involve

a journal of mathematics

mathematical sciences publishers

2011 vol. 4, no. 1

involve

pjm.math.berkeley.edu/involve

See inside back cover or http://pjm.math.berkeley.edu/involve for submission instructions.

The subscription price for 2011 is US $100/year for the electronic version, and $130/year (+$35 shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to Mathematical Sciences Publishers, Department of Mathematics, University of California, Berkeley, CA 94704-3840, USA.

Involve (ISSN 1944-4184 electronic, 1944-4176 printed) at Mathematical Sciences Publishers, Department of Mathematics, University of California, Berkeley, CA 94720-3840 is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

Involve peer review and production are managed by EditFLOW™ from Mathematical Sciences Publishers.

PUBLISHED BY

mathematical sciences publishers

http://msp.org/

A NON-PROFIT CORPORATION

Typeset in LaTeX

INVOLVE 4:1(2011)

msp

The arithmetic of trees

Adriano Bruno and Dan Yasaki

(Communicated by Robert W. Robinson)

The arithmetic of the natural numbers $\mathbb{N}$ can be extended to arithmetic operations on planar binary trees. This gives rise to a noncommutative arithmetic theory. In this exposition, we describe this *arithmetree*, first defined by Loday, and investigate prime trees.

1. Introduction

J.-L. Loday [2002] published a paper *Arithmetree*, in which he defines arithmetic operations on the set $\mathbb{Y}$ of groves of planar binary trees. These operations extend the usual addition and multiplication on the natural numbers $\mathbb{N}$ in the sense that there is an embedding $\mathbb{N} \hookrightarrow \mathbb{Y}$, and the multiplication and addition he defines become the usual ones when restricted to $\mathbb{N}$. Loday's reasons for introducing these notions have to do with intricate algebraic structures known as dendriform algebras [Loday et al. 2001].

Since the arithmetic extends the usual operations on $\mathbb{N}$, one can ask many of the same questions that arise in the natural numbers. In this exposition, we examine notions of primality, specifically studying *prime trees*. We will see that all trees of prime degree must be prime, but many trees of composite degree are also prime. One should not be misled by the idea that arithmetree is an extension of the usual arithmetic on $\mathbb{N}$. Indeed, away from the image of $\mathbb{N}$ in $\mathbb{Y}$, the arithmetic operations $+$ and $\times$ are noncommutative. Both operations are associative, but multiplication is only distributive on the left with respect to $+$. In the end it is somewhat surprising that there is a very natural copy of $\mathbb{N}$ inside $\mathbb{Y}$.

The paper is organized as follows. Sections 2–6 summarize without proofs the results that we need from [Loday 2002]. Specifically, basic definitions are given in Section 2 to set notation. The embedding $\mathbb{N} \hookrightarrow \mathbb{Y}$ is given in Section 3, and Section 4 discusses the basic operations on groves. Sections 5 and 6 define the

MSC2000: primary 05C05; secondary 03H15.

Keywords: arithmetree, planar binary trees.

These results grew out of an REU project in the summer of 2007 at the University of Massachusetts at Amherst; the authors are grateful for this support.

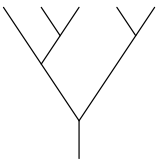
arithmetic on $\mathbb{Y}$. Finally, Section 8 discusses some new results and Section 9 gives a few final remarks.

2. Background

In this section, we give the basic definitions and set notation.

Definition 2.1. A *planar binary tree* is an oriented planar graph drawn in the plane with one root, $n+1$ leaves, and n interior vertices, all of which are trivalent.

Henceforth, by *tree*, we will mean a planar binary tree. We consider trees to be the same if they can be moved in the plane to each other. Thus we can always represent a tree by drawing a root and then having it "grow" upward. The *degree* is the number of internal vertices. Here is an example of a tree of degree four, with five leaves:



Let Y_n be the set of trees of degree n. For example,

$$Y_0 = \{\,|\,\}, \quad Y_1 = \{\curlyvee\}, \quad Y_2 = \{\curlyvee\curlyvee, \curlyvee\curlyvee\}, \quad Y_3 = \{\curlyvee\curlyvee\curlyvee, \curlyvee\curlyvee\curlyvee, \curlyvee\curlyvee\curlyvee, \curlyvee\curlyvee\curlyvee, \curlyvee\curlyvee\curlyvee\}.$$

One can show that the cardinality of Y_n is given by the n-th *Catalan* number,

$$c_n = \frac{1}{n+1}\binom{2n}{n} = \frac{(2n)!}{(n+1)!n!}.$$

The Catalan numbers arise in a variety of combinatorial problems [Stanley 2007].[1]

Definition 2.2. A nonempty subset of Y_n is called a *grove*. The set of all groves of degree n is denoted by $\mathbb{Y}_n$.

For example,

$$\mathbb{Y}_0 = \{\,|\,\}, \quad \mathbb{Y}_1 = \{\curlyvee\}, \quad \mathbb{Y}_2 = \{\curlyvee\curlyvee, \curlyvee\curlyvee, \curlyvee\curlyvee \cup \curlyvee\curlyvee\}.$$

Notice that we are omitting the braces around the sets in $\mathbb{Y}_n$ and use instead $\cup$ to denote the subsets. For example we write $\curlyvee\curlyvee \cup \curlyvee\curlyvee$ as opposed to $\{\curlyvee\curlyvee, \curlyvee\curlyvee\}$ to denote the grove in $\mathbb{Y}_2$ consisting of both trees of degree 2. Let $\mathbb{Y} = \bigcup_{n\in\mathbb{N}} \mathbb{Y}_n$ denote the set of all groves. By definition groves consist of trees of the same degree; hence we get a well-defined notion of degree

$$\deg : \mathbb{Y} \to \mathbb{N}. \tag{1}$$

[1] He currently gives 161 combinatorial interpretations of c_n.

n	$\#Y_n$	$\#\mathbb{Y}_n$
1	1	1
2	2	3
3	5	31
4	14	16383
5	42	4398046511103
6	132	54445178707350154154139937189082913833295
7	429	$\sim 1.386 \times 10^{129}$

Table 1. Number of trees and groves of degree $n \leq 7$.

The Catalan numbers c_n grow rapidly. Since $\mathbb{Y}_n$ is the set of subsets of Y_n, we see that the cardinality $\#\mathbb{Y}_n = 2^{c_n} - 1$ grows extremely fast (Table 1), necessitating the use of computers even for computations on trees of fairly small degree.

3. The natural numbers

In this section we give an embedding of $\mathbb{N}$ into $\mathbb{Y}$. There is a distinguished grove for each degree given by set of all trees of degree n.

Definition 3.1. The *total grove of degree n* is defined by $\underline{n} = \bigcup_{x \in Y_n} x$.

For example,

$$\underline{0} = \mid, \quad \underline{1} = \curlyvee, \quad \underline{2} = \curlyvee \cup \curlyvee, \quad \underline{3} = \curlyvee \cup \curlyvee \cup \curlyvee \cup \curlyvee \cup \curlyvee.$$

This gives an embedding $\mathbb{N} \hookrightarrow \mathbb{Y}$. It is clear that the degree map is a one-sided inverse in the sense that $\deg(\underline{n}) = n$ for all $n \in \mathbb{N}$. We will see in Section 7 that under this embedding, *arithmetree* can be viewed as an extension of arithmetic on $\mathbb{N}$.

4. Basic operations

In this section we define a few operations that will be used to define the arithmetic on $\mathbb{Y}$.

4.1. *Grafting.*

Definition 4.1. We say that a tree z is obtain as the *graft* of x and y (notation: $z = x \curlyvee y$) if z is gotten by attaching the root of x to the left leaf and the root of y to the right leaf of $\curlyvee$.

For example, $\curlyvee = \curlyvee \curlyvee \mid$ and $\curlyvee = \curlyvee \curlyvee \curlyvee$. It is clear that every tree x of degree greater than 1 can be obtained as the graft of trees x^l and x^r of degree less than n.

Specifically, we have that $x = x^l \vee x^r$. We refer to these subtrees as the *left* and *right parts* of x.

Given a tree x of degree n, then one can create a tree of degree $n+1$ that carries much of the structure of x by grafting on $\underline{0} = \,|\,$. Indeed, there are two such trees, $x \vee \underline{0}$ and $\underline{0} \vee x$. We will say that such trees are *inherited*.

Definition 4.2. A tree x is said to be *left-inherited* if $x^r = \underline{0}$ and *right-inherited* if $x^l = \underline{0}$. A grove is *left-inherited* (resp. *right-inherited*) if each of its member trees is *left-inherited* (resp. *right-inherited*).

We single out two special sequences of trees L_n and R_n.

Definition 4.3. Let $L_1 = R_1 = \underline{1}$. For $n > 1$, set $L_n = L_{n-1} \vee \underline{0}$ and $R_n = \underline{0} \vee R_{n-1}$. We will call such trees *primitive*.

Notice that L_n is the left-inherited tree such that $L_n^l = L_{n-1}$. Similarly, R_n is the right-inherited tree such that $R_n^r = R_{n-1}$.

4.2. *Over and under.*

Definition 4.4. For $x \in Y_p$ and $y \in Y_q$ the tree x/y (read x *over* y) in Y_{p+q} is obtained by identifying the root of x with the leftmost leaf of y. Similarly, the tree $x\backslash y$ (read x *under* y) in Y_{p+q} is obtained by identifying the rightmost leaf of x with the root of y.

For example, $\curlyvee/\vee = \curlyvee$ and $\curlyvee\backslash\vee = \curlyvee$.

4.3. ***Involution.*** The symmetry around the axis passing through the root defines an involution σ on Y. For example, $\sigma(\curlyvee) = \curlyvee$ and $\sigma(\curlyvee) = \curlyvee$. The involution can be extended to an involution on $\mathbb{Y}$, by letting σ act on each tree in the grove. One can easily check that for trees x, y:

(i) $\sigma(x \vee y) = \sigma(y) \vee \sigma(x)$,

(ii) $\sigma(x/y) = \sigma(y)\backslash\sigma(x)$,

(iii) $\sigma(x\backslash y) = \sigma(y)/\sigma(x)$.

We will see that this involution also respects the arithmetic of groves.

5. Addition

Before we define addition, we first put a partial ordering on Y_n.

5.1. ***Partial ordering.*** We say that the inequality $x < y$ holds if y is obtained from x by moving edges of x from left to right over a vertex. This induces a partial ordering on Y_n by imposing:

(i) $(x \vee y) \vee z \leq x \vee (y \vee z)$.

(ii) If $x < y$ then $x \vee z < y \vee z$ and $z \vee x < z \vee y$ for all $z \in Y_n$.

For example, $\curlyvee < \curlyvee\curlyvee < \curlyvee\curlyvee\curlyvee < \curlyvee\curlyvee\curlyvee\curlyvee$. Note that the primitive trees are extremal elements with respect to this ordering.

5.2. *Sum.*

Definition 5.1. The *sum* of two trees x and y is the following disjoint union of trees

$$x + y := \bigcup_{x/y \le z \le x\backslash y} z\ .$$

All the elements in the sum have the same degree, namely $\deg(x) + \deg(y)$. The definition of addition extends to groves by distributing. Namely, for groves $x = \bigcup_i x_i$ and $y = \bigcup_j y_i$,

$$x + y := \bigcup_{ij} (x_i + y_j). \tag{2}$$

We remark that it is not immediate that the result of the sum is a grove since it is not obvious that the trees arising in the union are all distinct. Loday shows that this is indeed the case for total groves

$$\underline{n} + \underline{m} = \underline{n + m},$$

and deduces the general case from this as every grove is a subset of some total grove.

Proposition 5.2 (Recursive property of addition). *Let $x = x^l \vee x^r$ and $y = y^l \vee y^r$ be nonzero trees. Then*

$$x + y = x^l \vee (x^r + y) \cup (x + y^l) \vee y^r.$$

The recursive property of addition says that the sum of two trees x and y is naturally a union of two sets, which we call the *left* and *right sum* of x and y:

$$x \dashv y = x^l \vee (x^r + y) \quad \text{and} \quad x \vdash y = (x + y^l) \vee y^r.^2 \tag{3}$$

Note that $x + y = x \dashv y \cup x \vdash y$. You can think about this as splitting the plus sign $+$ into two signs $\dashv$ and $\vdash$. From (2) and the definition, we see that the definition for left sum and right sum can also be extended to groves by distributing.

With the definition of inherited trees/groves and (3), one can easily check that left (respectively right) inheritance is passed along via right (respectively left) sums. More precisely,

Lemma 5.3. *Let y be a left-inherited tree. Then $x \vdash y$ is left-inherited. Similarly, if x is right-inherited, then $x \dashv y$ is right-inherited.*

[2] We set $x \vdash \underline{0} = \underline{0} \dashv y = \underline{0}$.

5.3. ***Universal expression.*** It turns out that every tree can expressed as a combination of left and right sums of $\vee$. This expression is unique modulo the failure of left and right sum to be associative. More precisely,

Proposition 5.4. *Every tree x of degree n can be written in as an iterated Left and Right sum of n copies of $\vee$. This is called the* universal expression *of x, and we denote it by $w_x(\vee)$. This expression is unique modulo*:

(i) $(x \dashv y) \dashv z = x \dashv (y + z)$,

(ii) $(x \vdash y) \dashv z = x \vdash (y \dashv z)$,

(iii) $(x + y) \vdash z = x \vdash (y \vdash z)$.

For example,

$$\curlyvee = \vee \vdash \vee \quad \text{and} \quad \curlyvee = \vee \vdash \vee \dashv \vee.$$

Loday gives a algorithm for computing the universal expression of a tree x.

Proposition 5.5 (Recursive property for universal expression). *Let x be a tree of degree greater than* 1. *The algorithm for determining $w_x(\vee)$ is given through the recursive relation*

$$w_x(\vee) = w_{x^l}(\vee) \vdash \vee \dashv w_{x^r}(\vee).$$

6. Multiplication

Essentially, we define the multiplication to distribute on the left over the universal expression.

Definition 6.1. The product $x \times y$ is defined by

$$x \times y = w_x(y).$$

This means to compute the product $x \times y$, first compute the universal expression for x, then replace each occurrence of $\vee$ by the tree y, then compute the resulting Left and Right sums. For example, one can easily check that $\curlyvee = \vee \vdash \vee$. This means for any tree y, $\curlyvee \times y = y \vdash y$. In particular,

$$\curlyvee \times \curlyvee = \curlyvee \vdash \curlyvee$$

is the tree shown in the figure on page 2.

Note that the definition of $x \times y$ as stated still makes sense if y is a grove. We can further extend the definition of multiplication to the case when x is a grove by declaring multiplication to be distributive on the left over disjoint unions:

$$(x \cup x') \times y = x \times y \cup x' \times y = w_x(y) \cup w_{x'}(y).$$

7. Properties

We list a few properties of *arithmetree*.

- The addition $+ : \mathbb{Y} \times \mathbb{Y} \to \mathbb{Y}$ is associative, but not commutative.
- The multiplication $\times : \mathbb{Y} \times \mathbb{Y} \to \mathbb{Y}$ is associative, but not commutative. It is distributive on the left with respect to $+$, but it is not right distributive.
- There is an injective map $\mathbb{N} \hookrightarrow \mathbb{Y}$, $n \mapsto \underline{n}$ (defined in Section 3) that respects the arithmetic. Namely,

$$\underline{m+n} = \underline{m} + \underline{n} \quad \text{and} \quad \underline{mn} = \underline{m} \times \underline{n} \quad \text{for all } m, n \in \mathbb{N}.$$

- Degree gives a surjective map $\deg : \mathbb{Y} \to \mathbb{N}$ that respects the arithmetic and is a one-sided inverse to the injection above . For every $x, y \in \mathbb{Y}$,

$$\deg(x+y) = \deg(x) + \deg(y) \quad \text{and} \quad \deg(x \times y) = \deg(x)\deg(y).$$

- $\deg(\underline{n}) = n$ for all $n \in \mathbb{N}$.
- The neutral element for $+$ is $\underline{0} = \mid$.
- The neutral element for $\times$ is $\underline{1} = \curlyvee$.
- The involution σ satisfies

$$\sigma(x+y) = \sigma(y) + \sigma(x) \quad \text{and} \quad \sigma(x \times y) = \sigma(x) \times \sigma(y).$$

8. Results

The recursive properties of addition and multiplication allowed us to implement arithmetree on a computer using PARI/GP [2005]. The computational experimentation was done using Loday's [2002] naming convention for trees.

8.1. ***Counting trees.*** Since each grove $x \in \mathbb{Y}$ is just a subset of trees, there is another measure of the "size" of x other than degree.

Definition 8.1. Let $x \in \mathbb{Y}$ be a grove. The *count* of x, denoted $C(x)$ is defined as the cardinality of x.

It turns out that count function gives a coarse measure of how complicated a grove x is in terms of arithmetree. Namely, if x is the sum (resp. product) of other groves, then the count of x is at least as large as the count of any of the summands (resp. factors).

Lemma 8.2. *Let $x, y \in \mathbb{Y}$ be two nonzero groves. Then*

(i) $C(x \dashv y) \geq C(x)C(y)$, *with equality if and only if x is a left-inherited grove.*

(ii) $C(x \vdash y) \geq C(x)C(y)$, *with equality if and only if y is a right-inherited grove.*

Proof. We first consider Lemma 8.2(i). Since $\dashv$ is distributive over unions, it suffices to prove the case when x and y are trees. Namely, we must show that for all nonzero trees x and y, $C(x \dashv y) \geq 1$, with equality if and only if x is a left-inherited tree. It is immediate that $C(x \dashv y) \geq 1$; it remains to show that equality is only attained when x is left-inherited. From the definition of left sum, $x \dashv y = x^l \vee (x^r + y)$. If x is not left-inherited, then $x^r \neq \underline{0}$ and

$$\begin{aligned} C(x \dashv y) &= C(x^l \vee (x^r + y)) = C(x^r + y) \\ &= C(x^r \dashv y \cup x^r \vdash y) = C(x^r \dashv y) + C(x^r \vdash y) \\ &> 1. \end{aligned}$$

On the other hand, if x is left-inherited, then $x^r = \underline{0}$ and

$$C(x \dashv y) = C(x^l \vee (x^r + y)) = C(x^l \vee y) = 1.$$

Item (ii) follows similarly. □

Proposition 8.3. *Let $x, y \in \mathbb{Y}$ be two nonzero groves. Then*

(i) $C(x + y) \geq 2C(x)C(y)$, *with equality if and only if x is a left-inherited and y is right-inherited.*

(ii) $C(x \times y) \geq C(x)C(y)^{\deg(x)}$.

Proof. Since $x + y = x \dashv y \cup x \vdash y$, Proposition 8.3(i) follows immediately from Lemma 8.2. For Proposition 8.3(ii), we note that multiplication is left distributive over unions, and so it suffices to prove the case when x is a tree. Namely we must show that for a tree x and a grove y, $C(x \times y) \geq C(y)^{\deg(x)}$.

Let w_x be the universal expression of the tree x. Then $x \times y = w_x(y)$ is some combination of left and right sums of y. By distributivity of left and right sum over unions and repeated usage of Lemma 8.2, the result follows. □

8.2. *Primes.*

Definition 8.4. A grove x is said to be *prime* if x is not the product of two groves different from $\underline{1}$.

Since $\deg(x \times y) = \deg(x)\deg(y)$ for all groves x, y, it is immediate that any grove of prime degree is prime. However, there are also prime groves of composite degree. For example, by taking all possible products of elements of $\mathbb{Y}_2$, one can check by hand that the primitive tree L_4 is a prime grove of degree 4.

We turn our focus to prime trees, which are prime groves with count 1. It turns out that composite trees have a nice description in terms of inherited trees. Namely, a composite tree must have an inherited tree as a right factor and a primitive tree as a left factor.

Theorem 8.5. *Let z be a composite tree of degree n. Then there exists a proper divisor $d \neq 1$ of n and a tree $T \in Y_{d-1}$ such that*

$$z = L_{n/d} \times (\underline{0} \vee T) \quad \text{or} \quad z = R_{n/d} \times (T \vee \underline{0}).$$

Proof. Let $z = x \times y$ be a composite tree of degree n. By Proposition 8.3, x and y must also be trees. Since $n = \deg(z) = \deg(x)\deg(y)$, it follows that there exists a proper divisor $d \neq 1$ of n such that $\deg(y) = d$ and $\deg(x) = n/d$.

We proceed by induction on the degree of x. Suppose x is a tree of degree 2. Then $x = \vee \dashv \vee$ or $x = \vee \vdash \vee$. If $x = \vee \vdash \vee$, then $x = L_2$ is primitive and

$$1 = C(x \times y) = C(y \vdash y).$$

From Proposition 8.3, it follows that y is right-inherited. Similarly, if $x = \, | \, \dashv \, | \,$, then $x = R_2$ and y is left-inherited.

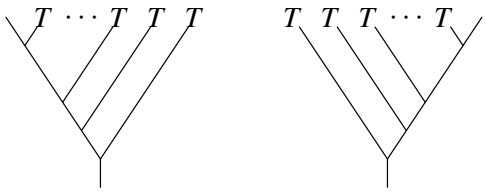
Now suppose x is a tree of degree k such that $x \times y$ is a tree of degree n. From Proposition 5.5 and the definition of multiplication, it follows that

$$\begin{aligned} x \times y &= w_x(y) \\ &= w_{x^l}(y) \vdash y \dashv w_{x^r}(y) \\ &= (x^l \times y) \vdash y \dashv (x^r \times y). \end{aligned}$$

Suppose $x^r \neq \underline{0}$. Then $x^r \times y \neq \underline{0}$ and $C(y \dashv (x^r \times y)) = 1$. Then by Proposition 8.3, y is left-inherited. Let $T = y \dashv (x^r \times y)$. By Lemma 5.3, T is also left-inherited. Since $C((x^l \times y) \vdash T) = 1$ and $T \neq \underline{0}$, we must have that either T is also right-inherited, or $(x^l \times y) = \underline{0}$. The only tree that is both left and right-inherited is the tree $\underline{1} = \, | \,$. It follows that $(x^l \times y) = \underline{0}$, and hence $x^l = \underline{0}$. By the inductive hypothesis, x^r is a right-primitive tree, and hence $x = R_k$.

Now suppose $x^r = \underline{0}$. Then $x^l \neq \underline{0}$, and an analogous argument shows that y is left-inherited and $x = L_k$. □

From this theorem, we get a nice picture of the possible shapes of composite trees:



Indeed, one computes that the product $L_k \times (\underline{0} \vee T)$ has the form shown on the left, and and $R_k \times (T \vee \underline{0})$ the form on the right.

It follows that the primitive trees (L_k and R_k) and the inherited trees ($\underline{0} \vee T$ and $T \vee \underline{0}$) are prime. More precisely:

Proposition 8.6. *A nonzero tree is either* $\vee$, *prime*, *or the product of exactly two prime trees. Furthermore*, *the factors are exactly the ones given in Theorem 8.5*, *and can be read off from the shape of the tree.*

The following combinatorial formula is a consequence of Proposition 8.6:

Corollary 8.7. *Let* a_n *denote the number of composite trees of degree* n. *Then*

$$\frac{a_n}{2} = -c_1 - c_n + \sum_{d|n} c_d, \quad \text{where } c_d \text{ is the } d\text{-th Catalan number.}$$

9. Final remarks

9.1. ***Unique factorization.*** Loday [2002] conjectures that arithmetree possesses unique factorization. Namely, when a grove x is written as a product of prime groves, the ordered sequence of factors is unique. Very narrowly interpreted, this statement is false. For example since multiplication in $\mathbb{N}$ is commutative and multiplication in $\mathbb{Y}$ extends arithmetic on $\mathbb{N}$, we see that for $n \in \mathbb{N}$, if $n = p_1 p_2 \cdots p_k$, then

$$\underline{n} = \underline{p_{\sigma(1)}} \times \underline{p_{\sigma(2)}} \times \cdots \times \underline{p_{\sigma(k)}},$$

for any permutation σ. However, away from the image of $\mathbb{N}$ in $\mathbb{Y}$, it appears that this narrow interpretation is true. Specifically, computer experimentation on groves of degree up to 12 yielded a unique ordered sequence of prime factors for each grove outside of the image of $\mathbb{N}$ in $\mathbb{Y}$.

If we interpret the image of $\mathbb{N}$ in $\mathbb{Y}$ in terms of the count function, we see that it is precisely the set of groves with maximal count:

$$\mathbb{Y}^{\max} = \bigcup_{n\in\mathbb{N}} \{x \in \mathbb{Y}_n \mid C(x) = c_n\}.$$

This subset $\mathbb{Y}^{\max}$ possesses unique factorization up to permutation of the factors. On the other extreme, the trees are precisely the set of groves with minimal count;

$$\mathbb{Y}^{\min} = \bigcup_{n\in\mathbb{N}} \{x \in \mathbb{Y}_n \mid C(x) = 1\}.$$

It follows from Proposition 8.6 that $\mathbb{Y}^{\min}$ possesses unique factorization in the narrow sense. The question of unique factorization for all of $\mathbb{Y}$ is open.

9.2. ***Additively irreducible.*** From Proposition 8.3 we see that not every grove can be written as a sum of groves. In fact it is easy to see that every tree is *additively irreducible* in the sense that it cannot be written as the sum of two groves. It would be interesting to study additively irreducible groves. In an analogue to the question of unique factorization, one could ask if arithmetree possesses *unique partitioning*. Namely, when a grove is written as a sum of additively irreducible elements, is the ordered sequence of summands unique?

Acknowledgements

Yasaki thanks Paul Gunnells for introducing him to this very interesting topic, and for all the help with typesetting and computing. The authors thank the referee for helpful comments.

References

[Loday 2002] J.-L. Loday, "Arithmetree", *J. Algebra* **258**:1 (2002), 275–309. MR 2004c:05053 Zbl 1063.16044

[Loday et al. 2001] J.-L. Loday, A. Frabetti, F. Chapoton, and F. Goichot, *Dialgebras and related operads*, Lecture Notes in Mathematics **1763**, Springer, Berlin, 2001. MR 2002e:00012 Zbl 0970.00010

[PARI 2005] The PARI Group, "PARI/GP", version 2.1.6, 2005, http://pari.math.u-bordeaux.fr/.

[Stanley 2007] R. P. Stanley, "Catalan addendum", 2007, available at http://www-math.mit.edu/~rstan/ec/catadd.pdf.

Received: 2008-06-03 Revised: 2011-05-19 Accepted: 2011-05-20

bruno@math.umass.edu *Department of Mathematics and Statistics, Lederle Graduate Research Tower, The University of Massachusetts at Amherst, Amherst, Massachusetts 01003-9305, United States*

d_yasaki@uncg.edu *Department of Mathematics and Statistics, The University of North Carolina at Greensboro, Greensboro, North Carolina 27402-6170, United States*

INVOLVE 4:1(2011)

msp

Vertical transmission in epidemic models of sexually transmitted diseases with isolation from reproduction

Daniel Maxin, Timothy Olson and Adam Shull

(Communicated by Suzanne Lenhart)

We describe a population logistic model exposed to a mild life-long sexually transmitted disease, that is, without significant increased mortality among infected individuals and providing no immunity/recovery. We then modify this model to include groups isolated from sexual contact and analyze their potential effect on the dynamics of the population. We are interested in how the isolated class may curb the growth of the infected group while keeping the healthy population at acceptable levels. In particular, we analyze the connection between vertical transmission and isolation from reproduction on the long term behavior of the disease. A comparison with similar effects caused by vaccination and quarantine is also provided.

1. Introduction

The dynamics of a population depends on the relation between reproduction and mortality. One factor that we analyze in this paper is the long-term effect on the population growth caused by the segregation of portions of the general (reproductive) population into a nonreproductive class that really consists of individuals of two very different kinds: *sexually active but nonprocreating*, such as infertile individuals, and *sexually inactive*, consisting of individuals who by choice or medical reasons refrain from sexual contact for life. The influence of the nonreproductive group on general population dynamics has been analyzed for several exponential and logistic models in [Milner 2005]. It has been shown that the nonreproductive group can indeed alter the population trend and may even make an exponentially increasing population stagnate or decline. A similar result holds for logistic models. Maxin and Milner [2007] extended these models to incorporate a sexually transmitted disease (STD) without recovery that does not increase mortality. It has been

MSC2000: primary 92D30; secondary 92D25.

Keywords: sexually transmitted diseases, abstaining individuals, vertical transmission, isolation from reproduction.

shown that the abstaining groups have the ability to induce a stable disease-free equilibrium (DFE) in an endemic situation. This is quite different from quarantine since the sexually isolated individuals do not reproduce and, by this, the number of susceptibles decreases since no vertical transmission is assumed.

In this paper we extend the logistic model from [Maxin and Milner 2007] — a reference we henceforth abbreviate as [MM 2007] — to include vertical transmission which assumes that the newborn can acquire the disease from an infected mother. It is intuitively obvious that, with vertical transmission, there is a new source of newly infected individuals in the population and the conditions for disease clearance will become more restrictive. Our goal in this paper is to show that, even in this case, a stable, disease-free steady state is possible and may be caused primarily by isolation from reproduction.

The paper is structured as follows. In Section 2, we introduce the model and analyze the extinction and the disease-free equilibrium, and correlate these results with the ones obtained in [MM 2007]. We then compute a threshold condition on the nonreproductive rates that describes how the isolated class induces a disease-free equilibrium in an endemic situation caused by vertical transmission. In Section 3, we analyze a particular model that assumes total vertical transmission when all the newborn from infected people are infected at birth and that leads to the existence of a *total endemic* steady state when the entire healthy population vanishes. While this is not realistic for known diseases, the stability condition of the endemic equilibrium suggests that, contrary to what might be expected, a higher isolation rate of infected leads to an endemic equilibrium (where healthy and infected individuals coexist) regardless of how big the infection rate may be. We conclude in Section 4 with a brief comparison between our model and a similar S-I type model with vaccination and quarantine to show that the previous result may not be possible in the absence of isolation from reproduction. We conclude our paper with several thoughts on further avenues of research.

2. The logistic model with abstaining groups and vertical transmission

Maxin and Milner [MM 2007] introduced several exponential and logistic STD models that incorporate an abstaining class A of people who are isolated from sexual contact. Here we consider their model with logistic mortality and assume that each newborn from an infected individual has a probability ϵ of being healthy at birth. Thus, if β is the per capita birth rate and I is the infected class, the rate at which individuals are born already infected is $\beta(1-\epsilon)I$. The system becomes

$$\begin{cases} S' = \beta S + \beta\epsilon I - \lambda SI - \bar{\mu} S - \nu_1 S, \\ I' = \beta(1-\epsilon)I + \lambda SI - \bar{\mu} I - \nu_2 I, \\ A' = \nu_1 S + \nu_2 I - \bar{\mu} A, \end{cases} \tag{1}$$

where $\bar{\mu} = \mu + bP$ with $P = S + I + A$.

The meaning of the remaining parameters is as follows:

- S and A denote the susceptible and the abstaining class. Note that, since the abstaining individuals do not reproduce and do not participate in the infection process, we can include both the infected and healthy isolated people into a single group A in order to keep the dimension of the system as small as possible. Whenever the disease is cleared (such as in the case of a stable disease-free equilibrium) A will contain healthy isolated individuals only.
- $\bar{\mu}$ is the logistic death rate and b is the logistic linear coefficient that captures the total population effect on the death rate.
- λ represents the infection rate using the mass-action law corresponding to a homogeneous population.
- ν_1 and ν_2 represent the transition rates from susceptibles and infected into the isolated class A.
- P will denote, throughout this paper, the total population.

When $\epsilon = 1$, this system is identical with the one analyzed in [MM 2007].

It is reasonable to assume that the isolation rate of infected individuals is greater, since some infected individuals may choose to quarantine themselves in order to avoid spreading the disease. Thus, we will assume throughout this paper that

$$\nu_1 < \nu_2.$$

The model always admits an extinction equilibrium $(0, 0, 0)$.

If $\beta - \mu - \nu_1 > 0$ there is also a disease-free equilibrium (S_*, I_*, A_*) where

$$\begin{cases} S_* = \Big(K - \dfrac{\nu_1}{b}\Big)\Big(1 - \dfrac{\nu_1}{\beta}\Big) = \Big(\dfrac{\beta - \mu - \nu_1}{b}\Big)\Big(1 - \dfrac{\nu_1}{\beta}\Big), \\ I_* = 0, \\ A_* = \Big(K - \dfrac{\nu_1}{b}\Big)\dfrac{\nu_1}{\beta} = \Big(\dfrac{\beta - \mu - \nu_1}{b}\Big)\dfrac{\nu_1}{\beta}, \end{cases} \tag{2}$$

with $K = (\beta - \mu)/b$. The endemic equilibrium will be analyzed in the context of complete vertical transmission in the following section.

Theorem 1 (stability of the boundary steady states). *The extinction equilibrium is locally asymptotically stable if*

$$\beta - \mu - \nu_1 < 0.$$

The disease-free equilibrium (S_*, I_*, A_*) *is locally asymptotically stable if*

$$\beta - \mu - \nu_1 > 0 \quad \textit{and} \quad \lambda < \frac{\beta\epsilon - \nu_1 + \nu_2}{\Big(1 - \dfrac{\nu_1}{\beta}\Big)\Big(K - \dfrac{\nu_1}{b}\Big)}.$$

Proof. The Jacobian of (1) is

$$J = \begin{pmatrix} \beta-\lambda I-\bar{\mu}-bS-\nu_1 & \beta\epsilon-\lambda S-bS & -bS \\ \lambda I-bI & \beta(1-\epsilon)+\lambda S-\bar{\mu}-bI-\nu_2 & -bI \\ \nu_1-bA & \nu_2-bA & -\bar{\mu}-bA \end{pmatrix}.$$

Evaluated at $(0, 0, 0)$ this is

$$J(0,0,0) = \begin{pmatrix} \beta-\mu-\nu_1 & \beta\epsilon & 0 \\ 0 & \beta(1-\epsilon)-\mu-\nu_2 & 0 \\ \nu_1 & \nu_2 & -\mu \end{pmatrix}.$$

It follows that the extinction equilibrium is locally asymptotically stable if

$$\beta-\mu-\nu_1<0 \quad \text{and} \quad \beta(1-\epsilon)-\mu-\nu_2<0.$$

However, the second inequality follows from the first, since $0<\epsilon<1$ and $\nu_1<\nu_2$:

$$\beta(1-\epsilon)<\beta<\mu+\nu_1<\mu+\nu_2.$$

Assuming now that $\beta-\mu-\nu_1>0$, and denoting

$$P_* = S_* + A_* = K - \frac{\nu_1}{b} = \frac{\beta-\mu-\nu_1}{b} > 0,$$

the Jacobian J evaluated at (S_*, I_*, A_*) is

$$\begin{pmatrix} \beta-\mu-bP_*-bS_*-\nu_1 & \beta\epsilon-\lambda S_*-bS_* & -bS_* \\ 0 & \beta(1-\epsilon)+\lambda S_*-\mu-bP_*-\nu_2 & 0 \\ \nu_1-bA_* & \nu_2-bA_* & -\mu-bP_*-bA_* \end{pmatrix}.$$

The eigenvalues are $\beta(1-\epsilon)+\lambda S_*-\mu-bP_*-\nu_2$ (this being the single nonzero entry on its row) together with the eigenvalues of the complementary minor,

$$M = \begin{pmatrix} \beta-\mu-bP_*-bS_*-\nu_1 & -bS_* \\ \nu_1-bA_* & -\mu-bP_*-bA_* \end{pmatrix}.$$

Since $\operatorname{Tr}(M) = -\mu-2bP_*<0$ and $\det M = b\beta S_*>0$, the eigenvalues of M have negative real parts. Thus local asymptotic stability holds for (S_*, I_*, A_*) if

$$\beta(1-\epsilon)+\lambda S_*-\mu-bP_*-\nu_2<0,$$

which is equivalent to

$$\lambda < \frac{\beta\epsilon-\nu_1+\nu_2}{\left(1-\frac{\nu_1}{\beta}\right)\left(K-\frac{\nu_1}{b}\right)}. \qquad \square$$

From [MM 2007] we know that, in the absence of the isolated class A, the disease is endemic if $\beta/K<\lambda$. Similarly, with vertical transmission, if there is no isolation from reproduction, the disease is endemic provided that $\beta\epsilon/K<\lambda$. The

double inequality below indicates the range of the infection rate λ that would cause an endemic situation in the absence of the isolated class A and a stable disease-free equilibrium in the presence of it:

$$\frac{\beta\epsilon}{K} < \lambda < \frac{\beta\epsilon - \nu_1 + \nu_2}{\left(1 - \frac{\nu_1}{\beta}\right)\left(K - \frac{\nu_1}{b}\right)}. \tag{3}$$

This condition resembles the similar one obtained in [MM 2007], with $\epsilon = 1$ (no vertical transmission):

$$\frac{\beta}{K} < \lambda < \frac{\beta - \nu_1 + \nu_2}{\left(1 - \frac{\nu_1}{\beta}\right)\left(K - \frac{\nu_1}{b}\right)}.$$

This means that the isolated class A, represented by the two isolation rates ν_1 and ν_2, has the ability to induce stability to the disease-free equilibrium in an otherwise endemic situation. With the addition of vertical transmission we notice another threshold effect which suggests that the vertical transmission alone can induce an endemic situation even in the case where the abstaining class satisfies the condition in [MM 2007]. This happens if the infection rate satisfies

$$\frac{\beta\epsilon - \nu_1 + \nu_2}{\left(1 - \frac{\nu_1}{\beta}\right)\left(K - \frac{\nu_1}{b}\right)} < \lambda < \frac{\beta - \nu_1 + \nu_2}{\left(1 - \frac{\nu_1}{\beta}\right)\left(K - \frac{\nu_1}{b}\right)}.$$

To summarize, the vertical transmission reduces the disease-free stability range of λ, which is to be expected with the additional infected newborns in the model.

In Figure 1 we plot two numerical examples to illustrate Theorem 1. The birth

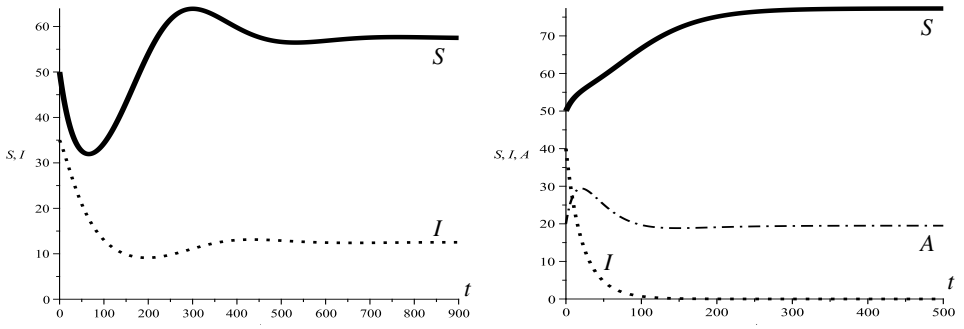


Figure 1. Example equilibria: endemic equilibrium (left) in the absence of abstainers ($\nu_1 = \nu_2 = 0$) and disease-free equilibrium (right) in their presence ($\nu_1 = 0.01$, $\nu_2 = 0.04$). In both cases, $\beta = 0.04962$, $\mu = 0.02026$, $\epsilon = 0.7$, $b = 0.0002$, and $\lambda = 0.0004$. The first inequality in (3) is satisfied in the first case, and both are in the second.

and death rates are those given in the *CIA World Factbook* for Niger in 2008, but all other parameter values are for illustration purposes only and do not reflect real data.

A major difference from the model treated in [MM 2007] appears when the vertical transmission rate is very high. Although not realistic, for theoretical purposes we will assume the extreme case, $\epsilon = 0$, which indicates 100% vertical transmission. We treat this case in greater detail in the following section.

3. Complete vertical transmission

Setting $\epsilon = 0$ in (1), we obtain:

$$\begin{cases} S' = \beta S - \lambda SI - \bar{\mu} S - \nu_1 S, \\ I' = \beta I + \lambda SI - \bar{\mu} I - \nu_2 I, \\ A' = \nu_1 S + \nu_2 I - \bar{\mu} A. \end{cases} \tag{4}$$

The system, in this form, allows us to explicitly compute the endemic equilibrium (a nontrivial task if $\epsilon \neq 0$):

$$S^* = \frac{\mu^* + \nu_2 - \beta}{\lambda}, \quad I^* = \frac{\beta - \mu^* - \nu_1}{\lambda}, \quad A^* = \frac{(\nu_2 - \nu_1)(\beta - \mu^*)}{\lambda \mu^*},$$

where $\mu^* = \mu + bP^*$. Adding the equations for S^*, I^*, and A^* together gives us

$$P^* = \frac{(\nu_2 - \nu_1)\beta}{\lambda \mu^*}.$$

For a biologically meaningful endemic equilibrium (EE) to exist (i.e., positive values) we need

$$\nu_1 < \beta - \mu^* < \nu_2,$$

or

$$\frac{\beta}{\mu^* + \nu_1} > 1 \quad \text{and} \quad \frac{\beta}{\mu^* + \nu_2} < 1.$$

This translates to a requirement that the reproductive number of the susceptibles must be greater than one, while the reproductive number of the infected population must be less than one.

In addition to the disease-free and endemic equilibria, (4) admits a third steady state in which the entire healthy population vanishes. We call this the susceptible extinction equilibrium (SEE):

$$\bar{S} = 0, \quad \bar{I} = \left(1 - \frac{\nu_2}{\beta}\right)\bar{P}, \quad \bar{A} = \frac{\nu_2}{\beta}\bar{P},$$

where $\bar{P} = (\beta - \mu - \nu_2)/b$.

We see that, for a positive SEE equilibrium, we need $\beta - \mu - \nu_2 > 0$.

Theorem 2 (existence and local stability conditions for EE and SEE). *If either*

$$\frac{\nu_2-\nu_1}{\left(1-\frac{\nu_1}{\beta}\right)\left(K-\frac{\nu_1}{b}\right)}<\lambda<\frac{\nu_2-\nu_1}{\left(1-\frac{\nu_2}{\beta}\right)\left(K-\frac{\nu_2}{b}\right)} \tag{5}$$

or

$$\frac{\nu_2-\nu_1}{\left(1-\frac{\nu_1}{\beta}\right)\left(K-\frac{\nu_1}{b}\right)}<\lambda \quad \textit{and} \quad \beta<\frac{\mu}{2}+\nu_2, \tag{6}$$

the endemic equilibrium (S^*, I^*, A^*) *exists and is locally asymptotically stable. If*

$$\beta>\mu+\nu_2 \quad \textit{and} \quad \lambda>\frac{\nu_2-\nu_1}{\left(1-\frac{\nu_2}{\beta}\right)\left(K-\frac{\nu_2}{b}\right)}, \tag{7}$$

the susceptible extinction equilibrium $(\bar{S}, \bar{I}, \bar{A})$ *exists and is locally asymptotically stable.*

Proof. First we show that the EE is stable whenever it exists. The Jacobian of (4), evaluated at (S^*, I^*, A^*), is

$$J(S^*, I^*, A^*)=\begin{pmatrix} -bS^* & -\lambda S^*-bS^* & -bS^* \\ \lambda I^*-bI^* & -bI^* & -bI^* \\ \nu_1-bA^* & \nu_2-bA^* & -\mu^*-bA^* \end{pmatrix}.$$

If the characteristic equation of this matrix is $x^3+p_1x^2+p_2x+p_3=0$, then

$$\begin{aligned} p_1 &= -\operatorname{Tr}(J)=\mu^*+bP^*, \\ p_2 &= \left(b^2S^*I^*+(\lambda^2-b^2)S^*I^*\right)+\left(bS^*(\mu^*+bA^*)+bS^*(\nu_1-bA^*)\right) \\ &\qquad +\left(bI^*(\mu^*+bA^*)+bI^*(\nu_2-bA^*)\right) \\ &= \lambda^2I^*S^*+b\nu_1S^*+b\nu_2I^*+b\mu^*(S^*+I^*), \\ p_3 &= -\operatorname{Det}(J)=\lambda S^*I^*(b\nu_2-b\nu_1+\lambda\mu^*+\lambda bA^*). \end{aligned}$$

Clearly $p_1>0$, $p_2>0$ and $p_3>0$, since $\nu_2>\nu_1$.

Replacing S^*, I^*, A^* and P^* with their corresponding values computed above, we also see that $p_1p_2>p_3$ since

$$p_1p_2-p_3=\frac{b\beta(\nu_2-\nu_1)\left(\lambda(\mu^*)^2+b\beta(\nu_2-\nu_1)\right)}{\lambda^2\mu^*}>0.$$

Hence, according to the Routh–Hurwitz criterion, the interior equilibrium is always stable whenever it exists. It remains now to interpret the positivity condition $\nu_1<\beta-\mu^*<\nu_2$ in terms of the original parameters.

To this end, we solve for μ^* using the following equation:

$$P^* = \frac{\mu^* - \mu}{b} = \frac{\beta(\nu_2 - \nu_1)}{\lambda\mu^*}.$$

There is a unique positive solution

$$\mu^* = \frac{\mu\lambda + \sqrt{\mu^2\lambda^2 + 4b\beta\lambda(\nu_2 - \nu_1)}}{2\lambda},$$

and the existence condition above becomes

$$2(\beta - \nu_2) - \mu < \frac{1}{\lambda}\sqrt{\mu^2\lambda^2 + 4b\beta\lambda(\nu_2 - \nu_1)} < 2(\beta - \nu_1) - \mu. \tag{8}$$

Consider the second inequality first. Its right side is positive, since our standing assumption is that $\beta > \mu + \nu_1$, to avoid total population extinction. Therefore squaring both sides leads to an equivalent inequality,

$$\frac{1}{\lambda^2}\big(\mu^2\lambda^2 + 4b\beta\lambda(\nu_2 - \nu_1)\big) < 4(\beta - \nu_1)^2 + \mu^2 - 4\mu(\beta - \nu_1),$$

which after simplification becomes, in terms of $K = \dfrac{\beta - \mu}{b}$, the condition

$$\lambda > \frac{\nu_2 - \nu_1}{\left(1 - \frac{\nu_1}{\beta}\right)\left(K - \frac{\nu_1}{b}\right)}.$$

Thus the second inequality in (8) amounts to precisely the opposite of the condition for disease-free stability at the end of the statement of Theorem 1, in the case $\epsilon = 0$.

There remains to study the first inequality in (8). It is certainly satisfied if its left side is negative, that is, if

$$\beta < \frac{\mu}{2} + \nu_2.$$

In the opposite case, $\beta \geq \dfrac{\mu}{2} + \nu_2$, we can square both sides to obtain the equivalent condition

$$\lambda < \frac{\nu_2 - \nu_1}{\left(1 - \frac{\nu_2}{\beta}\right)\left(K - \frac{\nu_2}{b}\right)}. \tag{9}$$

In other words, the endemic equilibrium exists and it is stable if conditions (5) and (6) are satisfied.

The Jacobian of (4) evaluated at $(\bar{S}, \bar{I}, \bar{A})$ is

$$\begin{pmatrix} -\lambda\bar{I} + \nu_2 - \nu_1 & 0 & 0 \\ (\lambda - b)\bar{I} & -b\bar{I} & -b\bar{I} \\ \nu_1 - b\bar{A} & \nu_2 - b\bar{A} & -\bar{\mu} - b\bar{A} \end{pmatrix},$$

where $\bar{\mu}$ here denotes $\mu + b(\bar{S} + \bar{I} + \bar{A})$.

It is clear that one of the eigenvalues is negative when $\lambda \bar{I} > \nu_2 - \nu_1$, which is equivalent to the second condition in (7):

$$\lambda > \frac{\nu_2 - \nu_1}{\left(1 - \frac{\nu_2}{\beta}\right)\left(K - \frac{\nu_2}{b}\right)}. \tag{10}$$

Removing the row and column containing that eigenvalue leaves us with a 2×2 matrix whose determinant is always positive ($b\beta\bar{I} > 0$) and whose trace is always negative ($-\bar{\mu} - b\bar{P} < 0$). Thus, the susceptible extinction equilibrium is locally asymptotically stable, with λ satisfying condition (7) . □

Remark 1. Condition (6) has an interesting consequence. First, if $\beta < \mu/2 + \nu_2$, then $\beta < \mu + \nu_2$ also, so the susceptible extinction equilibrium does not exist in this case. This means that if ν_2 is big enough, the susceptible class never goes extinct and the endemic equilibrium is stable *regardless* of how big the infection rate λ may be. This emphasizes the epidemiological role of isolation from reproduction.

Remark 2. If one excludes the fact that the isolated class A does not reproduce, then the model (4) resembles an $S - I$ type model with vaccination (ν_1) and quarantine (ν_2). Thus, in order to sustain the previous remark that the Susceptible Extinction Equilibrium may be eliminated by the isolation from reproduction we need to investigate whether this result holds for a similar model where the vaccinated and quarantined classes do reproduce. In the next section we show that the answer to this question is negative meaning that the result obtained for our original model is indeed primarily due to the isolation from reproduction.

We provide some numerical examples to illustrate Theorem 2. In Figure 2 we

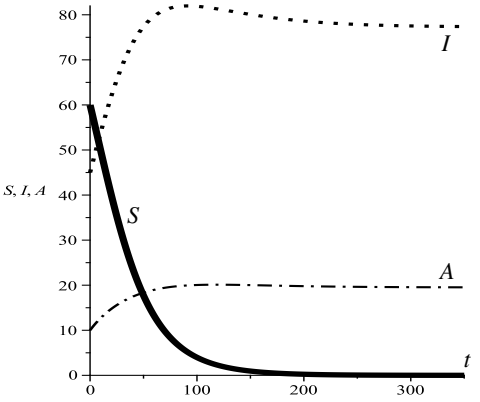


Figure 2. Example of a susceptible extinction equilibrium: $\beta = 0.04962$, $\mu = 0.02026$, $\nu_1 = 0.005$, $\nu_2 = 0.01$, $\epsilon = 0$, $b = 0.0002$, $\lambda = 0.0004$. Inequality (10) is satisfied.

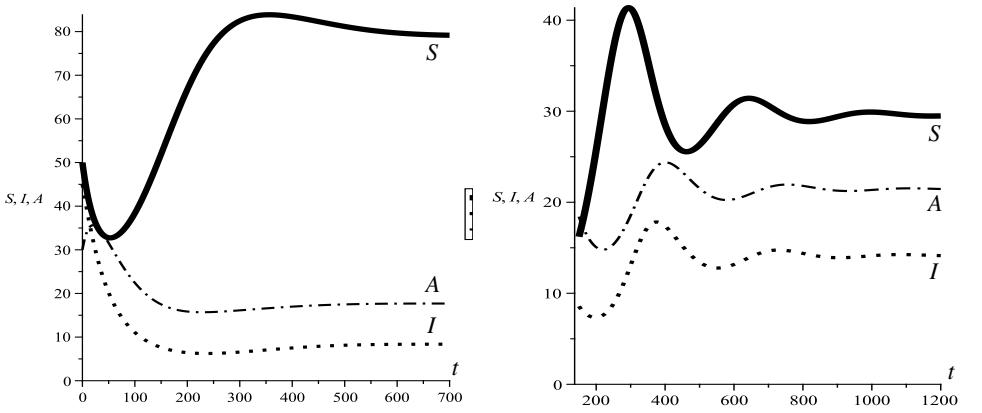


Figure 3. Examples of endemic equilibria: $\beta = 0.04962$, $\mu = 0.02026$, $\nu_1 = 0.005$, $\nu_2 = 0.04$, $\epsilon = 0$, $b = 0.0002$, $\lambda = 0.0004$ (left) or $\lambda = 0.0008$ (right). Inequalities (6) are satisfied.

show an example when the SEE is stable. In Figure 3 we illustrate the case of a stable EE satisfying (6). In Figure 3, right, we double the value of λ while keeping the other parameters the same as in the left half of the figure, to illustrate that under condition (6) the stability of EE is maintained regardless of how big the infection rate is.

4. A model with complete vertical transmission, vaccination and quarantine

The model proposed in this section is intended to eliminate the ambiguity concerning the epidemiological role of the isolated class A. In other words, we would like to see if the result in the previous section is due to the nonreproduction or perhaps due to vaccination and quarantine (which are other possible interpretations for the transition rates ν_1 and ν_2). To this end, we assume now that the model resembles an S-I type dynamics with vaccination and quarantine. Another main difference is that all individuals reproduce, including the quarantined. Since the isolated classes reproduce and one needs to track the infected and healthy newborns, we must separate the isolated class A into two classes: V, the vaccinated individuals and Q the quarantined infected people. Furthermore, we denote by η the transition rate from vaccinated individuals back to the susceptible class S to account for a possible imperfect vaccine where some individuals loose the acquired immunity.

The model is as follows:

$$\begin{cases} S' = \beta(S+V) - \lambda SI - \bar{\mu}S - \nu_1 S + \eta V, \\ I' = \beta(I+Q) + \lambda SI - \bar{\mu}I - \nu_2 I, \\ V' = \nu_1 S - \bar{\mu}V - \eta V, \\ Q' = \nu_2 I - \bar{\mu}Q, \end{cases} \tag{11}$$

where $\bar{\mu} = \mu + b(S + I + V + Q)$.

Remark 3. In this model we assumed the same reproduction rate for all individuals. A possible interpretation is that, in the case of sexually transmitted diseases, quarantine can be viewed as abstaining from sexual contact with healthy people only. This is true sometimes for diseases such as herpes simplex type 2 (HSV-2) when infected individuals search for partners among groups already infected. In reality, due to these considerations, the quarantined class will always exhibit a certain degree of isolation from reproduction. However, the main purpose of model (11) is to verify the results in the previous sections under the assumption that no isolation from reproduction occurs with transitions from one class to another.

Notice that there is no endemic equilibrium where the healthy and infected individuals coexist as shown below:

Substituting $V = \nu_1 S/(\bar{\mu} + \eta)$ and $Q = \nu_2 I/\bar{\mu}$ into the first two equations, we obtain

$$\lambda I = (\beta - \bar{\mu})\Big(1 + \frac{\nu_1}{\bar{\mu} + \eta}\Big) \quad \text{and} \quad \lambda S = (\bar{\mu} - \beta)\Big(1 + \frac{\nu_2}{\bar{\mu}}\Big).$$

Clearly, it is impossible for both of them to be positive since $\lambda S > 0$ implies $\beta < \bar{\mu}$ which, in turn, implies $\lambda I < 0$.

Adding the equations of (11) we obtain a logistic equation for the total population P:

$$P' = \beta P - \bar{\mu} P = (\beta - \mu - bP)P.$$

Therefore,

$$\lim_{t \to \infty} P(t) = \frac{\beta - \mu}{b} := K,$$

provided that $\beta > \mu$. If $\beta < \mu$ the population declines to zero.

Thus there are three steady states:

- the extinction equilibrium: $(0, 0, 0, 0)$,
- the susceptible extinction equilibrium (SEE):

$$\bar{S} = 0, \quad \bar{I} = \frac{\beta K}{\beta + \nu_2}, \quad \bar{V} = 0, \quad \bar{Q} = \frac{\nu_2 K}{\beta + \nu_2},$$

- the disease-free equilibrium (DFE):

$$S^* = \frac{(\beta + \eta)K}{\beta + \nu_1 + \eta}, \quad \bar{I} = 0, \quad \bar{V} = \frac{\nu_1 K}{\beta + \nu_1 + \eta}, \quad \bar{Q} = 0.$$

Theorem 3. *If $\beta > \mu$, the SEE is locally asymptotically stable and the DFE is unstable* (*whenever it exists*).

Proof. The Jacobian of (11) is

$$J = \begin{pmatrix} \beta-\lambda I-\bar{\mu}-bS-\nu_1 & -(\lambda+b)S & -bS+\beta+\eta & -bS \\ (\lambda-b)I & \beta+\lambda S-\bar{\mu}-bI-\nu_2 & -bI & -bI+\beta \\ \nu_1-bV & -bV & -\bar{\mu}-bV-\eta & -bV \\ -bQ & \nu_2-bQ & -bQ & -\bar{\mu}-bQ \end{pmatrix}.$$

We evaluate first the characteristic polynomial of $J(\bar{S}, \bar{I}, \bar{V}, \bar{Q})$, which is

$$f(x) = \big(x^2 + (b\bar{I} + b\bar{Q} + \beta + \nu_2)x + b(\bar{I} + \bar{Q})(\beta + \nu_2)\big) \times \big(x^2 + (\beta + \nu_1 + \lambda\bar{I} + \eta)x + \lambda\bar{I}(\beta + \eta)\big).$$

Since all its coefficients are positive then the real parts of all eigenvalues are negative and the susceptible extinction equilibrium is locally asymptotically stable whenever it exists.

On the contrary, for the disease-free equilibrium, the characteristic polynomial of $J(S^*, I^*, V^*, Q^*)$ is

$$g(x) = \big(x^2 + (bS^* + bV^* + \beta + \eta + \nu_1)x + b(S^* + V^*)(\beta + \eta + \nu_1)\big) \times \big(x^2 + (\beta + \nu_2 - \lambda S^*)x - \beta\lambda S^*\big)$$

and the real part of one of its eigenvalues is always positive: from the second quadratic factor of $g(x)$ we see that the product of its roots is given by

$$x_1 x_2 = -\beta\lambda S^* < 0. \qquad \square$$

Thus, the DFE is always unstable and the possibility of eliminating the disease is not possible through quarantine and vaccination alone when the population is faced with complete vertical transmission. In Figure 4 we provide a numerical example using the same parameter values as those in Figure 3 to illustrate that, in the absence of isolation from reproduction, the SEE is stable and the healthy population vanishes.

5. Conclusions

We modified the epidemic model with sexually abstaining groups introduced in [MM 2007] to include vertical transmission. We found that previous results claiming that the isolated class may induce the stability of the disease-free equilibrium in an endemic situation are still valid in the presence of vertical transmission although the range of the infection rate when this is possible is more restrictive.

One major difference appears when the vertical transmission rate is very high, that is, close to 100%. To simplify our analysis we actually considered a complete vertical transmission situation where every newborn from infected parents is infected as well. In this case, under certain conditions on the vital parameters, we

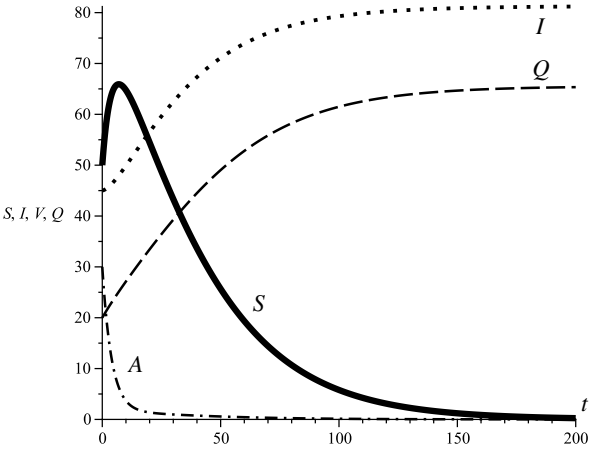


Figure 4. The susceptible extinction equilibrium in the absence of isolation from reproduction. Parameter values: $\beta = 0.04962$, $\mu = 0.02026$, $\nu_1 = 0.005$, $\nu_2 = 0.04$, $\epsilon = 0$, $b = 0.0002$, $\lambda = 0.0004$, $\eta = 0.2$.

found that the model admits a steady state (SEE) when the entire susceptible population vanishes, in addition to the disease-free and interior (endemic) steady states. The local stability analysis for the endemic equilibrium shows that both the infected and the healthy groups may coexist and that the total endemic situation when the healthy population declines to zero can be avoided by isolation from reproduction alone. A comparison with model (11) shows that this result may indeed be due to isolation from reproduction and not due to vaccination or quarantine, which are other possible interpretations for the transition rates ν_1 and ν_2.

One important limitation of our work is given by the use of one-sex models. Since we were interested in showing that there is an important correlation between vertical transmission and isolation from reproduction, we chose the simplest possible model to sustain our argument and to keep the mathematical details as simple as possible. Our next objective related to the present research is to investigate if similar results can be obtained using two-sex models. The influence of sexually abstaining groups on STD dynamics has been analyzed in [Maxin 2009; Maxin and Milner 2009] using a gender structured logistic model. We intend to extend that model to include vertical transmission. This research is currently underway and will be reported later.

References

[Maxin 2009] D. Maxin, "The influence of sexually active non-reproductive groups on persistent sexually transmitted diseases", *J. Biol. Dyn.* **3**:5 (2009), 532–550. MR 2010m:92068

[Maxin and Milner 2007] D. Maxin and F. A. Milner, "The effect of nonreproductive groups on persistent sexually transmitted diseases", *Math. Biosci. Eng.* **4**:3 (2007), 505–522. MR 2008c:92077

[Maxin and Milner 2009] D. Maxin and F. A. Milner, "The role of sexually abstained groups in two-sex demographic and epidemic logistic models with non-linear mortality", *J. Theor. Biol.* **258**:3 (2009), 389–402.

[Milner 2005] F. A. Milner, "How do nonreproductive groups affect population growth?", *Math. Biosci. Eng.* **2**:3 (2005), 579–590. MR 2007a:92058 Zbl 1082.92037

Received: 2009-09-10 Revised: 2010-12-16 Accepted: 2011-03-04

daniel.maxin@valpo.edu *Department of Mathematics and Computer Science, Valparaiso University, 1900 Chapel Drive, Valparaiso, IN 46383, United States*
http://faculty.valpo.edu/dmaxin/

Timothy.Olson@valpo.edu *Department of Mathematics and Computer Science, Valparaiso University, 1900 Chapel Drive, Valparaiso, IN 46383, United States*

adam.shull@valpo.edu *Department of Mathematics, Indiana University Bloomington, 831 East Third Street, Bloomington, IN 47405, United States*

INVOLVE 4:1(2011)

msp

On the maximum number of isosceles right triangles in a finite point set

Bernardo M. Ábrego, Silvia Fernández-Merchant and David B. Roberts

(Communicated by Kenneth S. Berenhaut)

Let Q be a finite set of points in the plane. For any set P of points in the plane, $S_Q(P)$ denotes the number of similar copies of Q contained in P. For a fixed n, Erdős and Purdy asked for the maximum possible value of $S_Q(P)$, denoted by $S_Q(n)$, over all sets P of n points in the plane. We consider this problem when $Q = \triangle$ is the set of vertices of an isosceles right triangle. We give exact solutions when $n \leq 9$, and provide new upper and lower bounds for $S_\triangle(n)$.

1. Introduction

Paul Erdős and George Purdy [1971; 1975; 1976] posed the question: *Given a finite set of points Q, what is the maximum number $S_Q(n)$ of similar copies that can be contained in an n-point set in the plane*? This problem remains open in general. However, there has been some progress regarding the order of magnitude of this maximum as a function of n. Elekes and Erdős [1994] noted that $S_Q(n) \leq n(n-1)$ for any pattern Q and they also gave a quadratic lower bound for $S_Q(n)$ when $|Q| = 3$ or when all the coordinates of the points in Q are algebraic numbers. They also proved a slightly subquadratic lower bound for all other patterns Q. Later, Laczkovich and Ruzsa [1997] characterized precisely those patterns Q for which $S_Q(n) = \Theta(n^2)$. In spite of this, the coefficient of the quadratic term is not known for any nontrivial pattern; it is not even known if $\lim_{n\to\infty} S_Q(n)/n^2$ exists!

Apart from being a natural question in discrete geometry, this problem also arose in connection with the optimization of algorithms designed to look for patterns among data obtained from scanners, digital cameras, telescopes, and so on [Brass 2002; Brass et al. 2005; Brass and Pach 2005].

Our paper considers the case where Q is the set of vertices of an isosceles right triangle. The case where Q is the set of vertices of an equilateral triangle has been considered in [Ábrego and Fernández-Merchant 2000]. To avoid redundancy, we

MSC2000: primary 52C10; secondary 05C35.
Keywords: Erdős problems, similar triangles, isosceles right triangles.
Supported in part by CURM, BYU, and the NSF (DMS–0636648).

refer to an isosceles right triangle as an IRT for the remainder of the paper. We begin with some definitions. Let P denote a finite set of points in the plane. We define $S_{\triangle}(P)$ to be the number of triplets in P that are the vertices of an IRT. Furthermore, let

$$S_{\triangle}(n) = \max_{|P|=n} S_{\triangle}(P).$$

As mentioned before, Elekes and Erdős established that $S_{\triangle}(n) = \Theta(n^2)$ and it is implicit from their work that $1/18 \leq \liminf_{n\to\infty} S_{\triangle}(n)/n^2 \leq 1$. The main goal of this paper is to derive improved constants that bound the function $S_{\triangle}(n)/n^2$. Specifically, in Sections 2 and 3, we prove:

Theorem 1. $$0.433064 < \liminf_{n\to\infty} \frac{S_{\triangle}(n)}{n^2} \leq \frac{2}{3} < 0.66667.$$

We proceed to determine in Section 4 the exact values of $S_{\triangle}(n)$ when $3 \leq n \leq 9$. Several ideas for the proofs of these bounds come from the equivalent bounds for equilateral triangles in [Ábrego and Fernández-Merchant 2000].

2. Lower bound

For $z \in P$, let $R_{\pi/2}(z, P)$ be the $\pi/2$ counterclockwise rotation of P with center z. Let $\deg_{\pi/2}(z)$ be the number of isosceles right triangles in P such that z is the right-angle vertex of the triangle. If $z \in P$, then $\deg_{\pi/2}(z)$ can be computed by simply rotating our point set P by $\pi/2$ about z and counting the number of points in the intersection other than z. Therefore,

$$\deg_{\pi/2}(z) = \left|P \cap R_{\pi/2}(z, P)\right| - 1. \tag{1}$$

Since an IRT has only one right angle,

$$S_{\triangle}(P) = \sum_{z\in P} \deg_{\pi/2}(z).$$

That is, the sum computes the number of IRTs in P. From this identity an initial lower bound of $\frac{5}{12}$ can be derived for $\liminf_{n\to\infty} S_{\triangle}(n)/n^2$ using the set

$$P = \left\{(x, y) \in \mathbb{Z}^2 : 0 \leq x \leq \sqrt{n}, 0 \leq y \leq \sqrt{n}\right\}.$$

We now improve this bound.

The following theorem generalizes our method for finding a lower bound. We denote by Λ the lattice generated by the points $(1, 0)$ and $(0, 1)$, and we refer to points in Λ as *lattice points*. The next result provides a formula for the leading term of $S_{\triangle}(P)$ when our points in P are lattice points enclosed by a given shape. This theorem, its proof, and notation, are similar to those of Theorem 2 in [Ábrego and Fernández-Merchant 2000], where we obtained a similar result for equilateral triangles in place of IRTs.

Theorem 2. *Let K be a compact set with finite perimeter and area* 1. *Define*

$$f_K : \mathbb{C} \to \mathbb{R}^+ \quad \text{as } f_K(z) = \text{Area}(K \cap R_{\pi/2}(z, K)), \quad \text{where } z \in K.$$

If K_n is a similar copy of K intersecting Λ in exactly n points, then

$$S_\triangle(K_n \cap \Lambda) = \left(\int_K f_K(z)\, dz\right) n^2 + O(n^{3/2}).$$

Proof. Given a compact set L with finite area and perimeter, we have

$$|rL \cap \Lambda| = \text{Area}(rL) + O(r) = r^2 \text{Area}(L) + O(r),$$

where rL is the scaling of L by a factor r. Therefore,

$$\begin{aligned} S_\triangle(K_n \cap \Lambda) &= \sum_{z \in K_n \cap \Lambda} |(\Lambda \cap K_n) \cap R_{\pi/2}(z, (K_n \cap \Lambda))| - 1 \\ &= \sum_{z \in K_n \cap \Lambda} \text{Area}(K_n \cap R_{\pi/2}(z, K_n)) + O(\sqrt{n}). \end{aligned}$$

We see that each error term in the sum is bounded by the perimeter of K_n, which is finite by hypothesis. Thus,

$$\begin{aligned} S_\triangle(K_n \cap \Lambda) &= n^2 \sum_{z \in K_n \cap \Lambda} \frac{1}{n^2} \text{Area}(K_n \cap R_{\pi/2}(z, K_n)) + O(n^{3/2}) \\ &= n^2 \sum_{z \in K_n \cap \Lambda} \frac{1}{n} \text{Area}\left(\frac{1}{\sqrt{n}}(K_n \cap R_{\pi/2}(z, K_n))\right) + O(n^{3/2}) \\ &= n^2 \sum_{z \in K_n \cap \Lambda} \frac{1}{n} \text{Area}\left(\frac{1}{\sqrt{n}} K_n \cap R_{\pi/2}\left(\frac{z}{\sqrt{n}}, \frac{1}{\sqrt{n}} K_n\right)\right) + O(n^{3/2}). \end{aligned}$$

The last sum is a Riemann approximation for the function $f_{(1/\sqrt{n})K_n}$ over the region $(1/\sqrt{n})K_n$; thus

$$S_\triangle(K_n \cap \Lambda) = n^2 \left(\int_{\frac{1}{\sqrt{n}} K_n} f_{\frac{1}{\sqrt{n}} K_n}(z)\, dz + O\left(\frac{1}{\sqrt{n}}\right)\right) + O(n^{3/2}).$$

Since

$$\begin{aligned} \text{Area}\left(\frac{1}{\sqrt{n}} K_n\right) &= \frac{1}{n} \text{Area}(K_n) = \frac{1}{n}(n + O(\sqrt{n})) \\ &= 1 + O\left(\frac{1}{\sqrt{n}}\right) = \text{Area}(K) + O\left(\frac{1}{\sqrt{n}}\right), \end{aligned}$$

it follows that

$$\int_{\frac{1}{\sqrt{n}} K_n} f_{\frac{1}{\sqrt{n}} K_n}(z)\, dz = \int_K f_K(z)\, dz + O\left(\frac{1}{\sqrt{n}}\right).$$

As a result,

$$S_\triangle(K_n \cap \Lambda) = n^2 \int_{\frac{1}{\sqrt{n}} K_n} f_{\frac{1}{\sqrt{n}} K_n}(z)\, dz + O(n^{3/2}) = n^2 \int_K f_K(z)\, dz + O(n^{3/2}). \quad \square$$

The importance of this theorem can be seen immediately. Although our lower bound of $\frac{5}{12}$ for $\liminf_{n\to\infty} S_\triangle(n)/n^2$ was derived by summing the degrees of each point in a square lattice, the same result can be obtained by letting K be the square $\{(x, y) : |x| \le \frac{1}{2}, |y| \le \frac{1}{2}\}$. It follows that $f_K(x, y) = (1 - |x| - |y|)\big(1 - \big||x| - |y|\big|\big)$ and

$$S_\triangle(K_n \cap \Lambda) = \left(\int_K f_K(z)\, dz\right) n^2 + O(n^{3/2}) = \tfrac{5}{12} n^2 + O(n^{3/2}).$$

An improved lower bound will follow provided that we find a set K such that the value for the integral in Theorem 2 is larger than $\frac{5}{12}$. We get a larger value for the integral by letting K be the circle $\{z \in \mathbb{C} : |z| \le 1/\sqrt{\pi}\}$. In this case

$$f_K(z) = \frac{2}{\pi} \arccos\Big(\frac{\sqrt{2\pi}}{2}|z|\Big) - |z| \sqrt{\frac{2}{\pi} - |z|^2} \tag{2}$$

and

$$S_\triangle(K_n \cap \Lambda) = \left(\int_K f_K(z)\, dz\right) n^2 + O(n^{3/2}) = \left(\frac{3}{4} - \frac{1}{\pi}\right) n^2 + O(n^{3/2}).$$

It was conjectured in [Ábrego and Fernández-Merchant 2000] that not only does $\lim_{n\to\infty} E(n)/n^2$ exist, but it is attained by the uniform lattice in the shape of a circle. ($E(n)$ denotes the maximum number of equilateral triangles determined by n points in the plane.) The corresponding conjecture in the case of the isosceles right triangle turns out to be false. That is, if $\lim_{n\to\infty} S_\triangle(n)/n^2$ exists, then it must be strictly greater than $\frac{3}{4} - \pi^{-1}$. Define $\overline{\Lambda}$ to be the translation of Λ by the vector $(\frac{1}{2}, \frac{1}{2})$. The following lemma will help us to improve our lower bound.

Lemma 3. *If* $(j, k) \in \mathbb{R}^2$ *and* $\Lambda' = \Lambda$ *or* $\Lambda' = \overline{\Lambda}$, *then*

$$R_{\pi/2}((j, k), \Lambda') \cap \Lambda' = \begin{cases} \Lambda' & \text{if } (j, k) \in \Lambda \cup \overline{\Lambda}, \\ \varnothing & \text{else.} \end{cases}$$

Proof. Observe that

$$R_{\pi/2}((j, k), (s, t)) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} s-j \\ t-k \end{pmatrix} + \begin{pmatrix} j \\ k \end{pmatrix} = \begin{pmatrix} k-t+j \\ s-j+k \end{pmatrix}.$$

First suppose $(s, t) \in \Lambda$. Since $s, t \in \mathbb{Z}$, then $(k-t+j, s-j+k) \in \Lambda$ if and only if $k-j \in \mathbb{Z}$ and $k+j \in \mathbb{Z}$. This can only happen when either both j and k are half-integers (i.e., $(j, k) \in \overline{\Lambda}$), or both j and k are integers (i.e., $(j, k) \in \Lambda$). Now suppose $(s, t) \in \overline{\Lambda}$. In this case, because both s and t are half-integers, we conclude

that $(k-t+j, s-j+k) \in \overline{\Lambda}$ if and only if both $k-j \in \mathbb{Z}$ and $k+j \in \mathbb{Z}$. Once again this occurs if and only if $(j,k) \in \Lambda \cup \overline{\Lambda}$. □

Recall that if K denotes the circle of area 1, then $(\frac{3}{4} - \pi^{-1})n^2$ is the leading term of $S_\triangle(K_n \cap \Lambda)$. The previous lemma implies that, if we were to adjoin a point $z \in \mathbb{R}^2$ to $K_n \cap \Lambda$ such that z has half-integer coordinates and is located near the center of the circle formed by the points of $K_n \cap \Lambda$, then $\deg_{\pi/2}(z)$ will approximately equal $|K_n \cap \Lambda|$. We obtain the next theorem by further exploiting this idea.

Theorem 4. $$0.43169 \approx \frac{3}{4} - \frac{1}{\pi} < 0.433064 < \liminf_{n\to\infty} \frac{S_\triangle(n)}{n^2}.$$

Proof. Let K be the circle of area 1 and set $A = K_{m_1} \cap \Lambda$, $B = K_{m_2} \cap \overline{\Lambda}$. Position B so that its points are centered on the circle formed by the points in A (see Figure 1). We let $n = m_1 + m_2 = |A \cup B|$ and $m_2 = x \cdot m_1$, where $0 < x < 1$ is a constant to be determined.

We proceed to maximize the leading coefficient of $S_\triangle(A \cup B)$ as x varies from 0 to 1. By Lemma 3, there cannot exist an IRT whose right-angle vertex lies in A while one $\pi/4$ vertex lies in A and the other lies in B. Similarly, there cannot exist an IRT whose right angle-vertex lies in B while one $\pi/4$ vertex lies in A and the other lies in B. Therefore, each IRT with vertices in $A \cup B$ must fall under one of four cases:

Case 1: All three vertices in A. By Theorem 2, there are $(\frac{3}{4} - \pi^{-1})m_1^2 + O(m_1^{3/2})$ IRTs in this case. Since $m_1 = n/(1+x)$, the number of IRTs in terms of n equals

$$\left(\frac{3}{4} - \frac{1}{\pi}\right)\frac{n^2}{(1+x)^2} + O(n^{3/2}). \tag{3}$$

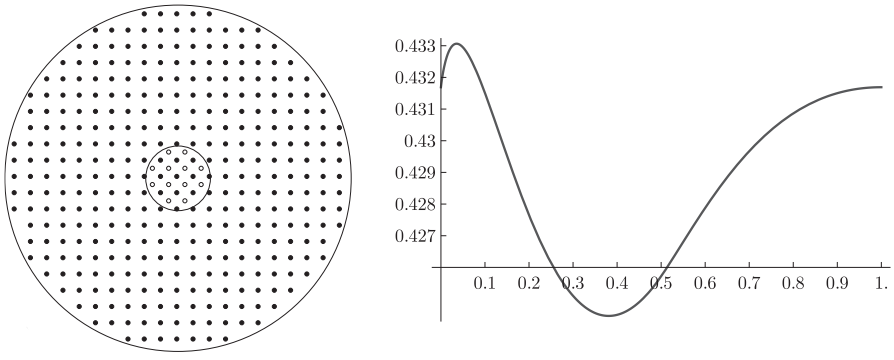


Figure 1. Left: set B (open dots) centered on set A (black dots). Right: plot of the n^2 coefficient of $S_\triangle(A \cup B)$ for x from 0 to 1.

Case 2: All three vertices in B. By Theorem 2, there are $(\frac{3}{4}-\pi^{-1})m_2^2+O(m_2^{3/2})$ IRTs in this case. This time $m_2 = nx/(1+x)$ and the number of IRTs in terms of n equals

$$\left(\frac{3}{4}-\frac{1}{\pi}\right)\frac{n^2x^2}{(1+x)^2}+O(n^{3/2}). \tag{4}$$

Case 3: Right-angle vertex in B, $\pi/4$ vertices in A. The relationship given by Lemma 3 allows us to slightly adapt the proof of Theorem 2 in order to compute the number of IRTs in this case. The integral approximation to the number of IRTs in this case is given by

$$\sum_{z\in K_{m_2}\cap\overline{\Lambda}} \left|(K_{m_1}\cap\Lambda)\cap R_{\pi/2}(z,(K_{m_1}\cap\Lambda))\right| = m_1^2\left(\int_{\frac{1}{\sqrt{m_1}}K_{m_2}} f_{\frac{1}{\sqrt{m_1}}K_{m_1}}(z)\,dz\right)+O(m_1^{3/2}).$$

But

$$\text{Area}\left(\frac{1}{\sqrt{m_1}}K_{m_2}\right) = \text{Area}\left(\sqrt{\frac{m_2}{m_1}}K\right)+O(\sqrt{m_1}),$$

so

$$m_1^2\left(\int_{\frac{1}{\sqrt{m_1}}K_{m_2}} f_{\frac{1}{\sqrt{m_1}}K_{m_1}}(z)\,dz\right)+O(m_1^{3/2}) = m_1^2\left(\int_{\sqrt{\frac{m_2}{m_1}}K} f_K(z)\,dz\right)+O(m_1^{3/2}).$$

Expressing this value in terms of n gives

$$\left(\int_{\sqrt{x}K} f_K(z)\,dz\right)\frac{n^2}{(1+x)^2}+O(n^{3/2}). \tag{5}$$

Case 4: Right-angle vertex in A, $\pi/4$ vertices in B. As in Case 3, the number of IRTs is given by

$$\sum_{z\in K_{m_1}\cap\Lambda} \left|(K_{m_2}\cap\overline{\Lambda})\cap R_{\pi/2}(z,(K_{m_2}\cap\overline{\Lambda}))\right| = m_2^2\left(\int_{\frac{1}{\sqrt{m_2}}K_{m_1}} f_{\frac{1}{\sqrt{m_2}}K_{m_2}}(z)\,dz\right)+O(m_2^{3/2}). \tag{6}$$

Now recall that $f_{(1/\sqrt{m_2})K_{m_2}}(z) = \text{Area}\left((1/\sqrt{m_2})K_{m_2}\cap R_{\pi/2}(z,(1/\sqrt{m_2})K_{m_2})\right)$. It follows that $f_{(1/\sqrt{m_2})K_{m_2}}(z_0)=0$ if and only if z_0 is farther than $\sqrt{2/\pi}$ from the center of $(1/\sqrt{m_2})K_{m_2}$. Thus for small enough values of m_2, the region of integration in (6) is actually $(\sqrt{2/m_2})K_{m_2}$, so it does not depend on m_1. We consider two subcases.

First, if $x\le\frac{1}{2}$ (i.e., $m_2\le m_1/2$), then

$$\sqrt{\frac{2}{\pi}} = \frac{1}{\sqrt{m_2}}\frac{\sqrt{2m_2}}{\sqrt{\pi}} \le \frac{1}{\sqrt{m_2}}\frac{\sqrt{2}}{\sqrt{\pi}}\frac{\sqrt{m_1}}{\sqrt{2}} = \frac{1}{\sqrt{m_2}}\sqrt{\frac{m_1}{\pi}}.$$

The left side of this inequality is the radius of $(\sqrt{2/m_2})K_{m_2}$, while the right side is the radius of $(1/\sqrt{m_2})K_{m_1}$; thus the region of integration where $f_{(1/\sqrt{m_2})K_{m_2}}$ is nonzero equals $(\sqrt{2/m_2})K_{m_2}$. Hence, the number of IRTs equals

$$m_2^2\left(\int_{\sqrt{\frac{2}{m_2}}K_{m_2}} f_{\frac{1}{\sqrt{m_2}}K_{m_2}}(z)\,dz\right)+O(m_2^{3/2}) = m_2^2\left(\int_{\sqrt{2}K} f_K(z)\,dz\right)+O(m_2^{3/2}) \quad (7)$$

$$= \left(\int_{\sqrt{2}K} f_K(z)\,dz\right)\frac{n^2x^2}{(1+x)^2}+O(n^{3/2}).$$

Now we consider the case $x > \frac{1}{2}$ (i.e., $m_2 > m_1/2$). In this case, $f_{(1/\sqrt{m_2})K_{m_2}}$ is nonzero for all points in $(1/\sqrt{m_2})K_{m_1}$. Thus the number of IRTs is then

$$m_2^2\left(\int_{\frac{1}{\sqrt{m_2}}K_{m_1}} f_{\frac{1}{\sqrt{m_2}}K_{m_2}}(z)\,dz\right)+O(m_2^{3/2}) = m_2^2\left(\int_{\sqrt{\frac{m_1}{m_2}}K} f_K(z)\,dz\right)+O(m_2^{3/2}) \quad (8)$$

$$= \left(\int_{\sqrt{\frac{1}{x}}K} f_K(z)\,dz\right)\frac{n^2x^2}{(1+x)^2}+O(n^{3/2}).$$

By (2), we have, for $t > 0$,

$$\int_{tK} f_K(z)\,dz = 2\pi\int_0^{t/\sqrt{\pi}}\left(\frac{2}{\pi}\arccos\left(\frac{\sqrt{2\pi}}{2}r\right)-r\sqrt{\frac{2}{\pi}-r^2}\right)r\,dr$$

$$= \frac{1}{2\pi}\left(4t^2\arccos\frac{t}{\sqrt{2}}+2\arcsin\frac{t}{\sqrt{2}}-t(t^2+1)\sqrt{2-t^2}\right).$$

Therefore, putting all four cases together — i.e., expressions (3)–(5), and either (7) or (8) — we obtain that the n^2 coefficient of $S_\triangle(A\cup B)$ equals

$$\frac{1}{4\pi(x+1)^2}\left(8x\arccos\sqrt{\frac{x}{2}}+4\arcsin\sqrt{\frac{x}{2}}+(5\pi-4)x^2+(3\pi-4)\right.$$
$$\left.-2(x+1)\sqrt{2x-x^2}\right)$$

if $0 < x \le \frac{1}{2}$, or

$$\frac{1}{4\pi(x+1)^2}\left(8x\left(\arccos\sqrt{\frac{x}{2}}+\arccos\sqrt{\frac{1}{2x}}\right)+4\arcsin\sqrt{\frac{x}{2}}+4x^2\arcsin\sqrt{\frac{1}{2x}}\right.$$
$$\left.+(3\pi-4)(x^2+1)-2(x+1)\left(\sqrt{2x-x^2}+\sqrt{2x-1}\right)\right),$$

if $\frac{1}{2} < x < 1$. Letting x vary from 0 to 1, this coefficient is maximized (see Figure 1) when $x \approx 0.0356067$, corresponding to a radius of B approximately 18.87% of the radius of A. Letting x equal this value gives 0.433064 as a decimal approximation to the maximum value attained by the n^2 coefficient. □

At this point, one might be tempted to further increase the quadratic coefficient by placing a third set of lattice points arranged in a circle and centered on the circle formed by B. It turns out that forming such a configuration does not improve the results in the previous theorem. This is due to Lemma 3. More specifically, given our construction from the previous theorem, there is no place to adjoin a point z to the center of $A \cup B$ such that $z \in \Lambda$ or $z \in \overline{\Lambda}$. Hence, if we were to add the point z to the center of $A \cup B$, then any new IRTs would have their right-angle vertex located at z with one $\pi/4$ vertex in A and the other $\pi/4$ vertex in B. Doing so can produce at most $2m_2 = 2xm_1 \approx 0.0712m_1$ new IRTs (recall that $x \approx 0.0356066$ in our construction). On the other hand, adding z to the perimeter of A, gives us $m_1 f_K(1/\sqrt{\pi}) \approx 0.1817m_1$ new IRTs.

3. Upper bound

We now turn our attention to finding an upper bound for $S_\triangle(n)/n^2$. It is easy to see that $S_\triangle(n) \leq n^2 - n$, since any pair of points can be the vertices of at most six IRTs. Our next theorem improves this bound. The idea is to prove that there exists a point in P that does not belong to many IRTs. First, we need the following definition.

For every $z \in P$, let $R^+_{\pi/4}(z, P)$ and $R^-_{\pi/4}(z, P)$ be the dilations of P, centered at z, with factor $\sqrt{2}$ and $1/\sqrt{2}$, respectively, followed by a $\pi/4$ counterclockwise rotation with center z. Let $\deg^+_{\pi/4}(z)$ and $\deg^-_{\pi/4}(z)$ be the number of isosceles right triangles zxy with $x, y \in P$ such that zxy occur in counterclockwise order, and zy, respectively zx, is the hypotenuse of the triangle zxy.

Much like the case of $\deg_{\pi/2}$, $\deg^+_{\pi/4}$ and $\deg^-_{\pi/4}$ can be computed with the identities

$$\deg^+_{\pi/4}(z) = \left|P \cap R^+_{\pi/4}(z, P)\right| - 1, \quad \deg^-_{\pi/4}(z) = \left|P \cap R^-_{\pi/4}(z, P)\right| - 1.$$

Theorem 5. $S_\triangle(n) \leq \left\lfloor \frac{2}{3}(n-1)^2 - \frac{5}{3} \right\rfloor$ *for* $n \geq 3$.

Proof. By induction on n. If $n = 3$, then $S_\triangle(3) \leq 1 = \left\lfloor \frac{1}{3}(2{\cdot}4-5) \right\rfloor$. Now suppose the theorem holds for $n = k$. We must show this implies the theorem holds for $n = k+1$. Suppose that there is a point $z \in P$ such that $\deg_{\pi/2}(z) + \deg^+_{\pi/4}(z) + \deg^-_{\pi/4}(z) \leq \left\lfloor \frac{1}{3}(4n-5) \right\rfloor$. Then, by induction,

$$\begin{aligned} S_\triangle(k+1) &\leq \deg_{\pi/2}(z) + \deg^+_{\pi/4}(z) + \deg^-_{\pi/4}(z) + S_\triangle(k) \\ &\leq \left\lfloor \tfrac{1}{3}(4k-1) \right\rfloor + \left\lfloor \tfrac{2}{3}(k-1)^2 - \tfrac{5}{3} \right\rfloor = \left\lfloor \tfrac{2}{3}k^2 - \tfrac{5}{3} \right\rfloor. \end{aligned}$$

The last equality can be verified by considering the three possible residues of k when divided by 3. Hence, our theorem is proved if we can find a point $z \in P$ with the desired property.

Let $x, y \in P$ be points such that x and y form the diameter of P. In other words, if $w \in P$, then the distance from w to any other point in P is less than or equal to the distance from x to y. We now prove that either x or y is a point with the desired property mentioned above. We begin by analyzing $\deg^-_{\pi/4}$. We use the notation from [Ábrego and Fernández-Merchant 2000, Theorem 1].

Define $N_x = P \cap R^-_{\pi/4}(x, P)\backslash\{x\}$ and $N_y = P \cap R^-_{\pi/4}(y, P)\backslash\{y\}$. It follows from our identities that, $\deg^-_{\pi/4}(x) = |N_x|$ and $\deg^-_{\pi/4}(y) = |N_y|$. Furthermore, by the inclusion-exclusion principle for finite sets, we have

$$|N_x| + |N_y| = |N_x \cup N_y| + |N_x \cap N_y|.$$

We shall prove by contradiction that $|N_x \cap N_y| \leq 1$. Suppose that there are two points $u, v \in N_x \cap N_y$. This means that there are points $u_x, v_x, u_y, v_y \in P$ such that the triangles xu_xu, xv_xv, yu_yu, yv_yv are IRTs oriented counterclockwise with right angle at either u or v.

But notice that the line segments u_xu_y and v_xv_y are simply the $(\pi/2)$-counterclockwise rotations of xy about centers u and v, respectively. Hence, $u_xu_yv_xv_y$ is a parallelogram with two sides having length xy as shown in Figure 2, left. This is a contradiction since one of the diagonals of the parallelogram is longer than any of its sides. Thus, $|N_x \cap N_y| \leq 1$. Furthermore, $x \notin N_y$ and $y \notin N_x$, so $|N_x \cup N_y| \leq n-2$ and thus $\deg^-_{\pi/4}(x) + \deg^-_{\pi/4}(y) = |N_x \cup N_y| + |N_x \cap N_y| \leq n - 2 + 1 = n - 1$. This also implies that $\deg^+_{\pi/4}(x) + \deg^+_{\pi/4}(y) \leq n-1$, since we can follow the same argument applied to the reflection of P about the line xy.

We now look at $\deg_{\pi/2}(x)$ and $\deg_{\pi/2}(y)$. We claim that, for every $p \in P$, at most one of $R_{\pi/2}(x, p)$ or $R_{\pi/2}(y, p)$ belongs to P. Indeed, let $p_x = R_{\pi/2}(x, p)$ and $p_y = R_{\pi/2}(y, p)$ (see Figure 2, right). The distance p_xp_y is exactly the distance xy but scaled by $\sqrt{2}$. This contradicts the fact that xy is the diameter of P.

Define a graph G with vertex set $V(G) = P\backslash\{x, y\}$ and edge set given by saying that $uv \in E(G)$ if and only if $v = R_{\pi/2}(x, u)$ or $v = R_{\pi/2}(y, u)$. We show that

$$0 \leq \deg_{\pi/2}(x) + \deg_{\pi/2}(y) - |E(G)| \leq 1. \tag{9}$$

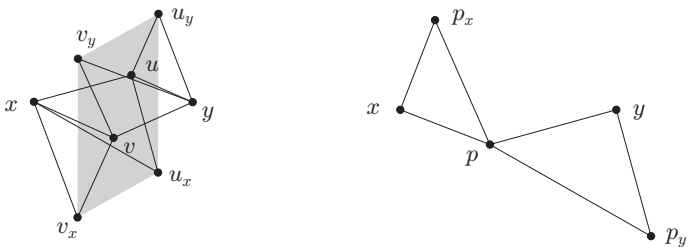


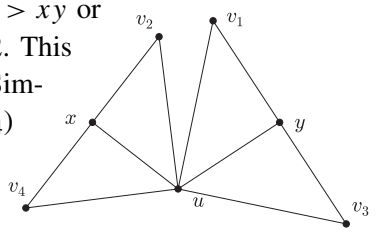
Figure 2. Proof of Theorem 5.

The left inequality follows from the fact each edge counts an IRT in either $\deg_{\pi/2}(x)$ or $\deg_{\pi/2}(y)$ and possibly in both. However, if uv is an edge of G so that $v = R_{\pi/2}(x, u)$ and $u = R_{\pi/2}(y, v)$, then $xuyv$ is a square, so this can only happen for at most one edge.

Now, let $\deg_G(u)$ be the number of edges in $E(G)$ incident to u. We claim that

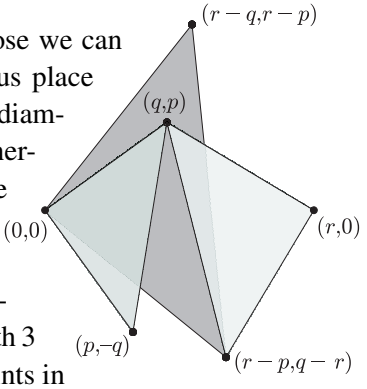
$$\deg_G(u) \leq 2 \quad \text{for every } u \in V(G). \tag{10}$$

Indeed, take $uv_1 \in E(G)$; we can assume, without loss, that $u = R_{\pi/2}(y, v_1)$. If $v_3 = R_{\pi/2}(y, u) \in P$, then we conclude that $xv_3 > xy$ or $xv_1 > xy$, because $\angle xyv_3 \geq \pi/2$ or $\angle xyv_1 \geq \pi/2$. This contradicts the fact that xy is the diameter of P. Similarly, if v_2 and v_4 are defined as $u = R_{\pi/2}(x, v_4)$ and $v_2 = R_{\pi/2}(x, u)$, then at most one of v_2 or v_4 can be in P.



Claim. *All paths in G have length at most* 2.

Proof. We prove this claim by contradiction. Suppose we can have a path of length 3 or more. To assist us, let us place our points on a cartesian coordinate system with our diameter xy relabeled as the points $(0, 0)$ and $(r, 0)$, furthermore, assume $p, q \geq 0$ and that the four vertices of the path of length 3 are $(p, -q)$, (q, p), $(r-p, q-r)$, and $(r-q, r-p)$. Our aim is to show that the distance between $(r-q, r-p)$ and $(r-p, q-r)$ contradicts that r is the diameter of P. Now, if paths of length 3 were possible, the distance between every pair of points in the figure on the right must be less than or equal to r. Since $d((p, -q), (q, p)) \leq r$, we have $p^2 + q^2 \leq r^2/2$.



Now let us analyze the square of the distance from $(r-q, r-p)$ to $(r-p, q-r)$. Because $2(p^2+q^2) \geq (p+q)^2$, it follows that

$$\begin{aligned} d^2\big((r-q, r-p), (r-p, q-r)\big) &= (-q+p)^2 + (2r-p-q)^2 \\ &= 4r^2 - 4r(p+q) + 2(p^2+q^2) \\ &\geq 4r^2 - 4\sqrt{2}r\sqrt{p^2+q^2} + 2(p^2+q^2) \\ &= \big(2r - \sqrt{2(p^2+q^2)}\big)^2. \end{aligned}$$

But $\sqrt{2(p^2+q^2)} \leq r$, so $\big(2r - \sqrt{2(p^2+q^2)}\big) \geq r$ and thus

$$d^2\big((r-q, r-p), (r-p, q-r)\big) \geq r^2.$$

Equality occurs if and only if $p = r/2$ and $q = r/2$; otherwise, the distance between

$(r-q, r-p)$ and $(r-p, q-r)$ is strictly greater than r, contradicting the fact that the diameter of P is r. Therefore if $p \neq r/2$ or $q \neq r/2$ then there is no path of length 3. In the case that $p = r/2$ and $q = r/2$ the points (q, p) and $(r-q, r-p)$ become the same and so do the points $(p, -q)$ and $(r-p, q-r)$. Thus we are left with a path of length 1. □

It follows from (10) and the Claim that all paths of length 2 are disjoint. Thus, G is the union of disjoint paths of length at most 2. If a denotes the number of paths of length 2 and b the number of paths of length 1, then

$$|E(G)| = 2a + b \quad \text{and} \quad 3a + 2b \leq n - 2.$$

Recall from (9) that either

$$\deg_{\pi/2}(x) + \deg_{\pi/2}(y) = |E(G)| \quad \text{or} \quad \deg_{\pi/2}(x) + \deg_{\pi/2}(y) = |E(G)| + 1.$$

If $\deg_{\pi/2}(x) + \deg_{\pi/2}(y) = |E(G)|$, then

$$2\,|E(G)| = 4a + 2b \leq n - 2 + a \leq n - 2 + \frac{n-2}{3},$$

so $\deg_{\pi/2}(x) + \deg_{\pi/2}(y) = |E(G)| \leq \frac{2}{3}(n-2)$. Moreover, if

$$\deg_{\pi/2}(x) + \deg_{\pi/2}(y) = |E(G)| + 1,$$

then $b \geq 1$ and we get a minor improvement,

$$2\,|E(G)| = 4a + 2b \leq n - 2 + a \leq n - 4 + \frac{n-2}{3},$$

so $\deg_{\pi/2}(x) + \deg_{\pi/2}(y) = |E(G)| + 1 \leq (2n-7)/3 < \frac{2}{3}(n-2)$.

We are now ready to put everything together. Between the two points x and y, we derived the bounds:

$$\begin{aligned}
\deg_{\pi/2}(x) + \deg_{\pi/2}(y) &\leq \tfrac{2}{3}(n-2), \\
\deg^{+}_{\pi/4}(x) + \deg^{+}_{\pi/4}(y) &\leq (n-1), \\
\deg^{-}_{\pi/4}(x) + \deg^{-}_{\pi/4}(y) &\leq (n-1).
\end{aligned}$$

Because the degree of a point must take on an integer value, it must be the case that either x or y satisfies $\deg_{\pi/2} + \deg^{+}_{\pi/4} + \deg^{-}_{\pi/4} \leq \lfloor (4n-5)/3 \rfloor$. □

4. Small cases

In this section we determine the exact values of $S_{\triangle}(n)$ when $3 \leq n \leq 9$.

Theorem 6. *For* $3 \leq n \leq 9$, $S_{\triangle}(3) = 1$, $S_{\triangle}(4) = 4$, $S_{\triangle}(5) = 8$, $S_{\triangle}(6) = 11$, $S_{\triangle}(7) = 15$, $S_{\triangle}(8) = 20$, *and* $S_{\triangle}(9) = 28$.

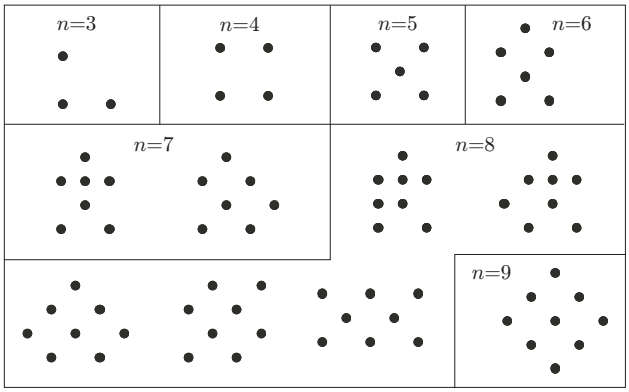


Figure 3. Optimal sets achieving equality for $S_\triangle(n)$.

The corresponding optimal sets are shown in Figure 3.

Proof. We begin with $n = 3$. Since three points uniquely determine a triangle, and there is an IRT with three points, shown in Figure 4(a), this situation becomes trivial and we conclude that $S_\triangle(3) = 1$.

Now let $n = 4$. In Figure 4(b) we show a point set P such that $S_\triangle(P) = 4$. This implies that $S_\triangle(4) \geq 4$. However, $S_\triangle(4)$ is also bounded above by $\binom{4}{3} = 4$. Hence, $S_\triangle(4) = 4$.

To continue with the proof for the remaining values of n, we need the following two lemmas.

Lemma 7. *Suppose* $|P| = 4$ *and* $S_\triangle(P) \geq 2$. *The sets in parts* (b)–(e) *of Figure 4 are the only possibilities for such a set* P, *not counting symmetric repetitions.*

Proof. Having $S_\triangle(P) \geq 2$ implies that we must always have more than one IRT in P. Hence, we can begin with a single IRT and examine the possible ways of adding a point and producing more IRTs. We accomplish this task in Figure 4(a). The 10 numbers in the figure indicate the location of a point, and the total number of IRTs after its addition to the set of black dots. All other locations not labeled with a number do not increase the number of IRTs. Therefore, except for symmetries, all the possibilities for P are shown in Figure 4(b)–(e). □

Lemma 8. *Let* P *be a finite set with* $|P| = n$. *Suppose that* $S_\triangle(A) \leq b$ *for all* $A \subseteq P$ *with* $|A| = k$. *Then*

$$S_\triangle(P) \leq \left\lfloor \frac{n\,(n-1)\,(n-2)\,b}{k\,(k-1)\,(k-2)} \right\rfloor.$$

Proof. Suppose that within P, every k-point configuration contains at most b IRTs. The number of IRTs in P can then be counted by adding all the IRTs in every k-point subset of P. However, in doing so, we end up counting a fixed IRT exactly

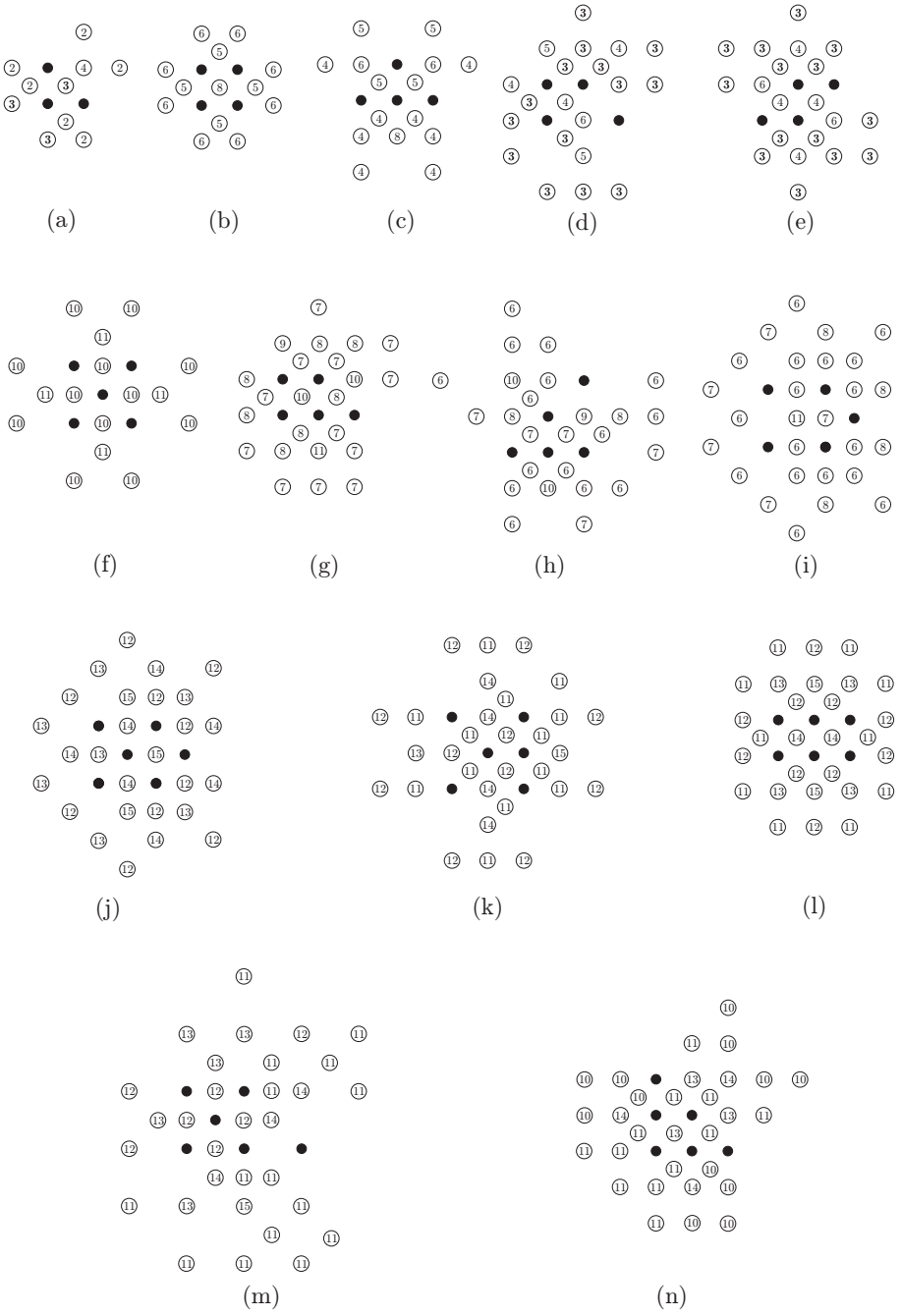


Figure 4. Proof of Theorem 6. Each circle with a number indicates the location of a point and the total number of IRTs resulting from its addition to the base set of black dots.

$\binom{n-3}{k-3}$ times. Because $S_\triangle(A) \le b$ we get,

$$\binom{n-3}{k-3} S_\triangle(P) = \sum_{\substack{A \subseteq P \\ |A|=k}} S_\triangle(A) \le \binom{n}{k} b.$$

Notice that $S_\triangle(P)$ can only take on integer values so,

$$S_\triangle(P) \le \left\lfloor \frac{\binom{n}{k} b}{\binom{n-3}{k-3}} \right\rfloor = \left\lfloor \frac{n(n-1)(n-2)b}{k(k-1)(k-2)} \right\rfloor. \qquad \square$$

Now suppose $|P| = 5$. If $S_\triangle(A) \le 1$ for all $A \subseteq P$ with $|A| = 4$, then by Lemma 8, $S_\triangle(P) \le 2$. Otherwise, by Lemma 7, P must contain one of the four sets shown in Figure 4(b)–(e). The result now follows by examining the possibilities for producing more IRTs by placing a fifth point in the four distinct sets. In Figure 4(b)–(e), we accomplish this task. Just as in Lemma 7, every number in a figure indicates the location of a point, and the total number of IRTs after its addition to the set of black dots. It follows that the maximum value achieved by placing a fifth point is 8 and so $S_\triangle(5) = 8$. The point set that uniquely achieves equality is shown in Figure 4(f). Moreover, there is exactly one set P with $S_\triangle(P) = 6$ (shown in Figure 4(g)), and two sets P with $S_\triangle(P) = 5$ (Figures 4(h) and 4(i)).

Now suppose $|P| = 6$. If $S_\triangle(A) \le 4$ for all $A \subseteq P$ with $|A| = 5$, then by Lemma 8, $S_\triangle(P) \le 8$. Otherwise, P must contain one of the sets in Figure 4(f)–(i). We now check all possibilities for adding more IRTs by joining a sixth point to our four distinct sets. This is shown in Figure 4(f)–(i). It follows that the maximum value achieved is 11 and so $S_\triangle(6) = 11$. The point set that uniquely achieves equality is shown in Figure 4(j). Also, except for symmetries, there are exactly three sets P with $S_\triangle(P) = 10$ (Figure 4(k)–(m)) and only one set P with $S_\triangle(P) = 9$ (Figure 4(n)).

Now suppose $|P| = 7$. If $S_\triangle(A) \le 8$ for all $A \subseteq P$ with $|A| = 6$, we have $S_\triangle(P) \le 14$, by Lemma 8. Otherwise, P must contain one of the sets in parts (j)–(n) of Figure 4. We now check all possibilities for adding more IRTs by joining a seventh point to our 5 distinct configurations. We complete this task in parts (j)–(n). Because the maximum value achieved is 15, we deduce that $S_\triangle(7) = 15$. In this case, there are exactly two point sets that achieve 15 IRTs.

The proof for the values $n = 8$ and $n = 9$ follows along the same lines, but there are many more intermediate sets to be considered. We omit the details. $\square$

Inspired by our method used for proving exact values of $S_\triangle(n)$, a computer algorithm was devised to construct the best one-point extension of a given base set. This algorithm, together with appropriate heuristic choices for some initial sets, led to the construction of point sets with many IRTs giving us our best lower

n	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
bound	35	43	52	64	74	85	97	112	124	139	156	176	192	210	229	252

Table 1. Best lower bounds for $S_\triangle(n)$.

bounds for $S_\triangle(n)$ when $10 \le n \le 25$. These lower bounds are shown in Table 1 and the point sets achieving them in Figure 5.

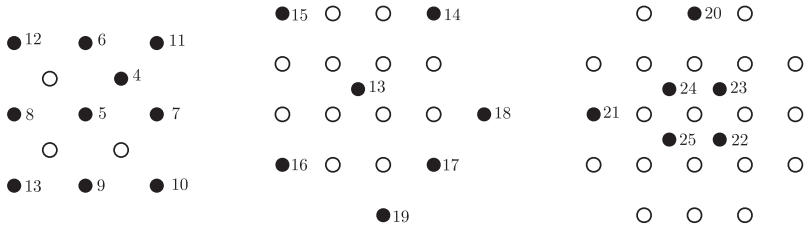


Figure 5. Best constructions A_n for $n \le 25$. Each set A_n is obtained as the union of the starting set (in white) and the points with label $\le n$. The value $S_\triangle(A_n)$ is given by Table 1.

Acknowledgements

We thank Virgilio Cerna who, as part of the CURM mini-grant that supported this project, helped to implement the program that found the best lower bounds for smaller values of n. We also thank an anonymous referee for some useful suggestions and improvements to the presentation.

References

[Ábrego and Fernández-Merchant 2000] B. M. Ábrego and S. Fernández-Merchant, "On the maximum number of equilateral triangles. I", *Discrete Comput. Geom.* **23**:1 (2000), 129–135. MR 2000i:05006 Zbl 0963.52007

[Brass 2002] P. Brass, "Combinatorial geometry problems in pattern recognition", *Discrete Comput. Geom.* **28**:4 (2002), 495–510. MR 2003m:68114 Zbl 1011.68117

[Brass and Pach 2005] P. Brass and J. Pach, "Problems and results on geometric patterns", pp. 17–36 in *Graph theory and combinatorial optimization*, edited by D. Avis et al., GERAD 25th Anniv. Ser. **8**, Springer, New York, 2005. MR 2006f:05051 Zbl 1083.68110

[Brass et al. 2005] P. Brass, W. Moser, and J. Pach, *Research problems in discrete geometry*, Springer, New York, 2005. MR 2006i:52001 Zbl 1086.52001

[Elekes and Erdős 1994] G. Elekes and P. Erdős, "Similar configurations and pseudo grids", pp. 85–104 in *Intuitive geometry* (Szeged, 1991), edited by G. F. Tâoth and K. Bèorèoczky, Colloq. Math. Soc. János Bolyai **63**, North-Holland, Amsterdam, 1994. MR 97b:52020

[Erdős and Purdy 1971] P. Erdős and G. Purdy, "Some extremal problems in geometry", *J. Combinatorial Theory Ser. A* **10** (1971), 246–252. MR 43 #1045 Zbl 0219.05006

[Erdős and Purdy 1975] P. Erdős and G. Purdy, "Some extremal problems in geometry, III", *Congressus Numerantium* **14** (1975), 291–308. MR 52 #13650 Zbl 0328.05018

[Erdős and Purdy 1976] P. Erdős and G. Purdy, "Some extremal problems in geometry, IV", *Congressus Numerantium* **17** (1976), 307–322. MR 55 #10292 Zbl 0345.52007

[Laczkovich and Ruzsa 1997] M. Laczkovich and I. Z. Ruzsa, "The number of homothetic subsets", pp. 294–302 in *The mathematics of Paul Erdős, II*, edited by R. Graham and J. Nesetril, Algorithms Combin. **14**, Springer, Berlin, 1997. MR 98e:52017 Zbl 0871.52012

Received: 2010-01-14 Revised: 2011-02-27 Accepted: 2011-02-27

bernardo.abrego@csun.edu *Department of Mathematics, California State University, 18111 Nordhoff Street, Northridge, CA 91330-8313, United States*
http://www.csun.edu/~ba70714

silvia.fernandez@csun.edu *Department of Mathematics, California State University, 18111 Nordhoff Street, Northridge, CA 91330-8313, United States*
http://www.csun.edu/~sf70713

david.roberts.0@csun.edu *Department of Mathematics, California State University, 18111 Nordhoff Street, Northridge, CA 91330-8313, United States*

INVOLVE 4:1(2011)

msp

Stability properties of a predictor-corrector implementation of an implicit linear multistep method

Scott Sarra and Clyde Meador

(Communicated by John Baxley)

We examine the stability properties of a predictor-corrector implementation of a class of implicit linear multistep methods. The method has recently been described in the literature as suitable for the efficient integration of stiff systems and as having stability regions similar to well known implicit methods. A more detailed analysis reveals that this is not the case.

1. Introduction

In an undergraduate research project that started as a senior capstone project, Meador [2009] became aware of an explicit ODE method that claimed to have desirable stability properties that are usually only enjoyed by implicit methods. The little known method seemed too good to be true. If it had the claimed stability properties, it deserved to be better known and more widely used in applications. In this work we describe what a more careful study of the method revealed. We calculate the correct stability regions of the methods and verify our claims with numerical experiments.

2. Linear multistep methods

A general s-step linear multistep method (LMM) for the numerical solution of the autonomous ordinary differential equation (ODE) initial value problem (IVP)

$$y' = F(y), \qquad y(0) = y_0 \tag{1}$$

is of the form

$$\sum_{m=0}^{s} \alpha_m y^{n+m} = \Delta t \sum_{m=0}^{s} \beta_m F(y^{n+m}), \qquad n = 0, 1, \ldots, \tag{2}$$

MSC2000: 65L04, 65L06, 65L20.

Keywords: linear multistep method, eigenvalue stability, numerical differential equations, stiffness.

where α_m and β_m are given constants. It is conventional to normalize (2) by setting $\alpha_s = 1$. When $\beta_s = 0$ the method is explicit. Otherwise, it is implicit. In order to start multistep methods, the first $s - 1$ time levels have to be calculated by a one-step method such as a Runge–Kutta method. Many of the properties of the method (2) can be described in terms of the characteristic polynomials

$$\rho(\omega) = \sum_{m=0}^{s} \alpha_m \omega^s \qquad \text{and} \qquad \sigma(\omega) = \sum_{m=0}^{s} \beta_m \omega^s. \tag{3}$$

The linear stability region of a numerical ODE method is determined by applying the method to the scalar linear equation

$$y' = \lambda y, \qquad y(0) = 1, \tag{4}$$

where λ is a complex number. The exact solution of (4) is $y(t) = e^{\lambda t}$, which approaches zero as $t \to \infty$ if and only if the real part of λ is negative. The set of all numbers $z = \Delta t \lambda$ such that $\lim_{n\to\infty} y^n = 0$ is called the linear stability region of the method. For z in the stability domain, the numerical method exhibits the same asymptotic behavior as (4). For stability, all the scaled eigenvalues of the coefficient matrix of a linear system of ODEs must lie in the stability region. For nonlinear systems, the scaled eigenvalues of the Jacobian matrix of the system must lie within the stability region. A numerical ODE method is A-stable if its region of absolute stability contains the entire left half-plane ($\mathrm{Re}(\Delta t \lambda) < 0$).

For LMMs, the boundary of the stability region is found by the boundary locus method which plots the parametric curve of the function

$$r(\theta) = \frac{\rho(e^{i\theta})}{\sigma(e^{i\theta})}, \qquad 0 \le \theta \le 2\pi, \tag{5}$$

that is, the ratio of the method's characteristic polynomials (3). Standard references on numerical ODEs can be consulted for more details [Butcher 2003; Hairer et al. 2000; Hairer and Wanner 2000; Iserles 1996; Lambert 1973]

3. Implicit LIL linear multistep methods

In this work we consider a class of LMM that has been referred to as local iterative linearization (LIL) in the literature. The s-stage implicit LIL method also has accuracy of order s. The LIL method has been applied to chaotic dynamical systems in [Danca and Chen 2004; Luo et al. 2007]. The convergence, accuracy, and stability properties of the LIL methods were examined in [Danca 2006].

In [Danca and Chen 2004; Danca 2006; Luo et al. 2007], both the implicit and predictor-corrector versions are referred to as LIL methods. However, the stability properties of the methods are very different and we distinguish between

the methods by calling the implicit method ILIL, and the predictor-corrector implementation PCLIL.

Using the notation $f^n = F(y^n)$, the first four ILIL formulas follow. The $s = 1$ ILIL formula

$$y^{n+1} - y^n = \Delta t f^{n+1} \tag{6}$$

coincides with the implicit Euler method. For $s = 2$ the ILIL algorithm is

$$y^{n+2} - \frac{4}{3}y^{n+1} + \frac{1}{3}y^n = \Delta t \left(\frac{25}{36} f^{n+2} - \frac{1}{18} f^{n+1} + \frac{1}{36} f^n \right); \tag{7}$$

for $s = 3$,

$$y^{n+3} - \frac{5}{3}y^{n+2} + \frac{13}{15}y^{n+1} - \frac{1}{5}y^n = \Delta t \left(\frac{26}{45} f^{n+3} - \frac{1}{9} f^{n+2} + \frac{4}{45} f^{n+1} - \frac{1}{45} f^n \right); \tag{8}$$

and for $s = 4$,

$$y^{n+4} - 2y^{n+3} + \frac{8}{5}y^{n+2} - \frac{26}{35}y^{n+1} + \frac{1}{7}y^n = \Delta t \left(\frac{6463}{12600} f^{n+4} - \frac{523}{3150} f^{n+3} + \frac{383}{2100} f^{n+2} - \frac{283}{3150} f^{n+1} + \frac{223}{12600} f^n \right). \tag{9}$$

The characteristic polynomial coefficients of the ILIL methods are listed in Table 1. The stability regions of the ILIL methods of orders 1 through 4 are shown in Figure 1 (left). The stability regions are exterior to the curves. The innermost curve is associated with the first-order method and the stability region shrinks as the order of the method increases. The first- and second-order methods are A-stable, while the third and fourth-order methods do not include all of the left half-plane. It is well known that the order of an A-stable LMM cannot exceed 2 [Lambert 1973].

	$s=1$	$s=2$	$s=3$	$s=4$
α_0/β_0	$-1/0$	$\frac{1}{3}/\frac{1}{36}$	$\frac{-1}{5}/\frac{-1}{45}$	$\frac{1}{7}/\frac{223}{12600}$
α_1/β_1	$1/1$	$\frac{-4}{3}/\frac{-1}{18}$	$\frac{13}{15}/\frac{4}{45}$	$\frac{-26}{35}/\frac{-283}{3150}$
α_2/β_2	-	$1/\frac{25}{36}$	$\frac{-5}{3}/\frac{-1}{9}$	$\frac{8}{5}/\frac{383}{2100}$
α_3/β_3	-	-	$1/\frac{26}{45}$	$-2/\frac{-523}{3150}$
α_4/β_4	-	-	-	$1/\frac{6463}{12600}$

Table 1. Coefficients of the characteristic polynomials (3) for the ILIL algorithms.

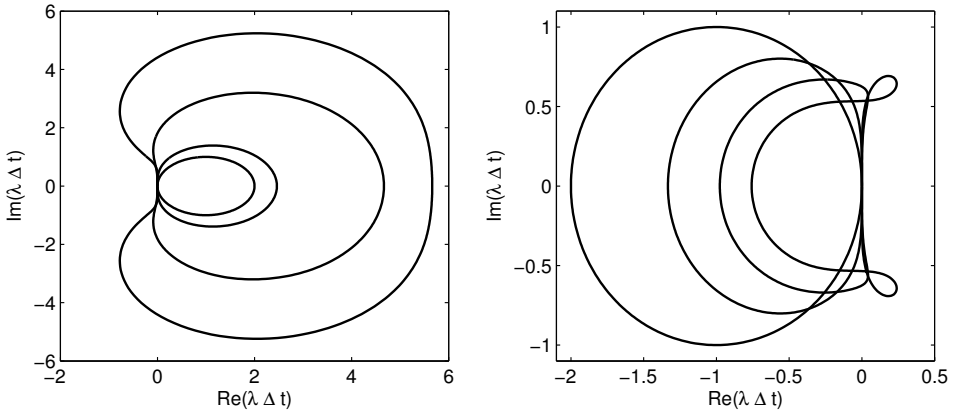


Figure 1. Left: Implicit LIL methods have stability regions consisting of the exterior of the plotted curves. Right: Predictor-corrector implemented LIL methods have bounded stability regions in the interior of the plotted curves.

4. LIL predictor-corrector

Two types of methods that are commonly used to solve the nonlinear difference equations of implicit methods are functional iteration and Newton's method. A third approach, which does not involve solving nonlinear equations, that can be used to implement an implicit ODE method is a predictor-corrector approach. An explicit formula, the predictor, is used to get a preliminary approximation $\hat{y}^{n+s}$ of y^{n+s}. Then the corrector step uses formulas like the implicit LIL methods (6)–(9), with $\hat{y}^{n+s}$ in place of y^{n+s} when calculating f^{n+s}, to get a more accurate approximation of y^{n+s}. The predictor-corrector approach turns the implicit method into one that is implemented in the manner of an explicit method. However, the stability properties of the predictor-corrector method will be inferior to those of the original implicit method. The predictors for the PCLIL methods are listed in Table 2.

s	order s LIL predictor
1	$\hat{y}^{n+1} = y^n$
2	$\hat{y}^{n+2} = 2y^{n+1} - y^n$
3	$\hat{y}^{n+3} = 3y^{n+2} - 3y^{n+1} + y^n$
4	$\hat{y}^{n+4} = 4y^{n+3} - 6y^{n+2} + 4y^{n+1} - y^n$

Table 2. The predictor stages for the predictor-corrector LIL algorithms.

Applying the PCLIL methods to the stability test problem (4) reveals that the α coefficients of the characteristic polynomial (3) remain the same as the implicit LIL methods. However, the β coefficients are modified to be $\hat{\beta}$ which lead to different stability regions. The $\hat{\beta}$ coefficients for the PCLIL methods are listed in Table 3. The details of finding the $\hat{\beta}$ coefficients are illustrated with the second-order PCLIL method:

$$\begin{aligned}\alpha_2 y^{n+2} + \alpha_1 y^{n+1} + \alpha_0 y^n &= \Delta t\left(\beta_2 f^{n+2} + \beta_1 f^{n+1} + \beta_0 f^n\right)\\ &= \Delta t\left(\beta_2\lambda(2y^{n+1} - y^n) + \beta_1\lambda y^{n+1} + \beta_0\lambda y^n\right)\\ &= \Delta t\left((\beta_1 + 2\beta_2)\lambda y^{n+1} + (\beta_0 - \beta_2)\lambda y^n\right)\\ &= \Delta t\left(\hat{\beta}_1 f^{n+1} + \hat{\beta}_0 f^n\right).\end{aligned}$$

The stability regions for the PCLIL methods of orders 1 through 4 are shown in the right image of Figure 1. Since the stability regions consist of the regions that are interior to the curves, PCLIL methods are not A-stable. It is well known that A-stable explicit LMMs do not exist [Nevanlinna and Sipilä 1974].

5. Numerical examples

Many problems arising from various fields result in systems of ODEs that have a property called stiffness. A formal definition can be formulated (see [Lambert 1973], for example), but the essence of a stiff problem can be explained by the fact the coefficient matrix of a linear ODE system (or Jacobian matrix of a nonlinear ODE system) has some eigenvalues with large negative real parts. Thus, explicit methods with their bounded stability regions may be required to take much smaller time steps for stability than are necessary for accuracy. Implicit methods, particularly A-stable methods, with their unbounded stability regions are well suited for stiff problems.

	$s=1$	$s=2$	$s=3$	$s=4$
$\hat{\beta}_0$	1	$-\frac{2}{3}$	$\frac{5}{9}$	$\beta_0 - \beta_4$
$\hat{\beta}_1$	0	$\frac{4}{3}$	$-\frac{74}{45}$	$\beta_1 + 4\beta_4$
$\hat{\beta}_2$	-	0	$\frac{73}{45}$	$\beta_2 - 6\beta_4$
$\hat{\beta}_3$	-	-	0	$\beta_3 + 4\beta_4$
$\hat{\beta}_4$	-	-	-	0

Table 3. Modified β coefficients of the characteristic polynomials (3) for the LIL algorithms implemented as predictor-correctors.

Linear example. We consider the linear ODE system

$$\begin{aligned} y_1' &= -21y_1 + 19y_2 - 20y_3, \quad & y_1(0) &= 1, \\ y_2' &= 19y_1 - 21y_2 + 20y_3, \quad & y_2(0) &= 0, \\ y_3' &= 40y_1 - 40y_2 - 40y_3, \quad & y_3(0) &= -1, \end{aligned} \tag{10}$$

which may be considered stiff. The coefficient matrix

$$A = \begin{bmatrix} -21 & 19 & -20 \\ 19 & -21 & 20 \\ 40 & -40 & -40 \end{bmatrix} \tag{11}$$

has eigenvalues $\lambda_1 = -2$, $\lambda_2 = -40 + 40i$, and $\lambda_3 = -40 - 40i$.

In Figure 2 the stability region of the third-order ILIL is the outside of the dashed curve and the stability region of the third-order PCLIL is the interior of solid curve. The eigenvalues of the linear ODE system (10) scaled by $\Delta t = 0.017$ are in the left image and scaled by $\Delta t = 0.012$ in the right image.

The unstable PCLIL solution of the $y_1(t)$ component of the system using $\Delta t = 0.017$ is shown in the left image in Figure 3 and the stable solution using $\Delta t = 0.012$ is shown on the right. The system can be integrated with the implicit LIL methods with any size time step and the method will remain stable.

Note that for linear problems it is possible to derive an explicit expression from the implicit LIL formulas and that an iterative method is not required. For example,

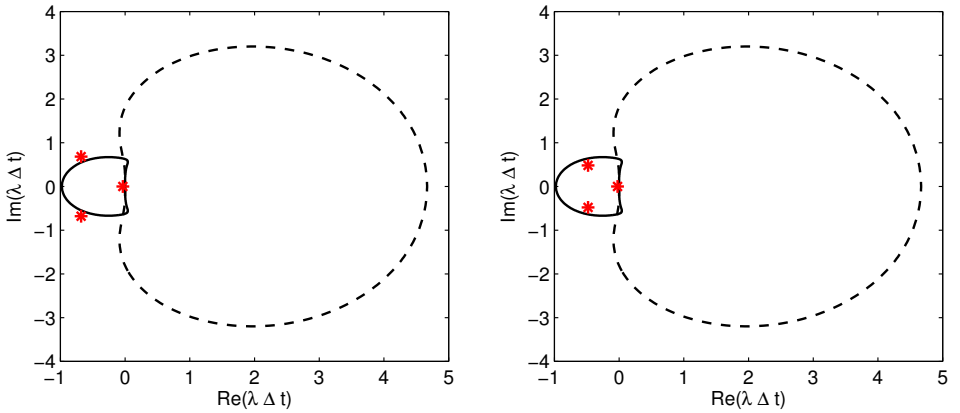


Figure 2. Color dots indicate the eigenvalues of the linear ODE system (10) scaled by $\Delta t = 0.017$ (left) and $\Delta t = 0.012$ (right). The third-order ILIL is stable for eigenvalues outside the dashed curve, and the third-order PCLIL for those inside the solid curve.

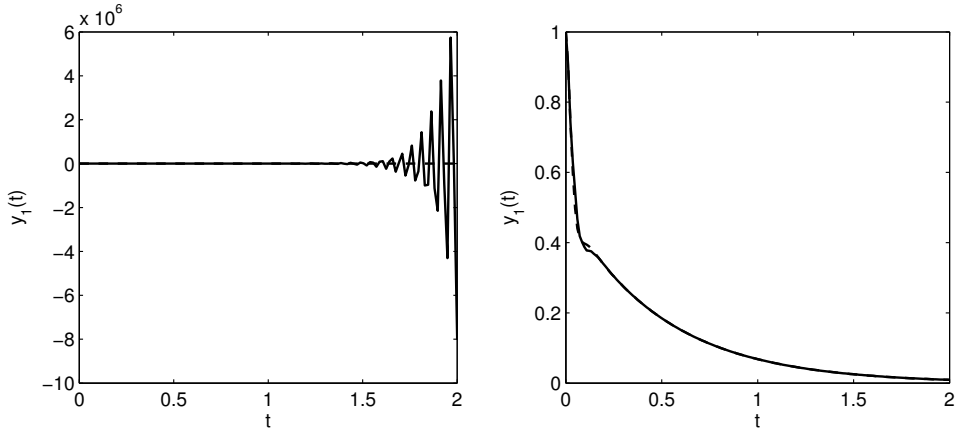


Figure 3. Left: unstable PCLIL solution of the $y_1(t)$ component of the system (10) using $\Delta t = 0.017$. Right: stable solution using $\Delta t = 0.012$.

the second-order implicit LIL method applied to the linear ODE system (10) can be evaluated as

$$y^{n+1} = \Big(I - \frac{25\Delta t}{36}A\Big)^{-1}\Big(\frac{4}{3}I - \frac{\Delta t}{18}A\Big)y^n + \Big(I - \frac{25\Delta t}{36}A\Big)^{-1}\Big(\frac{-1}{3}I - \frac{\Delta t}{36}A\Big)y^{n-1},$$

where I is the 3×3 identity matrix.

Nonlinear example. We consider the Rabinovich–Fabrikant (RF) equations, a set of differential equations in three variables with two constant parameters a and b:

$$\begin{aligned} x' &= y(z - 1 + y^2) + ax, \\ y' &= x(3z + 1 - x^2) + ay, \\ z' &= -2z(b + xy). \end{aligned}$$

PCLIL methods have been used extensively in the study of this system [Danca and Chen 2004; Luo et al. 2007; Danca 2006].

In our numerical work, we encountered severe stability issues while using the PCLIL methods with certain settings of the parameters. For instance, with $a = 0.33$ and $b = 0.5$, a very small step size of $\Delta t = 0.0001$ was needed to stably integrate the system to $t = 200$ with the fourth-order PCLIL method. The resulting attractor is shown in Figure 4. The fourth-order ILIL method was implemented and was an improvement in many cases. However, due to the method not being A-stable, we still had stability problems for some parameter settings.

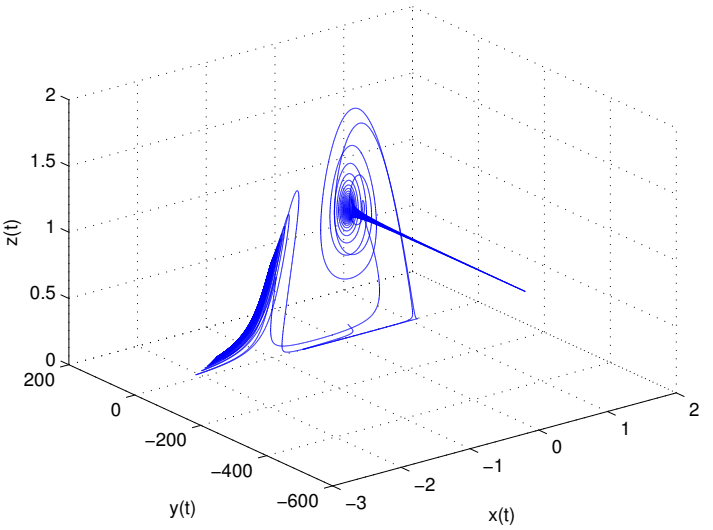


Figure 4. Phase plots of the Rabinovich–Fabrikant equations for parameter settings $a = 0.33$ and $b = 0.5$.

We note that the most efficient method that we found for our numerical exploration of the RF system was an implicit Runge–Kutta method. Using the 4-stage, eighth-order accurate, A-stable Gauss method [Butcher 1964; Ehle 1968; Hairer and Wanner 2000; Sanz-Serna and Calvo 1994], we were able to accurately approximate the attractor in Figure 4 with a step size as large as $\Delta t = 0.2$.

6. Conclusions

Previously, the predictor-corrector implementation of the LIL method has been analyzed in [Danca 2006] where of the PCLIL method it was said that "The time stability of LIL method is more efficient than that of other known algorithms and is comparable with time stability of the Gear's algorithm" and that the LIL method is suitable for stiff problems. Additionally, in [Danca and Chen 2004; Luo et al. 2007] the PCLIL was applied to chaotic dynamical systems that had stiff characteristics and was presented as a method well suited to this type of problem. As we have shown here, this is not the case. The PCLIL methods are explicit and have bounded stability regions that decrease in area as the order of the method increases. The PCLIL methods are not well suited for stiff problems as they will require very small time steps in order to remain stable. It is possible that in the previous application to nonlinear chaotic systems that very small time steps were always used for accuracy purposes and thus stability issues were not encountered.

References

[Butcher 1964] J. C. Butcher, "Implicit Runge–Kutta processes", *Math. Comp.* **18** (1964), 50–64. MR 28 #2641 Zbl 0123.11701

[Butcher 2003] J. C. Butcher, *Numerical methods for ordinary differential equations*, John Wiley & Sons Ltd., Chichester, 2003. MR 2004e:65069 Zbl 1040.65057

[Danca 2006] M.-F. Danca, "A multistep algorithm for ODEs", *Dyn. Contin. Discrete Impuls. Syst. Ser. B Appl. Algorithms* **13**:6 (2006), 803–821. MR 2007k:65097 Zbl 1111.65065

[Danca and Chen 2004] M.-F. Danca and G. Chen, "Bifurcation and chaos in a complex model of dissipative medium", *Internat. J. Bifur. Chaos Appl. Sci. Engrg.* **14**:10 (2004), 3409–3447. MR 2107556 Zbl 1129.37314

[Ehle 1968] B. L. Ehle, "High order A-stable methods for the numerical solution of systems of D.E.'s", *Nordisk Tidskr. Informationsbehandling* (*BIT*) **8** (1968), 276–278. MR 39 #1119

[Hairer and Wanner 2000] E. Hairer and G. Wanner, *Solving ordinary differential equations, II: Stiff and differential-algebraic problems*, Spinger Series in Computational Math. **14**, Springer, 2000.

[Hairer et al. 2000] E. Hairer, S. Norsett, and G. Wanner, *Solving ordinary differential equations, I: Nonstiff problems*, Spinger Series in Computational Math. **8**, Springer, 2000.

[Iserles 1996] A. Iserles, *A first course in the numerical analysis of differential equations*, Cambridge Texts in Applied Mathematics, Cambridge University Press, Cambridge, 1996. MR 1384977 (97m:65003)

[Lambert 1973] J. D. Lambert, *Computational methods in ordinary differential equations*, Wiley, New York, 1973. MR 54 #11789 Zbl 0258.65069

[Luo et al. 2007] X. Luo, M. Small, M.-F. Danca, and G. Chen, "On a dynamical system with multiple chaotic attractors", *Internat. J. Bifur. Chaos Appl. Sci. Engrg.* **17**:9 (2007), 3235–3251. MR 2008k:37081 Zbl 1185.37081

[Meador 2009] C. Meador, "A comparison of two 4th-order numerical ordinary differential equation methods applied to the Rabinovich–Fabrikant equations", 2009, http://www.scottsarra.org/math/papers/ClydeMeador_SeniorCapstone_2009.pdf.

[Nevanlinna and Sipilä 1974] O. Nevanlinna and A. H. Sipilä, "A nonexistence theorem for explicit A-stable methods", *Math. Comp.* **28** (1974), 1053–1056. MR 50 #1515 Zbl 0293.65055

[Sanz-Serna and Calvo 1994] J. M. Sanz-Serna and M. P. Calvo, *Numerical Hamiltonian problems*, Applied Mathematics and Mathematical Computation **7**, Chapman & Hall, London, 1994. MR 95f:65006 Zbl 0816.65042

Received: 2010-03-01 Revised: 2011-03-23 Accepted: 2011-05-07

sarra@marshall.edu *Department of Mathematics, Marshall University, One John Marshall Drive, Huntington, WV 25755-2560, United States* http://www.scottsarra.org/

meador16@marshall.edu *Department of Mathematics, Marshall University, One John Marshall Drive, Huntington, WV 25755-2560, United States*

INVOLVE 4:1(2011)

msp

Five-point zero-divisor graphs determined by equivalence classes

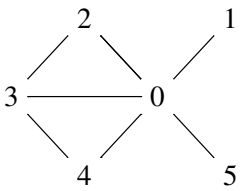
Florida Levidiotis and Sandra Spiroff

(Communicated by Scott Chapman)

We study condensed zero-divisor graphs (those whose vertices are equivalence classes of zero-divisors of a ring R) having exactly five vertices. In particular, we determine which graphs with exactly five vertices can be realized as the condensed zero-divisor graph of a ring. We provide the rings for the graphs which are possible, and prove that the rest of graphs can not be realized via any commutative ring. There are 34 graphs in total which contain exactly five vertices.

1. Introduction

Beck [1988] introduced, for a commutative ring R, a graph whose vertices are the elements of R and whose edges are given by the rule that two vertices r and s share an edge if and only if $rs = 0$. Thus, for the ring $R = \mathbb{Z}/6\mathbb{Z} = \{0, 1, 2, 3, 4, 5\}$, the associated graph is this:



This is by definition a simple graph (no loops or multiple edges) and it is clearly connected with diameter at most two,[1] since all vertices share an edge with 0.

Anderson and Livingston [1999] later introduced the *zero-divisor graph* $\Gamma(R)$ of a commutative R, by taking the subgraph of Beck's graph consisting of all zero-divisors[2] together with the edges they share — in other words, by discarding from

MSC2000: primary 13A99; secondary 05C99.

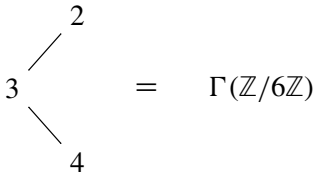
Keywords: condensed zero-divisor graphs, equivalence classes of zero-divisors.

This work is based on Levidiotis' undergraduate honor's thesis project [2010] under the supervision of the second author.

[1] See Definition 2.2 for terms from graph theory.

[2] See Definition 2.1 for terms from ring theory.

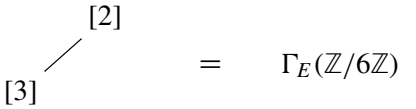
Beck's graph the vertex 0 and all vertices that are not zero-divisors. For instance, the zero-divisors of the ring $\mathbb{Z}/6\mathbb{Z}$ are $\{2, 3, 4\}$, so $\Gamma(\mathbb{Z}/6\mathbb{Z})$ is this graph:



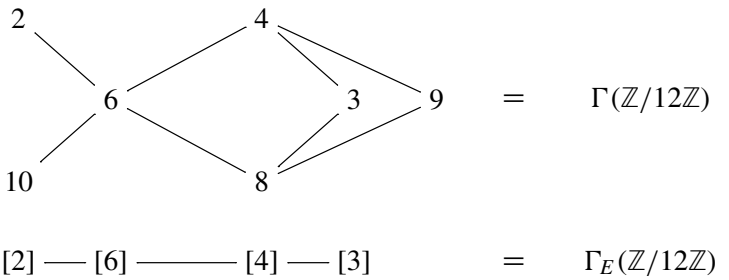
It turns out that the zero-divisor graph, too, is always connected, and its diameter is at most three.

Mulay [2002, (3.5)] demonstrated how a graph $\Gamma_E(R)$ could be constructed from $\Gamma(R)$ by collapsing into equivalence classes zero-divisors that have the same annihilator ideal. Thus, the equivalence class $[r]$ of an element $r \in R$ is the set of zero-divisors s such that $\operatorname{ann}_R(r) = \operatorname{ann}_R(s)$; and such equivalence classes form the vertices of $\Gamma_E(R)$. We call $\Gamma_E(R)$ the *condensed zero-divisor graph* of R. (In [Spiroff and Wickham 2011; Coykendall et al. 2012] the term used was "zero-divisor graph determined by equivalence classes".) Once again, these graphs are simple and connected; the diameter is at most three.

Example 1.1. The equivalence classes of zero-divisors of the ring $R = \mathbb{Z}/6\mathbb{Z}$ are $\{[2], [3]\}$. Note that $\operatorname{ann}_R(2) = \operatorname{ann}_R(4)$, hence $[2] = [4]$.



Example 1.2 [Spiroff and Wickham 2011, Example 1.11]. To illustrate the relation between the zero-divisor graph $\Gamma(R)$ and its condensed counterpart $\Gamma_E(R)$, consider $R = \mathbb{Z}/12\mathbb{Z}$:



To motivate the study of $\Gamma_E(R)$, we provide an additional example. The ring $(\mathbb{Z}/6\mathbb{Z})[X]$, consisting of polynomials in the variable X with coefficients from $\mathbb{Z}/6\mathbb{Z}$, contains infinitely many elements and zero-divisors. However, there are still just two equivalence classes of zero-divisors, and the graph takes the same form as that in Example 1.1.

The goal of this project is to examine the five-point condensed zero-divisor graphs and to determine which of them are possible. This work grew out of [Spiroff and Wickham 2011]; we rely on the results there and provide some answers to questions that arose during that initial study. A subsequent paper [Coykendall et al. 2012] generalizes some of the results in this project.

For those graphs that can be constructed from equivalence classes, we provide an associated ring. For those graphs that can not be constructed from equivalence classes, we prove that no ring exists such that $\Gamma_E(R)$ takes the necessary form. The list of all thirty-four graphs with exactly five vertices can be found in [Harary 1969, pages 216–217]. The connected ones are all shown in this paper at the relevant places, and are labeled (1)–(21).

2. Definitions and basic results

Throughout, R will be a commutative ring with identity that satisfies the ascending chain condition on ideals. A good general reference for the ring theory needed here is [Dummit and Foote 1991]. For zero-divisor graphs, see [Anderson and Livingston 1999].

Definition 2.1. Some definitions from ring theory are collected here:

(1) A *zero divisor* of R is a nonzero element r of R for which there is another nonzero element s of R such that $rs = 0$.

(2) The *annihilator ideal of* r in R, denoted by $\operatorname{ann}_R(r)$, is the set of all elements a in R such that $ar = 0$.

(3) A *unit* in R is a nonzero element u that has a multiplicative inverse; that is, $uu^{-1} = 1$ for some u^{-1} in R.

(4) An ideal J of R is *maximal* if, whenever $J \subseteq I$ for any proper ideal I of R, then $J = I$.

(5) An *equivalence relation* on R is a binary relation $\sim$ that is reflexive, symmetric, and transitive.

Definition 2.2. Some definitions from graph theory are collected here:

(1) A *graph* consists of a set of vertices, a set of edges, and an incidence relation, describing which pairs of vertices are joined by an edge. Two vertices joined by an edge are called *adjacent*.

(2) A *path of length* n between two vertices v and w is a finite sequence of vertices $u_0, u_1, \ldots, u_n$ such that $v = u_0$, $w = u_n$, and u_{i-1} and u_i are adjacent for all $1 \leq i \leq n$.

(3) A graph is said to be *connected* if there is a path between every pair of vertices of the graph.

(4) The *distance* between two vertices v and w in a connected graph is the length of the shortest path between them.

(5) The *diameter* of a connected graph G is the greatest distance between any two vertices.

(6) A graph is said to be *complete* if every vertex in the graph is adjacent to every other vertex in the graph.

Definition 2.3. The condensed zero-divisor graph of a ring R, denoted by $\Gamma_E(R)$, is the graph associated to R whose vertices are the classes of zero-divisors, where a pair of distinct classes $[r], [s]$ is adjacent if and only if $[r] \cdot [s] = 0$, where $[r] \cdot [s] := [rs]$.

Remark 2.4 [Mulay 2002, (3.5)]. Multiplication is well-defined: let $[r_1] = [r_2]$ and $[s_1] = [s_2]$; that is, $\operatorname{ann}_R(r_1) = \operatorname{ann}_R(r_2)$ and $\operatorname{ann}_R(s_1) = \operatorname{ann}_R(s_2)$. Then $r_1 s_1 = 0$ if and only if $s_1 \in \operatorname{ann}_R(r_1) = \operatorname{ann}_R(r_2)$, if and only if $r_2 s_1 = 0$, if and only if $r_2 \in \operatorname{ann}_R(s_1) = \operatorname{ann}_R(s_2)$, if and only if $r_2 s_2 = 0$.

Proposition 2.5 [Mulay 2002, (3.5); Spiroff and Wickham 2011, Propositions 1.4, 1.5, 1.8]. *For any ring R, $\Gamma_E(R)$:*

(a) *is connected*;

(b) *has diameter at most three*;

(c) *is not a cycle graph*; *that is*, *does not take the form of an n-gon*, *for any n*;

(d) *is not complete if it has at least three vertices.*

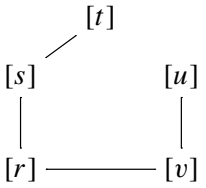
Lemma 2.6. *If u is a unit in R and r is a zero-divisor in R, then* $\operatorname{ann}_R(ur) = \operatorname{ann}_R(r)$.

Proof. If $s \in \operatorname{ann}_R(r)$, then $s(ur) = u(sr) = 0$, hence $s \in \operatorname{ann}_R(ur)$. Conversely, if $s \in \operatorname{ann}_R(ur)$, then $0 = s(ur) = u(sr)$ implies $u^{-1} \cdot 0 = u^{-1} \cdot u(sr)$, and hence $0 = sr$. Thus, $s \in \operatorname{ann}_R(r)$. □

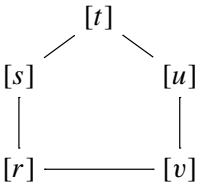
3. Negative results

In this section, we prove that all but four of the five-point graphs can not be realized as the condensed zero-divisor graph of a ring. (Recall that we are assuming that all rings are commutative with identity and satisfy the ascending chain condition on ideals.) By part (a) of Proposition 2.5, only connected graphs need to be considered. By parts (b)–(d) of the same proposition, graphs of types (1)–(3) are not

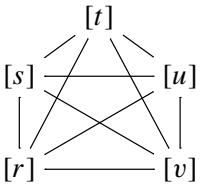
possible:



(1)



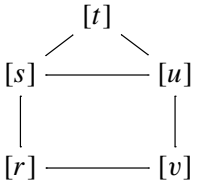
(2)



(3)

The rest of the arguments proceed by contradiction. Namely, we assume that there exists R such that $\Gamma_E(R)$ has exactly the graph in question, which means, in particular, that R has exactly five distinct equivalence classes as represented by the graph. Then from the classes and relations, we show that there must be, in fact, a distinct sixth class, and hence arrive at a contradiction.

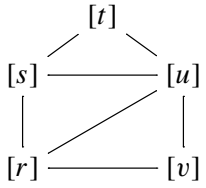
Consider this graph:



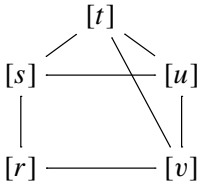
(4)

We show that the element $t+v$ determines a sixth class. First, $t+v$ is annihilated by u, but not by s: indeed, $s(t+v) = 0 + sv \neq 0$, as there is no edge between s and v. Likewise, r does not annihilate $t+v$. However, based on the graph, every class is annihilated by $[r]$ or $[s]$. Thus, $[t+v]$ is not represented by any vertex, and hence must determine a new class.

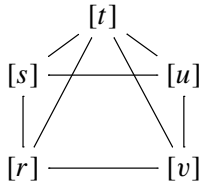
The proofs for graphs (5)–(8) below proceed along the same lines: in (5) and (6), the element $t+v$ determines a new class, and in (7) and (8), the elements $u+v$ and rt determine a new class, respectively.



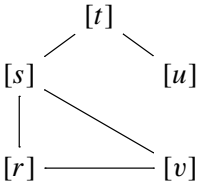
(5)



(6)



(7)

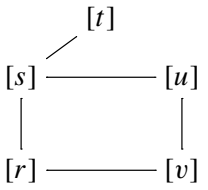


(8)

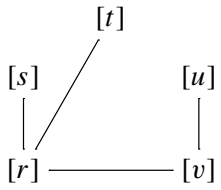
The remaining proofs rely on two key strategies.

Strategy I. *If two points on the condensed zero-divisor graph are adjacent to the same set of vertices, but are not adjacent to one another, then at least one is self-annihilating; otherwise, the two points would represent the same class.*

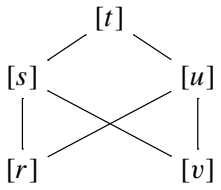
Consider graph (9). One can assume that $r^2 = 0$, else $[r] = [u]$. Then the element $r + v$, which is annihilated by r, but not s or u, determines a new class since, based on the graph, every class is annihilated by $[s]$ or $[u]$. Similarly, for (10), we have $s^2 = 0$, else $[s] = [t]$, hence sv determines a new class. In (11) we have $u^2 = 0$, else $[s] = [u]$, and $v^2 = 0$, else $[r] = [v]$, hence $s + v$ determines a new class.



(9)



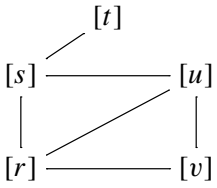
(10)



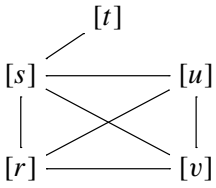
(11)

Strategy II. *If two points on the condensed zero-divisor graph are adjacent to the same set of vertices and are also adjacent to one another, then at least one of the points must not annihilate itself*; *otherwise, the two points would represent the same class.*

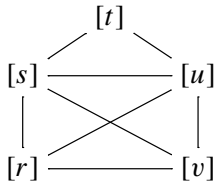
More specifically, in graph (12), one can assume that $r^2 \neq 0$, else $[r] = [u]$. Then the element $r + v$, which is annihilated by u, but not r or s, determines a new class since, based on the graph, every class is annihilated by $[r]$ or $[s]$. Similarly, in (13), $r^2 \neq 0$ and $v^2 \neq 0$, else $[r] = [u] = [v]$; hence $r + v$ determines a new class; and in (14), $r^2 \neq 0$, else $[r] = [v]$ and $s^2 \neq 0$, else $[s] = [u]$; hence $r + s$ determines a new class.



(12)



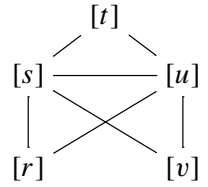
(13)

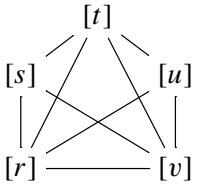


(14)

The proofs for graphs (15) and (16), shown on the next page, use both strategies. In (15), one can assume that $v^2 = 0$, by Strategy I, else $[r] = [v]$, and that $s^2 \neq 0$, by Strategy II, else $[s] = [u]$. Then the element $s + v$, which is annihilated by u and v, but not r, s or t, determines a new class since, based on the graph, every class is annihilated by $[r]$, $[s]$ or $[t]$. Similarly, in (16), one can assume that $u^2 = 0$, else $[s] = [u]$, and that $v^2 \neq 0$, else $[r] = [v]$; hence the element $u + v$ determines a new class.

(15)

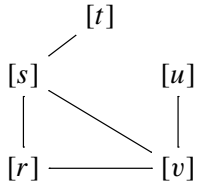




(16)

The last negative case is more complicated.

Proposition 3.1. *The graph in* (17) *can not be realized as* $\Gamma_E(R)$ *for any ring* R.



(17)

Proof. Suppose that R is a ring such that $\Gamma_E(R)$ takes the form in (17). Note that $su \neq 0$, but $rsu = tsu = vsu = 0$, hence $[su] = [s]$. As a result, $su^2 \neq 0$, and hence $u^2 \neq 0$. By symmetry, $[tv] = [v]$ and $t^2 \neq 0$. Next, consider $s + v$, which is annihilated by r, but not t or u. The only candidate for $[s + v]$ is $[r]$, which means that r is self-annihilating. Moreover, it implies the same of s and v, since $0 = rs = (s + v)s$ and $0 = rv = (s + v)v$.

Consider tu, which is annihilated by s and v. We will show that $[tu]$ must represent a new class. The candidates for $[tu]$ are $[r]$, $[s]$, and $[v]$. By symmetry, we need only consider $[tu] = [r]$ and $[tu] = [s]$.

Case I: $[tu] = [r]$. This means that $t^2u \neq 0$, $tu^2 \neq 0$, but $t^2u^2 = 0$ since r is self-annihilating. Here we are using the fact from [Mulay 2002, (3.5), page 3552] that if $y \in [x]$ and $x^n = 0$, then $y^n = 0$ as well. Now $[t^2] \neq [v]$ since t^2 is not annihilated by u; likewise, $[u^2] \neq [s]$. Thus, $[t^2] \neq [t]$, else $t^2u^2 = 0$ implies that $[u^2] = [s]$. Next, if $[t^2] = [r]$, then $t^2v = 0$, which contradicts $[tv] = [v]$, and for the same reason, $[t^2] \neq [s]$. Finally, t^2 is annihilated by s, hence $[t^2] \neq [u]$. Thus, $[tu]$ determines a new class; contradiction.

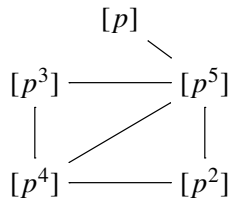
Case II: $[tu] = [s]$. This means that $t^2u = 0$. Thus $[t^2] = [v]$, and hence $t^2v = 0$ since v is self-annihilating. But this contradicts the fact that $[tv] = [v]$. □

4. Positive results

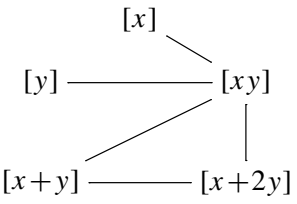
The graphs in this section, labeled (18)–(21), can be realized as condensed zero-divisor graphs. In Proposition 4.1 we prove that when $R = \mathbb{Z}/p^6\mathbb{Z}$, for any prime number p, we get (18) for $\Gamma_E(R)$. In Proposition 4.2 we show that the ring

$$\frac{(\mathbb{Z}/3\mathbb{Z})[\![X, Y]\!]}{(X^2, Y^2)} \tag{$*$}$$

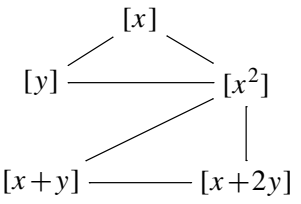
has a graph of the form (19), where lowercase letters match the corresponding uppercase letters in the quotient rings; that is, $x = X + (X^2, Y^2)$ in the ring $(*)$.



(18)



(19)



(20)

The graph (20) is the condensed zero-divisor graph of the ring

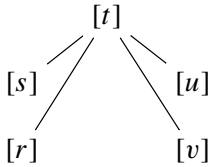
$$\frac{(\mathbb{Z}/3\mathbb{Z})[\![X, Y]\!]}{(X^3, Y^3, XY, X^2 + 2Y^2)}.$$

This was first reported in [Spiroff and Wickham 2011, Example 3.9], but without details; we supply the details in Proposition 4.3.

Finally, the graded ring

$$R = \frac{A[T]}{(T^3, T^2x, T^2y, Txy)}, \quad \text{where} \quad A = \frac{(\mathbb{Z}/2\mathbb{Z})[\![X, Y]\!]}{(X^2, Y^2)}$$

and x and y represent the cosets of X and Y in A, has the graph shown in (21); a summary of the proof is given in Proposition 4.4. This is an example of a *star graph* or *fan graph*; such graphs are studied in our context in [Coykendall et al. 2012, Section 2], and we refer the interested reader to that paper for a full proof that this ring has the graph shown.



(21)

Proposition 4.1. *If $R = \mathbb{Z}/p^6\mathbb{Z}$, then $\Gamma_E(R)$ has the graph* (18).

Proof. Every nonzero element $\bar{r} = r + p^6\mathbb{Z}$ in R is either a unit, in which case $\gcd(r, p) = 1$, or a zero-divisor, in which case $\bar{r} = \overline{up^k}$, where $\bar{u}$ is a unit, and $k \in \{1, 2, 3, 4, 5\}$. By Lemma 2.6, $\operatorname{ann}_R(\bar{u}\,\overline{p^k}) = \operatorname{ann}_R(\overline{p^k})$, therefore the elements p, p^2, p^3, p^4, and p^5 represent the classes. They are all distinct since $\overline{p^i} \in \operatorname{ann}_R(\overline{p^{6-i}})$, but $\overline{p^i} \notin \operatorname{ann}_R(\overline{p^{6-j}})$, for $j > i$. From this the relations follow. □

Proposition 4.2. *If $R = \dfrac{(\mathbb{Z}/3\mathbb{Z})[\![X, Y]\!]}{(X^2, Y^2)}$, then $\Gamma_E(R)$ has the graph shown in* (19).

Proof. The ring has a unique maximal ideal $\mathfrak{m} = (x, y)$. Note that $\mathfrak{m}^2 = (x^2, xy, y^2)$ and hence $\mathfrak{m}^2 = (xy)$ since xy is the only nonzero generator. Moreover, $\mathfrak{m}^3 = 0$ in R; that is, both x and y, annihilate xy. Therefore, a general element of R looks like $a + bx + cy + dxy$, where the coefficients a, b, c, d lie in $\{0, 1, 2\}$. However, whenever $a \neq 0$, this element is a unit since the other terms all lie in $\mathfrak{m}$; see, for instance, [Matsumura 1989, page 3]. We have shown this:

The only possible zero-divisors live in $\mathfrak{m}$ *and have the form* $bx + cy + dxy$.

We now proceed to describe each class.

***First class:* $[xy]$.** $\operatorname{Ann}_R(dxy) = \mathfrak{m}$, for any $d \neq 0$. To see this, note by $\operatorname{ann}_R(xy) \subseteq \mathfrak{m}$, by the statement proved immediately above. On the other hand, since both generators of $\mathfrak{m}$ annihilate xy, $\mathfrak{m} \subseteq \operatorname{ann}_R(xy)$. Thus, $\operatorname{ann}_R(xy) = \mathfrak{m}$. Also, since 2 is a unit in R, Lemma 2.6 implies that $[2xy] = [xy]$.

***Second class:* $[x]$.** $\operatorname{Ann}_R(bx + dxy) = (x)$, for $b \neq 0$.

Let $b'x + c'y + d'xy \in \operatorname{ann}_R(bx + dxy)$. Then

$$0 = (bx + dxy)(b'x + c'y + d'xy) = bc'xy,$$

which is zero if and only if $bc' = 0$. Since b, c' are elements of a field and $b \neq 0$, we must have $c' = 0$. Therefore, the annihilators of $bx + dxy$ have the form

$$b'x + d'xy = x(b' + d'y) = x(b' + b''x + d'y + d''xy),$$

for any $b', b'', d', d'' \in \mathbb{Z}/3\mathbb{Z}$; that is, $\operatorname{ann}_R(bx + dxy) = (x)$.

***Third class:* $[y]$.** An analogous argument shows that $\operatorname{ann}_R(cy + dxy) = (y)$, for $c \neq 0$.

***Fourth class:* $[x + y]$.** $\operatorname{Ann}_R(bx + by + dxy) = (x + 2y)$, for $b \neq 0$.

Let $b'x + c'y + d'xy \in \operatorname{ann}_R(bx + by + dxy)$. Then

$$0 = (bx + by + dxy)(b'x + c'y + d'xy) = bc'xy + bb'xy = b(b' + c')xy,$$

which is zero if and only if $b(b' + c') = 0$. Since $b \neq 0$, we must have $b' + c' \equiv 0$ in $\mathbb{Z}/3\mathbb{Z}$. Therefore, the elements that annihilate $bx + by + dxy$ are $d'xy$, $x + 2y + d'xy$ and $2x + y + d'xy$. However, these last two differ by a unit, for example, $2x + y = 2(x + 2y)$, and $d'xy = d'y(x + 2y)$, hence only $x + 2y$ is necessary as a generator. Thus, $\operatorname{ann}_R(bx + by + dxy) = (x + 2y)$.

***Fifth class:* $[x + 2y]$.** A similar analysis shows that $\operatorname{ann}_R(bx + 2by + dxy) = (x + y)$, where $b \neq 0$. □

Proposition 4.3 [Spiroff and Wickham 2011, Example 3.9]. *If*

$$R = \frac{(\mathbb{Z}/3\mathbb{Z})[\![X, Y]\!]}{(X^3, Y^3, XY, (X+Y)(X+2Y))},$$

then $\Gamma_E(R)$ *has the graph shown in* (20).

Proof. The ring has unique maximal ideal $\mathfrak{m} = (x, y)$. The nonzero generators of $\mathfrak{m}^2$ are (x^2, y^2) and $\mathfrak{m}^3 = 0$ in R; that is, both x and y, annihilate every element in $\mathfrak{m}^2$. Therefore, a general element of R looks like $a + bx + cy + dx^2 + ey^2$, where the coefficients a, b, c, d, and e, are all either 0, 1 or 2. However, whenever $a \neq 0$, this polynomial is a unit since the other terms all lie in $\mathfrak{m}$; see [Matsumura 1989, page 3]. Moreover, the relation $(x+y)(x+2y) = 0$ simplifies to $x^2 = y^2$. Therefore, *the only possible zero-divisors live in* $\mathfrak{m}$ *and have the form* $bx + cy + dx^2$.

First class: $[x^2]$. $\operatorname{Ann}_R(dx^2) = \mathfrak{m}$, $d \neq 0$.

To see this, we first note that $\operatorname{ann}_R(x^2) \subseteq \mathfrak{m}$. On the other hand, since both generators of $\mathfrak{m}$ annihilate x^2, $\mathfrak{m} \subseteq \operatorname{ann}_R(x^2)$. Thus, $\operatorname{ann}_R(x^2) = \mathfrak{m}$. Moreover, since 2 is a unit in R, Lemma 2.6 implies that $[2x^2] = [x^2]$.

Second class: $[x]$. $\operatorname{Ann}_R(bx + dx^2) = (y)$, for $b \neq 0$.

Let $b'x + c'y + d'x^2 \in \operatorname{ann}_R(bx + dx^2)$. Then $0 = (bx + dx^2)(b'x + c'y + d'x^2) = bb'x^2$, which is zero if and only if $bb' = 0$. Since b, b' are elements of a field and $b \neq 0$, we must have $b' = 0$. Therefore, the annihilators of $bx + dx^2$ have the form $c'y + d'x^2$, or $c'y + d'y^2$, since $x^2 = y^2$, and $c'y + d'y^2 = y(c' + d'y) = y(c' + b''x + d'y + d''x^2)$, for any $c', b'', d', d'' \in \mathbb{Z}/3\mathbb{Z}$; that is, $\operatorname{ann}_R(bx + dx^2) = (y)$.

Third class: $[y]$. An analogous argument shows that $\operatorname{ann}_R(cy + dx^2) = (x)$, for $c \neq 0$.

Fourth class: $[x + y]$. $\operatorname{Ann}_R(bx + by + dx^2) = (x + 2y)$, for $b \neq 0$.

Let $b'x + c'y + d'x^2 \in \operatorname{ann}_R(bx + by + dx^2)$. Then

$$0 = (bx + by + dx^2)(b'x + c'y + d'x^2) = bb'x^2 + bc'y^2 = b(b' + c')x^2,$$

which is zero if and only if $b(b' + c') = 0$. Since $b \neq 0$, we must have $b' + c' \equiv 0$ in $\mathbb{Z}/3\mathbb{Z}$. Therefore, the elements that annihilate $bx + by + dx^2$ are $d'x^2$, $x + 2y + d'x^2$ and $2x + y + d'x^2$. However, these last two differ by a unit, for example, $2x + y = 2(x + 2y)$, and $d'x^2 = d'x(x + 2y)$, hence only $x + 2y$ is necessary as a generator. Thus, $\operatorname{ann}_R(bx + by + dx^2) = (x + 2y)$.

Fifth class: $[x + 2y]$. A similar analysis shows that $\operatorname{ann}_R(bx + 2by + dxy) = (x + y)$, where $b \neq 0$. □

Proposition 4.4. *If*

$$R = \frac{A[T]}{(T^3, T^2x, T^2y, Txy)}, \qquad \text{where } A = \frac{(\mathbb{Z}/2\mathbb{Z})[\![X, Y]\!]}{(X^2, Y^2)},$$

then $\Gamma_E(R)$ *has the graph shown in* (21).

Outline of proof. (See [Coykendall et al. 2012] for details.) The ring A is similar to the ring in Proposition 4.2, but with a smaller coefficient ring, and an analogous argument to the one there shows that zero-divisors in A take the form $bx+cy+dxy$, where $b, c, d \in \mathbb{Z}/2\mathbb{Z}$, and there are four distinct classes, given by $\operatorname{ann}_A(x) = (x)$, $\operatorname{ann}_A(y) = (y)$, $\operatorname{ann}_A(xy) = (x, y)$, and $\operatorname{ann}_A(x + y) = (x + y)$. In fact, these determine four distinct classes in R. Note that R has the direct sum decomposition

$$A \oplus \frac{A}{(xy)} \cdot t \oplus \frac{A}{(x, y)} \cdot t^2$$

as an abelian group. We describe the first four classes in R and the last class, determined by t.

***First class:* [xy].** $\operatorname{Ann}_R(xy + \overline{\gamma}t^2) = (x, y)A \oplus \dfrac{A}{(xy)} \cdot t \oplus \dfrac{A}{(x, y)} \cdot t^2$, for $\overline{\gamma}$ in $A/(x, y)$.

***Second class:* [x].** $\operatorname{Ann}_R(x + \overline{\gamma}t^2) = (x)A \oplus \dfrac{(x, y)}{(xy)} \cdot t \oplus \dfrac{A}{(x, y)} \cdot t^2$.

***Third class:* [y].** $\operatorname{Ann}_R(y + \overline{\gamma}t^2) = (y)A \oplus \dfrac{(x, y)}{(xy)} \cdot t \oplus \dfrac{A}{(x, y)} \cdot t^2$.

***Fourth class:* [$x + y$].** $\operatorname{Ann}_R(x + y + \overline{\gamma}t^2) = (x + y)A \oplus \dfrac{(x, y)}{(xy)} \cdot t \oplus \dfrac{A}{(x, y)} \cdot t^2$.

***Fifth class:* t.** $\operatorname{Ann}_R(t + \overline{\gamma}t^2) = (xy)A \oplus \dfrac{(x, y)}{(xy)} \cdot t \oplus \dfrac{A}{(x, y)} \cdot t^2$.

Remark 4.5. The (nonzero) elements $\alpha + \overline{\beta}t$, where $\alpha \in (x, y)A$ and $\overline{\beta} \in A/(xy)$, fall into the above categories. If $\alpha = 0$ and $\beta \in (x, y)A$, then the element is in the first class; if $\alpha \neq 0$ and $\beta \in (x, y)A$, then the element is in $[\alpha]$; finally, if $\beta \notin (x, y)A$, then the element is in $[t]$. □

References

[Anderson and Livingston 1999] D. F. Anderson and P. S. Livingston, "The zero-divisor graph of a commutative ring", *J. Algebra* **217**:2 (1999), 434–447. MR 2000e:13007 Zbl 0941.05062

[Beck 1988] I. Beck, "Coloring of commutative rings", *J. Algebra* **116**:1 (1988), 208–226. MR 89i:13006 Zbl 0654.13001

[Coykendall et al. 2012] J. Coykendall, S. Sather-Wagstaff, L. Sheppardson, and S. Spiroff, "On zero divisor graphs", in *Progress in commutative algebra, II: Closures, finiteness and factorization*, edited by C. Francisco et al., de Gruyter, Berlin, 2012.

[Dummit and Foote 1991] D. S. Dummit and R. M. Foote, *Abstract algebra*, Prentice Hall, Englewood Cliffs, NJ, 1991. MR 92k:00007 Zbl 0751.00001

[Harary 1969] F. Harary, *Graph theory*, Addison-Wesley, Menlo Park, CA, 1969. MR 41 #1566 Zbl 0182.57702

[Levidiotis 2010] F. Levidiotis, *Five point zero divisor graphs determined by equivalence classes of zero divisors*, Undergraduate honors thesis, University of Mississippi, 2010.

[Matsumura 1989] H. Matsumura, *Commutative ring theory*, 2nd ed., Cambridge Studies in Advanced Mathematics **8**, Cambridge University Press, 1989. MR 90i:13001 Zbl 0666.13002

[Mulay 2002] S. B. Mulay, "Cycles and symmetries of zero-divisors", *Comm. Algebra* **30**:7 (2002), 3533–3558. MR 2003j:13007a Zbl 1087.13500

[Spiroff and Wickham 2011] S. Spiroff and C. Wickham, "A zero divisor graph determined by equivalence classes of zero divisors", *Comm. Alg.* **39**:7 (2011), 2338–2348.

Received: 2010-06-17 Revised: 2011-02-11 Accepted: 2011-02-23

flevidiotis@luc.edu *Department of Mathematics and Statistics, Loyola University Chicago, 26 E Pearson Street, #2305, Chicago, IL 60611, United States*

spiroff@olemiss.edu *Department of Mathematics, University of Mississippi, Hume Hall 305, P.O. Box 1848, University, MS 38677-1848, United States*
http://home.olemiss.edu/depts/mathematics/spiroff.htm

INVOLVE 4:1(2011)

msp

A note on moments in finite von Neumann algebras

Jon Bannon, Donald Hadwin and Maureen Jeffery

(Communicated by David R. Larson)

By a result of the second author, the Connes embedding conjecture (CEC) is false if and only if there exists a self-adjoint noncommutative polynomial $p(t_1, t_2)$ in the universal unital C^*-algebra $\mathcal{A} = \langle t_1, t_2 : t_j = t_j^*,\ 0 < t_j \leq 1 \text{ for } 1 \leq j \leq 2\rangle$ and positive, invertible contractions x_1, x_2 in a finite von Neumann algebra $\mathcal{M}$ with trace τ such that $\tau(p(x_1, x_2)) < 0$ and $\mathrm{Tr}_k(p(A_1, A_2)) \geq 0$ for every positive integer k and all positive definite contractions A_1, A_2 in $M_k(\mathbb{C})$. We prove that if the real parts of all coefficients but the constant coefficient of a self-adjoint polynomial $p \in \mathcal{A}$ have the same sign, then such a p cannot disprove CEC if the degree of p is less than 6, and that if at least two of these signs differ, the degree of p is 2, the coefficient of one of the t_i^2 is nonnegative and the real part of the coefficient of $t_1 t_2$ is zero then such a p disproves CEC only if either the coefficient of the corresponding linear term t_i is nonnegative or both of the coefficients of t_1 and t_2 are negative.

1. Introduction

The Connes embedding conjecture (CEC) is true if every separable type II_1 factor $\mathcal{M}$ embeds in a tracial ultrapower $\mathcal{R}^\omega$ of the amenable type II_1 factor $\mathcal{R}$. This question concerns the matricial approximation of the elements of a type II_1 factor $\mathcal{M}$ with faithful normal trace state τ in the sense we now recall. For an N-tuple $(x_1, \ldots, x_N)$ of self-adjoint elements in $\mathcal{M}$, $R > 0$, $n, k \in \mathbb{N}$ and $\varepsilon > 0$, we let

$$\Gamma_R(x_1, \ldots, x_N : n, k, \varepsilon)$$

denote the set of tuples $(A_1, \ldots, A_N)$ of those $k \times k$ self-adjoint matrices over $\mathbb{C}$ of operator norm at most R satisfying

$$\left| \tau(x_{i_1} x_{i_2} \ldots x_{i_p}) - \frac{1}{k} \mathrm{Tr}(A_{i_1} A_{i_2,} \ldots A_{i_p}) \right| < \varepsilon,$$

MSC2000: primary 46L10; secondary 46L54.

Keywords: von Neumann algebras, noncommutative moment problems, Connes embedding conjecture.

Jeffery is an undergraduate at Siena College in Loudonville, New York.

whenever $1 \leq p \leq n$ and $(i_1, i_2, \ldots, i_p) \in \{1, 2, \ldots, N\}^p$. We call the elements of $\Gamma_R(x_1, \ldots, x_N : n, k, \varepsilon)$ *approximating microstates* for $(x_1, \ldots, x_N)$ of precision (n, ε) using $k \times k$ matrices of norm at most R. A separable type II_1 factor $\mathcal{M}$ embeds in an ultrapower $\mathcal{R}^\omega$ if and only if for all tuples $(x_1, \ldots, x_N)$ of self-adjoint elements in $\mathcal{M}$, all $n \in \mathbb{N}$ and all $\varepsilon > 0$, it is possible to find $k \in \mathbb{N}$ and $R > 0$ such that $\Gamma_R(x_1, \ldots, x_N : n, k, \varepsilon) \neq \varnothing$. In [Rădulescu 1999] it is proved that this statement is true under the restriction that $n \in \{2, 3\}$, and that if the statement were true for $n = 4$, the CEC would follow.

Our paper concerns the following reformulation of the CEC:

Theorem 1.1 [Hadwin 2001, Corollary 2.3]. *Let $\mathcal{H}$ be a separable Hilbert space. The Connes embedding conjecture is false if and only if there is a positive integer n, a noncommutative polynomial $p(t_1, t_2, \ldots, t_n)$ in the universal unital C^*-algebra $\mathcal{A}_n = \langle t_1, t_2, \ldots, t_n : t_j = t_j^*, -1 < t_j \leq 1 \text{ for } 1 \leq j \leq n\rangle$ and an n-tuple $(x_1, \ldots, x_n)$ of self-adjoint contractions in $B(H)$ such that*

(i) $\mathrm{Tr}_k(p(A_1, A_2, \ldots, A_n)) \geq 0$ *for every positive integer k and every n-tuple $(A_1, \ldots, A_n)$ of self-adjoint contractions $A_1, A_2, \ldots, A_n$ in $M_k(\mathbb{C})$, and*

(ii) *$W^*(x_1, x_2, \ldots, x_n)$ has a faithful tracial state τ and $\tau(p(x_1, x_2, \ldots, x_n)) < 0$.*

It is well known that a separable type II_1 factor $\mathcal{M}$ embeds in an $\mathcal{R}^\omega$ if and only if $\mathcal{M} \otimes M_k(\mathbb{C})$ does for all $k \in \mathbb{N}$. If $\mathcal{M}$ is generated by k self-adjoint elements then $\mathcal{M} \otimes M_k(\mathbb{C})$ is generated by two self-adjoint elements [Sinclair and Smith 2008, Proposition 16.1.1]. Whenever $x \in B(H)$ is a self-adjoint contraction and $\varepsilon > 0$, it follows (e.g., by the continuous functional calculus for x) that

$$\frac{(1+\varepsilon)+x}{2+\varepsilon}$$

is a positive invertible contraction. Therefore, if we replace $\mathcal{A}_n$ by

$$\mathcal{A} = \big\langle t_1, t_2 : t_j = t_j^*,\ 0 < t_j \leq 1 \text{ for } 1 \leq j \leq 2\big\rangle,$$

and repeat the argument in [Hadwin 2001, Section 2], we obtain the following.

Theorem 1.2. *Let $\mathcal{H}$ be a separable Hilbert space. The Connes embedding conjecture is false if and only if there is a noncommutative polynomial $p(t_1, t_2)$ in the universal unital C^*-algebra $\mathcal{A} = \langle t_1, t_2 : t_j = t_j^*$, with $0 < t_j \leq 1$ for $1 \leq j \leq 2\rangle$, and positive, invertible contractions x_1 and x_2 in $B(H)$ such that*

(i) $\mathrm{Tr}_k(p(A_1, A_2)) \geq 0$ *for every positive integer k and all positive definite contractions A_1 and A_2 in $M_k(\mathbb{C})$, and*

(ii) *$W^*(x_1, x_2)$ has a faithful tracial state τ and $\tau(p(x_1, x_2)) < 0$.*

Also note that if a polynomial $p \in \mathcal{A}$ satisfies (i) and (ii) in the theorem, then so does the polynomial $p + p^*$. We may therefore assume that the polynomial appearing in the theorem is self-adjoint.

Note that, even if we restrict our attention in Theorem 1.1 (or Theorem 1.2) to the case where the degree of p is less than or equal to 3, we cannot use [Rădulescu 1999] to rule out the possibility of finding such a p that will disprove the CEC, because existing methods only allow us to use, when $R' < R$, the existence of a microstate in $\Gamma_R(x_1, \ldots, x_N : n, k, \varepsilon)$ to guarantee the existence of a microstate in $\Gamma_{R'}(x_1, \ldots, x_N : n', k, \varepsilon')$, where $\varepsilon' < \varepsilon$ and $n' > n$ — that is, decreasing R comes at the expense of increasing n. See, for example, Proposition 2.4 of [Voiculescu 1994] or Lemma 4 of [Dostál and Hadwin 2003]. Even if this difficulty were overcome, there is no guarantee that the matrices in any approximating microstates found would be positive definite. It behooves us, therefore, to either look for a noncommutative polynomial that may be used to disprove the CEC as prescribed in Theorem 1.1, or to proceed inductively, by degree, to show that such a polynomial cannot exist.

In Section 2 of this paper we prove, in Corollary 2.5 and Theorem 2.6 that if the real parts of all coefficients but the constant coefficient of a self-adjoint noncommutative polynomial $p \in \mathcal{A}$ share the same sign, then such a p cannot disprove the CEC if the degree of p is less than 6. We prove in Section 3 that if the degree of a self-adjoint noncommutative polynomial $p \in \mathcal{A}$ is 2, the real part of the coefficient of $t_1 t_2$ is zero and the coefficient of one of the t_i^2 is nonnegative, then such a p disproves the CEC only if either the coefficient of the corresponding linear term t_i is nonnegative or if both of the coefficients of t_1 and t_2 are negative.

From here on in this paper, the symbols t_1 and t_2 will denote the standard generators of the universal C^*-algebra

$$\mathcal{A} = \langle t_1, t_2 : t_j = t_j^*,\ 0 < t_j \leq 1 \text{ for } 1 \leq j \leq 2 \rangle.$$

We refer the reader to [Kadison and Ringrose 1983; Sinclair and Smith 2008] for the basic theory of finite von Neumann algebras.

2. τ-symmetrizable monomials

We prove that if the real parts of all coefficients but the constant coefficient of self-adjoint $p \in \mathcal{A}$ share the same sign, and the constant coefficient is positive, then p cannot disprove the CEC if its degree is less than six. Let $\mathcal{M}$ be a finite von Neumann algebra with faithful trace state τ, and $0 < x_1, x_2 \leq 1$ self-adjoint contractions in $\mathcal{M}$.

Definition 2.1. A *symmetric expression* in x_1, x_2 is a finite sequence

$$(w_0, w_1, \ldots, w_{N-1}, w_N)$$

of elements in $\mathcal{M}$, where $N \in \mathbb{N}$, $w_k = x_i^s$ with $i \in \{1, 2\}$, $s \in \{1, 1/2\}$ and $w_k = w_{N-k}$ for all $k \in \{0, 1, \ldots, N\}$. A monic monomial $m(x_1, x_2) = x_{i_1} x_{i_2} \ldots x_{i_l} \in \mathcal{M}$ with $i_j \in \{1, 2\}$ for $j \in \{1, 2, \ldots, l\}$ is τ-symmetrizable if there exists a symmetric expression $(w_0, w_1, \ldots, w_{N-1}, w_N)$ in x_1, x_2 such that

$$\tau(x_{i_1} x_{i_2} \ldots x_{i_l}) = \tau(w_0 w_1 \ldots w_{N-1} w_N).$$

The element $w_0 w_1 \ldots w_{N-1} w_N \in \mathcal{M}$ is called the element associated to the symmetric expression $(w_0, w_1, \ldots, w_{N-1}, w_N)$.

Lemma 2.2. *If $(w_0, w_1, \ldots, w_{N-1}, w_N)$ is a symmetric expression in x_1, x_2, then the associated element $w_0 w_1 \ldots w_{N-1} w_N$ in $\mathcal{M}$ is a nonnegative contraction.*

Proof. We prove this by induction on $N + 1$. If $N + 1 = 1$, then $N = 0$ and the result is clear from the assumptions on the x_i.

Assume now that the result holds for $N + 1 \leq l$, that is, for all symmetric expressions $(w_0, w_1, \ldots, w_{j-1}, w_j)$ in x_1, x_2 with $j < l$. Let $(w_0, w_1, \ldots, w_{l-1}, w_l)$ be a symmetric expression in x_1, x_2. Then so is $(w_1, \ldots, w_{l-1})$. By the induction hypothesis, $w_1 \ldots w_{l-1} \in \mathcal{M}$ is a nonnegative contraction. Since $w_0 = w_l = x_i^s$ for some $i \in \{1, 2\}$ and $s \in \{1, \frac{1}{2}\}$, we have

$$0 \leq w_0 w_1 \ldots w_{l-1} w_l = x_i^s w_1 \ldots w_{l-1} x_i^s \leq x_i^{2s} \leq x_i \leq 1. \qquad \square$$

Remark 2.3. It is a straightforward exercise to verify that every monic noncommutative monomial $m(x_1, x_2)$ of degree less than six is τ-symmetrizable in any finite von Neumann algebra $\mathcal{M}$ with faithful trace state τ. (Here, of course, it is essential that $0 < x_1, x_2 \leq 1$!)

Corollary 2.4. *If $m(x_1, x_2) = x_{i_1} x_{i_2} \ldots x_{i_l} \in \mathcal{M}$ is a τ-symmetrizable monic monomial, then $1 - \tau(m(x_1, x_2)) \geq 0$.*

Proof. Since m is τ-symmetrizable, there exists a symmetric expression

$$(w_0, w_1, \ldots, w_{N-1}, w_N)$$

in x_1, x_2 such that

$$\tau(x_{i_1} x_{i_2} \ldots x_{i_l}) = \tau(w_0 w_1 \ldots w_{N-1} w_N).$$

By Lemma 2.2 and the fact that τ is a state, $\tau(w_0 w_1 \ldots w_{N-1} w_N) \leq 1$. $\square$

In the following two results, $J = J \setminus \{0\}$ denotes a finite index set, and for all $j \in J$, $c_j \in \mathbb{C}$, and $m_j(t_1, t_2) \neq 1$ denotes a monic monomial in $\mathcal{A}$.

Corollary 2.5. *If $0 < x_1, x_2 \leq 1$ in $\mathcal{M}$ and $p(t_1, t_2) = c_0 1 + \sum_{j \in J} c_j m_j(t_1, t_2)$ is a self-adjoint noncommutative polynomial in $\mathcal{A}$ such that, such that $c_0 > 0$, $\mathrm{Re}(c_j) \geq 0$ for all $j \in J$, $p(1, 1) \geq 0$ and $m_j(x_1, x_2)$ is τ-symmetrizable for every $j \in J$, then $\tau(p(x_1, x_2)) \geq 0$.*

Proof. This is trivial application of Corollary 2.4. □

Theorem 2.6. *If $0 < x_1, x_2 \leq 1$ in $\mathcal{M}$ and $p(t_1, t_2) = c_0 1 + \sum_{j \in J} c_j m_j(t_1, t_2)$ is a self-adjoint noncommutative polynomial in $\mathcal{A}$ such that $c_0 > 0$, $\mathrm{Re}(c_j) < 0$ for all $j \in J$, $p(1, 1) \geq 0$ and $m_j(x_1, x_2)$ is τ-symmetrizable for every $j \in J$, then $\tau(p(x_1, x_2)) \geq 0$.*

Proof. Suppose $p(t_1, t_2)$ satisfies the hypotheses. We have

$$p(1, 1) = c_0 1 + \sum_{j \in J} c_j \geq 0,$$

and therefore

$$\tau(p(x_1, x_2)) \geq \sum_{j \in J} c_j (m_j(x_1, x_2) - 1) \geq 0.$$ □

3. Degrees 1 and 2

In degree 1 it is convenient to consider the statement of Theorem 1.1 above. The next result rules out the possibility of finding a polynomial p of degree 1 that will disprove the CEC via Theorem 1.1. Observe that if $p(s, t) = c_0 + c_1 s + c_2 t = \bar{c}_0 + \bar{c}_1 s + \bar{c}_2 t$ for any real numbers $-1 \leq s, t \leq 1$ and that $p(s, t) \geq 0$ for any such s and t, then $c_0 \geq |c_1 + c_2|$.

Theorem 3.1. *Let $\mathcal{H}$ be a separable Hilbert space. Let x_1 and x_2 be self-adjoint contraction operators in $B(H)$ such that $W^*(x_1, x_2)$ has a faithful trace state τ. If $p(t_1, t_2) = c_0 + c_1 t_1 + c_2 t_2 = \bar{c}_0 + \bar{c}_1 t_1 + \bar{c}_2 t_2$ is a self-adjoint polynomial in $\mathcal{A}$ with $c_0 \geq |c_1 + c_2|$ then $\tau(p(x_1, x_2)) \geq 0$.*

Proof. Observe that $\tau(c_0 + c_1 x_1 + c_2 x_2) = c_0 + c_1 \tau(x_1) + c_2 \tau(x_2) \geq c_0 - |c_1 + c_2|$, since $-1 \leq \tau(x_i) \leq 1$ for $i \in \{1, 2\}$. □

We now turn to degree 2. We first prove in Theorem 3.4 that if

$$p(t_1, t_2) = c_0 + c_1 t_1 + c_2 t_2 + c_3 t_1^2 + c_4 t_1 t_2 + \bar{c}_4 t_2 t_1 + c_5 t_2^2$$

is a quadratic, self-adjoint noncommutative polynomial such that either c_4 is the only nonzero degree 2 term with $2\,\mathrm{Re}(c_4) \neq 0$ or one of c_3 or c_5 is positive, then whenever $p(s, t)$ is nonnegative for all real numbers $0 < s, t \leq 1$, it follows that $\mathrm{Tr}_k(p(A, B)) \geq 0$ for all positive definite contractions A and B in $M_k(\mathbb{C})$, for any $k \in \mathbb{N}$.

To prove the result above, we shall need the fact that any positive definite square matrix has strictly positive entries on its main diagonal. This is a direct consequence of Sylvester's minorant criterion for positive definiteness.

Lemma 3.2. *Let $A = (A_{ij})_{i=1}^k \in M_k(\mathbb{C})$ be positive definite. Then $A_{ii} > 0$ for all $i \in \{1, 2, \ldots, k\}$.*

Proof. We prove this by induction on k. Recognize that the case $k = 1$ is clear. Assume the claim holds for $k = l$, and that $A = (A_{ij})_{i=1}^{l+1}$ is a positive definite matrix. By Sylvester's criterion, $A = (A_{ij})_{i=1}^{l}$ is also positive definite, and therefore, by the induction hypothesis, $A_{ii} > 0$ if $i \in \{1, 2, \ldots l\}$. We need only show $A_{(l+1)(l+1)} > 0$. Let $v \in \mathbb{C}^{l+1}$ be the vector with 1 in its $(l+1)$-st row and zero elsewhere. Then $\langle Av, v\rangle = A_{(l+1)(l+1)} > 0$ by the positive definiteness of A. □

We now observe that if a polynomial is nonnegative on $(0, 1] \times (0, 1]$, then its constant term must be nonnegative.

Lemma 3.3. *If* $p(s, t) = c_0 + c_1 s + c_2 t + c_3 s^2 + 2\,\mathrm{Re}(c_4) st + c_5 t^2 \geq 0$ *for all real numbers* $0 < s, t \leq 1$, *then* $c_0 \geq 0$.

Proof. For any $\varepsilon > 0$ we have

$$0 < p(\varepsilon, \varepsilon) = c_0 + \big(c_1 + c_2 + (c_3 + 2\,\mathrm{Re}(c_4) + c_5)\varepsilon\big)\varepsilon;$$

hence $c_0 \geq 0$. □

Theorem 3.4. *Let* $p(t_1, t_2) = c_0 + c_1 t_1 + c_2 t_2 + c_3 t_1^2 + c_4 t_1 t_2 + \bar{c}_4 t_2 t_1 + c_5 t_2^2$ *be a self-adjoint noncommutative polynomial in* $\mathcal{A}$. *Suppose*

$$p(s, t) = c_0 + c_1 s + c_2 t + c_3 s^2 + 2\,\mathrm{Re}(c_4) st + c_5 t^2 \geq 0$$

for all real numbers $0 < s, t \leq 1$, *and either* $c_3 = 0$, $c_5 = 0$ *and* $2\,\mathrm{Re}(c_4) \neq 0$ *or* $c_3 > 0$ *or* $c_5 > 0$. *Then* $\mathrm{Tr}_k(p(A, B)) \geq 0$ *for any positive definite contractions* A, B *in* $M_k(\mathbb{C})$.

Proof. For simplicity, let us assume $c_5 \geq 0$. Let A, B be positive definite contractions in $M_k(\mathbb{C})$. By the spectral theorem, we may assume $A = \mathrm{diag}(A_i)_{i=1}^{k}$ is diagonal. A simple computation establishes that, for all $i \in \{1, 2, \ldots, k\}$,

$$(p(A, B))_{ii} = p(A_i, B_{ii}) + \sum_{j \in \{1,2,\ldots,k\}\setminus\{i\}} c_5 |B_{ij}|^2.$$

Since A is a positive definite contraction, each A_i satisfies $0 < A_i \leq 1$. If we could establish that the matrix $B_0 := \mathrm{diag}(B_{ii})_{i=1}^{k}$ is a positive definite contraction, then each $p(A_i, B_{ii})$ would follow nonnegative by assumption and therefore $\mathrm{Tr}_k(p(A, B)) = \sum_{i=1}^{k} (p(A, B))_{ii} \geq 0$. Positivity of B_0 is a simple consequence of the positive definiteness of B, since every diagonal entry of a positive definite matrix is strictly positive by Lemma 3.2. It remains to show that B_0 is a contraction, which is equivalent to proving that $I - B_0$ is positive semidefinite. We know, however, that $I - B$ is positive semidefinite, and hence that for all $\varepsilon > 0$ that $(I + \varepsilon) - B$ is positive definite. Again as a consequence of Sylvester's criterion, $((I + \varepsilon) - B)_{ii} > 0$ for all $i \in \{1, 2, \ldots, n\}$, therefore for all such i it follows that $1 + \varepsilon > B_{ii}$, and hence $1 \geq B_{ii}$. It follows that $I - B_0$ is positive semidefinite, hence B_0 is a contraction. □

Let $\mathcal{M}$ be a von Neumann algebra with faithful trace state τ. Below,

$$\langle x, y\rangle_2 = \tau(y^*x) \quad \text{and} \quad \|x\|_2^2 = \tau(x^*x)^{1/2}, \quad \text{for } x, y \in \mathcal{M}.$$

Let $n \in \mathbb{N}$ and x_1, x_2 be positive invertible contractions in $\mathcal{M}$. For every $k \in \mathbb{N}$, there are spectral projections $\{P_i^{(k)}\}_{i=1}^k$ in $\{1, x_1\}''$ such that $\tau(P_i^{(k)}) = 1/k$ for each i and

$$\Big\| x_1 - \sum_{i=1}^{k} \frac{i-1}{k} P_i^{(k)} \Big\| < \frac{1}{k}.$$

If $i = j$, let $V_{ij}^{(k)} = P_i^{(k)}$, and if $i \neq j$, let $V_{ij}^{(k)}$ be a partial isometry in $\mathcal{M}$ with initial projection $P_j^{(k)}$ (meaning that $V_{ij}^{(k)}(V_{ij}^{(k)})^* = P_j$) and final projection $P_i^{(k)}$ (meaning that $(V_{ij}^{(k)})^* V_{ij}^{(k)} = P_i^{(k)}$). We now prove that if x_2 is sufficiently close (in $\|\cdot\|_2$) to a positive definite element in the type I subfactor of $\mathcal{M}$ generated by $\{V_{ij}^{(k)}\}_{i,j=1}^k$, then $\tau(p(x_1, x_2)) \geq 0$ when p satisfies the hypotheses of Theorem 3.4. In the statement of the theorem, we regard x_2 as an operator matrix and compare it entry-wise to the element $(b_{ij} V_{ij}^{(k)})_{i,j=1}^k$.

Theorem 3.5. *Let $\mathcal{M}$ be a finite von Neumann algebra with faithful trace state τ, let x_1, x_2 be positive, invertible elements in $\mathcal{M}$, and adopt the notation in the previous paragraph. Let $p(t_1, t_2) = c_0 + c_1 t_1 + c_2 t_2 + c_3 t_1^2 + c_4 t_1 t_2 + \bar{c}_4 t_2 t_1 + c_5 t_2^2$ be a self-adjoint noncommutative polynomial in $\mathcal{A}$. Suppose that*

$$p(s, t) = c_0 + c_1 s + c_2 t + c_3 s^2 + 2\,\mathrm{Re}(c_4) st + c_5 t^2 \geq 0$$

for all real numbers $0 < s, t \leq 1$, that either $c_3 = 0$, $c_5 = 0$ and $2\,\mathrm{Re}(c_4) \neq 0$ or $c_3 > 0$ or $c_5 > 0$, and that for all $k \in \mathbb{N}$ there exists a type I subfactor of $\mathcal{M}$ generated by $\{V_{ij}^{(k)}\}_{i,j=1}^k$ as in the previous paragraph, and a positive definite contraction $(b_{ij})_{i,j=1}^k \in M_k(\mathbb{C})$ such that

$$\big\| P_i^{(k)} x_2 P_j^{(k)} - b_{ij} V_{ij}^{(k)} \big\|_2 < \frac{1}{k^{100}}, \quad \textit{for all } i, j \in \{1, 2, \ldots, k\}.$$

Then $\tau(p(x_1, x_2)) \geq 0$.

Proof. Let $D_k = \sum_{i=1}^{k} \frac{i-1}{k} P_i^{(k)}$ and $B_k = \sum_{i,j=1}^{k} b_{ij} V_{ij}^{(k)}$. Writing $x_1 = D_k + (x_1 - D_k)$ and $x_2 = B_k + (x_2 - B_k)$, we have

$$\begin{aligned}
&\tau(p(x_1), p(x_2)) \\
&\quad = c_0 + c_1 \tau\big(D_k + (x_1 - D_k)\big) + c_2 \tau\big(B_k + (x_2 - B_k)\big) + c_3 \tau\big((D_k + (x_1 - D_k))^2\big) \\
&\qquad + 2\,\mathrm{Re}(c_4) \tau\big((D_k + (x_1 - D_k))(B_k + (x_2 - B_k))\big) + c_5 \tau\big((B_k + (x_2 - B_k))^2\big) \\
&\quad = p(\tau(D_k), \tau(B_k)) + c_1 \tau(x_1 - D_k) + c_2 \tau(x_2 - B_k) \\
&\qquad + 2c_3 \tau(D_k(x_1 - D_k)) + c_3 \tau(x_1 - D_k)^2 + 2\,\mathrm{Re}(c_4) \tau(D_k(x_2 - B_k)) \\
&\qquad + 2\,\mathrm{Re}(c_4) \tau(B_k(x_1 - D_k)) + 2\,\mathrm{Re}(c_4) \tau\big((x_1 - D_k)(x_2 - B_k)\big) \\
&\qquad + 2c_5 \tau\big(B_k(x_2 - B_k)\big) + c_5 \tau\big((x_2 - B_k)^2\big).
\end{aligned}$$

Therefore, by the triangle and Cauchy–Schwartz inequalities and the fact that the operator norm dominates the $\|\cdot\|_2$-norm,

$$\big|\tau(p(x_1),\, p(x_2)) - p(\tau(D_k),\, \tau(B_k))\big| \le \big(|c_1| + |c_2| + 3|c_3| + 6\,\mathrm{Re}(c_4) + 3c_5\big)\frac{1}{k}.$$

Since $W^*(D_k,\, B_k) \cong W^*(\mathrm{diag}((i-1)/k,\, i \in \{1,\ldots,k\}),\, (b_{ij})_{i,j=1}^{k}) \subseteq M_k(\mathbb{C})$ via the obvious trace-preserving $*$-isomorphism, it follows that

$$\tau(p(x_1),\, p(x_2)) \ge 0. \qquad \square$$

Proposition 3.6. *Let $p(t_1, t_2) = c_0 + c_1 t_1 + c_2 t_2 + c_3 t_1^2 + c_4 t_1 t_2 + \bar{c}_4 t_2 t_1 + c_5 t_2^2$ be a self-adjoint noncommutative polynomial in $\mathcal{A}$ satisfying the hypotheses of Theorem 3.4, and let $\mathcal{M}$ be a finite von Neumann algebra with faithful trace state τ. If $0 < x_1, x_2 \le 1$ in $\mathcal{M}$ then $\tau(p(x_1, x_2)) < 0$ if and only if*

$$c_5\|x_2 - \tau(x_2)\|_2^2 + c_3\|x_1 - \tau(x_1)\|_2^2 + 2\,\mathrm{Re}(c_4)\langle x_1 - \tau(x_1),\, x_2 - \tau(x_2)\rangle_2 < -p(\tau(x_1),\, \tau(y_1)).$$

Proof. Writing each $\tau(x_i x_j)$ as $\tau\big((x_i - \tau(x_i)1)(x_j - \tau(x_j)1)\big) + \tau(x_i)\tau(x_j)$, we see that

$$\tau(p(x_1, x_2)) = p(\tau(x_1),\, \tau(y_1)) + c_5\|x_2 - \tau(x_2)\|_2^2 + c_3\|x_1 - \tau(x_1)\|_2^2 + 2\,\mathrm{Re}(c_4)\langle x_1 - \tau(x_1),\, x_2 - \tau(x_2)\rangle_2.$$

The result follows. $\square$

In the rest of this section, we narrow down the possibilities for disproving the CEC using polynomials satisfying the hypotheses of Theorem 3.4 in the nonrotated case, where $\mathrm{Re}(c_4) = 0$. We point out that if $p(t_1, t_2) = c_0 + c_1 t_1 + c_2 t_2 + c_3 t_1^2 + c_5 t_2^2$ is a self-adjoint noncommutative polynomial in $\mathcal{A}$ satisfying the hypotheses of Theorem 3.4 with both $c_5 \ge 0$ and $c_3 \ge 0$, then $\tau(p(x_1, x_2)) \ge 0$ by the proof of Proposition 3.6.

Theorem 3.7. *Let $p(t_1, t_2) = c_0 + c_1 t_1 + c_2 t_2 + c_3 t_1^2 + c_5 t_2^2$ be a self-adjoint noncommutative polynomial in $\mathcal{A}$ satisfying the hypotheses of Theorem 3.4 with $c_3 > 0$, $c_5 < 0$ and such that $c_1 \ge 0$ and $c_2 \le 0$. Then, for any finite von Neumann algebra $\mathcal{M}$ with faithful trace state τ, we have*

$$\tau(p(x_1, x_2)) \ge 0,$$

for any positive definite contractions x_1 and x_2 in $\mathcal{M}$.

Proof. Assume that $p(t_1, t_2)$ satisfies the hypotheses. Suppose that there exists a finite von Neumann algebra $\mathcal{M}$ with faithful trace state τ and positive definite

contractions x_1 and x_2 such that $\tau(p(x_1, x_2)) < 0$. If $c_1 \geq 0$ and $c_2 \leq 0$, then

$$p(t_1, t_2) = c_0 + c_1 t_1 + c_2 t_2 + c_3 t_1^2 + c_5 t_2^2,$$

so $c_0 + (c_1 + c_3\varepsilon)\varepsilon + c_2 + c_5 \geq 0$ for every $\varepsilon > 0$, and hence $c_0 \geq -c_5 - c_2$. Thus

$$\begin{aligned} 0 &> c_0 + c_1 t_1 + c_2 t_2 + c_3 t_1^2 + c_5 t_2^2 \\ &\geq -c_5 - c_2 + c_1 t_1 + c_2 t_2 + c_3 t_1^2 + c_5 t_2^2 \\ &= -c_5(1 - t_2^2) + c_3 t_1^2 + c_1 t_1 - c_2(1 - t_2), \end{aligned}$$

and

$$0 > -c_5\tau(1 - x_2^2) + c_3\tau(x_1^2) + c_1\tau(x_1) - c_2\tau(1 - x_2) \geq 0.$$

This is a contradiction. □

Theorem 3.8. *Let $p(t_1, t_2) = c_0 + c_1 t_1 + c_2 t_2 + c_3 t_1^2 + c_5 t_2^2$ be a self-adjoint non-commutative polynomial in $\mathcal{A}$ satisfying the hypotheses of Theorem 3.4 with $c_3 > 0$, $c_5 < 0$ and such that $c_1 < 0$ and $c_2 = 0$. Then for any finite von Neumann algebra $\mathcal{M}$ with faithful trace state τ,*

$$\tau(p(x_1, x_2)) \geq 0,$$

for any positive definite contractions x_1 and x_2 in $\mathcal{M}$.

Proof. Assume that $p(t_1, t_2)$ satisfies the hypotheses. Let $\mathcal{M}$ be a finite von Neumann algebra with faithful trace state τ and let x_1 and x_2 be positive definite contractions. If $c_1 < 0$ and $c_2 = 0$, then for every $\varepsilon > 0$ letting $t_1 = \varepsilon - c_1/(2c_3)$,

$$c_0 + c_3\varepsilon^2 - \frac{c_1^2}{4c_3} + c_5 \geq 0,$$

and therefore $c_0 \geq \dfrac{c_1^2}{4c_3} - c_5$. Then

$$\begin{aligned} p(t_1, t_2) &= c_0 + c_3\Big(t_1 + \frac{c_1}{2c_3}\Big)^2 - \frac{c_1^2}{4c_3} + c_5 t_2^2 \\ &\geq \frac{c_1^2}{4c_3} - c_5 + c_3\Big(t_1 + \frac{c_1}{2c_3}\Big)^2 - \frac{c_1^2}{4c_3} + c_5 t_2^2 = -c_5(1 - t_2^2) + c_3\Big(t_1 + \frac{c_1}{2c_3}\Big)^2. \end{aligned}$$

Therefore

$$\tau(p(x_1, x_2)) = -c_5\tau(1 - x_2^2) + c_3\tau\left(\left(x_1 + \frac{c_1}{2c_3}\right)^2\right) \geq 0.$$ □

The previous two theorems establish that any polynomial $p(t_1, t_2) = c_0 + c_1 t_1 + c_2 t_2 + c_3 t_1^2 + c_5 t_2^2$ in $\mathcal{A}$ that has a chance to disprove the CEC must satisfy either $c_2 > 0$ or both $c_1 < 0$ and $c_2 < 0$.

References

[Dostál and Hadwin 2003] M. Dostál and D. Hadwin, "An alternative to free entropy for free group factors", *Acta Math. Sin.* (*Engl. Ser.*) **19**:3 (2003), 419–472. MR 2005a:46136 Zbl 1115.46054

[Hadwin 2001] D. Hadwin, "A noncommutative moment problem", *Proc. Amer. Math. Soc.* **129**:6 (2001), 1785–1791. MR 2003a:46101 Zbl 0982.44005

[Kadison and Ringrose 1983] R. V. Kadison and J. R. Ringrose, *Fundamentals of the theory of operator algebras, I: Elementary theory*, Pure and Applied Mathematics **100-I**, Academic Press, New York, 1983. MR 85j:46099

[Rădulescu 1999] F. Rădulescu, "Convex sets associated with von Neumann algebras and Connes' approximate embedding problem", *Math. Res. Lett.* **6**:2 (1999), 229–236. MR 2000c:46119

[Sinclair and Smith 2008] A. M. Sinclair and R. R. Smith, *Finite von Neumann algebras and masas*, London Math. Soc. Lecture Note Series **351**, Cambridge University Press, 2008. MR 2009g:46116 Zbl 1154.46035

[Voiculescu 1994] D. Voiculescu, "The analogues of entropy and of Fisher's information measure in free probability theory, II", *Invent. Math.* **118**:3 (1994), 411–440. MR 96a:46117 Zbl 0820.60001

Received: 2010-07-09 Accepted: 2011-02-26

jbannon@siena.edu *Department of Mathematics, Siena College, Loudonville, NY 12211, United States*

don@math.unh.edu *Department of Mathematics and Statistics, The University of New Hampshire, Durham, NH 03824, United States*

me03jeff@siena.edu *Department of Mathematics, Siena College, Loudonville, NY 12211, United States*

INVOLVE 4:1(2011)

msp

Combinatorial proofs of Zeckendorf representations of Fibonacci and Lucas products

Duncan McGregor and Michael Jason Rowell

(Communicated by Arthur T. Benjamin)

In 1998, Filipponi and Hart introduced many Zeckendorf representations of Fibonacci, Lucas and mixed products involving two variables. In 2008, Artz and Rowell proved the simplest of these identities, the Fibonacci product, using tilings. This paper extends the work done by Artz and Rowell to many of the remaining identities from Filipponi and Hart's work. We also answer an open problem raised by Artz and Rowell and present many Zeckendorf representations of mixed products involving three variables.

1. Preliminaries

Definition 1.1. The *n-th Fibonacci number* is the term f_n of the *Fibonacci sequence* defined recursively by

$$f_0 = 1, \quad f_1 = 1, \quad f_n = f_{n-1} + f_{n-2}.$$

This definition is shifted relative to the standard Fibonacci sequence, which begins at 0. This is done to ensure that the combinatorial interpretation matches our sequence without having to shift indices.

Benjamin and Quinn [2003] presented a combinatorial interpretation for the Fibonacci sequence: f_n is the number of possible tilings of an $1 \times n$ board with 1×2 dominoes and 1×1 squares.[1] They also gave a combinatorial interpretation for a related sequence introduced by Edouard Lucas:

Definition 1.2. The *n-th Lucas number* is the term L_n of the *Lucas sequence*, defined recursively by

$$L_0 = 2, \quad L_1 = 1, \quad L_n = L_{n-1} + L_{n-2}.$$

MSC2000: 05A19, 11B39.

Keywords: number theory, Fibonacci numbers, Zeckendorf representations, combinatorics.

[1] The $1 \times n$ board, or *n-board*, is divided into 1×1 squares, called *cells*. In a *tiling*, the board is entirely covered by tiles without overlap. (A *tile* is either a domino or a square.) Two tilings are equivalent if, given any pair of cells, they belong to the same tile in one tiling if and only if they belong to the same tile in the other.

L_n is the number of possible square-and-domino tilings of an *n-bracelet*, that is, an n-board with ends identified. (One can think of such a board as a ring of curved cells.) We do *not* consider as equivalent tilings superimposable by a rotation or reflection; the equivalence relation is the same as for a linear board (see note 1). An n-bracelet has a designated starting cell and ending cell. If these two cells are covered by the same domino, we say that the board is *out of phase*. Otherwise, the board is *in phase*.

The combinatorial interpretation of f_n and L_n given by Benjamin and Quinn is easy to prove by induction. (For instance, in the linear case, consider the first cell of the n-board: either it's covered by a domino, in which case there are, by the induction assumption, f_{n-2} possible tilings of the $n-2$ leftover cells, or it's covered by a square, in which case there are f_{n-1} possibilities.) Since the introduction of these interpretations, many Fibonacci and Lucas identities have been proved combinatorially. Some identities are presented below and will be used repeatedly throughout the paper.

Lemma 1.1. *For any positive integer $n \geq 0$,*

$$f_n = \begin{cases} f_0 + f_2 + \cdots + f_{n-1} & \text{for } n \text{ odd}, \\ f_1 + f_3 + \cdots + f_{n-1} + 1 & \text{for } n \text{ even}. \end{cases}$$

A combinatorial proof of the odd case of Lemma 1.1 appears as Identity 2 in [Benjamin and Quinn 2003]. The even case can be proved similarly.

In the next proof and later one, we say that a tiling *has a fault at m* if the m-th and $(m+1)$-st cells belong to different tiles.

Lemma 1.2. *For any positive integers $m, n \geq 1$,*

$$f_{m+n} - f_m f_n = f_{m-1} f_{n-1}.$$

Proof. Consider the tilings of an $(m+n)$-board; we know there are f_{m+n} of them. Divide the board into an m-board and an n-board. For tilings that have a fault at m, there are f_m possibilities for the m-board and f_n for the n-board, for a total of $f_m f_n$ possibilities. The complementary case is where there is a domino straddling tiles m and $m+1$. Then we're left with subboards of lengths $m-1$ and $n-1$, and there are $f_{m-1} f_{n-1}$ such possibilities. □

Lemma 1.3. *For any positive integer $n \geq 2$,*

$$L_n = f_n + f_{n-2}.$$

A combinatorial proof of this appears under Identity 32 in [Benjamin and Quinn 2003]. We will repeatedly apply this lemma in our identities that involve Lucas products so that we can work with n-boards rather than bracelets. For example,

$$L_m L_n = f_m f_n + f_{m-2} f_n + f_m f_{n-2} + f_{m-2} f_{n-2}.$$

Each of the four terms on the right-hand side are each of the combinations of two bracelets either being in or out of phase.

Edouard Zeckendorf, an amateur mathematician and a doctor in the Belgian army, proved [1972] an interesting property of Fibonacci numbers (here $\mathbb{N}$ stands for the natural numbers, not including 0):

Theorem 1.4. *Every $N \in \mathbb{N}$ can be expressed uniquely as a sum*

$$\sum_{i=1}^{M} f_{a_i} = N,$$

where $M \in \mathbb{N}$, $a_i \in \mathbb{N}$ for $1 \le i \le M$, and $a_{i+1} > a_i + 1$ for $1 \le i < M$.

We call this decomposition the *Zeckendorf representation of N*. Note that, since $a_{i+1} > a_i + 1$, repeated or consecutive Fibonacci numbers cannot appear in the representation.

An open exercise in [Benjamin and Quinn 2003] lists a number of identities involving Zeckendorf representations of multiples of Fibonacci numbers and asks for combinatorial proofs:

$$\begin{aligned} 2f_n &= f_{n-2} + f_{n+1}, \\ 3f_n &= f_{n-2} + f_{n+2}, \\ 4f_n &= f_{n-2} + f_n + f_{n+2}, \\ &\vdots \end{aligned}$$

Wood [2007] provided combinatorial proofs for several of these identities, but without a unified method. Gerdemann [2009] gave a combinatorial algorithm for finding the Zeckendorf representation of any particular mf_n, but it does not give a general closed-form representation.

Artz and Rowell [2009] found combinatorial proofs of certain Zeckendorf representations of $f_m f_n$ originally proved in [Filipponi and Hart 1998] by other means:

Theorem 1.5. *For $n > 2k + 1$,*

$$f_{2k+1} f_n = \sum_{i=1}^{k+1} f_{n-2k-4+4i}.$$

Theorem 1.6. *For $n > 2k$,*

$$f_{2k} f_n = f_{n-2k} + \sum_{i=1}^{k} f_{n-2k-1+4i}.$$

To sketch the proof for the case of $f_{2k+1} f_n$, one must break the set of all tilings of an $(n + 2k + 1)$-board with a fault at n into many disjoint sets where the closest square is i dominoes away from the fault at n. Further our closest square can be no further than k dominoes away from the fault; therefore, $0 \le i \le k$.

In Sections 2 and 3 we provide combinatorial proofs of additional Zeckendorf representations of Fibonacci and Lucas products given in [Filipponi and Hart 1998], namely those for $2f_m f_n$ and $L_m L_n$. In Section 4 we answer an open problem from [Artz and Rowell 2009] and present many new Fibonacci and Lucas product Zeckendorf representations.

2. The Zeckendorf representation of $2f_m f_n$

A Zeckendorf representation for $2f_m f_n$ was given in [Filipponi and Hart 1998]. We provide a combinatorial proof for this identity, extending the combinatorial methods from [Artz and Rowell 2009].

Theorem 2.1. *For integers k and n such that $n > 2k+1 > 0$,*

$$2f_{2k+1} f_n = f_{n+2k+1} + \sum_{i=1}^{k} f_{n+2k+3-4i} + f_{n-2k-2}.$$

Proof. The tilings of an $(n+2k+1)$-board having a fault at n make up a $f_n f_{2k+1}$-element set. We will partition this set into a union of four sequences of subsets R_i, S_i, T_i, and U_i, for $0 \le i \le k$, according to Figure 1. Specifically, given a $(n+2k+1)$-board tiling having a fault at n, let i be the number of dominos between the fault and a square closest to the fault: then $i \le k$ (there is at least one square in the $(2k+1)$-board to the right of the fault). Next assign this tiling to the set

R_i if there are i dominos adjacent to the fault on each side, followed by a square on each side;

S_i if there are i dominos adjacent to the fault on each side, followed by yet another domino on the left and a square on the right;

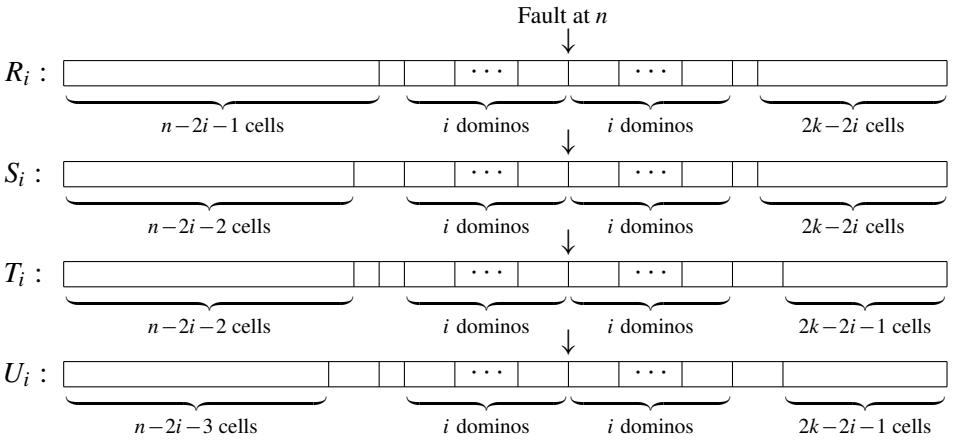


Figure 1. Configurations charaterizing membership in the sets R_i, S_i, T_i and U_i.

T_i if there are i dominos adjacent to the fault on each side, followed by two squares on the left and a domino on the right;

U_i if there are i dominos adjacent to the fault on each side, followed by a square and a domino on the left and a domino on the right.

(Note that T_k and U_k are empty.) Thus, the sets R_i, S_i, T_i, U_i for $0 \le i \le k$ account exactly once for each tiling having a fault at n.

Further, we take a second copy of each of these sets, denoting them by R_i^*, S_i^*, T_i^*, and U_i^*, and we define

$$A_i = R_i \cup R_i^* \cup S_i \cup T_i \cup T_i^* \cup U_i, \qquad B_i = S_i^* \cup U_i^*.$$

It follows that the sets A_i and B_i, for $0 \le i \le k$, account exactly twice for each tiling having a fault at n. Therefore

$$\sum_{i=0}^{k} |A_i \cup B_i| = 2 f_n f_{2k+1},$$

by the first sentence of the proof. To complete the proof, we will show the following equalities:

$$|A_0| = f_{n+2k+1};$$
$$|A_i \cup B_{i-1}| = f_{n+2k+3-4i} \quad \text{for } 1 \le i \le k;$$
$$|B_k| = f_{n-2k-2}.$$

We prove each equality by exhibiting a bijection from the set of tilings of a board of the appropriate size to the set in the left-hand side of the equality. For instance, to show that $|A_0| = f_{n+2k+1}$, we start from the set of all tilings of the $(n+2k+1)$-board; this set, as we know, has f_{n+2k+1} elements. So consider any tiling of the $(n+2k+1)$-board.

- If the tiling has a fault at n and a square next to the fault, on either or both sides, do nothing. This gives an element of $R_0 \cup S_0 \cup T_0 \cup U_0$.
- If the tiling has a fault at n and a domino on both sides of the fault, replace the domino to the left of the fault with two squares, obtaining an element of T_0^*.
- If the tiling does not have a fault at n, split the domino covering cells n and $n+1$ into two squares, obtaining an element of R_0^*.

Since $A_0 = R_0 \cup R_0^* \cup S_0 \cup T_0 \cup T_0^* \cup U_0$ and all elements of the component sets are accounted for, we have shown that $|A_0| = f_{n+2k+1}$.

Next we show that $|A_i \cup B_{i-1}| = f_{n+2k+3-4i}$ for $1 \le i \le k$. Consider any tiling of an $(n+2k+3-4i)$-board, and remove the last tile. Suppose first that the removed tile was a domino, which leaves an $(n+2k+1-4i)$-board.

- If the tiling has a fault at $n-2i$ and a square next to the fault, on either or both sides, insert $2i$ dominos at the fault. This gives an element of $R_i \cup S_i \cup T_i \cup U_i$.
- If the tiling has a fault at $n-2i$ and a domino on both sides of the fault, replace the domino to the left of the fault with two squares and insert $2i$ dominos at the fault, obtaining an element of T_i^*.
- If the tiling does not have a fault at $n-2i$, replace the domino covering the fault with two squares and insert $2i$ dominos between the two squares, obtaining an element of R_i^*.

This accounts for each element of A_i once. Now suppose instead that the tile we removed was a square, which leaves an $(n+2k+2-4i)$-board.

- If the tiling has a fault at $n-2i$, insert $2i-1$ dominos followed by a square at the fault, obtaining an element of S_{i-1}^*.
- If the tiling does not have a fault at $n-2i$, insert a square followed by $2i-1$ dominos just before the domino that covers cell $n-2i$. This gives an element of U_{i-1}^*.

This accounts for each element of B_{i-1} once. Thus $A_i \cup B_{i-1}$ is in bijection with the set of tilings of the $(n+2k+3-4i)$-board.

Lastly, we must show that $|B_k| = f_{n-2k-2}$. Given any tiling of an $(n-2k-2)$-board, append $2k+1$ dominos followed by a square at the right edge, to obtain an element of $B_k = S_k^*$ (recall that U_k^* is empty). This concludes the proof. □

We only present, but do not prove, the case $2f_{2k}f_n$. Its proof is similar to the case presented above and is left to the interested reader.

Theorem 2.2. *For integers k and n such that $n > 2k+1 > 0$,*

$$2f_{2k}f_n = f_{n+2k} + \sum_{i=1}^{k} f_{n+2k+2-4i} + f_{n-2k}.$$

3. Zeckendorf representations of $L_m L_n$

Also given in [Filipponi and Hart 1998] is a Zeckendorf representation of $L_m L_n$. We again extend the notion of squares closest to a given fault to prove our theorem combinatorially.

Lemma 3.1. *Let m and n be positive integers such that $n > m > 1$. Then*

$$f_n f_{m-2} - f_{n-1} f_{m-1} = (-1)^m f_{n-m}.$$

Proof. Let $A^{\{n+m-2,n\}}$ be the set of all tilings of an $(n+m-2)$-board with a fault at n.

For $0 \le i \le \lfloor (m-2)/2 \rfloor$, let $A_{2i}^{\{n+m-2,n\}}$ be the set of all tilings of an $(n+m-2)$-board with a fault at n, i dominos on both sides of the fault and a square at cell

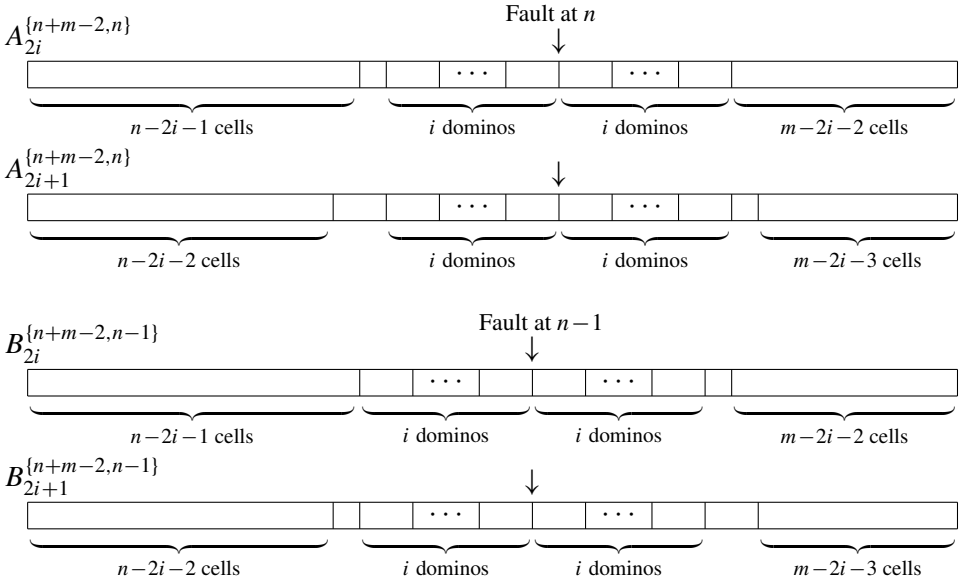


Figure 2. Configurations characterizing membership in various sets.

$n-2i$. For $0 \le i \le \lfloor (m-3)/2 \rfloor$, let $A_{2i+1}^{\{n+m-2,n\}}$ be the set of all tilings of an $(n+m-2)$-board with a fault at n, i dominos on either side of the fault, a domino at cell $n-2i-1$ and a square at cell $n+2i+1$. See Figure 2. For m odd we have

$$A^{\{n+m-2,n\}} = \bigcup_{i=0}^{m-2} A_i^{\{n+m-2,n\}}.$$

If m is even, we need one more set to complete our construction of $A^{\{n+m-2,n\}}$. Let $A_{m-1}^{\{n+m-2,n\}}$ be the set of all tilings of an $(n+m-2)$-board with a fault at n, $m/2-1$ dominos on the right side of the fault and $m/2$ dominos on the left side of the fault. Then

$$A^{\{n+m-2,n\}} = \bigcup_{i=0}^{m-1} A_i^{\{n+m-2,n\}}.$$

Let $B^{\{n+m-2,n-1\}}$ be the set of all tilings of an $(n+m-2)$-board with a fault at $n-1$.

For $0 \le i \le \lfloor (m-2)/2 \rfloor$, let $B_{2i}^{\{n+m-2,n-1\}}$ be the set of all tilings of an $(n+m-2)$-board with a fault at $n-1$, i dominos on either side of the fault and a square at cell $n+2i$. For $0 \le i \le \lfloor (m-3)/2 \rfloor$, let $B_{2i+1}^{\{n+m-2,n-1\}}$ be the set of all tilings of an $(n+m-2)$-board with a fault at $n-1$, i dominos on either side of the fault, a square at cell $n-2i-1$ and a domino at cell $n+2i$. See again Figure 2. For m even we have

$$B^{\{n+m-2,n-1\}} = \bigcup_{i=0}^{m-2} B_i^{\{n+m-2,n-1\}}.$$

If m is odd, we need one more set to complete our construction of $B^{\{n+m-2,n-1\}}$. Let $B_{m-1}^{\{n+m-2,n-1\}}$ be the set of all tilings of an $(n+m-2)$-board with a fault at $n-1$ and $(m-1)/2$ dominos on either side of the fault. Then

$$B^{\{n+m-2,n-1\}} = \bigcup_{i=0}^{m-2} B_i^{\{n+m-2,n-1\}}.$$

Note that $\left|A_i^{\{n+m-2,n\}}\right| = \left|B_i^{\{n+m-2,n-1\}}\right|$ for $0 \le i \le m-2$, since the cardinality of each of these sets is just $f_{n-i-1} f_{m-i-2}$. Thus

$$\left|A^{\{n+m-2,n\}}\right| - \left|B^{\{n+m-2,n-1\}}\right| = \begin{cases} \left|A_{m-1}^{\{n+m-2,n\}}\right| & \text{if } m \text{ is even,} \\ -\left|B_{m-1}^{\{n+m-2\}}\right| & \text{if } m \text{ is odd.} \end{cases}$$

Noting that

$$\left|A^{\{n+m-2,n\}}\right| = f_n f_{m-2}, \quad \left|B^{\{n+m-2,n-1\}}\right| = f_{n-1} f_{m-1},$$
$$\left|A_{m-1}^{\{n+m-1,n\}}\right| = \left|B_{m-1}^{\{n+m-2\}}\right| = f_{n-m},$$

we see that

$$\begin{aligned} f_n f_{m-2} - f_{n-1} f_{m-1} &= \left|A^{\{n+m-2,n\}}\right| - \left|B^{\{n+m-2,n-1\}}\right| \\ &= \begin{cases} \left|A_{m-1}^{\{n+m-2,n\}}\right| & \text{if } m \text{ is even,} \\ -\left|B_{m-1}^{\{n+m-2\}}\right| & \text{if } m \text{ is odd,} \end{cases} \\ &= (-1)^m f_{n-m}. \end{aligned}$$

□

We present four corollaries helpful in proving the Zeckendorf representation of $L_m L_n$. In each of them, an application of Lemma 1.2 is used.

Corollary 3.2. *For integers k and n such that $n > 2k > 1$,*

$$f_n f_{2k-2} - (f_{n+2k} - f_n f_{2k}) = f_{n-2k}.$$

Proof. Let $m \to 2k$ in Lemma 3.1 and note that

$$f_{n-1} f_{2k-1} = f_{n+2k} - f_n f_{2k}.$$

□

Corollary 3.3. *For integers k and n such that $n-2 > 2k > 1$,*

$$f_{n-2} f_{2k-2} - (f_{n+2k-2} - f_{n-2} f_{2k}) = f_{n-2k-2}.$$

Proof. Let $m \to 2k$ and $n \to n-2$ in Lemma 3.1 and note that

$$f_{n-3} f_{2k-1} = f_{n+2k-2} - f_{n-2} f_{2k}.$$

□

Corollary 3.4. *For integers k and n such that $n-1 > 2k+2 > 1$,*

$$(f_{n+2k+1} - f_n f_{2k+1}) - f_{n-2} f_{2k+1} = f_{n-2k-3}.$$

Proof. Let $m \to 2k+2$ and $n \to n-1$ in Lemma 3.1 and note that

$$f_{n-1}f_{2k} = f_{n+2k+1} - f_n f_{2k+1}. \qquad \square$$

Corollary 3.5. *For integers k and n such that $n-1 > 2k > 1$,*

$$(f_{n+2k-1} - f_n f_{2k-1}) - f_{n-2}f_{2k-1} = f_{n-2k-1}.$$

Proof. Let $m \to 2k$ and $n \to n-1$ in Lemma 3.1 and note that

$$f_{n-1}f_{2k-2} = f_{n+2k-1} - f_n f_{2k-1}. \qquad \square$$

Theorem 3.6. *For integers k and n such that $n-2 > 2k > 1$,*

$$L_{2k}L_n = f_{n+2k} + f_{n+2k-2} + f_{n-2k} + f_{n-2k-2}.$$

Proof. By Lemma 1.3 we know that

$$L_{2k}L_n = f_n f_{2k} + f_n f_{2k-2} + f_{n-2}f_{2k} + f_{n-2}f_{2k-2}.$$

Rearranging terms we see that our theorem can be rewritten as

$$f_n f_{2k-2} - (f_{n+2k} - f_n f_{2k}) + f_{n-2}f_{2k-2} - (f_{n+2k-2} - f_{n-2}f_{2k}) = f_{n-2k} + f_{n-2k-2}.$$

Applying Corollaries 3.2 and 3.3 concludes our proof. $\square$

Before moving on to the case $L_n L_{2k+1}$, we need another lemma:

Lemma 3.7. *For integers k and n such that $n+2 > 2k-1 > 0$,*

$$f_{n-2k-4} + f_{n-2k-1} + f_{n+2k+1} + \sum_{j=1}^{2k-1} f_{n-2k+2j} = f_{n+2k+1} + f_{n+2k-1} - f_{n-2k-1} - f_{n-2k-3}.$$

Proof. We will first turn our eye to the summation on the left-hand side of our identity. Applying Lemma 1.1 we can collapse this sum to two terms:

$$\begin{aligned}\sum_{j=1}^{2k-1} f_{n-2k+2j} &= (f_0 + f_2 + \cdots + f_{n+2k-2}) - (f_0 + f_2 + \cdots + f_{n-2k}) \\ &= f_{n+2k-1} - f_{n-2k+1}.\end{aligned}$$

It is left to show that

$$f_{n-2k-4} + f_{n-2k-1} + f_{n+2k+1} + f_{n+2k-1} - f_{n-2k+1} = f_{n+2k+1} + f_{n+2k-1} - f_{n-2k-1} - f_{n-2k-3},$$

or, equivalently,

$$f_{n-2k-4} + f_{n-2k-3} + f_{n-2k-1} = f_{n-2k+1} - f_{n-2k-1}. \tag{3-1}$$

We do this by showing that both sides of our identity count the total number of ways of tiling an $(n-2k)$-board.

On the left-hand side of (3-1) we have all the tilings of an $(n-2k-4)$-board, an $(n-2k-3)$-board and an $(n-2k-1)$-board. To each of the tilings of length $n-2k-4$ add two dominos at the end of the board. To those of length $n-2k-3$ add a square followed by a domino at the end of the board. To the tilings of length $n-2k-1$ add a square at the end of the board. This constructs all tilings of length $n-2k$.

On the right-hand side of (3-1) we have all the tilings of an $(n-2k+1)$-board and an $(n-2k-1)$-board. If we append a domino to all of our tilings of length $n-2k-1$, we see that our right-hand side can be interpreted as all tilings of length $n-2k+1$ that do not end in a domino. Thus, we are counting all tilings of length $n-2k+1$ that end in a square. Removing the square in each of the tilings leaves us with all tilings of length $n-2k$. □

Theorem 3.8. *For integers k and n such that $n-3 > 2k > 1$,*

$$L_{2k+1}L_n = f_{n-2k-4} + f_{n-2k-1} + f_{n+2k+1} + \sum_{j=1}^{2k-1} f_{n-2k+2j}.$$

Proof. Applying Lemmas 1.3 and 3.7, we can rewrite this equality as

$$f_n f_{2k+1} + f_{n-2} f_{2k+1} + f_n f_{2k-1} + f_{n-2} f_{2k-1} = f_{n+2k+1} + f_{n+2k-1} - f_{n-2k-1} - f_{n-2k-3}.$$

Rearranging terms, we see that this is equivalent to

$$f_{n-2k-3} + f_{n-2k-1} = (f_{n+2k+1} - f_n f_{2k+1}) - f_{n-2} f_{2k+1} + (f_{n+2k-1} - f_n f_{2k-1}) - f_{n-2} f_{2k-1}.$$

Applying Corollaries 3.4 and 3.5 concludes our proof. □

4. Answering an open problem and new Zeckendorf representations

In [Artz and Rowell 2009], the following theorem was given and an open problem was posed to find a combinatorial proof. The following proof gives an answer to the open question.

Theorem 4.1. *For integers m and n such that $n > m > 0$,*

$$(f_{m+1} + f_{m-1}) f_n = f_{n+m+1} - (-1)^m f_{n-m-1}.$$

Proof. Let $m \to 2k+1$ in Lemma 3.1. Then

$$f_n f_{2k-1} - f_{n-1} f_{2k} = -f_{n-2k-1}.$$

Applying Lemma 1.2, we see that this is equivalent to

$$f_n f_{2k-1} - (f_{n+2k+1} - f_n f_{2k+1}) = -f_{n-2k-1}.$$

Rearranging terms we see that this proves the case m odd of our theorem. Similarly, we use Corollary 3.2 to prove the case m even. □

Filipponi and Hart introduced Zeckendorf representations of mixed triple products including both Fibonacci and Lucas numbers, namely of the form $f_m^2 L_n$ and $L_m^2 f_n$. We extend their work and present the Zeckendorf representations of a mixed products including three variables. In each of the following identities we assume that our variables take on appropriate integer values.

The remainder of this section was motivated almost entirely by the even case of Theorem 4.1. For sufficiently large values of n, we can ensure that our Zeckendorf representations do not overlap.

Theorem 4.2. *For $n > 2j > m$ and $n > 2j + m$,*

$$f_m L_{2j} f_n = \begin{cases} f_{n+2j-m} + f_{n-2j-m} + \sum_{i=1}^{m/2} f_{n+2j+m-4i-1} + \sum_{i=1}^{m/2} f_{n-2j+m-4i-1} & \text{for } m \text{ even,} \\ \sum_{i=1}^{(m+1)/2} f_{n+2j-m-3+4i} + \sum_{i=1}^{(m+1)/2} f_{n-2j-m-3+4i} & \text{for } m \text{ odd.} \end{cases}$$

Proof. We begin with the first case, say $m = 2k$ for some positive integer k. Applying Theorem 4.1 with $m \to 2j$, followed by Theorem 1.6 with $n \to n+2j$ and $n \to n-2j$, we get

$$\begin{aligned} f_{2k} L_{2j} f_n &= f_{2k}(f_{n+2j} + f_{n-2j}) \\ &= f_{n+2j-2k} + f_{n-2j-2k} + \sum_{i=1}^{k-1} f_{n+2j+2k-4i-1} + \sum_{i=1}^{k-1} f_{n-2j+2k-4i-1}. \end{aligned}$$

Next let $m = 2k+1$ instead. Apply Theorem 4.1 with $m \to 2j$, followed by Theorem 1.5 with $n \to n+2j$ and $n \to n-2j$ to see that

$$\begin{aligned} f_{2k+1} L_{2j} f_n &= f_{2k+1} f_{n+2j} + f_{2k+1} f_{n-2j} \\ &= \sum_{i=1}^{k+1} f_{n+2j-2k-4+4i} + \sum_{i=1}^{k+1} f_{n-2j-2k-4+4i}. \end{aligned}$$ □

Noting that $L_m = f_{m-2} + f_m$, it is easy to extend our previous theorem to the following:

Theorem 4.3. *For $n > 2j > m$ and $n > 2j + m$*

$$L_m L_{2j} f_n = \begin{cases} f_{n-2j-m} + f_{n-2j+m} + f_{n+2j-m} + f_{n+2j+m} & \text{for } m \text{ even,} \\ \sum_{i=1}^{m-1} f_{n+2j-m-1+2i} + \sum_{i=1}^{m} f_{n-2j-m-1+2i} & \text{for } m \text{ odd.} \end{cases}$$

Proof. Let $m = 2k$ for some positive integer k. Applying Theorem 4.1 with $m \to 2j$, followed by Theorem 4.1 twice more with $m \to n + 2j$ and $m \to n - 2j$, we see that

$$L_{2k}L_{2j}f_n = L_{2k}(f_{n+2j} + f_{n-2j}) = f_{n-2j-2k} + f_{n-2j+2k} + f_{n+2j-2k} + f_{n+2j+2k}.$$

If instead $m = 2k+1$, rewriting L_{2k+1} as $f_{2k+1} + f_{2k-1}$ and applying Theorem 4.2 twice yields our result. □

We next consider the Zeckendorf representation of a Lucas triple product.

Lemma 4.4. *For $k > 1$,*

$$2\sum_{i=1}^{k} f_{n+2i-2} = f_{n-2} + f_{n+2k} + \sum_{i=1}^{k-2} f_{n+2i}.$$

Proof. Noting $2f_m = f_{m-2} + f_{m+1}$ [Benjamin and Quinn 2003, Identity 16, page 13], we see that

$$\begin{aligned} 2\sum_{i=1}^{k} f_{n+2i-2} &= \sum_{i=1}^{k} 2f_{n+2i-2} = \sum_{i=1}^{k} f_{n+2i-4} + f_{n+2i-1} = \sum_{i=1}^{k} f_{n-4+2i} + \sum_{i=1}^{k} f_{n+2i-1} \\ &= f_{n-2} + \sum_{i=1}^{k-1} f_{n+2i-2} + \sum_{i=1}^{k-1} f_{n+2i-1} + f_{n+2k-1}. \end{aligned}$$

Finally, noting that $f_m = f_{m-1} + f_{m-2}$, we see that

$$\begin{aligned} 2\sum_{i=1}^{k} f_{n+2i-2} &= f_{n-2} + \sum_{i=1}^{k-1} f_{n+2i-2} + \sum_{i=1}^{k-1} f_{n+2i-1} + f_{n+2k-1} \\ &= f_{n-2} + \sum_{i=1}^{k-1} f_{n+2i} + f_{n+2k-1} = f_{n-2} + \sum_{i=1}^{k-2} f_{n+2i} + f_{n+2k-2} + f_{n+2k-1} \\ &= f_{n-2} + \sum_{i=1}^{k-2} f_{n+2i} + f_{n+2k}. \end{aligned}$$

□

Theorem 4.5. *For $n > 2j > m$ and $n > 2j + m + 2$*

$$L_m L_{2j} L_n = \begin{cases} f_{n-2j-m} + f_{n-2j+m} + f_{n+2j-m} + f_{n+2j+m} + f_{n-2j-m-2} \\ \quad + f_{n-2j+m-2} + f_{n+2j-m-2} + f_{n+2j+m-2} & \text{for } m \text{ even,} \\ f_{n+2j-m-3} + f_{n+2j-m} + \sum_{i=1}^{m} f_{n+2j-m+2i+1} \\ \quad + f_{n-2j-m-3} + f_{n-2j-m} + \sum_{i=1}^{m} f_{n-2j-m+2i+1} & \text{for } m \text{ odd.} \end{cases}$$

Proof. Let $m = 2k$ for some positive integer k. Rewriting L_n as $f_n + f_{n+2}$ and applying Theorem 4.3 twice yields the result for m even.

Let $m = 2k+1$. Rewriting L_n as $f_n + f_{n-2}$ and applying Theorem 4.3 twice we see that

$$L_{2k+1}L_{2j}L_n = f_{n+2j-2k-2} + f_{n+2j+2k} + 2\sum_{i=1}^{2k} f_{n+2j-2k-2+2i} + f_{n-2j-2k-2} + f_{n-2j+2k} + 2\sum_{i=1}^{2k} f_{n-2j-2k-2+2i}.$$

Applying Lemma 4.4 to each of our series with $n \to n+2j-2k$, $k \to 2k$ and $n \to n-2j-2k$, $k \to 2k$, respectively, yields,

$$L_{2k+1}L_{2j}L_n = 2f_{n+2j-2k-2} + f_{n+2j+2k} + f_{n+2j+2k+2} + \sum_{i=1}^{2k-1} f_{n+2j-2k+2i} + 2f_{n-2j-2k-2} + f_{n-2j+2k} + f_{n-2j+2k+2} + \sum_{i=1}^{2k-1} f_{n-2j-2k+2i}.$$

Finally, we will apply Theorem 1.6 with $2k \to 2$ and with $n \to n+2j-2k-2$ and $n \to n-2j-2k-2$, respectively. □

We present our last Zeckendorf representation of a triple product,

Theorem 4.6. *For $n > m > 2j$ and $n > m+2j$,*

$$L_{2j}f_m f_n = \begin{cases} f_{n-m+2j-1} + f_{n+m-2j} + \sum_{i=1}^{j} f_{n-m-2j-3+4i} + \sum_{i=1}^{j} f_{n+m-2j-1+4i} \\ \qquad + \sum_{i=1}^{m/2-j} f_{n-m+2j+4i} \quad \textit{for } m \textit{ odd}, \\ f_{n-m-2j} + f_{n+m-2j} + \sum_{i=1}^{j} f_{n-m-2j-1+4i} + \sum_{i=1}^{j} f_{n+m-2j-1+4i} \\ \qquad + \sum_{i=1}^{m/2-j} f_{n-m+2j-2+4i} \quad \textit{for } m \textit{ even}. \end{cases}$$

Proof. Let $m = 2k$ for some positive integer k. Applying Theorem 1.6 we see that

$$L_{2j}f_{2k}f_n = L_{2j}\big(f_{n-2k} + \sum_{i=1}^{k} f_{n-2k-1+4i}\big).$$

Now distribute L_{2j} and apply Theorem 4.1 to each term. Rearranging terms we see that

$$\begin{aligned} L_{2j}f_{2k}f_n &= f_{n-2k-2j} + f_{n-2k+2j} + \sum_{i=1}^{k} (f_{n-2k-2j-1+4i} + f_{n-2k+2j-1+4i}) \\ &= f_{n-2k-2j} + f_{n-2k+2j} + \sum_{i=1}^{j} f_{n-2k-2j-1+4i} + 2\sum_{i=1}^{k-j} f_{n-2k+2j-1+4i} \\ &\qquad + \sum_{i=1}^{j} f_{n+2k-2j-1+4i}. \end{aligned}$$

We can now apply Theorem 1.6, with $2k \to 2$. Recalling that $f_n = f_{n-1} + f_{n-2}$, we obtain

$$\begin{aligned}
L_{2j} f_{2k} f_n &= f_{n-2k-2j} + f_{n-2k+2j} + \sum_{i=1}^{j} f_{n-2k-2j-1+4i} + \sum_{i=1}^{j} f_{n+2k-2j-1+4i} \\
&\qquad + \sum_{i=1}^{k-j} (f_{n-2k+2j+4i} + f_{n-2k+2j-3+4i}) \\
&= f_{n-2k-2j} + f_{n-2k+2j} + \sum_{i=1}^{j} f_{n-2k-2j-1+4i} + \sum_{i=1}^{j} f_{n+2k-2j-1+4i} \\
&\qquad + f_{n-2k+2j+1} + f_{n+2k-2j} + \sum_{i=1}^{k-j-1} (f_{n-2k+2j+4i} + f_{n-2k+2j+1+4i}) \\
&= f_{n-2k-2j} + f_{n+2k-2j} + \sum_{i=1}^{j} f_{n-2k-2j+4i-1} + \sum_{i=1}^{j} f_{n+2k-2j+4i-1} \\
&\qquad + \sum_{i=1}^{k-j} f_{n-2k+2j+4i-2}.
\end{aligned}$$

We turn to the case m odd, $m = 2k+1$. Applying Theorem 1.5 we can see that

$$L_{2j} f_{2k+1} f_n = L_{2j}\Big(\sum_{i=1}^{k+1} f_{n-2k-4+4i} \Big).$$

Now distribute L_{2j} and apply Theorem 4.1 to each term. Rewriting terms reveals

$$L_{2j} f_{2k+1} f_n = \sum_{i=1}^{j} f_{n-2k-2j-4+4i} + \sum_{i=1}^{j} f_{n+2k-2j+4i} + 2 \sum_{i=1}^{k-j+1} f_{n-2k+2j-4+4i}.$$

We now apply Theorem 1.6 with $2k \to 2$, recalling the recursion relation of the Fibonacci sequence, which shows

$$\begin{aligned}
L_{2j} f_{2k+1} f_n &= \sum_{i=1}^{j} f_{n-2k-2j-4+4i} + \sum_{i=1}^{j} f_{n+2k-2j+4i} \\
&\qquad + \sum_{i=1}^{k-j+1} f_{n-2k+2j-3+4i} + f_{n-2k+2j-6+4i} \\
&= f_{n-2k+2j-2} + f_{n+2k-2j+1} + \sum_{i=1}^{j} f_{n-2k-2j-4+4i} + \sum_{i=1}^{j} f_{n+2k-2j+4i} \\
&\qquad + \sum_{i=1}^{k-j} f_{n-2k+2j-1+4i}. \qquad \square
\end{aligned}$$

5. Conclusions and future work

Having proved the Zeckendorf representation of $2 f_n f_m$, we can see that we can prove individual cases of $k f_n f_m$ using similar methods. Further, Lemma 3.1 seems to hold the key to many interesting Zeckendorf representations involving Lucas numbers. We find it especially intriguing that it led to mixed products of three variables involving even Lucas numbers. We did, however, have little luck finding closed form Zeckendorf representation of $f_p L_m f_n$ where m is odd.

The Zeckendorf representations in Section 4 are proved using many combinatorial mappings of our boards and bracelets to produce their Zeckendorf representations. We believe much insight into the problem could be found by proving each with a single mapping.

References

[Artz and Rowell 2009] J. Artz and M. Rowell, "A tiling approach to Fibonacci product identities", *Involve* **2**:5 (2009), 581–587. MR 2601578 Zbl 1194.05008

[Benjamin and Quinn 2003] A. T. Benjamin and J. J. Quinn, *Proofs that really count: the art of combinatorial proof*, The Dolciani Mathematical Expositions **27**, Mathematical Association of America, Washington, DC, 2003. MR 2004f:05001 Zbl 1044.11001

[Filipponi and Hart 1998] P. Filipponi and E. L. Hart, "The Zeckendorf decomposition of certain Fibonacci–Lucas products", *Fibonacci Quart.* **36**:3 (1998), 240–247. MR 99d:11006 Zbl 0942.11012

[Gerdemann 2009] D. Gerdemann, "Combinatorial proofs of Zeckendorf family identities", *Fibonacci Quart.* **46/47**:3 (2009), 249–261. MR 2010j:11025 Zbl 05614043

[Wood 2007] P. M. Wood, "Bijective proofs for Fibonacci identities related to Zeckendorf's theorem", *Fibonacci Quart.* **45**:2 (2007), 138–145. MR 2009b:05032 Zbl 1162.11014

[Zeckendorf 1972] E. Zeckendorf, "Représentation des nombres naturels par une somme de nombres de Fibonacci ou de nombres de Lucas", *Bull. Soc. Roy. Sci. Liège* **41** (1972), 179–182. MR 46 #7147 Zbl 0252.10011

Received: 2010-08-10 Accepted: 2010-10-24

mcgr5577@pacificu.edu *Department of Mathematics and Computer Science, Pacific University, 2043 College Way, Forest Grove, OR 97116, United States*

rowell@pacificu.edu *Department of Mathematics and Computer Science, Pacific University, 2043 College Way, Forest Grove, OR 97116, United States*
http://www.pacificu.edu/as/math/

INVOLVE 4:1(2011)

msp

A generalization of even and odd functions

Micki Balaich and Matthew Ondrus

(Communicated by Vadim Ponomarenko)

We generalize the concepts of even and odd functions in the setting of complex-valued functions of a complex variable. If $n > 1$ is a fixed integer and r is an integer with $0 \leq r < n$, we define what it means for a function to have type r mod n. When $n = 2$, this reduces to the notions of even ($r = 0$) and odd ($r = 1$) functions respectively. We show that every function can be decomposed in a unique way as the sum of functions of types-0 through $n - 1$. When the given function is differentiable, this decomposition is compatible with the differentiation operator in a natural way. We also show that under certain conditions, the type r component of a given function may be regarded as a real-valued function of a real variable. Although this decomposition satisfies several analytic properties, the decomposition itself is largely algebraic, and we show that it can be explained in terms of representation theory.

1. Introduction

1.1. ***Background.*** The notions of *even* and *odd* functions are well-known to most students of high school and college algebra. A function $f : \mathbb{R} \to \mathbb{R}$ is even if $f(-x) = f(x)$ for all $x \in \mathbb{R}$ and is odd if $f(-x) = -f(x)$ for all $x \in \mathbb{R}$. These concepts are important in many areas of analysis, and there are numerous useful examples of even or odd functions. For example, the function $f(x) = \cos x$ is even, as is any polynomial in x whose nonzero coefficients all correspond to even powers of x. Although there are numerous functions that are neither even nor odd, every function $f : \mathbb{R} \to \mathbb{R}$ decomposes in a unique way as $f = f_e + f_o$, where f_e is even and f_o is odd. For instance, the equation $e^x = \cosh x + \sinh x$ can be thought of as the decomposition of the exponential function e^x into its even and odd parts.

To motivate the following work, we revisit the definitions of even and odd functions and express the defining equations slightly differently. Let $f : \mathbb{R} \to \mathbb{R}$ be a function, and write $\epsilon = -1 \in \mathbb{R}$. Then f is even if

$$f(\epsilon x) = \epsilon^0 f(x) \quad \text{for all } x \in \mathbb{R}, \tag{1}$$

MSC2010: primary 30A99; secondary 20C15.
Keywords: complex function, group, representation.

and f is odd if

$$f(\epsilon x) = \epsilon^1 f(x) \quad \text{for all } x \in \mathbb{R}. \tag{2}$$

In other words, a function is even (or odd) if it satisfies a certain functional equation involving a square root of unity. Note that this definition also makes sense if we replace the field $\mathbb{R}$ with the field $\mathbb{C}$ of complex numbers.

1.2. ***Summary of results.*** In the following, we let $f : \mathbb{C} \to \mathbb{C}$ be a function and fix an integer $n > 1$. If $r \in \mathbb{Z}$ with $0 \leq r < n$, we say that the function f is of type r mod n if f satisfies a certain functional equation (depending upon r) for every n-th root of unity. In the special case that $n = 2$, this definition reduces to the usual notions of even and odd functions. In Theorem 5, we show (for arbitrary n) that every function $f : \mathbb{C} \to \mathbb{C}$ decomposes as $f = f_0 + \cdots + f_{n-1}$, where f_r is of type r mod n. Moreover, we show in Theorem 6 that this decomposition is unique. The set of all functions $f : \mathbb{C} \to \mathbb{C}$ may be regarded as a vector space, and the set of all functions of type r mod n may be regarded as a subspace. Thus we also explain how the decomposition $f = f_0 + \cdots + f_{n-1}$ may be thought of in terms of projections from a vector space onto various subspaces.

We show in Section 3 (Corollary 15) that if a given complex function $f : \mathbb{C} \to \mathbb{C}$ is real (i.e., $f(\mathbb{R}) \subseteq \mathbb{R}$), then under certain assumptions, the functions f_r (in the decomposition $f = f_0 + \cdots + f_{n-1}$) are also real. This explains, for example, why the functions $\cosh : \mathbb{C} \to \mathbb{C}$ and $\sinh : \mathbb{C} \to \mathbb{C}$ produce real outputs when the inputs are restricted to real numbers. In the classical setting of even and odd functions, it is well-known that the derivative of an even (resp. odd) function is odd (resp. even), and in Section 4 we prove several analogous results that apply in our setting. In the case of the exponential function e^z, these analytic results lead to solutions to a familiar real differential equation, and we address this connection between our framework and differential equations in Example 20.

In Section 5, we show that some of these results may be explained by working in the algebraic setting of representation theory. We replace the set of n-th roots of unity with a finite group G, and we replace the field of complex numbers with a $\mathbb{C}[G]$-module, where $\mathbb{C}[G]$ denotes the complex group of G. Under certain conditions, a function $f : V \to W$ decomposes as a sum of functions that satisfy various functional equations analogous to those of Section 2. These conditions are easily satisfied in the setting of a function $f : \mathbb{C} \to \mathbb{C}$.

2. Definitions and basic results

Fix an integer $n > 1$. A complex number ϵ is an n-th root of unity if $\epsilon^n = 1$. We now generalize the definitions in (1) and (2) in the setting where $f : \mathbb{C} \to \mathbb{C}$ is a complex-valued function.

Definition 1. Suppose $f : \mathbb{C} \to \mathbb{C}$ is a function. Fix an integer $n > 1$ and an integer r with $0 \le r < n$. We say that f is of *type* r mod n if

$$f(\epsilon z) = \epsilon^r f(z) \tag{3}$$

for every $z \in \mathbb{C}$ and every n-th root of unity $\epsilon \in \mathbb{C}$.

If there is no danger of ambiguity regarding n, we may shorten the notation and say that a function $f : \mathbb{C} \to \mathbb{C}$ has type r. If, for example $n = 3$, then a function $f : \mathbb{C} \to \mathbb{C}$ may have type 0, type 1, or type 2. The third roots of unity are $\epsilon = 1$, $e^{2\pi i/3}$, or $e^{4\pi i/3}$. Thus a type-0 function satisfies the equations

$$\begin{aligned} f(1z) &= 1 f(z), \\ f(e^{2\pi i/3} z) &= f(z), \\ f(e^{4\pi i/3} z) &= f(z), \end{aligned}$$

a type-1 function satisfies

$$\begin{aligned} f(1z) &= 1 f(z), \\ f(e^{2\pi i/3} z) &= e^{2\pi i/3} f(z), \\ f(e^{4\pi i/3} z) &= e^{4\pi i/3} f(z), \end{aligned}$$

and a type-2 function satisfies

$$\begin{aligned} f(1z) &= 1 f(z), \\ f(e^{2\pi i/3} z) &= e^{4\pi i/3} f(z), \\ f(e^{4\pi i/3} z) &= e^{2\pi i/3} f(z). \end{aligned}$$

If $\epsilon \in \mathbb{C}$ is an n-th root of unity and $\epsilon^k \ne 1$ for all $1 \le k < n$, we say that ϵ is a *primitive* n-th root of unity. For example, in case $n = 3$ above, the primitive third roots of unity are $e^{2\pi i/3}$ and $e^{4\pi i/3}$, whereas 1 is not a primitive third root of unity. The next lemma shows that we do not need to check that a given function satisfies (3) for every n-th root of unity. Rather, it is enough to know that (3) holds for at least one primitive n-th root ϵ.

Lemma 2. *Suppose $f : \mathbb{C} \to \mathbb{C}$ is a function. Fix an integer $n > 1$ and an integer $0 \le r < n$. Let $\epsilon \in \mathbb{C}$ be a primitive n-th root of unity. If f has the property that $f(\epsilon z) = \epsilon^r f(z)$ for all $z \in \mathbb{C}$, then $f(\omega z) = \omega^r f(z)$ for every $z \in \mathbb{C}$ and every n-th root of unity $\omega \in \mathbb{C}$.*

Proof. Note that $\omega = \epsilon^k$ for some integer $0 \le k < n$. It follows that $f(\omega z) = f(\epsilon^k z)$, and we see that $f(\epsilon^k z) = \epsilon^r f(\epsilon^{k-1} z) = \epsilon^r \epsilon^r f(\epsilon^{k-2} z) = \cdots = (\epsilon^r)^{k-1} f(\epsilon^1 z) = (\epsilon^r)^k f(z)$. Hence $f(\omega z) = f(\epsilon^k z) = (\epsilon^r)^k f(z) = (\epsilon^k)^r f(z) = \omega^r f(z)$. □

The following construction gives rise to a function of type r mod n defined in terms of some given function $f : \mathbb{C} \to \mathbb{C}$. We shall see in Theorem 5 that this construction leads to a way to decompose f as a sum of functions of types $0, 1, \ldots,$ and $n-1$.

Definition 3. Suppose $f : \mathbb{C} \to \mathbb{C}$ is a function. Given an integer $n > 1$, a primitive n-th root of unity $\epsilon \in \mathbb{C}$, and $r \in \mathbb{Z}$ with $0 \le r < n$, define $f_{(r,\epsilon)} : \mathbb{C} \to \mathbb{C}$ by

$$f_{(r,\epsilon)}(z) = \frac{1}{n} \sum_{k=0}^{n-1} \epsilon^{-kr} f(\epsilon^k z). \tag{4}$$

Theorem 4. *Suppose $f : \mathbb{C} \to \mathbb{C}$ is a function. Given an integer $n > 1$, a primitive n-th root of unity $\epsilon \in \mathbb{C}$, and $r \in \mathbb{Z}$ with $0 \le r < n$, define $f_{(r,\epsilon)}(z)$ as in Definition 3. Then $f_{(r,\epsilon)}$ is of type r* mod *n.*

Proof. By Lemma 2, it suffices to show that $f_{(r,\epsilon)}(\epsilon z) = \epsilon^r f_{(r,\epsilon)}(z)$. Note that

$$f_{(r,\epsilon)}(\epsilon z) = \frac{1}{n} \sum_{k=0}^{n-1} \epsilon^{-kr} f(\epsilon^{k+1} z) = \frac{1}{n} \sum_{l=1}^{n} \epsilon^{-r(l-1)} f(\epsilon^l z), \quad \text{where } l = k+1.$$

Also note that $\epsilon^{-r(n-1)} f(\epsilon^n z) = \epsilon^{-r(0-1)} f(\epsilon^0 z)$, so

$$f_{(r,\epsilon)}(\epsilon z) = \frac{1}{n} \sum_{l=0}^{n-1} \epsilon^{-r(l-1)} f(\epsilon^l z) = \frac{\epsilon^r}{n} \sum_{l=0}^{n-1} \epsilon^{-rl} f(\epsilon^l z) = \epsilon^r f_{(r,\epsilon)}(z). \qquad \square$$

Theorem 5. *Suppose $f : \mathbb{C} \to \mathbb{C}$ is a function. Fix an integer $n > 1$, and let $\epsilon \in \mathbb{C}$ be a primitive n-th root of unity. Then $f = \sum_{r=0}^{n-1} f_{(r,\epsilon)}$, where $f_{(r,\epsilon)}$ is given by Definition 3.*

Proof. Note that

$$\sum_{r=0}^{n-1} f_{(r,\epsilon)}(z) = \sum_{r=0}^{n-1} \frac{1}{n} \sum_{k=0}^{n-1} \epsilon^{-kr} f(\epsilon^k z) = \frac{1}{n} \sum_{k=0}^{n-1} \left(\sum_{r=0}^{n-1} \epsilon^{-kr} \right) f(\epsilon^k z).$$

Since

$$\sum_{r=0}^{n-1} \epsilon^{-kr} = \sum_{r=0}^{n-1} (\epsilon^{-k})^r = \begin{cases} \dfrac{1-(\epsilon^{-k})^n}{1-\epsilon^{-k}} = 0 & \text{if } 0 < k < n, \\ n & \text{if } k = 0, \end{cases}$$

it follows that

$$\sum_{r=0}^{n-1} f_{(r,\epsilon)}(z) = \frac{1}{n} \sum_{k=0}^{n-1} \left(\sum_{r=0}^{n-1} \epsilon^{-kr} \right) f(\epsilon^k z) = \frac{n}{n} f(\epsilon^0 z) = f(z). \qquad \square$$

Although Theorem 5 asserts that every function can be written as a sum of functions of types 0 through $n-1$, it does not preclude the possibility that this can be done in several ways. Theorem 6 addresses this issue.

Theorem 6. *Suppose $f : \mathbb{C} \to \mathbb{C}$ is a function, and fix an integer $n > 1$. If $f = f_0 + \cdots + f_{n-1}$ and $f = g_0 + \cdots + g_{n-1}$ where f_r and g_r have type r mod n for $0 \le r < n$, then $f_r = g_r$ for all r.*

Proof. Suppose that $f = f_0 + \cdots + f_{n-1}$ and $f = g_0 + \cdots + g_{n-1}$ where f_r and g_r have type r. Then $h_0 + \cdots + h_{n-1} = 0$ where $h_r = f_r - g_r$ has type r for all r. Thus it is sufficient to prove that if $h_0 + \cdots + h_{n-1} = 0$, where h_r has type r, then $h_r = 0$ for all r.

Suppose the result is false. There exists a strictly increasing sequence $r_1, \ldots, r_k$, with $r_i \in \{0, 1, \ldots, n-1\}$ for all i, along with functions q_{r_i} $(i = 1, \ldots, k)$ so that q_{r_i} is a nonzero function of type r_i and

$$q_{r_1} + \cdots + q_{r_k} = 0. \tag{5}$$

Furthermore, we may suppose we have chosen such a counterexample with k minimal.

Let ϵ be a primitive n-th root of unity. Evaluating both sides of (5) at ϵz implies that $0 = (q_{r_1} + \cdots + q_{r_k})(\epsilon z) = \epsilon^{r_1} q_{r_1}(z) + \cdots + \epsilon^{r_k} q_{r_k}(z)$, while multiplying both sides of (5) by ϵ^{r_1} yields $\epsilon^{r_1} q_{r_1} + \epsilon^{r_1} q_{r_2} + \cdots + \epsilon^{r_1} q_{r_k} = 0$. After subtracting these two equations, we see that

$$(\epsilon^{r_2} - \epsilon^{r_1}) q_{r_2} + \cdots + (\epsilon^{r_k} - \epsilon^{r_1}) q_{r_k} = 0.$$

By assumption $q_{r_i} \neq 0$ and $\epsilon^{r_i} - \epsilon^{r_1} \neq 0$ since ϵ is a primitive n-th root of unity and $r_i \neq r_1$. Hence $r_2 \cdots r_k$ is a strictly increasing sequence with $r_2, \ldots, r_k \in \{0, 1, \ldots, n-1\}$ such that $(\epsilon^{r_i} - \epsilon^{r_1}) q_{r_i}$ $(2 \le i \le n-1)$ is a nonzero function of type r_i with $(\epsilon^{r_2} - \epsilon^{r_1}) q_{r_2} + \cdots + (\epsilon^{r_k} - \epsilon^{r_1}) q_{r_k} = 0$, contradicting the fact that k was minimal. It follows that $q_{r_i} = 0$ for all i. □

Corollary 7. *Suppose $f : \mathbb{C} \to \mathbb{C}$ is a function. Fix an integer $n > 1$, and let $r \in \mathbb{Z}$ with $0 \le r < n$. Let $\epsilon, \omega \in \mathbb{C}$ be primitive n-th roots of unity. Then $f_{(r,\epsilon)} = f_{(r,\omega)}$.*

Proof. By Theorem 4 and Lemma 2 it follows that $f_{(r,\epsilon)}(\epsilon z) = \epsilon^r f_{(r,\epsilon)}(z)$ and $f_{(r,\omega)}(\epsilon z) = \epsilon^r f_{(r,\omega)}(z)$ for all $z \in \mathbb{C}$. From Theorem 5 we know that $f = \sum_{r=0}^{n-1} f_{(r,\epsilon)}$ and $f = \sum_{r=0}^{n-1} f_{(r,\omega)}$. Theorem 6 implies that $f_{(r,\epsilon)} = f_{(r,\omega)}$ for all r. □

Remark 8. We have shown that $f_{(r,\epsilon)} = f_{(r,\omega)}$ whenever $\epsilon, \omega \in \mathbb{C}$ are primitive n-th roots of unity. Thus it is unambiguous to define the notation f_r by the equation

$$f_r = f_{(r,\epsilon)}, \tag{6}$$

where ϵ is any primitive n-th root of unity and $f_{(r,\epsilon)}$ is given by Definition 3.

An obvious corollary of Theorem 6 is that there is a unique way to write the zero function as a sum of functions of various types. This can essentially be regarded as the statement that functions of differing types (mod n) are linearly independent, and thus it makes sense to phrase these results in terms of linear algebra.

Definition 9. Let F be the vector space of all functions $f : \mathbb{C} \to \mathbb{C}$. Fix an integer $n > 1$, and let $r \in \mathbb{Z}$ with $0 \leq r < n$. Define $F_r \subseteq F$ by

$$F_r = \{f \in F \mid f \text{ has type } r \text{ mod n}\}.$$

It is straightforward to show that if $f, g \in F_r$ and $c \in \mathbb{C}$, then $cf + g \in F_r$. Thus the subset F_r is in fact a vector subspace of F. Note that Theorem 5 and Theorem 6 may be summarized by noting that F decomposes as

$$F = F_0 \oplus \cdots \oplus F_{n-1}.$$

Definition 10. Let $f \in F$. Fix an integer $n > 1$, and let $r \in \mathbb{Z}$ with $0 \leq r < n$. Define $\pi_r(f)$ to be the unique type-r summand that corresponds to writing f as a sum of functions of types 0 through $n - 1$.

In light of the decomposition $F = F_0 \oplus \cdots \oplus F_{n-1}$, the map π_r is well-defined and may be regarded as the projection from F onto the subspace F_r. From Theorem 5 and Theorem 6, it follows that $\pi_r(f) = f_r$, where f_r is defined as in (6). Although this equation could be used as a definition for $\pi_r : F \to F_r$, Definition 10 has the advantage of that the next results follow almost immediately from this definition.

Lemma 11. *Let $f \in F$. Fix an integer $n > 1$, and let $r \in \mathbb{Z}$ with $0 \leq r < n$. Then $(f_r)_r = f_r$.*

Proof. This is equivalent to the assertion that $\pi_r \circ \pi_r = \pi_r$. Since $\pi_r(f)$ is of type r, Definition 10 implies that $\pi_r(\pi_r(f)) = \pi_r(f)$. □

Lemma 12. *Let $f \in F$. Fix an integer $n > 1$, and let $r, s \in \mathbb{Z}$ with $0 \leq r, s < n$. Then $(f_r)_s = 0$ if $r \neq s$.*

Proof. If $f_r \in F_r$ is decomposed according to the direct sum $F = F_0 \oplus \cdots \oplus F_{n-1}$, then the r-th component of f_r is f_r, and every other component is 0, so it follows that $(f_r)_s = 0$ when $r \neq s$. □

3. Relationship to real-valued functions

Several important complex-valued functions $f : \mathbb{C} \to \mathbb{C}$ have the property that $\overline{f(z)} = f(\overline{z})$ for all $z \in \mathbb{C}$. For example, the functions e^z, $\sin z$, $\cos z$, $\sinh z$, and $\cosh z$ have this property, as do all polynomial functions with real coefficients. In this section, we show that this property carries over to the type-r component of f.

Lemma 13. *Suppose $f : \mathbb{C} \to \mathbb{C}$ is a function with the property that $\overline{f(z)} = f(\overline{z})$ for all $z \in \mathbb{C}$. Fix an integer $n > 1$, and let $r \in \mathbb{Z}$ with $0 \le r < n$. Define π_r as in Definition 10. Then $\overline{\pi_r(f)(z)} = \pi_r(f)(\overline{z})$ for all $z \in \mathbb{C}$.*

Proof. Let ϵ be a primitive n-th root of unity. Since $\overline{z_1 + z_2} = \overline{z_1} + \overline{z_2}$ and $\overline{z_1 \cdot z_2} = \overline{z_1} \cdot \overline{z_2}$ for all $z_1, z_2 \in \mathbb{C}$, it follows that

$$\overline{\pi_r(f)(z)} = \overline{\frac{1}{n}\sum_{k=0}^{n-1} \epsilon^{-kr} f(\epsilon^k z)} = \frac{1}{n}\sum_{k=0}^{n-1} \overline{\epsilon^{-kr}}\, f(\overline{\epsilon^k z}) = \frac{1}{n}\sum_{k=0}^{n-1} \overline{\epsilon^{-kr}}\, f(\overline{\epsilon}^k \overline{z}).$$

Observe that $\overline{\epsilon} = \epsilon^{-1}$ because $|\epsilon| = 1$, and thus

$$\overline{\pi_r(f)(z)} = \frac{1}{n}\sum_{k=0}^{n-1} \overline{\epsilon}^{-kr} f(\overline{\epsilon}^k \overline{z}) = \frac{1}{n}\sum_{k=0}^{n-1} \omega^{-kr} f(\omega^k \overline{z}),$$

where $\omega = \overline{\epsilon}$. Since $\omega = \epsilon^{-1}$, ω is also a primitive n-th root of unity, whence $\overline{\pi_r(f)(z)} = f_{(r,\omega)}(\overline{z}) = \pi_r(f)(\overline{z})$ for all $z \in \mathbb{C}$ by Remark 8. □

Recall that a complex function $f : \mathbb{C} \to \mathbb{C}$ is said to be *real* if $f(x) \in \mathbb{R}$ whenever $x \in \mathbb{R}$. The next lemma provides a criterion to show that a function is real, and a proof can be found in [Churchill and Brown 2008, page 87].

Lemma 14. *Suppose $f : \mathbb{C} \to \mathbb{C}$ is a function. If f has the property that $\overline{f(z)} = f(\overline{z})$ for all $z \in \mathbb{C}$, then f is real.*

The following result is now obvious in light of Lemma 14 and Lemma 13.

Corollary 15. *Suppose $f : \mathbb{C} \to \mathbb{C}$ is a function with the property that $\overline{f(z)} = f(\overline{z})$ for all $z \in \mathbb{C}$. Fix an integer $n > 1$, and let $r \in \mathbb{Z}$ with $0 \le r < n$. If $z \in \mathbb{R}$, then $f_r(z) \in \mathbb{R}$.*

Since $\cosh z$ and $\sinh z$ may be regarded as $\pi_0(e^z)$ and $\pi_1(e^z)$ (with $n = 2$), we recover the obvious facts that $\cosh x, \sinh x \in \mathbb{R}$ if $x \in \mathbb{R}$. More interestingly ($n = 3$), we see for example that if $\epsilon = e^{2\pi i/3} = -\frac{1}{2} + \frac{\sqrt{3}}{2} i$ and $r \in \{0, 1, 2\}$, then $\frac{1}{3}\left(e^x + \epsilon^{-r} e^{\epsilon x} + \epsilon^{-2r} e^{\epsilon^2 x}\right) \in \mathbb{R}$ for all $x \in \mathbb{R}$.

It is not immediately obvious whether the condition that $f : \mathbb{C} \to \mathbb{C}$ is real is sufficient to guarantee that f_r is real whenever $0 \le r < n$. Define $f : \mathbb{C} \to \mathbb{C}$ by

$$f(z) = \begin{cases} 0 & \text{if } z \in \mathbb{R}, \\ i & \text{if } z \in \mathbb{C} \setminus \mathbb{R}. \end{cases}$$

Then, if $n = 3$, it is straightforward to compute that

$$f_0(1) = \frac{2i}{3} \quad \text{and} \quad f_1(1) = f_2(1) = \frac{i}{3}\left(e^{2\pi i/3} + e^{4\pi i/3}\right) = -\frac{i}{3},$$

which implies that $f_r(1) \notin \mathbb{R}$ for $r = 0, 1, 2$. In particular, this shows that f must satisfy a stronger condition (than the condition that f is real) in order to guarantee that f_r is real for $0 \leq r < n$.

4. Relationship to the derivative

Recall that F denotes the space of all functions $f : \mathbb{C} \to \mathbb{C}$.

Definition 16. Define the vector space $\mathscr{F}$ by

$$\mathscr{F} = \{f \in F \mid f \text{ is holomorphic}\}.$$

Definition 17. Let $f \in F$. Fix an integer $n > 1$, and let $r \in \mathbb{Z}$ with $0 \leq r < n$. Define the subspace $\mathscr{F}_r$ by

$$\mathscr{F}_r = \mathscr{F} \cap F_r.$$

If $f : \mathbb{C} \to \mathbb{C}$ is a holomorphic function, the following theorem establishes a relationship between the projection maps π_r and the differentiation operator.

Theorem 18. *Fix an integer $n > 1$, and let $r \in \mathbb{Z}$ with $0 \leq r < n$. Define π_r and π_{r-1} as in Definition 10, and let $\frac{d}{dz} : \mathscr{F} \to \mathscr{F}$ denote the differentiation operator. Then, for $f \in \mathscr{F}$, we have*

$$\left(\frac{d}{dz} \circ \pi_r\right)(f) = \left(\pi_{r-1} \circ \frac{d}{dz}\right)(f),$$

where we read the integer $r - 1$ modulo n.

Proof. Let $f \in \mathscr{F}$ and fix a primitive n-th root of unity $\epsilon \in \mathbb{C}$. Note that by definition

$$\left(\pi_{r-1} \circ \frac{d}{dz}\right)(f)(z) = \pi_{r-1}(f')(z) = \sum_{k=0}^{n-1} \epsilon^{-k(r-1)} f'(\epsilon^k z).$$

From the chain rule, the derivative of the function $z \mapsto f(\epsilon^k z)$ is the function $z \mapsto \epsilon^k f'(\epsilon^k z)$, so we have

$$\left(\frac{d}{dz} \circ \pi_r\right)(f)(z) = \sum_{k=0}^{n-1} \epsilon^k \epsilon^{-kr} f'(\epsilon^k z) = \sum_{k=0}^{n-1} \epsilon^{-k(r-1)} f'(\epsilon^k z). \qquad \square$$

The following corollary generalizes the fact that the derivative of an odd (resp. even) function is even (resp. odd). Although it can be demonstrated directly from the definition [Ahlfors 1979, page 24] of the complex derivative, we prove the result using Theorem 18.

Corollary 19. *Let $f \in \mathscr{F}$. Fix an integer $n > 1$, and let $r \in \mathbb{Z}$ with $0 \leq r < n$. If $f \in \mathscr{F}_r$, then $f' \in \mathscr{F}_{r-1}$, where we read the integer $r - 1$ modulo n.*

Proof. By Theorem 18, $\frac{d}{dz}(f) = \frac{d}{dz}(\pi_r(f)) = \pi_{r-1}\left(\frac{d}{dz}(f)\right) \in \mathscr{F}_{r-1}$. $\square$

Example 20. Fix an integer $n > 1$, and let $f(z) = e^z$. We saw in Corollary 15 that for $0 \le k < n$, $f_k(x) \in \mathbb{R}$ whenever $x \in \mathbb{R}$. Moreover Theorem 18 implies that $df_r/dz = \pi_{r-1}(df/dz) = f_{r-1}$. Thus if we let $f_k|_{\mathbb{R}}$ denote the restriction of f_k to the real numbers, it follows that

$$\frac{d}{dx}(f_r|_{\mathbb{R}}) = f_{r-1}|_{\mathbb{R}},$$

where d/dx denotes the real differentiation operator and the subscripts r and $r-1$ are read modulo n. Thus the function $f_r|_{\mathbb{R}}$ is a solution to the (real) differential equation $d^n y/dx^n = y$. If, for example, $n = 3$, it is straightforward to check that the functions $f_0|_{\mathbb{R}}$, $f_1|_{\mathbb{R}}$, and $f_2|_{\mathbb{R}}$ form a basis for the solution space of $d^n y/dx^n = y$.

5. Relationship to representation theory

The previous setting can be generalized considerably. For a fixed integer $n > 1$, the set G of all n-th roots of unity in $\mathbb{C}$ forms a multiplicative group. This group acts on the space $\mathbb{C}$ as follows. For $g \in G \subseteq \mathbb{C}$ and $z \in \mathbb{C}$, the action is given by $g.z = gz$. (Here, we use the dot notation for group actions, as in [Fulton and Harris 1991].) Thus the domain and codomain of a function $f : \mathbb{C} \to \mathbb{C}$ are G-modules. Because of this, it is natural to conjecture that the above results can explained module-theoretically. Indeed, many of the previous concepts may be regarded as special cases of module-theoretic results. For example, Definition 21 is a module-theoretic analogue of Definition 3, and Corollary 28 yields Theorem 5 as a special case.

If G is a finite group, we define the group algebra $\mathbb{C}[G]$ as in [Isaacs 1976]. We define the notions of a module, a simple module, and a module homomorphism as in [Isaacs 1993] or any other standard text. Note that the function $f : V \to W$ in Definition 21 need not be linear.

Definition 21. Let G be a finite group, and V and W be $\mathbb{C}[G]$-modules. Suppose $f : V \to W$ is a function and $\phi : G \to G$ is a homomorphism. Then define $f_\phi : V \to W$ by

$$f_\phi(v) = \frac{1}{|G|} \sum_{h \in G} \phi(h^{-1}).f(h.v).$$

Note that if G is the group of n-th roots of unity in $\mathbb{C}$ and $\phi : G \to G$ is given by $\phi(g) = g^r$, then the function f_ϕ is exactly the function $f_{(r,\epsilon)}$ given in Definition 3. The following theorem states that not only does f_ϕ generalize $f_{(r,\epsilon)}$, but it behaves in a manner that generalizes Theorem 4.

Theorem 22. *Let G be a finite group and V and W be $\mathbb{C}[G]$-modules. Suppose $f : V \to W$ is a function and $\phi : G \to G$ is a homomorphism. Then $f_\phi(g.v) = \phi(g).f_\phi(v)$.*

Proof. From the definition of f_ϕ, we have, with $u = hg$,

$$f_\phi(g.v) = \frac{1}{|G|}\sum_{h\in G}\phi(h^{-1}).f(hg.v) = \frac{1}{|G|}\sum_{u\in G}\phi(gu^{-1}).f(u.v)$$
$$= \phi(g)\frac{1}{|G|}\sum_{u\in G}\phi(u^{-1}).f(u.v) = \phi(g).f_\phi(v). \qquad \square$$

In the case where G is the n-th roots of unity in $\mathbb{C}$ and $\phi(g) = g^r$ then the properties of the homomorphism f_ϕ are identical to those of the function $f_{(r,\epsilon)}$. The following Theorem shows that the property of $f_{(r,\epsilon)}$ shown in Lemma 11 not only holds under these conditions, but also in the more abstract setting of Theorem 22.

Theorem 23. *Let G be a finite group, and V and W be $\mathbb{C}[G]$-modules. Suppose $f : V \to W$ is a function and $\phi : G \to G$ is a homomorphism. Then $(f_\phi)_\phi = f_\phi$.*

Proof. By definition $f_\phi(v) = \frac{1}{|G|}\sum_{h\in G}\phi(h^{-1}).f(h.v)$. It follows that

$$((f_\phi)_\phi)(v) = \frac{1}{|G|}\sum_{h\in G}\phi(h^{-1}).f_\phi(h.v) = \frac{1}{|G|}\sum_{h\in G}\phi(h^{-1})\phi(h).f_\phi(v)$$
$$= \frac{1}{|G|}\sum_{h\in G}\phi(h^{-1}h).f_\phi(v) = \frac{1}{|G|}(nf_\phi(v)) = f_\phi(v). \qquad \square$$

When G is cyclic of order n, every homomorphism from G to G is determined by the image of some generator of G. For $0 \le r < n$, define $\phi_r : G \to G$ by $\phi_r(x) = x^r$ for all $x \in G$. Then the set of homomorphisms $G \to G$ is $\{\phi_r \mid 0 \le r < n\}$. As Corollary 24 shows, this new setting allows us to generalize the property of $f_{(r,\epsilon)}$ from Theorem 4 in slightly more specific terms than those of Theorem 22.

Corollary 24. *Let G be a finite cyclic group and V and W be $\mathbb{C}[G]$-modules. Suppose $f : V \to W$ is a function, and let $\phi_r : G \to G$ be the homomorphism given by $\phi_r(x) = x^r$. Then $f_{\phi_r}(x.v) = x^r.f_{\phi_r}(v)$ for all $v \in V$ and $x \in G$.*

If G is cyclic and V and W are $\mathbb{C}[G]$-modules with W simple, then it is possible to generalize Theorem 5. To demonstrate this, we rely on the following well-known fact, whose proof can be found in [Isaacs 1976].

Lemma 25 (Schur's Lemma). *Let G be a finite group, and suppose V and W are simple $\mathbb{C}[G]$-modules and $\phi : V \to W$ is a module homomorphism.*

(1) *Either ϕ is an isomorphism or $\phi = 0$.*

(2) *If $V = W$ then $\phi : W \to W$ is a scalar multiple of the identity function.*

If $x \in G$ is central in G, then the function $f_x : W \to W$ defined by $f_x(w) = x.w$ is a $\mathbb{C}[G]$-module homomorphism. Thus Schur's Lemma implies that every central element of G acts by a scalar on any simple module W. With G cyclic, every

element in G is central. In particular, the generator $g \in G$ is central, so there must be some scalar ξ by which g acts on the elements of simple modules. Furthermore, G is finite so $|G| = n$ for some integer n. The next lemma shows that this integer allows us to be somewhat precise about the value of $\xi \in \mathbb{C}$.

Lemma 26. *Let G be a finite cyclic group with generator g and $|G| = n$. If W is a simple $\mathbb{C}[G]$-module, then g acts on W as multiplication by an n-th root of unity.*

Proof. The group G is abelian, so by Schur's Lemma, there exists $\xi \in \mathbb{C}$ so that $g.w = \xi w$ for all $w \in W$. This implies that an arbitrary element $g^k \in G$ acts by the scalar ξ^k. Since $|G| = n$, g^n is the identity element of G, and it follows that for $w \in W$, $w = g^n.w = \xi^n w$, which forces $\xi^n = 1$. □

In light of Theorem 23 and Corollary 24, it is reasonable to conjecture that there is some module-theoretic analogue of Theorem 5. The following theorem establishes a formula for the sum of the functions $f_{\phi_0}, f_{\phi_1}, \ldots, f_{\phi_{n-1}}$. As a consequence of working in this more general setting, the resulting formula is more complicated than the formula in Theorem 5.

Theorem 27. *Let G be a finite cyclic group with generator g and $|G| = n$. Let V, W be $\mathbb{C}[G]$-modules with W simple, and let $f : V \to W$. If g acts on all $w \in W$ by the scalar ξ having multiplicative order d, then for all $v \in V$,*

$$\sum_{r=0}^{n-1} f_{\phi_r}(v) = \sum_{\substack{0 \le k < n \\ d \mid k}} f(g^k.v).$$

Proof. For $v \in V$,

$$\sum_{r=0}^{n-1} f_{\phi_r}(v) = \frac{1}{n} \sum_{r=0}^{n-1} \sum_{k=0}^{n-1} (g^{-k})^r . f(g^k.v) = \frac{1}{n} \sum_{k=0}^{n-1} \left(\sum_{r=0}^{n-1} (\xi^{-k})^r \right) f(g^k.v).$$

Observe that

$$\sum_{r=0}^{n-1} (\xi^{-k})^r = \begin{cases} \dfrac{1-(\xi^{-}k)^n}{1-\xi^{-k}} = 0 & \text{if } d \nmid k \\ n & \text{if } d \mid k. \end{cases}$$

Hence

$$\frac{1}{n} \sum_{k=0}^{n-1} \left(\sum_{r=0}^{n-1} (\xi^{-k})^r \right) f(g^k.v) = \frac{1}{n} \sum_{\substack{0 \le k < n \\ d \mid k}} n f(g^k.v) = \sum_{\substack{0 \le k < n \\ d \mid k}} f(g^k.v),$$

and the desired result follows. □

Lemma 26 does not make it clear which n-th root of unity ξ is. If ξ happens to be primitive, then $|\xi| = |G| = n$. Applying this reasoning to Theorem 27 leads directly to the following module-theoretic generalization of Theorem 5.

Corollary 28. *Let G be a finite cyclic group with generator g and $|G| = n$. Let V and W be $\mathbb{C}[G]$-modules with W simple, and $f : V \to W$. Let $\xi \in \mathbb{C}$ be the n-th root of unity with the property that $g.w = \xi w$ for all $w \in W$. If ξ is a primitive n-th root of unity, then $f = \sum_{r=0}^{n-1} f_{\phi_r}$.*

Proof. Theorem 27 implies that $\sum_{r=0}^{n-1} f_{\phi_r}(v) = \sum_{k\in\Delta} f(g^k.v)$, where

$$\Delta = \{0 \le k < n \mid n \text{ divides } k\}.$$

But $\Delta = \{0\}$, so it follows that $\sum_{r=0}^{n-1} f_{\phi_r}(v) = f(v)$. □

This framework obviously applies in the setting of a function $f : \mathbb{C} \to \mathbb{C}$, and thus many of the results of Section 2 may be regarded as consequences of representation theory. With the current perspective, it is, for example, possible to decompose functions of the form $f : V \to \mathbb{C}$, where V is any module for the group G of complex n-th roots of unity. For instance, if V is the set of all $m \times m$ matrices with complex entries, then G acts on V by entry-wise multiplication. Alternatively, if V is taken to be the group algebra $\mathbb{C}[G]$, then G acts on V via the regular action, and this setting applies to functions $f : \mathbb{C}[G] \to \mathbb{C}$.

Acknowledgments

The authors thank the referee for numerous valuable suggestions that improved the quality of this paper.

References

[Ahlfors 1979] L. V. Ahlfors, *Complex analysis: an introduction to the theory of analytic functions of one complex variable*, 3rd ed., McGraw-Hill, New York, 1979. MR 80c:30001 Zbl 0395.30001

[Churchill and Brown 2008] R. V. Churchill and J. W. Brown, *Complex variables and applications*, 8th ed., McGraw-Hill, New York, 2008.

[Fulton and Harris 1991] W. Fulton and J. Harris, *Representation theory: a first course*, Grad. Texts in Math. **129**, Springer, New York, 1991. MR 93a:20069

[Isaacs 1976] I. M. Isaacs, *Character theory of finite groups*, Pure and Applied Math. **69**, Academic Press, New York, 1976. Reprinted Dover, New York, 1994. MR 57 #417 Zbl 0337.20005

[Isaacs 1993] I. M. Isaacs, *Algebra: a graduate course*, Brooks/Cole, Pacific Grove, CA, 1993. MR 95k:00003 Zbl 1157.00004

Received: 2010-09-14 Revised: 2011-05-02 Accepted: 2011-05-25

mickibalaich@weber.edu *Mathematics Department, Weber State University, 1702 University Circle, Ogden, UT 84408, United States*

mattondrus@weber.edu *Mathematics Department, Weber State University, 1702 University Circle, Ogden, UT 84408, United States*

Guidelines for Authors

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, Corel Draw, 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@mathscipub.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

2011 vol. 4 no. 1

1944-4176(2011)4:1;1-E