\tau_2)u^\Delta(\tau_2).$$

Continuing in this manner, we have, for $m > 0$,

$$u(\tau_m) = -\sum_{i=1}^m \mu(\tau_i)u^\Delta(\tau_i). \tag{11}$$

And, given our relabeling between the τ's and the t's, we can conclude that for each $m > 0$ there exists some $k > 0$ such that

$$u(\tau_m) = u(t_k) = -\sum_{i=k}^\infty \mu(\tau_i)u^\Delta(\tau_i).$$

We also have

$$u^{\Delta\Delta}(t_1) = \frac{u^\Delta(t_2) - u^\Delta(t_1)}{\mu(t_1)} = \frac{u^\Delta(t_2) - u^\Delta(0)}{\mu(t_1)} = \frac{u^\Delta(t_2)}{\mu(t_1)},$$

and from (9) we have $u^{\Delta\Delta}(t_1) = -\tilde{h}(t_2, u(t_2), u^\Delta(t_1))$, which yields

$$u^\Delta(t_2) = -\mu(t_1)\tilde{h}(t_2, u(t_2), u^\Delta(t_1)).$$

Similarly, we have

$$u^{\Delta\Delta}(t_2) = \frac{u^\Delta(t_3) - u^\Delta(t_2)}{\mu(t_2)} = -\tilde{h}(t_3, u(t_3), u^\Delta(t_2)),$$

and via substitution of $u^\Delta(t_2)$ and simply solving for $u^\Delta(t_3)$, we have

$$u^\Delta(t_3) = -\mu(t_1)\tilde{h}(t_2, u(t_2), u^\Delta(t_1)) - \mu(t_2)\tilde{h}(t_3, u(t_3), u^\Delta(t_2)).$$

Continuing in this manner, we conclude that

$$u^\Delta(t_j) = -\sum_{i=2}^j \mu(t_{i-1})\tilde{h}(t_i, u(t_i), u^\Delta(t_{i-1})). \tag{12}$$

By substituting (12) into (11), we see that for $k > 0$

$$u(t_k) = -\sum_{j=k}^\infty \mu(t_j)\left(-\sum_{i=2}^j \mu(t_{i-1})\tilde{h}(t_i, u(t_i), u^\Delta(t_{i-1}))\right)$$

$$= \sum_{j=k}^\infty \mu(t_j)\sum_{i=2}^j \mu(t_{i-1})\tilde{h}(t_i, u(t_i), u^\Delta(t_{i-1})) = (\mathcal{T}u)(t_k).$$

We now assume that u is a fixed point of $\mathcal{T}$, i.e., $u = \mathcal{T}u$. Then,

$$u(t_k) = (\mathcal{T}u)(t_k) = \sum_{j=k}^{\infty} \mu(t_j) \sum_{i=2}^{j} \mu(t_{i-1})\tilde{h}(t_i, u(t_i), u^{\Delta}(t_{i-1})).$$

Also,

$$u^{\Delta}(t_{k-1}) = \frac{u(t_k) - u(t_{k-1})}{\mu(t_{k-1})}$$

$$= \frac{\left(\sum_{j=k}^{\infty} \mu(t_j) \sum_{i=2}^{j} \mu(t_{i-1})\tilde{h}(t_i, u(t_i), u^{\Delta}(t_{i-1}))\right)}{\mu(t_{k-1})}$$

$$- \frac{\left(\sum_{j=k-1}^{\infty} \mu(t_j) \sum_{i=2}^{j} \mu(t_{i-1})\tilde{h}(t_i, u(t_i), u^{\Delta}(t_{i-1}))\right)}{\mu(t_{k-1})}$$

$$= -\frac{\mu(t_{k-1}) \sum_{i=2}^{k-1} \mu(t_{i-1})\tilde{h}(t_i, u(t_i), u^{\Delta}(t_{i-1}))}{\mu(t_{k-1})}$$

$$= -\sum_{i=2}^{k-1} \mu(t_{i-1})\tilde{h}(t_i, u(t_i), u^{\Delta}(t_{i-1})),$$

and

$$u^{\Delta\Delta}(t_{k-1})$$

$$= \frac{u^{\Delta}(t_k) - u^{\Delta}(t_{k-1})}{\mu(t_{k-1})}$$

$$= \frac{\left(-\sum_{i=2}^{k} \mu(t_{i-1})\tilde{h}(t_i, u(t_i), u^{\Delta}(t_{i-1}))\right)}{\mu(t_{k-1})} - \frac{\left(-\sum_{i=2}^{k-1} \mu(t_{i-1})\tilde{h}(t_i, u(t_i), u^{\Delta}(t_{i-1}))\right)}{\mu(t_{k-1})}$$

$$= -\frac{\mu(t_{k-1})\tilde{h}(t_k, u(t_k), u^{\Delta}(t_{k-1}))}{\mu(t_{k-1})}$$

$$= -\tilde{h}(t_k, u(t_k), u^{\Delta}(t_{k-1})).$$

Thus, u solves (9).

We need now only consider the boundary conditions (2) in order to complete Step 3 of this proof. To this end, we recall the construction of the time scale $\mathbb{T}$ and notice the following based on what was just derived as the formula for u^{Δ}:

$$u^{\Delta}(0) = u^{\Delta}(t_1) = -\sum_{i=2}^{1} \mu(t_{i-1})\tilde{h}(t_i, u(t_i), u^{\Delta}(t_{i-1})) = 0.$$

We now turn our attention over to $t = 1$ and recall from the construction of $\mathbb{T}$ that $t_{\infty} = 1$. Also note that standard convention when discussing time scales of this sort is $\sigma(t) = t$ if $\mathbb{T}$ has a maximum t, or for our purposes $\mu(t) = 0$ if $\mathbb{T}$ has a

maximum t. As such,

$$u(1) = u(t_\infty) = \sum_{j=\infty}^{\infty} \mu(t_j) \sum_{i=2}^{j} \mu(t_{i-1}) \tilde{h}(t_i, u(t_i), u^\Delta(t_{i-1}))$$

$$= \mu(t_\infty) \sum_{i=2}^{\infty} \mu(t_{i-1}) \tilde{h}(t_i, u(t_i), u^\Delta(t_{i-1}))$$

$$= \mu(1) \sum_{i=2}^{\infty} \mu(t_{i-1}) \tilde{h}(t_i, u(t_i), u^\Delta(t_{i-1}))$$

$$= 0 \cdot \sum_{i=2}^{\infty} \mu(t_{i-1}) \tilde{h}(t_i, u(t_i), u^\Delta(t_{i-1})) = 0.$$

Therefore, we get that u solves (9), (2) and Step 3 is complete.

Step 4: The remaining piece we need to show is that solutions of (9), (2) satisfy

$$\alpha(t) \le u(t) \le \beta(t), \quad t \in \mathbb{T}.$$

To this end, without loss of generality consider the case of obtaining $u(t) \le \beta(t)$, and let $v(t) = u(t) - \beta(t)$. For the purpose of establishing a contradiction, assume that $\max\{v(t) : t \in \mathbb{T}\} := v(l) > 0$. From (2) and (7), we see that l must be an interior point in $\mathbb{T}$; i.e., $l := t_j \in \mathbb{T} \setminus \{0, 1\}$. With t_j necessarily being an interior point, t_{j-1} and t_{j+1} are well-defined, and we have

$$v(t_{j-1}) \le v(t_j) \quad \text{and} \quad v(t_{j+1}) \le v(t_j).$$

Consequently,

$$v^\Delta(t_{j-1}) \ge 0 \quad \text{and} \quad v^\Delta(t_j) \le 0.$$

Further, we now know also that

$$v^{\Delta\Delta}(t_{j-1}) = \frac{v^\Delta(t_j) - v^\Delta(t_{j-1})}{\mu(t_{j-1})} \le 0.$$

Therefore,

$$u^{\Delta\Delta}(t_{j-1}) - \beta^{\Delta\Delta}(t_{j-1}) \le 0. \tag{13}$$

On the other hand, since h is nonincreasing in its third variable, we have from (9) and (8) that

$$u^{\Delta\Delta}(t_{j-1}) - \beta^{\Delta\Delta}(t_{j-1}) = -\tilde{h}(t_j, u(t_j), u^\Delta(t_{j-1})) - \beta^{\Delta\Delta}(t_{j-1})$$

$$= -\left(\tilde{h}(t_j, \beta(t_j), \beta^\Delta(t_{j-1})) - \frac{u(t_j) - \beta(t_j)}{u(t_j) - \beta(t_j) + 1} \right) - \beta^{\Delta\Delta}(t_{j-1})$$

$$= -\tilde{h}(t_j, \beta(t_j), \beta^\Delta(t_{j-1})) + \frac{v(t_j)}{v(t_j) + 1} - \beta^{\Delta\Delta}(t_{j-1})$$

$$\ge \beta^{\Delta\Delta}(t_{j-1}) + \frac{v(l)}{v(l) + 1} - \beta^{\Delta\Delta}(t_{j-1}) = \frac{v(l)}{v(l) + 1} > 0.$$

Hence we have a contradiction to (13) and we conclude that $\max\{v(t) : t \in \mathbb{T}\} \leq 0$. Thus, $v(t) \leq 0$ for all $t \in \mathbb{T}$, or rather

$$u(t) \leq \beta(t) \quad \text{for all } t \in \mathbb{T}.$$

A similar argument shows that $\alpha(t) \leq u(t)$ for all $t \in \mathbb{T}$.

Thus, our conclusion holds and the proof is complete. $\qquad\square$

4. Main result

In this section, we make use of Theorem 3.3 to obtain positive solutions to the singular problem (1), (2). In particular, in applying Theorem 3.3, we deal with a sequence of regular perturbations of (1), (2). Ultimately, we obtain a desired solution by passing to the limit on a sequence of solutions for the perturbations.

Theorem 4.1. *Assume conditions* (A), (B), *and* (C) *hold, along with the following*:

(D) *There exists $c \in (0, \infty)$ so that $f(t, c, y) \leq 0$ for all $t \in \mathbb{T}$ and $y \in \mathbb{R}$.*

(E) *$f(t, x, y)$ is nonincreasing in its y-variable for all $t \in \mathbb{T}$ and $x \in (0, c)$.*

(F) *$\lim_{x \to 0^+} f(t, x, y) = \infty$ for $t \in \mathbb{T}$ and $y \in (-c, c)$.*

Then, (1), (2) has a solution u satisfying

$$0 < u(t) \leq c, \quad t \in \mathbb{T} \setminus \{1\}.$$

Proof. We proceed through this proof via a sequence of steps.

<u>Step 1</u>: For $k > 0$, $t \in \mathbb{T}$, and $y \in \mathbb{R}$, define

$$f_k(t, x, y) = \begin{cases} f(t, |x|, y) & \text{if } |x| \geq 1/k, \\ f(t, 1/k, y) & \text{if } |x| < 1/k. \end{cases}$$

Then, f_k is continuous on $\mathbb{T} \times \mathbb{R}^2$.

Assumption (F) implies that there exists k_0 such that, for all $k \geq k_0$,

$$f_k(t, 0, y) = f\left(t, \frac{1}{k}, y\right) > 0 \quad \text{for all } t \in \mathbb{T}, y \in \mathbb{R}.$$

We now consider

$$u^{\Delta\Delta}(t_{i-1}) + f_k(t_i, u(t_i), u^\Delta(t_{i-1})) = 0, \quad t \in \mathbb{T}. \tag{14}$$

Now, let $\alpha(t) = 0$ and $\beta(t) = c$. Then, for each $k \geq k_0$, α and β are lower and upper solutions of (14), (2), respectively. Also, $\alpha(t) \leq \beta(t)$ for $t \in \mathbb{T}$. Thus, by Theorem 3.3, for each $k \geq k_0$, there exists a solution u_k to each problem (14), (2) that satisfies $0 \leq u_k(t) \leq c$ for $t \in \mathbb{T}$.

Consequently, for all $t_i \in \mathbb{T}$,

$$|u^\Delta(t_i)| \leq c \cdot \mu(t_{i-1}). \tag{15}$$

<u>Step 2</u>: For $k \geq k_0$, let $\delta \in (0, 1)$ and consider the time scale $\mathbb{T}_1 := \mathbb{T} \cap [0, \delta]$. Since u_k solves (14), we get from work similar to that in the proof of Theorem 3.3 that

$$u_k^{\Delta}(t_j) = -\sum_{i=2}^{j} \mu(t_{i-1}) f_k(t_i, u(t_i), u^{\Delta}(t_{i-1})). \qquad (16)$$

We use this version of u_k^{Δ} as follows:

By assumption (F), there exists $\varepsilon_1 \in (0, 1/k_0)$ such that for all $k \geq 1/\varepsilon_1$

$$f_k(t_2, x, y) > c, \quad x \in (0, \varepsilon_1), \ y \in (-c, c). \qquad (17)$$

For the sake of establishing a contradiction, assume that for $k \geq 1/\varepsilon_1$ we have $u_k(t_2) < \varepsilon_1$. Then, by (16) and (17),

$$u_k^{\Delta}(t_2) = -\sum_{i=2}^{2} \mu(t_{i-1}) f_k(t_i, u(t_i), u^{\Delta}(t_{i-1}))$$

$$= -\mu(t_1) f_k(t_2, u(t_2), u^{\Delta}(t_1)) < -\mu(t_1) \cdot c.$$

However, this contradicts (15). Thus, $u_k(t_2) \geq \varepsilon_1$ for all $k \geq 1/\varepsilon_1$.

Continuing, also by assumption (F), there now exists $\varepsilon_2 \in (0, \varepsilon_1)$ such that for all $k \geq 1/\varepsilon_2$

$$f_k(t_3, x, y) > c + m_1, \quad x \in (0, \varepsilon_2), \ y \in (-c, c), \qquad (18)$$

where $m_1 = \max\{|f_k(t_2, x, y)| : x \in [\varepsilon_1, c], y \in (-c, c)\}$. For the sake of establishing a contradiction, assume that for $k \geq 1/\varepsilon_2$ we have $u_k(t_3) < \varepsilon_2$. Then, by (16) and (18),

$$u_k^{\Delta}(t_3) = -\sum_{i=2}^{3} \mu(t_{i-1}) f_k(t_i, u(t_i), u^{\Delta}(t_{i-1}))$$

$$= -\mu(t_1) f_k(t_2, u(t_2), u^{\Delta}(t_1)) - \mu(t_2) f_k(t_3, u(t_3), u^{\Delta}(t_2))$$

$$\leq \mu(t_1) \cdot m_1 - \mu(t_2)(c + m_1) < \mu(t_2) \cdot c.$$

However, this contradicts (15). Thus, $u_k(t_3) \geq \varepsilon_2$ for all $k \geq 1/\varepsilon_2$.

We continue in this manner, proceeding across the interval $\mathbb{T}_1$ for $j = 3, 4, 5,$ $\dots, l-1$ and we create a nested sequence of epsilons, $0 < \varepsilon_{l-1} < \cdots < \varepsilon_2 < \varepsilon_1$, where $u_k(t_j) \geq \varepsilon_{j-1}$ when $k \geq 1/\varepsilon_{j-1}$.

Continuing, by assumption (F), there exists $\varepsilon_l \in (0, \varepsilon_{l-1})$ such that for all $k \geq 1/\varepsilon_l$

$$f_k(t_{l+1}, x, y) > c + \sum_{i=1}^{l-1} m_i, \quad x \in [\varepsilon_i, c], \ y \in (-c, c), \qquad (19)$$

where $m_i = \max\{|f_k(t_{i+1}, x, y)| : x \in [\varepsilon_i, c], y \in (-c, c)\}$. For the sake of establishing a contradiction, assume that for $k \geq 1/\varepsilon_l$ we have $u_k(t_{l+1}) < \varepsilon_l$. Then, by

(16) and (19),

$$u_k^\Delta(t_{l+1}) = -\sum_{i=2}^{l+1} \mu(t_{i-1}) f_k(t_i, u(t_i), u^\Delta(t_{i-1}))$$

$$= -\mu(t_1) f_k(t_2, u(t_2), u^\Delta(t_1)) - \mu(t_2) f_k(t_3, u(t_3), u^\Delta(t_2))$$

$$- \cdots - \mu(t_l) f_k(t_{l+1}, u(t_{l+1}), u^\Delta(t_l))$$

$$\leq \mu(t_1) m_1 - \cdots - \mu(t_{l-1}) m_{l-1} - \mu(t_l) \left(c + \sum_{i=1}^{l-1} m_i \right)$$

$$< \mu(t_l) \cdot c.$$

However, this contradicts (15). Thus, $u_k(t_{l+1}) \geq \varepsilon_l$ for all $k \geq 1/\varepsilon_l$.

Now, recall that we are on the interval $\mathbb{T}_1$, which just so happens to be an interval with a finite number of points included in its time scale. Call the largest of these points t_M, and based on the previous argument, there exists $\varepsilon_{m-1} > 0$ so that $u_k(t_M) \geq \varepsilon_{m-1}$ for $k \geq 1/\varepsilon_{m-1}$. Choose $\varepsilon = \varepsilon_{m-1}/2$ and note that

$$0 < \varepsilon \leq u_k(t) \leq c \quad \text{for all } t \in \mathbb{T}_1, \ k \geq \frac{1}{\varepsilon}.$$

We now need only discuss what happens when

$$t \in \mathbb{T}_2 := \mathbb{T} \setminus \mathbb{T}_1 = \mathbb{T} \cap [\delta, 1].$$

Note that for each δ as $\delta \to 1$, via previous arguments, we have for sufficiently large k that $u_k(t) > 0$, $t \in \mathbb{T}_2$. Also note that for sufficiently large k, as $\delta \to 1$, we have $u_k(t) \geq 0$. This leads to the fact that for sufficiently large k, we get $u_k(t) > 0$ for $t \in \mathbb{T} \setminus \{1\}$ and, as our boundary condition states, $u_k(1) = 0$.

We now choose a subsequence $\{u_{k_n}(t)\} \subseteq \{u_k(t)\}$ so that

$$\lim_{n \to \infty} u_{k_n}(t) = u(t), \quad t \in \mathbb{T},$$

and note that $u(t) \in E$, where E is defined as in the proof of Theorem 3.3.

Moreover, (16) yields, for sufficiently large n,

$$u_{k_n}^\Delta(t_j) = -\sum_{i=2}^{j} \mu(t_{i-1}) f(t_i, u_{k_n}(t_i), u_{k_n}^\Delta(t_{i-1})),$$

and so letting $n \to \infty$ and from the continuity of f we get

$$u^\Delta(t_j) = -\sum_{i=2}^{j} \mu(t_{i-1}) f(t_i, u(t_i), u^\Delta(t_{i-1})).$$

Consequently,

$$u^{\Delta\Delta}(t_{i-1}) = -f(t_i, u(t_i), u^\Delta(t_{i-1})). \qquad \square$$

5. Example

Let $\mathbb{T}$ be as given in Definition 2.1. Let $\alpha \in [0, \infty)$, $c, \beta \in (0, \infty)$, and $a : \mathbb{T} \to \mathbb{R}$. Then, by Theorem 4.1, the problem

$$u^{\Delta\Delta}(t_{i-1}) + \big(a(t_i) + (u(t_i))^{\alpha} + (u(t_i))^{-\beta}\big)(c - u(t_i)) - (u^{\Delta}(t_{i-1}))^3 = 0, \quad t_{i-1} \in \mathbb{T},$$

along with the boundary conditions (2), has a solution u satisfying the desired inequality.

References

[Agarwal and O'Regan 1999] R. P. Agarwal and D. O'Regan, "Singular discrete boundary value problems", *Appl. Math. Lett.* **12**:4 (1999), 127–131. MR Zbl

[Bao et al. 2012] G. Bao, X. Xu, and Y. Song, "Positive solutions for three-point boundary value problems with a non-well-ordered upper and lower solution condition", *Appl. Math. Lett.* **25**:4 (2012), 767–770. MR Zbl

[Bohner and Peterson 2001] M. Bohner and A. Peterson, *Dynamic equations on time scales*, Birkhäuser, Boston, 2001. MR Zbl

[Henderson and Kunkel 2006] J. Henderson and C. J. Kunkel, "Singular discrete higher order boundary value problems", *Int. J. Difference Equ.* **1**:1 (2006), 119–133. MR Zbl

[Kelley and Peterson 1991] W. G. Kelley and A. C. Peterson, *Difference equations: an introduction with applications*, Academic Press, Boston, 1991. MR Zbl

[Kunkel 2006] C. J. Kunkel, "Singular second order boundary value problems for differential equations", pp. 119–124 in *Proceedings of neural, parallel, and scientific computations, III* (Atlanta, 2006), edited by G. S. Ladde et al., Dynamic, Atlanta, 2006. MR Zbl

[Kunkel 2008] C. J. Kunkel, "Singular second order boundary value problems on purely discrete time scales", *J. Difference Equ. Appl.* **14**:4 (2008), 411–420. MR Zbl

[Precup 2016] R. Precup, "Abstract method of upper and lower solutions and application to singular boundary value problems", *Stud. Univ. Babeş-Bolyai Math.* **61**:4 (2016), 443–451. MR Zbl

[Rachůnková and Rachůnek 2006] I. Rachůnková and L. Rachůnek, "Singular discrete second order BVPs with p-Laplacian", *J. Difference Equ. Appl.* **12**:8 (2006), 811–819. MR Zbl

[Rachůnková and Rachůnek 2009] I. Rachůnková and L. Rachůnek, "Singular discrete and continuous mixed boundary value problems", *Math. Comput. Modelling* **49**:3-4 (2009), 413–422. MR Zbl

[Zeidler 1986] E. Zeidler, *Nonlinear functional analysis and its applications, I: Fixed-point theorems*, Springer, 1986. MR Zbl

ckunkel@utm.edu   *Department of Mathematics and Statistics,*
                                 University of Tennessee at Martin, Martin, TN, United States

alejlanc@ut.utm.edu   *Department of Mathematics and Statistics,*
                                 University of Tennessee at Martin, Martin, TN, United States

Guidelines for Authors

Submissions in all mathematical areas are encouraged. All manuscripts accepted for publication in *Involve* are considered publishable in quality journals in their respective fields, and include a minimum of one-third student authorship. Submissions should include substantial faculty input; faculty co-authorship is strongly encouraged. Authors may submit manuscripts in PDF format on-line at the Submission page at the Involve website.

Originality. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles in *Involve* are usually in English, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

Format. Authors are encouraged to use LaTeX but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@msp.org with details about how your graphics were generated.

White space. Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Proofs. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

involve