# involve

## a journal of mathematics





msp

2019 vol. 12, no. 3

# involve

msp.org/involve


## INVOLVE YOUR STUDENTS IN RESEARCH

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




See inside back cover or msp.org/involve for submission instructions. The subscription price for 2019 is US $195/year for the electronic version, and $260/year (+$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP.

Involve (ISSN 1944-4184 electronic, 1944-4176 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o 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**

**nonprofit scientific publishing**

http://msp.org/

INVOLVE 12:3 (2019)
dx.doi.org/10.2140/involve.2019.12.361

# Darboux calculus

Marco Aldi and Alexander McCleary

(Communicated by Kenneth S. Berenhaut)

We introduce a formalism to analyze partially defined functions between ordered sets. We show that our construction provides a uniform and conceptual approach to all the main definitions encountered in elementary real analysis including Dedekind cuts, limits and continuity.

## 1. Introduction

Following the pioneering work of Bolzano and Weierstrass, "$(\varepsilon, \delta)$-definitions" are at the heart of textbook presentations of elementary analysis; see, e.g., [Rudin 1953]. While with practice the motivated student quickly becomes proficient in this language, it is natural to ask if fundamental notions such as limits, continuity and integrals could perhaps be defined more conceptually.

In the present paper we develop a rather general framework, which we refer to as *Darboux calculus*, whose specialization to the context of real analysis provides a unified and conceptual approach to all the main definitions encountered in, say, single variable calculus. Our starting point is the observation that the completeness of the ordered set of extended real numbers $\widehat{\mathbb{R}} = \{\pm\infty\} \cup \mathbb{R}$ is equivalent to the validity of the following.

**Lemma 1.1.** *Let $\mathcal{O}$ be a (partially) ordered set, let $\mathcal{S} \subseteq \mathcal{O}$ be any subset and let $\psi : \mathcal{S} \to \widehat{\mathbb{R}}$ be an order-preserving function. Then the set of order-preserving functions $f : \mathcal{O} \to \widehat{\mathbb{R}}$ whose restriction to $\mathcal{S}$ coincides with $\psi$ has a maximum and a minimum.*

In particular, such an order-preserving function $\psi$ singles out a distinguished subset $\mathrm{Dar}(\psi) \subseteq \mathcal{O}$, the *Darboux set of* $\psi$, of elements on which the maximum and minimum extensions of $\psi$ coincide. Equivalently, $\mathrm{Dar}(\psi)$ can be thought of as the subset to which $\psi$ extends canonically. We denote this canonical extension by $\mathrm{ex}_\psi$.

The prototypical example of this construction is provided by the Darboux integral. Let $\mathcal{O}$ denote the set of all bounded functions on an interval $[a, b] \subseteq \mathbb{R}$, let $\mathcal{S}$ be the

*MSC2010:* 06A06, 06A11, 18B35, 26A06, 97I10.
*Keywords:* partially ordered sets, Kan extensions, foundations of real analysis.

subset of step functions and let $\psi$ be the function that to each step function assigns its integral defined naively in terms of signed areas of rectangles. In this case, as shown in Example 7.9 below, $\mathrm{Dar}(\psi)$ coincides with the set of Darboux integrable functions on $[a, b]$ and $\mathrm{ex}_\psi$ is the Darboux integral.

This approach to the Darboux integral exemplifies the philosophy of this paper: naturally occurring pairs $(\mathcal{X}, \varphi)$ consisting of a class $\mathcal{X}$ of $\widehat{\mathbb{R}}$-valued functions and an order-preserving function $\varphi : \mathcal{X} \to \widehat{\mathbb{R}}$ are of the form $(\mathrm{Dar}(\psi), \mathrm{ex}_\psi)$ for a suitable order-preserving function $\psi$ defined on a subset $\mathcal{S} \subseteq \mathcal{X}$ of functions that "obviously belong to $\mathcal{X}$".

For instance, let $\mathcal{O}$ be the set of all sequences of real numbers, let $\mathcal{S}$ be the subset of sequences that are eventually constant and let $\psi$ be the function that to each sequence $\eta \in \mathcal{S}$ assigns the only value that $\eta$ attains infinitely many times. Then, as shown in as shown in Example 7.5 below, $\mathrm{Dar}(\psi)$ coincides with the set of convergent (possibly to $\pm\infty$) sequences and $\mathrm{ex}_\psi(f) = \lim_n f(n)$ for every $f \in \mathrm{Dar}(\psi)$. The advantage here is that instead of having to come up with a clever $(\varepsilon, \delta)$-definition of limit of a sequence we only need to prescribe the obvious limit of an eventually constant sequence and the formalism of Darboux calculus automatically takes care of the general case.

Similarly, let $\mathcal{O}$ be the set of all functions $f : \mathbb{R} \to \mathbb{R}$ and fix $x_0 \in \mathbb{R}$. It is shown in Example 7.6 that if $\mathcal{S}$ denotes the set of all functions that are constant on some open neighborhood of $x_0$ and $\psi$ is the function that to each $\eta \in \mathcal{S}$ assigns $\psi(\eta) = \eta(x_0)$, then $\mathrm{Dar}(\psi)$ is the set of functions that are continuous at $x_0$ and $\mathrm{ex}_\psi(f) = f(x_0)$ for all $f \in \mathrm{Dar}(\psi)$. Once again, given as only input the set of functions that are obviously continuous at $x_0$, our machinery returns the set of functions that are continuous at $x_0$ as output. We view this as an intuitive alternative to the standard $(\varepsilon, \delta)$-definition of continuity.

The statement of Lemma 1.1 holds more generally if $\widehat{\mathbb{R}}$ is replaced with any ordered set that is complete in the sense that every subset has a least upper bound and a greatest lower bound. Furthermore, the inclusion $\iota : \mathcal{S} \hookrightarrow \mathcal{O}$ can be replaced with an arbitrary embedding of ordered sets. In fact, the reader familiar with category theory will easily recognize the maximum and minimum extensions of $\psi$ in Lemma 1.1 as, respectively, the right and left Kan extensions [Mac Lane 1971] (assuming they exist) of $\psi$ along $\iota$. Similarly, the Darboux set of $\psi$ can be thought of as the equalizer of the left and right Kan extensions. Here we are implicitly using the standard interpretation of an ordered set $\mathcal{O}$ as a category whose objects are the elements $x \in \mathcal{O}$ and such that $\mathrm{Hom}(x, y)$ consists of a single element if $x \leq y$ and is empty otherwise. From the vantage point of category theory, the present paper can be summarized as the observation that equalizers of left and right Kan extensions arise naturally in elementary analysis. While some of our propositions and theorems are particular instances of much more general results about left and

right Kan extensions, we choose to give self-contained proofs in the case of ordered sets. In this way, we hope to provide evidence of the effectiveness of Darboux calculus as a stand-alone approach to the foundations of analysis that might be one day used to teach the subject at the undergraduate level.

An example of the flexibility of categorical thinking in this context comes from looking at the Yoneda embedding of an ordered set $\mathcal{O}$ into the set of order-preserving functions from $\mathcal{O}$ to the unique (up to a unique isomorphism) nontrivial ordered set with two elements. As it turns out, the Darboux set of the identity function of the image of the Yoneda embedding essentially coincides with the Dedekind–MacNeille completion of $\mathcal{O}$. While the idea of understanding Dedekind cuts in terms of presheaves is not new, see, e.g., [Taylor 1999], our emphasis is on the fact that Darboux sets are not only effective in isolating interesting classes of $\mathbb{R}$-valued functions but can be used to construct $\mathbb{R}$ itself! In fact we show that with a little more effort, the field structure of $\mathbb{R}$ can also be recovered from that of $\mathbb{Q}$ in terms of Kan extensions. Our exposition appears to be somewhat more succinct, direct and self-contained than previous treatments of elementary analysis based on category theory; see, e.g., [Univalent Foundations 2013; Taylor 2010; Edalat and Lieutier 2004]. It would be interesting to carry out a detailed comparison between these approaches and the one presented here.

The paper is organized as follows. Section 2 contains basic material on ordered sets and order-preserving functions. In Section 3 we introduce the main concepts used in this paper, including Darboux sets and Darboux extensions. Section 4 is devoted to the notion of completeness defined here in terms of extensions of partially defined order-preserving functions. As we show, our definition, which we refer to as Darboux completeness, is in fact equivalent to the more familiar notion of Dedekind completeness. In Sections 5–6 we discuss the Yoneda embedding and the Darboux completion of an arbitrary ordered set. In particular in Section 5 we use Darboux extensions to prove that completely integrally closed subgroups of automorphisms of a complete ordered set lift to automorphisms of the completion, a result that we use to construct the field operations on $\mathbb{R}$. Our strategy here can be thought of as a Darboux-theoretic version of the approach used in [Fuchs 1963] to establish similar results directly at the level of Dedekind cuts. Once the real numbers are constructed, in Section 7 we shift our attention to ordered sets of $\mathbb{R}$-valued functions. We prove that an $\mathbb{R}$-valued function $f$ has limit with respect to some filter basis $F$ (in the sense that each $\varepsilon$-neighborhood of the limit contains the image $f(\mathcal{S})$ of some $\mathcal{S} \in F$) if and only if $f$ is in the Darboux set of the partial function defined by assigning to each function constant on some $\mathcal{S} \in F$ the only value that it attains on $\mathcal{S}$. This characterization of convergence with respect to a filter basis yields at once Darboux-theoretic formulations of several $(\varepsilon, \delta)$-definitions such as limits of sequences, limits of functions of one real variable and continuity. After discussing

Darboux integrability (after which the general notion of Darboux set is modeled), we use Darboux calculus to prove a theorem which simultaneously generalizes the usual linearity theorems for limits, continuous functions and integrals. In fact, all the major theorems of elementary real analysis (e.g., the intermediate value theorem, the extreme value theorem and the fundamental theorem of calculus) can be proved conceptually using the language of Darboux calculus. We hope to come back to this point elsewhere and ultimately provide an exhaustive and fully self-contained treatment of elementary real analysis in the language of this paper.

## 2. Preliminaries on ordered sets

**Definition 2.1.** A (*partially*) *ordered set* is a set $\mathcal{O}$ together with a reflexive, anti-symmetric, and transitive relation, which we denote by $\leq$.

**Example 2.2.** If $\mathcal{O}$ is an ordered set, every subset $\mathcal{S} \subseteq \mathcal{O}$ inherits an induced order. For every $x, y \in \mathcal{O}$ such that $x \leq y$, the *interval with endpoints x and y* is the (ordered) subset $[x, y]$ of all $z \in \mathcal{O}$ such that $x \leq z \leq y$.

**Example 2.3.** A *discrete* set is an ordered set with the trivial order with respect to which $x \leq y$ if and only if $x = y$. If $\mathcal{O}$ is an ordered set, we denote by $|\mathcal{O}|$ its underlying discrete set.

**Remark 2.4.** If $\mathcal{O}$ is an ordered set, we denote by $\mathcal{O}^{\mathrm{op}}$ the *opposite* ordered set such that $|\mathcal{O}^{\mathrm{op}}| = |\mathcal{O}|$ and $x \leq y$ in $\mathcal{O}^{\mathrm{op}}$ if and only if $y \leq x$ in $\mathcal{O}$.

**Example 2.5.** Given two ordered sets $\mathcal{O}_1, \mathcal{O}_2$, we denote by $\mathcal{O}_1 \times \mathcal{O}_2$ the ordered set such that $|\mathcal{O}_1 \times \mathcal{O}_2| = |\mathcal{O}_1| \times |\mathcal{O}_2|$ with order such that $(x_1, x_2) \leq (y_1, y_2)$ if and only if $x_1 \leq y_1$ and $x_2 \leq y_2$.

**Definition 2.6.** Let $\mathcal{O}$ and $\mathcal{P}$ be ordered sets. The set of *order-preserving functions from $\mathcal{O}$ to $\mathcal{P}$* is

$$\mathrm{OP}(\mathcal{O}, \mathcal{P}) = \{f : |\mathcal{O}| \to |\mathcal{P}| \mid f(x) \leq f(y) \text{ if } x \leq y\}.$$

We view $\mathrm{OP}(\mathcal{O}, \mathcal{P})$ as an ordered set such that $f \leq g$ if and only if $f(x) \leq g(x)$ for all $x \in \mathcal{O}$. We use the shorthand notation $\mathrm{OP}(\mathcal{O}) = \mathrm{OP}(\mathcal{O}, \mathcal{O})$. We also say that $f \in \mathrm{OP}(\mathcal{O}, \mathcal{P})$ is an *embedding* if for any $x, y \in \mathcal{O}$, $f(x) \leq f(y)$ implies $x \leq y$. An *isomorphism* is a surjective embedding. Given an ordered set $\mathcal{O}$, we denote by $\mathrm{Aut}(\mathcal{O})$ the group of all isomorphisms in $\mathrm{OP}(\mathcal{O})$.

**Definition 2.7.** If $\mathcal{O}$ is an ordered set, we define its *augmentation* to be the ordered set $\widehat{\mathcal{O}}$ such that

(1) $|\widehat{\mathcal{O}}| = |\mathcal{O}| \cup \{-\infty, +\infty\}$;

(2) the canonical inclusion of $|\mathcal{O}|$ into $|\widehat{\mathcal{O}}|$ defines an embedding of $\mathcal{O}$ into $\widehat{\mathcal{O}}$;

(3) $\widehat{\mathcal{O}} = [-\infty, +\infty]$.

**Definition 2.8.** Let $\mathcal{O}$ and $\mathcal{P}$ be ordered sets. A *partial function* $\psi : \mathcal{O} \rightharpoonup \mathcal{P}$ *from* $\mathcal{O}$ *to* $\mathcal{P}$ is an order-preserving function $\psi : \mathrm{dom}(\psi) \to \mathcal{P}$ defined on an ordered subset $\mathrm{dom}(\psi) \subseteq \mathcal{O}$ called the *domain* of $\psi$. The ordered set $\mathrm{im}(\psi) = \psi(\mathrm{dom}(\psi))$ is called the *image* of $\psi$. An *extension of* $\psi$ *to* $\mathcal{O}$ is an order-preserving function $f : \mathcal{O} \to \mathcal{P}$ whose restriction $f|_{\mathrm{dom}(\psi)}$ to $\mathrm{dom}(\psi)$ coincides with $\psi$.

**Example 2.9.** Let $\mathcal{O}$ be an ordered set and let $\mathbf{1}$ be the unique (up to a unique isomorphism) ordered set with one element. Then $\mathcal{O}$ is canonically identified with $\mathrm{OP}(\mathbf{1}, \mathcal{O})$.

**Definition 2.10.** Let $\mathcal{O}$ and $\mathcal{P}$ be ordered sets. A set $\Psi$ of partial functions from $\mathcal{O}$ to $\mathcal{P}$ is *compatible* if for any $\psi', \psi'' \in \Psi$, the restrictions of $\psi'$ and $\psi''$ to $\mathrm{dom}(\psi') \cap \mathrm{dom}(\psi'')$ coincide. If $\Psi$ is compatible, we define its *common extension* to be the partial function $\psi : \mathcal{O} \rightharpoonup \mathcal{P}$ such that

$$\mathrm{dom}(\psi) = \bigcup_{\psi' \in \Psi} \mathrm{dom}(\psi')$$

and $\psi(x) = \psi'(x)$ for every $x \in \mathrm{dom}(\psi')$ and for every $\psi' \in \Psi$.

**Remark 2.11.** Let $\mathcal{O}$ and $\mathcal{P}$ be ordered sets. If $\psi$ is the common extension of a compatible set $\Psi$ of partial functions from $\mathcal{O}$ to $\mathcal{P}$, then $f : \mathcal{O} \to P$ is an extension of $\psi$ to $\mathcal{O}$ if and only if it is an extension of $\psi'$ to $\mathcal{O}$ for each $\psi' \in \Psi$.

## 3. Darboux sets and Darboux extensions

**Definition 3.1.** Let $\mathcal{O}$ and $\mathcal{P}$ be ordered sets. A partial function $\psi : \mathcal{O} \rightharpoonup \mathcal{P}$ is *extremizable* if there exist extensions $\mathrm{lex}_\psi, \mathrm{uex}_\psi : \mathcal{O} \to \mathcal{P}$ of $\psi$ to $\mathcal{O}$ such that $\mathrm{lex}_\psi \leq f \leq \mathrm{uex}_\psi$ for all extensions $f : \mathcal{O} \to \mathcal{P}$ of $\psi$ to $\mathcal{O}$. If this is the case, we call $\mathrm{lex}_\psi$ and $\mathrm{uex}_\psi$ the *lower and upper extensions* of $\psi$, respectively.

**Remark 3.2.** Let $\psi : \mathcal{O} \to \mathcal{P}$ be an extremizable partial function. If $x \in \mathrm{dom}(\psi)$, then $\mathrm{lex}_\psi(x) = \mathrm{uex}_\psi(x)$. Therefore, if $f : \mathcal{O} \to \mathcal{P}$ is an order-preserving function such that $\mathrm{lex}_\psi \leq f \leq \mathrm{uex}_\psi$, then $f$ is automatically an extension of $\psi$.

**Example 3.3.** Let $\mathcal{O}$ and $\mathcal{P}$ be ordered sets and let $\psi : \mathcal{O} \rightharpoonup \widehat{\mathcal{P}}$ be a partial function such that $\mathrm{dom}(\psi) = \{x\}$. Then $\psi$ is extremizable. Moreover, $\mathrm{lex}_\psi(y) = \psi(x)$ if $x \leq y$ and $\mathrm{lex}_\psi(y) = -\infty$ otherwise. Similarly, $\mathrm{uex}_\psi(y) = \psi(x)$ if $y \leq x$ and $\mathrm{uex}_\psi(y) = +\infty$ otherwise.

**Definition 3.4.** Let $\mathcal{O}$ and $\mathcal{P}$ be ordered sets. For each extremizable partial function $\psi : \mathcal{O} \rightharpoonup \mathcal{P}$, we define the *Darboux set of* $\psi$ to be

$$\mathrm{Dar}(\psi) = \{x \in \mathcal{O} \mid \mathrm{lex}_\psi(x) = \mathrm{uex}_\psi(x)\}.$$

Moreover, we denote by $\mathrm{ex}_\psi : \mathcal{O} \rightharpoonup \mathcal{P}$ the *Darboux extension of* $\psi$, i.e., the restriction of $\mathrm{uex}_\psi$ (or equivalently of $\mathrm{lex}_\psi$) to $\mathrm{Dar}(\psi)$.

**Definition 3.5.** Let $\mathcal{O}$, $\mathcal{P}$ be ordered sets and let $\psi$ be a partial function from $\mathcal{O}$ to $\mathcal{P}$. We say that $x \in \mathcal{O}$ is *$\psi$-bounded* if $y \leq x \leq z$ for some $y, z \in \mathrm{dom}(\psi)$. We denote the set of $\psi$-bounded elements of $\mathcal{O}$ by $\mathrm{B}(\psi)$. We say that $\psi$ is *encompassing* if every element of $\mathcal{O}$ is $\psi$-bounded. Moreover, for each extremizable $\psi : \mathcal{O} \rightharpoonup \mathcal{P}$ we define the *bounded Darboux set of* $\psi$ to be the subset $\mathrm{BDar}(\psi)$ of all $\psi$-bounded elements of $\mathrm{Dar}(\psi)$.

**Remark 3.6.** Let $\mathcal{O}$, $\mathcal{P}$ be ordered sets and let $\psi$ be the common extension of a compatible set $\Psi$ of partial functions from $\mathcal{O}$ to $\mathcal{P}$. If any $\psi' \in \Psi$ is encompassing, then $\mathrm{dom}(\psi') \subseteq \mathrm{dom}(\psi)$ implies that $\psi$ is also encompassing.

**Remark 3.7.** Let $\mathcal{O}$ and $\mathcal{P}$ be ordered sets and let $\Psi$ be a compatible set of extremizable partial functions from $\mathcal{O}$ to $\mathcal{P}$. If the common extension $\psi$ of $\Psi$ is also extremizable, then Remark 2.11 implies that $f \in [\mathrm{lex}_\psi, \mathrm{uex}_\psi]$ if and only if $f \in [\mathrm{lex}_{\psi'}, \mathrm{uex}_{\psi'}]$ for each $\psi' \in \Psi$. In particular,

$$\{\mathrm{uex}_\psi\} = \bigcap_{\psi' \in \Psi} [\mathrm{uex}_\psi, \mathrm{uex}_{\psi'}] \quad \text{and} \quad \{\mathrm{lex}_\psi\} = \bigcap_{\psi' \in \Psi} [\mathrm{lex}_{\psi'}, \mathrm{lex}_\psi].$$

**Remark 3.8.** Let $\mathcal{O}$ and $\mathcal{P}$ be ordered sets and let $\psi$ be an extremizable partial function from $\mathcal{O}$ to $\mathcal{P}$. If $f : \mathcal{O} \to \mathcal{P}$ is an extension of $\mathrm{ex}_\psi$ to $\mathcal{O}$, then its restriction to $\mathrm{dom}(\psi)$ coincides with $\psi$ and thus $\mathrm{lex}_\psi \leq f \leq \mathrm{uex}_\psi$. Since by construction $\mathrm{lex}_\psi$ and $\mathrm{uex}_\psi$ restrict to $\mathrm{ex}_\psi$ on $\mathrm{Dar}(\psi)$, it follows that the set of extensions of $\mathrm{ex}_\psi$ to $\mathcal{O}$ coincides with the set of extensions of $\psi$ to $\mathcal{O}$. In particular, $\mathrm{Dar}(\mathrm{ex}_\psi) = \mathrm{Dar}(\psi)$ and $\mathrm{ex}_{\mathrm{ex}_\psi} = \mathrm{ex}_\psi$.

**Definition 3.9.** Let $\mathcal{O}_1$, $\mathcal{O}_2$ and $\mathcal{O}_3$ be ordered sets. The partial functions $\psi_1 : \mathcal{O}_1 \rightharpoonup \mathcal{O}_2$ and $\psi_2 : \mathcal{O}_2 \rightharpoonup \mathcal{O}_3$ are *composable* if $\mathrm{dom}(\psi_2) \cap \mathrm{im}(\psi_1)$ is nonempty. If this is the case, their *composition* is the partial function $\psi_2 \circ \psi_1 : \mathcal{O}_1 \rightharpoonup \mathcal{O}_3$ such that $(\psi_2 \circ \psi_1)(x) = \psi_2(\psi_1(x))$ for each $x$ in

$$\mathrm{dom}(\psi_2 \circ \psi_1) = \{x \in \mathrm{dom}(\psi_1) \mid \psi_1(x) \in \mathrm{dom}(\psi_2)\}.$$

**Proposition 3.10.** *Let $\mathcal{O}_1$, $\mathcal{O}_2$ and $\mathcal{O}_3$ be ordered sets. Let $\psi_1 : \mathcal{O}_1 \rightharpoonup \mathcal{O}_2$ and $\psi_2 : \mathcal{O}_2 \rightharpoonup \mathcal{O}_3$ be partial functions such that*

(i) $\mathrm{dom}(\psi_2) \subseteq \mathrm{im}(\psi_1)$;

(ii) *$\psi_1$, $\psi_2$ and $\psi_2 \circ \psi_1$ are extremizable.*

*Then*

(1) $\mathrm{lex}_{\psi_2 \circ \psi_1} \leq \mathrm{lex}_{\psi_2} \circ \mathrm{lex}_{\psi_1} \leq \mathrm{uex}_{\psi_2} \circ \mathrm{uex}_{\psi_1} \leq \mathrm{uex}_{\psi_2 \circ \psi_1}$;

(2) $\mathrm{ex}_{\psi_1}\big(\mathrm{Dar}(\psi_2 \circ \psi_1) \cap \mathrm{Dar}(\psi_1)\big) \subseteq \mathrm{Dar}(\psi_2)$;

(3) $(\mathrm{ex}_{\psi_2} \circ \mathrm{ex}_{\psi_1})(x) = \mathrm{ex}_{\psi_2 \circ \psi_1}(x)$ *for all* $x \in \mathrm{Dar}(\psi_1) \cap \mathrm{Dar}(\psi_2 \circ \psi_1)$.

*Proof.* Item (1) is a consequence of the fact that $\mathrm{uex}_{\psi_2} \circ \mathrm{uex}_{\psi_1}$ and $\mathrm{lex}_{\psi_2} \circ \mathrm{lex}_{\psi_1}$ are extensions of $\psi_2 \circ \psi_1$ to $\mathcal{O}_1$. If $x \in \mathrm{Dar}(\psi_2 \circ \psi_1) \cap \mathrm{Dar}(\psi_1)$, then (1) implies

$$\mathrm{ex}_{\psi_2 \circ \psi_1}(x) = \mathrm{lex}_{\psi_2}(\mathrm{ex}_{\psi_1}(x)) = \mathrm{uex}_{\psi_2}(\mathrm{ex}_{\psi_1}(x)),$$

which proves (2) and (3). □

**Remark 3.11.** Since $\mathrm{dom}(\psi_2) = \mathcal{O}_2$ implies $\mathrm{lex}_{\psi_2} = \psi_2 = \mathrm{uex}_{\psi_2}$, we know that $\mathrm{Dar}(\psi_2 \circ \psi_1) \subseteq \mathrm{Dar}(\psi_1)$ whenever the partial function $\psi_2$ in the statement of Proposition 3.10 is an embedding and thus $\mathrm{ex}_{\psi_1}(\mathrm{Dar}(\psi_2 \circ \psi_1)) \subseteq \mathrm{Dar}(\psi_2)$.

**Lemma 3.12.** *Let $\mathcal{O}$, $\mathcal{P}$, $\mathcal{P}'$ be ordered sets and let $\psi : \mathcal{O} \rightharpoonup \mathrm{OP}(\mathcal{P}, \mathcal{P}')$ be an extremizable partial function. For each $p \in \mathcal{P}$, let $\mathrm{ev}_p : \mathrm{OP}(\mathcal{P}, \mathcal{P}') \to \mathcal{P}'$ be the order-preserving function that to each $f : \mathcal{P} \to \mathcal{P}'$ assigns its evaluation $\mathrm{ev}_p(f) = f(p)$ at $p$. If $\mathrm{ev}_p \circ \psi : \mathcal{O} \rightharpoonup \mathcal{P}'$ is extremizable for every $p \in \mathcal{P}$, then*

$$\mathrm{ev}_p \circ \mathrm{uex}_\psi = \mathrm{uex}_{\mathrm{ev}_p \circ \psi} \quad \textit{and} \quad \mathrm{ev}_p \circ \mathrm{lex}_\psi = \mathrm{lex}_{\mathrm{ev}_p \circ \psi}\,.$$

*Proof.* Using Proposition 3.10, $\mathrm{ev}_p \circ \mathrm{uex}_\psi = \mathrm{uex}_{\mathrm{ev}_p} \circ \mathrm{uex}_\psi \leq \mathrm{uex}_{\mathrm{ev}_p \circ \psi}$. Consider the order-preserving function $g : \mathcal{O} \to \mathrm{OP}(\mathcal{P}, \mathcal{P}')$ such that $(g(x))(p) = \mathrm{uex}_{\mathrm{ev}_p \circ \psi}$ for every $x \in \mathcal{O}$ and for every $p \in \mathcal{P}$. Then $(g(\eta))(p) = \mathrm{uex}_{\mathrm{ev}_p \circ \psi}(\eta) = (\psi(\eta))(p)$ for every $\eta \in \mathrm{dom}(\psi)$. Therefore, $g \leq \mathrm{uex}_\psi$ and thus $\mathrm{uex}_{\mathrm{ev}_p \circ \psi} = \mathrm{ev}_p \circ g \leq \mathrm{ev}_p \circ \mathrm{uex}_\psi$. Hence, $\mathrm{ev}_p \circ \mathrm{uex}_\psi = \mathrm{uex}_{\mathrm{ev}_p \circ \psi}$. The second equality is proved in a similar way. □

**Lemma 3.13.** *Let $\mathcal{O}$, $\mathcal{P}_1$, $\mathcal{P}_2$ be ordered sets, let $\psi : \mathcal{O} \rightharpoonup \mathcal{P}_1 \times \mathcal{P}_2$ be a partial function and for $i = 1, 2$ let $\pi_i : \mathcal{P}_1 \times \mathcal{P}_2 \to \mathcal{P}_i$ be the (order-preserving) projection onto the respective factor. Then $\psi$ is extremizable if and only if $\pi_i \circ \psi : \mathcal{O} \rightharpoonup \mathcal{P}_i$ is extremizable for each $i = 1, 2$. If this is the case, then $\pi_i \circ \mathrm{uex}_\psi = \mathrm{uex}_{\pi_i \circ \psi}$ and $\pi_i \circ \mathrm{lex}_\psi = \mathrm{lex}_{\pi_i \circ \psi}$ for each $i = 1, 2$.*

*Proof.* Assume that $\psi$ is extremizable. Then $\pi_i \circ \mathrm{lex}_\psi$ and $\pi_i \circ \mathrm{uex}_\psi$ are extensions of the partial function $\pi_i \circ \psi : \mathcal{O} \rightharpoonup \mathcal{P}_i$ (with domain $\mathrm{dom}(\psi)$) for each $i = 1, 2$. Furthermore, if $f_1 : \mathcal{O} \to \mathcal{P}_1$ and $f_2 : \mathcal{O} \to \mathcal{P}_2$ are, respectively, extensions of $\pi_1 \circ \psi$ and $\pi_2 \circ \psi$, then $(f_1, f_2) : \mathcal{O} \to \mathcal{P}_1 \times \mathcal{P}_2$ is an extension of $\psi$. By assumption, this implies $\mathrm{lex}_\psi \leq (f_1, f_2) \leq \mathrm{uex}_\psi$ and thus

$$\pi_i \circ \mathrm{lex}_\psi \leq f_i \leq \pi_i \circ \mathrm{uex}_\psi$$

for each $i = 1, 2$. Hence $\pi \circ \psi_i$ is extremizable, $\pi_i \circ \mathrm{uex}_\psi = \mathrm{uex}_{\pi_i \circ \psi}$ and $\pi_i \circ \mathrm{lex}_\psi = \mathrm{lex}_{\pi_i \circ \psi}$ for each $i = 1, 2$. Conversely, assume that $\pi_1 \circ \psi$ and $\pi_2 \circ \psi$ are extremizable. Then $(\mathrm{lex}_{\pi_1 \circ \psi}, \mathrm{lex}_{\pi_2 \circ \psi})$ and $(\mathrm{uex}_{\pi_1 \circ \psi}, \mathrm{uex}_{\pi_2 \circ \psi})$ are both extensions of $\psi = (\pi_1 \circ \psi, \pi_2 \circ \psi)$. Moreover, if $f : \mathcal{O} \to \mathcal{P}_1 \times \mathcal{P}_2$ is any extension of $\psi$, then $\pi_i \circ f$ is an extension of $\pi_i \circ \psi$ for each $i = 1, 2$. Since $f = (\pi_1 \circ f, \pi_2 \circ f)$, this implies

$$(\mathrm{lex}_{\pi_1 \circ \psi}, \mathrm{lex}_{\pi_2 \circ \psi}) \leq f \leq (\mathrm{uex}_{\pi_1 \circ \psi}, \mathrm{uex}_{\pi_2 \circ \psi})$$

and thus $\psi$ is extremizable with lower extension equal to $(\mathrm{lex}_{\pi_1\circ\psi}, \mathrm{lex}_{\pi_2\circ\psi})$ and upper extension equal to $(\mathrm{uex}_{\pi_1\circ\psi}, \mathrm{uex}_{\pi_2\circ\psi})$. □

**Remark 3.14.** Let $\mathcal{O}$ be a nonempty ordered set and let $\varnothing : \mathcal{O} \rightharpoonup \mathcal{O}$ be *the empty partial function of* $\mathcal{O}$, i.e., the unique partial function from $\mathcal{O}$ to itself whose domain is the empty set. Since the set of extensions of $\varnothing$ to $\mathcal{O}$ coincides with $\mathrm{OP}(\mathcal{O})$, if $\varnothing$ is extremizable, then in particular $\mathrm{lex}_\varnothing \le x \le \mathrm{uex}_\varnothing$ for every constant function $x : \mathcal{O} \to \mathcal{O}$. In other words, the lower and upper Darboux extensions of the empty partial function are constant and (with a slight abuse of notation) $\mathcal{O} = [\mathrm{lex}_\varnothing(\mathcal{O}), \mathrm{uex}_\varnothing(\mathcal{O})]$.

## 4. Darboux-complete ordered sets

**Definition 4.1.** An ordered set $\mathcal{P}$ is *Darboux complete* if every partial function from $\widehat{\mathcal{P}}$ to itself is extremizable.

**Example 4.2.** Since by Example 3.3 each partial function $f : \widehat{\varnothing} \rightharpoonup \widehat{\varnothing}$ is extremizable, the empty ordered set $\varnothing$ is Darboux complete.

**Lemma 4.3.** *Let $\mathcal{S}$ be a nonempty subset of a Darboux-complete ordered set $\mathcal{P}$. If $\mathrm{id}_\mathcal{S} : \widehat{\mathcal{P}} \rightharpoonup \widehat{\mathcal{P}}$ is the identity function on $\mathcal{S}$ and $J = [\mathrm{uex}_{\mathrm{id}_\mathcal{S}}(-\infty), \mathrm{lex}_{\mathrm{id}_\mathcal{S}}(+\infty)]$, then*

(1) $\mathcal{S} \subseteq J$;

(2) *$J$ is the intersection of all intervals of $\widehat{\mathcal{P}}$ that contain $\mathcal{S}$.*

*Proof.* Since $\mathcal{P}$ is Darboux complete, $\mathrm{lex}_{\mathrm{id}_\mathcal{S}}$ and $\mathrm{uex}_{\mathrm{id}_\mathcal{S}}$ exist. For every $s \in \mathcal{S}$

$$\mathrm{uex}_{\mathrm{id}_\mathcal{S}}(-\infty) \le \mathrm{uex}_{\mathrm{id}_\mathcal{S}}(s) = \mathrm{lex}_{\mathrm{id}_\mathcal{S}}(s) \le \mathrm{lex}_{\mathrm{id}_\mathcal{S}}(+\infty),$$

which implies (1). If $x, y \in \widehat{\mathcal{P}}$ are such that $\mathcal{S} \subseteq [x, y]$, let $\psi : \widehat{\mathcal{P}} \rightharpoonup \widehat{\mathcal{P}}$ be the partial function with domain $\widehat{\mathcal{S}}$ whose restriction to $\mathcal{S}$ is the identity and such that $\psi(-\infty) = x$ and $\psi(+\infty) = y$. Then

$$x = \mathrm{uex}_\psi(-\infty) \le \mathrm{uex}_{\mathrm{id}_\mathcal{S}}(-\infty) \le \mathrm{lex}_{\mathrm{id}_\mathcal{S}}(+\infty) \le \mathrm{lex}_\psi(+\infty) = y. \qquad \square$$

**Proposition 4.4.** *Let $\mathcal{P}$ be an ordered set. The following are equivalent*:

(1) *$\mathcal{P}$ is Darboux complete.*

(2) *Every partial function with codomain $\widehat{\mathcal{P}}$ is extremizable.*

(3) *For every ordered set $\mathcal{O}$, every partial function with codomain $\mathrm{OP}(\mathcal{O}, \widehat{\mathcal{P}})$ is extremizable.*

*Proof.* Assume that $\mathcal{P}$ is Darboux complete. Let $\psi$ be a partial function from an ordered set $\mathcal{O}$ to $\widehat{\mathcal{P}}$ and let $x \in \mathcal{O}$. Consider the subsets

$$\mathcal{S}_x = \{\psi(y) \mid y \le x \text{ and } y \in \mathrm{dom}(\psi)\} \subseteq \widehat{\mathcal{P}}, \tag{1}$$

$$\mathcal{S}^x = \{\psi(y) \mid x \le y \text{ and } y \in \mathrm{dom}(\psi)\} \subseteq \widehat{\mathcal{P}} \tag{2}$$

together with their identity functions $\mathrm{id}_{\mathcal{S}_x}, \mathrm{id}_{\mathcal{S}^x} : \widehat{\mathcal{P}} \rightharpoonup \widehat{\mathcal{P}}$. Define $l, u : \mathcal{O} \to \widehat{\mathcal{P}}$ such that

$$l(x) = \mathrm{lex}_{\mathrm{id}_{\mathcal{S}_x}}(+\infty) \quad \text{and} \quad u(x) = \mathrm{uex}_{\mathrm{id}_{\mathcal{S}^x}}(-\infty)$$

for all $x \in \mathcal{O}$. To see that $l$ and $u$ are indeed order-preserving, assume that $x, y \in \mathcal{O}$ are such that $x \leq y$. Since $\mathcal{S}_x \subseteq \mathcal{S}_y$, $\mathrm{lex}_{\mathrm{id}_{\mathcal{S}_y}}$ is an extension of $\mathrm{id}_{\mathcal{S}_x}$ and thus $l$ is order-preserving. Similarly, $u$ is order-preserving because $\mathcal{S}^y \subseteq \mathcal{S}^x$ implies that the restriction of $\mathrm{lex}_{\mathrm{id}_{\mathcal{S}_x}}$ to $\mathcal{S}^y$ coincides with $\mathrm{id}_{\mathcal{S}^y}$. Moreover $l$ is an extension of $\psi$ to $\mathcal{O}$ since for every $x \in \mathrm{dom}(\psi)$, $\mathcal{S}_x \subseteq [-\infty, \psi(x)]$ and Lemma 4.3 implies

$$\psi(x) = \mathrm{lex}_{\mathrm{id}_{\mathcal{S}_x}}(\psi(x)) \leq l(x) \leq \psi(x).$$

On the other hand, $\psi(y) = f(y) \leq f(x)$ for any extension $f$ of $\psi$ to $\mathcal{O}$ and for any $\psi(y) \in \mathcal{S}_x$. Therefore, $\mathcal{S}_x \subseteq [-\infty, f(x)]$ and thus (using again Lemma 4.3), $l(x) \leq f(x)$. Together with a similar argument involving $u$, this proves (2). Assume that (2) holds and let $\mathcal{O}$, $\mathcal{O}'$ be arbitrary ordered sets. Consider the canonical embedding $\alpha$ that to each partial function $\psi : \mathcal{O}' \rightharpoonup \mathrm{OP}(\mathcal{O}, \widehat{\mathcal{P}})$ assigns the partial function $\alpha(\psi) : \mathcal{O}' \times \mathcal{O} \rightharpoonup \widehat{\mathcal{P}}$ such that $(\alpha(\psi))(x', x) = (\psi(x'))(x)$ for all $(x', x) \in \mathrm{dom}(\alpha(\psi)) = \mathrm{dom}(\psi) \times \mathcal{O}$. The Darboux completeness of $\mathcal{P}$ ensures that $\alpha(\psi)$ is extremizable and thus $\mathrm{lex}_{\alpha(\psi)} \leq \alpha(f) \leq \mathrm{uex}_{\alpha(\psi)}$ for each extension $f$ of $\psi$ to $\mathcal{O}'$. Since the restriction of $\alpha$ to the subset of order-preserving functions $\mathcal{O}' \to \mathrm{OP}(\mathcal{O}, \widehat{\mathcal{P}})$ is an isomorphism, $\mathrm{lex}_{\psi} = \alpha^{-1}(\mathrm{lex}_{\alpha(\psi)})$ and $\mathrm{uex}_{\psi} = \alpha^{-1}(\mathrm{uex}_{\alpha(\psi)})$, which proves (3). Example 2.9 shows that (1) is a particular case of (3), which concludes the proof. □

**Remark 4.5.** Let $\mathcal{P}$ be an ordered set. Assume $\mathcal{P}$ is a Darboux-complete ordered set, and $\mathcal{S} \subseteq \mathcal{P}$ is nonempty and bounded, i.e., $\mathcal{S} \subseteq [x, y]$ for some $x, y \in \mathcal{P}$. Then Lemma 4.3 implies that $\mathrm{lex}_{\mathrm{id}_{\mathcal{S}}}(+\infty)$ and $\mathrm{uex}_{\mathrm{id}_{\mathcal{S}}}(-\infty)$ are respectively the least upper bound $\sup(\mathcal{S})$ and the greatest lower bound $\inf(\mathcal{S})$ of $\mathcal{S}$. Therefore, $\mathcal{P}$ is Dedekind complete. Conversely, suppose that the least upper bound and the greatest lower bound of every nonempty bounded subset of $\mathcal{P}$ exist. Given any partial function $\psi : \widehat{\mathcal{P}} \to \widehat{\mathcal{P}}$, let $\mathcal{S}_x$ and $\mathcal{S}^x$ be defined as in (1) and (2) respectively. Then the same argument as in the proof of Proposition 4.4 shows that $\psi$ is extremizable with $\mathrm{lex}_{\psi}(x) = \sup(\mathcal{S}_x)$ and $\mathrm{uex}_{\psi}(x) = \inf(\mathcal{S}^x)$ for all $x \in \mathcal{O}$. Hence, $\mathcal{P}$ is Darboux complete if and only if $\mathcal{P}$ is Dedekind complete. While these two notions of completeness are equivalent, the point of view of this paper is that Darboux completeness allows for a more direct and conceptual route to the foundations of elementary analysis.

**Corollary 4.6.** *Let $\mathcal{O}$ be an ordered set, let $\mathcal{P}$ be a Darboux-complete ordered set and let $N$ be a positive integer. Every encompassing partial function from $\mathcal{O}$ to $\mathcal{P}^N$ is extremizable.*

*Proof.* By Lemma 3.13, it suffices to prove the $N = 1$ case. Let $\varphi : \mathcal{O} \rightharpoonup \mathcal{P}$ be encompassing and let $\iota : \mathcal{P} \to \widehat{\mathcal{P}}$. By assumption, for each $z \in \mathcal{O}$ there exist

$x, y \in \mathrm{dom}(\varphi)$ such that $x \leq z \leq y$ and thus

$$\varphi(x) = \iota(\varphi(x)) \leq \mathrm{lex}_{\iota\circ\varphi}(z) \leq \mathrm{uex}_{\iota\circ\varphi}(z) \leq \iota(\psi(y)) = \varphi(y).$$

Therefore, $\mathrm{lex}_{\iota\circ\varphi}$ and $\mathrm{uex}_{\iota\circ\varphi}$ have their image contained in $\mathcal{P}$ and thus are extensions of $\varphi$ to $\mathcal{O}$. Moreover, $\mathrm{lex}_{\iota\circ\varphi}(x) \leq f(x) \leq \mathrm{uex}_{\iota\circ\varphi}(x)$ for every extension $f$ of $\varphi$ to $\mathcal{O}$ and for every $x \in \mathcal{P}$. Hence $\varphi$ is extremizable and $\mathrm{lex}_{\varphi}(x) = \mathrm{lex}_{\iota\circ\varphi}(x)$, $\mathrm{uex}_{\varphi}(x) = \mathrm{uex}_{\iota\varphi}(x)$ for all $x \in \mathcal{O}$. □

**Example 4.7.** We define the *free cocompletion* of an ordered set $\mathcal{O}$ to be the ordered set $\mathcal{O}^\vee = \mathrm{OP}(\mathcal{O}^{\mathrm{op}}, \widehat{\varnothing})$. Let $\mathcal{P} = \mathcal{O}^\vee \setminus \{\pm\infty\}$, where $\pm\infty$ denotes the constant function such that $\mathrm{im}(\pm\infty) = \pm\infty$. Combining Example 4.2 with Proposition 4.4 shows that every partial function with codomain $\mathcal{O}^\vee = \widehat{\mathcal{P}}$ is extremizable. Using Proposition 4.4 again, we conclude that $\mathcal{P}$ is Darboux complete.

**Example 4.8.** Let $\mathcal{O}$ be an ordered set, let $\mathcal{P}$ be a Darboux-complete ordered set and let $\mathcal{S}$ be a nonempty subset of $\mathcal{O}$. Furthermore, let $\psi_{\mathcal{S}} : \mathrm{OP}(\mathcal{O}, \mathcal{P}) \rightharpoonup \widehat{\mathcal{P}}$ be the partial function with domain the subset of functions that are constant on $\mathcal{S}$ and such that $\psi_{\mathcal{S}}(f) = f(x)$ for every $f \in \mathrm{dom}(\psi_{\mathcal{S}})$ and every $x \in \mathcal{S}$. For each $x \in \mathcal{S}$, $\mathrm{ev}_x$ coincides with $\psi_{\mathcal{S}}$ on $\mathrm{dom}(\psi_{\mathcal{S}})$ and thus $\mathrm{ev}_x \in [\mathrm{lex}_{\psi_{\mathcal{S}}}, \mathrm{uex}_{\psi_{\mathcal{S}}}]$. In particular, if $f : \mathcal{O} \to \mathcal{P}$ is in the Darboux set of $\psi_{\mathcal{S}}$, then $\mathrm{ev}_x \circ f = \mathrm{ev}_y \circ f$ for every $x, y \in \mathcal{S}$, i.e., $f$ is constant on $\mathcal{S}$. Hence, $\mathrm{dom}(\psi_{\mathcal{S}}) = \mathrm{Dar}(\psi_{\mathcal{S}})$.

**Remark 4.9.** Using the notation of Example 4.8, assume furthermore that $\mathcal{O}$ is discrete. For every order-preserving function $f : \mathcal{O} \to \mathcal{P}$ and for every $y \in \mathcal{P}$, let $f_y \in \mathrm{dom}(\psi_{\mathcal{S}})$ be the function whose restriction to $\mathcal{O} \setminus \mathcal{S}$ coincides with $f$ and such that $f_y(x) = y$ for all $x \in \mathcal{S}$. In particular, if there exists $y, z \in \mathcal{P}$ such that $f(x) \in [y, z]$ for all $x \in \mathcal{S}$, then $f \in [f_y, f_z]$ and thus $[\mathrm{lex}_{\psi_{\mathcal{S}}}, \mathrm{uex}_{\psi_{\mathcal{S}}}] \subseteq [y, z]$. Moreover, Corollary 4.6 implies that the restriction $\varphi_{\mathcal{S}} : \mathrm{B}(\psi_{\mathcal{S}}) \rightharpoonup \mathcal{P}$ of $\psi_{\mathcal{S}}$ to $\mathrm{B}(\psi)$ is extremizable.

## 5. Completely integrally closed subgroups

**Proposition 5.1.** *Let $\mathcal{O}, \mathcal{O}'$ be ordered sets, let $\mathcal{P}$ be a Darboux-complete ordered set and consider the composition of ordered functions $\mu : \mathrm{OP}(\widehat{\mathcal{P}}) \times \mathrm{OP}(\widehat{\mathcal{P}}) \to \mathrm{OP}(\widehat{\mathcal{P}})$ defined by setting $\mu(\varphi, \varphi') = \varphi \circ \varphi'$ for all $\varphi, \varphi' \in \mathrm{OP}(\widehat{\mathcal{P}})$. If $\psi : \mathcal{O} \rightharpoonup \mathrm{OP}(\widehat{\mathcal{P}})$ and $\psi' : \mathcal{O}' \rightharpoonup \mathrm{OP}(\widehat{\mathcal{P}})$ are partial functions with images in* $\mathrm{Aut}(\widehat{\mathcal{P}})$, *then*

$$\mu \circ (\mathrm{uex}_{\psi} \times \mathrm{uex}_{\psi'}) = \mathrm{uex}_{\mu\circ(\psi\times\psi')} \quad \textit{and} \quad \mu \circ (\mathrm{lex}_{\psi} \times \mathrm{lex}_{\psi'}) = \mathrm{lex}_{\mu\circ(\psi\times\psi')}\,.$$

*Proof.* Since $\mu \circ (\mathrm{uex}_{\psi} \times \mathrm{uex}_{\psi'})$ is an extension of $\mu \circ (\psi \times \psi')$ to $\mathcal{O} \times \mathcal{O}'$, we have $\mu \circ (\mathrm{uex}_{\psi} \times \mathrm{uex}_{\psi'}) \leq \mathrm{uex}_{\mu\circ(\psi\times\psi')}$. On the other hand, if $\eta \in \mathrm{dom}(\psi)$ is fixed, then

$$(\psi(\eta))^{-1} \circ \mathrm{uex}_{\mu\circ(\psi\times\psi')}(\eta, \eta') = \psi'(\eta')$$

for every $\eta' \in \mathrm{dom}(\psi')$. Using the assumption that $\psi'$ is extremizable, it follows that

$$(\psi(\eta))^{-1} \circ \mathrm{uex}_{\mu\circ(\psi\times\psi')}(\eta, x') \leq \mathrm{uex}_{\psi'}(x')$$

and thus

$$\begin{aligned}\mathrm{uex}_{\mu\circ(\psi\times\psi')}(\eta, x') &\leq \psi(\eta) \circ \mathrm{uex}_{\psi'}(x') \\ &= (\mu \circ (\mathrm{uex}_{\psi} \times \mathrm{uex}_{\psi'}))(\eta, x') \leq \mathrm{uex}_{\mu\circ(\psi\times\psi')}(\eta, x')\end{aligned}$$

for all $(\eta, x') \in \mathrm{dom}(\psi) \times \mathcal{O}'$. Setting $q = (\mathrm{uex}_{\psi'}(x'))(p)$ yields

$$\mathrm{ev}_p \circ \mathrm{uex}_{\mu\circ(\psi\times\psi')}(\eta, x) = (\psi(\eta))(q) = \mathrm{uex}_{\mathrm{ev}_q\circ\psi}(\eta)$$

for every $p \in \widehat{\mathcal{P}}$ and for every $\eta \in \mathrm{dom}(\psi)$. Lemma 3.12 then implies

$$\begin{aligned}\mathrm{ev}_p \circ \mathrm{uex}_{\mu\circ(\psi\times\psi')}(x, x') &\leq \mathrm{uex}_{\mathrm{ev}_q\circ\psi}(x) = \mathrm{ev}_q \circ \mathrm{uex}_{\psi}(x) \\ &= \mathrm{ev}_p \circ \mu \circ (\mathrm{uex}_{\psi} \times \mathrm{uex}_{\psi'})(x, x')\end{aligned}$$

for all $p \in \widehat{\mathcal{P}}$ and for all $(x, x') \in \mathcal{O} \times \mathcal{O}'$. This proves the first half of the proposition, the second equality is proved in a similar way. □

**Remark 5.2.** Given any ordered set $\mathcal{O}$, the ordered set $\mathrm{OP}(\mathcal{O})$ of order-preserving functions $f : \mathcal{O} \to \mathcal{O}$ is a monoid with respect to composition.

**Definition 5.3.** Let $\mathcal{O}$ be an ordered set. A subgroup (that is, a submonoid closed under inverses) $\mathcal{A}$ of $\mathrm{OP}(\mathcal{O})$ is *completely integrally closed* if for every $a, a' \in \mathcal{A}$, $a^n \leq a'$ for all $n \in \mathbb{N}$ implies $a \leq \mathrm{id}_{\mathcal{O}}$.

**Remark 5.4.** Completely integrally closed subgroups are a particular instance of the more general notion of (abstract) completely integrally closed ordered groups, which plays a key role in the classical study [Fuchs 1963] of embeddings in Dedekind-complete ordered groups. The remainder of this section can be thought of as an alternate construction of these embeddings formulated in the equivalent language of Darboux-complete ordered sets. Our main application is the self-contained construction of the field structure on the ordered set of real numbers described in Section 6.

**Proposition 5.5.** *Let $\mathcal{P}$ be a Darboux-complete ordered set. If $\mathcal{A}$ is a completely integrally closed subgroup of* $\mathrm{OP}(\widehat{\mathcal{P}})$, *then* $\mathrm{BDar}(\mathrm{id}_{\mathcal{A}})$ *is a subgroup of* $\mathrm{OP}(\widehat{\mathcal{P}})$.

*Proof.* Since $\mathrm{lex}_{\mathrm{id}_{\mathcal{A}}} \circ \mu$ and $\mathrm{uex}_{\mathrm{id}_{\mathcal{A}}} \circ \mu$ are extensions of $\mu \circ (\mathrm{id}_{\mathcal{A}} \times \mathrm{id}_{\mathcal{A}})$ to $(\mathrm{OP}(\widehat{\mathcal{P}}))^2$, we obtain

$$\mathrm{lex}_{\mu\circ(\mathrm{id}_{\mathcal{A}}\times\mathrm{id}_{\mathcal{A}})} \leq \mathrm{lex}_{\mathrm{id}_{\mathcal{A}}} \circ \mu \leq \mathrm{uex}_{\mathrm{id}_{\mathcal{A}}} \circ \mu \leq \mathrm{uex}_{\mu\circ(\mathrm{id}_{\mathcal{A}}\times\mathrm{id}_{\mathcal{A}})}\,. \tag{3}$$

By Proposition 5.1, we conclude that these inequalities restrict to equalities on $(\mathrm{Dar}(\mathrm{id}_{\mathcal{A}}))^2$. Hence $\mathrm{Dar}(\mathrm{id}_{\mathcal{A}})$ is closed under composition. We conclude that $\mathrm{Dar}(\mathrm{id}_{\mathcal{A}})$, which contains the submonoid $\mathcal{A}$ of $\mathrm{OP}(\widehat{\mathcal{P}})$, is itself a submonoid of

$\mathrm{OP}(\widehat{\mathcal{P}})$. Given $\varphi_1, \varphi_2 \in \mathrm{BDar}(\mathrm{id}_{\mathcal{A}})$, by definition there exist $a_i, a_i' \in \mathcal{A}$ such that $a_i \leq \varphi_i \leq a_i'$ for $i = 1, 2$. Therefore $a_1 \circ a_2 \leq \varphi_1 \circ \varphi_2 \leq a_1' \circ a_2'$ and thus $\mathrm{BDar}(\psi)$ is also a submonoid. In order to construct inverses, consider the partial function $\psi : (\mathrm{OP}(\widehat{\mathcal{P}}))^{\mathrm{op}} \rightharpoonup \mathrm{OP}(\widehat{\mathcal{P}})$ with domain $\mathcal{A}$ and such that $\psi(a) = a^{-1}$ for every $a \in \mathcal{A}$. By Proposition 5.1,

$$\mathrm{lex}_{\psi}(\varphi) \circ \varphi = \mathrm{lex}_{\mu\circ(\psi\times\mathrm{id}_{\mathcal{A}})}(\varphi, \varphi) \leq \mathrm{uex}_{\mu\circ(\psi\times\mathrm{id}_{\mathcal{A}})}(\varphi, \varphi)$$

for all $\varphi \in \mathrm{Dar}(\mathrm{id}_{\mathcal{A}})$. Since $\mathrm{im}(\psi) = \mathcal{A}$,

$$\mathrm{id}_{\mathcal{A}} \circ \mu(\psi \times \mathrm{id}_{\mathcal{A}}) = \mu(\psi \times \mathrm{id}_{\mathcal{A}})$$

and thus, using Proposition 3.10,

$$\mathrm{lex}_{\psi}(\varphi) \circ \varphi \leq \mathrm{lex}_{\mathrm{id}_{\mathcal{A}}}(\mathrm{lex}_{\mu\circ(\psi\times\mathrm{id}_{\mathcal{A}})}(\varphi, \varphi)) \leq \mathrm{lex}_{\mathrm{id}_{\mathcal{A}}}(\mathrm{uex}_{\mu\circ(\psi\times\mathrm{id}_{\mathcal{A}})}(\varphi, \varphi)). \tag{4}$$

Let $\varphi \in \mathrm{BDar}(\mathrm{id}_{\mathcal{A}})$, $a \in \mathcal{A}$ such that $a \leq \varphi$, and $a' \in \mathcal{A}$ such that $a' \leq \mathrm{uex}_{\mu\circ(\psi\times\mathrm{id}_{\mathcal{A}})}(\varphi, \varphi)$. Then we have

$$a \circ a' \leq a \circ \mathrm{uex}_{\mu\circ(\psi\times\mathrm{id}_{\mathcal{A}})}(\varphi, \varphi) \leq a \circ \mathrm{uex}_{\mu\circ(\psi\times\mathrm{id}_{\mathcal{A}})}(a, \varphi) = \varphi.$$

Iterating the same argument with $a$ replaced by $a \circ (a')^{n-1}$ yields $a \circ (a')^n \leq \varphi$ for all $n \in \mathbb{N}$. Since $\mathcal{A}$ is completely integrally closed, this implies $a' \leq \mathrm{id}_{\widehat{\mathcal{P}}}$. Together with a similar argument involving $\mathrm{lex}_{\mu\circ(\psi\times\mathrm{id}_{\mathcal{A}})}$, we conclude that

$$\mathrm{lex}_{\mathrm{id}_{\mathcal{A}}}(\mathrm{uex}_{\mu\circ(\psi\times\mathrm{id}_{\mathcal{A}})}(\varphi, \varphi)) \leq \mathrm{id}_{\mathcal{P}} \leq \mathrm{uex}_{\mathrm{id}_{\mathcal{A}}}(\mathrm{lex}_{\mu\circ(\psi\times\mathrm{id}_{\mathcal{A}})}(\varphi, \varphi)).$$

Therefore, applying $\mathrm{uex}_{\mathrm{id}_{\mathcal{A}}}$ to both sides of (4) and using Proposition 3.10 yields

$$\mathrm{id}_{\widehat{\mathcal{P}}} \leq \mathrm{uex}_{\mathrm{id}_{\mathcal{A}}}(\mathrm{lex}_{\mu\circ(\psi\times\mathrm{id}_{\mathcal{A}})}(\varphi, \varphi)) \leq \mathrm{uex}_{\mathrm{id}_{\mathcal{A}}}(\mathrm{lex}_{\psi}(\varphi)) \circ \varphi \leq \mathrm{uex}_{\mathrm{id}_{\mathcal{A}}}(\mathrm{id}_{\widehat{\mathcal{P}}}) = \mathrm{id}_{\widehat{\mathcal{P}}},$$

where the last equality follows from the fact that $\mathrm{id}_{\widehat{\mathcal{P}}}$ is an element of $\mathcal{A}$. Hence $\varphi$ has a left inverse. A similar argument shows that it has right inverse and concludes the proof. □

**Corollary 5.6.** *Let $\mathcal{P}$ be a Darboux-complete ordered set and let $\mathcal{A} \subseteq \mathrm{OP}(\widehat{\mathcal{P}})$ be a commutative completely integrally closed subgroup. Then* $\mathrm{BDar}(\mathrm{id}_{\mathcal{A}})$ *is a commutative group.*

*Proof.* Let $\mu' : \mathrm{OP}(\widehat{\mathcal{P}}) \times \mathrm{OP}(\widehat{\mathcal{P}}) \to \mathrm{OP}(\widehat{\mathcal{P}})$ denote composition in reverse order; i.e., $\mu'(\varphi, \varphi') = \varphi' \circ \varphi$ for all $\varphi, \varphi' \in \mathrm{OP}(\widehat{\mathcal{P}})$. Since the restrictions of $\mathrm{lex}_{\mathrm{id}_{\mathcal{A}}} \circ \mu'$ and $\mathrm{uex}_{\mathrm{id}_{\mathcal{A}}} \circ \mu'$ to $\mathcal{A} \times \mathcal{A}$ coincide with $\mu \circ (\mathrm{id}_{\mathcal{A}} \times \mathrm{id}_{\mathcal{A}})$, we obtain

$$\mathrm{lex}_{\mu\circ(\mathrm{id}_{\mathcal{A}}\times\mathrm{id}_{\mathcal{A}})} \leq \mathrm{lex}_{\mathrm{id}_{\mathcal{A}}} \circ \mu' \leq \mathrm{uex}_{\mathrm{id}_{\mathcal{A}}} \circ \mu' \leq \mathrm{uex}_{\mu\circ(\mathrm{id}_{\mathcal{A}}\times\mathrm{id}_{\mathcal{A}})}.$$

Together with (3), this implies the commutativity of the monoid $\mathrm{Dar}(\mathrm{id}_{\mathcal{A}})$, which contains $\mathrm{BDar}(\mathrm{id}_{\mathcal{A}})$. □

## 6. The Darboux completion

**Remark 6.1.** Let $\mathcal{O}$ be an ordered set. For each $x \in \mathcal{O}$, let $\delta_x : \mathcal{O}^{\mathrm{op}} \rightharpoonup \widehat{\varnothing}$ be the partial function such that $\mathrm{dom}(\delta_x) = \{x\}$ and $\delta_x(x) = +\infty$. Then $\mathrm{lex}_{\delta_x}(y) = +\infty$ if and only if $y \leq x$. Let us define $Y(x) = \mathrm{lex}_{\delta_x}$ for every $x \in \mathcal{O}$. If $f \in \mathcal{O}^\vee$ then $f(x) = +\infty$ if and only if $Y(x) \leq f$. Moreover, $f \leq g$ in $\mathcal{O}^\vee$ if and only if $Y(x) \leq f$ implies $Y(x) \leq g$. In particular, $Y(x) \leq Y(y)$ if and only if $x \leq y$. Hence the assignment $x \mapsto Y(x)$ defines an order-preserving embedding $Y : \mathcal{O} \to \mathcal{O}^\vee$ called the *Yoneda embedding of* $\mathcal{O}$.

**Proposition 6.2.** *Let $\mathcal{O}$ be an ordered set, let $\varphi : \mathcal{O}^\vee \rightharpoonup \mathcal{O}^\vee$ be the identity function of the image of the Yoneda embedding of $\mathcal{O}$ and let* $\mathrm{Dar}(\mathcal{O})$ *denote the Darboux set of $\varphi$. Then*

(1) $\mathrm{lex}_\varphi = \mathrm{id}_{\mathcal{O}^\vee}$;

(2) *if* $g \in \mathrm{OP}(\mathrm{Dar}(\mathcal{O}))$ *restricts to the identity on* $Y(\mathcal{O})$, *then* $g = \mathrm{id}_{\mathrm{Dar}(\mathcal{O})}$;

(3) $\mathrm{uex}_\varphi(\mathcal{O}^\vee) \subseteq \mathrm{Dar}(\mathcal{O})$;

(4) *the empty partial function of* $\mathrm{Dar}(\mathcal{O})$ *is extremizable.*

*Proof.* Since $\mathrm{id}_{\mathcal{O}^\vee}$ restricts to $\varphi$ on $Y(\mathcal{O})$, we know $\mathrm{lex}_\varphi(f) \leq f$ for every $f \in \mathcal{O}^\vee$. On the other hand, $Y(x) \leq f$ implies $Y(x) = \mathrm{lex}_\varphi(Y(x)) \leq \mathrm{lex}_\varphi(f)$. Using Remark 6.1, this proves (1). Item (2) follows immediately from (1) and the definition of $\mathrm{Dar}(\mathcal{O})$. Proposition 3.10 and (1) yield

$$\mathrm{uex}_\varphi = \mathrm{uex}_\varphi \circ \mathrm{lex}_\varphi \leq \mathrm{uex}_\varphi \circ \mathrm{uex}_\varphi \leq \mathrm{uex}_{\varphi\circ\varphi} = \mathrm{uex}_\varphi,$$

which readily implies (3). Since $+\infty \leq \mathrm{uex}_\varphi(+\infty) \leq +\infty$, we have $+\infty \in \mathrm{Dar}(\mathcal{O})$. If $-\infty \neq \mathrm{Dar}(\mathcal{O})$, then by Remark 6.1 there exists $x \in \mathcal{O}$ such that $Y(x) \leq \mathrm{uex}_\varphi(-\infty)$. By Lemma 4.3 this implies $x \leq y$ for all $y \in \mathcal{O}$. Therefore, the empty partial function of $\mathrm{Dar}(\mathcal{O})$ is extremizable, $\mathrm{uex}_\varnothing(\mathcal{O}) = +\infty$ and $\mathrm{lex}_\varnothing(\mathcal{O})$ is the function that takes the value $-\infty$ on the complement of a set of cardinality at most 1. □

**Definition 6.3.** Using the notation of Proposition 6.2 and Remark 3.14, we define the *Darboux completion of an ordered set* $\mathcal{O}$ to be the ordered set

$$\mathrm{Dar}'(\mathcal{O}) = \mathrm{Dar}(\mathcal{O}) \setminus \{\mathrm{lex}_\varnothing(\mathrm{Dar}(\mathcal{O})), \mathrm{uex}_\varnothing(\mathrm{Dar}(\mathcal{O}))\}.$$

**Corollary 6.4.** *The Darboux completion of an ordered set is Darboux complete.*

*Proof.* Let $\mathcal{O}$ be an ordered set and let $\iota : \mathrm{Dar}(\mathcal{O}) \to \mathcal{O}^\vee$ be the inclusion. For any partial function $\psi : \mathrm{Dar}(\mathcal{O}) \rightharpoonup \mathrm{Dar}(\mathcal{O})$, Example 4.7 ensures that $\iota\circ\psi$ is extremizable. By Proposition 6.2, $\mathrm{uex}_\varphi \circ \mathrm{uex}_{\iota\circ\psi}$ and $\mathrm{uex}_\varphi \circ \mathrm{lex}_{\iota\circ\psi}$ are order-preserving functions in $\mathrm{OP}(\mathrm{Dar}(\mathcal{O}))$ that restrict to $\psi$ on $\mathrm{dom}(\psi)$. On the other hand, $\mathrm{lex}_{\iota\circ\psi} \leq \iota \circ g \leq \mathrm{uex}_{\iota\circ\psi}$ for any extension $g$ of $\psi$ to $\mathrm{Dar}(\mathcal{O})$. Since $\mathrm{uex}_\varphi \circ \iota \circ g = g$, this implies

$\mathrm{uex}_\varphi \circ \mathrm{lex}_{\psi'} \leq g \leq \mathrm{uex}_\varphi \circ \mathrm{uex}_{\psi'}$ and thus $\psi$ is extremizable. This concludes the proof, since by construction $\mathrm{Dar}(\mathcal{O})$ is canonically isomorphic to $\widehat{\mathrm{Dar}'(\mathcal{O})}$. □

**Remark 6.5.** From now on we use the Yoneda embedding to canonically identify $\mathcal{O}$ with a subset of $\mathrm{Dar}(\mathcal{O}) \subseteq \mathcal{O}^\vee$. In particular, this provides a canonical embedding of $\mathrm{OP}(\mathcal{O})$ into the set of partial functions $\mathrm{Dar}(\mathcal{O}) \rightharpoonup \mathrm{Dar}(\mathcal{O})$.

**Example 6.6.** We define the set of *real numbers* to be the Darboux completion $\mathbb{R}$ of the ordered set $\mathbb{Q}$ of rational numbers. Moreover, $\mathrm{Dar}(\mathbb{Q})$ is canonically identified with the set of *extended real numbers* $\widehat{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$.

**Example 6.7.** Let $\mathbb{Q}_{>0} \subseteq \mathbb{Q}$ be the ordered set of positive rational numbers and let $\mathbb{R}_{>0} = \mathrm{Dar}'(\mathbb{Q}_{>0})$. Extending each function in $\mathrm{Dar}'(\mathbb{Q}_{>0})$ by $+\infty$ to $\mathbb{Q} \setminus \mathbb{Q}_{>0}$ yields a canonical embedding of $\mathbb{R}_{>0}$ into $\mathbb{R}$ whose image consists of real numbers that are greater than $Y(0)$. Moreover, the composition of this embedding with the canonical embedding of $\mathbb{Q}_{>0}$ into $\mathrm{Dar}'(\mathbb{Q}_{>0})$ coincides with the restriction of the canonical embedding of $\mathbb{Q}$ into $\mathbb{R}$. Keeping in mind the above canonical identifications, it makes sense to write equalities such as $\mathbb{Q}_{>0} = \mathbb{Q} \cap \mathbb{R}_{>0}$.

**Remark 6.8.** Let $\mathcal{O}$ be an ordered set, let $\mathcal{P}$ be a complete ordered set and let $\psi : \mathrm{Dar}(\mathcal{O}) \rightharpoonup \widehat{\mathcal{P}}$ be an embedding with domain $Y(\mathcal{O})$ and inverse $\psi' : \widehat{\mathcal{P}} \rightharpoonup \mathrm{Dar}(\mathcal{O})$. Since $\mathrm{uex}_{\psi'} \circ \mathrm{uex}_\psi$ restricts to the identity on $Y(\mathcal{O})$, it is equal to $\mathrm{id}_{\mathrm{Dar}(\mathcal{O})}$ by Proposition 6.2. Therefore, $\mathrm{uex}_\psi : \mathrm{Dar}(\mathcal{O}) \to \widehat{\mathcal{P}}$ is an embedding. In particular, it can attain the values $\pm\infty$ at most once, which implies that $\mathrm{uex}_\psi$ restricts to an embedding $f : \mathrm{Dar}'(\mathcal{O}) \to \mathcal{P}$. By Remark 4.5 this implies that $\mathrm{Dar}'(\mathcal{O})$ also satisfies the universal property of the Dedekind–MacNeille completion of $\mathcal{O}$ and is therefore canonically isomorphic to it. In particular, this shows that our definition of $\mathbb{R}$ is canonically isomorphic to the ordered set $\mathbb{R}'$ of Dedekind cuts of $\mathbb{Q}$. In fact, in this case it is easy to see directly that $\mathrm{uex}_{\psi'} : \widehat{\mathbb{R}}' \to \widehat{\mathbb{R}}$ is injective since it maps the cut associated to a rational number $x$ to $Y(x)$ and $(\mathrm{uex}_{\psi'}(C))^{-1}(+\infty) = C$ for any irrational cut $C$.

**Proposition 6.9.** *There exists a canonical embedding* $\alpha : \mathrm{Aut}(\mathcal{O}) \to \mathrm{Aut}(\mathrm{Dar}(\mathcal{O}))$. *Moreover,* $\alpha$ *is a group homomorphism.*

*Proof.* Let $\varphi \in \mathrm{Aut}(\mathcal{O})$. Using the convention of Remark 6.5, we may think of $\varphi$ as a partial function $\mathrm{Dar}(\mathcal{O}) \rightharpoonup \mathrm{Dar}(\mathcal{O})$. Then by Remark 6.8

$$\mathrm{lex}_{\varphi^{-1}} \circ \mathrm{uex}_\varphi = \mathrm{id}_{\mathrm{Dar}(\mathcal{O})} = \mathrm{uex}_\varphi \circ \mathrm{lex}_{\varphi^{-1}} .$$

This implies that $\mathrm{uex}_\varphi$ is invertible and $\mathrm{uex}_\varphi \leq \mathrm{lex}_\varphi \circ \mathrm{lex}_{\varphi^{-1}} \circ \mathrm{uex}_\varphi \leq \mathrm{lex}_\varphi$. Therefore, $\mathrm{ex}_\varphi \in \mathrm{Aut}(\mathrm{Dar}(\mathcal{O}))$. Let $\alpha(\varphi) = \mathrm{ex}_\varphi$ for all $\varphi \in \mathrm{Aut}(\mathcal{O})$. Combining Remark 6.8 and Proposition 3.10, we conclude that $\alpha$ is an injective group homomorphism and the proposition is proved. □

**Example 6.10.** Addition in $\mathbb{Q}$ defines an embedding $\lambda : \mathbb{Q} \to \mathrm{Aut}(\mathbb{Q})$ such that $(\lambda(r))(s) = r + s$ for all $r, s \in \mathbb{Q}$. Composing with $\alpha$ we obtain an embedding $\beta : \mathbb{Q} \to \mathrm{Aut}(\widehat{\mathbb{R}})$. Since every (order-preserving) automorphism of $\widehat{\mathbb{R}}$ necessarily fixes $\pm\infty$, we have a canonical identification of $\mathrm{Aut}(\widehat{\mathbb{R}})$ with $\mathrm{Aut}(\mathbb{R})$. In particular, $(\beta(x))(\pm\infty) = \pm\infty$ for all $x \in \mathbb{Q}$.

**Proposition 6.11.** $\mathbb{R}$ *is canonically isomorphic to* $\mathrm{BDar}(\mathrm{id}_{\beta(\mathbb{Q})})$.

*Proof.* Considering the embedding $\beta$ constructed in Example 6.10 as a partial function $\mathbb{R} \rightharpoonup \mathrm{OP}(\widehat{\mathbb{R}})$ (which is extremizable by Proposition 4.4), we obtain order-preserving functions $\mathrm{lex}_\beta, \mathrm{uex}_\beta : \mathbb{R} \to \mathrm{OP}(\widehat{\mathbb{R}})$. The order-preserving function $\mathrm{ev}_0 : \mathrm{BDar}(\mathrm{id}_{\beta(\mathbb{Q})}) \to \mathbb{R}$ is surjective by Remark 6.8 since $\mathrm{ev}_0 \circ \mathrm{lex}_\beta$ and $\mathrm{ev}_0 \circ \mathrm{uex}_\beta$ both restrict to the identity on $\mathbb{Q}$. Since $\mathrm{lex}_\beta \circ \mathrm{ev}_0$ and $\mathrm{uex}_\beta \circ \mathrm{ev}_0$ both restrict to the identity on $\beta(\mathbb{Q})$, they both equal the identity on $\mathrm{BDar}(\mathrm{id}_{\beta(\mathbb{Q})})$. Therefore $\mathrm{ev}_0$ is invertible with inverse $\mathrm{ex}_\beta$. □

**Remark 6.12.** Combining Proposition 6.11 with Corollary 5.6, we conclude that $\mathbb{R}$ has a canonical structure of commutative group. Alternatively, this structure can be understood as follows. Let $+$ be the addition operation on $\mathbb{Q}$, thought of as a partial function $\mathbb{R} \times \mathbb{R} \rightharpoonup \mathbb{R}$. Since $(\mathrm{ex}_\beta(r))(s) = r + s$ for all $r, s \in \mathbb{Q}$, we obtain

$$\mathrm{lex}_+(x, y) \leq (\mathrm{ex}_\beta(x))(y) \leq \mathrm{uex}_+(x, y) \tag{5}$$

for all $x, y \in \mathbb{R}$. On the other hand, for every $r \in \mathbb{Q}$ both $\mathrm{ex}_\beta(r)^{-1} \circ \mathrm{uex}_+(r, -)$ and $\mathrm{ex}_\beta(r)^{-1} \circ \mathrm{lex}_+(r, -)$ restrict to the identity of $\mathbb{Q}$. By Remark 6.8, this implies $\mathrm{lex}_+(r, -) = \mathrm{ex}_\beta(r) = \mathrm{uex}_+(r, -)$ for all $r \in \mathbb{Q}$ and thus $\mathrm{lex}_\beta(x) \leq \mathrm{lex}_+(x, -) \leq \mathrm{uex}_+(x, -) \leq \mathrm{uex}_\beta(x)$ for all $x \in \mathbb{R}$. Hence the inequalities of (5) are actually equalities for all $x, y \in \mathbb{R}$.

**Remark 6.13.** A similar argument shows that the multiplication on $\mathbb{Q}_{>0}$ thought of as a partial function $\bullet : \mathbb{R}_{>0} \times \mathbb{R}_{>0} \rightharpoonup \mathbb{R}_{>0}$ defines a partial function $\gamma : \mathbb{R}_{>0} \rightharpoonup \mathrm{Aut}(\mathbb{R}_{>0})$ such that $\mathrm{dom}(\gamma) = \mathbb{Q}_{>0}$ and $(\mathrm{ex}_\gamma(x))(y) = \mathrm{ex}_\bullet(x, y)$ for all $x, y \in \mathbb{R}_{>0}$.

**Theorem 6.14.** $(\mathbb{R}_{>0}, \mathrm{ex}_+, \mathrm{ex}_\bullet)$ *is a semifield.*

*Proof.* Let $\psi : (\mathbb{R}_{>0})^3 \rightharpoonup \mathbb{R}_{>0}$ be the partial function with domain $(\mathbb{Q}_{>0})^3$ and such that $\psi(r, s, t) = r(s + t)$ for all $r, s, t \in \mathbb{Q}_{>0}$. Since $\mathrm{ex}_\bullet(r, \mathrm{ex}_+(s, t)) = \psi(r, s, t)$ for all $r, s, t \in \mathbb{Q}_{>0}$,

$$\mathrm{lex}_\psi(x, y, z) \leq \mathrm{ex}_\bullet(x, \mathrm{ex}_+(y, z)) \leq \mathrm{uex}_\psi(x, y, z) \tag{6}$$

for all $x, y, z \in \mathbb{R}_{>0}$. On the other hand, since $\gamma(s+t)$ agrees with both $\mathrm{lex}_\psi(-, s, t)$ and $\mathrm{uex}_\psi(-, s, t)$ on $\mathbb{Q}_{>0}$ for all $s, t \in \mathbb{Q}_{>0}$, they also agree on $\mathbb{R}_{>0}$. Using that $(\gamma(s + t))(x) = \mathrm{ex}_\bullet(x, s + t) = (\mathrm{ex}_\gamma(x))(s + t)$, we obtain

$$\mathrm{lex}_+(y, z) \leq (\mathrm{ex}_\gamma(x))^{-1} \mathrm{lex}_\psi(x, y, z) \leq (\mathrm{ex}_\gamma(x))^{-1} \mathrm{uex}_\psi(x, y, z) \leq \mathrm{uex}_+(y, z).$$

Since $\mathrm{ex}_\bullet(x, \mathrm{ex}_+(y,z)) = (\mathrm{ex}_\gamma(x))(\mathrm{ex}_+(y,z))$ for all $x, y, z \in \mathbb{R}_{>0}$, we conclude that the inequalities of (6) are in fact equalities. It follows from the distributivity of $\bullet$ over $+$ on $\mathbb{Q}_{>0}$ that $\mathrm{ex}_+(\mathrm{ex}_\bullet(r,t), \mathrm{ex}_\bullet(r,t)) = \psi(r,s,t)$ for all $r, s, t \in \mathbb{Q}_{>0}$. Hence, $\mathrm{ex}_\bullet$ distributes over $\mathrm{ex}_+$ on $\mathbb{R}_{>0}$ and the theorem is proved. □

**Remark 6.15.** A standard argument shows that $\mathrm{ex}_\bullet$ can be canonically extended to an operation $\cdot$ on $\mathbb{R}$ (which is not order-preserving) in such a way that $(\mathbb{R}, \mathrm{ex}_+, \cdot)$ is a field. With a slight abuse of notation, from now on we write $+$ for $\mathrm{ex}_+$. Since $(\beta(x))(\pm\infty) = \pm\infty$ for all $x \in \mathbb{Q}$, we set $x + (\pm\infty) = \pm\infty$ for all $x \in \mathbb{R}$.

**Remark 6.16.** For each ordered set $\mathcal{O}$, the set $\mathrm{OP}(\mathcal{O}, \mathbb{R})$ inherits a canonical structure of $\mathbb{R}$-algebra with operations defined pointwise on $\mathcal{O}$. In particular, $f_1 \le f_2$ implies $f_1 + f_3 \le f_2 + f_3$ for any $f_1, f_2, f_3 \in \mathrm{OP}(\mathcal{O}, \mathbb{R})$ and $f_1 f_3 \le f_2 f_3$ whenever $0 \ge f_3$.

## 7. Limits and integrals

**Definition 7.1.** A *filter basis* on an ordered set $\mathcal{O}$ is a collection $F$ of nonempty subsets of $\mathcal{O}$ that is closed under finite intersections. To each filter basis $F$ of $\mathcal{O}$ we associate the partial function $\psi_F : \mathrm{OP}(\mathcal{O}, \mathbb{R}) \rightharpoonup \widehat{\mathbb{R}}$ such that

$$\mathrm{dom}(\psi_F) = \bigcup_{\mathcal{S} \in F} \mathrm{dom}(\psi_{\mathcal{S}}),$$

where $\psi_{\mathcal{S}}$ is defined as in Example 4.8 and $\psi_F(f) = \psi_{\mathcal{S}}(f)$ for each $f \in \mathrm{dom}(\psi_{\mathcal{S}})$ and for each $\mathcal{S} \in F$.

**Definition 7.2.** Let $\mathcal{O}$ be a discrete set and let $F$ be a filter basis on $\mathcal{O}$. An order-preserving function $f : \mathcal{O} \to \mathbb{R}$ is *$F$-convergent* if there exists $\lim_F(f) \in \widehat{\mathbb{R}}$ such that for every $\varepsilon > 0$ there exists $\mathcal{S} \in F$ such that $f(x) \in [\lim_F(f) - \varepsilon, \lim_F(f) + \varepsilon]$ for all $x \in \mathcal{S}$.

**Theorem 7.3.** *Let $\mathcal{O}$ be a discrete set and let $F$ be a filter basis on $\mathcal{O}$. An order-preserving function $f : \mathcal{O} \to \mathbb{R}$ is $F$-convergent if and only if $f \in \mathrm{Dar}(\psi_F)$. Moreover, $\mathrm{ex}_{\psi_F}(f) = \lim_F(f)$ for all $f \in \mathrm{Dar}(\psi_F)$.*

*Proof.* Assume that $f \in \mathrm{Dar}(\psi_F)$. Then by Remark 2.11 for every $\varepsilon > 0$ there exist $\mathcal{S}', \mathcal{S}'' \in F$ such that $[\mathrm{ex}_{\psi_F}(f), \mathrm{ex}_{\psi_{\mathcal{S}'}}(f)] \subseteq [\mathrm{ex}_{\psi_F}(f), \mathrm{ex}_{\psi_F}(f) + \varepsilon]$ and $[\mathrm{ex}_{\psi_{\mathcal{S}''}}(f), \mathrm{ex}_{\psi_F}(f)] \subseteq [\mathrm{ex}_{\psi_F}(f) - \varepsilon, \mathrm{ex}_{\psi_F}(f)]$. Therefore, setting $\mathcal{S} = \mathcal{S}' \cap \mathcal{S}''$ we obtain

$$\mathrm{ex}_{\psi_F}(f) - \varepsilon \le \mathrm{lex}_{\psi_{\mathcal{S}}}(f) \le f(x) \le \mathrm{uex}_{\psi_{\mathcal{S}}}(f) \le \mathrm{ex}_{\psi_F}(f) + \varepsilon$$

for every $x \in \mathcal{S}$. Hence $f$ is $F$-convergent and $\lim_F(f) = \mathrm{ex}_{\psi_F}(f)$. Conversely, if $f$ is $F$-convergent, for every $\varepsilon > 0$ there exists $\mathcal{S} \in F$ such that

$$\lim_F(f) - \varepsilon \le f(x) \le \lim_F(f) + \varepsilon$$

for all $x \in \mathcal{S}$. Using Remark 4.9, this implies

$$[\mathrm{lex}_{\psi_F}(f), \mathrm{uex}_{\psi_F}(f)] \subseteq [\mathrm{lex}_{\psi_S}(f), \mathrm{uex}_{\psi_S}(f)] \subseteq [\lim_F(f) - \varepsilon, \lim_F(f) + \varepsilon]$$

for every $\varepsilon > 0$. Hence, $f \in \mathrm{Dar}(\psi_F)$ and $\lim_F(f) = \mathrm{ex}_{\psi_F}(f)$. □

**Remark 7.4.** In terms of the philosophy outlined in Section 1, the functions in $\mathrm{dom}(\psi_F)$, i.e., the functions that are constant on some element of $F$, are "obviously $F$-convergent" and $\psi_F$ is their "obvious limit". Feeding the machinery of Darboux calculus with this information results in a construction of general $F$-convergent functions that is alternative to the $(\varepsilon, \delta)$-definition given in Definition 7.2.

**Example 7.5.** Let $\mathbb{N}$ be the set of natural numbers with its usual order. Let $\mathcal{O} = |\mathbb{N}|$ and let $F = \{\mathbb{N} \setminus [1, n]\}_{n \in \mathbb{N}}$. Then $\mathrm{OP}(\mathcal{O}, \mathbb{R})$ is the set of all sequences, $\mathrm{dom}(\psi_F)$ is the set of sequences that are eventually constant and $\psi_F(f)$ is the function that to such a sequence assigns its obvious limit, i.e., the only value that $f$ attains infinitely many times. Moreover, $\mathrm{Dar}(\psi_F)$ is precisely the set of all convergent sequences (including those converging to $\pm\infty$) and $\mathrm{ex}_{\psi_F}$ is their limit.

**Example 7.6.** Let $\mathcal{O} = |\mathbb{R}|$, let $x_0 \in \mathbb{R}$ and let $F$ be the collection of all subsets of the form $[x_0 - \delta, x_0 + \delta]$ for some $\delta > 0$. Then $\mathrm{OP}(\mathcal{O}, \mathbb{R})$ is the set of all real-valued functions of one real variable, $\mathrm{dom}(\psi_F)$ is the subset of functions that are constant in a neighborhood of $x_0$ and $\psi_F(f) = f(x_0)$ for all $f \in \mathrm{dom}(\psi_F)$. Moreover, $\mathrm{Dar}(\psi_F)$ is precisely the set of all functions that are continuous at $x_0$ and $\mathrm{ex}_{\psi_F}(f) = f(x_0)$ for all $f \in \mathrm{dom}(\psi_F)$.

**Example 7.7.** In the notation of Example 7.6, we could also consider $F$ to be the collection of all subsets of the form $[x_0 - \delta, x_0 + \delta] \setminus \{x_0\}$ for some $\delta > 0$. Then $f \in \mathrm{Dar}(\psi_F)$ if and only if $f$ has a limit at $x_0$, in which case $\mathrm{ex}_{\psi_F}(f)$ equals the limit. We leave the obvious variations leading to left and right limits to the reader.

**Definition 7.8.** We denote by $\mathrm{Int}(\mathcal{O})$ the ordered set of all intervals with endpoints in $\mathcal{O}$ with order given by $[a, b] \leq [c, d]$ if and only if $c \leq a \leq b \leq d$. We write $\mathrm{int}(\mathcal{O})$ for the collection of nonempty subsets of $\mathcal{O}$ of the form $[x, z] \setminus \{x, z\}$, for some $[x, z] \in \mathrm{Int}(\mathcal{O})$ (also ordered by inclusion).

**Example 7.9.** For any $J \in \mathrm{Int}(\mathbb{R})$ let $m : \mathrm{int}(J) \to \mathbb{R}_{>0}$ be the order-preserving function defined by $m([x, z] \setminus \{x, z\}) = z - x$ whenever $x \leq z$ and 0 otherwise. Given $J \in \mathrm{Int}(\mathbb{R})$, let $\mathrm{Par}(J)$ be the collection of *partitions* of $J$, i.e., finite collections $P \subseteq \mathrm{int}(J)$ of mutually disjoint subsets such that $J \setminus \bigcup_{I \in P} I$ is finite. For each $I \in \mathrm{int}(J)$, let $\psi_I : \mathrm{OP}(|J|, \mathbb{R}) \rightharpoonup \widehat{\mathbb{R}}$ be the partial function associated to the nonempty subset $I$ as in Example 4.8. In particular, the set of all bounded $\mathbb{R}$-valued functions on $J$ coincides with

$$\mathcal{O} = B(\psi_J) = \bigcap_{I \in P} B(\psi_I)$$

for any $P \in \mathrm{Par}(J)$. Let $\varphi_I : \mathrm{B}(\psi_I) \rightharpoonup \mathbb{R}$ be the extremizable partial function defined as in Remark 4.9 for each $I \in \mathrm{int}(J)$ and let $U_P, L_P : \mathcal{O} \to \mathbb{R}$ be the order-preserving functions defined by

$$U_P = \sum_{I \in P} m(I)\,\mathrm{uex}_{\varphi_I} \quad \text{and} \quad L_P = \sum_{I \in P} m(I)\,\mathrm{lex}_{\varphi_I}\,.$$

By Remark 4.9 it is clear that for each $f \in \mathcal{O}$, $U_P(f)$ coincides with the usual *upper Darboux sums* of $f$ and $L_P(f)$ coincides with the usual *lower Darboux sums* of $f$; see, e.g., [Rudin 1953]. On the other hand, for each $P \in \mathrm{Par}(J)$ consider the partial function $\varphi_P : \mathcal{O} \rightharpoonup \mathbb{R}$ defined by

$$\varphi_P(f) = \sum_{I \in P} m(I)\varphi_I(f) \tag{7}$$

for each element $f$ of

$$\mathrm{dom}(\varphi_P) = \bigcap_{I \in P} \mathrm{dom}(\varphi_I).$$

Since $\varphi_P$ is clearly encompassing, it is also extremizable by Corollary 4.6. Moreover,

$$\mathrm{lex}_{\varphi_P} \le L_P \le U_P \le \mathrm{uex}_{\varphi_P}, \tag{8}$$

as each term in the above chain of inequalities restricts to $\varphi_P$ on $\mathrm{dom}(\varphi_P)$. Since each function in $\mathcal{O}$ attains only finitely many values, $\mathrm{dom}(\varphi_P) \cong \mathbb{R}^N$ for some integer $N$. In particular, Corollary 4.6 ensures that if $\rho_P : \mathcal{O} \rightharpoonup \mathrm{dom}(\varphi_P)$ denotes the identity function on $\mathrm{dom}(\varphi_P)$ then $\rho_P$ is extremizable. Therefore

$$\mathrm{uex}_{\varphi_P} \le \varphi_P \circ \mathrm{uex}_{\rho_P} = \sum_{I \in P} m(I)(\varphi_I \circ \mathrm{uex}_{\rho_P}) \le \sum_{I \in P} m(I)\,\mathrm{uex}_{\varphi_I \circ \rho_P} = U_P. \tag{9}$$

Combined with an analogous estimate for $\mathrm{lex}_{\varphi_P}$ and (8), (9) shows that $\mathrm{uex}_{\varphi_P} = U_P$ and $\mathrm{lex}_{\varphi_P} = L_P$. Since $m(I) = m(I_1) + m(I_2)$ whenever $I \setminus (I_1 \cup I_2)$ is finite and $I \subseteq I'$ implies $\varphi_I(f) = \varphi_{I'}(f)$ for each $f \in \mathrm{dom}(\varphi'_I)$, we have for each $f \in \mathrm{dom}(\varphi_P) \cap \mathrm{dom}(\varphi_{P'})$

$$\varphi_P(f) = \sum_{I \in P} m(I)\varphi_I(f) = \sum_{I \in P} \sum_{I' \in P'} m(I \cap I')\varphi_{I \cap I'}(f) = \sum_{I' \in \mathcal{P}'} m(I')\varphi_{I'}(f) = \varphi_{P'}(f).$$

Let $\varphi$ be the common extension of the compatible set $\{\varphi_P\}_{P \in \mathrm{Par}(J)}$. In particular, $\mathrm{dom}(\varphi)$ is the set of *step functions on* $J$, i.e., the set of all functions on $|J|$ that are constant on each interval of some partition of $J$. Combining Remark 3.6 with Corollary 4.6 we conclude that $\varphi$ is encompassing and thus extremizable. By Remark 3.7, an order-preserving function $g : \mathcal{O} \to \mathbb{R}$ restricts to $\varphi$ on $\mathrm{dom}(\varphi)$ if and only if

$$g \in \bigcap_{P \in \mathrm{Par}(J)} [\mathrm{lex}_{\varphi_P}, \mathrm{uex}_{\varphi_P}] = \bigcap_{P \in \mathrm{Par}(J)} [L_P, U_P].$$

Hence, $\mathrm{lex}_\varphi$ and $\mathrm{uex}_\varphi$ coincide with the lower and upper integrals of $f$ on $J$, respectively. In particular, $\mathrm{Dar}(\varphi)$ coincides with the set of functions on $J$ that are integrable in the sense of Riemann and $\mathrm{ex}_\varphi$ is the Riemann integral. This is in fact the motivating example for the philosophy of Section 1: to define the Riemann integral it is sufficient to feed the "obvious" definition for step functions, given by (7), into the machinery of Darboux calculus to automatically obtain the correct general definition.

**Theorem 7.10** (linearity). *Let* $\psi : \mathrm{OP}(\mathcal{O}, \mathbb{R}) \rightharpoonup \mathbb{R}$ *be a partial function such that*

(1) $\psi$ *is encompassing*;

(2) $\mathrm{dom}(\psi)$ *is an* $\mathbb{R}$*-linear subspace of* $\mathrm{OP}(\mathcal{O}, \mathbb{R})$;

(3) $\psi$ *is an* $\mathbb{R}$*-linear transformation.*

*Then for every* $f_1, f_2 \in \mathrm{OP}(\mathcal{O}, \mathbb{R})$ *and for every nonnegative* $a_1, a_2 \in \mathbb{R}$

$$\mathrm{uex}_\psi(a_1 f_1 + a_2 f_2) \leq a_1 \,\mathrm{uex}_\psi(f_1) + a_2 \,\mathrm{uex}_\psi(f_2) \tag{10}$$

*and similarly*

$$a_1 \,\mathrm{lex}_\psi(f_1) + a_2 \,\mathrm{lex}_\psi(f_2) \leq \mathrm{lex}_\psi(a_1 f_1 + a_2 f_2). \tag{11}$$

*Moreover*

$$-\mathrm{lex}_\psi(f) = \mathrm{uex}_\psi(-f) \tag{12}$$

*for every* $f \in \mathrm{OP}(\mathcal{O}, \mathbb{R})$. *In particular*, $\mathrm{Dar}(\psi)$ *is an* $\mathbb{R}$*-linear subspace of* $\mathrm{OP}(\mathcal{O}, \mathbb{R})$ *and* $\mathrm{ex}_\psi$ *is* $\mathbb{R}$*-linear.*

*Proof.* Since $\psi$ is encompassing, it is extremizable. By the additivity of $\psi$, the assignment $f_1 \mapsto \mathrm{uex}_\psi(f_1 + \eta_2) - \psi(\eta_2)$ coincides with $\psi$ on $\mathrm{dom}(\psi)$ for each fixed $\eta_2 \in \mathrm{dom}(\psi)$. Therefore

$$\mathrm{uex}_\psi(f_1 + \eta_2) \leq \mathrm{uex}_\psi(f_1) + \psi(\eta_2) \tag{13}$$

for every $f_1 \in \mathrm{OP}(\mathcal{O}, \mathbb{R})$ and for every $\eta_2 \in \mathrm{dom}(\psi)$. In particular

$$\mathrm{uex}_\psi(f_1) + \psi(\eta_2) = \mathrm{uex}_\psi((f_1 + \eta_2) + (-\eta_2)) + \psi(\eta_2) \leq \mathrm{uex}_\psi(f_1 + \eta_2)$$

and thus the inequality in (13) is actually an equality. Therefore, for each $f_1 \in \mathrm{OP}(\mathcal{O}, \mathbb{R})$ the assignment $f_2 \mapsto \mathrm{uex}_\psi(f_1 + f_2) - \mathrm{uex}_\psi(f_1)$ coincides with $\psi$ on $\mathrm{dom}(\psi)$. This proves (10) when $a_1 = a_2 = 1$. Since $\psi$ is compatible with scalar multiplication, $a$ is a positive real number and the assignment $f \mapsto a^{-1} \,\mathrm{uex}_\psi(af)$ coincides with $\psi$ on $\mathrm{dom}(\psi)$. Therefore, $\mathrm{uex}_\psi(af) \leq a \,\mathrm{uex}_\psi(f)$, which in turn implies

$$a \,\mathrm{uex}_\psi(f) = a \,\mathrm{uex}_\psi(a^{-1}(af)) \leq \mathrm{uex}_\psi(af).$$

As a result, $a \,\mathrm{uex}_\psi(f) = \mathrm{uex}_\psi(af)$ for every positive real number $a$ and for every $f \in \mathrm{OP}(\mathcal{O}, \mathbb{R})$. This proves (10) and (11) is proved similarly. Since $\psi$ is odd,

the assignments $f \mapsto -\operatorname{lex}_\psi(-f)$ and $f \mapsto -\operatorname{uex}_\psi(-f)$ both restrict to $\psi$ on $\operatorname{dom}(\psi)$ and thus

$$\operatorname{lex}_\psi(f) \leq -\operatorname{uex}_\psi(-f) \leq -\operatorname{lex}_\psi(-f) \leq \operatorname{uex}_\psi(f) \tag{14}$$

for every $f \in \mathrm{OP}(\mathcal{O}, \mathbb{R})$. The first inequality in (14) implies $\operatorname{uex}_\psi(-f) \leq -\operatorname{lex}_\psi(f)$, while the last inequality of (14) implies $-\operatorname{lex}_\psi(f) \leq \operatorname{uex}_\psi(-f)$ and thus (12). The last statement is a straightforward consequence of (10)–(12). □

**Example 7.11.** Specializing Theorem 7.10 to Examples 7.5–7.9 we obtain the well-known linearity theorems for limits of sequences, continuous functions, limits of functions of real variable and integrals.

## References

[Edalat and Lieutier 2004] A. Edalat and A. Lieutier, "Domain theory and differential calculus (functions of one variable)", *Math. Structures Comput. Sci.* **14**:6 (2004), 771–802. MR Zbl

[Fuchs 1963] L. Fuchs, *Partially ordered algebraic systems*, Pergamon, Oxford, 1963. MR Zbl

[Mac Lane 1971] S. Mac Lane, *Categories for the working mathematician*, Graduate Texts in Math. **5**, Springer, 1971. MR Zbl

[Rudin 1953] W. Rudin, *Principles of mathematical analysis*, McGraw-Hill, New York, 1953. MR Zbl

[Taylor 1999] P. Taylor, *Practical foundations of mathematics*, Cambridge Studies in Advanced Math. **59**, Cambridge Univ. Press, 1999. MR Zbl

[Taylor 2010] P. Taylor, "A lambda calculus for real analysis", *J. Log. Anal.* **2** (2010), art. id. 5. MR Zbl

[Univalent Foundations 2013] The Univalent Foundations Program, *Homotopy type theory: univalent foundations of mathematics*, Inst. Advanced Study, Princeton, NJ, 2013. MR Zbl

Received: 2016-05-06 Revised: 2017-12-20 Accepted: 2018-05-22

maldi2@vcu.edu *Department of Mathematics and Applied Mathematics, Virginia Commonwealth University, Richmond, VA, United States*

mcclearyaj@vcu.edu *Department of Mathematics, Colorado State University, Fort Collins, CO, United States*

INVOLVE 12:3 (2019)
dx.doi.org/10.2140/involve.2019.12.381

# A countable space with an uncountable fundamental group

Jeremy Brazas and Luis Matos

(Communicated by Józef H. Przytycki)

Traditional examples of spaces that have an uncountable fundamental group (such as the Hawaiian earring space) are path-connected compact metric spaces with uncountably many points. We construct a $T_0$ compact, path-connected, locally path-connected topological space $H$ with countably many points but with an uncountable fundamental group. The construction of $H$, which we call the "coarse Hawaiian earring" is based on the construction of the usual Hawaiian earring space $\mathbb{H} = \bigcup_{n\geq 1} C_n$ where each circle $C_n$ is replaced with a copy of the four-point "finite circle".

## 1. Introduction

Since fundamental groups are defined in terms of maps from the unit interval $[0, 1]$, students are often surprised to learn that spaces with finitely many points can be path connected and have nontrivial fundamental groups. In fact, it has been known since the 1960s that the homotopy theory of finite spaces is quite rich [McCord 1966; Stong 1966]. The algebraic topology of finite topological spaces has gained significant interest since Peter May's Research Experience for Undergraduates (REU) Summer Program at the University of Chicago in 2003; see [May 2003a; 2003b; 2003c]. For more recent theory and applications of the algebraic topology of finite spaces, we refer to [Barmak 2011; Barmak and Minian 2008; Cianci and Ottina 2016].

While it is reasonable to expect that all finite connected spaces have finitely generated fundamental groups, it is rather remarkable that for every finitely generated group $G$ one can construct a finite space $X$ so that $\pi_1(X, x_0) \cong G$. In fact, every finite simplicial complex is weakly homotopy equivalent to a finite space [McCord 1966]. In the same spirit, we consider fundamental groups of spaces with coarse topologies.

*MSC2010:* primary 54D10, 55Q52; secondary 57M05, 57M10.

*Keywords:* fundamental group, countable topological space, finite topological space, Hawaiian earring, coarse Hawaiian earring.

It is well known that there are connected, locally path-connected compact metric space whose fundamental groups are uncountable [Cannon and Conner 2000]. Since finite spaces can only have finitely generated fundamental groups, we must extend our view to spaces with countably many points. We prove the following theorem.

**Theorem 1.** *There exists a connected, locally path-connected, compact, $T_0$ topological space $H$ with countably many points such that $\pi_1(H, w_0)$ is uncountable.*

Since countable simplicial complexes have countable fundamental groups, Theorem 1 shows that countable spaces need not be weakly homotopy equivalent to countable simplicial complexes. Thus the relationship between weak homotopy types of finite spaces and finite simplicial complexes cannot be fully generalized to countable spaces.

To construct the space $H$ in Theorem 1, we must consider spaces which are not locally finite, that is, spaces which have a point such that every neighborhood of that point contains infinitely many other points. Additionally, since our example must be path connected, the following lemma demands that such a space cannot have the $T_1$ separation axiom.

**Lemma 2.** *Every countable $T_1$ space is totally path disconnected.*

*Proof.* If $X$ is countable and $T_1$ and $\alpha : [0, 1] \to X$ is a nonconstant path, then $\{\alpha^{-1}(x) \mid x \in X\}$ is a nontrivial, countable partition of $[0, 1]$ into closed sets. However, it is a classical result in general topology that such a partition of $[0, 1]$ is impossible [Sierpinski 1918]. □

Ultimately, we construct the space $H$ by modeling the construction of the traditional Hawaiian earring space $\mathbb{H}$, which is the prototypical space that fails to be semilocally simply connected and which does not admit a traditional universal covering space. The fundamental group of the Hawaiian earring is an uncountable group which plays a key role in the homotopy classification of one-dimensional Peano continua given in [Eda 2010]. Due to the similarities between $\mathbb{H}$ and $H$, we call $H$ the *coarse Hawaiian earring*.

## 2. Fundamental groups

Let $X$ be a topological space with basepoint $x_0 \in X$. A *path* in $X$ is a continuous function $\alpha : [0, 1] \to X$. We say $X$ is *path connected* if every pair of points $x, y \in X$ can be connected by a path $p : [0, 1] \to X$ with $p(0) = x$ and $p(1) = y$. All spaces in this paper will be path connected.

We say a path $p$ is a *loop* based at $x_0$ if $\alpha(0) = \alpha(1)$. Let $\Omega(X, x_0)$ be the set of continuous functions $p : [0, 1] \to X$ such that $p(0) = p(1) = x_0$. Let $\alpha^- : [0, 1] \to X$ be the reverse path of $\alpha$ defined as $\alpha^-(t) = \alpha(1 - t)$. If $\alpha$ and $\beta$ are paths in $X$

satisfying $\alpha(1) = \beta(0)$, let $\alpha \cdot \beta$ be the concatenation defined piecewise as

$$\alpha \cdot \beta(t) = \begin{cases} \alpha(2t), & 0 \le t \le \frac{1}{2}, \\ \beta(2t-1), & \frac{1}{2} \le t \le 1. \end{cases}$$

More generally, if $\alpha_1, \ldots, \alpha_n$ is a sequence of paths such that $\alpha_i(1) = \alpha_{i+1}(0)$ for $i = 1, \ldots, n-1$, let $\prod_{i=1}^{n} \alpha_i$ be the path defined as $\alpha_i$ on the interval $[(i-1)/n, i/n]$.

Two loops $\alpha$ and $\beta$ based at $x_0$ are said to be *homotopic* if there is a map $H : [0, 1] \times [0, 1] \to X$ such that $H(s, 0) = \alpha(s)$, $H(s, 1) = \beta(s)$ and $H(0, t) = H(1, t) = x_0$ for all $s, t \in [0, 1]$. We write $\alpha \simeq \beta$ if $\alpha$ and $\beta$ are homotopic. Homotopy $\simeq$ is an equivalence relation on the set of loops $\Omega(X, x_0)$. The equivalence class $[\alpha]$ of a loop $\alpha$ is called the *homotopy class* of $\alpha$. The set of homotopy classes $\pi_1(X, x_0) = \Omega(X, x_0)/\simeq$ is called the *fundamental group* of $X$ at $x_0$. It is a group when it has multiplication $[\alpha] * [\beta] = [\alpha \cdot \beta]$ and $[\alpha]^{-1} = [\alpha^-]$ is the inverse of $[\alpha]$ [Munkres 2000]. A space $X$ is *simply connected* if $X$ is path connected and $\pi_1(X, x_0)$ is isomorphic to the trivial group. Finally, a map $f : X \to Y$ such that $f(x_0) = y_0$ induces a well-defined homomorphism $f_* : \pi_1(X, x_0) \to \pi_1(Y, y_0)$ given by $f_*([\alpha]) = [f \circ \alpha]$.

Fundamental groups are often studied using maps called covering maps. For this theory and other aspects of algebraic topology, we refer to [Munkres 2000; Spanier 1966], taking our conventions primarily from the former.

**Definition 3.** Let $p : \widetilde{X} \to X$ be a map. An open set $U \subseteq X$ is *evenly covered* by $p$ if $p^{-1}(U) \subseteq \widetilde{X}$ is the disjoint union $\bigsqcup_{\lambda \in \Lambda} V_\lambda$, where $V_\lambda$ is open in $\widetilde{X}$ and $p|_{V_\lambda} : V_\lambda \to U$ is a homeomorphism for every $\lambda \in \Lambda$. A *covering map* is a map $p : \widetilde{X} \to X$ such that every point $x \in X$ has an open neighborhood which is evenly covered by $p$. The space $\widetilde{X}$ is called a *covering space* of $X$. We call $p$ a *universal covering map* if $\widetilde{X}$ is simply connected.

**Remark 4.** An alternative definition of universal covering map appears in [Spanier 1966] where a covering map $p : \widetilde{X} \to X$ is defined to be universal if it is an initial object in the category of coverings over $X$, that is, if $\widetilde{X}$ is a covering space of every covering space of $X$. For general spaces (even locally path-connected compact metric spaces) the two definitions differ. Example 18 in Chapter 2 of [Spanier 1966] describes the twin cone $C\mathbb{H} \vee C\mathbb{H}$ over the Hawaiian earring $\mathbb{H}$ (sometimes called the Griffiths twin cone), which is a non-simply connected space whose only covering space is itself. Thus the identity map of the twin cone is a universal covering in the sense of [Spanier 1966] but not in the sense of [Munkres 2000]. On the other hand, one can use the covering space theory developed in [Munkres 2000] to confirm that the two definitions of "universal covering map" agree when $X$ is locally path connected and semilocally simply connected. Since we only consider covering maps over such spaces in this paper, we will not need to worry about the difference in the definitions.

An important property of covering maps is that for every path $\alpha : [0, 1] \to X$ such that $\alpha(0) = x_0$ and point $y \in p^{-1}(x_0)$ there is a unique path $\tilde{\alpha}_y : [0, 1] \to \widetilde{X}$ (called a *lift* of $\alpha$) such that $p \circ \tilde{\alpha}_y = \alpha$ and $\tilde{\alpha}_y(0) = y$.

**Lemma 5** [Munkres 2000, Theorem 54.6]. *A covering map $p : \widetilde{X} \to X$ such that $p(y) = x_0$ induces an injective homomorphism $p_* : \pi_1(\widetilde{X}, y) \to \pi_1(X, x_0)$. If $\alpha : [0, 1] \to X$ is a loop based at $x_0$, then $[\alpha] \in p_*(\pi_1(\widetilde{X}, y))$ if and only if $\tilde{\alpha}_y(1) = y$.*

A covering map $p : \widetilde{X} \to X$ induces a *lifting correspondence map* $\phi : \pi_1(X, x_0) \to p^{-1}(x_0)$ from the fundamental group of $X$ to the fiber over $x_0$ defined by the formula $\phi([\alpha]) = \tilde{\alpha}_y(1)$.

**Lemma 6** [Munkres 2000, Theorem 54.4]. *If $p : \widetilde{X} \to X$ is a covering map, then the lifting correspondence $\phi : \pi_1(X, x_0) \to p^{-1}(x_0)$ is surjective. If $p$ is a universal covering map, then $p$ is bijective.*

**Example 7.** Let $S^1 = \{(x, y) \mid x^2 + y^2 = 1\}$ be the unit circle and $b_0 = (1, 0)$. The exponential map $\epsilon : \mathbb{R} \to S^1$, $\epsilon(t) = (\cos(2\pi t), \sin(2\pi t))$, defined on the real line is a covering map such that $\epsilon^{-1}(b_0) = \mathbb{Z}$ is the set of integers. The lifting correspondence for this covering map $\phi : \pi_1(S^1, b_0) \to \epsilon^{-1}(b_0) = \mathbb{Z}$ is an isomorphism when $\mathbb{Z}$ is the additive group of integers. See [Munkres 2000, Theorem 54.5] for a proof.

## 3. Some basic finite spaces

A *finite space* is a topological space $X = \{x_1, x_2, \ldots, x_n\}$ with finitely many points.

**Example 8.** The *coarse interval* is the three-point space $I = \left\{0, \frac{1}{2}, 1\right\}$ with topology generated by the basic sets the sets $\{0\}$, $\{1\}$, and $I$ (See Figure 1). In other words, the topology of $I$ is $T_I = \{I, \{0\}, \{1\}, \{0, 1\}, \varnothing\}$.

The coarse interval clearly satisfies the $T_0$ separation axiom. It is also path connected since we can define a continuous surjection $p : [0, 1] \to I$ by

$$p(t) = \begin{cases} 0, & t \in \left[0, \frac{1}{2}\right), \\ \frac{1}{2} & t = \frac{1}{2}, \\ 1, & t \in \left(\frac{1}{2}, 1\right], \end{cases}$$

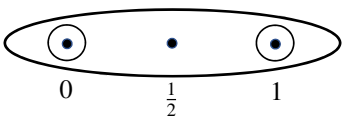


**Figure 1.** The coarse interval $I$. A basic open set is illustrated here as a bounded region whose interior contains the points of the set.

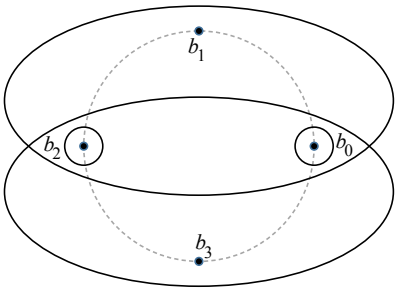


**Figure 2.** The coarse circle $S$ and it's basic open sets.

and the continuous image of a path-connected space is path connected. A space $X$ is contractible if the identity map $\mathrm{id} : X \to X$ is homotopic to a constant map $X \to X$. Every contractible space is simply connected.

**Lemma 9.** *The coarse interval $I$ is contractible.*

*Proof.* To show $I$ is contractible we define a continuous map $G : I \times [0, 1] \to I$ such that $G(x, 0) = x$ for $x \in I$ and $G(x, 1) = \frac{1}{2}$. The set $C = \big(\{0, 1\} \times \big[\frac{1}{2}, 1\big]\big) \cup \big(\big\{\frac{1}{2}\big\} \times [0, 1]\big)$ is closed in $I \times [0, 1]$. Define $G$ by

$$G(s, t) = \begin{cases} 0, & (s, t) \in \{0\} \times \big[0, \frac{1}{2}\big), \\ \frac{1}{2}, & (s, t) \in C, \\ 1, & (s, t) \in \{1\} \times \big[0, \frac{1}{2}\big). \end{cases}$$

This function is well-defined and continuous since $\{0\}$ and $\{1\}$ are open in $I$. □

**Corollary 10.** *$I$ is simply connected.*

For $n = 0, 1, 2, 3$, let $b_n = \big(\cos\big(\frac{1}{2}n\pi\big), \sin\big(\frac{1}{2}n\pi\big)\big) \in S^1$ be the points of the unit circle on the coordinate axes; i.e., $b_0 = (1, 0)$, $b_1 = (0, 1)$, $b_2 = (-1, 0)$, and $b_3 = (0, -1)$.

**Example 11.** The *coarse circle* is the four-point set $S = \{b_i \mid i = 0, 1, 2, 3\}$ with the topology generated by the basic sets $\{b_0, b_1, b_2\}$, $\{b_2, b_3, b_0\}$, $\{b_0\}$, and $\{b_2\}$ (see Figure 2). The entire topology of $S$ may be written as

$$T_S = \{S, \{b_0, b_1, b_2\}, \{b_2, b_3, b_0\}, \{b_0, b_2\}, \{b_0\}, \{b_2\}, \varnothing\}.$$

Observe that the open sets $U_1 = \{b_0, b_1, b_2\}$ and $U_2 = \{b_2, b_3, b_0\}$ are homeomorphic to $I$ when they are given the subspace topology. Since $S$ is the union of two path-connected subsets with nonempty intersection, $S$ is also path connected.

**Remark 12.** The spaces $I$ and $S$ have appeared many times in the literature. The space $S$ is sometimes called the "finite circle". We use the term "coarse circle" since we are considering it within the broader context of infinite spaces with non-$T_1$

topologies. The space $S$ is the smallest finite space having the same weak homotopy type as the usual circle $S^1$. In fact, this is a special case of a more general result on minimal $(2n+2)$-point models of $n$-spheres (see [Barmak 2011, Chapter 3]): there exists a space with $2n+2$ points weakly homotopy equivalent to $S^n$ and moreover any finite space that is weakly homotopy equivalent to the $n$-sphere $S^n$ must have at least $2n+2$ points.

## 4. The coarse line as a covering space

As indicated in Remark 12, it follows from the much more sophisticated theory in [Stong 1966] that $S$ has the weak homotopy type of $S^1$. To keep the current paper self-contained, we devote this section to a direct proof of the fact that $\pi_1(S, b_0)$ is isomorphic to the infinite cyclic group $\mathbb{Z}$, i.e., the additive group of integers, by constructing a map $g : S^1 \to S$ that induces an isomorphism on fundamental groups.

**Example 13.** The *coarse line* is the set $L = \left\{ \frac{1}{4}n \in \mathbb{R} \,\middle|\, n \in \mathbb{Z} \right\}$ with the topology generated by the basis consisting of the sets $A_n = \left\{ \frac{1}{2}n \right\}$ and $B_n = \left\{ \frac{1}{2}n,\ \frac{1}{4}(2n+1),\ \frac{1}{2}(n+1) \right\}$ for each $n \in \mathbb{Z}$ (see Figure 3). Even though $L$ is not a finite space, it is a countable space with a $T_0$ but non-$T_1$ topology.

**Lemma 14.** *$L$ is simply connected.*

*Proof.* The set $L_n = L \cap \left[ -\frac{1}{2}n, \frac{1}{2}n \right]$ is open in $L$ since it is the union of the basic sets $B_k = \left\{ \frac{1}{2}k,\ \frac{1}{4}(2k+1),\ \frac{1}{2}(k+1) \right\}$, $k = -n, \ldots, n-1$, with the subspace topology of $L$.

It follows from the classical van Kampen theorem [Munkres 2000, Theorem 70.2] that if $X = U \cup V$, where $U$, $V$ are open in $X$ and $U$, $V$, $U \cap V$ are simply connected, then $X$ is simply connected. We will apply this fact inductively to prove that $L_n$ is simply connected for all $n \geq 1$.

Since $B_k \cong I$ for each $k$, we know $B_k$ is simply connected for each $k$. Observe that $L_1 = B_{-1} \cup B_0$, where $B_{-1} \cap B_0 = \{0\}$ is simply connected since it only has one point. Thus $L_1$ is simply connected by the van Kampen theorem. Now suppose $L_n$ is simply connected. Since $L_n$, $B_n$, and $L_n \cap B_n = \left\{ \frac{1}{2}n \right\}$ are all simply connected, $L_n \cup B_n$ is simply connected by the van Kampen theorem. Similarly, since $L_n \cup B_n$, $B_{-n-1}$, and $(L_n \cup B_n) \cap B_{-k-1} = \left\{ -\frac{1}{2}n \right\}$ are all simply connected, $L_{n+1} = B_{-n-1} \cup L_n \cup B_n$ is simply connected by the van Kampen theorem. Thus $L_n$ is simply connected for all $n \geq 1$.

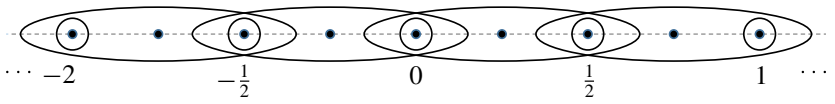


**Figure 3.** The basic open sets generating the topology of the coarse line $L$.

Since $L$ is the union of the path-connected sets $L_n$, all of which contain 0, it follows that $L$ is path connected. Now suppose $\alpha : [0, 1] \to L$ is a path such that $\alpha(0) = \alpha(1)$. Since $[0, 1]$ is compact, the image $\alpha([0, 1])$ is compact. But $\{L_n \mid n \geq 1\}$ is an open cover of $L$ such that $L_n \subseteq L_{n+1}$. Since $\alpha$ must have image in a finite subcover of $\{L_n \mid n \geq 1\}$, we must have $\alpha([0, 1]) \subseteq L_n$ for some $n$. But $L_n$ is simply connected, showing that $\alpha$ is homotopic to the constant loop at 0. This proves $\pi_1(L, 0)$ is the trivial group; i.e., $L$ is simply connected. □

Just like the usual covering map $\epsilon : \mathbb{R} \to S^1$ used to compute $\pi_1(S^1, b_0)$, we define a similar covering map in the coarse situation.

**Example 15.** Consider the function $p : L \to S$ from the coarse line to the coarse circle which is the restriction of the covering map $\epsilon : \mathbb{R} \to S^1$. More directly, define $p\left(\frac{1}{4}n\right) = b_{n \bmod 4}$. We check that the preimage of each basic open set in $S$ can be written as a union of basic open sets in $L$. Since

- $p^{-1}(\{b_0\}) = \mathbb{Z} = \bigcup_{k\in\mathbb{Z}} A_{2k}$,
- $p^{-1}(\{b_2\}) = \frac{1}{2} + \mathbb{Z} = \bigcup_{k\in\mathbb{Z}} A_{2k+1}$,
- $p^{-1}(U_1) = \bigcup_{k\in\mathbb{Z}} B_{2k}$,
- $p^{-1}(U_2) = \bigcup_{k\in\mathbb{Z}} B_{2k+1}$,

we can conclude that $p$ is continuous.

**Lemma 16.** *The function $p : L \to S$ is a covering map.*

*Proof.* We claim that the sets $U_1, U_2$ are evenly covered by $p$. Notice that $p^{-1}(U_1) = \bigcup_{k\in\mathbb{Z}} B_{2k}$ is a disjoint union where each $B_{2k}$ is open. Recall that both $B_{2k}$ and $U_1$ are homeomorphic to $I$; specifically $p|_{B_{2k}} : B_{2k} \to U_1$ is a homeomorphism. Thus $U_1$ is evenly covered. Similarly, $p^{-1}(U_2)$ is the disjoint union $\bigcup_{k\in\mathbb{Z}} B_{2k+1}$ where each $B_{2k+1}$ is open and is mapped homeomorphically on to $U_2$ by $p$. □

Since $p : L \to S$ is a covering map and $L$ is simply connected, $p$ is a universal covering map. The proof of the following theorem is similar to the proof that the lifting correspondence for $\epsilon$ is a group isomorphism. We remark that even though $L$ is not a topological group, the shift map $\sigma_n : L \to L$, $\sigma(t) = t + n$, is a homeomorphism satisfying $p \circ \sigma_n = p$ for each $n \in \mathbb{Z}$.

**Theorem 17.** *The lifting correspondence $\phi : \pi_1(S, b_0) \to p^{-1}(b_0) = \mathbb{Z}$ is a group isomorphism where $\mathbb{Z}$ has the usual additive group structure.*

*Proof.* Since $p : L \to S$ is a covering map and $L$ is simply connected, $\phi$ is bijective by Lemma 6. Suppose $\alpha, \beta : [0, 1] \to S$ are loops based at $b_0$. Respectively, let $\tilde{\alpha}_0 : [0, 1] \to L$ and $\tilde{\beta}_0 : [0, 1] \to L$ be the unique lifts of $\alpha$ and $\beta$ starting at 0. Since $\tilde{\alpha}_0(1) \in p^{-1}(b_0) = \mathbb{Z}$, we have $\phi([\alpha]) = \tilde{\alpha}_0(1) = n$ for some integer $n$. Similarly, $\phi([\beta]) = \tilde{\beta}_0(1) = m$ for some integer $m$.

Consider the concatenated path $\gamma = \tilde{\alpha}_0 \cdot (\sigma_n \circ \tilde{\beta}_0) : [0, 1] \to L$ from 0 to $m+n$. Since $p \circ \sigma_n = p$, we have

$$\begin{aligned} p \circ \gamma &= p \circ (\tilde{\alpha}_0 \cdot (\sigma_n \circ \tilde{\beta}_0)) \\ &= (p \circ \tilde{\alpha}_0) \cdot (p \circ \sigma_n \circ \tilde{\beta}_0) \\ &= (p \circ \tilde{\alpha}_0) \cdot (p \circ \tilde{\beta}_0) = \alpha \cdot \beta, \end{aligned}$$

which means that $\gamma$ is a lift of $\alpha \cdot \beta$ starting at 0. Since lifts are unique, this means $\gamma = \widetilde{\alpha \cdot \beta}_0$. It follows that $\phi([\alpha][\beta]) = \phi([\alpha \cdot \beta]) = \widetilde{\alpha \cdot \beta}_0(1) = \gamma(1) = m+n$. This proves $\phi$ is a group homomorphism. □

Both $\pi_1(S^1, b_0)$ and $\pi_1(S, b_0)$ are isomorphic to the infinite cyclic group $\mathbb{Z}$. In fact, we can define maps which induce the isomorphism between the two fundamental groups.

Let $f : \mathbb{R} \to L$ be the map defined so that $f\big(\big(\frac{1}{2}n - \frac{1}{4}, \frac{1}{2}n + \frac{1}{4}\big)\big) = \frac{1}{2}n$ and $f\big(\frac{1}{2}n + \frac{1}{4}\big) = \frac{1}{2}n + \frac{1}{4}$ for each $n \in \mathbb{Z}$. Notice that $p \circ f$ is constant on each fiber $\epsilon^{-1}(x)$, $x \in S^1$. Therefore, there is an induced map $g : S^1 \to S$ such that $g \circ \epsilon = p \circ f$.

As mentioned at the end of the previous section, the following proposition is a special case of more general results on weak homotopy types of finite spaces in [Stong 1966].

**Proposition 18.** *The induced homomorphism $g_* : \pi_1(S^1, b_0) \to \pi_1(S, b_0)$ is a group isomorphism.*

*Proof.* Recall that $\epsilon^{-1}(b_0) = \mathbb{Z}$ and $p^{-1}(b_0) = \mathbb{Z}$ and notice that the restriction to the fibers $f|_{\mathbb{Z}} : \mathbb{Z} \to \mathbb{Z}$ is the identity map. Let $i : [0, 1] \to \mathbb{R}$ be the inclusion and note $f \circ i : [0, 1] \to L$ is a path from 0 to 1. The group $\pi_1(S^1, b_0)$ is freely generated by the homotopy class of $\alpha = \epsilon \circ i$ and $\pi_1(S, b_0)$ is freely generated by the homotopy class of $p \circ f \circ i$. Since $g_*([\epsilon \circ i]) = [g \circ \epsilon \circ i] = [p \circ f \circ i]$, the homomorphism $g_*$ maps one free generator to the other and it follows that $g_*$ is an isomorphism. □

## 5. The coarse Hawaiian earring

Let $C_n = \{(x, y) \in \mathbb{R}^2 \mid (x - 1/n)^2 + y^2 = 1/n^2\}$ be the circle of radius $1/n$ centered at $(1/n, 0)$. The *Hawaiian earring* is the countably infinite union $\mathbb{H} = \bigcup_{n \geq 1} C_n$ of these circles over the positive integers (see Figure 4, left). We construct our countable version of $\mathbb{H}$ by replacing the usual circle with the coarse circle studied in the previous sections.

Let $w_0 = (0, 0)$, and for integers $n \geq 1$ define $x_n = (1/n, -1/n)$, $y_n = (2/n, 0)$, and $z_n = (1/n, 1/n)$. Let $D_n = \{w_0, x_n, y_n, z_n\}$ and $H = \bigcup_{n \geq 1} D_n$. Note that $H$ is a countable subset of $\mathbb{H}$ (see Figure 4, right).

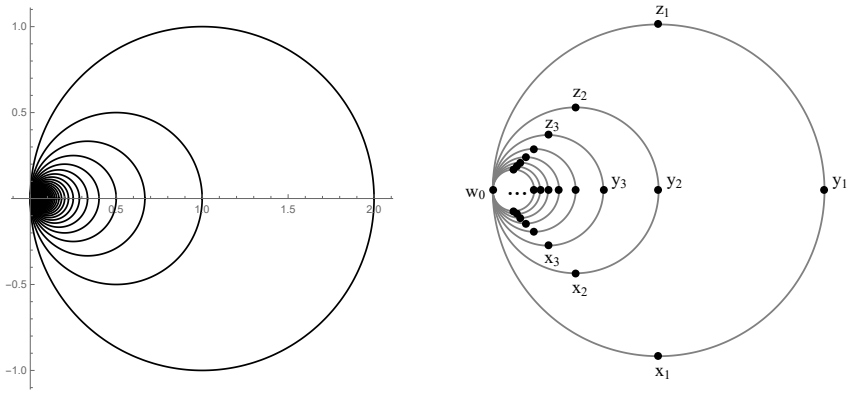


**Figure 4.** Left: the Hawaiian earring $\mathbb{H}$. Right: the underlying set of $H$ as a subset of $\mathbb{H}$. The intersection of the $n$-th circle $C_n$ and $H$ is the four-point set $D_n = \{w_0, x_n, y_n, z_n\}$.

**Proposition 19.** *Let $\mathscr{B}$ be the collection of subsets of $H$ of the form $\{x_n\}$, $\{z_n\}$, $\{x_n, y_n, z_n\}$, and $V_n = \bigcup_{j\geq n} D_j \cup \{x_n \mid n \geq 1\} \cup \{z_n \mid n \geq 1\}$ for $n \geq 1$. Then $\mathscr{B}$ is a basis for a topology on $H$.*

*Proof.* Since $H = V_1$, it is clear that every element of $H$ is contained in at least one element of $\mathscr{B}$. Suppose $x \in B_1 \cap B_2$, where $B_1, B_2 \in \mathscr{B}$. We must show there exists $B_3 \in \mathscr{B}$ such that $x \in B_3 \subseteq B_1 \cap B_2$. We complete the proof by defining $B_3$ for all possible cases of intersection:

(1) If one of $B_1$ or $B_2$ is of the form $\{x_n\}$ or $\{z_n\}$ then we may take $B_3$ to be this singleton.

(2) If $B_1 = \{x_m, y_m, z_m\}$ and $B_2 = \{x_n, y_n, z_n\}$, then we must have $n = m$ since these sets are disjoint if $n \neq m$. Set $B_3 = \{x_m, y_m, z_m\}$.

(3) Note that $V_n \subseteq V_m$ if $n \geq m$. Thus if $B_1 = V_m$ and $B_2 = V_n$, we may set $B_3 = V_m \cap V_n = V_{\max\{m,n\}} \in \mathscr{B}$.

(4) If $B_1 = \{x_m, y_m, z_m\}$ and $B_2 = V_n$, then $B_1 \cap B_2 = \{x_m, z_m\}$ and we may take $B_3$ to be the singleton (either $\{x_m\}$ or $\{z_m\}$) containing $x$. □

**Definition 20.** The *coarse Hawaiian earring* is the set $H$ with the topology generated by the basis consisting of subsets of the form $\{x_n\}$, $\{z_n\}$, $\{x_n, y_n, z_n\}$, and $V_n$ for $n \geq 1$.

A topological space whose topology is closed under arbitrary intersection is called an *Alexandroff space* [Arenas 1999]. Such spaces were introduced by P. Alexandroff [1937] and may also appear in modern literature under the name "A-space" or "Alexandroff-discrete space". Regarding the coarse Hawaiian earring, notice that

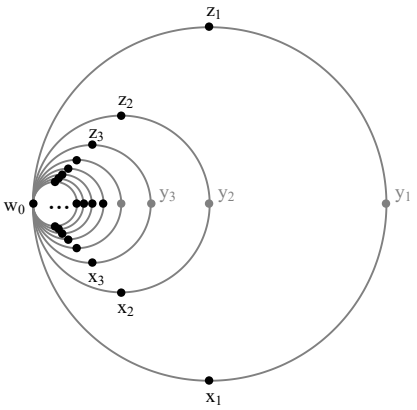


**Figure 5.** The basic open neighborhood $V_5$ of $w_0$ contains all points of $H$ except $y_1, y_2, y_3, y_4$, which are shaded lighter. In particular, $V_5$ contains the four-point set $D_n$ for all $n \geq 5$.

for all $n \geq 2$, the basic open neighborhood $V_n = H \setminus \{y_1, \ldots, y_{n-1}\}$ contains all but finitely many of the coarse circles $D_j$ (see Figure 5). These sets form a neighborhood base at $w_0$ so that $H$ is not an Alexandroff space.

**Remark 21.** Notice that the four-point subset $D_n \subset H$ is homeomorphic to the coarse circle $S$ when equipped with the subspace topology inherited from $H$. An explicit homeomorphism $S \to D_n$ is given by taking $b_0 \mapsto x_n$, $b_1 \mapsto w_0$, $b_2 \mapsto z_n$, and $b_3 \mapsto y_n$.

**Proposition 22.** *$H$ is a path-connected, locally path-connected, compact, $T_0$ space which is not $T_1$.*

*Proof.* Since $D_n$ is homeomorphic to $S$, we know $D_n$ is path connected for all $n \geq 1$. Moreover, since $w_0 \in \bigcap_{n\geq 1} D_n$ and $H = \bigcup_{n\geq 1} D_n$, it follows that $H$ is path connected. To see that $H$ is locally path connected, we check that every basic open set is path connected. Certainly, $\{x_n\}$ and $\{z_n\}$ are path connected. Since $\{x_n, y_n, z_n\}$ is homeomorphic to $I$ when it is given the subspace topology of $H$, this basic open set is path connected. Additionally, the subspace $\{w_0, x_n, y_n\} \subseteq H$ is homeomorphic to $I$ and is path connected. Therefore, since $V_n$ is the union $\bigcup_{j\geq n} D_n \cup \bigcup_{n\geq 1}\{w_0, x_n, y_n\}$ of sets all of which are path connected and contain $w_0$, we can conclude that $V_n$ is path connected. This proves $H$ is locally path connected.

To see that $H$ is compact let $\mathscr{U}$ be an open cover of $H$. Since the only basic open sets containing $w_0$ are the sets $V_n$, there must be a $U_0 \in \mathscr{U}$ such that $w_0 \in V_n \subseteq U_0$ for some $n$. For $i = 1, \ldots, n-1$, find a set $U_i \in \mathscr{U}$ such that $y_i \in U_i$. Now $\{U_0, U_1, \ldots, U_{n-1}\}$ is a finite subcover of $\mathscr{U}$. This proves $H$ is compact.

To see that $H$ is $T_0$, we pick two points $a, b \in H$. If $a = w_0$ and $b = y_n$, then $a \in V_{n+1}$ but $b \notin V_{n+1}$. If $a = w_0$ and $b \in \{x_n, z_n\}$, then $b$ lies in the open set

$\{x_n, y_n, z_n\}$ but $a$ does not. If $a \in \{x_n, z_n\}$ and $a \neq b$, then $\{a\}$ is open and does not contain $b$. This concludes all the possible cases of distinct pairs of points in $H$, proving that $H$ is $T_0$.

Lastly, $H$ is not $T_1$ since the every open neighborhood $V_n$ of $w_0$ contains the infinite set $\bigcup_{n\geq 1}\{w_0, x_n, z_n\}$. □

Since $D_n \cong S$, we have by Theorem 17 that $\pi_1(D_n, w_0) \cong \mathbb{Z}$ for all $n \geq 1$. Recall that if $A$ is a subspace of $X$, then a retraction is a map $r : X \to A$ such that the restriction $r|_A : A \to A$ is the identity map.

**Proposition 23.** *For each $n \geq 1$, the function $r_n : H \to D_n$ which is the identity on $D_n$ and collapses $\bigcup_{j\neq n} D_j$ to $w_0$ is a retraction.*

*Proof.* Since $D_n$ is a subspace of $H$, it suffices to show $r_n$ is continuous. We have

$$r_n^{-1}(\{x_n\}) = \{x_n\}, \quad r_n^{-1}(\{x_n, y_n, z_n\}) = \{x_n, y_n, z_n\},$$
$$r_n^{-1}(\{z_n\}) = \{z_n\}, \quad r_n^{-1}(\{w_0, x_n, y_n\}) = \{x_n\} \cup \{y_n\} \cup V_{n+1} \cup \bigcup_{j<n}\{x_j, y_j, z_j\}.$$

Since the pullback of each basic open set in $D_n$ is the union of basic open sets in $H$, $r_n$ is continuous. □

**Corollary 24.** *$H$ is not semilocally simply connected at $w_0$.*

*Proof.* Fix $n \geq 1$. We show that $V_n$ contains a loop $\alpha$ which is not null-homotopic in $H$. Let $\alpha : [0, 1] \to D_n$ be any loop based at $w_0$ such that $[\alpha]$ is not the identity element of $\pi_1(D_n, w_0)$. Let $i : D_n \to H$ be the inclusion map so that $r_n \circ i = \mathrm{id}_{D_n}$ is the identity map. Since $\pi_1$ is a functor, $(r_n)_* \circ i_* = (r_n \circ i)_* = \mathrm{id}_{\pi_1(D_n, w_0)}$ is the identity homomorphism of $\pi_1(D_n, w_0)$. In particular, $i \circ \alpha$ is a loop in $H$ with image in $D_n \subseteq V_n$ such that $(r_n)_*([i \circ \alpha]) = [\alpha]$ is not the identity of $\pi_1(D_n, w_0)$. Since homomorphisms preserve identity elements, $[i \circ \alpha]$ cannot be the identity element of $\pi_1(H, w_0)$. □

**Definition 25.** The *infinite product* of a sequence of groups $G_1, G_2, \ldots$ is denoted by $\prod_{n\geq 1} G_n$ and consists of all infinite sequences $(g_1, g_2, \ldots)$ with $g_n \in G_n$ for each $n \geq 1$. Group multiplication and inversion are evaluated componentwise. If $G_n = \mathbb{Z}$ for each $n \geq 1$, then the group $\prod_{n\geq 1} \mathbb{Z}$ consisting of sequences $(n_1, n_2, \ldots)$ of integers is called the *Baer–Specker group*.

Infinite products of groups have the useful property that if $G$ is a fixed group and $f_n : G \to G_n$ is a sequence of homomorphisms, then there is a well-defined homomorphism $f : G \to \prod_{n\geq 1} G_n$ given by $f(g) = (f_1(g), f_2(g), \ldots)$.

**Lemma 26.** *The infinite product $\prod_{n\geq 1} \pi_1(D_n, w_0)$ is uncountable.*

*Proof.* If each $G_n$ is nontrivial, then $G_n$ contains at least two elements. Therefore the product $\prod_{n\geq 1} G_n$ is uncountable since the Cantor set $\{0, 1\}^{\mathbb{N}} = \prod_{n\geq 1}\{0, 1\}$ can

be injected as a subset. In particular, the Baer–Specker group is uncountable. Since $\pi_1(D_n, w_0) \cong \mathbb{Z}$ for each $n \geq 1$, the infinite product $\prod_{n\geq 1} \pi_1(D_n, w_0)$ is isomorphic to the Baer–Specker group and is therefore uncountable. □

Let $\lambda_n : [0, 1] \to D_n$ be the loop defined as

$$\lambda_n(t) = \begin{cases} w_0, & t \in \{0, 1\}, \\ x_n, & t \in \left(0, \frac{1}{2}\right), \\ y_n, & t = \frac{1}{2}, \\ z_n, & t \in \left(\frac{1}{2}, 1\right). \end{cases}$$

This function is continuous and therefore a loop in $D_n$. In particular, our description of the universal covering of $S$ in the previous section shows that the homotopy class $[\lambda_n]$ is a generator of the cyclic group $\pi_1(D_n, b_0)$.

**Definition 27.** Suppose for each $n \geq 1$ we have a continuous loop $\alpha_n : [0, 1] \to H$ based at $w_0$ with image in $D_n$. The *infinite concatenation* of this sequence of loops is the loop $\alpha_\infty : [0, 1] \to H$ defined as follows: for each $n \geq 1$, the restriction of $\alpha_\infty$ to $[(n-1)/n, n/(n+1)]$ is the path $\alpha_n$ and $\alpha_\infty(1) = w_0$.

**Lemma 28.** *The loop $\alpha_\infty$ is continuous and $[r_n \circ \alpha_\infty] = [\alpha_n]$ for all $n \geq 1$.*

*Proof.* Since each loop $\alpha_n$ is continuous and each concatenation $\alpha_n \cdot \alpha_{n+1}$ is continuous, it is enough to show that $\alpha_\infty$ is continuous at 1. Consider a basic open neighborhood $V_n$ of $\alpha_\infty(1) = w_0$. Since $\alpha_i$ has image in $V_n$ for each $i \geq n$, we have $\alpha_\infty([(n-1)/n, 1]) \subseteq V_n$. In particular, $1 \in ((n-1)/n, 1] \subseteq f^{-1}(V_n)$. This proves that $\alpha_\infty$ is continuous.

Notice that $r_1 \circ \alpha_\infty$ is defined to be $\alpha_1$ on $\left[0, \frac{1}{2}\right]$ and is constant at $w_0$ on $\left[\frac{1}{2}, 1\right]$. If $n \geq 2$, then $r_n \circ \alpha_\infty$ is defined as $\alpha_n$ on $[(n-1)/n, n/(n+1)]$ and is constant at $w_0$ on $[0, (n-1)/n] \cup [n/(n+1), 1]$. Thus for all $n \geq 1$, we have $r_n \circ \alpha_\infty$ is homotopic to $\alpha_n$. □

**Theorem 29.** *The fundamental group $\pi_1(H, w_0)$ is uncountable.*

*Proof.* We have a sequence of homomorphisms $(r_n)_* : \pi_1(H, w_0) \to \pi_1(D_n, w_0)$ induced by the retractions $r_n$. Together, these induce a homomorphism $r : \pi_1(H, w_0) \to \prod_{n\geq 1} \pi_1(D_n, w_0)$ given by

$$r([\alpha]) = ((r_1)_*([\alpha]), (r_2)_*([\alpha]), \dots) = ([r_1 \circ \alpha], [r_2 \circ \alpha], \dots).$$

By Lemma 26, the infinite product $\prod_{n\geq 1} \pi_1(D_n, w_0)$ is uncountable. We claim that $r$ is onto.

Suppose $(g_1, g_2, \dots) \in \prod_{n\geq 1} \pi_1(D_n, w_0)$, where $g_n \in \pi_1(D_n, w_0)$. Since $g_n$ is an element of the infinite cyclic group $\pi_1(D_n, w_0)$ generated by $[\lambda_n]$, we may write $g_n = [\lambda_n]^{m_n}$ for some integer $m_n \in \mathbb{Z}$.

For each $n \geq 1$, define the loop $\alpha_n$ by

$$\alpha_n = \begin{cases} \prod_{i=1}^{m_n} \lambda_n & \text{if } m_n > 0, \\ \text{constant at } w_0 & \text{if } m_n = 0, \\ \prod_{i=1}^{|m_n|} \lambda_n^- & \text{if } m_n < 0. \end{cases}$$

Notice that $\alpha_n$ is defined so that $g_n = [\lambda_n]^{m_n} = [\alpha_n]$. Let $\alpha_\infty : [0, 1] \to H$ be the loop based at $w_0$ which is the infinite concatenation as in Definition 27. By Lemma 28, we have $[r_n \circ \alpha_\infty] = [\alpha_n] = g_n$ for each $n \geq 1$. Therefore, $r([\alpha_\infty]) = (g_1, g_2, \dots)$. This proves that $r$ is onto.

Thus, since $\pi_1(H, b_0)$ surjects onto an uncountable group, it must also be uncountable. □

We conclude that there is a $T_0$ space with countably many points but which has an uncountable fundamental group.

## References


[Alexandroff 1937] P. Alexandroff, "Diskrete Räume", *Rec. Math. Moscou* (*N.S.*) **2**:3 (1937), 501–519. Zbl

[Arenas 1999] F. G. Arenas, "Alexandroff spaces", *Acta Math. Univ. Comenian.* (*N.S.*) **68**:1 (1999), 17–25. MR Zbl

[Barmak 2011] J. A. Barmak, *Algebraic topology of finite topological spaces and applications*, Lecture Notes in Mathematics **2032**, Springer, 2011. MR Zbl

[Barmak and Minian 2008] J. A. Barmak and E. G. Minian, "Simple homotopy types and finite spaces", *Adv. Math.* **218**:1 (2008), 87–104. MR Zbl

[Cannon and Conner 2000] J. W. Cannon and G. R. Conner, "The combinatorial structure of the Hawaiian earring group", *Topology Appl.* **106**:3 (2000), 225–271. MR Zbl

[Cianci and Ottina 2016] N. Cianci and M. Ottina, "Smallest homotopically trivial non-contractible spaces", preprint, 2016. arXiv

[Eda 2010] K. Eda, "Homotopy types of one-dimensional Peano continua", *Fund. Math.* **209**:1 (2010), 27–42. MR Zbl

[May 2003a] J. P. May, "Finite groups and finite spaces", notes for REU, University of Chicago, 2003, available at http://www.math.uchicago.edu/~may/MISC/finitegroups.pdf.

[May 2003b] J. P. May, "Finite spaces and simplicial complexes", notes for REU, University of Chicago, 2003, available at http://www.math.uchicago.edu/~may/MISC/SimpCxes.pdf.

[May 2003c] J. P. May, "Finite topological spaces", notes for REU, University of Chicago, 2003, available at http://www.math.uchicago.edu/~may/MISC/FiniteSpaces.pdf.

[McCord 1966] M. C. McCord, "Singular homology groups and homotopy groups of finite topological spaces", *Duke Math. J.* **33** (1966), 465–474. MR Zbl

[Munkres 2000] J. R. Munkres, *Topology*, 2nd ed., Prentice Hall, Upper Saddle River, NJ, 2000. MR Zbl

[Sierpinski 1918] W. Sierpinski, "Un théorème sur les continus", *Tôhoku Math. J.* **13** (1918), 300–305. Zbl

[Spanier 1966] E. H. Spanier, *Algebraic topology*, McGraw-Hill, New York, 1966. MR Zbl

[Stong 1966] R. E. Stong, "Finite topological spaces", *Trans. Amer. Math. Soc.* **123** (1966), 325–340. MR Zbl

Received: 2017-04-08 Revised: 2018-06-12 Accepted: 2018-09-09

jbrazas@wcupa.edu *Department of Mathematics, West Chester University, West Chester, PA, United States*

93luismatos@gmail.com *School of Mathematics, Georgia Institute of Technology, Atlanta, GA, United States*

INVOLVE 12:3 (2019)
dx.doi.org/10.2140/involve.2019.12.395

# Toeplitz subshifts with trivial centralizers and positive entropy

Kostya Medynets and James P. Talisse

(Communicated by David Royal Larson)

Given a dynamical system $(X, G)$, the centralizer $C(G)$ denotes the group of all homeomorphisms of $X$ which commute with the action of $G$. This group is sometimes called the automorphism group of the dynamical system $(X, G)$. We generalize the construction of Bułatek and Kwiatkowski (1992) to $\mathbb{Z}^d$-Toeplitz systems by identifying a class of $\mathbb{Z}^d$-Toeplitz systems that have trivial centralizers. We show that this class of $\mathbb{Z}^d$-Toeplitz systems with trivial centralizers contains systems with positive topological entropy.

## 1. Introduction

Toeplitz dynamical systems were first introduced by Jacobs and Keane [1969]. They provided a classical definition for a Toeplitz sequence over $\{0, 1\}$. Markley [1975] studied these sequences and showed the equivalence of various definitions of them. The orbit closure of a Toeplitz sequence is regarded as a Toeplitz flow. Markley and Paul [1979] showed that these flows were exactly almost one-to-one extensions of odometers, or the group of $p$-adic integers. See [Hewitt and Ross 1979] for a general discussion of the group-theoretic properties of the group of $p$-adic integers. For a general survey of symbolic dynamics, we refer the reader to [Kitchens 1998]. For a good survey on $\mathbb{Z}$-odometers and Toeplitz flows, the reader is referred to [Downarowicz 2005]. Recently the definition of Toeplitz flows was extended to flows over $\mathbb{Z}^d$ by Cortez [2006], and then to flows over general groups in [Cortez and Petite 2008; Krieger 2010].

The centralizer of a dynamical system is the group of all homeomorphisms of the system which commute with the group action. Sometimes called the *automorphism group* of the dynamical system in the literature, the centralizer of a dynamical system has an intricate relationship with its parent dynamical system. For example,

*MSC2010:* 37B05, 37B40, 37B50.

*Keywords:* topological dynamics, symbolic dynamics, automorphism group, centralizer, topological entropy.

The research of Medynets was supported by NSA Young Investigator Grant H98230258656.

in [Boyle et al. 1988], Boyle, Lind and Rudolph studied the centralizers of shifts of finite type and showed that they are countable, residually finite and contain the free group on two generators. Several results have been shown by Cyr and Kra [2015; 2016a; 2016b] which relate varying levels of complexity of symbolic dynamical systems to algebraic properties of their centralizers. We notice that systems with positive entropy tend to have very large centralizers. For example, the centralizer of the full shift contains every finite group and the free group on two generators. On the other hand, Donoso, Durand and Petite [Donoso et al. 2016] showed that some classes of low complexity symbolic dynamical systems have very small centralizers, in the sense that they consist only of powers of $T$. Bułatek and Kwiatkowski [1990; 1992] studied the centralizer of a class of high-complexity Toeplitz systems. The centralizer of multidimensional symbolic dynamical systems was studied by Hochman [2010]. For example, he showed that the centralizer of a positive-entropy multidimensional shift of finite type contains a copy of every finite group.

The main question this paper seeks to answer is whether there are multidimensional systems with a trivial centralizer and positive entropy. Following the ideas of [Bułatek and Kwiatkowski 1992], which developed this result in one dimension, we establish this result with a constructive proof. We note that there are several constructions of *one-dimensional* Toeplitz systems with trivial centralizers and positive entropy; see, for example, [Donoso et al. 2017].

In Section 3 we present main facts with proofs regarding general $G$-odometers, where $G$ is a residually finite group. For the reader's convenience, we include the proofs, otherwise scattered across multiple sources. In particular we show that the centralizer group of $\mathbb{Z}^d$-Toeplitz systems embeds into the centralizer group of its maximal equicontinuous factor, which is a $\mathbb{Z}^d$-odometer, and so is Abelian. This result was originally established in [Auslander 1963, Theorem 9] using the techniques of enveloping semigroups. The proof we present in this note follows the approach developed in [Olli 2013].

In Section 4, we construct a class of $\mathbb{Z}^d$-Toeplitz systems that have trivial centralizers. Then in Section 5, we show that this class contains systems of positive entropy, and we provide an explicit construction of a two-dimensional Toeplitz system of positive entropy.

## 2. Definitions and background

By a *dynamical system* we mean a pair $(X, G)$, where $X$ is a compact topological space and $G$ is a countable discrete group acting on $X$ by homeomorphisms. The action of a group element $g \in G$ on $x \in X$ will be denoted by $g \cdot x = g(x)$. The set $\{g \cdot x \mid g \in G\}$ is called the orbit of the point $x$. If every orbit of $(X, G)$ is dense, we call the system *minimal*. A system $(X, G)$ is called *equicontinuous* if for all $\varepsilon > 0$

there exists $\delta > 0$ such that for all $x, y \in X$ if $d(x, y) < \varepsilon$, then $d(g \cdot x, g \cdot y) < \delta$ for all $g \in G$. Let $(X, G)$, and $(Y, G)$ be two minimal systems. If there exists a continuous surjection $\pi : X \to Y$ which preserves the action of $G$, we say that $X$ is an *extension* of $Y$, and that $Y$ is a *factor* of $X$. We call $\pi$ a *factor map*. Given two factor maps $\pi$ and $\pi'$, we say that $\pi$ is *larger* than $\pi'$ if there exists a third factor map $\pi''$ such that $\pi' = \pi'' \circ \pi$. As such, we can discuss the *maximal* factor of a system. It is a known fact that every dynamical system has a maximal equicontinuous factor.

In this paper we are interested in symbolic dynamical systems. We start with a finite set $\Sigma$ called the alphabet. Say $|\Sigma| = n$. The set of all bi-infinite sequences over $\Sigma$ is called the full $n$-shift and is denoted by $\Sigma^{\mathbb{Z}}$. In general, we denote the full $d$-dimensional $n$-shift by $\Sigma^{\mathbb{Z}^d}$. This set is endowed with the product topology from the discrete topology in each coordinate. Cylinder sets in which we fix a finite number of coordinates form a basis for the topology. For $x \in \Sigma^{\mathbb{Z}^d}$ we write $x = \{x(v)\}_{v \in \mathbb{Z}^d}$. We call $x$ a $\mathbb{Z}^d$-array. The group $\mathbb{Z}^d$ acts on $\Sigma^{\mathbb{Z}^d}$, denoted by $T^z(x)$ for $z \in \mathbb{Z}^d$ and $x \in \Sigma^{\mathbb{Z}^d}$ as follows: $T^z(x) = \{x(z+v)\}_{v \in \mathbb{Z}^d}$. The orbit of an array is $\{T^v(x) \mid v \in \mathbb{Z}^d\}$. A closed subset $X \subseteq \Sigma^{\mathbb{Z}^d}$ is called a *subshift* if it is closed under the action of $\mathbb{Z}^d$.

For the sake of completeness, we note that symbolic dynamics can be studied over general, discrete groups. In this case, let $G$ be a discrete group. Then $\Sigma^G$ is acted on by the group $G$. While in this paper we restrict our study of symbolic dynamics to $\mathbb{Z}^d$-systems, we note that many of the results can be extended to $G$-systems for more general groups $G$.

The topological spaces discussed in this note will be topological zero-dimensional compact metric spaces without isolated points, i.e., Cantor sets. Notice that by a theorem of Brouwer [1910] every Cantor set is homeomorphic to the middle-thirds Cantor set, and so all Cantor sets are homeomorphic.

## 3. Odometers

In this section, we will recall some basic facts about odometers and their almost one-to-one extensions. In particular, we show that the centralizer of an odometer is Abelian, and the centralizer of the almost one-to-one extension of an odometer is also Abelian. These results are mostly known, but are scattered. In particular, the proof of Lemma 3.11 appears in [Veech 1970] and the proof of Proposition 3.12 appears in [Olli 2013]. We present slightly modified proofs for clarity and the reader's convenience.

**Definition 3.1.** A group $G$ is called *residually finite* if the intersection of all its finite-index normal subgroups is trivial.

**Definition 3.2.** Let $G$ be a residually finite group and $G = G_0 \supseteq G_1 \supseteq G_2 \supseteq \cdots$ be nested normal subgroups such that $\bigcap G_n = \{0\}$. Let $\pi_n$ be the natural homomorphism from $G/G_n$ onto $G/G_{n-1}$; i.e., $\pi_n(hG_n) = hG_{n-1}$ for $h \in G$. The

*G-odometer*, $\overline{G}$, is the inverse limit

$$\overline{G} = \varprojlim(G/G_i; \pi_i) = \left\{(g_k)_{k=0}^{\infty} \in \prod_{k=0}^{\infty} G/G_k \;\middle|\; \pi_n(g_n) = g_{n-1} \text{ for all } n \geq 1\right\}.$$

An element $g \in G$ acts on an element $y = (y_i)_{i=0}^{\infty} \in \overline{G}$ as $g \cdot y = (g \cdot y_i)_{i=0}^{\infty}$.
First we prove that $G$ embeds into $\overline{G}$.

**Lemma 3.3.** *Let $\varphi : G \to \overline{G}$ be defined as $g \mapsto (gG_1, gG_2, \ldots)$. Then $\varphi$ is an embedding.*

*Proof.* Notice that $\varphi$ is a homomorphism. Let $g_1, g_2 \in G$. Suppose

$$\varphi(g_1) = (g_1G_1, g_1G_2, g_1G_3, \ldots) = (g_2G_1, g_2G_2, g_2G_3, \ldots) = \varphi(g_2).$$

So $g_1G_i = g_2G_i$ for all $i$. Therefore $g_1^{-1}g_2 \in G_i$ for all $i$, and so $g_1^{-1}g_2 \in \bigcap G_i = \{0\}$. Thus $g_1 = g_2$, which implies that $\varphi$ is an embedding. □

So we have shown that $G$ embeds into $\overline{G}$ in a natural way. In what follows, we will identify the group $G$ with its image $\varphi(G)$. We now prove that $(\overline{G}, G)$ is minimal.

**Lemma 3.4.** *The system $(\overline{G}, G)$ is minimal.*

*Proof.* Consider the identity element, $e \in \overline{G}$. In particular, $e = (G_1, G_2, G_3, \ldots)$. Let $y = (y_i)_{i=0}^{\infty} \in \overline{G}$. So, for each $n$, we have $y_n = \bar{y}_n G_n$, where $\bar{y}_n \in G$ is a representative of the coset. Note

$$\begin{aligned}\bar{y}_n \cdot e &= \bar{y}_n(G_1, G_2, G_3, \ldots, G_n, \ldots)\\ &= (\bar{y}_nG_1, \bar{y}_nG_2, \bar{y}_nG_3, \ldots, \bar{y}_nG_n, \ldots)\\ &= (\bar{y}_1G_1, \bar{y}_2G_2, \ldots, \bar{y}_nG_n, \ldots) = (y_1, y_2, \ldots, y_n, \ldots).\end{aligned}$$

So $\bar{y}_n \cdot e$ agrees with $y$ in the first $n$ coordinates. And so we can get arbitrarily close to $y$ as we increase $n$. It follows that the orbit of $e$ is dense.

Now let $a, b \in \overline{G}$. Note we can find a sequence $b_n$ of elements of $G \subset \overline{G}$ such that $b_n \cdot e \to ab^{-1}$, since $e$ has a dense orbit. Then $(b_n \cdot e) \cdot b \to a$ so $b_n \cdot b \to a$. Therefore $b$ has a dense orbit. □

**Definition 3.5** (centralizer). Let $(X, G)$ be a dynamical system. The *centralizer*, $C(G)$, is defined as

$$C(G) = \{\varphi \in \text{Homeo}(X) \mid g\varphi = \varphi g \text{ for all } g \in G\}.$$

That is, the centralizer of a system consists of all homeomorphisms of the system which commute with the group action. It can be checked that this is a group under composition.

Next we show that elements of the centralizer of an odometer act as translations of the odometer.

**Lemma 3.6.** *Let $\varphi \in C(\overline{G}, G)$. There exists $g_0 \in \overline{G}$ such that $\varphi(x) = x \cdot g_0$ for all $x \in \overline{G}$.*

*Proof.* Set $g_0 = \varphi(e)$. Let $x \in \overline{G}$. Since the orbit of $e$ is dense, by Lemma 3.4, there exists a sequence $\{g_n\} \subseteq G$ such that $g_n \cdot e \to x$. Since $\varphi$ is continuous, $\varphi(g_n \cdot e) \to \varphi(x)$. Since $g_n \cdot e \to x$, we have $g_n \to x$. So $\varphi(g_n \cdot e) = \varphi(g_n) \cdot \varphi(e) \to x \cdot \varphi(e)$. Therefore $\varphi(x) = x \cdot \varphi(e) = x \cdot g_0$. □

We are now ready to prove the following proposition. In the following, $G$ is an Abelian group.

**Proposition 3.7.** *The centralizer $C(\overline{G}, G) = \{\varphi : \overline{G} \to \overline{G} \mid \varphi g = g\varphi \text{ for all } g \in G\}$ of an odometer $\overline{G}$ is isomorphic to $\overline{G}$.*

*Proof.* Define $\psi : C(\overline{G}, G) \to \overline{G}$ as $\psi(\varphi) = \varphi(e)$ for all $\varphi \in C(\overline{G}, G)$. Let $\varphi_1, \varphi_2 \in C(\overline{G}, G)$. Then

$$\psi(\varphi_1 \circ \varphi_2) = \varphi_1 \circ \varphi_2(e) = \varphi_1(\varphi_2(e)) = \varphi_2(e) \cdot \varphi_1(e) = \varphi_1(e) \cdot \varphi_2(e) = \psi(\varphi_1)\psi(\varphi_2).$$

So $\psi$ is a homomorphism. Let $y \in \overline{G}$. Let $\varphi_y(x) = x \cdot y$ for all $x \in \overline{G}$. Note, for $g \in \overline{G}$, we have $\varphi_y(gx) = g\varphi_y(x)$ so $\varphi_y \in C(\overline{G}, G)$. Also, $\psi(\varphi_y) = y$, so $\psi$ is onto. Suppose $\psi(\varphi_1) = \psi(\varphi_2)$. Then $\varphi_1(e) = \varphi_2(e)$. Using Lemma 3.6, we get that for any $x \in \overline{G}$, $\varphi_1(x) = x \cdot \varphi_1(e) = x \cdot \varphi_2(e) = \varphi_2(x)$. Therefore $\psi$ is an isomorphism. □

We now turn our attention to almost one-to-one extensions of odometers.

**Definition 3.8.** We say $(X, G)$ is an *almost one-to-one extension* of $(Y, G)$ if there is a factor map $\pi : X \to Y$ such that there is at least one $y \in Y$ so that $\pi^{-1}y$ is singleton. Almost one-to-one extensions of odometers are also called *Toeplitz systems*.

We make use of the following commutative diagram:

$$\begin{array}{ccc} X & \xrightarrow{G} & X \\ {\scriptstyle\pi}\downarrow & & \downarrow{\scriptstyle\pi} \\ Y & \xrightarrow[G]{} & Y \end{array}$$

Sometimes the context will deem the action of $G$ on $X$ or $Y$ ambiguous, so we will use $T^g x$ to denote the action of the group element $g \in G$ on $x \in X$ and $S^g y$ to denote the action of $g$ on $y \in Y$. In particular, $\pi \circ T^g = S^g \circ \pi$. If the context is clear, the action of $g$ on a point $x$ will be denoted by $g \cdot x$.

If $(X, G)$ is a minimal almost one-to-one extension of a minimal equicontinuous system $(Y, G)$, then it is known that $(Y, G)$ is the maximal equicontinuous factor

of $(X, G)$ [Auslander 1988]. As such, the odometer of which a Toeplitz system $(X, G)$ is an almost one-to-one extension is its maximal equicontinuous factor.

We will be considering almost one-to-one extensions of $\mathbb{Z}^d$-odometers. In this context, we will need the following proposition.

**Proposition 3.9.** *The centralizer $C(G)$ of the almost one-to-one extension of a $\mathbb{Z}^d$-odometer is Abelian.*

To prove Proposition 3.9, we show that the centralizer of the almost one-to-one extension of an odometer embeds into the centralizer of its maximal equicontinuous factor, which we have already shown to be isomorphic to the odometer, which is Abelian in the case of $G = \mathbb{Z}^d$.

**Definition 3.10** [Veech 1970]. Given a dynamical system $(X, G)$ and a metric $d$ compatible with the topology on $X$, two points $x_1, x_2 \in X$ are called *proximal* if

$$\inf_{g\in G} d(g \cdot x_1, g \cdot x_2) = 0.$$

**Lemma 3.11.** *Let $(X, G)$ be an almost one-to-one extension of an odometer $(\overline{G}, G)$ via the factor map $\pi$. Then points of $X$ are proximal if and only if they are in the same $\pi$-fiber.*

*Proof.* Let $x_1, x_2 \in X$ be in the same $\pi$-fiber; i.e., $\pi(x_1) = \pi(x_2)$. Let $y \in \overline{G}$ be such that $\pi^{-1}y$ is a singleton. Since $(\overline{G}, G)$ is minimal, there exists a sequence $\{g_n\}$ of elements in $G$ such that $\lim_{n\to\infty} S^{g_n}\pi x_1 = y$ and so $\lim_{n\to\infty} S^{g_n}\pi x_2 = y$. Since $X$ is compact, there is a subsequence $\{g_k\}$ of $\{g_n\}$ such that $T^{g_k}x_1$ and $T^{g_k}x_2$ converge. Suppose $\lim_{k\to\infty} T^{g_k}x_1 = z$. Applying $\pi$, we have

$$\pi z = \lim_{k\to\infty} \pi T^{g_k}x_1 = \lim_{k\to\infty} S^{g_k}\pi x_1 = y.$$

So we also have

$$\lim_{k\to\infty} \pi T^{g_k}x_2 = \lim_{k\to\infty} S^{g_k}\pi x_2 = y.$$

Since $\pi^{-1}y$ is a singleton, we get that $\lim_{k\to\infty} T^{g_k}x_2 = z$. Now,

$$\begin{aligned}\limsup_{k\to\infty} d(T^{g_k}x_1, T^{g_k}x_2) &\le \limsup_{k\to\infty}(d(T^{g_k}x_1, z) + d(z, T^{g_k}x_2)) \\ &\le \limsup_{k\to\infty} d(T^{g_k}x_1, z) + \limsup_{k\to\infty} d(z, T^{g_k}x_2) = 0.\end{aligned}$$

So the points $x_1$ and $x_2$ are proximal.

Now suppose $x_1, x_2 \in X$ are proximal. Then there is a sequence $\{g_n\} \subseteq G$ such that $\lim_{n\to\infty} T^{g_n}x_1 = \lim_{n\to\infty} T^{g_n}x_2 = z$. Applying $\pi$, we have $\lim_{n\to\infty} \pi T^{g_n}x_1 = \lim_{n\to\infty} \pi T^{g_n}x_2 = \pi z$. So $\lim_{n\to\infty} S^{g_n}\pi x_1 = \lim_{n\to\infty} S^{g_n}\pi x_2$, which implies $\pi x_1$ and $\pi x_2$ are proximal in $\overline{G}$. But $\overline{G}$ has no proximal points, so $\pi x_1 = \pi x_2$. □

Finally, we prove that the centralizers of Toeplitz systems embed in the centralizers of the underlying odometers.

**Proposition 3.12.** *Let $(X, G)$ be an almost one-to-one extension of a $G$-odometer $(Y, G)$. Every element $\varphi \in C(X, G)$ determines $\psi_\varphi \in C(Y, G)$ such that the following diagram commutes*:

$$\begin{array}{ccc} X & \xrightarrow{\varphi} & X \\ {\scriptstyle\pi}\downarrow & & \downarrow{\scriptstyle\pi} \\ Y & \xrightarrow[\psi_\varphi]{} & Y \end{array}$$

*Additionally, this relationship is an embedding; i.e., $\psi_{\varphi_1} = \psi_{\varphi_2} \Rightarrow \varphi_1 = \varphi_2$.*

*Proof.* Let $\varphi \in C(X, G)$. Let $x_1, x_2 \in X$ be proximal. So $\pi x_1 = \pi x_2$. Since $x_1$ and $x_2$ are proximal, $\inf_{g\in G} d(g \cdot x_1, g \cdot x_2) = 0$. Thus $\inf_{g\in G} d(\varphi(g \cdot x_1), \varphi(g \cdot x_2)) = 0$, which, by Lemma 3.11, implies that $\varphi(x_1), \varphi(x_2)$ are proximal. So $\varphi$ preserves the proximal relationship, and so it preserves the $\pi$-fibers. Define $\psi_\varphi : Y \to Y$ as $\psi_\varphi = \pi \circ \varphi \circ \pi^{-1}$. This map is well-defined because $\varphi$ preserves the $\pi$-fibers. Suppose $\psi_\varphi(y_1) = \psi_\varphi(y_2)$ for $y_1, y_2 \in Y$. So $\pi \circ \varphi \circ \pi^{-1}(y_1) = \pi \circ \varphi \circ \pi^{-1}(y_2)$, and thus $\varphi \circ \pi^{-1}(y_1)$ and $\varphi \circ \pi^{-1}(y_2)$ are in the same $\pi$-fibers. Since $\varphi$ preserves the $\pi$-fibers, $\pi^{-1}(y_1)$ and $\pi^{-1}(y_2)$ are in the same $\pi$-fibers, and so it is clear that $y_1 = y_2$. Therefore $\psi_\varphi$ is one-to-one. Also, $\psi_\varphi$ is continuous, so it is a homeomorphism; i.e., $\psi_\varphi \in C(Y, G)$.

Now suppose $\psi_{\varphi_1} = \psi_{\varphi_2}$. Let $y \in Y$ be such that $\pi^{-1}y = \{x\}$ is a singleton. Then $\varphi_1(x) = \pi^{-1}(\psi_{\varphi_1}(y))$ and $\varphi_2(x) = \pi^{-1}(\psi_{\varphi_2}(y))$. Since $\varphi_i$ preserves $\pi$-fibers for $i \in \{1, 2\}$, these are singletons. In particular, $\varphi_1(x) = \varphi_2(x)$. So it is clear then that $g \cdot \varphi_1(x) = g \cdot \varphi_2(x)$ for all $g \in G$, and so $\varphi_1(g \cdot x) = \varphi_2(g \cdot x)$ for all $g \in G$. But every orbit is dense, so $\varphi_1$ and $\varphi_2$ agree on a dense subset of $X$, and hence agree everywhere. □

Finally we prove Proposition 3.9.

*Proof.* We have shown in Proposition 3.12 that $C(X, G)$ embeds into $C(Y, G)$ and by Proposition 3.7 $C(Y, G)$ is Abelian, so $C(X, G)$ is Abelian. □

## 4. $\mathbb{Z}^d$-Toeplitz systems

In this section, we study Toeplitz systems over $\mathbb{Z}^d$ and generalize the construction of Bułatek and Kwiatkowski. In particular, we present a class of Toeplitz systems over $\mathbb{Z}^d$ with a trivial centralizer and positive entropy.

Let $x \in \Sigma^{\mathbb{Z}^d}$. Note that the topological closure of the orbit of $x$, $\overline{O(x)}$, is closed and $T$-invariant. So $(\overline{O(x)}, T)$ is a subshift. This is called the *orbit closure* of $x$.

**Definition 4.1.** The centralizer of a symbolic dynamical system is called *trivial* if every element of the centralizer is $T^g$ for some $g \in \mathbb{Z}^d$.

For $x \in \Sigma^{\mathbb{Z}^d}$, $\sigma \in \Sigma$, and a subgroup $Z \subset \mathbb{Z}^d$, define

$$\mathrm{Per}(x, Z, \sigma) = \{w \in \mathbb{Z}^d \mid x(w+z) = \sigma \text{ for all } z \in Z\},$$
$$\mathrm{Per}(x, Z) = \bigcup_{\sigma \in \Sigma} \mathrm{Per}(x, Z, \sigma).$$

We say that $x \in \Sigma^{\mathbb{Z}^d}$ is a *Toeplitz array* if for every $v \in \mathbb{Z}^d$, there exists a finite-index subgroup $Z \subseteq \mathbb{Z}^d$ (note that $Z$ is necessarily isomorphic to $\mathbb{Z}^d$) such that $v \in \mathrm{Per}(x, Z)$.

It can be shown that the orbit closure of a Toeplitz array is an almost one-to-one extension of a $\mathbb{Z}^d$-odometer. For details, the reader is referred to Theorem 7 and Proposition 21 in [Cortez 2006]. In fact, almost one-to-one extensions of odometers are exactly those systems which are orbit closures of Toeplitz arrays. In particular, defining a Toeplitz system as the orbit closure of a Toeplitz array is equivalent to Definition 3.8.

**Definition 4.2.** Given a finite alphabet $\Sigma$, a *patch* is a pair $(P, \mathcal{L})$, where $P \subseteq \mathbb{Z}^d$ and $\mathcal{L} : P \to \Sigma$ is a labeling of $P$. For the purposes of this paper, we will only consider rectangular patches (blocks) which can be defined by $d$ vectors parallel to the coordinate axes.

Given a patch $(P, \mathcal{L})$, we denote the coordinate closest to the origin in Cartesian space by $P[0]$. Any other location in the patch is denoted by $P[i]$, where $i \in \mathbb{Z}^d$ is a vector pointing to that location, as referenced from $P[0]$. A square block within $P$ is denoted by $P[i-l, i+k]$, where $k, l \in \mathbb{Z}$ and is the (hyper)cube in $P$ located between $P[i - l\bar{1}]$ and $P[i + k\bar{1}]$, where $\bar{1} = (1, 1, \ldots, 1)$.

For a finite block $D$ in $d$ dimensions, we denote the size of $D$ along the $i$-th dimension as $|D|_i$. Note that the left-most and bottom-most entry of $D$ is identified with $D(0, 0, \ldots, 0)$.

We now show how Toeplitz arrays can be constructed over an alphabet $\Sigma$ borrowing ideas from [Downarowicz 2005].

Let $\{p_{t,i}\}_{t=0}^{\infty}$, $1 \le i \le d$, be $d$ sequences of positive integers such that $p_{0,i} \ge 2$ and $p_{t,i}$ divides $p_{t+1,i}$ for all $1 \le i \le d$. Define $\lambda_{t+1,i} = p_{t+1,i}/p_{t,i}$ and $\lambda_{0,i} = p_{0,i}$ for all $1 \le i \le d$ and $t \ge 0$.

Specify blocks $A_t$ as follows:

(1) $|A_t|_i = p_{t,i}$.

(2) Some spaces in $A_t$ are filled with elements from $\Sigma$ and others are left unfilled. The unfilled spaces are called *holes*.

(3) The block $A_{t+1}$ is the concatenation of $\lambda_{t+1,i}$ copies of $A_t$ along the $i$-th dimension for all $1 \le i \le d$, where some holes are filled by symbols from $\Sigma$.

(4) For every $(i_1, i_2, \ldots, i_d) \in \mathbb{N}^d$ there exists a $t \geq 0$ such that $A_t(i_1, i_2, \ldots, i_d) \in \Sigma$ and $A_t(p_{t,1}-i_1, p_{t,2}-i_2, \ldots, p_{t,d}-i_d) \in \Sigma$.

Denote by $\omega_t$ the periodic tiling of $\mathbb{Z}^d$ by the block $A_t$ with the bottom-left corner of $A_t$ appearing at the origin. Set $\omega = \lim \omega_t$. The fourth condition assures that $\omega \in \Sigma^{\mathbb{Z}^d}$. We will additionally assume that $p_t$ is the smallest period of $\omega_t$, which ensures that $\omega$ is a nonperiodic Toeplitz array.

Essentially, in this construction we build finite blocks, each of which contains multiple copies of the block built in the previous step. As we copy these blocks, we fill in some of the holes, and leave some of them as holes. As we continue this process forever, we will have a Toeplitz array covering $\mathbb{Z}^d$.

**Example 4.3** (one-dimensional Toeplitz array [Downarowicz 2005]). We will construct a Toeplitz array over $\mathbb{Z}$ from the alphabet $\Sigma = \{0, 1\}$. Let $\{p_t\} = \{2, 4, 8, 16, \ldots\}$ and so $\lambda_t = 2$ for all $t \geq 0$. Let $A_0 = 0\_$, where the _ symbol indicates a hole. To get $A_1$, we copy $A_0$ twice and fill in some of the holes. Say $A_1 = 0\underline{1}0\_$. The underline indicates a hole that was filled in at that step. In each step we will have two holes. For this construction, at each step we will alternately fill in the first hole with 0 and 1. Let the limiting sequence of this process be $\omega$. Continuing, we have

$$A_2 = 010\underline{0}010\_,$$

$$A_3 = 01000101\underline{0}0100010\_,$$

$$A_4 = 0100010101000100\underline{0}010001010100010\_,$$

$$\vdots$$

$$\omega = 0100010101000100010001010100010101000101010001000100010101000100\ldots,$$

and so we have a Toeplitz array $\omega$. The orbit closure of this point is a Toeplitz system.

**Example 4.4** (two-dimensional Toeplitz array). Again we will use the alphabet $\Sigma = \{0, 1\}$ and we will construct a Toeplitz array over $\mathbb{Z}^2$. Let $\{p_{t,1}\} = \{p_{t,2}\} = \{2, 4, 8, 16, \ldots\}$. Then $\lambda_{t,1} = \lambda_{t,2} = 2$ for all $t \geq 0$.

Let

$$A_0 = \begin{array}{|c|c|}\hline 1 & 1 \\ \hline 0 & \blacksquare \\ \hline \end{array}, \qquad A_1 = \begin{array}{|c|c|c|c|}\hline 1 & 1 & 1 & 1 \\ \hline 0 & \blacksquare & 0 & 0 \\ \hline 1 & 1 & 1 & 1 \\ \hline 0 & 1 & 0 & \blacksquare \\ \hline \end{array}, \qquad A_2 = \begin{array}{|c|c|c|c|c|c|c|c|}\hline 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \hline 0 & \blacksquare & 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \hline 0 & 1 & 0 & \blacksquare & 0 & 1 & 0 & 1 \\ \hline 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & \blacksquare & 0 & 0 \\ \hline 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \hline 0 & 1 & 0 & 0 & 0 & 1 & 0 & \blacksquare \\ \hline \end{array}.$$

The black squares indicate where the holes are. Continuing this process, we will have a coloring of the whole plane, which will be a Toeplitz array, say $\omega$.

We call subblocks of $A_{t+1}$ which coincide with indices of the location of concatenated $A_t$ blocks *t-blocks*. We note that $\omega$ consists of the concatenation of $t$-blocks in all directions for any $t$, where all $t$-blocks agree in all locations except for where the holes were. In Example 4.4, the thick lines in $A_1$ indicate the 0-blocks, and the thick lines in $A_2$ indicate the 1-blocks.

Now we introduce a condition on constructing Toeplitz arrays which will give rise to Toeplitz systems with a trivial centralizer.

**Condition (∗).** We say a Toeplitz array satisfies the condition (∗) if:

- Every $t$-block in $A_{t+1}$ is composed of either $A_t$ where no hole remaining from $A_t$ is filled in or $A_t$ with all holes filled.
- The perimeter of $A_{t+1}$ is composed of $t$-blocks which are all filled in.
- For every $i \in \mathbb{Z}^d$ such that $A_t[i]$ is a hole, there are two $t$-blocks $B_1$ and $B_2$ with $B_1[i] \neq B_2[i]$.

Let $e_1, e_2, \ldots, e_d$ be the generators of $\mathbb{Z}^d$. For $1 \leq i \leq d$, let $T_i$ denote a shift by the vector $e_i$. In this context, the shift action on the system can be considered $d$ independent shift actions; i.e., $T^g = T^{(g_1,g_2,\ldots,g_d)} = T_1^{g_1} \times T_2^{g_2} \times \cdots \times T_d^{g_d}$.

**Theorem 4.5.** *Let $\omega$ be a Toeplitz array satisfying the condition (∗). Then the centralizer $C(T)$ of $(\overline{O(\omega)}, T)$ is trivial.*

*Proof.* Let $(\bar{G}, T_1 \times T_2 \times \cdots \times T_d)$ be the maximal equicontinuous factor of $(\overline{O(\omega)}, T)$. Denote by $\pi : (\overline{O(\omega)}, T) \to (\bar{G}, T_1 \times T_2 \times \cdots \times T_d)$ the almost one-to-one factor map. Let $S \in C(T)$. By Proposition 3.12, this determines an element $S' \in C(\bar{G}, T_1 \times T_2 \times \cdots \times T_d)$ which acts as a translation by some element $h \in \bar{G}$, by Lemma 3.6. By a result of [Hedlund 1969], we note $S$ is determined by a block code $f$ of window size $k \in \mathbb{N}$. In particular, if $u \in \overline{O(\omega)}$ and $z = S(u)$, then

$$z[i] = f(u[i-k, i+k]) \quad \text{for all } i \in \mathbb{Z}^d. \tag{1}$$

In particular, the automorphism determines what to put in a specific location by looking at a block around that location in the preimage. Increasing $k$ if necessary, we can assume that $S^{-1}$ is also determined by a block code of the same window size $k$.

Note $\bar{G}$ is a product odometer, so $h = (h_1, h_2, \ldots, h_d)$, where $h_i = \sum_{t=0}^{\infty} h_{t,i}\, p_{t-1,i}$ for $1 \leq i \leq d$ with $0 \leq h_{t,i} \leq \lambda_{t,i} - 1$, $p_{-1,i} = 1$. Each $h_i$ is an element of the one-dimensional odometer occurring in the $i$-th coordinate of $h$. Let $m_{t,i} = \sum_{j=0}^{t} h_{j,i}\, p_{j-1,i}$ and $m_t = (m_{t,1}, m_{t,2}, \ldots, m_{t,d}) \in \mathbb{Z}^d$. For each $g \in G$, denote by $X_g$ the preimage $\pi^{-1}(\{g\})$ under the factor map $\pi$. Then $S(X_g) = X_{g+h}$.

We claim that for all $1 \leq i \leq d$ either $m_{t,i} \leq k$ or $m_{t,i} \geq p_{t,i} - k - 1$.

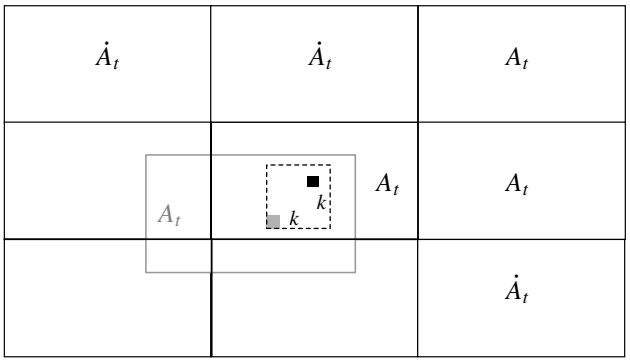


**Figure 1.** The two-dimensional case of the argument in the proof of Theorem 4.5. $\dot{A}_t$ indicates $A_t$ blocks with all holes filled and the solid black and gray squares indicate holes in $x_{t+1}$ and $y_{t+1}$, respectively.

Let $x \in \overline{O(\omega)}$ and $y = S(x)$. Suppose that $x$ has a $t$-block appearing at a location $x[i]$. Then by the construction of Toeplitz subshifts and almost one-to-one extensions, $y$ necessarily has a $t$-block at the location $y[i - m_t]$. Note that for every $t \geq 1$ the array $x$ can be written as a concatenation of $(t+1)$-blocks, which are made of $t$-blocks. Recall that all $t$-blocks are the same, except they may disagree where the holes are located. Denote by $x_{t+1}$ the $d$-dimensional array over the alphabet $\Sigma \cup \{\text{hole}\}$ consisting of copies of the block $A_{t+1}$ at the locations where they appear in $x$. Note that $\lim x_t = x$. Similarly, we can define $y_{t+1}$, a $(t+1)$-block skeleton of the array $y$. Note that $x_{t+1}$ and $y_{t+1}$ are $p_{t+1}$-periodic and shifted by the vector $m_{t+1}$ relative to each other. Let $A$ denote any $(t+1)$-block of $y$. This block looks like $A_{t+1}$, which in turn is the concatenation of $A_t$ blocks both filled and not. Thus, we can view both $y_{t+1}$ and $x_{t+1}$ as the concatenations of copies of $A_t$ and copies of $\dot{A}_t$, filled versions of $A_t$. The $t$-blocks in $x_{t+1}$ and $y_{t+1}$ appear shifted by the vector $m_t$.

Fix $t > 0$ such that $p_{t-1,i} > 2k + 1$ for every $i = 1, \ldots, d$. Fix $j \in \mathbb{Z}^d$ such that $y_{t+1}[j]$ is a hole. Note that this hole would correspond to a hole in $A_t$. Then the hypercube $x_{t+1}[j-k, j+k]$ must also contain a hole. For otherwise, $x_{t+1}[j-k, j+k]$ would be the same for every $j$ with $j \equiv i \bmod p_t$ and, thus, $y[j] = S(x)[j] = f(x[j-k, j+k])$ would be the same for every such $j$. This would contradict the last property of the condition $(*)$. Applying the same argument to $S^{-1}$, we see that if $x_{t+1}[j]$ is a hole for some $j \in \mathbb{Z}^d$, then $y[j-k, j+k]$ must also contain a hole.

This argument is demonstrated for the two-dimensional case and for the forward-looking centralizers in Figure 1.

Now, the $t$-blocks in the arrays $x_{t+1}$ and $y_{t+1}$ are shifted by the vector $m_t$ relative to each other. At the same time, by the argument above, the filled $t$-blocks $\dot{A}_t$ in $x_{t+1}$ and $y_{t+1}$ must appear under each other and can be shifted by a vector of length at

most $k$. Since $p_t$ is the smallest period of $\omega_t$, we conclude that for all $1 \le i \le d$ either $m_{t,i} \le k$ or $m_{t,i} \ge p_{t,i} - k - 1$. It follows that $h_i$ is an integer for every $i = 1, \dots, d$. So, $h = S'(0) = T^h(0)$; i.e., $S'$ and $T^h$ agree on one point. Furthermore, $S'$ agrees with the action of $T^h$ on the entire orbit of 0, which is dense. Therefore, $S' = T^h$.

Let $\alpha$ be in the orbit of $\omega$ in $(\overline{O(\omega)}, T)$; i.e., $\alpha = T^g\omega$ for some $g \in \mathbb{Z}^d$. Note

$$\pi S(\alpha) = \pi S(T^g\omega) = S'\pi(T^g\omega) = S'T^g(0) = T^hT^g(0) = \pi T^hT^g\omega = \pi T^h(\alpha).$$

So $S(\alpha)$ and $T^h(\alpha)$ are in the same $\pi$-fiber. Since $\alpha$ is in the orbit of $\omega$, it has a unique preimage under $\pi$. Therefore $S(\alpha) = T^h(\alpha)$. And so $S$ and $T^h$ agree on the entire orbit of $\omega$, which is dense. So $S = T^h$. □

## 5. Positive-entropy Toeplitz subshift

We now construct an explicit example of a two-dimensional Toeplitz subshift which has positive entropy. This example is constructed so that it obeys the condition $(*)$, thus ensuring that it has a trivial centralizer.

Let $h > 0$ and choose an integer $l_0$ such that $\log(l_0 - 1) \le h \le \log(l_0)$. For $i \ge 0$, let $\varepsilon_i > 0$ and $\{\varepsilon_i\}$ be such that $\sum_{i=0}^{\infty} \varepsilon_i < h/2$.

We note that for any $l$ and any $\varepsilon > 0$, there exists $n \in \mathbb{N}$ sufficiently large such that

$$\frac{\log(l^{n^2})}{(n+2)^2} \ge \log(l) - \varepsilon \tag{2}$$

since $(n/(n+2))^2 \to 1$.

Let $q_0$ be chosen so that

$$\frac{\log(l_0{}^{q_0^2})}{(q_0+2)^2} \ge \log(l_0) - \frac{\varepsilon_0}{2}.$$

Also require $q_0^2 \ge l_0$. Define $l_1 = l_0^{q_0^2}$. We notice that there are $l_0^{q_0^2}$ square blocks of side length $q_0$ over the alphabet $\{0, 1, \dots, l_0 - 1\}$. We enumerate these blocks as $B_i^{(0)}$ for $0 \le i \le l_1 - 1$. Furthermore, we require that $B_0^{(0)}$ and $B_1^{(0)}$ contain every letter from the alphabet. Let $C_i^{(0)}$ be the square block of side length $q_0 + 2$ with the block $B_i^{(0)}$ surrounded by a 0 in the top left corner, a 1 in the bottom right corner, and 0's below the main diagonal and 1's above it, as in the diagram below. We will denote this as $C_i^{(0)} = 0B_i^{(0)}1$ for $0 \le i \le l_1 - 1$:

$$C_i^{(0)} = \begin{array}{|c|c|c|c|}\hline 0 & 1 & \cdots & 1 \\ \hline 0 & \multicolumn{2}{c|}{\multirow{2}{*}{$B_i^{(0)}$}} & \vdots \\ \cline{1-1}\cline{4-4} \vdots & & & 1 \\ \hline 0 & \cdots & 0 & 1 \\ \hline \end{array}.$$

For $k \geq 1$, define $l_k = l_{k-1}^{q_{k-1}^2}$ and let $q_k$ be such that

$$\frac{\log(l_k{}^{q_k^2})}{(q_k+2)^2} \geq \log(l_k) - \frac{\varepsilon_k}{2}. \tag{3}$$

Additionally, require that $q_k^2 \geq l_k$. Let $B_i^{(k)}$ be all the square blocks of side length $q_k$ over the alphabet $\{0, 1, \ldots, l_k - 1\}$ for $0 \leq i \leq l_{k+1} - 1$. Require that $B_0^{(k)}$ and $B_1^{(k)}$ contain every letter from the alphabet. Let $C_i^{(k)} = 0B_i^{(k)}1$ for $0 \leq i \leq l_{k+1} - 1$.

Consider the following operation on finite blocks. Let $\{A_1, A_2, \ldots, A_n\}$ be square blocks of the same side length $|A|$ over some alphabet. Let $B$ be a square block over an alphabet containing $\{1, 2, \ldots, n\}$. We define the block

$$C = \{A_1, A_2, \ldots, A_n\} * B$$

as $C[i, j] = A_{B[i.j]}$. In particular, $C$ will be a square block of side length $|B| \cdot |A|$.

Our goal is to construct a sequence of blocks $\{A_t\}$ that defines a system of $k$-blocks and that satisfies the condition $(*)$. We proceed as follows: Let $A_i^{(0)} = C_i^{(0)}$ and

$$A_i^{(k)} = \{A_0^{(k-1)}, A_1^{(k-1)}, \ldots, A_{l_k-1}^{(k-1)}\} * C_i^{(k)}.$$

We note that since $C_0^{(1)}$ and $C_1^{(1)}$ have every letter of the alphabet $\{0, 1, \ldots, l_1 - 1\}$, the blocks $A_0^{(1)}$ and $A_1^{(1)}$ will have every 0-block as a subblock. Similarly, $C_0^{(2)}$ and $C_1^{(2)}$ contain every letter in $\{0, 1, \ldots, l_2 - 1\}$ and so the blocks $A_0^{(2)}$ and $A_1^{(2)}$ will contain every 1-block as a subblock. In general, we note that each block $A_i^{(k)}$ for $i = 0, 1$ has every $(k-1)$-block as a subblock.

We let

$$A_0 = \begin{array}{|c|c|c|c|}\hline 0 & 1 & \cdots & 1 \\ \hline 0 & \multicolumn{2}{c|}{\multirow{2}{*}{$-$}} & \vdots \\ \cline{1-1}\cline{4-4} \vdots & \multicolumn{2}{c|}{} & 1 \\ \hline 0 & \cdots & 0 & 1 \\ \hline \end{array},$$

where the side length of the square box $A_0$ is $q_0 + 2$, and the dash in the center square indicates a square of side length $q_0$ consisting of all holes. We note that $A_i^{(0)}$, $i = 0, \ldots, l_1 - 1$, are 0-blocks corresponding to $A_0$.

Inductively, define

$$A_{k+1} = \begin{array}{|c|c|c|c|c|}\hline A_0^{(k)} & A_1^{(k)} & \cdots & A_1^{(k)} & A_1^{(k)} \\ \hline A_0^{(k)} & A_k & \cdots & A_k & A_1^{(k)} \\ \hline \vdots & \vdots & \ddots & \vdots & \vdots \\ \hline A_0^{(k)} & A_k & \cdots & A_k & A_1^{(k)} \\ \hline A_0^{(k)} & A_0^{(k)} & \cdots & A_0^{(k)} & A_1^{(k)} \\ \hline \end{array},$$

where there is a square block consisting of $q_k^2$ copies of $A_k$ surrounded by $4q_k+4$ copies of $A_i^{(k)}$ for $i=0$ or 1 on each side positioned similarly to 0's and 1's in $A_0$. Notice that $A_0^{(k)}$ and $A_1^{(k)}$ have no holes, so all the holes are contained in the middle block of $A_k$ blocks. Note that $A_i^{(k)}$, $i=0,\dots,l_k-1$, are the $k$-blocks corresponding to the pattern $A_k$.

Let $\omega$ be the limiting array from the above process. Note that $\omega\in\{0,\dots,l_0-1\}^{\mathbb{Z}^2}$ and $\omega$ satisfies the condition $(*)$.

**Proposition 5.1.** *The Toeplitz system* $(\overline{O(\omega)},T)$ *has positive entropy.*

*Proof.* Define $\lambda_k=q_k+2$ and $p_k=\lambda_1\lambda_2\cdots\lambda_k$. Let $h_\omega$ be the entropy of $(\overline{O(\omega)},T)$ and let $\Theta(n)$ be the number of square blocks of side length $n$ appearing in $\omega$. We note that

$$h_\omega=\lim_{n\to\infty}\frac{\log(\Theta(n))}{n^2}=\lim_{k\to\infty}\frac{\log(\Theta(p_k))}{p_k^2},$$

by switching to a subsequence.

There are $l_{k+1}$ many $k$-blocks. We note that every block $A_k$ contains every $(k-1)$-block as a subblock. This is because the blocks $C_i^{(k)}$ for $i=0$ or $i=1$ contain every letter of the alphabet in them. This means that as we do the shuffling process described above, the blocks $A_i^{(k)}$ for $i=0$ or $i=1$ contain every single block $A_i^{(k-1)}$ for $0\le i\le l_k-1$. Furthermore, since $k$-blocks are squares of side length $p_k$, there are at least as many blocks of side length $p_k$ occurring in $\omega$ as there are $k$-blocks. Specifically, square blocks of length $p_k$ can occur at any position within $\omega$, while $k$-blocks only occur at specific positions. Hence we have

$$\Theta(p_k)\ge l_{k+1}. \tag{4}$$

So we have

$$h_\omega\ge\limsup_{k\to\infty}\frac{\log(l_{k+1})}{p_k^2}. \tag{5}$$

By (3) we have

$$\frac{\log(l_{k+1})}{\lambda_k^2}\ge\log(l_k)-\frac{\varepsilon_k}{2}.$$

It then follows by (2) that

$$\frac{\log(l_{k+1})}{p_k^2}\ge\frac{\lambda_k^2(\log(l_k)-\varepsilon_k/2)}{p_k^2}=\frac{\log(l_k)-\varepsilon_k/2}{p_{k-1}^2}\ge\frac{\log(l_k)}{p_{k-1}^2}-\varepsilon_k.$$

Continuing, we have

$$\frac{\log(l_{k+1})}{p_k^2}\ge h-\sum_{i=0}^{k}\varepsilon_i.$$

Taking the limit as $k\to\infty$, from (5), we have $h_\omega\ge h/2>0$.

It is a basic fact that every Toeplitz system is minimal, so this system is minimal. It is either finite or uncountable, and since it has positive entropy, it cannot be finite. So this is an infinite minimal Toeplitz system. □

## References


[Auslander 1963] J. Auslander, "Endomorphisms of minimal sets", *Duke Math. J.* **30**:4 (1963), 605–614. MR Zbl

[Auslander 1988] J. Auslander, *Minimal flows and their extensions*, North-Holland Mathematics Studies **153**, North-Holland, Amsterdam, 1988. MR Zbl

[Boyle et al. 1988] M. Boyle, D. Lind, and D. Rudolph, "The automorphism group of a shift of finite type", *Trans. Amer. Math. Soc.* **306**:1 (1988), 71–114. MR Zbl

[Brouwer 1910] L. E. J. Brouwer, "On the structure of perfect sets of points", *KNAW, Proc.* **12** (1910), 785–794.

[Bułatek and Kwiatkowski 1990] W. Bułatek and J. Kwiatkowski, "The topological centralizers of Toeplitz flows and their $Z_2$-extensions", *Publ. Mat.* **34**:1 (1990), 45–65. MR Zbl

[Bułatek and Kwiatkowski 1992] W. Bułatek and J. Kwiatkowski, "Strictly ergodic Toeplitz flows with positive entropies and trivial centralizers", *Studia Math.* **103**:2 (1992), 133–142. MR Zbl

[Cortez 2006] M. I. Cortez, "$\mathbb{Z}^d$ Toeplitz arrays", *Discrete Contin. Dyn. Syst.* **15**:3 (2006), 859–881. MR Zbl

[Cortez and Petite 2008] M. I. Cortez and S. Petite, "$G$-odometers and their almost one-to-one extensions", *J. Lond. Math. Soc.* (2) **78**:1 (2008), 1–20. MR Zbl

[Cyr and Kra 2015] V. Cyr and B. Kra, "The automorphism group of a shift of linear growth: beyond transitivity", *Forum Math. Sigma* **3** (2015), art. id. e5. MR Zbl

[Cyr and Kra 2016a] V. Cyr and B. Kra, "The automorphism group of a minimal shift of stretched exponential growth", *J. Mod. Dyn.* **10** (2016), 483–495. MR Zbl

[Cyr and Kra 2016b] V. Cyr and B. Kra, "The automorphism group of a shift of subquadratic growth", *Proc. Amer. Math. Soc.* **144**:2 (2016), 613–621. MR Zbl

[Donoso et al. 2016] S. Donoso, F. Durand, A. Maass, and S. Petite, "On automorphism groups of low complexity subshifts", *Ergodic Theory Dynam. Systems* **36**:1 (2016), 64–95. MR Zbl

[Donoso et al. 2017] S. Donoso, F. Durand, A. Maass, and S. Petite, "On automorphism groups of Toeplitz subshifts", *Discrete Anal.* (2017), art. id. 11. MR

[Downarowicz 2005] T. Downarowicz, "Survey of odometers and Toeplitz flows", pp. 7–37 in *Algebraic and topological dynamics* (Bonn, 2004), edited by S. Kolyada et al., Contemp. Math. **385**, Amer. Math. Soc., Providence, RI, 2005. MR Zbl

[Hedlund 1969] G. A. Hedlund, "Endomorphisms and automorphisms of the shift dynamical system", *Math. Systems Theory* **3** (1969), 320–375. MR Zbl

[Hewitt and Ross 1979] E. Hewitt and K. A. Ross, *Abstract harmonic analysis, I: Structure of topological groups, integration theory, group representations*, 2nd ed., Grundlehren der Mathematischen Wissenschaften **115**, Springer, 1979. MR Zbl

[Hochman 2010] M. Hochman, "On the automorphism groups of multidimensional shifts of finite type", *Ergodic Theory Dynam. Systems* **30**:3 (2010), 809–840. MR Zbl

[Jacobs and Keane 1969] K. Jacobs and M. Keane, "0-1-sequences of Toeplitz type", *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **13** (1969), 123–131. MR Zbl

[Kitchens 1998] B. P. Kitchens, *Symbolic dynamics: one-sided, two-sided and countable state Markov shifts*, Springer, 1998. MR Zbl

[Krieger 2010] F. Krieger, "Toeplitz subshifts and odometers for residually finite groups", pp. 147–161 in *École de Théorie Ergodique* (Marseilles, 2006), edited by Y. Lacroix et al., Sémin. Congr. **20**, Soc. Math. France, Paris, 2010. MR Zbl

[Markley 1975] N. G. Markley, "Substitution-like minimal sets", *Israel J. Math.* **22**:3-4 (1975), 332–353. MR Zbl

[Markley and Paul 1979] N. G. Markley and M. E. Paul, "Almost automorphic symbolic minimal sets without unique ergodicity", *Israel J. Math.* **34**:3 (1979), 259–272. MR Zbl

[Olli 2013] J. Olli, "Endomorphisms of Sturmian systems and the discrete chair substitution tiling system", *Discrete Contin. Dyn. Syst.* **33**:9 (2013), 4173–4186. MR Zbl

[Veech 1970] W. A. Veech, "Point-distal flows", *Amer. J. Math.* **92** (1970), 205–242. MR Zbl

Received: 2017-05-03 Revised: 2017-06-13 Accepted: 2018-06-25

medynets@usna.edu *Mathematics Department, United States Naval Academy, Annapolis, MD, United States*

jptalisse@gmail.com *Mathematics Department, United States Naval Academy, Annapolis, MD, United States*

INVOLVE 12:3 (2019)
dx.doi.org/10.2140/involve.2019.12.411

# Associated primes of $h$-wheels

Corey Brooke, Molly Hoch, Sabrina Lato, Janet Striuli and Bryan Wang

(Communicated by Kenneth S. Berenhaut)

We study the associated primes of the powers of the cover ideal of $h$-wheels. The main result generalizes a theorem of Kesting, Pozzi, and Striuli (2011).

Several pieces of information about an ideal $I$ in a commutative noetherian ring $R$ are enclosed in its primary decomposition: Given an ideal $I$ we can write $I = \bigcap_{i=1}^{\ell} Q_i$, where the radical ideal of each ideal $Q_i$ is given by a prime ideal $P_i$ of the ring $R$. The prime ideals $P_i$ for $i = 1, \ldots, \ell$ are called associated primes of the ideal $I$. The finiteness conditions imposed by a noetherian ring not only allow the decomposition of an ideal into primary components, but also have stronger repercussions, as shown in the following statement proved by Brodmann [1979] in which the set $\operatorname{Ass}(R/I)$ denotes the set of all the associated primes of $I$:

*Let $I$ be an ideal in a commutative noetherian ring*; *then the set*

$$\bigcup_{i=1}^{\infty} \operatorname{Ass}(R/I^i)$$

*is finite. Moreover, there exists an integer $m$ such that for all $k \geq m$ the equality* $\operatorname{Ass}(R/I^m) = \operatorname{Ass}(R/I^k)$ *holds.*

The positive integer $m$ identified by Brodmann's theorem is called the index of stability for the associated primes of $I$, denoted by $\operatorname{astab}(I)$. Despite the simplicity of the statement, the value of $\operatorname{astab}(I)$ remains generally unknown.

Much work has been done recently for graded ideals in polynomial rings. While a large upper bound for $\operatorname{astab}(I)$ for monomial ideals was given in [Hoa 2006] in terms of properties of the ideal itself, a lot of recent work supports the conjecture that in a polynomial ring $\mathsf{k}[x_1, \ldots, x_d]$ the uniform bound $\operatorname{astab}(I) \leq d$ for every graded ideal $I \subseteq \mathsf{k}[x_1, \ldots, x_d]$ holds; see for example [Herzog and Asloob Qureshi 2015, Theorem 4.1] for polymatroid ideals.

More cases for which the conjecture holds true come from ideals that arise from graphs. In this paper, a graph $G$ is given by a set of vertices $V_G = \{x_1, \ldots, x_d\}$ and a set of edges $E_G$; elements of $E_G$ are subsets of $V_G$ of cardinality 2. In particular,

*MSC2010:* primary 13F55, 05C25; secondary 05C38, 05E99.
*Keywords:* graph, polynomial ring, cover ideal, associated primes.

if $\{x_i, x_j\}$ is an edge then we say that $x_i$ and $x_j$ are adjacent vertices. Given such a graph $G$, the *edge ideal* of $G$ is an ideal of the polynomial ring $\mathsf{k}[x_1, \ldots, x_d]$ generated by the monomials $x_i x_j$ such that $\{x_i, x_j\} \in E_G$.

The conjecture is verified for edge ideals. It follows from [Simis et al. 1994, Theorem 5.9] that $\operatorname{astab}(I)$ is equal to 1 for edge ideals of bipartite graphs. In [Chen et al. 2002, Proposition 4.3], the authors show the conjecture, and in fact a stronger statement, holds for edge ideals of nonbipartite graphs.

The authors of [Francisco et al. 2011] look at cover ideals of graphs (in fact the paper deals with the more general notion of a hypergraph). We define the cover ideal later, but in Corollary 4.9 of the paper above, the authors prove that if $J$ is the cover ideal of a simple graph then $\operatorname{astab}(J) \leq \chi(G) - 1$, where $\chi(G)$ is the coloring number of the graph (which is bounded above by the number of vertices of a graph). Further, they fully characterize prime ideals that appear as associated primes of the second power of the cover ideal.

In line with this work, in [Kesting et al. 2011] the authors study which prime ideals appear as associated primes of the third power of the cover ideal. They prove that the *wheel* corresponds to an element of $\operatorname{Ass}(R/J^3)$.

In this paper we generalize the work of [Kesting et al. 2011]. Given an integer $h$, we define the $h$-wheel and prove the following:

**0.1. Theorem.** *Let $G$ be graph with vertex set $V_G = \{x_1, \ldots, x_d\}$ that is an $h$-wheel. Denote by $J_G \subseteq \mathsf{k}[x_1, \ldots, x_d]$ the cover ideal of $G$. Then the prime ideal $(x_1, \ldots, x_d)$ belongs to $\operatorname{Ass}(R/J^n)$ if and only if $n \geq h + 2$.*

As a corollary, for every integer $d \geq 6$ we deliver an ideal $I_d$ in a polynomial ring with $d$ variables such that $\operatorname{astab}(I_d) \geq d - 3$.

## 1. Definitions

We now introduce the notation and give the definitions used in the paper.

**1.1.** Given a graph $G$ with vertex set $V_G = \{x_1, \ldots, x_d\}$, we consider the polynomial ring $\mathsf{k}[x_1, \ldots, x_d]$, which we often denote by $\mathsf{k}[V_G]$. If $S$ is a subset of $V_G$, then the prime monomial ideal $P_S$ is the ideal generated by the variables $x \in S$. If $S = V_G$, then we denote $P_S$ by $\mathfrak{m}_G$, the maximal homogeneous ideal in $\mathsf{k}[V_G]$. It is worth noting that a prime monomial ideal is always generated by a subset of the variables. In this setting, given a monomial $\boldsymbol{m} \in \mathsf{k}[x_1, \ldots, x_d]$ we can write $\boldsymbol{m} = \prod_{i=1}^{d} x_i^{\alpha_i}$, where $\alpha_i \geq 0$. The support of $\boldsymbol{m}$ is the set of variables $\{x_i \mid \alpha_i > 0\}$ and it is denoted as $\operatorname{supp}(\boldsymbol{m})$. We denote by $\operatorname{ver}(\boldsymbol{m})$ the subset of $V_G$ of vertices labeled by the variables appearing in $\operatorname{supp}(\boldsymbol{m})$.

**1.2. Definition.** Given a graph $G$ with vertex set $V_G = \{x_1, \ldots, x_d\}$ and edge set $E_G$, a *cover* of $G$ is a subset $S$ of $V_G$ such that each edge in $E_G$ has a nonempty intersection with $S$.

The cover ideal $J_G \subset \mathsf{k}[x_1, \ldots, x_d]$ is the monomial ideal generated by monomials $\boldsymbol{m}$ such that $\operatorname{ver}(\boldsymbol{m})$ is a cover of $G$.

The following definition is a particular case of the definition of associated prime given in [Eisenbud 1995, page 89].

**1.3. Definition.** Let $I$ be a monomial ideal of the polynomial ring $\mathsf{k}[x_1, \ldots, x_d]$ and let $P = (x_{i_1}, \ldots, x_{i_\ell})$ be a monomial prime ideal containing $I$. We say that $P$ is an associated prime of $I$, and we write $P \in \operatorname{Ass}(R/I)$, if there exists a monomial $\boldsymbol{w} \in \mathsf{k}[x_1, \ldots, x_d]$ such that $\boldsymbol{w} \notin I$, $x_i \boldsymbol{w} \in I$ for $i = i_1, \ldots, i_\ell$, but $x_i \boldsymbol{w} \notin I$ for $i \neq i_1, \ldots, i_\ell$.

The monomial $\boldsymbol{w}$ is called a witness of $P$ for the ideal $I$.

As shown in [Eisenbud 1995, Theorem 3.10], the associated primes of a monomial ideal $I$ defined in the previous definition are exactly the prime ideals that are radical ideals in a minimal primary decomposition of $I$.

Let $G$ be a connected graph with vertex set $\{x_1, \ldots, x_d\}$. The edge ideal and the cover ideal of $G$ are dual to each other with respect to the Alexander duality; see for a proof [Bruns and Herzog 1993, Chapter 5] or consult [Van Tuyl 2013] for a quicker introduction to the subject. This fact implies that a prime ideal $P$ is an associated prime of the cover ideal if and only if $P = (x_i, x_j)$, where $\{x_i, x_j\}$ is in $E_G$.

The following theorem extends the knowledge of associated primes to second powers of the cover ideal [Francisco et al. 2010, Corollary 3.4].

**1.4.** Let $G$ be a connected graph, let $S$ be a subset of the vertex set $V_G$, and let $R = \mathsf{k}[V_G]$. A prime ideal $P_S \subset \mathsf{k}[V_G]$ belongs to $\operatorname{Ass}(R/J_G^2)$ if and only if the induced subgraph generated by $S$ is an odd cycle in $G$ or $S$ is an edge.

We concentrate our attention on a family of graphs called $h$-wheels, whose definition is given below. First we need the following notion:

**1.5.** Let $G$ be a graph with vertex set $V_G$. Given a vertex $x \in V_G$ and a subset $S \subseteq V_G$ of vertices of $G$, we denote by $N_S(x)$ the subset of $S$ consisting of adjacent vertices to $x$. If $S$ is the set of all vertices in $G$ then we use $N(x)$ to denote the set of all vertices adjacent to $x$.

**1.6. Definition.** A graph $G$ with vertex set $V_G$ is an $h$-wheel if $V_G$ can be written as the union of two disjoint sets, the set of rim vertices $R^G$ and the set of center vertices $C^G$, such that the following conditions hold:

(1) The subgraph induced by $C^G$ is the complete graph on $h$ vertices.

(2) The subgraph induced by $R^G$ is an odd cycle.

(3) There exist $x_1, \ldots, x_k \in R^G$ with $k \geq 3$ such that $N_{R^G}(y) = \{x_1, \ldots, x_k\}$ for all $y \in C^G$.

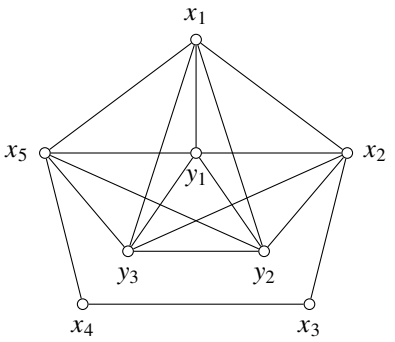


**Figure 1.** A 3-wheel.

(4) For every $y \in C^G$, the vertex $y$ belongs to at least two odd cycles in the subgraph induced by $y$ and $N_{R^G}(y)$.

We call $k$ the radial number for $G$. For each $i = 1, \ldots, k-1$, set $\ell_i$ as the length of the path along the subgraph induced by $R^G$ from $x_i$ to $x_{i+1}$, and set $\ell_k$ as the length from $x_k$ to $x_1$. The positive integers $\ell_1, \ldots, \ell_k$ are called the radial lengths.

In [Kesting et al. 2011], the authors studied the 1-wheel, which we call a wheel for simplicity. Notice that given an $h$-wheel $G$ and a vertex $y \in C^G$, the subgraph induced by $y$ and $R^G$ is a wheel.

**1.7. Example.** Figure 1 is a representation of a 3-wheel $G$. We have

$$C^G = \{y_1, y_2, y_3\}, \quad R^G = \{x_1, x_2, x_3, x_4, x_5\},$$
$$N_{R^G}(y_1) = N_{R^G}(y_2) = N_{R^G}(y_3) = \{x_1, x_2, x_3\}.$$

In the rest of the paper we rely on the following constructions.

**1.8. Definition.** Given a graph $G$ and a vertex $x \in V_G$, the *contraction* of $G$ via $x$ is a new graph obtained from $G$ by deleting $x$ and connecting all the vertices in $N(x)$ to each other.

**1.9. Definition.** Given a graph $G$, let $x_1$ and $x_2$ be two adjacent vertices in $G$. A *subdivision* of $G$ via the edge $\{x_1, x_2\}$ is a graph obtained from $G$ by deleting the edge $\{x_1, x_2\}$, adding a new vertex $y$, and adding two new edges $\{x_1, y\}$ and $\{x_2, y\}$.

## 2. Preliminary lemmas

We now prove several lemmas that are used to prove our main result.

The first lemma describes necessary conditions for a monomial to be a witness for a power of the cover ideal of a graph $G$.

**2.1. Lemma.** *Let $G$ be a graph with vertex set $V_G$, and let $J_G$ be the cover ideal of $G$ in the ring $R = \mathsf{k}[V_G]$. Let $S \subseteq V_G$, and assume that $P_S \in \operatorname{Ass}(R/J_G^n)$. Let $\boldsymbol{w}$ be a witness for $P_S$. Then $x^n$ does not divide $\boldsymbol{w}$ for any $x \in S$.*

*Proof.* By the definition of witness, $\boldsymbol{w} \notin J_G^n$.

Suppose toward contradiction that there exists $x \in S$ such that $x^n$ divides $\boldsymbol{w}$. Since the monomial $x\boldsymbol{w}$ is in $J_G^n$, there exist $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_n \in J_G$ such that $x\boldsymbol{w} = \boldsymbol{m}_1 \cdots \boldsymbol{m}_n$. Moreover, since $x^n \mid \boldsymbol{w}$, by the pigeonhole principle we know that there exists an integer $s$ such that $1 \leq s \leq n$ and $x^2$ divides $\boldsymbol{m}_s$. Let $\boldsymbol{m}_s'$ be the monomial $\boldsymbol{m}_s/x$. Since $\boldsymbol{m}_s \in J_G$, it follows that $\operatorname{ver}(\boldsymbol{m}_s)$ is a cover for $G$. Since $\operatorname{supp}(\boldsymbol{m}_s) = \operatorname{supp}(\boldsymbol{m}_s')$, we know $\operatorname{ver}(\boldsymbol{m}_{s'})$ is a cover for $G$, and it follows that $\boldsymbol{m}_s' \in J_G$. In particular $\boldsymbol{w}$ can be written as the product of the $n$ monomials $\boldsymbol{m}_1 \cdots \boldsymbol{m}_{s-1}\boldsymbol{m}_s' \cdots \boldsymbol{m}_n$, which shows that $\boldsymbol{w} \in J_G^n$. □

In the rest of the paper, if $\boldsymbol{m} = \prod_{i=1}^d x_i^{\alpha_i}$ is a monomial in the ring $\mathsf{k}[x_1, \ldots, x_d]$, then $\deg_{\boldsymbol{m}} x_i = \alpha_i$, while the total degree of $\boldsymbol{m}$ is given by $\sum_{i=1}^d \alpha_i$ and is denoted by $\operatorname{tot\,deg} \boldsymbol{m}$.

The following corollary is an immediate consequence of the previous lemma.

**2.2. Corollary.** *Let $G$ be a graph with vertex set $V_G$ of cardinality larger than 2. Let $J_G$ be the cover ideal of $G$ in the polynomial ring $\mathsf{k}[V_G]$. Assume that $\{x_1, x_2\}$ is an edge of $G$ and assume that $\mathfrak{m}_G \in \operatorname{Ass}(R/J_G^n)$. If $\boldsymbol{w}$ is a witness of $\mathfrak{m}_G$, then $x_1, x_2 \in \operatorname{supp} \boldsymbol{w}$. Moreover, $\deg_{\boldsymbol{w}} x_1 + \deg_{\boldsymbol{w}} x_2 \geq n$.*

*Proof.* Assume for the sake of contradiction that $x_2$ does not divide $\boldsymbol{w}$. Let $x \in V_G \setminus \{x_1, x_2\}$. The monomial $x\boldsymbol{w}$ can be written as the product of $n$ monomials $\boldsymbol{m}_1 \cdots \boldsymbol{m}_n$ such that $\boldsymbol{m}_i \in J_G$ for all $i = 1, \ldots, n$. By Lemma 2.1 $\deg_{\boldsymbol{w}} x_1 \leq n - 1$, and therefore we can conclude that there exists an $i \in \{1, \ldots, n\}$ such that $x_1$ does not divide $\boldsymbol{m}_i$. Since $x_2$ does not divide $\boldsymbol{w}$, it follows that $x_2$ does not divide $\boldsymbol{m}_i$. In particular, $\operatorname{ver}(\boldsymbol{m}_i)$ cannot be a cover of $G$, as neither $x_1$ nor $x_2$ are in $\operatorname{supp}(\boldsymbol{m}_i)$, while $\{x_1, x_2\}$ forms an edge.

Notice that either $x_1$ or $x_2$ divides $\boldsymbol{m}_i$, as $\boldsymbol{m}_i \in J_G$ for all $i = 1, \ldots, n$, verifying the final statement. □

In the following $K_h$ denotes the complete graph in $h$ vertices. Notice that every cover of $K_h$ contains at least $h - 1$ vertices.

**2.3. Lemma.** *Let $G$ be a graph with vertex set $V_G$. Let $J_G$ be the cover ideal in the polynomial ring $R = \mathsf{k}[V_G]$. If $G$ contains the complete graph $K_h$ as an induced subgraph but $G \neq K_h$, then $\mathfrak{m}_G \notin \operatorname{Ass}(R/J_G^n)$ for all integers $n$ such that $n \leq h - 1$.*

*Proof.* Suppose $G$ contains $K_h$ as an induced subgraph. Without loss of generality we may label the vertices of $K_h$ with the variables $\{x_1, \ldots, x_h\}$. Towards contradiction, assume that $\mathfrak{m}_G \in \operatorname{Ass}(R/J_G^n)$ with $n \leq h - 1$, and let $\boldsymbol{w}$ be a witness. For every monomial $\boldsymbol{c} \in J_G$, we have that $\boldsymbol{c} \in J_{K_h}$. This implies that at least

$h-1$ variables among $x_1, \ldots, x_h$ belong to $\operatorname{supp} \boldsymbol{c}$. Therefore, if $\boldsymbol{c} \in J_G^n$ then $\sum_{i=1}^h \deg_{\boldsymbol{c}} x_i \geq n(h-1) = nh - n$.

However, we know from Lemma 2.1 that for each variable $x_i$ the inequality $\deg_{\boldsymbol{w}} x_i \leq n-1$ holds, so that $\sum_{i=1}^h \deg_{\boldsymbol{w}} x_i \leq h(n-1) = hn - h$.

If $x \in V_G$ and $x \neq x_i$ for $i = 1, \ldots, h$, then $x\boldsymbol{w} \in J_G^n$, as $\boldsymbol{w}$ is a witness of $\mathfrak{m}_G$, which yields

$$n(h-1) \leq \sum_{i=1}^{h} \deg_{x_j \boldsymbol{w}} x_i = \sum_{i=1}^{h} \deg_{\boldsymbol{w}} x_i \leq h(n-1).$$

This gives us the desired contradiction $h \leq n$. □

In the following lemma, under proper assumptions, we can be more specific about the degree formula presented in Corollary 2.2.

**2.4.** A monomial $\boldsymbol{n} \in \mathsf{k}[x_1, \ldots, x_d]$ is said square-free if for all $i = 1, \ldots, d$ the monomial $x_i^2$ does not divide $\boldsymbol{n}$. For a graph $G$ with cover ideal $J_G$, given a monomial $\boldsymbol{m} \in J_G$, one can always find a square-free monomial $\boldsymbol{n} \in J_G$ such that $\boldsymbol{n}$ divides $\boldsymbol{m}$. In particular for a product of $n$ monomials $\boldsymbol{m} = \boldsymbol{m}_1 \cdots \boldsymbol{m}_n$ such that $\boldsymbol{m}_i \in J_G$ for all $i = 1, \ldots, n$ and $\deg_{\boldsymbol{m}} x_j \leq n-1$ for all $j = 1, \ldots, d$, we may assume that each $\boldsymbol{m}_i$ is square-free.

**2.5. Lemma.** *Let $G$ be a graph with vertex set $V_G$ of cardinality bigger than 4. Let $J_G$ be the cover ideal of $G$ in the polynomial ring $\mathsf{k}[V_G]$. Assume that there are $x_1, x_2, x_3, x_4 \in V_G$ such that $N(x_2) = \{x_1, x_3\}$ and $N(x_3) = \{x_2, x_4\}$. Assume further that, for a given positive integer $n$, $\mathfrak{m}_G \in \operatorname{Ass}(R/J_G^n)$ with witness $\boldsymbol{w}$. If $\deg_{\boldsymbol{w}} x_1 = n-1$, then $\deg_{\boldsymbol{w}} x_2 + \deg_{\boldsymbol{w}} x_3 = n$.*

*Proof.* Since $\boldsymbol{w}$ is a witness for the ideal $J_G^n$, we know that $\deg_{\boldsymbol{w}} x_2 + \deg_{\boldsymbol{w}} x_3 \geq n$ by the adjacency assumption and Corollary 2.2.

Since $\boldsymbol{w}$ is a witness for $\mathfrak{m}_G$, we have $x_2\boldsymbol{w} = \boldsymbol{m}_1 \cdots \boldsymbol{m}_n$, where $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_n \in J_G$. By Lemma 2.1, $\deg_{\boldsymbol{w}} x_i \leq n-1$, so we may assume that the monomial $\boldsymbol{m}_j$ is square-free for all $j = 1, \ldots, n$; see 2.4.

Suppose for contradiction that $\deg_{\boldsymbol{w}} x_2 + \deg_{\boldsymbol{w}} x_3 \geq n+1$, which implies that $\deg_{x_2\boldsymbol{w}} x_2 + \deg_{x_2\boldsymbol{w}} x_3 \geq n+2$.

By Corollary 2.2, both $x_2$, and $x_3$ are in $\operatorname{supp} \boldsymbol{w}$. This implies that $x_3^2$ divides $x_2\boldsymbol{w}$, as $\deg_{x_2\boldsymbol{w}} x_2 \leq n$, and therefore there exist two integers $i_1$ and $i_2$ such that $x_2$ and $x_3$ belong to $\operatorname{supp} \boldsymbol{m}_{i_1}$ and $\operatorname{supp} \boldsymbol{m}_{i_2}$. If also $x_1$ belongs to $\operatorname{supp} \boldsymbol{m}_{i_j}$ for some $j = 1, 2$, then $\boldsymbol{m}_{i_j}/x_2 \in J_G$, since $x_1x_3$ divides $\boldsymbol{m}_{i_j}/x_2$. Thus, in this case,

$$\boldsymbol{w} = \frac{x_2\boldsymbol{m}}{x_2} = \boldsymbol{m}_1 \cdots \frac{\boldsymbol{m}_{i_j}}{x_2} \cdots \boldsymbol{m}_n \in J_G^n,$$

a contradiction, since $\boldsymbol{w}$ is a witness. Thus we may assume that $x_1$ does not divide $\boldsymbol{m}_{i_1}$ and $\boldsymbol{m}_{i_2}$, which implies that $\deg_{\boldsymbol{w}} x_1 < n-1$, contradicting the hypothesis. □

The careful analysis of the degrees of the witnesses allows us to draw useful conclusions about when $\mathfrak{m}_G$ is an associated prime after contracting a vertex.

**2.6. Lemma.** *Let $G$ be a graph with vertex set $V_G$. Let $J_G$ be the cover ideal of $G$ in the polynomial ring $R = \mathsf{k}[V_G]$. Assume $x_1, y_1, y_2, x_2 \in V_G$ such that $N(y_1) = \{x_1, x_2\}$ and $N(y_2) = \{y_1, x_2\}$. Assume that $\mathfrak{m}_G \in \operatorname{Ass}(R/J_G^n)$ for some integer $n$ and that there exists a witness $\boldsymbol{w}$ such that $\deg_{\boldsymbol{w}} x_1 = n-1$. Obtain $G'$ by contracting $y_1$ and $y_2$. Then $\mathfrak{m}_{G'}$ belongs to $\operatorname{Ass}(\mathsf{k}[V_{G'}]/J_{G'}^n)$.*

*Proof.* Set $a_1 = \deg_{\boldsymbol{w}} y_1$ and let $a_2 = \deg_{\boldsymbol{w}} y_2$. We prove that the monomial $\boldsymbol{w}' = \boldsymbol{w}/(y_1^{a_1} y_2^{a_2})$ is a witness for the ideal $\mathfrak{m}_{G'}$, and thus $\mathfrak{m}_{G'}$ is an element of $\operatorname{Ass}(R/J_{G'}^k)$.

First, we show by contradiction that $\boldsymbol{w}' \notin J_{G'}^n$; toward this end, suppose that $\boldsymbol{w}' = \boldsymbol{m}_1 \cdots \boldsymbol{m}_n$ such that $\boldsymbol{m}_i \in J_{G'}$. For every $x \in V_{G'} \subset V_G$, we have $\deg_{\boldsymbol{w}'} x = \deg_{\boldsymbol{w}} x \leq n-1$, where the inequality is the content of Lemma 2.1. Therefore, by 2.4, we may assume that, for each $x \in V_{G'}$, $x^2$ does not divide $\boldsymbol{m}_j$ for $j = 1, \ldots, n$. For $1 \leq i \leq n$, define the monomial $\boldsymbol{n}_i$ as

$$\boldsymbol{n}_i = \begin{cases} \boldsymbol{m}_i & \text{if } x_1, x_2 \in \operatorname{supp} \boldsymbol{m}_i, \\ y_1 \boldsymbol{m}_i & \text{if } x_1 \notin \operatorname{supp} \boldsymbol{m}_i, \\ y_2 \boldsymbol{m}_i & \text{if } x_2 \notin \operatorname{supp} \boldsymbol{m}_i. \end{cases}$$

Since $\boldsymbol{m}_i \in J_{G'}$ and $\{x_1, x_2\}$ is an edge of the graph $G'$, each $\boldsymbol{m}_i$ is divisible by at least one of $x_1$ or $x_2$, so that our construction of $\boldsymbol{n}_i$ is well-defined. Moreover, for the same reason, for each $i$ such that $1 \leq i \leq n$, if $y_1 \in \operatorname{supp} \boldsymbol{n}_i$ or $y_2 \in \operatorname{supp} \boldsymbol{n}_i$ then $\boldsymbol{n}_i \in J_G$.

Denote by $\boldsymbol{w}''$ the product $\boldsymbol{n}_1 \cdots \boldsymbol{n}_n$ and set $b_i = \deg_{\boldsymbol{w}''} y_i$ for $i = 1, 2$. There are $n - b_1 - b_2$ monomials among the $\boldsymbol{n}_i$ such that $y_1, y_2 \notin \operatorname{supp} \boldsymbol{n}_i$ and therefore there are $n - b_1 - b_2$ monomials among the $\boldsymbol{n}_i$ such that $\operatorname{ver}(\boldsymbol{n}_i)$ are not covers of $G$ as $\{y_1, y_2\}$ is an edge in $G$. We may assume, by renaming the $\boldsymbol{n}_i$, that

$$\begin{cases} \boldsymbol{n}_i \notin J_G, & i = 1, \ldots, n - b_1 - b_2, \\ \boldsymbol{n}_i \in J_G, & i = n - b_1 - b_2 + 1, \ldots, n. \end{cases}$$

Since $\deg_{\boldsymbol{m}_i} x \leq 1$ for every $x \in V_{G'}$, we have $\deg_{\boldsymbol{w}''} y_j = n - \deg_{\boldsymbol{w}''} x_j$ for $j = 1, 2$. In particular,

$$b_j = \deg_{\boldsymbol{w}''} y_j = n - \deg_{\boldsymbol{w}''} x_j = n - \deg_{\boldsymbol{w}} x_j \leq \deg_{\boldsymbol{w}} y_j = a_i,$$

where the inequality follows from Corollary 2.2, the fact that $\boldsymbol{w}$ is a witness for $\mathfrak{m}_G$ in $\operatorname{Ass}(R/J_G^n)$, and the assumption that $\{x_j, y_j\}$ is an edge of $G$ for $j = 1, 2$. As $\deg_{\boldsymbol{w}''} x = \deg_{\boldsymbol{w}} x$ for all $x \in V_{G'}$, we know $\boldsymbol{w}''$ divides $\boldsymbol{w}$ and $\boldsymbol{w} = y_1^{a_1 - b_1} y_2^{a_2 - b_2} \boldsymbol{w}''$.

Notice that for each $i = 1, \ldots, n - b_1 - b_2$, and for each $j = 1, 2$, the monomial $y_j \boldsymbol{n}_i$ is in $J_G$. Since $a_1 + a_2 = n$ by Lemma 2.5, $y_1^{a_1 - b_1} y_2^{a_2 - b_2} \boldsymbol{n}_1 \cdots \boldsymbol{n}_{n - b_1 - b_2} \in J_G^{n - b_1 - b_2}$, so that $\boldsymbol{w} = y_1^{a_1} y_2^{a_2} \boldsymbol{w}'' = y_1^{a_1 - b_1} y_2^{a_2 - b_2} \boldsymbol{n}_1 \cdots \boldsymbol{n}_k \in J_G^n$, a contradiction to our assumption about $\boldsymbol{w}$ being a witness. Thus, we conclude that $\boldsymbol{w}'$ could not have been in $J_{G'}^n$ to begin with, completing the first section of the proof.

Next, we show that for $x \in V_{G'}$, we have $x\boldsymbol{w}' \in J_{G'}^n$. But $x\boldsymbol{w} \in J_G^n$, and in particular $x\boldsymbol{w} = \boldsymbol{m}_1 \cdots \boldsymbol{m}_n$, where $\boldsymbol{m}_i \in J_G$ for $1 \leq i \leq n$. Since $a_1 + a_2 = n$ by Lemma 2.5, and since each $\boldsymbol{m}_i$ must be divisible by at least one of $y_1$ or $y_2$ (since $\{y_1, y_2\} \in E_G$), it must be the case that each $\boldsymbol{m}_i$ contains precisely one of $y_1$ or $y_2$. This implies that $y_1 \in \operatorname{supp} \boldsymbol{m}_i$ if and only if $x_2 \in \operatorname{supp} \boldsymbol{m}_i$, and $y_2 \in \operatorname{supp} \boldsymbol{m}_i$ if and only if $x_1 \in \operatorname{supp} \boldsymbol{m}_i$, since $\operatorname{ver}(\boldsymbol{m}_i)$ is a cover for $G$. Thus either $x_1$ or $x_2$ belong to $\operatorname{supp}(\boldsymbol{m}_i)$ for every $i = 1, \ldots, n$. For this reason the monomials defined as

$$\boldsymbol{m}_i' = \begin{cases} \boldsymbol{m}_i / y_1 & \text{if } y_1 \in \operatorname{supp} \boldsymbol{m}_i, \\ \boldsymbol{m}_i / y_2 & \text{if } y_2 \in \operatorname{supp} \boldsymbol{m}_i \end{cases}$$

have the property that $\operatorname{ver}(\boldsymbol{m}_i')$ is a cover for $G'$ for all $i = 1, \ldots, n$. Therefore we have $x\boldsymbol{w}' = \boldsymbol{m}_1' \cdots \boldsymbol{m}_n' \in J_{G'}^n$, as desired. □

The following lemma gives instances for which a variable appears with maximal degree in a witness.

**2.7. Lemma.** *Let $G$ be a graph with vertex set $V_G$. Let $J_G$ be the cover ideal for $G$ in the polynomial ring $R = \mathsf{k}[V_G]$. Assume that there exists a positive integer $n$ such that $\mathfrak{m}_G \in \operatorname{Ass}(R/J_G^n)$ with witness $\boldsymbol{w}$. Suppose $G$ contains a proper induced subgraph $K$ that is a complete graph in $n+1$ vertices with one edge $\{y_1, y_2\}$ removed. Then $\deg_{\boldsymbol{w}}(y_1) = n - 1$.*

*Proof.* Label the vertices in $V_K$ as $y_1, y_2, \ldots, y_{n+1}$. By Lemma 2.1, we know $\deg_{\boldsymbol{w}}(y_1) \leq n - 1$, so it remains to show that $\deg_{\boldsymbol{w}}(y_1) \geq n - 1$. Suppose for the sake of contradiction that $\deg_{\boldsymbol{w}}(y_1) < n - 1$, and let $x$ be a vertex of $G$ but not a vertex of the proper subgraph $H$. Since $x\boldsymbol{w} \in J_G^n$, we can write $x\boldsymbol{w} = \boldsymbol{m}_1 \cdots \boldsymbol{m}_n$, with $\boldsymbol{m}_i \in J_G$. This implies that for each $i = 1, \ldots, n$, $\operatorname{ver}(\boldsymbol{m}_i)$ is a cover of $G$ and therefore a cover for $K$.

Since $\deg_{\boldsymbol{w}}(y_1) < n - 1$, suppose without loss of generality that $y_1 \nmid \boldsymbol{m}_{n-1}$ and $y_1 \nmid \boldsymbol{m}_n$. Then $y_j \in \operatorname{supp}(\boldsymbol{m}_i)$ for $3 \leq j \leq n+1$ and $i = n-1, n$ since $\{y_1, y_j\}$ is an edge of $H$ and therefore $G$. In particular $y_3 \cdots y_{n+1} \mid \boldsymbol{m}_{n-1}$ and $y_3 \cdots y_{n+1} \mid \boldsymbol{m}_n$. Again by Lemma 2.1, we know that $\deg_{\boldsymbol{w}}(y_j) \leq n - 1$, so $y_j$ can divide at most $n - 3$ of the monomials $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_{n-2}$ for $3 \leq j \leq n+1$. Thus,

$$\sum_{j=3}^{n+1} \sum_{i=1}^{n-2} \deg_{\boldsymbol{m}_i}(y_j) \leq \sum_{j=3}^{n+1} (n-3) = n^2 - 4n + 3.$$

On the other hand, each $\boldsymbol{m}_i$ must cover $H$ and so contains at least all but one of $y_3, \ldots, y_{n+1}$, whence

$$\sum_{i=1}^{n-2} \sum_{j=3}^{n+1} \deg_{\boldsymbol{m}_i}(y_j) \geq \sum_{i=1}^{n-2} (n-2) = n^2 - 4n + 4,$$

which is obviously a contradiction. Thus we conclude that $\deg_{\boldsymbol{w}}(y_1) = n-1$, as desired. □

In the rest of the paper, given a finite set $S$, we denote by $|S|$ its cardinality.

**2.8. Lemma.** *Let $G$ be an $h$-wheel with rim $R^G$ and center $C^G$. Let $k$ be its radial number and $\ell_1, \ldots, \ell_k$ its radial lengths. If $W$ is a vertex cover for $G$ that contains all the vertices in $C^G$, then*

$$|W| \geq \tfrac{1}{2}(|G| - h + 1) + h.$$

*If $W$ is a vertex cover for $G$ missing one vertex from $C^G$, then*

$$|W| \geq k + h - 1 + \left\lfloor \tfrac{1}{2}(\ell_1 - 1) \right\rfloor + \cdots + \left\lfloor \tfrac{1}{2}(\ell_k - 1) \right\rfloor.$$

*Moreover*,

$$k + h - 1 + \left\lfloor \tfrac{1}{2}(\ell_1 - 1) \right\rfloor + \cdots + \left\lfloor \tfrac{1}{2}(\ell_k - 1) \right\rfloor \geq \tfrac{1}{2}(|G| - h + 1) + h.$$

*Proof.* Assume that $W$ contains $C^G$. The vertex set $W \cap R^G$ has to be a vertex cover for $R^G$. Since $R^G$ is an odd hole, the cardinality of $W \cap R^G$ has to be at least

$$\tfrac{1}{2}(|R^G| + 1) = \tfrac{1}{2}(|G| - h + 1).$$

Therefore the cardinality of $W$ is at least

$$\tfrac{1}{2}(|G| - h + 1) + h.$$

Assume now that $W$ does not contain all the center vertices. If $G$ were a 1-wheel, we know from [Kesting et al. 2011, Lemma 2.1] that the cover not containing the center would have cardinality of at least

$$k + \left\lfloor \tfrac{1}{2}(\ell_1 - 1) \right\rfloor + \cdots + \left\lfloor \tfrac{1}{2}(\ell_k - 1) \right\rfloor,$$

which is also the number of vertices that $W$ needs to have to cover the subgraph induced by the 1-wheel with the center not in $W$. The cover $W$ needs to contain further the other $h-1$ centers, so that the following inequality holds:

$$|W| \geq k + h - 1 + \left\lfloor \tfrac{1}{2}(\ell_1 - 1) \right\rfloor + \cdots + \left\lfloor \tfrac{1}{2}(\ell_k - 1) \right\rfloor.$$

We now need to show that this value is greater than $\frac{1}{2}(|G| - h + 1) + h$. Denote by $C$ a subgraph of $G$ isomorphic to a 1-wheel. We know that

$$k + \left\lfloor \tfrac{1}{2}(\ell_1 - 1) \right\rfloor + \cdots + \left\lfloor \tfrac{1}{2}(\ell_k - 1) \right\rfloor \geq \tfrac{1}{2}|C| + 1,$$

as shown in [Kesting et al. 2011]. This implies

$$k + \left\lfloor \tfrac{1}{2}(\ell_1 - 1) \right\rfloor + \cdots + \left\lfloor \tfrac{1}{2}(\ell_k - 1) \right\rfloor \geq \tfrac{1}{2}(|G| - h + 1) + 1,$$

as $|G| - h + 1$ is the cardinality of a subgraph of $G$ isomorphic to a 1-wheel. It follows that

$$k + h - 1 + \left\lfloor \tfrac{1}{2}(\ell_1 - 1) \right\rfloor + \cdots + \left\lfloor \tfrac{1}{2}(\ell_k - 1) \right\rfloor \geq \tfrac{1}{2}(|G| - h + 1) + h. \qquad \square$$

## 3. Main theorems

We first prove that if $G$ is an $h$-wheel then $\mathfrak{m}_G$ appears as an associated prime of low powers of the cover ideal.

**3.1. Theorem.** *Let $G$ be an $h$-wheel, and let $J_G$ be the cover ideal of $G$ in the ring $R = \mathsf{k}[V_G]$. Then $\mathfrak{m}_G \notin \operatorname{Ass}(R/J_G^n)$ if $n \leq h + 1$.*

*Proof.* Let $y_1, \ldots, y_h$ label the vertices in $C^G$, let $x_1, x_2, \ldots, x_k$ label the radial vertices, and let $\ell_i$ be the radial lengths for $i = 1, \ldots, k$. Denote by $x_{ij}$, for $j = 1, \ldots, \ell_i - 1$, the vertices between $x_i$ and $x_{i+1}$ if $i < k$ and the vertices between $x_k$ and $x_1$ if $i = k$.

Because the centers and one radial vertex form a complete graph in $h+1$ vertices, Lemma 2.3 implies that $G \notin \operatorname{Ass}(R/J^n)$ for every integer $n$ such that $n \leq h$.

We next show that $G \notin \operatorname{Ass}(R/J_G^{h+1})$, and to do so we consider two cases.

Case 1: Assume that there are two radial vertices, say $x_t$ and $x_{t+1}$, such that $\{x_t, x_{t+1}\}$ is an edge. In this case we can conclude that $G \notin \operatorname{Ass}(R/J^{h+1})$ by a direct application of Lemma 2.3 since $x_t$, $x_{t+1}$, and the centers of the $h$-wheel $G$ form a complete $(h+2)$-graph.

Case 2: Assume that $G$ is an $h$-wheel with no two radial vertices adjacent. We know by the definition of an $h$-wheel that there exist an $x_t$ and an $x_{t+1}$ such that the path from $x_t$ to $x_{t+1}$ is odd. By relabeling the vertices of $G$ we may assume that $t = 1$. Suppose for a contradiction that there exists a witness $\boldsymbol{w}$ for the maximal ideal $\mathfrak{m}_G$ to be in $\operatorname{Ass}(R/J^{h+1})$. Using Lemma 2.7 with $K$ being the induced subgraph by $C^G$, and the vertices $x_1, x_2$, we can conclude that the $\deg_{\boldsymbol{w}} x_1 = h$. Thus from Lemma 2.5, we have that $\deg_w x_{11} + \deg_w x_{12} = h + 1$. Further, by an application of Lemma 2.6, we can contract $x_{11}$ and $x_{12}$ to form a new graph $G'$ such that $\mathfrak{m}_{G'} \in \operatorname{Ass}(\mathsf{k}[V_{G'}]/J_{G'}^{h+1})$. Because the path from $x_1$ to $x_2$ along the subgraph induced by $R^G$ is odd, we can perform this operation until $x_1$ is adjacent to $x_2$ and conclude the proof by an application of Case 1. $\square$

**3.2. Theorem.** *Let $G$ be an $h$-wheel and let $J_G$ be the cover ideal of $G$ in the ring $R = \mathsf{k}[V_G]$. Then $\mathfrak{m}_G \in \operatorname{Ass}(R/J_G^{h+2})$.*

*Proof.* Label with $y_1, \ldots, y_h$ the vertices in $C^G$, and with $x_1, \ldots, x_k$ the radial vertices, where $k$ is the radial number. Let $\ell_i$ denote the radial lengths for $i = 1, \ldots, k$. Label by $x_{ij}$, for $j = 1, \ldots, \ell_i - 1$, the vertices between $x_i$ and $x_{i+1}$ if

$i < k$ and the vertices between $x_k$ and $x_1$ if $i = k$. The subgraph $R^G$ is an odd cycle. We set $d$ to be the size of $R^G$. Notice that $\ell_1 + \cdots + \ell_k = d$.

We prove that $\mathfrak{m}_G$ is in $\operatorname{Ass}(R/J_G^{h+2})$ by providing a witness. Let $\boldsymbol{w}$ be the monomial

$$\boldsymbol{w} = \Big(\prod_{i=1,\dots,h} y_i^{h+1}\Big)\Big(\prod_{i=1,\dots,k} x_i^{h+1}\Big)\Big(\prod_{\substack{i=1,\dots,k\\ j=1,\dots,\ell_i-1}} x_{ij}^{a}\Big),$$

where $a = 1$ if $j$ is odd, and $a = h+1$ if $j$ is even.

To show that $\boldsymbol{w}$ is the desired monomial, we first prove that

$$\operatorname{tot deg}(\boldsymbol{w}) = hk + h(h+1) + n + h\big(\lfloor \tfrac{1}{2}(l_1-1)\rfloor + \cdots + \lfloor \tfrac{1}{2}(l_k-1)\rfloor\big).$$

In computing the $\deg(\boldsymbol{w})$, the contribution from the variables $y_m$ and $x_i$, for $m = 1, \dots, h$ and $i = 1, \dots, k$, is given by $(h+1)h + (h+1)k$. For $i = 1, \dots, k-1$, between $x_i$ and $x_{i+1}$, there are $\ell_i - 1$ vertices, and there are $\ell_k - 1$ vertices between $x_k$ and $x_1$. Given an integer $s$, there are $\lfloor \frac{1}{2}s \rfloor$ even integers and $\lceil \frac{1}{2}s \rceil$ odd integers between 1 and $s$. Therefore, in computing $\operatorname{tot deg}(\boldsymbol{w})$, the contributions from the variables $x_{ij}$ are given by

$$(h+1)\big(\lfloor \tfrac{1}{2}(l_1-1)\rfloor + \cdots + \lfloor \tfrac{1}{2}(l_k-1)\rfloor\big) + \lceil \tfrac{1}{2}(l_1-1)\rceil + \cdots + \lceil \tfrac{1}{2}(l_k-1)\rceil.$$

The total degree of the monomial $\boldsymbol{w}$ is therefore equal to

$$\begin{aligned}
\operatorname{tot deg}(\boldsymbol{w}) &= (h+1)k + (h+1)h + \sum_{i=1}^{k} \lceil \tfrac{1}{2}(\ell_i-1)\rceil + (h+1)\sum_{i=1}^{k} \lfloor \tfrac{1}{2}(\ell_i-1)\rfloor \\
&= (h+1)h + (h+1)k + h\sum_{i=1}^{k} \lfloor \tfrac{1}{2}(\ell_i-1)\rfloor + \sum_{i=1}^{k} \big(\lfloor \tfrac{1}{2}(\ell_i-1)\rfloor + \lceil \tfrac{1}{2}(\ell_i-1)\rceil\big) \\
&= hk + h(h+1) + k + \sum_{i=1}^{k} (\ell_i - 1) + h\sum_{i=1}^{k} \lfloor \tfrac{1}{2}(\ell_i-1)\rfloor \\
&= hk + h(h+1) + \sum_{i=1}^{k} \ell_i + h\sum_{i=1}^{k} \lfloor \tfrac{1}{2}(\ell_i-1)\rfloor \\
&= hk + h(h+1) + d + h\sum_{i=1}^{k} \lfloor \tfrac{1}{2}(\ell_i-1)\rfloor.
\end{aligned}$$

To prove that $\boldsymbol{w}$ does not belong to $J_G^{h+2}$, we first show that

$$\operatorname{tot deg}(\boldsymbol{w}) < 2\big(\tfrac{1}{2}(|G| - h + 1) + h\big) + h\Big(k + h - 1 + \sum_{i=1}^{k} \lfloor \tfrac{1}{2}(\ell_i - 1)\rfloor\Big). \quad (3.2.1)$$

Supposing this inequality is not satisfied, we have

$$2\big(\tfrac{1}{2}(|G|-h+1)+h\big)+hk+h^2-h+h\sum_{i=1}^{k}\big\lfloor\tfrac{1}{2}(\ell_i-1)\big\rfloor \leq hk+h^2+h+d+h\sum_{i=1}^{k}\big\lfloor\tfrac{1}{2}(\ell_i-1)\big\rfloor,$$

which implies

$$h+d\geq 2\big(\tfrac{1}{2}(|G|-h+1)+h\big)-h,$$

or $h+d\geq |G|+1$. But $|G|=|C^G|+h=d+h$. Thus

$$d+h\geq d+h+1,$$

which is impossible. Thus the inequality holds.

Now we show that this inequality implies $\boldsymbol{w}\notin J_G^{h+2}$. Assume otherwise. Then we can write $\boldsymbol{w}=\boldsymbol{h}\boldsymbol{m}_1\cdots\boldsymbol{m}_{h+2}$ such that for each $i=1,\dots,h+2$ not only the monomial $\boldsymbol{m}_i\in J_G$ but also $\operatorname{ver}(\boldsymbol{m}_i)$ is a minimal cover for $G$. The total degree of each $\boldsymbol{m}_i$ is equal to $|\operatorname{ver}(\boldsymbol{m}_i)|$. Therefore, by Lemma 2.8, we have

$$\operatorname{tot}\deg(\boldsymbol{m}_i)\geq \tfrac{1}{2}(|C|-h+1)+h$$

if $\operatorname{ver}(\boldsymbol{m}_i)$ is a cover containing the vertices of $C^G$, or

$$\operatorname{tot}\deg(\boldsymbol{m}_i)\geq k+h-1+\big\lfloor\tfrac{1}{2}(\ell_1-1)\big\rfloor+\cdots+\big\lfloor\tfrac{1}{2}(\ell_k-1)\big\rfloor$$

if $\operatorname{ver}(\boldsymbol{m}_i)$ is a cover that does not contain all vertices of $C^G$.

Notice that $\sum_{i=1}^{h}\deg_{\boldsymbol{w}} y_i=h(h+1)$. If $\operatorname{ver}(\boldsymbol{m}_i)$ is a cover that contains all the vertices of $C^G$ for each $i=1,\dots,h-2$ then $\sum_{i=1}^{h}\deg_{\boldsymbol{w}} y_i\geq h(h+2)$, which is a contradiction. In particular, there are least $h$ monomials among the monomials $\boldsymbol{m}_i$ that correspond to covers not containing all vertices in $C^G$. An application of Lemma 2.8, yields the inequality

$$\begin{aligned}\operatorname{tot}\deg(\boldsymbol{w})&=\operatorname{tot}\deg(\boldsymbol{h})+\operatorname{tot}\deg(\boldsymbol{m}_1)+\cdots+\operatorname{tot}\deg(\boldsymbol{m}_{h+2})\\&\geq 2\big(\tfrac{1}{2}(|C|-h+1)+h\big)+h\big(k+h-1+\big\lfloor\tfrac{1}{2}(l_1-1)\big\rfloor+\cdots+\big\lfloor\tfrac{1}{2}(l_k-1)\big\rfloor\big).\end{aligned}$$

This contradicts inequality (3.2.1) and shows that $\boldsymbol{w}\notin J_G^{h+2}$.

To finish the proof, we need to show that for every vertex $x\in V_G$ the monomial $x\boldsymbol{w}$ is in $J_G^{h+2}$.

For every $i=1,\dots,h$, let $C_i$ be the induced subgraph isomorphic to the 1-wheel with center in $y_i$. In [Kesting et al. 2011, Theorem 2.2], the authors prove that

$$\boldsymbol{w}_i=y_i^2\prod_{i=1,\dots,k}x_i^2\prod_{j\,\text{odd}}x_{ij}\prod_{j\,\text{even}}x_{ij}^2 \tag{3.2.2}$$

is a witness for $\mathfrak{m}_{C_i} \in \operatorname{Ass}(\mathsf{k}[V_{C_i}]/J_{C_i}^3)$. Pick a vertex $x \in V_G$. Without loss of generality we may assume that $x \in V_{C_1}$. Then $x\boldsymbol{w}_1 \in J_{C_1}^3$, so $y_2^3 \cdots y_h^3 x\boldsymbol{w}_1 \in J_G^3$.

Define $\boldsymbol{m} = \prod_{i=1,\dots,k} x_i^2 \prod_{j \text{ odd}} x_{ij} \prod_{j \text{ even}} x_{ij}^2$ and notice that

$$\boldsymbol{w} = \frac{y_1^{h-1} y_2^{h+1} \cdots y_h^{h+1} \boldsymbol{w}_1 \cdot \boldsymbol{m}^{h-1}}{\prod_{i,j} x_i^{h-1} x_{ij}^{h-1}}.$$

Define

$$\boldsymbol{m}_i = \frac{y_1 \cdots y_{i-1} y_{i+1} \cdots y_h \cdot \boldsymbol{m}}{\prod_{i,j} x_i x_{ij}}$$

for each $i = 2, \dots, h$. It is easy to see that $\operatorname{ver}(\boldsymbol{m}_i)$ is a cover for $G$ for every $i = 2, \dots, h$. The following equality shows that $x\boldsymbol{w} \in J_G^{h+2}$:

$$x\boldsymbol{w} = (y_2^3 \cdots y_h^3 x\boldsymbol{w}_1)\boldsymbol{m}_2 \cdots \boldsymbol{m}_h. \qquad \square$$

Finally we prove that if $G$ is an $h$-wheel then $\mathfrak{m}_G$ is an associated prime in high powers of the cover ideal.

**3.3. Theorem.** *Let $G$ be an $h$-wheel and let $J_G$ be the cover ideal of $G$ in the ring $R = \mathsf{k}[V_G]$. Then $\mathfrak{m}_G \in \operatorname{Ass}(R/J_G^n)$ for all $n \geq h+2$.*

*Proof.* Fix an integer $n \geq h+2$ and let $t$ satisfy $n = h+2+t$. Let $S$ be the cover of $G$ that has all the vertices in $C^G$ and every other vertex in $R^G$. In particular $|S| = h + \frac{1}{2}(|R^G|+1)$.

Consider the monomial $\tilde{\boldsymbol{w}} = (\boldsymbol{m})^t \boldsymbol{w}$, where $\boldsymbol{w}$ is the witness constructed in the proof of Theorem 3.2 and $\boldsymbol{m}$ is the squarefree monomial such that $\operatorname{ver}(\boldsymbol{m}) = S$. In particular, $\operatorname{tot deg} \boldsymbol{m} = h + \frac{1}{2}(|R^G|+1) = h + \frac{1}{2}(|G| - h + 1)$. Using the inequality (3.2.1) we obtain

$$\begin{aligned}
&\operatorname{tot deg}(\tilde{\boldsymbol{w}}) \\
&< t\big(\tfrac{1}{2}(|G|-h+1)+h\big) + 2\big(\tfrac{1}{2}(|G|-h+1)+h\big) + h\left(k+h-1+\sum_{i=1}^{k}\left\lfloor \tfrac{1}{2}(\ell_1-1)\right\rfloor\right) \\
&= (n-h)\big(\tfrac{1}{2}(|G|-h+1)+h\big) + h\left(k+h-1+\sum_{i=1}^{k}\left\lfloor \tfrac{1}{2}(\ell_1-1)\right\rfloor\right).
\end{aligned}$$

We claim that $\tilde{\boldsymbol{w}}$ is a witness for $\mathfrak{m}_G \in \operatorname{Ass}(\mathsf{k}[V_G]/(J_G^n))$. If, toward contradiction, $\tilde{\boldsymbol{w}} \in J_G^n$, then we can write $\tilde{\boldsymbol{w}} = \boldsymbol{h}\boldsymbol{m}_1 \cdots \boldsymbol{m}_n$ such that, for each $i = 1, \dots, n$, not only the monomial $\boldsymbol{m}_i \in J_G$ but also $\operatorname{ver}(\boldsymbol{m}_i)$ is a minimal cover for $G$. As $\sum_{i=1}^{h} \deg_{\tilde{\boldsymbol{w}}} y_i = th + h(h+1) = (n-1)h$, there are at least $h$ covers among $\operatorname{ver}(\boldsymbol{m}_i)$ that do not contain all of $C^G$. This implies

$$\begin{aligned}
\operatorname{tot deg}(\tilde{\boldsymbol{w}}) &= \operatorname{tot deg}(\boldsymbol{h}) + \operatorname{tot deg}(\boldsymbol{m}_1) + \cdots + \operatorname{tot deg}(\boldsymbol{m}_n) \\
&\geq (n-h)\big(\tfrac{1}{2}(|G|-h+1)+h\big) + h\left(k+h-1+\sum_{i=1}^{k}\left\lfloor \tfrac{1}{2}(\ell_1-1)\right\rfloor\right),
\end{aligned}$$

contradicting the inequality above. To finish, let $x \in V_G$. Then $x\tilde{\boldsymbol{w}} = (\boldsymbol{m})^t x\boldsymbol{w} \in J_G^{t+h+2}$, since $x\boldsymbol{w} \in J_G^{h+2}$, as we showed in the proof of Theorem 3.2, and $\boldsymbol{m} \in J_G$ by assumption. □

We conclude the paper with the following:

**3.4. Corollary.** *For every integer $d$ there exists an ideal $I_d \subset \mathsf{k}[x_1, \ldots, x_d]$ such that* $\operatorname{astab}(I_d) = d - 3$.

*Proof.* Consider the $h$-wheel with $h = d - 5$ such that the graph induced on $R^G$ is a 5-cycle. Theorems 3.2 and 3.3 show that $\operatorname{astab}(I_d) = d - 5 + 2 = d - 3$. □

## Acknowledgements

We would like to acknowledge the support of the NSF grant number 1358454 and thank the referee for the useful comments that improved the exposition of the paper.

## References

[Brodmann 1979] M. Brodmann, "Asymptotic stability of $\operatorname{Ass}(M/I^nM)$", *Proc. Amer. Math. Soc.* **74**:1 (1979), 16–18. MR Zbl

[Bruns and Herzog 1993] W. Bruns and J. Herzog, *Cohen–Macaulay rings*, Cambridge Studies in Advanced Mathematics **39**, Cambridge University Press, 1993. MR Zbl

[Chen et al. 2002] J. Chen, S. Morey, and A. Sung, "The stable set of associated primes of the ideal of a graph", *Rocky Mountain J. Math.* **32**:1 (2002), 71–89. MR Zbl

[Eisenbud 1995] D. Eisenbud, *Commutative algebra: with a view toward algebraic geometry*, Graduate Texts in Mathematics **150**, Springer, 1995. MR Zbl

[Francisco et al. 2010] C. A. Francisco, H. T. Hà, and A. Van Tuyl, "Associated primes of monomial ideals and odd holes in graphs", *J. Algebraic Combin.* **32**:2 (2010), 287–301. MR Zbl

[Francisco et al. 2011] C. A. Francisco, H. T. Hà, and A. Van Tuyl, "Colorings of hypergraphs, perfect graphs, and associated primes of powers of monomial ideals", *J. Algebra* **331** (2011), 224–242. MR Zbl

[Herzog and Asloob Qureshi 2015] J. Herzog and A. Asloob Qureshi, "Persistence and stability properties of powers of ideals", *J. Pure Appl. Algebra* **219**:3 (2015), 530–542. MR Zbl

[Hoa 2006] L. T. Hoa, "Stability of associated primes of monomial ideals", *Vietnam J. Math.* **34**:4 (2006), 473–487. MR Zbl

[Kesting et al. 2011] K. Kesting, J. Pozzi, and J. Striuli, "On the associated primes of the third order of the cover ideal", *Involve* **4**:3 (2011), 263–270. MR Zbl

[Simis et al. 1994] A. Simis, W. V. Vasconcelos, and R. H. Villarreal, "On the ideal theory of graphs", *J. Algebra* **167**:2 (1994), 389–416. MR Zbl

[Van Tuyl 2013] A. Van Tuyl, "A beginner's guide to edge and cover ideals", pp. 63–94 in *Monomial ideals, computations and applications*, edited by A. M. Bigatti et al., Lecture Notes in Math. **2083**, Springer, 2013. MR Zbl

Received: 2017-07-02 Revised: 2018-03-12 Accepted: 2018-05-22

cbrooke@uoregon.edu *St. Olaf College, Northfield, MN, United States*
*Current address:* *University of Oregon, Eugene, OR, United States*
mhoch@wellesley.edu *Wellesley College, Wellesley, MA, United States*
smlato@uwaterloo.ca *Carthage University, Kenosha, WI, United States*
*Current address:* *University of Waterloo, Waterloo, ON, Canada*
jstriuli@fairfield.edu *Department of Mathematics, Fairfield University, Fairfield, CT, United States*
bryan.wang@berkeley.edu *University of California, Berkeley, CA, United States*

INVOLVE 12:3 (2019)
dx.doi.org/10.2140/involve.2019.12.427

# An elliptic curve analogue to the Fermat numbers

Skye Binegar, Randy Dominick,
Meagan Kenney, Jeremy Rouse and Alex Walsh

(Communicated by Bjorn Poonen)

The Fermat numbers have many notable properties, including order universality, coprimality, and definition by a recurrence relation. We use rational points of infinite order on elliptic curves to generate sequences that are analogous to the Fermat numbers. We demonstrate that these sequences have many of the same properties as the Fermat numbers, and we discuss results about the prime factors of sequences generated by specific curves and points.

## 1. Introduction

In August 1640, Fermat wrote a letter to Frénicle [Fermat 1894, p. 205] recounting his discovery that if $n$ is not a power of 2, then $2^n + 1$ is composite. Fermat also stated that if $n$ is a power of 2, then $2^n + 1$ is prime. As examples, he listed the first seven numbers in this sequence, $F_n = 2^{2^n} + 1,\ n \geq 0$, now called the sequence of Fermat numbers.

In 1732, Euler discovered that Fermat's observation was incorrect, and that 641 divides $F_5 = 4294967297$. Indeed, it is now known that $F_n$ is composite for $5 \leq n \leq 32$. Very little is known about whether any $F_n$ are prime; heuristics suggest that only finitely many of them are prime. However, mathematicians have been unable to prove that there are infinitely many composite Fermat numbers.

The primality of the Fermat numbers is connected with the classical problem of constructing a regular polygon with $n$ sides using only an unmarked straightedge and a compass. In 1801, Gauss proved that if a positive integer $n$ is a power of 2 multiplied by a product of distinct Fermat primes, then a regular $n$-gon is constructible with a ruler and compass. The converse of this result was proven by Wantzel in 1837. (For a modern proof, see [Dummit and Foote 2004, p. 602].)

Elliptic curves are central objects in modern number theory and have led to novel methods of factoring [Lenstra 1987b], proofs that numbers are prime [Atkin

*MSC2010:* primary 11G05; secondary 11B37, 11G15, 11Y11.
*Keywords:* elliptic curves, Fermat numbers, duplication formula.
All authors were supported by NSF grant DMS-1461189.

and Morain 1993], and cryptography [Koblitz 1987; Miller 1986]. They have also played a role in a number of important theoretical developments, the most spectacular of which is the "modular method" that led to the solution of Fermat's last theorem [Wiles 1995]. Other such developments include the determination of all integer solutions to $x^2 + y^3 = z^7$ with $\gcd(x, y, z) = 1$ [Poonen et al. 2007] and the determination of all perfect powers in the Fibonacci sequence [Bugeaud et al. 2006]. The present paper relies on both elliptic curves and the sequence of Fermat numbers. We work with elliptic curves in the form $E : y^2 = x^3 + ax^2 + bx + c$. We begin with our central definition:

**Definition 1.** For an elliptic curve $E$ and a point $P \in E(\mathbb{Q})$ of infinite order, let $2^k P = (m_k/e_k^2, n_k/e_k^3)$ denote $P$ added to itself $2^k$ times under the group law on $E(\mathbb{Q})$. Here $m_k, n_k, e_k \in \mathbb{Z}$ with $e_k \geq 1$ and $\gcd(m_k, e_k) = \gcd(n_k, e_k) = 1$. We define the sequence of *elliptic Fermat numbers* $\{F_k(E, P)\}$ by $F_k(E, P) = n_k$.

Fermat's observation that if $n$ is not a power of 2, then $2^n + 1$ is not prime can be explained as follows. If $b$ is an odd divisor of $n$, and $q$ is a prime divisor of $2^{n/b} + 1$, then $2^{n/b} \equiv -1 \pmod{q}$ (so $2^{n/b}$ has order 2 in $\mathbb{F}_p^\times = \mathbb{G}_m(\mathbb{F}_p)$). Then $2^n \equiv (-1)^b \equiv -1 \pmod{q}$ and so $q \mid 2^n + 1$. Since $q \leq 2^{n/b} + 1 < 2^n + 1$, the number $2^n + 1$ cannot be prime.

We are essentially replacing $\mathbb{G}_m$ with an elliptic curve $E$. If $P \in E(\mathbb{Q})$ is a point on $E$, $p$ is a prime of good reduction for $E$, and $nP = (a_n/b_n^2, c_n/b_n^3)$, then $nP \in E(\mathbb{F}_p)$ has order 2 if and only if the $y$-coordinate of $nP$ reduces to 0 mod $p$, that is, $p \mid c_n$. As above, if $b$ is an odd divisor of $n$ and there is a prime $q$ of good reduction for $E$ so that $q \mid |c_{n/b}|$, then $q \mid c_n$. It follows that $c_n$ cannot be prime unless $|c_{n/b}| = |c_n|$, or all prime factors of $c_{n/b}$ are in $S$, the set of primes of bad reduction for $E$.

The growth rate of the numbers $c_n$ implies that $|c_{n/b}| = |c_n|$ for only finitely many $n$. The group law on $E$ implies that if all prime factors of $c_{n/b}$ are in $S$, then $2(n/b)P$ is an $S$-integral point, of which there are only finitely many on $E$ (and in some cases, none).

It follows that possibilities for $c_n$ to be prime when $n$ has an odd divisor are very constrained. For this reason, we choose to focus on the case where $n$ does not have any odd divisors, namely when $n$ is a power of 2. This leads directly to our definition of elliptic Fermat numbers above.

Our goal is to show that the sequence $\{F_k(E, P)\}$ strongly resembles the classic Fermat sequence. We do so by adapting properties of the classic Fermat numbers and proving that they hold for the elliptic Fermat numbers. It is well known, for example, that any two distinct classic Fermat numbers are relatively prime, as Goldbach proved in a 1730 letter to Euler. The elliptic Fermat numbers have a similar property:

**Theorem 2.** *For all $k \neq \ell$, if $p$ is a prime that divides* $\gcd(F_k(E, P), F_\ell(E, P))$, *then $p$ is a prime of bad reduction for $E : y^2 = x^3 + ax^2 + bx + c$.*

The classic Fermat numbers also have the useful property that for any non-negative integer $N$, 2 has order $2^{k+1}$ in $(\mathbb{Z}/N\mathbb{Z})^\times$ if and only if $N \mid F_0 \cdots F_k$ and $N \nmid F_0 \cdots F_{k-1}$. This property, which we call *order universality*, provides a powerful connection between order and divisibility. A close parallel applies to the elliptic Fermat numbers:

**Theorem 3.** *Let $\Delta(E)$ be the discriminant of $E$ and suppose that $N$ is a positive integer with* $\gcd(N, 6\Delta(E)) = 1$. *Then $P$ has order $2^{k+1}$ in $E(\mathbb{Z}/N\mathbb{Z})$ if and only if $N \mid F_0(E, P) \cdots F_k(E, P)$ and $N \nmid F_0(E, P) \cdots F_{k-1}(E, P)$.*

In the case where $N = p$ for some odd prime $p$, we can make this statement stronger. For the classic Fermat numbers, we know that 2 has order $2^{k+1}$ in $\mathbb{F}_p^\times$ if and only if $p \mid F_k$. The elliptic Fermat numbers yield the following result:

**Corollary 4.** *For any odd prime $p \nmid 6\Delta(E)$, $P$ has order $2^{k+1}$ in $E(\mathbb{F}_p)$ if and only if $p \mid F_k(E, P)$.*

This corollary plays a role in several important results in the paper.

Additionally, and quite interestingly, the classic Fermat numbers can be defined by several different recurrence relations. In Section 4, we present the following analogous result:

**Theorem 5.** *Let $E : y^2 = x^3 + ax^2 + bx + c$ be an elliptic curve, and let $P \in E(\mathbb{Q})$ be a point of infinite order. There is a sequence of integers $\{\tau_k\}$ so that*

$$m_k(E, P) = \frac{1}{\tau_k^2}(m_{k-1}^4 - 2bm_{k-1}^2 e_{k-1}^4 - 8cm_{k-1}e_{k-1}^6 + b^2 e_{k-1}^8 - 4ace_{k-1}^8), \quad (1)$$

$$F_k(E, P) = \frac{1}{\tau_k^3}\big(-2am_{k-1}m_k e_{k-1}^2\tau_k^2 - 4bm_{k-1}e_{k-1}^4 F_{k-1}^2 - bm_k e_{k-1}^4\tau_k^2 - 8ce_{k-1}^6 F_{k-1}^2 + 4m_{k-1}^3 F_{k-1}^2 - 3m_{k-1}^2 m_k\tau_k^2\big), \quad (2)$$

$$e_k(E, P) = \frac{1}{\tau_k}(2F_{k-1}e_{k-1}). \quad (3)$$

Unlike the various classic Fermat recurrence relations, which only depend on previous terms, the elliptic Fermat recurrence relation we have discovered relies on several other sequences of integers, namely $m_k$, $e_k$, and $\tau_k$.

This equation follows naturally from the definition of $F_k(E, P)$ and the duplication formula, which we will see in Section 2. In order to have a true recurrence relation, however, we need a way to explicitly calculate $|\tau_k|$. Luckily, we know the following fact:

**Theorem 6.** *The $|\tau_k|$ are eventually periodic, and there is an algorithm to compute $|\tau_k|$ for all $k$.*

In Section 5, we address one of the most famous aspects of the classic Fermat numbers: the question of their primality. Whereas the primality of the Fermat numbers remains an open question, the following result gives conditions under which the elliptic Fermat numbers are always composite. In this result, "the egg" refers to the nonidentity component of the real points of the elliptic curve:

**Theorem 7.** *For an elliptic curve* $E : y^2 = x^3 + ax^2 + bx$, *assume the following*:

(i) $E(\mathbb{Q}) = \langle P, T \rangle$, *where* $P$ *has infinite order and* $T = (0, 0)$ *is a rational point of order* 2.

(ii) $E$ *has an egg.*

(iii) $T$ *is on the egg.*

(iv) $T$ *is the only integral point on the egg.*

(v) $P$ *is not integral.*

(vi) $\gcd(b, m_0) = 1$.

(vii) *The equation* $x^4 + ax^2y^2 + by^4 = \pm 1$ *has no integer solutions where* $y \notin \{0, \pm 1\}$.

*Then* $F_k(E, P)$ *is composite for all* $k \geq 1$.

**Remark.** There are many theorems in the literature about the compositeness of coordinates of rational points on elliptic curves that are in the image of an isogeny; see for example the main theorem of [Everest et al. 2004], and Theorem 1.4 of [Everest et al. 2008]. One feature of the result above in contrast with others is that we give an explicit set of conditions which guarantees that $F_k$ is composite for all $k$.

**Remark.** We wish to note that given a rank-1 curve $E$ and a point $P \in E(\mathbb{Q})$, there is an algorithm that can check whether the conditions in the theorem are satisfied. The condition that $x^4 + ax^2y^2 + by^4 = 1$ has no integer solutions where $y \notin \{0, \pm 1\}$ can also be checked with finitely many calculations, as this is a Thue equation. Such an equation has finitely many solutions [Thue 1909], and the solutions can be found effectively [Tzanakis and de Weger 1989].

There are choices of $E$ for which all seven of the above conditions are satisfied. For example, we can take $E : y^2 = x^3 - 199x^2 - x$. Note that $\Delta(E)$ is positive and thus $E$ has an egg [Silverman 1994, p. 420]. The only integral point on the curve is $T = (0, 0)$, which must be on the egg because 0 is in-between the $x$-coordinates of the other two roots of the polynomial. Also, $2T = (0 : 1 : 0)$ and thus $T$ is a rational point of order 2 on $E$. The generating point of the curve is $P = \left(\frac{2809}{9}, \frac{89623}{27}\right)$, and $\gcd(-1, 2809) = 1$. Finally, Magma [Bosma et al. 1997] can be used to solve Thue equations in order to conclude that there are no integer solutions to $x^4 - 199x^2y^2 - y^4 = \pm 1$ where $y \notin \{0, \pm 1\}$. Thus this example satisfies the conditions for the theorem, and so $F_k$ is composite for all $k$.

Section 6 focuses on the growth rate of the elliptic Fermat numbers. Much like the classic Fermat numbers, the elliptic Fermat numbers grow at a doubly exponential rate:

**Theorem 8.** *Let $F_k$ be the $k$-th elliptic Fermat number in the sequence generated by the elliptic curve $E$ and the point $P = (m_0/e_0^2, n_0/e_0^3)$. If $\hat{h}(P)$ denotes the canonical height of $P$, then*

$$\lim_{k\to\infty} \frac{\log(F_k)}{4^k} = \tfrac{3}{2}\hat{h}(P).$$

The proof is straightforward and is based on the properties of the $\{\tau_k\}$ sequence and the theory of height functions.

Finally, in Section 7, we examine the curve $E : y^2 = x^3 - 2x$ and the elliptic Fermat sequence generated by the point $P = (2, 2)$. It is a theorem of Lucas that a prime divisor of the Fermat sequence is congruent to 1 mod $2^{n+2}$. Upon examination of the factorization of the numbers in the sequence $\{F_n(E, P)\}$, we arrive at a pleasing congruence analogue:

**Theorem 9.** *Let $E : y^2 = x^3 - 2x$ and consider the point $P = (2, 2)$ and the elliptic Fermat sequence $(F_n(E, P))$. For any prime $p$ such that $p \mid F_n(E, P)$ for some $n$, we have*

$$p \equiv \begin{cases} 1 \pmod{2^{n+1}} & \text{if } p \equiv 1 \pmod 4, \\ -1 \pmod{2^{n+1}} & \text{if } p \equiv -1 \pmod 4. \end{cases}$$

In addition to this congruence result, we have a partial converse that tells us about the presence of Fermat and Mersenne primes in $(F_n(E, P))$:

**Theorem 10.** *For $E : y^2 = x^3 - 2x$, consider the point $P = (2, 2)$. Let $F_k = 2^{2^k} + 1$ be a Fermat prime and $F_k \neq 5, 17$. Then $F_k$ divides $F_n(E, P)$ for some $n \leq 2^{k-1} - 2$.*

**Theorem 11.** *For $E : y^2 = x^3 - 2x$, consider the point $P = (2, 2)$. Let $q = 2^p - 1 \geq 31$ be a Mersenne prime. Then $q$ divides $F_n(E, P)$ for some $n \leq p - 4 \in \mathbb{N}$.*

## 2. Background

We begin with some general background on elliptic curves. For the purposes of this paper, an elliptic curve is a nonsingular cubic curve defined over $\mathbb{Q}$ that has the form $y^2 = x^3 + ax^2 + bx + c$ for some $a, b, c \in \mathbb{Z}$. When we say $E$ is nonsingular, we mean that there are no singular points on the curve. We will often think of $E$ as living in $\mathbb{P}^2$ and represent it with the homogeneous equation $y^2z = x^3 + ax^2z + bxz^2 + cz^3$. A *singular point* is a point $P = (x : y : z)$ at which there is not a well-defined tangent line. These points occur when the following equations are equal to 0:

$$F(x, y, z) = y^2z - x^3 - ax^2z - bx^2z - cz^3,$$
$$\frac{\partial F}{\partial x} = -3x^2 - 2azx - bz^2, \quad \frac{\partial F}{\partial y} = 2yz, \quad \frac{\partial F}{\partial z} = y^2 - ax^2 - 2bxz - 3cz^2. \tag{4}$$

We write $E(\mathbb{Q})$ to denote the set of rational points on $E$ along with the point at infinity, $(0:1:0)$. Using the following binary operation, we can give $E(\mathbb{Q})$ a group structure: for $P, Q \in E(\mathbb{Q})$, draw a line through $P$ and $Q$ and let $R = (x, y)$ be the third intersection point of the line with the curve. Then $P + Q = (x, -y)$. This operation gives an abelian group structure on $E(\mathbb{Q})$ with $(0:1:0)$ as the identity.

Any $P \in E(\mathbb{Q})$ can be expressed in projective space as $P = (m/e^2 : n/e^3 : 1) = (me : n : e^3)$ for some $m, n, e \in \mathbb{Z}$ with $\gcd(m, e) = \gcd(n, e) = 1$. From this, there is a well-defined map from $E(\mathbb{Q}) \to E(\mathbb{F}_p)$ that takes $(me : n : e^3)$ to $(me \bmod p : ne \bmod p : e^3 \bmod p)$; this map is a homomorphism if $E/\mathbb{F}_p$ is nonsingular. We have $P \equiv (0:1:0) \pmod{p}$ if and only if $p \mid e$.

Let $\mathbb{Q}_p$ be the field of $p$-adic numbers. The following sets are subgroups of $E(\mathbb{Q}_p)$:

$$\begin{aligned} E_0(\mathbb{Q}_p) &= \{P \in E(\mathbb{Q}_p) \mid P \text{ reduces to a nonsingular point}\}, \\ E_1(\mathbb{Q}_p) &= \{P \in E(\mathbb{Q}_p) \mid P \text{ reduces to } (0:1:0) \bmod p\}. \end{aligned} \tag{5}$$

We have $E_1(\mathbb{Q}_p) \subseteq E_0(\mathbb{Q}_p) \subseteq E(\mathbb{Q}_p)$, and the index $[E(\mathbb{Q}_p) : E_0(\mathbb{Q}_p)]$ is finite and is called the *Tamagawa number* of $E$ at $p$.

The *discriminant* of an elliptic curve $E$ is defined as

$$\Delta(E) = 64a^3c + 16a^2b^2 + 288abc - 64b^3 - 432c^2.$$

The set $E(\mathbb{R})$ can have one or two components depending on whether or not $\Delta(E) < 0$ or $\Delta(E) > 0$ [Silverman 1994, p. 420]. We refer to the connected component of the identity as the *nose*. If there is a second component, we refer to it as the *egg*. For a curve with two components, let $P_{\text{egg}}$, $Q_{\text{egg}}$ be points on the egg, and let $P_{\text{nose}}$, $Q_{\text{nose}}$ be points on the nose. Then $P_{\text{egg}} + Q_{\text{egg}}$ and $P_{\text{nose}} + Q_{\text{nose}}$ are on the nose, while $P_{\text{egg}} + P_{\text{nose}} = P_{\text{nose}} + P_{\text{egg}}$ is on the egg.

Since our definition of the elliptic Fermat numbers involves doubling points, it is convenient to use the notation $2^k P = (m_k/e_k^2, n_k/e_k^3)$. We also rely on the *duplication formula* expressing the $x$-coordinate of $2Q$ in terms of that of $Q$. In particular, if $2^{k-1}P = (x_{k-1}, y_{k-1})$, [Silverman and Tate 1992, p. 39] gives

$$X(2^k P) = \frac{x_{k-1}^4 - 2bx_{k-1}^2 - 8cx_{k-1} + b^2 - 4ac}{4(x_{k-1}^3 + ax_{k-1}^2 + bx_{k-1} + c)}.$$

Letting $2^{k-1}P = (m_{k-1}/e_{k-1}^2, n_{k-1}/e_{k-1}^3)$, we can put this in terms of $m_{k-1}$, $e_{k-1}$, and $n_{k-1}$:

$$X(2^k P) = \frac{m_{k-1}^4 - 2bm_{k-1}^2 e_{k-1}^4 - 8cm_{k-1}e_{k-1}^6 + b^2 e_{k-1}^8 - 4ace_{k-1}^8}{4n_{k-1}^2 e_{k-1}^2}. \tag{6}$$

We will refer to the unreduced numerator and denominator in the above equation as $A$ and $B$, respectively; i.e.,

$$A = m_{k-1}^4 - 2bm_{k-1}^2 e_{k-1}^4 - 8cm_{k-1}e_{k-1}^6 + b^2 e_{k-1}^8 - 4ace_{k-1}^8, \tag{7}$$

$$B = 4n_{k-1}^2 e_{k-1}^2. \tag{8}$$

One last aspect of elliptic curves that will prove useful in Section 7 is the concept of complex multiplication. We say that an elliptic curve has *complex multiplication* if its endomorphism ring is isomorphic to an order in an imaginary quadratic field. In other words, $E$ is equipped with more maps than simple integer multiplication of a point, and composition of these maps is similar to multiplication in an imaginary quadratic field.

Complex multiplication is relevant to our work because it allows us to count the points on the curve over finite fields. In the final section, we will study the curve $E : y^2 = x^3 - 2x$, and our results rely on having a good understanding of $|E(\mathbb{F}_p)|$. As a special case of Proposition 8.5.1 from [Cohen 2007, p. 566], we have the following fact about our curve $E$:

**Proposition 12.** *Let $E : y^2 = x^3 - 2x$ be an elliptic curve and let $p$ be an odd prime. Then $|E(\mathbb{F}_p)| = p + 1 - a_p(E)$, where $a_p(E)$ is known as the **trace of Frobenius** of an elliptic curve modulo $p$. When $p \equiv 3 \pmod 4$, we have $a_p(E) = 0$. If $p \equiv 1 \pmod 4$, then*

$$a_p(E) = 2\left(\frac{2}{p}\right)\begin{cases} -a, & \text{if } 2^{(p-1)/4} \equiv 1 \pmod p, \\ a, & \text{if } 2^{(p-1)/4} \equiv -1 \pmod p, \\ -b, & \text{if } 2^{(p-1)/4} \equiv -a/b \pmod p, \\ b, & \text{if } 2^{(p-1)/4} \equiv a/b \pmod p, \end{cases}$$

*where $a$ and $b$ are integers such that $p = a^2 + b^2$ with $a \equiv -1 \pmod 4$.*

## 3. Coprimality and order universality

We begin by proving Corollary 4 and then use this to prove Theorem 2, that is, $\gcd(F_k(E, P), F_\ell(E, P))$ can only be a multiple of primes of bad reduction.

*Proof of Corollary 4.* If $p \nmid \Delta(E)$, then $p$ is a prime of good reduction for $E$. We have $p \mid F_k(E, P)$ if and only if $2^k P$ reduces modulo $p$ to a nonsingular point with $y \equiv 0 \pmod p$. This occurs if and only if $2^{k+1} P \equiv (0 : 1 : 0) \pmod p$ and since $2^k P \not\equiv (0 : 1 : 0) \pmod p$ it follows that the order of $P \in E(\mathbb{F}_p)$ is $2^k$. □

Now, we prove Theorem 2.

*Proof.* Suppose that $p$ is a prime that divides $\gcd(F_k(E, P), F_\ell(E, P))$. If $p$ is a prime of good reduction for $E$, the previous corollary gives that $p \mid F_k(E, P)$

implies that $P \in E(\mathbb{F}_p)$ must have order exactly $2^{k+1}$, and $p \mid F_\ell(E, P)$ implies that $P \in E(\mathbb{F}_p)$ must have order exactly $2^{\ell+1}$. This is a contradiction if $k \neq \ell$. □

Note that the nonsingularity of $E$ mod $p$ is necessary in both of the proofs above. If $E : y^2 = x^3 + x^2 + 67x + 79$, then $E$ is singular mod 43. The point $P = (10, 43) \in E(\mathbb{Q})$ has infinite order and $2P = \left(-\frac{3}{4}, \frac{43}{8}\right)$ has the property that $P$ and $2P$ (and in fact $2^k P$ for all $k \geq 1$) reduce to a singular point modulo 43, because $P \notin E_0(\mathbb{Q}_{43})$ and the Tamagawa number of $E$ at 43 is 3. It follows that $F_k(E, P)$ is a multiple of 43 for all $k$.

To embark on the proof of Theorem 3, we must make sense of reducing points on an elliptic curve modulo an arbitrary integer $N$, and for this reason we need to recall some results from the theory of elliptic curves over arbitrary rings. Our treatment comes from that of [Lenstra 1987a]. Given a commutative ring $R$, we say that a finite collection of elements $(a_i)$ is *primitive* if it generates $R$ as an $R$-ideal. That is, $(a_i)$ is primitive if there exist $b_i \in R$ such that $\sum a_i b_i = 1$.

Lenstra showed that there is a natural way to define a group structure on the points on $E$ in $\mathbb{P}^2(R)$ provided $6\Delta(E)$ is a unit in $R$, and for any primitive $m \times n$ matrix with entries in $R$ whose $2 \times 2$ subdeterminants are all zero, there exists a linear combination of the rows that is primitive in $R$. This second condition holds in any finite ring and also in any PID, and so Lenstra's construction works in $\mathbb{Z}/N\mathbb{Z}$ if $\gcd(6\Delta(E), N) = 1$.

Given points $S = (x_1 : y_1 : z_1)$ and $T = (x_2 : y_2 : z_2)$ in $E(\mathbb{Z}/N\mathbb{Z})$, Lenstra described *three* families of polynomials in the six variables $(x_1, y_1, z_1, x_2, y_2, z_2)$ such that $S + T$ can be given by any of $(q_1 : r_1 : s_1)$, $(q_2 : r_2 : s_2)$, $(q_3 : r_3 : s_3)$, provided one of these points is primitive. Lenstra showed that the $3 \times 3$ matrix made with the polynomials as its entries has vanishing $2 \times 2$ subdeterminants, and is primitive. It follows that some linear combination $(q_0 : r_0 : s_0)$ of the rows gives a formula for $S + T$ in $E(\mathbb{Z}/N\mathbb{Z})$. This construction works not just over $\mathbb{Z}/N\mathbb{Z}$, but also over $R = \mathbb{Z}[1/(6|\Delta(E)|)]$ and gives $E$ the structure of a group scheme over this ring. It follows from Proposition 3.2 of Chapter IV of [Silverman 1994] that the reduction map $E(R) \to E(\mathbb{Z}/N\mathbb{Z})$ is a homomorphism. By thinking of a point in $E(\mathbb{Q})$, namely $(m/e^2, n/e^3)$ as $(me : n : e^3) \in E(R)$, we get that the reduction mod $N$ map $E(\mathbb{Q}) \to E(\mathbb{Z}/N\mathbb{Z})$ is a homomorphism. It is worth noting that $(m/e^2, n/e^3)$ reduces to $(0 : 1 : 0)$ modulo $N$ if and only if $e \equiv 0 \pmod{N}$. From this, it follows that if $N = p_1^{e_1} p_2^{e_2} \cdots p_k^{e_k}$, then the natural map

$$E(\mathbb{Z}/N\mathbb{Z}) \to \prod_{i=1}^{k} E(\mathbb{Z}/p_i^{e_i}\mathbb{Z})$$

is an isomorphism. Now we prove Theorem 3.

*Proof.* Let $P \in E(\mathbb{Q})$ be a point of infinite order and $k$ a nonnegative integer. Recall that we define $2^k P = (m_k e_k : n_k : e_k^3)$ for $m_k, n_k, e_k \in \mathbb{Z}$ with $\gcd(m_k, e_k) =$

$\gcd(n_k, e_k) = 1$. We consider first the case where $N = p^r$ is an odd prime power. In that situation, we have that if $p^r \mid F_k(E, P)$, then $2^k P \equiv (x : 0 : 1) \pmod{p^r}$ and so the order of $P$ in $E(\mathbb{Z}/p^r\mathbb{Z})$ is $2^{k+1}$. Conversely, if the order of $P \in E(\mathbb{Z}/p^r\mathbb{Z})$ is $2^{k+1}$, then $e_{k+1}$ is a multiple of $p^r$. However, the duplication formula shows that $e_{k+1} \mid 2n_k e_k$. Since $2^k P = (m_k e_k : n_k : e_k^3)$ has order 2 in $E(\mathbb{Z}/p^r\mathbb{Z})$, it also has order 2 in $E(\mathbb{Z}/p\mathbb{Z})$ and so $p \mid n_k$, which implies that $p \nmid e_k$. Thus, $p^r \mid 2n_k e_k$ but $\gcd(p, e_k) = 1$ and so $p^r \mid n_k = F_k(E, P)$. Theorem 2 gives that $N \mid F_k(E, P)$ if and only if $N \mid F_0(E, P)F_1(E, P)\cdots F_k(E, P)$ but $N \nmid F_0(E, P)F_1(E, P)\cdots F_{k-1}(E, P)$. The desired result follows.

Now, we consider the general case. If $N = \prod_{i=1}^{\ell} p_i^{e_i}$, we have the isomorphism

$$E(\mathbb{Z}/N\mathbb{Z}) \cong \prod_{i=1}^{\ell} E(\mathbb{Z}/p_i^{e_i}\mathbb{Z}).$$

It follows from this that $P$ has order equal to $2^{k+1}$ in $E(\mathbb{Z}/N\mathbb{Z})$ if and only if (i) for all prime powers $p_i^{e_i}$ the order of $P$ in $E(\mathbb{Z}/p_i^{e_i}\mathbb{Z})$ is equal to $2^j$ for some $j \le k+1$, and (ii) there is a prime power $p_j^{e_j}$ such that $P \in E(\mathbb{Z}/p_j^{e_j}\mathbb{Z})$ is $2^{k+1}$. Condition (i) means that $p_i^{e_i} \mid F_{j-1}(E, P)$ and condition (ii) means that $p_j^{e_j} \mid F_k(E, P)$ (and hence by Theorem 2 that $p_j \nmid F_\ell(E, P)$ for $\ell < k$). It follows that $P \in E(\mathbb{Z}/N\mathbb{Z})$ has order $2^{k+1}$ if and only if $N \mid F_0(E, P)\cdots F_k(E, P)$ but $N \nmid F_0(E, P)\cdots F_{k-1}(E, P)$. □

## 4. Recurrence

We will now explore the recurrence relation given by Theorem 5. Before continuing, we define the sequence $\{\tau_k\}$. If we write $2^{k-1}P = (m_{k-1}/e_{k-1}^2, F_{k-1}/e_{k-1}^3)$ with $m_{k-1}, F_{k-1}, e_{k-1} \in \mathbb{Z}$ with $e_{k-1} \ge 1$ and $\gcd(m_{k-1}, e_{k-1}) = \gcd(F_{k-1}, e_{k-1}) = 1$, then let

$$\tau_k(E, P) = \frac{2F_{k-1}e_{k-1}}{e_k}.$$

When the duplication formula is applied to compute the $x$-coordinate of $2^k P$, we obtain the formula

$$X(2^k P) = \frac{m_{k-1}^4 - 2bm_{k-1}^2 e_{k-1}^2 - 8cm_{k-1}e_{k-1}^6 + (b^2 - 4ac)e_{k-1}^8}{(2F_{k-1}e_{k-1})^2} = \frac{A}{B} = \frac{m_k}{e_k^2}.$$

Here $(2F_{k-1}e_{k-1})^2 = B$ is the "unreduced" denominator of $X(2^k P)$, and $e_k^2$ is the reduced denominator. So $e_k \mid 2F_{k-1}e_{k-1}$, and the number $\tau_k$ measures the discrepancy between the two quantities $e_k$ and $2F_{k-1}e_{k-1}$, that is, the amount of cancellation that occurs. It is clear then that $\tau_k^2 = \gcd(A, B)$.

We will now prove Theorem 5. For now, keep in mind that we can explicitly calculate $\tau_k$ for all $k$; we will prove this at the end of the section. We can see that

(3) is just a restatement of the definition of $\tau_k$ and (1) is just a restatement of the duplication formula.

**Lemma 13.** *Equation* (2) *is correct.*

*Proof.* From the formulas given in [Silverman 1986, p. 58–59], we can see that

$$Y(2^k P) = \frac{1}{2F_{k-1}e_{k-1}^3 e_k^2}\big(-2am_{k-1}m_k e_{k-1}^4 - bm_{k-1}e_{k-1}^4 e_k^2 \\ -bm_k e_{k-1}^6 - 2ce_{k-1}^6 e_k^2 + m_{k-1}^3 e_k^2 - 3m_{k-1}^2 m_k e_{k-1}^2\big).$$

Then since $Y(2^k P) = F_k/e_k^3$,

$$\begin{aligned} F_k &= Y(2^k P)\cdot e_k^3 \\ &= \frac{1}{2F_{k-1}e_{k-1}^3}\big(-2am_{k-1}m_k e_{k-1}^4 e_k - bm_{k-1}e_{k-1}^4 e_k^3 \\ &\qquad -bm_k e_{k-1}^6 e_k - 2ce_{k-1}^6 e_k^3 + m_{k-1}^3 e_k^3 - 3m_{k-1}^2 m_k e_{k-1}^2 e_k\big). \end{aligned}$$

Then using the fact that $e_k/e_{k-1} = 2F_{k-1}/\tau_k$, we can simplify this to

$$F_k(E, P) = \frac{1}{\tau_k^3}\big(-2am_{k-1}m_k e_{k-1}^2\tau^2 - 4bm_{k-1}e_{k-1}^4 F_{k-1}^2 \\ -bm_k e_{k-1}^4\tau_k^2 - 8ce_{k-1}^6 F_{k-1}^2 + 4m_{k-1}^3 F_{k-1}^2 - 3m_{k-1}^2 m_k \tau_k^2\big). \quad \square$$

We can now see that the recurrence relation is correct, thus proving Theorem 5. The remainder of this section will be devoted to developing a better understanding of $\tau_k$ and developing an algorithm to calculate the sequence.

Ayad [1992] studied the sequences obtained by taking a point $M$ on an elliptic curve, and evaluated the usual division polynomials at $M$ to compute

$$mM = \left(\frac{\phi_m(M)}{\psi_m^2(M)}, \frac{\omega_m(M)}{\psi_m^3(M)}\right).$$

Ayad [1992, Théorème A] proved that if $p$ is a prime, then there is an integer $n$ such that $\phi_n(M)$ and $\psi_n(M)$ both have positive $p$-adic valuation if and only if $M$ is singular modulo $p$, and moreover that in this case $\psi_m(M)$ is a multiple of $P$ for all $m \geq 2$. As a consequence of this, it follows that the only primes that can divide $\tau_k$ are the primes of bad reduction. Also, applying Ayad's theorem with $M = 2^{k-1}P$, if $p$ is an odd prime and $p \mid \tau_k$, then $2^{k-1}P$ is a singular point modulo $p$.

We next wish to obtain more precise information about the power of a prime of bad reduction that can divide $\tau_k$. In particular, for $E : y^2 = x^3 + ax^2 + bx + c$, we define $\Delta(E) = 16(-4a^3c + a^2b^2 + 18abc - 4b^3 - 27c^2)$. The primes for which this model of $E$ has bad reduction are precisely the primes that divide $\Delta(E)$. (We do not assume that $E : y^2 = x^3 + ax^2 + bx + c$ is a global minimal model for $E$.)

**Lemma 14.** *The number* $\tau_k^2$ *divides* $\frac{1}{4}\Delta(E)$.

*Proof.* Let

$$f(x) = x^3 + ax^2 + bx + c,$$
$$F(x) = 3x^3 - ax^2 - 5bx + 2ab - 27c,$$
$$\phi(x) = x^4 - 2bx^2 - 8cx + b^2 - 4ac,$$
$$\Phi(x) = -3x^2 - 2ax + a^2 - 4b.$$

Silverman and Tate [1992, p. 62] showed that $\frac{1}{16}\Delta(E) = f(x)F(x) + \phi(x)\Phi(x)$. Setting $x = X(2^{k-1}P)$, we obtain

$$\tfrac{1}{16}\Delta(E)e_{k-1}^{12}$$
$$= \left(e_{k-1}^6 f\left(\frac{m_{k-1}}{e_{k-1}^2}\right)\right)\left(e_{k-1}^6 F\left(\frac{m_{k-1}}{e_{k-1}^2}\right)\right) + \left(e_{k-1}^4 \Phi\left(\frac{m_{k-1}}{e_{k-1}^2}\right)\right)\left(e_{k-1}^8 \phi\left(\frac{m_{k-1}}{e_{k-1}^2}\right)\right).$$

Recall that $\tau_k^2 = \gcd(A, B)$ where $A$ and $B$ are given by (7) and (8). Rewriting this equation in terms of $A$ and $B$ gives

$$\tfrac{1}{16}\Delta(E)e_{k-1}^{12} = \frac{B}{4e_{k-1}^2}\left(e_{k-1}^6 F\left(\frac{m_{k-1}}{e_{k-1}^2}\right)\right) + \left(e_{k-1}^4 \Phi\left(\frac{m_{k-1}}{e_{k-1}^2}\right)\right)A.$$

Multiplying through by $4e_{k-1}^2$ gives that $A$ and $B$ both divide $\frac{1}{4}\Delta(E)e_{k-1}^{14}$. However, $\gcd(m_{k-1}, e_{k-1}) = 1$ implies that $\gcd(A, e_{k-1}) = 1$ and so $\tau_k^2 = \gcd(A, B)$ is relatively prime to $e_{k-1}$ and so $\tau_k^2 \mid \left(\frac{1}{4}\Delta(E)\right)$, as desired. □

As stated above, Ayad's theorem implies that if $p \mid \tau_k$, then $2^{k-1}P$ is a singular point modulo $p$. We will prove a converse to this result.

**Theorem 15.** *Let $p$ be an odd prime. Suppose that $2^{k-1}P$ and $2^kP$ both reduce to singular points mod $p$. Then $p \mid \tau_k$.*

*Proof.* Since $p$ is odd, singular points modulo $p$ have $y$-coordinate $\equiv 0 \pmod p$ and hence if $2^{k-1}P$ reduces to a singular point modulo $p$, then $p \mid F_{k-1}(E, P)$. On the other hand, $2^kP$ reducing to a singular point modulo $p$ means that $p \nmid e_k$ and hence $p \mid \tau_k = 2F_{k-1}e_{k-1}/e_k$. □

The results above apply for odd primes. Now, we consider the parity of $\tau_k$ and $F_k$.

**Theorem 16.** *If $2^kP \not\equiv (0:1:0) \pmod 2$, then $\tau_k$ is even. If $2^{k-1}P \equiv 2^kP \equiv (0:1:0) \pmod 2$, then $\tau_k$ is odd.*

*Proof.* If $2^kP \not\equiv (0:1:0) \pmod 2$, then $2 \nmid e_k$. Since $\tau_k(E, P) = 2F_{k-1}e_{k-1}/e_k$, the numerator is even and the denominator is odd, so $\tau_k$ is even.

If $2^{k-1}P \equiv 2^k P \equiv (0:1:0) \pmod 2$, then $e_{k-1}$ and $e_k$ are both even, while $m_{k-1}$ and $m_k$ are both odd. Considering the duplication formula

$$\frac{A}{B} = \frac{m_{k-1}^4 - 2bm_{k-1}^2e_{k-1}^2 - 8cm_{k-1}e_{k-1}^4 + (b^2-4ac)e_{k-1}^4}{(2F_{k-1}e_{k-1})^2},$$

one sees that $A$ is odd and $B$ is even, and since $\tau_k^2 = \gcd(A, B)$, it follows that $\tau_k$ is odd. □

Recalling that $E_1(\mathbb{Q}_p)$ denotes the set of points in $E(\mathbb{Q}_p)$ that reduce to the point at infinity modulo $p$, the above theorem gives that $\tau_k$ is even for all sufficiently large $k$ if and only if the order of $P \in E(\mathbb{Q}_2)/E_1(\mathbb{Q}_2)$ is not a power of 2, and $\tau_k$ is odd for all sufficiently large $k$ if and only if the order of $P \in E(\mathbb{Q}_2)/E_1(\mathbb{Q}_2)$ is a power of 2.

While it is nice to know all of these properties, we need to know exactly what $\tau_k$ is in order for the recurrence relations to be useful. In accordance with Theorem 6, we can calculate $|\tau_k|$ for all $k$ using the following algorithm. (The proof of the correctness of the algorithm will be given later in this section.)

(1) Find and factor the discriminant $\Delta(E)$.

(2) For each prime $p$ such that $p^2 \mid \Delta(E)$, complete the following:

(a) Find the smallest $\ell \in \mathbb{Z}^+$ such that $\ell P \equiv (0:1:0) \pmod p$.

(b) If $\ell$ is a power of 2, then $\operatorname{ord}_p(\tau_k) = 0$ for all $k \geq \ell + 1$.

(i) Move on to the next $p^2 \mid \Delta(E)$.

(c) If $\ell$ is not a power of 2, then $\operatorname{ord}_p(\tau_k) = \operatorname{ord}_p(2F_{k-1})$.

(i) Find some $r \in \mathbb{Z}^+$ such that $rP = (m/e^2, n/e^3)$ with $p^s \mid e$. Choose $s$ such that either $p^{2s} \,||\, \Delta(E)$ or $p^{2s+1} \,||\, \Delta(E)$. Here $p^n \,||\, a$ means that the prime power $p^n$ fully divides $a$; that is, $p^n \mid a$ but $p^{n+1} \nmid a$.

(ii) Now $\operatorname{ord}_p(Y(tP))$ depends only on $t \bmod r$. Find all possible values of $2^k \bmod r$ and note the lowest $k$ which generates each value.

(iii) Calculate $\operatorname{ord}_p(F_{k-1})$ for each $k$ noted in (ii). Use this to calculate $\operatorname{ord}_p(\tau_k)$.

(iv) Move on to the next $p^2 \mid \Delta(E)$.

(3) We now know $\operatorname{ord}_p(\tau_k)$ for all (but finitely many, in some cases) $k$ for each $p$ such that $p^2 \mid \Delta(E)$, which are all the $p$ that could divide $\tau_k$. Use this to calculate $|\tau_k|$.

Note that doing the above computations in $E(\mathbb{Q})$ can be challenging since the heights of points on elliptic curves grow quickly. Instead, doing the computations in $E(\mathbb{Q}_p)$, which is implemented in Sage [SageMath 2017], is more straightforward.

Now we will prove that this algorithm is correct. In order to do this, we must first prove the following theorem.

**Theorem 17.** *Let $E : y^2 = x^3 + ax^2 + bx + c$ be an elliptic curve. Assume $Q, R \in E(\mathbb{Q})$ are such that*

$$Q = (x_1, y_1) = \left(\frac{m_1}{e_1^2}, \frac{n_1}{e_1^3}\right), \quad p \nmid e_1,$$

$$R = (x_2, y_2) = \left(\frac{m_2}{e_2^2}, \frac{n_2}{e_2^3}\right), \quad p^k \,||\, e_2.$$

*Let*

$$Q + R = (x_3, y_3) = \left(\frac{m_3}{e_3^2}, \frac{n_3}{e_3^3}\right).$$

*Then*

$$X(Q + R) \equiv X(Q) \pmod{p^k}, \quad Y(Q + R) \equiv Y(Q) \pmod{p^k}.$$

The result above follows from the fact that the natural map from $E(\mathbb{Q}) \to E(\mathbb{Z}/p^k\mathbb{Z})$ is a homomorphism in the case when $p \nmid 6\Delta(E)$, but in light of the algorithm above, we are primarily interested in the case where $p \mid 6\Delta(E)$.

*Proof.* From [Silverman 1986, p. 58–59], we know that if we let

$$\lambda = \frac{y_2 - y_1}{x_2 - x_1} \quad \text{and} \quad v = \frac{y_1 x_2 - y_2 x_1}{x_2 - x_1},$$

then we have

$$x_3 = \lambda^2 - a - x_1 - x_2 = \frac{ax_2^2 + bx_2 + c - 2y_1y_2 + y_1^2 + 2x_1x_2^2 - x_1^2x_2}{x_2^2 - 2x_1x_2 + x_1^2} - a - x_1.$$

Now since $p^k \,||\, e_2$, we can let $x_2 = \tilde{x}_2 p^{-2k}$ and $y_2 = \tilde{y}_2 p^{-3k}$. Plugging this in yields

$$x_3 = \frac{a\tilde{x}_2^2 + b\tilde{x}_2 p^{2k} + cp^{4k} - 2y_1\tilde{y}_2 p^k + y_1^2 p^{4k} + 2x_1\tilde{x}_2^2 - x_1^2\tilde{x}_2 p^{2k}}{\tilde{x}_2^2 - 2x_1\tilde{x}_2 p^{2k} + x_1^2 p^{4k}} - a - x_1. \quad (9)$$

Reducing mod $p^k$ and mod $p^{2k}$ gives us

$$x_3 \equiv x_1 \pmod{p^k}, \quad (10)$$

$$x_3 \equiv x_1 - \frac{2y_1\tilde{y}_2 p^k}{\tilde{x}_2^2} \pmod{p^{2k}}. \quad (11)$$

Now that we have shown that $x_3 \equiv x_1 \pmod{p^k}$, we just need to show that $y_3 \equiv y_1 \pmod{p^k}$. Since $x_3 \equiv x_1 \pmod{p^k}$, we can write $x_3 = x_1 + rp^k$. And again using the definitions of $\lambda$ and $v$ given above, we have

$$y_3 = -\lambda x_3 - v = \frac{-n_1m_1e_2^3 + n_1m_2e_1^2e_2 - n_1e_1^2e_2^3rp^k + n_2e_1^5rp^k}{m_1e_1^3e_2^3 - m_2e_1^5e_2}.$$

Once again, since $p^k \,||\, e_2$, we can let $e_2 = \tilde{e}_2 p^k$. Then

$$y_3 = \frac{-n_1 m_1 \tilde{e}_2^3 p^{2k} + n_1 m_2 e_1^2 \tilde{e}_2 - n_1 e_1^2 \tilde{e}_2^3 r p^{3k} + n_2 e_1^5 r}{m_1 e_1^3 \tilde{e}_2^3 p^{2k} - m_2 e_1^5 \tilde{e}_2}.$$

Reducing mod $p^k$ gives us

$$y_3 \equiv \frac{-n_1}{e_1^3} - \frac{n_2 r}{m_2 \tilde{e}_2} \pmod{p^k}. \tag{12}$$

Now from (11), we know

$$r \equiv -\frac{2 y_1 \tilde{y}_2}{\tilde{x}_2^2} \pmod{p^k}.$$

Simple algebra allows us to see that

$$r \equiv \frac{-2 n_1 n_2 \tilde{e}_2}{m_2^2 e_1^3} \pmod{p^k}.$$

Plugging this into (12), we get

$$\begin{aligned} y_3 &\equiv \frac{-n_1}{e_1^3} - \frac{n_2}{m_2 \tilde{e}_2} \cdot \frac{-2 n_1 n_2 \tilde{e}_2}{m_2^2 e_1^3} \pmod{p^k} \\ &\equiv \frac{-n_1}{e_1^3} + \frac{2 n_1 (m_2^3 + a m_2^2 e_2^2 + b m e_2^4 + c e_2^6)}{m_2^3 e_1^3} \pmod{p^k}. \end{aligned}$$

And since $e_2 \equiv 0 \pmod{p^k}$, we have

$$y_3 \equiv \frac{-n_1}{e_1^3} + \frac{2 n_1 m_2^3}{m_2^3 e_1^3} \pmod{p^k} \equiv y_1 \pmod{p^k}, \tag{13}$$

completing the proof. □

Now we prove that the algorithm to calculate $\tau_k$ is correct.

*Proof.* From Lemma 14, we can conclude that for any $p$ dividing $\tau_k$, we must have $p^2 \mid \Delta(E)$. So we only need to consider primes $p$ which satisfy this condition. We now break this problem into two cases based on the smallest $\ell \in \mathbb{Z}^+$ so that $\ell P \equiv (0:1:0) \pmod{p}$.

Case I: If $\ell$ is a power of 2, then there exists $d \in \mathbb{Z}^+$ such that $2^d P \equiv (0:1:0) \pmod{p}$. First, if $p > 2$, then for $k \geq d+1$, we have $p \nmid F_k$ and since $e_k$ is a multiple of $e_{k-1}$, but $e_k$ is a divisor of $2F_{k-1}e_{k-1}$, it follows that $\mathrm{ord}_p(e_{k-1}) = \mathrm{ord}_p(e_k)$ and so $p \nmid \tau_k$. If $p = 2$, the desired result follows from Theorem 16.

Case II: If $\ell$ is not a power of 2, then $2^k P \not\equiv (0:1:0) \pmod{p}$ for any $k$. This implies that $p \nmid e_k$ for any $k$ and hence $\mathrm{ord}_p(\tau_k) = \mathrm{ord}_p(2F_{k-1})$. Choose $s$ such that either $p^{2s} \,||\, \Delta(E)$ or $p^{2s+1} \,||\, \Delta(E)$. Now, we can find some $r \in \mathbb{Z}^+$ such that

$rP = (m/e^2, n/e^3)$ with $p^s \mid e$. Then $rP \equiv (0:1:0) \pmod{p^s}$. Using Theorem 17, we can see that $jP + rP \equiv jP \pmod{p^s}$ and conclude that $\mathrm{ord}_p(Y(tP))$ depends only on $t \bmod r$. Then, since $2^k \bmod r$ will repeat, we can use a finite number of calculations to determine $\mathrm{ord}_p(Y(2^k P)) = \mathrm{ord}_p(F_k)$ for all $k \geq 1$. □

## 5. Primality

In this section, we prove Theorem 7. This theorem states the following. Suppose that $E: y^2 = x^3 + ax^2 + bx$ is an elliptic curve of rank 1 generated by $P$ with $x$-coordinate $m_0/e_0^2$ and the torsion subgroup of $E(\mathbb{Q}) \cong \mathbb{Z}/2\mathbb{Z}$ generated by $T = (0, 0)$, which lies on the egg and is the only integral point on the egg. Let $F_k(E, P)$ denote the sequence of elliptic Fermat numbers. Suppose that $\gcd(b, m_0) = 1$, and suppose that the Thue equation $x^4 + ax^2y^2 + by^4 = \pm 1$ has no integer solutions with $y \notin \{0, \pm 1\}$. Then all the elliptic Fermat numbers $F_k(E, P)$ are composite.

We start by proving two lemmas that will be useful in the proof of Theorem 7.

**Lemma 18.** *Assume that $E(\mathbb{Q}) \cong \mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ and $E(\mathbb{Q}) = \langle P, T \rangle$, where $P$ is a generator of $E(\mathbb{Q})$ and $T$ is a rational point of order 2. Assume that*:

(i) *$E$ has an egg.*

(ii) *$T$ is on the egg.*

(iii) *$T$ is the only integral point on the egg.*

(iv) *$P$ is not integral.*

*Then $T$ is the only integral point on $E$.*

*Proof.* Every point in $E(\mathbb{Q})$ is of the form $mP$ or $mP + T$. If $P$ is on the nose, then we have that for any $m \neq 0$, $mP$ is on the nose, and $mP$ is not integral because $P$ is not integral. We also have that $mP + T$ is on the egg and thus is not integral because $T$ is the only integral point on the egg by assumption. If $P$ is on the egg, then let $P' = P + T$. Then $P'$ is on the nose, and the proof is the same as before. □

**Lemma 19.** *Let $E$ be an elliptic curve of the form $y^2 = x^3 + ax^2 + bx$ and suppose* $\gcd(m_0, b) = 1$. *Then* $\gcd(m_k, b) = 1$ *for all $k$.*

*Proof.* We use induction. The base case $\gcd(m_0, b) = 1$ is true by assumption. Now assume that $\gcd(m_{k-1}, b) = 1$. Since $c = 0$, from our recurrence relations, we can see that

$$m_k = \frac{m_{k-1}^4 - 2bm_{k-1}^2 e_{k-1}^4 + b^2 e_{k-1}^8}{\tau_k^2}.$$

Now since $b$ divides the terms $-2bm_{k-1}^2 e_{k-1}^4$ and $b^2 e_{k-1}^8$ in the numerator but is coprime to the term $m_{k-1}^4$, we know $b$ is coprime to the numerator. Dividing by $\tau_k^2$ will not change this. Thus $\gcd(m_k, b) = 1$ for all $k$. □

With these two lemmas, we can now prove Theorem 7.

*Proof of Theorem 7.* Let $2^k P = (m_k/e_k^2, F_k/e_k^3)$ and let $2^k P + T = (m_T/e_T^2, n_T/e_T^3)$. Using the formulas for adding points given in [Silverman 1986, p. 58–59], we see

$$X(2^k P + T) = \frac{be_k^2}{m_k}, \quad Y(2^k P + T) = \frac{-bF_k e_k}{m_k^2}. \tag{14}$$

By the assumption that $\gcd(b, m_0) = 1$ and by Lemma 19, we know $\gcd(b, m_k) = 1$. And since $\gcd(m_k, e_k) = 1$, the first equation in (14) must be in lowest terms. This gives $e_T = \sqrt{|m_k|}$ and $n_T = -bF_k e_k/\sqrt{|m_k|}$. We find from this that

$$-\frac{n_T e_T}{be_k} = \frac{(bF_k e_k/\sqrt{|m_k|})\sqrt{|m_k|}}{be_k} = F_k.$$

Note that if $p$ is a prime and $p \mid e_k$ then $2^k P \equiv (0 : 1 : 0) \pmod{p}$, in which case $2^k P + T \equiv T \pmod{p}$. And since $T$ is not the point at infinity, $2^{k-1} P + T \not\equiv (0 : 1 : 0) \pmod{p}$. Therefore $p \nmid e_T$. Hence $\gcd(e_k, e_T) = 1$. Since $e_T = \sqrt{|m_k|}$ and $\gcd(b, m_k) = 1$, we get the factorization

$$F_k = \left(-\frac{n_T}{be_k}\right) e_T,$$

where both factors are integers. Therefore $F_k$ is composite as long as

$$\frac{n_T}{be_k} = -\frac{F_k}{\sqrt{|m_k|}} \neq \pm 1.$$

If we assume that $F_k = \pm\sqrt{|m_k|}$, then $2^k P = (m_k/e_k^2, F_k/e_k^3)$ being a point on $E$ gives $|m_k| = F_k^2 = m_k^3 + am_k^2 e_k^2 + bm_k e_k^4$, which yields $m_k^2 + am_k e_k^2 + be_k^4 = \pm 1$. But by assumption, this equation has no solutions where $e_k \notin \{0, \pm 1\}$. Therefore $F_k$ is composite for all $k \geq 1$. □

## 6. Growth rate

In this section, we will discuss the growth rate of the elliptic Fermat numbers and prove Theorem 8. In order to do so, we need a few more tools. The first new definition we need is the *height* of a point.

**Definition 20.** The *height* of a point $P = (m/e^2, n/e^3)$ on an elliptic curve is defined as

$$h(P) = \log(\max(|m|, e^2)).$$

The height of a point gives us a way to express how "complicated" the coordinates of the point are. We also need to make use of the *canonical height*.

**Definition 21.** The *canonical height* of a point $P$ on an elliptic curve is defined as

$$\hat{h}(P) = \lim_{k\to\infty} \frac{h(2^k P)}{4^k}.$$

Note that Theorem 8 can be summarized as saying that $F_k$ is approximately equal to $e^{4^k\cdot(3/2)\hat{h}(P)}$. So the elliptic Fermat sequences grow doubly exponentially, like the classic Fermat sequence, albeit much more quickly. The proof is as follows:

*Proof of Theorem 8.* First, recall that $|F_k| = |\tau_{k+1}|e_{k+1}/(2e_k)$. This relates the $y$-coordinate of $2^k P$ to the $x$-coordinate. We then have

$$\begin{aligned}
\lim_{k\to\infty} \frac{\log(|F_k(E,P)|)}{4^k} &= \lim_{k\to\infty} \frac{\log(|\tau_{k+1}|e_{k+1}/(2e_k))}{4^k} \\
&= \lim_{k\to\infty} \frac{\frac{1}{2}\log(e_{k+1}^2)}{4^k} - \lim_{k\to\infty} \frac{\frac{1}{2}\log(e_k^2)}{4^k} + \lim_{k\to\infty} \frac{\log\left(\frac{1}{2}|\tau_{k+1}|\right)}{4^k} \\
&= 2\lim_{k\to\infty} \frac{\log(e_{k+1}^2)}{4^{k+1}} - \tfrac{1}{2}\lim_{k\to\infty} \frac{\log(e_k^2)}{4^k} + 0 \\
&= 2\hat{h}(P) - \tfrac{1}{2}\hat{h}(P) = \tfrac{3}{2}\hat{h}(P).
\end{aligned}$$

□

## 7. Elliptic Fermat numbers for the curve $y^2 = x^3 - 2x$

In this section, we apply the hitherto developed theory of elliptic Fermat numbers to examine properties of the curve $E : y^2 = x^3 - 2x$ and the point $P = (2, 2) \in E(\mathbb{Q})$.

We begin with some remarks on $E$ and the point $P$. Recall that $E$ is equipped with complex multiplication and so Proposition 12 gives a formula for $|E(\mathbb{F}_p)|$ for all $p$. Elliptic curves with complex multiplication are the key to the elliptic curve primality proving algorithm of Atkin, Goldwasser, Kilian and Morain, and elliptic curve algorithms to prove primality of Fermat numbers and other special sequences have been considered previously in [Gross 2005; Denomme and Savin 2008; Tsumura 2011; Abatzoglou et al. 2016]. The last remark we make is about the elliptic Fermat sequence $\{F_n(E, P)\}$ and the appearance of Fermat primes and Mersenne primes, i.e, primes of the form $2^p - 1$ for a prime $p$, in the factorization of $F_k(E, P)$.

Table 1 provides factorizations of the first five elliptic Fermat numbers for $E$ at $P$, with known Fermat and Mersenne primes in bold. (The primes $p_6$, $p_7$ have 16 digits each, and $p_8$ and $p_9$ have 18 digits each.) In fact, every odd prime factor dividing $F_n(E, P)$ for $n \geq 2$ will have a congruence that is either Mersenne-like or Fermat-like. We now present the proof of Theorem 9, beginning with the congruence result for a prime divisor $p \equiv -1 \pmod 4$, which yields a tidy Mersenne-like congruence.

*Proof of Theorem 9 for $p \equiv 3 \pmod 4$.* By Theorem 3, $p \mid F_n(E, P)$ tells us that $P$ has order $2^{n+1}$ in $E(\mathbb{F}_p)$. Then by Lagrange's theorem and Proposition 12, $2^{n+1} \mid |E(\mathbb{F}_p)| = p + 1$, and so $p \equiv -1 \pmod{2^{n+1}}$. □

| $n$ | $F_n(E, P)$ |
|---|---|
| 0 | 2 |
| 1 | $-\mathbf{3} \cdot \mathbf{7}$ |
| 2 | $\mathbf{31} \cdot 113 \cdot \mathbf{257}$ |
| 3 | $-2113 \cdot 2593 \cdot 46271 \cdot 101281 \cdot 623013889$ |
| 4 | $\mathbf{127} \cdot \mathbf{65537} \cdot 33303551 \cdot 70639871 \cdot 364024274689 \cdot p_6 \cdot p_7 \cdot p_8 \cdot p_9$ |

**Table 1.** Factorizations of the first five elliptic Fermat numbers for $E$ at $P$.

Proving the congruence in the case of a prime divisor of an elliptic Fermat number congruent to 1 modulo 4 will require multiple steps. We will eventually show that such a prime divisor of $F_n(E, P)$ is congruent to 1 modulo $2^n$, but we begin by showing an initial congruence result:

**Lemma 22.** *Let $E : y^2 = x^3 - 2x$ be an elliptic curve, $P = (2, 2)$ a point of infinite order and $F_n(E, P)$ the $n$-th elliptic Fermat number associated to $E$ at the point $P$. Then for any odd prime divisor $p \equiv 1 \pmod 4$ of $F_n(E, P)$, $n \geq 3$, we have $p \equiv 1 \pmod{\max(2^{\lfloor (n+1)/2 \rfloor}, 8)}$.*

*Proof.* If $p \equiv 1 \pmod 4$, then $p = a^2 + b^2$, where $a \equiv -1 \pmod 4$. Recall that Proposition 12 gives a formula for the value of $|E(\mathbb{F}_p)|$ which depends on the quartic character of 2 modulo $p$. Let us first consider the case where 2 is a fourth power. Then $|E(\mathbb{F}_p)| = p + 1 - 2a$.

Like the proof of the previous theorem, we use Lagrange's theorem to show that $2^{n+1} \mid E(\mathbb{F}_p) = a^2 + b^2 + 1 - 2a = (a-1)^2 + b^2$. So $(a-1)^2 + b^2 \equiv 0 \pmod{2^{n+1}}$. Then $a - 1 \equiv b \equiv 0 \pmod{2^{\lfloor (n+1)/2 \rfloor}}$, giving $p = a^2 + b^2 \equiv 1^2 + 0^2 \pmod{2^{\lfloor (n+1)/2 \rfloor}}$. A symmetric argument follows when 2 is a quadratic residue but not a fourth power. In this situation we arrive at the equation $(a+1)^2 + b^2 \equiv 0 \pmod{2^{n+1}}$; however, the result is precisely the same.

To conclude, we rule out the case where 2 is not a quadratic residue modulo $p$. This would imply $|E(\mathbb{F}_p)| = p + 1 \pm 2b$. The same algebraic manipulation leads to a similar situation where $a^2 + (b \mp 1)^2 \equiv 0 \pmod{2^{n+1}}$, but this means $b \equiv \pm 1 \pmod{2^{\lfloor (n+1)/2 \rfloor}}$; however, $b$ is the even part of the two-square representation of $p$. So it cannot be the case that 2 is not a quadratic residue modulo 8, which happens only when $p \equiv 5 \pmod 8$. □

Because of the lemma, we have $p \equiv 1 \pmod 8$, and so we can make sense of $\sqrt{2}$ and $i$ modulo $p$. We now define the recklessly notated action $i$ on $E(\mathbb{F}_p)$ as $i(x, y) \mapsto (-x, iy)$, where the point $(-x, iy)$ uses $i$ as the square root of $-1$ modulo $p$. This action makes $E(\mathbb{F}_p)$ into a $\mathbb{Z}[i]$-module. We will prove one last lemma concerning the action of $(1 + i)$ before moving on to the full congruence.

**Lemma 23.** *Let $E : y^2 = x^3 - 2x$ be an elliptic curve, $P = (2, 2)$ a point of infinite order and $F_n(E, P)$ the n-th elliptic Fermat number associated to $E$ at the point $P$. Then for any odd prime factor $p \equiv 1 \pmod 4$ of $F_n(E, P)$, $n \geq 3$, we have $(1+i)^{2n+2}P = 0$ in $E(\mathbb{F}_p)$ and $(1+i)^{2n}P \neq 0$.*

*Proof.* Note that $(1+i)^k P = 2^k i^k P$. Recall that $P$ has order $2^{n+1}$, so

$$(1+i)^{2(n+1)}P = (2i)^{n+1}P = i^{n+1}(2^{n+1}P) = i^{n+1} \cdot 0 = 0.$$

It suffices to show that $(1+i)^x P \neq 0$ for $x \leq 2n$. Suppose not, and $(1+i)^x P = 0$. Then certainly $(1+i)^{2n}P = i^n 2^n P = 0$. The action of $i^{n-1}$ makes no difference on the identity. This implies that $2^n P = 0$, contradicting order universality since $P$ has order $2^{n+1}$. □

With this last lemma proven, we are ready to introduce the Fermat-like congruence in full regalia and finish Theorem 9.

*Proof of Theorem 9 for $p \equiv 1 \pmod 4$.* As a consequence of the lemma above, we have that either $(1+i)^{2n+2}P = 0$ or $(1+i)^{2n+1}P = 0$. We are able to bolster the $(2n+1)$-case by introducing a new point $Q = (-i(\sqrt{2}-2), (2-2i)(\sqrt{2}-1))$. It is routine point addition to see that $(1+i)Q = (2, 2) = P$. In either case we have $(1+i)^{2n+3}Q = 0$ and $(1+i)^{2n+1}Q \neq 0$.

Consider the $\mathbb{Z}[i]$-module homomorphism $\phi : \mathbb{Z}[i] \to E(\mathbb{F}_p)$ given by $\phi(x) = xQ$. The image of $\phi$ is $\mathbb{Z}[i]Q = \{(a+bi)Q \mid a, b \in \mathbb{Z}\}$, the orbit of $\mathbb{Z}[i]$ on $Q$. By the first isomorphism theorem, $\mathbb{Z}[i]Q$ is isomorphic to $\mathbb{Z}[i]/\ker(\phi)$. Since $(1+i)^{2n+1} \notin \ker(\phi)$ and $(1+i)^{2n+3} \in \ker(\phi)$, and $(1+i)$ is an irreducible ideal in $\mathbb{Z}[i]$, we know the kernel is either the ideal $((1+i)^{2n+2})$ or $((1+i)^{2n+3})$; hence $\mathbb{Z}[i]/\ker(\phi)$ is a group of size $2^k$, where $k = 2n+2$ or $k = 2n+3$.

Like the previous congruence results, we use Lagrange's theorem to assert $2^k \mid |E(\mathbb{F}_p)|$ and through the same reasoning as before, we arrive at

$$p \equiv 1 \pmod{2^{\lfloor k/2 \rfloor} = 2^{n+1}}.$$ □

We now present the proofs of Theorems 10 and 11, which give us information about sufficiently large Fermat and Mersenne primes dividing the elliptic Fermat sequence $\{F_n(E, P)\}$. First, we provide two lemmas.

**Lemma 24.** *Let $p \equiv \pm 1 \pmod{2^n}$ be an odd prime. Let $\zeta_\ell$ denote a primitive $\ell$-th root of unity in some extension of $\mathbb{F}_p$. Then $\zeta_{2^k} + \zeta_{2^k}^{-1}$ exists in $\mathbb{F}_p$ for all $k \leq n$.*

*Proof.* If $p \equiv 1 \pmod{2^k}$, then clearly there is a primitive $2^k$-th root of unity in $\mathbb{F}_p$.

If $p \equiv 3 \pmod 4$, then we employ methods from Galois theory. First, because $p \equiv -1 \pmod{2^k}$, we have $p^2 \equiv 1 \pmod{2^k}$. Then there is a primitive $2^k$-th root of unity in $\mathbb{F}_{p^2}$. Then we have that $\alpha = \zeta_{2^k} + \zeta_{2^k}^{-1}$ is in $\mathbb{F}_p$ if and only if $\sigma(\alpha) = \alpha$, where $\sigma(x) = x^p$ is the Frobenius endomorphism.

This says that $\alpha \in \mathbb{F}_p$ if and only if

$$\alpha^p = (\zeta_{2^k} + \zeta_{2^k}^{-1})^p = \zeta_{2^k}^p + \zeta_{2^k}^{-p} = \zeta_{2^k} + \zeta_{2^k}^{-1}.$$

We may write this equality as $\zeta_{2^k}^{2p} + \zeta_{2^k}^{p+1} + \zeta_{2^k}^{-p+1} + 1 = 0$. This factors into $(\zeta_{2^k}^p - \zeta_{2^k})(\zeta_{2^k}^p - \zeta_{2^k}^{-1}) = 0$. Then the equality holds if and only if $\zeta_{2^k}^p = \zeta_{2^k}$, meaning $p \equiv 1 \pmod{2^k}$, or $\zeta_{2^k}^p = \zeta_{2^k}^{-1}$; hence $p \equiv -1 \pmod{2^k}$. □

**Lemma 25.** *Let $p$ be a Fermat or Mersenne prime that is at least* 31. *Then there exists a $Q \in E(\mathbb{F}_p)$ such that $2Q = P$.*

*Proof.* From [Silverman and Tate 1992, p. 76], for $E$ we have its isogenous curve $E' : y^2 = x^3 + 8x$ and two homomorphisms, $\phi : E \to E'$ and $\psi : E' \to E$ given by

$$\phi(x, y) = \begin{cases} \left(\dfrac{y^2}{x^2}, \dfrac{y(x^2+2)}{x^2}\right) & \text{if } (x, y) \neq (0:0:1), (0:1:0), \\ (0:1:0) & \text{otherwise,} \end{cases}$$

$$\psi(x, y) = \begin{cases} \left(\dfrac{y^2}{4x^2}, \dfrac{y(x^2-8)}{8x^2}\right) & \text{if } (x, y) \neq (0:0:1), (0:1:0), \\ (0:1:0) & \text{otherwise.} \end{cases}$$

The maps hold the special property $\phi \circ \psi(S) = 2S$. The advantage of this framework is that we are able to break point-halving, a degree-4 affair, into solving two degree-2 problems. Another fact from [Silverman and Tate 1992, p. 85] is that $P = (x, y) \in \psi(E'(\mathbb{Q}))$ if and only if $x$ is a square.

We now use this to show there is a $Q \in E(\mathbb{F}_p)$ such that $2Q = P$. For brevity, let $z = \sqrt{2+\sqrt{2}}$, and we define the following ascending chain of fields: $\mathbb{Q}$, $K = \mathbb{Q}(\sqrt{2})$ and $L = K(z)$. Here $K$ is the minimal subfield where $P$ has a $\psi$ preimage $Q_1$ in $E'$, and $L$ is the minimal subfield where that preimage has its own $\phi$ preimage $Q$ in $E$. It is a quick check in Magma to verify that in $E(L)$, $P$ is divisible by 2. It then remains to verify that the elements $\sqrt{2}$ and $z = \sqrt{2+\sqrt{2}}$ are in $\mathbb{F}_p$.

First, we have that since 2 has order $p$, which is odd, there exists $h_k \in \mathbb{F}_p^\times$ such that $(h_k)^{2^k} = 2$. So any 2-power root of 2 is sure to exist.

For $z = \sqrt{2+\sqrt{2}}$ itself, we use Lemma 24 and $p \equiv \pm 1 \pmod{16}$ to show that we have an element $z = \zeta_{16} + \zeta_{16}^{-1} \in \mathbb{F}_p$, so we have all the necessary elements of $L$ in $E(\mathbb{F}_p)$ to show there exists a $Q \in E(\mathbb{F}_p)$ such that $2Q = P$. □

These two lemmas will allow us to sharpen the threshold to search for Fermat and Mersenne primes in the elliptic Fermat sequence. We now prove Theorem 10.

*Proof.* First, it is a quick computation in Magma to verify that for $p = 5, 17$, $P$ does not have a 2-power order in $E(\mathbb{F}_p)$, and so by Corollary 4, we have that 5 and 17 do not divide any elliptic Fermat number generated by $P$.

We rely on Proposition 12 and Lagrange's theorem. For a classical Fermat prime $F_n \neq 5, 17$, we have that 2 is a fourth power in $\mathbb{Z}/F_n\mathbb{Z}$. We can see this because for a generator $g$ of $\mathbb{Z}/F_n\mathbb{Z}$, we have $2 = g^k$, additionally, we have $g^{p-1} = g^{2^{2^n}} = 1$. We will show that $k \equiv 0 \pmod 4$. This is because 2 has order $2^{n+1} \in (\mathbb{Z}/F_n\mathbb{Z})^\times$, and so $2^{2^{n+1}} = (g^k)^{2^{n+1}} = 1$. Therefore, $2^{2^n} \mid k(2^{n+1})$, finally giving $2^{2^n-n-1} \mid k$, which is a multiple of 4 for $n \geq 3$.

Since 2 is a fourth power in $\mathbb{F}_p$, we know that $E : y^2 = x^3 - 2x$ is isomorphic to the curve $E' : y^2 = x^3 - x$. From [Denomme and Savin 2008], we also have $E'(\mathbb{F}_p) \cong \mathbb{Z}[i]/(1+i)^{2^n}$. Moreover, $\mathbb{Z}[i]/(1+i)^{2^n} = \mathbb{Z}[i]/2^{2^{n-1}} \cong (\mathbb{Z}/2^{2^{n-1}}\mathbb{Z}) \times (\mathbb{Z}/2^{2^{n-1}}\mathbb{Z})$, from which we can deduce that $E(\mathbb{F}_p) \cong (\mathbb{Z}/2^{2^{n-1}}\mathbb{Z}) \times (\mathbb{Z}/2^{2^{n-1}}\mathbb{Z})$. Thus the order of $P$ is a divisor of $2^{2^{n-1}}$.

By Lemma 25, we know there exists some $Q \in E(\mathbb{F}_p)$ such that $2Q = P$. In light of this we can tighten this initial upper bound by noting that all elements have order dividing $2^{2^{n-1}}$, and so $2^{2^{n-1}-1}P = 2^{2^{n-1}-1}(2Q) = 2^{2^{n-1}}Q = 0$. We conclude that $P$ has order dividing $2^{2^{n-1}-1}$ and so $p$ must divide $F_k(E, P)$ for some $k \leq 2^{n-1} - 2$ by Corollary 4. □

It remains to discuss the appearance of a Mersenne prime in the elliptic Fermat sequence. We prove Theorem 11.

*Proof.* The method we take to show this bound begins with the fact that $|E(\mathbb{F}_q)| = q + 1 = 2^p$. Additionally, we have $E(\mathbb{F}_q) \cong \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/mn\mathbb{Z}$, where $q \equiv 1 \pmod m$. We have that $E(\mathbb{F}_q)$ contains all three points of order 2 because these are $(0, 0)$ and $(\pm\sqrt{2}, 0)$ and $\sqrt{2} \in \mathbb{F}_q$ since $q \equiv 7 \pmod 8$. Combining this with $q \equiv -1 \pmod{2^p}$ we have $E(\mathbb{F}_q) \cong \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2^{p-1}\mathbb{Z}$. So the order of any point in $E(\mathbb{F}_p)$ must divide $2^{p-1}$. It suffices to exhibit a point $R$ such that $4R = P$, so that $2^{p-3}P = 2^{p-3}2^2R = 2^{p-1}R = 0$.

Continuing the methodology first used in the proof of Lemma 25, we will show that such an $R$ is in $E(\mathbb{F}_q)$ so that $2R = Q$, where $Q \in E(L)$ is the point found in Lemma 25 . To do this, we extend the fields from Lemma 25 and create

$$M = L(\sqrt{z(2+z)}) \quad \text{and} \quad N = M(\sqrt{\sqrt{2}(z-1)}).$$

Again, one may check in Magma that indeed $P$ is divisible by 4 in $E(N)$, so we just need to check for the existence of necessary elements.

We have already shown there is an element $z$ such that $z^2 = 2 + \sqrt{2}$, but we further assert that in $\mathbb{F}_q$, $2 + \sqrt{2}$ has odd order, and thus all 2-power roots exist. This is quick to see because $(2 + \sqrt{2})^{(q-1)/2} = (z^2)^{(q-1)/2} = z^{q-1} = 1$.

We now find $\sqrt{z(2+z)}$, which amounts to finding square roots of $z$ and $2 + z$. By the above, we already have a square root of $z$, so we just need to show the existence of the square root of $2 + z$. This is simple if we let $w = \zeta_{32} + \zeta_{32}^{-1}$ in $\mathbb{F}_p$, which we know to exist if $q \equiv -1 \pmod{32}$. Then $w^2 = 2 + z$.

It remains to find $\sqrt{\sqrt{2}(z-1)}$. Again it suffices to just find a square root of $z-1$. To show such a root exists, consider

$$(z-1)(-z-1) = -z^2 + 1 = 1 - \sqrt{2} = (-1)(1+\sqrt{2}).$$

Note that $z = \sqrt[4]{2}\sqrt{(1+\sqrt{2})}$, and that $1+\sqrt{2}$ is a square because $\sqrt[4]{2}$ and $z$ are squares, but $-1$ is not a square modulo $q$ since $q \equiv -1 \pmod 4$, so $(z-1)(-z-1)$ is not a square. This implies that exactly one of $(z-1)$ and $(-z-1)$ is a square. So we choose the appropriate $z'$ such that $z'-1$ is a square and we are done.

Since all adjoined elements exist in $\mathbb{F}_q$, we are good to construct points $R$ such that $4R = 2Q = P$. Similar to Theorem 10, this implies that we can tighten the condition that $|P|\,|\,2^{p-1}$ further by $|P|\,|\,2^{p-3}$, and so by Corollary 4, $p$ must divide $F_k(E, P)$ for some $k \leq p-4$. □

## Acknowledgements

We would like to thank the Wake Forest Department of Mathematics and Statistics for their hospitality and resources. The authors used Magma version 2.22-9 [Bosma et al. 1997] and Sage version 7.5.1 [SageMath 2017] for computations. The authors would like to thank Joe Silverman and Katherine Stange for helpful comments, and we would also like to thank the anonymous referee for helpful comments and references which have substantially improved the paper.

## References

[Abatzoglou et al. 2016] A. Abatzoglou, A. Silverberg, A. V. Sutherland, and A. Wong, "A framework for deterministic primality proving using elliptic curves with complex multiplication", *Math. Comp.* **85**:299 (2016), 1461–1483. MR Zbl

[Atkin and Morain 1993] A. O. L. Atkin and F. Morain, "Elliptic curves and primality proving", *Math. Comp.* **61**:203 (1993), 29–68. MR Zbl

[Ayad 1992] M. Ayad, "Points $S$-entiers des courbes elliptiques", *Manuscripta Math.* **76**:3-4 (1992), 305–324. MR Zbl

[Bosma et al. 1997] W. Bosma, J. Cannon, and C. Playoust, "The Magma algebra system, I: The user language", *J. Symbolic Comput.* **24**:3-4 (1997), 235–265. MR Zbl

[Bugeaud et al. 2006] Y. Bugeaud, M. Mignotte, and S. Siksek, "Classical and modular approaches to exponential Diophantine equations, I: Fibonacci and Lucas perfect powers", *Ann. of Math.* (2) **163**:3 (2006), 969–1018. MR Zbl

[Cohen 2007] H. Cohen, *Number theory, I: Tools and Diophantine equations*, Graduate Texts in Mathematics **239**, Springer, 2007. MR Zbl

[Denomme and Savin 2008] R. Denomme and G. Savin, "Elliptic curve primality tests for Fermat and related primes", *J. Number Theory* **128**:8 (2008), 2398–2412. MR Zbl

[Dummit and Foote 2004] D. S. Dummit and R. M. Foote, *Abstract algebra*, 3rd ed., John Wiley & Sons, Hoboken, NJ, 2004. MR Zbl

[Everest et al. 2004] G. Everest, V. Miller, and N. Stephens, "Primes generated by elliptic curves", *Proc. Amer. Math. Soc.* **132**:4 (2004), 955–963. MR Zbl

[Everest et al. 2008] G. Everest, P. Ingram, V. Mahé, and S. Stevens, "The uniform primality conjecture for elliptic curves", *Acta Arith.* **134**:2 (2008), 157–181. MR Zbl

[Fermat 1894] P. de Fermat, *Œuvres de Pierre Fermat, II*, edited by P. Tannery and C. Henry, Gauthier-Villars et Fils, Paris, 1894.

[Gross 2005] B. H. Gross, "An elliptic curve test for Mersenne primes", *J. Number Theory* **110**:1 (2005), 114–119. MR Zbl

[Koblitz 1987] N. Koblitz, "Elliptic curve cryptosystems", *Math. Comp.* **48**:177 (1987), 203–209. MR Zbl

[Lenstra 1987a] H. W. Lenstra, Jr., "Elliptic curves and number-theoretic algorithms", pp. 99–120 in *Proceedings of the International Congress of Mathematicians, I* (Berkeley, CA., 1986), edited by A. M. Gleason, Amer. Math. Soc., Providence, RI, 1987. MR Zbl

[Lenstra 1987b] H. W. Lenstra, Jr., "Factoring integers with elliptic curves", *Ann. of Math.* (2) **126**:3 (1987), 649–673. MR Zbl

[Miller 1986] V. S. Miller, "Use of elliptic curves in cryptography", pp. 417–426 in *Advances in cryptology: CRYPTO '85* (Santa Barbara, CA, 1985), edited by H. C. Williams, Lecture Notes in Comput. Sci. **218**, Springer, 1986. MR Zbl

[Poonen et al. 2007] B. Poonen, E. F. Schaefer, and M. Stoll, "Twists of $X(7)$ and primitive solutions to $x^2 + y^3 = z^7$", *Duke Math. J.* **137**:1 (2007), 103–158. MR Zbl

[SageMath 2017] The Sage Developers, *SageMath, the Sage Mathematics Software System*, 2017, available at http://www.sagemath.org. Version 7.5.1.

[Silverman 1986] J. H. Silverman, *The arithmetic of elliptic curves*, Graduate Texts in Mathematics **106**, Springer, 1986. MR Zbl

[Silverman 1994] J. H. Silverman, *Advanced topics in the arithmetic of elliptic curves*, Graduate Texts in Mathematics **151**, Springer, 1994. MR Zbl

[Silverman and Tate 1992] J. H. Silverman and J. Tate, *Rational points on elliptic curves*, Springer, 1992. MR Zbl

[Thue 1909] A. Thue, "Über Annäherungswerte algebraischer Zahlen", *J. Reine Angew. Math.* **135** (1909), 284–305. MR Zbl

[Tsumura 2011] Y. Tsumura, "Primality tests for $2^p \pm 2^{(p+1)/2} + 1$ using elliptic curves", *Proc. Amer. Math. Soc.* **139**:8 (2011), 2697–2703. MR Zbl

[Tzanakis and de Weger 1989] N. Tzanakis and B. M. M. de Weger, "On the practical solution of the Thue equation", *J. Number Theory* **31**:2 (1989), 99–132. MR Zbl

[Wiles 1995] A. Wiles, "Modular elliptic curves and Fermat's last theorem", *Ann. of Math.* (2) **141**:3 (1995), 443–551. MR Zbl

Received: 2017-08-12 Revised: 2018-07-17 Accepted: 2018-07-22

skye@gatech.edu *Department of Mathematics, Reed College, Portland, OR, United States*

randydominick1093@gmail.com *Department of Mathematics & Statistics, Texas Tech University, Lubbock, TX, United States*

mk6673@bard.edu *Department of Mathematics, Bard College, Annandale-on-Hudson, NY, United States*

rouseja@wfu.edu *Department of Mathematics and Statistics, Wake Forest University, Winston-Salem, NC, United States*

alexandra_walsh@brown.edu *Mathematics Department, Brown University, Providence, RI, United States*

INVOLVE 12:3 (2019)
dx.doi.org/10.2140/involve.2019.12.451

# Nilpotent orbits for Borel subgroups of $\mathrm{SO}_5(k)$

Madeleine Burkhart and David Vella

(Communicated by Kenneth S. Berenhaut)

Let $G$ be a quasisimple algebraic group defined over an algebraically closed field $k$ and $B$ a Borel subgroup of $G$ acting on the nilradical $\mathfrak{n}$ of its Lie algebra $\mathfrak{b}$ via the adjoint representation. It is known that $B$ has only finitely many orbits in only five cases: when $G$ is type $A_n$ for $n \leq 4$, and when $G$ is type $B_2$. We elaborate on this work in the case when $G = \mathrm{SO}_5(k)$ (type $B_2$) by finding the defining equations of each orbit. We use these equations to determine the dimension of the orbits and the closure ordering on the set of orbits. The other four cases, when $G$ is type $A_n$, can be approached the same way and are treated in a separate paper.

## 1. Introduction

Before specializing to $G = \mathrm{SO}_5(k)$, we make some general remarks in order to provide context and some motivation for our work. Let $k$ be an algebraically closed field and $G$ a quasisimple algebraic group over $k$. Fix a maximal torus $T$ of $G$, and let $\Phi$ denote the root system of $G$ relative to $T$ ($\Phi$ is irreducible since $G$ is quasisimple). Fix a set $\Delta$ of simple roots in $\Phi$, with corresponding set of positive roots $\Phi^+$, and let $B = TU$ ($U$ is the unipotent radical of $B$) be the Borel subgroup of $G$ determined by $\Phi^+$. Write the one-dimensional unipotent root group corresponding to a root $\alpha$ as $U_\alpha$. Denote the Lie algebra of $G$ by $\mathfrak{g}$, that of $T$ by $\mathfrak{h}$, and that of $B$ by $\mathfrak{b}$. Then the nilradical $\mathfrak{n} = \mathfrak{n}(\mathfrak{b})$ of $\mathfrak{b}$ is in fact the Lie algebra of $U$, and we have decompositions $\mathfrak{b} = \mathfrak{h} \oplus \mathfrak{n}$, and $\mathfrak{n} = \bigoplus_{\alpha \in \Phi^+} \mathfrak{g}_\alpha$ as vector spaces, where $\mathfrak{g}_\alpha$ is the root space of $\mathfrak{g}$ corresponding to $\alpha$ and is also the Lie algebra of $U_\alpha$. There is a corresponding decomposition of $U \approx \prod_{\alpha \in \Phi^+} U_\alpha$ as algebraic varieties over $k$. In particular, $U$ is generated as a group by the root groups $U = \langle U_\alpha \mid \alpha \in \Phi^+ \rangle$, a fact we use repeatedly in our calculations in Section 2 below.

$G$ acts on $\mathfrak{g}$ via the adjoint representation, and the orbits of this action have been intensely studied, partly because there are connections between the nilpotent orbit theory and the representation theory of $G$. It is known that there are only finitely many nilpotent $G$-orbits (a *nilpotent orbit* means the orbit of a nilpotent element

*MSC2010:* 17B08, 20G05.
*Keywords:* nilpotent orbits, Borel subgroups, modality.

of $\mathfrak{g}$). There are combinatorial indexing sets for these nilpotent orbits, and there are formulas to compute the dimension of each orbit. Also, it is known which orbits are in the Zariski closure of any given orbit (the *closure ordering*). Therefore, it is well understood how all the nilpotent orbits fit together to form a larger object, called the *nullcone* $\mathcal{N}$ of $\mathfrak{g}$, which is the union of the nilpotent orbits. For details of this classical theory, see [Collingwood and McGovern 1993] for the characteristic-0 case and [Carter 1985; Jantzen 2004] more generally.

The notion of a support variety of a module is one example of a connection between nilpotent $G$-orbits and representation theory, when the characteristic of $k$ is $p > 0$. In this case recall that there is a *p-th power map* $x \mapsto x^{[p]}$ on $\mathfrak{g}$ that makes $\mathfrak{g}$ into a restricted Lie algebra, and there is a *Frobenius map* $F : G \to G$, whose kernel is denoted by $G_1$, which is an infinitesimal group scheme whose rational representation theory coincides with the representation theory of $\mathfrak{g}$. Similarly, denote the kernel of $F : B \to B$ by $B_1$. By results of [Friedlander and Parshall 1986], the *cohomology variety* of $G_1$ (the maximal ideal spectrum of the even-degree cohomology ring $H^{2\bullet}(G_1, k)$) identifies naturally with a subvariety $\mathcal{N}_1 = \{x \in \mathcal{N} \mid x^{[p]} = 0\}$ of the nullcone $\mathcal{N}$ of $\mathfrak{g}$, and furthermore, for any finite-dimensional $\mathfrak{g}$-module $M$, there is an important subvariety of $\mathcal{N}_1$ called the *support variety* of $M$, denoted by $V_{\mathfrak{g}}(M)$ or $V_{G_1}(M)$. If $M$ is a rational $G$-module, then $V_{G_1}(M)$ is $G$-stable in $\mathcal{N}_1$ and is therefore a union of nilpotent $G$-orbits.

Aspects of the representation theory of $G_1$ are determined by this support variety (for example, $M$ is a projective module if and only if $V_{G_1}(M) = \{0\}$), so one would like to be able to compute these support varieties, and knowing they are unions of nilpotent $G$-orbits may be useful.

If $H$ is a closed subgroup of $G$ and $M$ is a rational $H$-module, denote the rational $G$-module induced from $M$ by $M|_H^G$. Now let $X(T)$ be the character group of $T$, and for $\lambda \in X(T)$, we also use the symbol $\lambda$ to denote the one-dimensional $T$-module on which $T$ acts via $\lambda : t \cdot v = \lambda(t)v$ for all $t \in T$. This rational $T$-module extends to a rational $B$-module by trivial $U$-action, also denoted by $\lambda$. The modules $\lambda|_B^G$ are (the duals of) the well-known *Weyl modules* of $G$. Another important class of modules are those of the form $Z(\lambda) = \lambda|_{B_1}^{G_1}$, sometimes called *baby Verma modules*. (The name comes from the fact that there is an alternate definition of $Z(\lambda)$ which is analogous to the definition of a Verma module for $\mathfrak{g}$, while the adjective "baby" can be interpreted as alluding to the infinitesimal subgroups in our definition using induction, or to the fact that the $Z(\lambda)$ are finite-dimensional and the usual Verma modules are not.)

One of the goals of the paper [Nakano et al. 2002, Theorem 6.21] was to calculate the support varieties $V_{G_1}(\lambda|_B^G)$ in order to prove the "Jantzen conjecture". A central strategy of that paper is to compare $V_{G_1}(\lambda|_B^G)$ to $V_{G_1}(Z(\lambda))$. Of course, $Z(\lambda)$ is not a $G$-module (only a $G_1$-module), so one would not expect $V_{G_1}(Z(\lambda))$ to be a

$G$-stable subset of $\mathcal{N}_1$. However, it turns out that this variety is stable under the action of $B$ and therefore $V_{G_1}(Z(\lambda))$ is a union of nilpotent $B$-orbits [Nakano et al. 2002, Proposition 7.1.1].

That nilpotent $B$-orbits (as well as nilpotent $G$-orbits) have a connection to the representation theory of $G$ provides a motivation for studying nilpotent $B$-orbits. Even without this explicit motivation, it is interesting to try to generalize what is known about nilpotent $G$-orbits to nilpotent $B$-orbits. However, the case of nilpotent $B$-orbits is not nearly as tidy as that of nilpotent $G$-orbits. One important difference is that most of the time there are infinitely many nilpotent $B$-orbits. Thus, finding a nice indexing set for the orbits could be difficult. In general, we would expect an indexing set to have a continuous piece as well as a discrete piece, as certain infinite families of orbits might be described by continuous parameters.

In this paper, our focus is on the case $G = \mathrm{SO}_5(k)$, which is one of the five cases where there are finitely many nilpotent $B$-orbits. For each of the orbits, we find the polynomial defining equations of the orbit. From these calculations, it is easy to determine both the dimension of each orbit as well as the closure ordering for the set of orbits.

## 2. Nilpotent $B$-Orbits in $\mathrm{SO}_5(k)$

Throughout this section we assume the characteristic of $k$ is either 0 or a prime $p \neq 2$. The following proposition is a basic fact about algebraic group actions. A proof can be found in [Borel 1991; Humphreys 1975].

**Proposition 1.** *Let $G$ be an algebraic group acting morphically on a nonempty variety $V$. Then each orbit is a locally closed, smooth variety, and its boundary is a union of orbits of strictly lower dimension.*

Thus, the orbit $G \cdot x$ is open and dense in its closure $\overline{G \cdot x}$, and hence has the same dimension as its closure.

We study the action of a Borel subgroup $B$ of $G = \mathrm{SO}_5(k)$ on the nilradical $\mathfrak{n}$ of its Lie algebra $\mathfrak{b} \subseteq \mathfrak{so}_5(k)$, via the adjoint representation. Our main results consist of finding the defining equations of each nilpotent $B$-orbit, which will exhibit each orbit explicitly as an intersection of an open set and a closed set. From there it will be an easy matter to determine the closure of each orbit, and thereby find the dimension of each orbit, as well as to determine which orbits comprise the boundary of a given orbit. That is, we will find the partial order determined by the orbit closures, which is defined as $G \cdot x \preceq G \cdot y$ if and only if $G \cdot x \subseteq \overline{G \cdot y}$.

Let $f$ be a polynomial. We use the standard notation that the zero set of $f$ is written as $Z(f)$ and that $Z(f, g) = Z(f) \cap Z(g)$ is the set of common zeros of polynomials $f$ and $g$. If we have a finite set of polynomials $f_1, \ldots, f_r$, then $Z(f_1, f_2, \ldots, f_r)$ is a Zariski-closed set, that is, it is an affine variety. The notation

$V(f)$ denotes the complement of $Z(f)$, the set of elements that are not zeros of $f$, and so $V(f)$ is a Zariski-open set. A locally closed set is an intersection of an open set and a closed set, and in this section the orbits will turn out to be locally closed sets of the form $V = Z(f_1, f_2, \ldots, f_r) \cap V(g_1) \cap V(g_2) \cap \cdots \cap V(g_t)$ for polynomials $f_i$ and $g_j \neq 0$ for all $j$. Observe that the closure of $V$ is then $Z(f_1, f_2, \ldots, f_r)$ and $V$ is open and dense in this closure, whence $\dim V = \dim Z(f_1, f_2, \ldots, f_r)$.

If $U\gamma$ is a root group of $G$, then $U_\gamma(t)$ denotes the image of $t$ under the standard isomorphism $k_{\text{add}} \approx U_\gamma$. In classical groups, the adjoint action on the Lie algebra is simply conjugation of matrices. The matrix $e_{ij}$ is the matrix with a 1 in the $ij$ position and 0 everywhere else. Now $\mathfrak{g} = \mathfrak{so}_5(k)$, and we take $T$ to be the set of diagonal matrices in $G$. More precisely, a typical element of the torus $T$ has the form

$$T(s,t) = \operatorname{diag}(1, s, t, s^{-1}, t^{-1}) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & s & 0 & 0 & 0 \\ 0 & 0 & t & 0 & 0 \\ 0 & 0 & 0 & s^{-1} & 0 \\ 0 & 0 & 0 & 0 & t^{-1} \end{bmatrix}$$

for $s$ and $t$ nonzero in $k$.

For root systems we use the standard notation one finds in [Humphreys 1972; Bourbaki 2002]. In type $B_2$, the simple roots are $\Delta = \{\alpha_1, \alpha_2\}$ and $\Phi^+ = \{\alpha_1, \alpha_2, \alpha_1+\alpha_2, \alpha_1+2\alpha_2\}$, so $\mathfrak{n} \approx k^4$ as an affine variety and vector space. The root vectors in $\mathfrak{n}$ are the matrices

$$\begin{aligned} x_{\alpha_1} &= e_{23} - e_{54}, \\ x_{\alpha_2} &= e_{15} - e_{31}, \\ x_{\alpha_1+\alpha_2} &= e_{14} - e_{21}, \\ x_{\alpha_1+2\alpha_2} &= e_{25} - e_{34}. \end{aligned}$$

Now each root space has a coordinate function, which we denote by a capital $X$ with the same subscript as the root space. In other words, a typical element of $\mathfrak{n}$ has the form of a linear combination of the four root vectors, $x = a_1 x_{\alpha_1} + a_2 x_{\alpha_2} + a_3 x_{\alpha_1+\alpha_2} + a_4 x_{\alpha_1+2\alpha_2}$, which we can write in coordinate form as $(a_1, a_2, a_3, a_4)$ (showing explicitly $\mathfrak{n} \approx k^4$), and the coordinate function just selects the appropriate coordinate; so $X_{\alpha_1}(x) = a_1$ and $X_{\alpha_1+2\alpha_2}(x) = a_4$, for example. Thus, the $B$-orbits in which we are interested are locally closed sets in $\mathfrak{n}$ or $k^4$ which are defined by polynomials in the four variables of the polynomial ring $k[X_{\alpha_1}, X_{\alpha_2}, X_{\alpha_1+\alpha_2}, X_{\alpha_1+2\alpha_2}]$.

To begin the calculations, let's determine the $B$-orbit of the highest root vector $x_{\alpha_1+2\alpha_2}$. Since the highest root vector has the highest weight of the adjoint representation, it is a maximal vector — it is fixed by $U$, the unipotent radical of $B$, and sent

to a multiple of itself by the torus $T$. It follows that $B \cdot x_{\alpha_1+2\alpha_2} = T \cdot U \cdot x_{\alpha_1+2\alpha_2} = T \cdot x_{\alpha_1+2\alpha_2} \subseteq \mathfrak{g}_{\alpha_1+2\alpha_2}$. Thus, we need only compute the $T$-orbit of this weight vector, which is easy by direct calculation. Abbreviating the root vector by $x$, we have

$$\begin{aligned} T(s,t)\cdot x &= T(s,t)xT(s,t)^{-1} \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & s & 0 & 0 & 0 \\ 0 & 0 & t & 0 & 0 \\ 0 & 0 & 0 & s^{-1} & 0 \\ 0 & 0 & 0 & 0 & t^{-1} \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & s^{-1} & 0 & 0 & 0 \\ 0 & 0 & t^{-1} & 0 & 0 \\ 0 & 0 & 0 & s & 0 \\ 0 & 0 & 0 & 0 & t \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & st \\ 0 & 0 & 0 & -st & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} = stx. \end{aligned}$$

In particular, taking $s = 1$, we obtain all elements of the form $tx$, $t \neq 0$, as part of this orbit. Since 0 is in an orbit by itself, this shows the orbit is precisely the set of all nonzero multiples of $x$. In terms of linear combinations of root vectors, or coordinates in $\mathfrak{n} \approx k^4$, this says an element $(a_1, a_2, a_3, a_4)$ belongs to the orbit $B \cdot x$ if and only if $a_1 = a_2 = a_3 = 0$ and $a_4 \neq 0$. In other words, this shows that the orbit is

$$B \cdot x = B \cdot x_{\alpha_1+2\alpha_2} = Z(X_{\alpha_1}, X_{\alpha_2}, X_{\alpha_1+\alpha_2}) \cap V(X_{\alpha_1+2\alpha_2}).$$

These are the defining equations of this orbit. Clearly, its closure is $\overline{B \cdot x} = Z(X_{\alpha_1}, X_{\alpha_2}, X_{\alpha_1+\alpha_2})$, the intersection of three coordinate hyperspaces in $n \approx k^4$, which is precisely the axis of the fourth coordinate $X_{\alpha_1+2\alpha_2}$. In particular, it is obvious that this orbit is one-dimensional (it is dense in the highest root space $\mathfrak{g}_{\alpha_1+2\alpha_2}$.)

In order to save space, we will eschew writing out the matrices from this point on, in favor of writing elements of $\mathfrak{n}$ as linear combinations of root vectors (or as ordered quadruples in $k^4$), and elements of $B$ as products of elements in $T$ and elements of the four one-dimensional root groups $U_\alpha$ for $\alpha \in \Phi^+$. Thus, the above calculation could be more compactly written as

$$T(s,t)U \cdot x = T(s,t) \cdot x = stx,$$

leaving the reader to check the actual matrix calculation. In subsequent calculations, it will be helpful to remember how the unipotent root groups act on weight vectors in rational $G$-modules:

**Lemma 2** [Humphreys 1975, Proposition 27.2]. *Let $\alpha \in \Phi$, and let $v \in V_\lambda$ be a weight vector in any rational $G$-module. Then each element $u \in U_\alpha$ acts on $v$ as*

*follows*: $u \cdot v = v + \sum_{k>0} v_{\lambda+k\alpha}$, *where* $v_{\lambda+k\alpha}$ *is a weight vector of weight* $\lambda + k\alpha$, *and* $k$ *is a positive integer.*

Next, consider the orbit of the root vector $x = x_{\alpha_1+\alpha_2}$ for the highest short root $\alpha_1 + \alpha_2$. If $\gamma$ is a positive root, then by Lemma 2, $U_\gamma(r) \cdot x = x + w$, where $w$ is a sum of root vectors for roots of the form $(\alpha_1 + \alpha_2) + k\gamma$ for some $k > 0$, but there are no roots of this form unless $k = 1$ and $\gamma = \alpha_2$. It follows that $w = 0$ for all positive roots $\gamma$ except $\gamma = \alpha_2$. In particular, $U_{\alpha_1}$, $U_{\alpha_1+\alpha_2}$, and $U_{\alpha_1+2\alpha_2}$ all fix the vector $x$, whence $U \cdot x = U_{\alpha_2} \cdot x$. Therefore, we have $B \cdot x = TU \cdot x = TU_{\alpha_2} \cdot x$. Now take arbitrary elements $T(s, t) \in T$ and $U_{\alpha_2}(r)$ and compute directly

$$T(s, t)U_{\alpha_2}(r) \cdot x = sx + rstx_{\alpha_1+2\alpha_2}. \tag{1}$$

It follows, since $s \neq 0$, that the $x = x_{\alpha_1+\alpha_2}$-coordinate is nonzero, while it is clear that for all $r \in k$ and $s, t \in k - \{0\}$ that the $x_{\alpha_1}$- and $x_{\alpha_2}$-coordinates are 0, whence

$$B \cdot x = T \cdot U_{\alpha_2} \cdot x \subseteq Z(X_{\alpha_1}, X_{\alpha_2}) \cap V(X_{\alpha_1+\alpha_2}).$$

To check the reverse containment, we start with an arbitrary element of the locally closed set on the right, and find an element of $B$ that carries $x$ to that element. So let $y \neq 0$ and $z \in k$ be arbitrary. Then the element $(0, 0, y, z) = yx + zx_{\alpha_1+2\alpha_2}$ is an arbitrary element of our locally closed set. But substitute the element $T(y, 1)U_{\alpha_2}(z/y)$ directly into (1) to obtain

$$T(y, 1)U_{\alpha_2}\left(\frac{z}{y}\right) \cdot x = yx + y \cdot 1 \cdot \frac{z}{y} x_{\alpha_1+2\alpha_2} = (0, 0, y, z).$$

This shows the reverse containment and so proves the orbit is $B \cdot x = B \cdot x_{\alpha_1+\alpha_2} = Z(X_{\alpha_1}, X_{\alpha_2}) \cap V(X_{\alpha_1+\alpha_2})$.

Next consider the orbit of $x = x_{\alpha_2}$. By Lemma 2, we only need consider the action of $U_{\alpha_1}$ and $U_{\alpha_1+\alpha_2}$. By direct calculation, we have

$$\begin{aligned} T(p, q)U_{\alpha_1}(s)U_{\alpha_1+\alpha_2}(r) \cdot x &= qx + psx_{\alpha_1+\alpha_2} - pqrx_{\alpha_1+2\alpha_2} \\ &= (0, q, ps, -pqr). \end{aligned} \tag{2}$$

Since $q \neq 0$, this shows $B \cdot x \subseteq Z(X_{\alpha_1}) \cap V(X_{\alpha_2})$. To see we have equality we again start with an arbitrary element $(0, w, y, z) \in Z(X_{\alpha_1}) \cap V(X_{\alpha_2})$ (so $w \neq 0$), and exhibit an element of $B$ which carries $x = (0, 1, 0, 0)$ to it. One such element is $T(1, w)U_{\alpha_1}(y)U_{\alpha_1+\alpha_2}(-z/w)$. Indeed by (2) we obtain

$$T(1, w)U_{\alpha_1}(y)U_{\alpha_1+\alpha_2}\left(-\frac{z}{w}\right) \cdot x = wx + yx_{\alpha_1+\alpha_2} + zx_{\alpha_1+2\alpha_2} = (0, w, y, z).$$

We have shown that $B \cdot x = B \cdot x_{\alpha_2} = Z(X_{\alpha_1}) \cap V(X_{\alpha_2})$.

Next consider the orbit of $x = x_{\alpha_1}$. By Lemma 2, the $U$-orbit of $x$ is the same as the $U_{\alpha_2}$-orbit. Thus, direct calculation yields

$$T(s,t)U_{\alpha_2}(r)\cdot x = \frac{s}{t}x - rsx_{\alpha_1+\alpha_2} - \frac{r^2st}{2}x_{\alpha_1+2\alpha_2} = \left(\frac{s}{t}, 0, -rs, -\frac{r^2st}{2}\right). \quad (3)$$

Note that because of the 2 in the denominator, we must avoid characteristic-2 fields. Since $s/t \neq 0$, this shows $B\cdot x \subseteq Z(X_{\alpha_2}) \cap V(X_{\alpha_1})$. However, unlike the above orbits, we do not have an equality in this case due to algebraic dependence relations among the coordinates.

Indeed, define $w, y, z$ by equating

$$\left(\frac{s}{t}, 0, -rs, -\frac{r^2st}{2}\right) = (w, 0, y, z),$$

and observe that $y^2 + 2wz = 0$ for every element of this form. This shows that, in fact, $B\cdot x \subseteq Z(X_{\alpha_2}, X_{\alpha_1+\alpha_2}^2 + 2X_{\alpha_1}X_{\alpha_1+2\alpha_2}) \cap V(X_{\alpha_1})$. We now claim we have an equality. Indeed an arbitrary element of this locally closed set has the form $(w, 0, y, z)$ with $y^2 + 2wz = 0$ and $w \neq 0$ But it follows from (3) that

$$\begin{aligned} T(w,1)U_{\alpha_2}\left(-\frac{y}{w}\right)\cdot x &= wx + yx_{\alpha_1+\alpha_2} - \frac{y^2}{2w}x_{\alpha_1+2\alpha_2} \\ &= \left(w, 0, y, -\frac{y^2}{2w}\right) = (w, 0, y, z), \end{aligned}$$

where the last equality follows because $y^2 + 2wz = 0$. This shows that $B\cdot x = B\cdot x_{\alpha_1} = Z(X_{\alpha_2}, X_{\alpha_1+\alpha_2}^2 + 2X_{\alpha_1}X_{\alpha_1+2\alpha_2}) \cap V(X_{\alpha_1})$.

So far we have determined the orbits of the four root vectors, but taken together they do not exhaust all of $\mathfrak{n}$. The remaining orbits can be taken to be orbits of certain sums of root vectors. For example, consider the element $x = x_{\alpha_1} + x_{\alpha_1+2\alpha_2}$. All the root groups $U_\gamma$ of $U$ fix $x$ except for $U_{\alpha_2}$ by Lemma 2. By direct computation we have

$$T(s,t)U_{\alpha_2}(r)\cdot x = \frac{s}{t}x_{\alpha_1} - rsx_{a_1+\alpha_2} + st\left(1-\frac{r^2}{2}\right)x_{\alpha_1+2\alpha_2} = \left(\frac{s}{t}, 0, rs, st\left(1-\frac{r^2}{2}\right)\right).$$

Now $s/t \neq 0$, so the orbit is contained in $Z(X_{\alpha_2}) \cap V(X_{\alpha_1})$. But also

$$X_{\alpha_1+\alpha_2}^2 + 2X_{\alpha_1}X_{\alpha_1+2\alpha_2} = (-rs)^2 + 2\frac{s}{t}\left(st\left(1-\frac{r^2}{2}\right)\right) = 2s^2 \neq 0.$$

So $B\cdot x \subseteq Z(X_{\alpha_2}) \cap V(X_{\alpha_1}) \cap V(X_{\alpha_1+\alpha_2}^2 + 2X_{\alpha_1}X_{\alpha_1+2\alpha_2})$. We now prove the reverse containment. Note that an arbitrary element of this locally closed set has the form $(w, 0, y, z)$ with $w \neq 0$, and $y^2 + 2wz \neq 0$. Since $k$ is not of characteristic 2, the

element $\frac{1}{2}(y^2+2wz)$ exists and is nonzero in $k$, and since $k$ is algebraically closed, its square root exists in $k$ and is also nonzero. Now by direct calculation we have

$$T\left(\sqrt{\frac{y^2+2wz}{2}}, \frac{1}{w}\sqrt{\frac{y^2+2wz}{2}}\right)U_{\alpha_2}\left(-y\sqrt{\frac{2}{y^2+2wz}}\right)\cdot x = (w, 0, y, z).$$

This proves $B\cdot x = B\cdot(x_{\alpha_1}+x_{\alpha_1+\alpha_2}) = Z(X_{\alpha_2})\cap V(X_{\alpha_1})\cap V(X_{\alpha_1+\alpha_2}^2+2X_{\alpha_1}X_{\alpha_1+2\alpha_2})$.

The last orbit we need to consider is the orbit of $x = x_{\alpha_1}+x_{\alpha_2}$. Only $U_{\alpha_1+2\alpha_2}$ fixes $x$, so we need to see how all three of the other root groups act. By direct matrix calculation we have

$$T(s,t)U_{\alpha_1}(a)U_{\alpha_2}(b)U_{\alpha_1+\alpha_2}(c)\cdot x = \left(\frac{s}{t}, t, (a-b)s, -st\left(\frac{b^2}{2}+c\right)\right). \quad (4)$$

Since $s, t \neq 0$, we have $B\cdot x \subseteq V(X_{\alpha_1})\cap V(X_{\alpha_2})$. We now show the reverse containment. Let $(w, u, y, z) \in V(X_{\alpha_1})\cap V(X_{\alpha_2})$ be arbitrary (so $w, u \neq 0$). Then using (4), we have

$$T(wu, u)U_{\alpha_1}\left(\frac{y}{wu}\right)U_{\alpha_2}(0)U_{\alpha_1+\alpha_2}\left(-\frac{z}{wu^2}\right)\cdot x = (w, u, y, z).$$

This shows $B\cdot x = V(X_{\alpha_1})\cap V(X_{\alpha_2})$ is an open, dense orbit in $\mathfrak{n}$, called the *regular* orbit. We are nearly finished with the proof of our main result:

**Theorem 3.** *Let $G = \mathrm{SO}_5(k)$, where $k$ is algebraically closed and not of characteristic 2, and let $B$ be a Borel subgroup acting on $\mathfrak{n}$ via the adjoint action. Then $B$ has just seven orbits as indicated in the following table along with their defining equations. The dimensions of these orbits are also indicated in the table, and the closure order is indicated by the Hasse diagram in Figure 1.*

| *element $x$ of $\mathfrak{n}$* | *defining equations for $B\cdot x$* | $\dim B\cdot x$ |
|---|---|---|
| $0$ | $Z(X_{\alpha_1}, X_{\alpha_2}, X_{\alpha_1+\alpha_2}, X_{\alpha_1+2\alpha_2})$ | 0 |
| $x_{\alpha_1+2\alpha_2}$ | $Z(X_{\alpha_1}, X_{\alpha_2}, X_{\alpha_1+\alpha_2})\cap V(X_{\alpha_1+2\alpha_2})$ | 1 |
| $x_{\alpha_1+\alpha_2}$ | $Z(X_{\alpha_1}, X_{\alpha_2})\cap V(X_{\alpha_1+\alpha_2})$ | 2 |
| $x_{\alpha_1}$ | $Z(X_{\alpha_2}, X_{\alpha_1+\alpha_2}^2+2X_{\alpha_1}X_{\alpha_1+2\alpha_2})\cap V(X_{\alpha_1})$ | 2 |
| $x_{\alpha_2}$ | $Z(X_{\alpha_1})\cap V(\alpha_2)$ | 3 |
| $x_{\alpha_1}+x_{\alpha_1+2\alpha_2}$ | $Z(X_{\alpha_2})\cap V(X_{\alpha_1})\cap V(X_{\alpha_1+\alpha_2}^2+2X_{\alpha_1}X_{\alpha_1+2\alpha_2})$ | 3 |
| $x_{\alpha_1}+x_{\alpha_2}$ | $V(X_{\alpha_1})\cap V(X_{\alpha_2})$ | 4 |

*In the Hasse diagram of the closure ordering in Figure 1, each orbit is indicated by its representative element from the first column of the table.*

*Proof.* We have already verified the entries in the first two columns of the table. Note that the orbit closures are just the closed sets from the defining equations. For example, since $B\cdot x_{\alpha_1+\alpha_2} = Z(X_{\alpha_1}, X_{\alpha_2})\cap V(X_{\alpha_1+\alpha_2})$, we have $\overline{B\cdot x_{\alpha_1+\alpha_2}}$

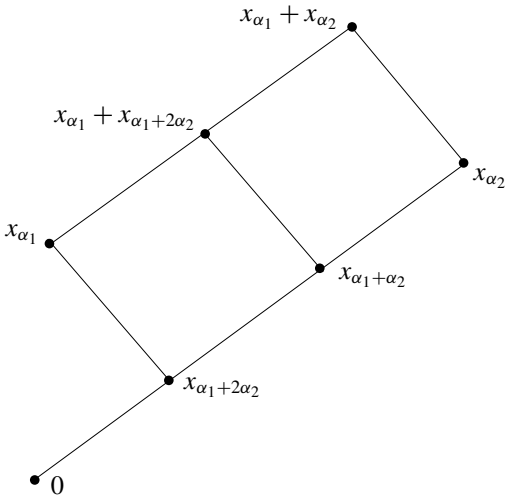


**Figure 1.** The closure order for nilpotent $B$-orbits in type $B_2$.

$= Z(X_{\alpha_1}, X_{\alpha_2})$. Using the closures, we can easily determine the dimensions in the third column as well as the closure ordering. Note that for polynomials $f_i$ in $r$ variables, the dimension of $Z(f_1, f_2, \ldots, f_k)$ is just $r-k$ provided that the $f_i$ are all algebraically independent. It should be clear that we found all the algebraic dependencies when we worked out the defining equations, so that the $f_i$ are algebraically independent in the closed sets in the second column of the table. Thus, since $r = \dim \mathfrak{n} = 4$, the dimensions in the third column are equal to $4-k$, where $k$ is the number of polynomials whose zero sets define the orbit closure.

The only nontrivial containment for the closure ordering is $B \cdot x_{a_1+2\alpha_2} \subseteq \overline{B \cdot x}_{\alpha_1}$, which happens if and only if $\overline{B \cdot x}_{\alpha_1+2\alpha_2} \subseteq \overline{B \cdot x}_{\alpha_1}$. So take an arbitrary element $x \in \overline{B \cdot x}_{\alpha_1+2\alpha_2} = Z(X_{\alpha_1}, X_{\alpha_2}, X_{\alpha_1+\alpha_2})$. Then, since both $X_{\alpha_1} = 0$ and $X_{\alpha_1+\alpha_2} = 0$, it follows that both $X^2_{\alpha_1+\alpha_2} = 0$ and $X_{\alpha_1} X_{\alpha_1+2\alpha_2} = 0$ when evaluated at $x$. Therefore, $X^2_{\alpha_1+\alpha_2} + 2X_{\alpha_1} X_{\alpha_1+2\alpha_2} = 0$ as well, so $x \in Z(X_{\alpha_2}, X^2_{\alpha_1+\alpha_2} + 2X_{\alpha_1} X_{\alpha_1+2\alpha_2}) = \overline{B \cdot x}_{\alpha_1}$, showing the desired containment. The other containments shown in the Hasse diagram follow similarly.

All that remains to show is that we have exhausted all the nilpotent orbits in $\mathfrak{n}$. So let $n = wx_{\alpha_1} + xx_{\alpha_2} + yx_{\alpha_1+\alpha_2} + zx_{\alpha_1+2\alpha_2} = (w, x, y, z)$ be an arbitrary element of $\mathfrak{n}$. We must show $n$ lies in one of these seven orbits. We will distinguish cases according to how many and which of the four coordinates are 0. If both $w$ and $x$ are nonzero, then $n$ is in $V(X_{\alpha_1}) \cap V(X_{\alpha_2}) = B \cdot (x_{\alpha_1} + x_{\alpha_2})$, the regular orbit. So it only remains to consider cases when one or both of $w, x$ are 0. First, suppose $w = 0$ but $x \neq 0$. Then $n = (0, x, y, z) \in Z(X_{\alpha_1}) \cap V(X_{\alpha_2}) = B \cdot x_{\alpha_2}$. On the other hand, suppose $w \neq 0$ and $x = 0$, so $n = (w, 0, y, z) \in Z(X_{\alpha_2}) \cap V(X_{\alpha_1})$. But then $n \in B \cdot (x_{\alpha_1} + x_{\alpha_1+2\alpha_2})$ or $n \in B \cdot x_{\alpha_1}$, depending on whether or not $y^2 + 2wz = 0$.

Lastly, we consider cases where $w = 0 = x$. In this case, $n = (0, 0, y, z) \in Z(X_{\alpha_1}) \cap Z(X_{\alpha_2})$. If $y \neq 0$, then $n \in Z(X_{\alpha_1}, X_{\alpha_2}) \cap V(X_{\alpha_1+\alpha_2}) = B \cdot x_{\alpha_1+\alpha_2}$. On the other hand, if $y = 0$, then $n = (0, 0, 0, z)$, which belongs to either $B \cdot 0$ or $B \cdot x_{\alpha_1+2\alpha_2}$, depending on whether or not $z = 0$. This covers all possible cases, and in each case, $n$ was in one of the above-mentioned orbits, whence the union of the seven orbits is all of $\mathfrak{n}$. □

## 3. Conclusions

The result that there are only finitely many nilpotent $B$-orbits for $\mathrm{SO}_5(k)$ can be phrased in terms of a general concept for algebraic group actions called modality (see [Popov and Röhrle 1997] for example).

Let $G$ be an arbitrary algebraic group acting morphically on a nonempty variety $V$. The *modality* of the action is

$$\mathrm{mod}(G, V) = \max_{Z} \min_{z \in Z} \mathrm{codim}_Z (G^0 \cdot z), \tag{5}$$

where $Z$ runs through all irreducible $G^0$-invariant subvarieties of $V$. Here, $G^0$ is the connected component of the identity in $G$. Informally, the modality is the maximum number of (continuous) parameters on which a family of $G$-orbits may depend.

Although we are mainly interested in nilpotent orbits for a Borel subgroup $B$ of $G$, much of the literature is written in terms of the more general case of orbits for a *parabolic* subgroup $P$, which is any closed subgroup containing a Borel subgroup. If $P$ is parabolic, denote its Lie algebra by $\mathfrak{p}$, and the nilradical of $\mathfrak{p}$ by $\mathfrak{n}(\mathfrak{p})$. Then $P$ acts on $\mathfrak{n}(\mathfrak{p})$ via the adjoint representation, and the *modality of* $P$ is defined to be $\mathrm{mod}(P, \mathfrak{n}(\mathfrak{p}))$. Thus the modality of $P$ is 0 precisely when there are only finitely many nilpotent $P$-orbits in $\mathfrak{n}(\mathfrak{p})$.

When $P = G$, the nilradical of $\mathfrak{g}$ is trivial since $\mathfrak{g}$ is simple, so the modality of $G$ is trivially 0. At the other extreme, the modality of $B$ is almost never 0. So one consequence of Theorem 3 is that if $p \neq 2$, then $B$ has modality 0 in type $B_2$. Of course, this is a well-known result. Based on earlier work in [Bürgstein and Hesselink 1987], in [Kashin 1990] all the Borel subgroups of modality 0 were determined in characteristic 0:

**Theorem 4** [Kashin 1990]. *Let $G$ be quasisimple over $k$, where $k$ has characteristic* 0, *and suppose $B$ is a Borel subgroup of $G$. The number of orbits of $B$ on $\mathfrak{n}$ is finite* (*that is*, *$B$ has modality* 0) *if and only if $G$ is type $A_n$ for $n \leq 4$, or $G$ is type $B_2$.*

Aside from the consequences of this theorem for our investigation on nilpotent $B$-orbits, Kashin's result launched an investigation into the modality of parabolic subgroups in general. For example, see [Röhrle 1996; 1999; Popov 1997; Popov

and Röhrle 1997; Hille and Röhrle 1999; Brüstle et al. 1999]. In fact, Theorem 1.1 in [Hille and Röhrle 1999] shows there is a strong connection between the modality of a parabolic subgroup and the length of a descending central series of $R_u(P)$, the unipotent radical of $P$, also called the *nilpotency class* of $R_u(P)$. Using this theorem, one can easily recover Kashin's original theorem, with the added benefit that the proof is valid in good prime characteristics for $G$ as well as for characteristic 0. In type $A$, all primes are good, and in type $B$, all primes are good except $p = 2$.

In a previous version of this paper, using similar techniques as here, we showed directly that there are finitely many nilpotent $B$-orbits for $G$ of types $A_1$, $A_2$, $A_3$ and $A_4$ without any restrictions on $p$, and used that information to determine the dimensions of the orbits and the closure ordering. A referee pointed out to us that the closure orderings for the four type-$A$ cases were already discussed in [Brüstle et al. 1999], making a lot of our work seem redundant. Note that the techniques used in that paper are quite different than ours — they are much more sophisticated than our matrix calculations. Their approach has some advantages, such as both being more elegant than our approach and also being closer in spirit to the way nilpotent $G$-orbits are classified. A possible advantage of our techniques, though, is that they yield the explicit polynomial defining equations of each orbit. It may be an advantage to knowing these defining equations in applying this work, perhaps to computing support varieties of baby Verma modules as discussed in Section 1, or perhaps for other applications. For this reason, we have uploaded our type-$A$ calculations [Burkhart and Vella 2017] on the arXiv so that despite the overlap with [Brüstle et al. 1999], our tables for these orbits are publicly available. Here we conclude by simply reminding the reader how many orbits there are in each case: two orbits in type $A_1$, five orbits in type $A_2$, 16 orbits in type $A_3$, and 61 orbits in type $A_4$. For the details of the defining equations, etc., consult [Burkhart and Vella 2017], and for the dimensions of each orbit and the Hasse diagrams of the closure order in these cases, valid for all characteristics, consult either [Brüstle et al. 1999] or [Burkhart and Vella 2017].

## References

[Borel 1991] A. Borel, *Linear algebraic groups*, 2nd ed., Graduate Texts in Mathematics **126**, Springer, 1991. MR Zbl

[Bourbaki 2002] N. Bourbaki, *Lie groups and Lie algebras: Chapters 4–6*, Springer, 2002. MR Zbl

[Brüstle et al. 1999] T. Brüstle, L. Hille, G. Röhrle, and G. Zwara, "The Bruhat–Chevalley order of parabolic group actions in general linear groups and degeneration for $\Delta$-filtered modules", *Adv. Math.* **148**:2 (1999), 203–242. MR Zbl

[Bürgstein and Hesselink 1987] H. Bürgstein and W. H. Hesselink, "Algorithmic orbit classification for some Borel group actions", *Compositio Math.* **61**:1 (1987), 3–41. MR Zbl

[Burkhart and Vella 2017] M. Burkhart and D. Vella, "Defining equations of nilpotent orbits for Borel subgroups of modality zero in type $A_n$", preprint, 2017. arXiv

[Carter 1985] R. W. Carter, *Finite groups of Lie type: conjugacy classes and complex characters*, John Wiley & Sons, New York, 1985. MR Zbl

[Collingwood and McGovern 1993] D. H. Collingwood and W. M. McGovern, *Nilpotent orbits in semisimple Lie algebras*, Van Nostrand Reinhold Co., New York, 1993. MR Zbl

[Friedlander and Parshall 1986] E. M. Friedlander and B. J. Parshall, "Support varieties for restricted Lie algebras", *Invent. Math.* **86**:3 (1986), 553–562. MR Zbl

[Hille and Röhrle 1999] L. Hille and G. Röhrle, "A classification of parabolic subgroups of classical groups with a finite number of orbits on the unipotent radical", *Transform. Groups* **4**:1 (1999), 35–52. MR Zbl

[Humphreys 1972] J. E. Humphreys, *Introduction to Lie algebras and representation theory*, Graduate Texts in Mathematics **9**, Springer, 1972. MR Zbl

[Humphreys 1975] J. E. Humphreys, *Linear algebraic groups*, Graduate Texts in Mathematics **21**, Springer, 1975. MR Zbl

[Jantzen 2004] J. C. Jantzen, "Nilpotent orbits in representation theory", pp. 1–211 in *Lie theory*, edited by J.-P. Anker and B. Orsted, Progr. Math. **228**, Birkhäuser, Boston, 2004. MR Zbl

[Kashin 1990] V. V. Kashin, "Orbits of an adjoint and co-adjoint action of Borel subgroups of a semisimple algebraic group", pp. 141–158 in *Problems in group theory and homological algebra*, edited by A. L. Onishchik, Yaroslav. Gos. Univ., Yaroslavl, 1990. In Russian. MR Zbl

[Nakano et al. 2002] D. K. Nakano, B. J. Parshall, and D. C. Vella, "Support varieties for algebraic groups", *J. Reine Angew. Math.* **547** (2002), 15–49. MR Zbl

[Popov 1997] V. L. Popov, "A finiteness theorem for parabolic subgroups of fixed modality", *Indag. Math.* (*N.S.*) **8**:1 (1997), 125–132. MR Zbl

[Popov and Röhrle 1997] V. Popov and G. Röhrle, "On the number of orbits of a parabolic subgroup on its unipotent radical", pp. 297–320 in *Algebraic groups and Lie groups*, edited by G. Lehrer et al., Austral. Math. Soc. Lect. Ser. **9**, Cambridge Univ. Press, 1997. MR Zbl

[Röhrle 1996] G. Röhrle, "Parabolic subgroups of positive modality", *Geom. Dedicata* **60**:2 (1996), 163–186. MR Zbl

[Röhrle 1999] G. Röhrle, "On the modality of parabolic subgroups of linear algebraic groups", *Manuscripta Math.* **98**:1 (1999), 9–20. MR Zbl

Received: 2017-08-16 Revised: 2018-02-08 Accepted: 2018-07-10

burkhm2@uw.edu *Mathematics Department, University of Washington, Seattle, WA, United States*

dvella@skidmore.edu *Mathematics and Statistics Department, Skidmore College, Saratoga Springs, NY, United States*

INVOLVE 12:3 (2019)
dx.doi.org/10.2140/involve.2019.12.463

# Homophonic quotients of linguistic free groups German, Korean, and Turkish

Herbert Gangl, Gizem Karaali and Woohyung Lee

(Communicated by Kenneth S. Berenhaut)

The homophonic quotient groups for French and English (i.e., the quotient of the free group generated by the French/English alphabet determined by relations representing standard pronunciation rules) were explicitly characterized by Mestre et al. (1993). We apply the same methodology to three different language systems: German, Korean, and Turkish. We argue that our results point to some interesting differences between these three languages (or at least their current script systems).

## 1. Introduction

Mestre et al. [1993] explicitly characterized the homophonic quotient groups for French and English (i.e., the quotient of the free group generated by the French/English alphabet determined by relations representing standard pronunciation rules). Some references mention an analogous characterization for Japanese, but that result does not seem to be easily accessible.

In this paper we apply the same methodology to three different language systems: German, Korean, and Turkish. The analysis for German was circulated in unpublished form for a while; the Korean and the Turkish analyses are new. As we suggest in the final section of this paper, our results may point to some interesting differences between these three languages (or at least their current script systems).

The paper is organized in a straightforward manner, with each numbered section presenting the analysis for one language. In particular Section 2 presents our results for German, Section 3 presents our results for Korean, and Section 4 presents our results for Turkish. A final section brings together these analyses and offers some thoughts on what we might gain from this comparative study.

*MSC2010:* primary 00A69; secondary 20F05.
*Keywords:* quotient groups, free groups, homophones.

## 2. German

In their phonetically calibrated paper [Mestre et al. 1993], Mestre, Schoof, Washington, and Zagier showed that the homophonic quotient of the free group on the 26 letters of the alphabet is trivial for both the French and the English languages. As already foreshadowed in that paper, we obtain the same answer for the German language.

Let $G$ be the quotient of the free group on 26 letters a, b, c,..., z by the relations $A = B$ provided there are words $A$ and $B$ in the German language whose pronunciations agree.

We justify the term "agree" by invoking standard dictionaries like [Duden 1986; 1990], whose name "Duden" has become synonymous with the official norm, as well as its online version http://www.duden.de/suchen/dudenonline. Alternatively, for most of the pairs of words below, we can use an automatic phonetic converter such as the one at http://familientagebuch.de/rainer/2007/38.html#4.

**Theorem 1.** *The group G is trivial.*

*Proof.* We successively eliminate letters using specific properties of spoken German. For homophonicity we need to distinguish in particular between long and short vowels as well as between voiced and unvoiced consonants.

***Vowels*** (methods of idempotents, see [Mestre et al. 1993], and of vanishing with 'h').

(a) For instance, 'aa', 'ah' and 'a' may often be pronounced alike, in particular they often have the same length, like "Waage" [scales] and "wage" [(I) dare] or "Wahl" [choice] and "Wal" [whale].

(e) Similarly, 'ee', 'eh' and sometimes 'e' can sound the same: "Meer" [the sea] and "mehr" [more].

(o) Both 'oo' and 'oh' are often used within words, and can both be pronounced like a single 'o' ("Boot" [boat] and "bot" [(he) offered], and "hohle" [hollow (pl.)] vs. "hole" [(I) fetch]).

We note that for 'i' and 'u' the corresponding identifications do not work; e.g., while both 'ie' and 'ih' indicate a long 'i', the former can never occur at the beginning of a word where it is instead replaced by the second one ("ihnen" [(to) them], "ihr" [her]), and 'ii' in a word (like "liieren" [(to) liaise]) is pronounced with a glottal stop between the 'i's. Similarly, the 'uu' in words like "Kontinuum" or "Trauung" indeed comes across as two 'u's, and there aren't any words with a 'uu' that would sound like a long 'u', say. Hence we need to treat these two vowels separately.

***Consonants***.

(g/b/n) (voiceless in the end) At the end of a word, a voiced consonant is pronounced in the same way as the corresponding unvoiced one (like "Bug" [(nautical) bow]

and “buk” [(he) baked], or “Alb” and “Alp” [both for nightmare]). Similarly, an ‘nn’ at times sounds like a single ‘n’ (“Mann” [man] and “man” [one/you (pronoun)]).

(v/w) (WVF?) The consonant ‘v’ is typically pronounced in one of two ways: like ‘f’ or like ‘w’, depending mostly on the etymological origin of the word (“viel” [many] vs. “fiel” [(he) fell] and “vage” [vague] vs. “wage” [(I) dare]).

(l/r/f/p/s) (idempotents) By combining certain consonants we can further minimize the influence of a single contributing consonant, so while it is hard to find the same sounds for ‘ll’ and ‘l’ at the end of a word, one can add a ‘t’ to it and succeed (“hallt” [(it) reverberates/echoes] vs. “Halt” [halt]). Similar comments apply to ‘rr’ and ‘r’ (“starrt” [(he) stares] vs. “Start” [start]), for ‘ff’ and ‘f’ (“schafft” [(he) manages] vs. “Schaft” [shaft]), as well as ‘pp’ and ‘p’ (“klappst” [(you) flap/fold] vs. ‘klapst’ [(he) claps lightly]; alternatively, “schnippst” and “schnipst”, both from schnipsen [(to) clip]). Furthermore, “fasst” [(he) catches] and “fast” [almost] are homophonic.

(t/d) (little ‘dt’ for ‘tt’) A related case is the combination ‘th’ which also often ensures that a preceding vowel is pronounced as a short one: e.g., “Zithern” [zithers] and “zittern” [(to) tremble] are pronounced the same way; another means to the same end is the use of ‘dt’ in place of ‘tt’, giving, e.g., that “Stadt” [city] and “statt” [instead of] are homophonic.

(m) A variant of the idempotent method, using also the voiced/unvoiced consonant at the end of a word, is “hemmt” [hinders] vs. “Hemd” [shirt].

(c) (departing of the ‘c’) Other constructs that make sure that a vowel is short are to follow it up with a ‘ck’ rather than a ‘k’; for example, the words “packt” [(he) packs] and “Pakt” [(a) pact] sound alike. Note, however, that in a very similar setting the words “hackt” [(he) hacks] and “hakt” [(he) hooks] are pronounced differently, as the latter ‘a’ then denotes a *long* vowel.

(z) A further peculiarity is the pronunciation of ‘z’, typically equivalent to the combination ‘t-s’ (with obvious exceptions for loanwords like “Jazz” where the educated citizen will make an attempt to sound more anglophonic), so we can identify the genitive “Kitts” of “Kitt” [glue] with “Kitz” [fawn].

(x) In the same vein as ‘z’, the letter ‘x’ is pronounced ‘k-s’ which is also the pronunciation of ‘chs’ (i.e., when ‘ch’ precedes ‘s’ it often becomes ‘k’), so we find “lax” [lax] to be homophonic to “Lachs” [salmon].

The remaining letters ‘k’, ‘u’, ‘i’, ‘y’, ‘j’ and ‘q’ are somewhat harder to trivialize, but modulo the above this is doable, albeit by using loanwords from different languages (English, Italian, Hungarian).

(k) The English word “Clip” for office equipment is often used and is homophonic to “klipp” (e.g., from “klipp und klar” [in no uncertain terms]).

(u) The Italian word “ciao” has been assimilated as “tschau”, both terms being used.

(i) The word “roien” [(to) row] is homophonic to “reuen” [(to) rue], the former being used mainly in “Niederdeutsch”, i.e., in the north of Germany. Alternatively, the loanword (from the English language) “beaten” [(to) make beat music] is acceptable according to [Duden 1990], and it is homophonic to “bieten” [(to) offer].

(y) The word “toi” from the saying “toi, toi, toi” [break a leg] sounds like “Toy” [sex toy]. Alternatively, a “Bayer” [Bavarian] can be spelled “Baier”. (We could also invoke the ambiguous spellings of “Yoghurt” and “Joghurt”. For yet another possibility, the Hungarian word “Gulyas” [goulash] has been assimilated also as “Gulasch”.)

(j) As to ‘j’, we use the word “Yak” [yak] (from the Tibetan “gyag”) and its similarity to “Jacke” [jacket], which are not homophonic as such, but their respective diminutives “Yäkchen” and “Jäckchen” (note the ensuing umlaut for either case) are.

(q) Finally, for the quite rare letter ‘q’, we can use the French word “clique” (which has been adapted into German with a short ‘i’), whose pronunciation agrees with that of “klicke” [(I) click]. Another possibility is to note that the letter “Q” itself can be used as a word (say, as the Q in a game of Scrabble) and is homophonic to “Kuh” [cow].

In the table below we successively eliminate the letters on the left using the homophonic ambiguity displayed on their right, completing the proof of the theorem:

| | | | |
|---|---|---|---|
| a | Waage — wage | s | fasst — fast |
| h | Wahl — Wal | t | Zittern — Zithern |
| e | Meer — mehr | d | Stadt — statt |
| o | Boot — bot | m | hemmt — Hemd |
| g | Bug — buk | c | packt — Pakt |
| b | Alb — Alp | z | Kitz — Kitts |
| n | Mann — man | x | lax — Lachs |
| v | viel — fiel | k | klipp — Clip |
| w | wage — vage | u | tschau — ciao |
| l | gewallt — Gewalt | i | roien — reuen |
| r | starrt — Start | y | Toy — toi |
| f | schafft — Schaft | j | Jäckchen — Yäkchen |
| p | klappst — klapst | q | Clique — klicke |

□

***Generalizations.*** One can also try to include the umlaute ‘ä’, ‘ö’, ‘ü’, and the “sharp s” ß into these investigations. The result remains the same. Our suggestion for the corresponding trivializations are the following: For ‘ä’ we invoke that in combination with ‘u’ the diphthongs ‘äu’ and ‘eu’ sound alike, for instance in the words “häutig” [of a skinny texture] and “heutig” [contemporarily]; alternatively, we can use that a long ‘ä’ can sound like the ‘ai’ for certain loanwords from the English language, for example in “Fähre” [ferry] and “faire” [fair]. For ‘ö’ we use that certain words are spelled with both the original French ‘eu’ and the assimilated German ‘ö’, like “Frisör” and “Friseur”. Furthermore, the pronunciation of ‘ü’ is often the same as that of ‘y’, like in the Greek letter “My” [mu] and “müh” [(I) labor], or, a far better one due to Martin Brandenburg, “Mythen” [myths] and “mühten” [(they) labored]. Finally, a ‘sharp s’ at the end of a word is typically preceded by a long vowel, and hence it is not difficult to construct word pairs like “aß” [(I) ate] and “Aas” [(rotten) carcass]:

| | |
|---|---|
| ä | häutig — heutig |
| ö | Frisör — Friseur |
| ü | müh – My |
| ß | aß — Aas |

## 3. Korean

What differentiates Korean from the languages discussed in [Mestre et al. 1993] is the number of alphabets and some fundamental grammar structures. Nevertheless, there exist many rules regarding homophones, so the first natural assumption would be that the resulting quotient group shouldn’t have too many elements. It turns out that this is indeed the case.

Here we note that this mathematical analysis of Korean does not describe the entire structure of the Korean language. It takes the phonetic aspect of the language and restructures the alphabets into a free group with a very specific and somewhat restrictive equivalence relation. Using such a structure, we inevitably lose a lot of information about the Korean language, but are, however, rewarded with a unique finite group that characterizes it.

Now, let us begin with describing some necessary concepts about the Korean language.

**3.1.** ***Some basics of Korean.*** Korean characters, like English, consist of *vowels* and *consonants*. The alphabet contains 19 consonants and 21 vowels. The exact list is shown below in Table 1.

Because of the complications arising from the unique structure of Korean, from here on, each of the above symbols in the table will be called a *character*. To show why such clarification is crucial, let us take a look at a Korean word that stands for

| | |
|---|---|
| consonants | ㄱ ㄲ ㄴ ㄷ ㄸ ㄹ ㅁ ㅂ ㅃ ㅅ ㅆ ㅇ ㅈ ㅉ ㅊ ㅋ ㅌ ㅍ ㅎ |
| vowels | ㅏ ㅐ ㅑ ㅒ ㅓ ㅔ ㅕ ㅖ ㅗ ㅘ ㅙ ㅚ ㅛ ㅜ ㅝ ㅞ ㅟ ㅠ ㅡ ㅢ ㅣ |

**Table 1.** Korean characters [KLI].

"number". It is written as 수. This word is composed of a single letter, and that letter is composed of a consonant and a vowel, which are, in this case, ㅅ and ㅜ. These letters form the bases of Korean words, as no single consonant or vowel is ever used alone without the other. However, this is not the end.

To add to the already complex structure, a single letter can be made up of multiple consonants and a vowel, up to three consonants and one vowel. Denoting vowels and consonants as $v$ and $c$ in respective order, the possible combinations are $\{c+v,\ c+v+c,\ c+v+c+c\}$. Henceforth, expressions of the form $c+v+\cdots$ will be called ordered decompositions. The fact that these are the only combinations, however, effectively erases the need to distinguish between letters and combinations of characters. We present the needed argument below.

**Theorem 2.** *Ordered decomposition uniquely encodes any formal composition of Korean letters or words. Equivalently, the formal expression of a Korean word is uniquely encoded in the ordered decomposition.*

*Proof.* A letter in Korean is always given as one of $c+v,\ c+v+c$, or $c+v+c+c$. In particular, it always begins with $c+v$, and hence identifying each $c+v$ in the ordered decomposition allows us to retrieve the unique formal expression of the corresponding Korean word. □

Now that we've established some basics we will examine the homophonic structure of the quotient group $G$ of the free group on 40 Korean characters, given by the equivalence relation $A = B$ whenever $A$ and $B$ have the same pronunciation in Korean. We will use a standard pronunciation guide such as [KLI] for reference. Also we will use 1 to denote the empty word as we analyze the group structure of Korean.

**3.2.** ***Triviality of consonants.*** We first show that all consonants are trivial. We do this in three steps.

**3.2.1.** *ㅇ is trivial.* To show this, let us take a look at the word 안일하다 [to be idle]. 안일하다 has exactly the same pronunciation as 아닐하다 [KLI]. Just by looking at the two words, 하다 is present on both sides, so it can be canceled out. Now the equivalence relation is between ㅇ+ㅏ+ㄴ+ㅇ+ㅣ+ㄹ and ㅇ+ㅏ+ㄴ+ㅣ+ㄹ. Clearly after canceling out, ㅇ = 1, and hence ㅇ is trivial.

**3.2.2.** *ㄱ = ㄲ = ㅋ, ㄷ = ㅅ = ㅆ = ㅈ = ㅊ = ㅌ, ㅂ = ㅍ.* To show the above equivalence relations, let us examine words containing letters of the form $c+v+c$. First,

we will examine the words 부엌 [kitchen] and 밖 [outside]. By the equivalence relation defined above, 부엌=부억 and 밖=박. By rewriting these relations in ordered decomposition, ㅂ+ㅜ+ㅇ+ㅓ+ㅋ=ㅂ+ㅜ+ㅇ+ㅓ+ㄱ and ㅂ+ㅏ+ㄲ=ㅂ+ㅏ+ㄱ. Now it is clear that ㄱ=ㅋ and ㄱ=ㄲ. By the transitive property ㅋ=ㄲ=ㄱ.

For the second part we can examine the equivalence relations, 낫 [scythe] =낟, 낮 [day] =낟, 낯 [face] =낟, 밭 [field] =받. Clearly ㄷ=ㅅ=ㅈ=ㅊ=ㅌ. Proving ㄷ=ㅆ is a bit more difficult as there are no single-letter words in Korean ending in ㅆ. To prove this we need to look at the two-letter word 불소 [fluorine]. By the equivalence relation 불소= 불쏘, we clearly have ㅅ=ㅆ. Since we already know ㅅ=ㄷ, by the transitive property, ㄷ= ㅆ, thus concluding our proof of the second equivalence relation.

For the last equivalence relation, we can look at 짚 [hay] = 집, and can conclude that ㅂ = ㅍ.

By proving these relations, we've reduced the set of consonants to {ㄱ,ㄴ,ㄷ,ㄸ, ㄹ,ㅁ,ㅂ,ㅃ,ㅉ,ㅎ}.

**3.2.3.** *Consonants are trivial.* To further reduce the set of consonants let us look at the equivalence relation 앞마당 [lawn] = 암마당. This shows that ㅁ = ㅍ, and ㅍ = ㅂ, so ㅁ = ㅂ. Additionally, 있는 [existing] = 인는, and so ㄴ = ㅆ = ㄷ. Also, 국물 [soup] = 궁물, and 놓는 [lay down] = 논는, so ㄱ is trivial and ㄷ = ㄴ = ㅎ. Observe that 숱하다 [to be in abundance] = 수타다, which shows that ㅎ is also trivial. Since ㄷ = ㅎ and ㅎ is trivial, ㄷ is also trivial. Now, there only remain five nontrivial consonants, {ㄸ,ㄹ,ㅂ,ㅃ,ㅉ}.

Let's look at the equivalence relation 웃다 [smile] = 울따, which in ordered decomposition is ㅇ + ㅜ + ㅅ + ㄷ + ㅏ = ㅇ + ㅜ +ㄷ +ㄸ + ㅏ. We know ㅇ,ㅅ = ㄷ are trivial, so ㅜ + ㅏ = ㅜ + ㄸ + ㅏ. Hence ㄸ = 1 and so ㄸ is also trivial. 약지 [ring finger] = 약찌; hence ㅉ = ㅈ = ㄷ = 1 and ㅉ is trivial. 막론 [whether] = 망논, which can be rewritten as ㅁ+ㅏ+ㄱ+ㄹ+ㅗ+ㄴ = ㅁ+ㅏ+ㅇ+ㄴ+ㅗ+ㄴ, and since ㄱ,ㄴ = ㄷ are both trivial, ㅁ+ㅏ+ㄹ+ㅗ = ㅁ+ㅏ+ㅗ, and so ㄹ is trivial. 국밥 [soup and rice] = 국빱 implies that ㅂ = ㅃ. There remains only one nontrivial consonant, ㅂ.

Lastly we need to examine a word with a letter of the form $c + v + c + c$. 넓다 [wide] = 널따, and we know that ㄴ,ㄹ,ㄷ,ㄸ are all trivial. So, after canceling both sides, we have ㅓ + ㅂ + ㅏ = ㅓ + ㅏ, and so ㅂ is trivial. Hence we've proved the triviality of all Korean consonants.

**3.3.** ***Vowels have two nontrivial elements*: *vowels* = {ㅏ,ㅗ}.** While in examining consonants we only needed to look at a single equivalence relation, vowels are not so easy. There are multiple equivalence relations between three or more vowels, so we need to sort through these relations to see how they can be reduced. Furthermore,

many words in the Korean language have reduced forms, where a letter of the form ㅇ$+v$ is merged with the previous letter, as witnessed in 되어 = 돼. For our analysis of Korean, we will also take such reductions as equivalence relations.

Let us begin examining the vowels by the equality 가지어 [have] = 가져 = 가저. All consonants are trivial, so we can simply look at ㅏ + ㅣ + ㅕ = ㅏ + ㅖ = ㅏ + ㅕ. From this we can conclude that ㅣ is trivial and ㅕ = ㅖ. Also, 통계 [statistics] = 통게 implies that ㅖ = ㅔ, and 희미 [faint] = 히미 implies ㅢ = ㅣ. Since ㅣ is trivial, ㅢ is trivial, and 금괴 [gold bar] = 금궤 [metal box] implies ㅚ = ㅞ. 누이다 = 뉘다 (to lay down) implies 위 can be identified with 우 + 이, so by the triviality of ㅣ, we have ㅟ = ㅜ. 되어 = 돼 implies ㅚ + ㅕ = ㅗ + ㅕ = ㅙ = ㅗ + ㅐ, 싸이다 = 쌔다 implies ㅏ + ㅣ = ㅏ = ㅐ, and 트이다 = 틔다 = 티다 implies that ㅡ is also trivial. Continuing we have 미아 = 먀 implies ㅏ = ㅑ and 쏘이다 = 쐬다 implies ㅗ + ㅣ = ㅗ = ㅚ. Also, ㅚ = ㅞ, so ㅗ = ㅚ = ㅞ. Lastly, there are consonants whose pronunciations are defined as the combination of two other consonants, such as ㅛ = ㅣ + ㅗ.

Listing these rules that do not overlap into an easily decipherable form we get:

(1) ㅣ is trivial.
(2) ㅕ = ㅖ.
(3) ㅖ = ㅔ.
(4) ㅢ = ㅣ.
(5) ㅚ = ㅞ.
(6) ㅟ = ㅜ + ㅣ $\iff$ ㅟ = ㅜ.
(7) ㅚ + ㅕ = ㅙ.
(8) ㅏ + ㅣ = ㅐ $\iff$ ㅏ = ㅐ.
(9) ㅡ + ㅣ = ㅣ $\iff$ ㅡ = 1.
(10) ㅣ + ㅏ = ㅑ $\iff$ ㅏ = ㅑ.
(11) ㅗ + ㅣ = ㅚ $\iff$ ㅗ = ㅚ.
(12) ㅢ = ㅔ $\iff$ ㅔ = 1.
(13) ㅗ + ㅐ = ㅙ.
(14) ㅝ = ㅜ + ㅕ.
(15) ㅛ = ㅣ + ㅗ.
(16) ㅠ = ㅣ + ㅜ.

Just by looking at these rules, we can reduce the set {ㅏ, ㅐ, ㅑ, ㅒ, ㅕ, ㅖ, ㅕ, ㅖ, ㅗ, ㅘ, ㅙ, ㅚ, ㅛ, ㅜ, ㅝ, ㅞ, ㅟ, ㅠ, ㅡ, ㅢ, ㅣ} of vowels into {ㅏ, ㅕ, ㅗ, ㅘ, ㅙ, ㅛ, ㅜ, ㅝ, ㅞ, ㅠ}.

Now we outline the rest of the process:

- (13) and (8) combine to show that ㅗ + ㅐ = ㅗ + ㅏ = ㅘ = ㅙ, implying ㅘ = ㅙ.
- (5) and (11) combine to show that ㅗ = ㅞ = ㅜ + ㅔ, and since (12) states that ㅔ is trivial, ㅗ = ㅜ.
- As a direct result of ㅗ = ㅜ, (14), (13), (11), (7) and (1), we have ㅝ = ㅜ + ㅕ = ㅗ + ㅕ = ㅗ + ㅣ + ㅕ = ㅚ + ㅕ = ㅙ = ㅗ + ㅐ. So ㅝ = ㅙ = ㅘ.
- (1) and (15) together show that ㅗ = ㅛ.
- Since ㅠ = ㅣ + ㅜ and we've concluded that ㅗ = ㅜ, we have ㅠ = ㅣ + ㅜ = ㅣ + ㅗ = ㅗ.

- Recall that from (7) and (13), we have ㅢ + ㅓ = ㅐ = ㅗ + ㅐ. However, ㅢ = ㅗ + ㅣ = ㅗ, so ㅗ + ㅓ = ㅗ+ ㅐ, and by canceling ㅗ, we have that ㅓ = ㅐ.
- (8) states that ㅏ = ㅐ, so with the above result, ㅏ = ㅓ.
- Since ㅘ = ㅗ+ ㅏ, we have ㅘ is generated by ㅗ and ㅏ.

***Conclusion.*** The homophonic quotient of the Korean language can be written in terms of two generators { ㅏ, ㅗ }.

In some sense, we have identified the two most fundamental characters in Korean as their pronunciations are not discarded in any Korean word they appear in. Furthermore because we have allowed for the equivalence of words and their reduced forms, it is unlikely that the set of distinct Korean characters under homophonic quotients can be further reduced. However, distinctions between pronunciations of certain vowels are becoming more obscure; hence it is possible that after appropriately adjusting the formal pronunciation rules to accommodate such trends, the homophonic quotient group of Korean is further reduced.

## 4. Turkish

In this section we determine the homophonic quotient group for Turkish. There are several Turkic languages, and alphabets encoding them have many commonalities. We will exclusively focus on the modern Turkish alphabet.

**4.1.** ***The sounds of Turkish.*** The modern Turkish alphabet was introduced in 1928 along with a wide-reaching literacy campaign. The Latin-based script was developed to replace the use of the Arabic script, and contains a total of 29 letters (8 vowels and 21 consonants) as seen in Table 2.

This set of letters was specifically selected to represent the sounds present in the spoken language of the time, taking the Istanbul dialect as the standard. Each letter is supposed to represent a unique sound of the spoken language (except the so-called "soft g", ğ, which tends to extend the vowel before it and blends it to the following vowel if there is one, but is otherwise completely silent; see [Logacev et al. 2014] for more on the "soft g"). For more on the sound system of modern Turkish, see [Yavuz and Balci 2011].

To this day the modern Turkish script retains most of its phonetic representativeness [Kopkalli-Yavuz 2010]. Indeed many hold that there are no homophones in

| | |
|---|---|
| consonants | b c ç d f g ğ h j k l m n p r s ş t v y z |
| vowels | a e ı i o ö u ü |

**Table 2.** Letters of the modern Turkish script (only lowercase letters are given).

Turkish; see for instance [Raman and Weekes 2005], where Turkish is described as a "completely transparent writing system" with "invariant and context-independent one-to-one mappings between orthography and phonology".

This suggests that the free group generated by the 29 sound representatives will not shrink much if at all when we try introducing homophonic equivalences. Nonetheless there are indeed some relations we might use if we consider "how words are actually pronounced by real live people".[1]

**4.2.** ***The "soft g" disappears.*** As noted above the "soft g" is often not a distinctly pronounced consonant but instead helps to accentuate or blend the surrounding vowels. Most native speakers would agree that we can identify the following encodings of the male name meaning "Khan":

$$\text{Kaan} = \text{Kağan}.$$

Thus in the quotient group we would identify the "soft g" with identity.

**4.3.** ***Other disappearing acts: 'h' and 't'.*** The standard pronunciation of the word "dershane" [classroom] overlaps with the pronunciation of "dersane", thus allowing us to conclude that 'h' too is trivial in the quotient. Similarly the double 't's in the words "Hacettepe" and "Anıttepe" [two location names in Ankara] are most commonly pronounced as if they were written as "Hacetepe" and "Anıtepe" respectively. Thus we can identify 'tt' with 't', trivializing 't'.

**4.4.** ***Vowel confusion: the transformations of 'a' and 'e' into 'ı' and 'i' and two final disappearing acts.*** The Turkish language captures the phrase "let me look" in the single word "bakayım". The native speaker pronounces the latter in the same way that she would read the letter collection "bakıyım". This allows us to identify $\mathrm{a} = \imath$. Similarly the phrase "içecek" [drink] is pronounced the same way that one would read "içicek" and so we identify $\mathrm{e} = \mathrm{i}$.

Finally the word "ağabey" [older brother] has an almost universally accepted informal spelling, "abi", representing the way people actually pronounce the word. Together with the sound equivalence of "ağa" [master, land owner] and "ağ" [network], this gives us two additional trivializations, of 'a' and 'y'.

Putting the above reductions together we conclude that the homophonic quotient group for Turkish is a free group on 22 generators:

$$\mathrm{b}\ ,\ \mathrm{c}\ ,\ \mathrm{ç}\ ,\ \mathrm{d}\ ,\ \mathrm{e}\ (=\mathrm{i})\ ,\ \mathrm{f}\ ,\ \mathrm{g}\ ,\ \mathrm{j}\ ,\ \mathrm{k}\ ,\ \mathrm{l}\ ,\ \mathrm{m}\ ,\ \mathrm{n}\ ,\ \mathrm{o}\ ,\ \mathrm{ö}\ ,\ \mathrm{p}\ ,\ \mathrm{r}\ ,\ \mathrm{s}\ ,\ \mathrm{ş}\ ,\ \mathrm{u}\ ,\ \mathrm{ü}\ ,\ \mathrm{v}\ ,\ \mathrm{z}$$

[1] In his MathSciNet review (MR1273406), James Wiegold notes that the authors of [Mestre et al. 1993] "have [perhaps deliberately?] neglected all considerations of how words are actually pronounced by real live people." Clearly if we were to take into consideration each native speaker's distinct pronunciation patterns, the homophonic quotients problem would become quite intractable. However we will indeed introduce some of this complication into our analysis of Turkish. This may be justified by the fact that there is deemed to be a standard spoken Turkish, and it is indeed distinct from most formal descriptions of the orthography/phonology correspondence for the language.

## 5. Final words: bringing the three threads together

In this paper we investigated three different languages and their writing systems. We believe that our results offer an interesting example of applied algebra. We explored how the writing system of a modern language and its correspondence with the sounds of that language can be encoded in group theory. Other algebraic structures have been identified in various symmetrical constructions of nature such as crystals, as well as in a range of sociological and anthropological contexts such as the kinship structure of the Warlpiri of Australia.[2]

It is important to note that our methods do not address the full phonetic structure of any single language. Our work only pertains to the relationship between orthography and phonology of a language, that is, the extent to which a single symbol may represent a multiplicity of sounds of a given language. A simplistic interpretation of our method would suggest that if the generating set for the resulting quotient group is small, there are, on average, more sounds represented by a single symbol.

We should also note that the complexity of the resulting group may be correlated not directly with the complexity of the sound system of a given language but perhaps more with the maturity of the particular writing system associated to it. Languages evolve, and oral traditions evolve much faster than written ones. Thus a young script like modern Turkish might be naturally more representative of the phonetical structure of the language and equivalently offer fewer homophones than a script which is more mature, such as the Korean one, which in turn may offer fewer homophones than an even older script such as the German one.

## Acknowledgment

The authors thank the reviewer for helpful suggestions. Gangl is grateful to the MPI Bonn for providing ideal working conditions and in particular to Don Zagier for setting the original challenge.

## References

[Ascher 1991] M. Ascher, *Ethnomathematics: a multicultural view of mathematical ideas*, Chapman & Hall, New York, 1991. MR Zbl

[Duden 1986] K. Duden, *Rechtschreibung der deutschen Sprache und der Fremdwörter*, edited by D. Berger and W. Scholze, Der Duden in 10 Bänden **1**, Duden, Mannheim, 1986.

[Duden 1990] K. Duden, *Duden Aussprachewörterbuch: Wörterbuch der deutschen Standardaussprache*, edited by M. Mangold, Der Duden in 10 Bänden **6**, Duden, Mannheim, 1990.

[KLI] "Romanization of Korean", website, The National Institute of Korean Language, available at http://www.mcst.go.kr/english/koreaInfo/language/romanization.jsp.

[2] As ethnomathematician Marcia Ascher describes in detail in her book, the kinship structure of the Warlpiri, an indigenous people in Australia, can be accurately and succinctly represented by the dihedral group of order 8. See Chapter 3 of [Ascher 1991] for details.

[Kopkalli-Yavuz 2010] H. Kopkalli-Yavuz, "The sound inventory of Turkish: consonants and vowels", in *Communication disorders in Turkish*, edited by S. Topbaş and M. S. Yavas, Communication disorders across languages **4**, Multilingual Matters, Bristol, 2010.

[Logacev et al. 2014] O. U. Logacev, S. Fuchs, and M. Żygis, "Soft 'g' in Turkish: evidence for sound change in progress", pp. 437–440 in *Proceedings of the 10th International Seminar on Speech Production* (Cologne, 2014), edited by S. Fuchs et al., University of Cologne, 2014.

[Mestre et al. 1993] J.-F. Mestre, R. Schoof, L. Washington, and D. Zagier, "Quotients homophones des groupes libres = Homophonic quotients of free groups", *Experiment. Math.* **2**:3 (1993), 153–155. MR Zbl

[Raman and Weekes 2005] I. Raman and B. S. Weekes, "Deep disgraphia in Turkish", *Behavioural Neurology* **16**:2-3 (2005), 59–69.

[Yavuz and Balci 2011] H. Yavuz and A. Balci, *Turkish phonology and morphology*, edited by Z. Balpinar, Anadolu University, Eskişehir, 2011.

Received: 2017-09-11 Revised: 2018-08-03 Accepted: 2018-08-04

herbert.gangl@durham.ac.uk *Department of Mathematical Sciences, Durham University, Durham, United Kingdom*

gizem.karaali@pomona.edu *Department of Mathematics, Pomona College, Claremont, CA, United States*

leew16@wfu.edu *Wake Forest University, Winston-Salem, NC, United States*

INVOLVE 12:3 (2019)
dx.doi.org/10.2140/involve.2019.12.475

# Effective moments of Dirichlet $L$-functions in Galois orbits

Rizwanur Khan, Ruoyun Lei and Djordje Milićević

(Communicated by Stephan Garcia)

Khan, Milićević, and Ngo evaluated the second moment of $L$-functions associated to certain Galois orbits of primitive Dirichlet characters to modulus a large power of any fixed odd prime $p$. Their results depend on $p$-adic Diophantine approximation and are ineffective, in the sense of computability. We obtain an effective asymptotic for this second moment in the case of $p = 3, 5, 7$.

## 1. Introduction

Dirichlet $L$-functions, introduced by Dirichlet in 1837, are the first generalization of the Riemann zeta function. They are extremely important in number theory, being used, for example, to study the number of primes in arithmetic progressions and the class number of certain number fields (via Dirichlet's class number formula). Given a primitive Dirichlet character $\chi$ with modulus $q$ (see [Davenport 2000] for further background), the associated $L$-function is defined for $\mathrm{Re}(s) > 1$ by the absolutely convergent series

$$L(s, \chi) = \sum_{n \geq 1} \frac{\chi(n)}{n^s}. \tag{1-1}$$

This has an Euler product

$$L(s, \chi) = \prod_p \left(1 - \frac{\chi(p)}{p^s}\right)^{-1}$$

and analytically continues to an entire function with functional equation

$$\Lambda(s, \chi) := \left(\frac{\pi}{q}\right)^{-(s+\kappa)/2} \Gamma\left(\frac{s+\kappa}{2}\right) L(s, \chi) = \frac{\tau(\chi)}{i^\kappa q^{1/2}} \Lambda(1-s, \bar{\chi}),$$

where $\tau(\chi)$ is the Gauss sum and

$$\kappa := \begin{cases} 0 & \text{if } \chi(-1) = 1, \\ 1 & \text{if } \chi(-1) = -1. \end{cases} \tag{1-2}$$

*MSC2010:* 11M20.
*Keywords:* $L$-functions, Dirichlet characters, moments.

As is typically the case, the line of symmetry $\mathrm{Re}(s) = \frac{1}{2}$ of the functional equation is where the $L$-function is most difficult to understand. Since the values at $s = \frac{1}{2}$ of $L$-functions often encode important arithmetic information, it is natural to consider the central values $L\left(\frac{1}{2}, \chi\right)$. From the adelic point of view, these may be considered as finite-place-twist analogs of the archimedean twist $\zeta\left(\frac{1}{2} + it\right)$, which is of classical interest in analytic number theory. For example, it is conjectured that the central value $L\left(\frac{1}{2}, \chi\right)$ is never zero, but only partial results exist in this direction [Bui 2012; Khan and Ngo 2016; Soundararajan 2000]. As another example, an analog of the Lindelöf conjecture asserts that $L\left(\frac{1}{2}, \chi\right) \ll q^{\epsilon}$ for any $\epsilon > 0$, but again only partial results exist [Burgess 1963; Conrey and Iwaniec 2000; Milićević 2016]. (Here and henceforth, $\epsilon$ will always be used to denote an arbitrarily small positive constant, but may not be the same from one occurrence to the next. All implicit constants may depend on $\epsilon$.)

Given the lack of "closed-form formulas" that would directly shed light on the values of individual $L\left(\frac{1}{2}, \chi\right)$, one often thinks of $L$-functions as embedded in families and of the central value $L\left(\frac{1}{2}, \chi\right)$ as a random variable whose distribution we are trying to understand. From probability theory, we know that one way to understand the distribution of a random variable is to find its moments. For example, given a large sample of test scores, the first moment tells us the average score, the second moment is related to the variance of the scores, and if, as is often the case for test scores, their distribution follows the bell curve, then the $n$-th moment of the observed scores should correspond to that of the (rescaled) normal distribution. This philosophy about computing moments is in fact a typical starting point in solving problems about nonvanishing and size in families of $L$-functions. We remark on the side that numerics, partial theoretical results including the known moments, as well as analogs over function fields support a general conjecture that families of $L$-functions exhibit random behavior in a suitable sense; see, for example, [Katz and Sarnak 1999].

The moments problem is to evaluate asymptotically (as $q \to \infty$)

$$\sum_{\chi \bmod q}^{*} L\left(\tfrac{1}{2}, \chi\right)^n$$

for all $n \in \mathbb{N}$, as well as

$$\sum_{\chi \bmod q}^{*} \left|L\left(\tfrac{1}{2}, \chi\right)\right|^n$$

for even values of $n$, where $\sum^{*}$ means that the summation is restricted to the primitive characters. The evaluation of the first and second moments ($n = 1, 2$) is classical and due to Paley [1931]. The third and fourth moments ($n = 3, 4$) are quite recent. The third moment was obtained by Zacharias [2017] for prime values of $q$. The fourth moment was first obtained by Heath-Brown [1981] for values of $q$ with a

restricted number of prime factors and by Soundararajan [2007] for all values of $q$, and an asymptotic with a power savings error term was given by Young [2011] for prime values of $q$; see also [Blomer and Milićević 2015] for factorable $q$ (including prime powers). No asymptotic is known for the fifth moment or higher ($n \geq 5$).

In this paper we are interested in moments over natural subsets of the primitive Dirichlet characters mod $q$, where $q$ is of a special form. Working over a smaller set gets us closer to the true asymptotic features of individual $L$-functions, but of course it also means that there are fewer "harmonics" available to average over, so the evaluation of the moments becomes more difficult. We now proceed to describe our set of characters.

Let $\xi$ be a primitive $\phi(q)$-th root of unity, where $\phi$ is the Euler totient function, and let $\mathbb{Q}(\xi)$ be the corresponding cyclotomic field, which is Galois over $\mathbb{Q}$. The group $G = \mathrm{Gal}(\mathbb{Q}(\xi)/\mathbb{Q})$ acts on the set of primitive Dirichlet characters modulo $q$ as follows. For $\sigma \in G$, we define $\chi^\sigma$ to be that character for which $\chi^\sigma(n) = \sigma(\chi(n))$ for all $(n, q) = 1$. The action under $G$ partitions the set of characters into orbits $\mathcal{O}$, which we usually refer to as *Galois orbits*. Thus, from an algebraic perspective, any two characters in a single orbit $\mathcal{O}$ are indistinguishable.

Several works have studied the average values of $L$-functions over these orbits [Chinta 2002; Greenberg 1985; Khan et al. 2016; Rohrlich 1984]. For the rest of the paper, we specialize to moduli of the form

$$q = p^k,$$

where $p$ is a fixed odd prime (thus $q \to \infty$ is equivalent to $k \to \infty$). For such moduli, the orbits under the action of $G$ are easy to describe. We have that $\chi_1$ and $\chi_2$ belong to the same orbit if and only if $\chi_1$ and $\chi_2$ have the same order in the group of characters mod $q$. The possible orders are $l = p^{k-1}d$ for $d \mid (p-1)$, and the primitive characters of order $l$ form an orbit $\mathcal{O}$ of cardinality $\phi(l)$; see Table 1 for an example. These facts are justified in [Khan et al. 2016].

In the course of studying nonvanishing of Dirichlet $L$-functions within the Galois orbits described above, Ngo and two of us proved in [Khan et al. 2016, Theorem 1.2b] the following asymptotic for the second moment (as $k \to \infty$): for any given orbit $\mathcal{O}$ and $\epsilon > 0$, we have

$$\frac{1}{|\mathcal{O}|} \sum_{\chi \in \mathcal{O}} \left|L\left(\tfrac{1}{2}, \chi\right)\right|^2 = \frac{p-1}{p}(\log q + C) + O(q^{-1/4+\epsilon}), \tag{1-3}$$

where

$$C = \frac{\Gamma'\left(\frac{1}{4}(1+2\kappa)\right)}{\Gamma\left(\frac{1}{4}(1+2\kappa)\right)} + 2\gamma + 2\frac{\log p}{p-1} - \log \pi,$$

log is the natural logarithm, $\gamma = 0.57721\ldots$ is the Euler constant, and $\kappa$ is defined in (1-2). The implicit constant in the error term of (1-3) is *ineffective*. This means

| $n \bmod 9$ | 1 | 2 | 4 | 5 | 7 | 8 | primitive? | order | orbit |
|---|---|---|---|---|---|---|---|---|---|
| $\chi_0(n)$ | 1 | 1 | 1 | 1 | 1 | 1 | | 1 | $\{\chi_0\}$ |
| $\chi_1(n)$ | 1 | $\xi$ | $\xi^2$ | $\xi^5$ | $\xi^4$ | $-1$ | ✓ | $3\cdot 2$ | $\{\chi_1, \chi_5\}$ |
| $\chi_2(n)$ | 1 | $\xi^2$ | $\xi^4$ | $\xi^4$ | $\xi^2$ | 1 | ✓ | $3\cdot 1$ | $\{\chi_2, \chi_4\}$ |
| $\chi_3(n)$ | 1 | $-1$ | 1 | $-1$ | 1 | $-1$ | | 2 | $\{\chi_3\}$ |
| $\chi_4(n)$ | 1 | $\xi^4$ | $\xi^2$ | $\xi^2$ | $\xi^4$ | 1 | ✓ | $3\cdot 1$ | $\{\chi_2, \chi_4\}$ |
| $\chi_5(n)$ | 1 | $\xi^5$ | $\xi^4$ | $\xi$ | $\xi^2$ | $-1$ | ✓ | $3\cdot 2$ | $\{\chi_1, \chi_5\}$ |

**Table 1.** Four of the six characters modulo $9 = 3^2$ are primitive. They fall into two Galois orbits, the orbit $\{\chi_2, \chi_4\}$ consisting of characters of order $3\cdot 1 = 3$, of size $\phi(3) = 2$, and the orbit $\{\chi_1, \chi_5\}$ consisting of characters of order $3\cdot 2 = 6$, of size $\phi(6) = 2$.

that the error term is $\le C' q^{-1/4+\epsilon}$ for *some* constant $C' = C'(p, \epsilon)$, but we have no way of computing $C'$ given the values of $p$ and $\epsilon$. In turn, this means that there is a constant $k_0$ such that for all $k > k_0$, the main term of (1-3) dominates the error term, but there is no way to give an explicit value for $k_0$. In other words, we do not know how large $k$ must be before the given main term is a useful estimate of the second moment. This ineffectivity is a side effect of the fact that the argument for (1-3) given in [Khan et al. 2016] hinges crucially on Roth's theorem in Diophantine approximation (more precisely, on the $p$-adic version of Roth's theorem due to Ridout [1958]), which is well known to be ineffective. The goal of this paper is to remedy this situation for natural towers of characters to powers of several primes $p$.

**Theorem 1.1.** *Let $p = 3, 5$ or $7$. For every $q = p^k$ ($k \ge 1$) and every Galois orbit $\mathcal{O}$ of characters modulo $q$, we have*

$$\frac{1}{|\mathcal{O}|}\sum_{\chi\in\mathcal{O}}\left|L\left(\tfrac{1}{2}, \chi\right)\right|^2 = \frac{p-1}{p}(\log q + C) + O(q^{-\lambda_p+\epsilon}),$$

*where $\lambda_3 = \frac{1}{2}$ and $\lambda_5 = \lambda_7 = \frac{1}{6}$. The implicit constant is computable.*

Our argument differs from that of [Khan et al. 2016] in that we do not appeal to Roth's theorem. The present argument yields the fully effective Theorem 1.1 (with computable bounds on the error term), and in fact in (5-11) we provide an explicit version with a specific constant depending on $\epsilon > 0$. Given the power saving error term, it should be possible to extend our main theorem to include a mollifier. This would give an effective version of the nonvanishing result given in [Khan et al. 2016, Theorem 1.2b], but only for $p = 3, 5, 7$ and with possibly smaller proportions of nonvanishing.

In the statement of Theorem 1.1 and for the rest of the paper, the asymptotic notations $f \ll g$ and $f = O(g)$ mean that $|f| \le Cg$ for some constant $C > 0$, which may depend on $\epsilon > 0$, but is always computable for any given value of $\epsilon$.

## 2. Preliminaries

We first state a result which follows directly from [Khan et al. 2016, Lemma 2.3]. This illustrates an orthogonality property within orbits.

**Lemma 2.1.** *Suppose $q = p^k$ for an odd prime $p$ and $k \ge 1$, $\mathcal{O}$ is a Galois orbit of primitive Dirichlet characters mod $q$, and $n$ and $m$ are integers coprime to $p$. Clearly, $\frac{1}{|\mathcal{O}|}\left|\sum_{\chi\in\mathcal{O}} \chi(n)\bar{\chi}(m)\right| \le 1$. But if*

$$n^{p-1} \not\equiv m^{p-1} \bmod p^{k-1},$$

*then*

$$\frac{1}{|\mathcal{O}|}\sum_{\chi\in\mathcal{O}} \chi(n)\bar{\chi}(m) = 0.$$

Next we state a standard result from analytic number theory, called the approximate functional equation. The approximate functional equation expresses the $L$-function at the central point, where (1-1) does not converge, in terms of essentially finite sums of the form resembling a truncated version of Dirichlet series like (1-1). This is standard so we do not reproduce the entire proof.

**Lemma 2.2.** *For a primitive Dirichlet character $\chi$ modulo $q$, let $\kappa \in \{0, 1\}$ be such that $\chi(-1) = (-1)^\kappa$, and let*

$$V(x) = \frac{1}{2\pi i}\int_{(2)} \frac{\Gamma\left(\frac{1}{2}(s+\kappa)+\frac{1}{4}\right)^2}{\Gamma\left(\frac{1}{2}\kappa+\frac{1}{4}\right)^2}(\pi x)^{-s}\frac{ds}{s}. \tag{2-1}$$

*We have*

$$V(x) \ll_N \min\{1, x^{-N}\} \tag{2-2}$$

*for any $x, N > 0$, and*

$$\left|L\left(\tfrac{1}{2}, \chi\right)\right|^2 = 2\sum_{nm\ge 1} \frac{\chi(n)\bar{\chi}(m)}{(nm)^{1/2}} V\left(\frac{nm}{q}\right). \tag{2-3}$$

*Proof.* See [Khan et al. 2016, Lemma 2.1]. For the estimate (2-2), shift the line of integration to $\mathrm{Re}(s) = N$ if $x > 1$, and to $\mathrm{Re}(s) = -\frac{1}{4}$ if $x \le 1$. The shift left crosses a simple pole at $s = 0$, with residue 1. □

By the decay property (2-2), the range of summation in the sum (2-3) is essentially $nm < q^{1+\epsilon}$. Note that the sum is restricted to $(nm, p) = 1$, for otherwise the character values vanish.

We conclude the preliminaries section with two known results in elementary number theory. The first of these, Hensel's lemma, describes solutions to polynomial congruences modulo prime powers. In Lemma 2.3, we have taken the first statement from [Rosen 1984, Theorem 4.15(i)], and the second one follows by induction on $k$.

**Lemma 2.3** (Hensel's lemma). *Suppose that $f(x)$ is a polynomial with integer coefficients, $k$ is an integer with $k \geq 2$, and $p$ is a prime.*

(1) *If $r$ is a solution of the congruence $f(x) \equiv 0 \pmod{p^{k-1}}$ such that $f'(r) \not\equiv 0 \pmod{p}$, then there is a unique integer $t$, $0 \leq t < p$, such that $f(r+tp^{k-1}) \equiv 0 \pmod{p^k}$.*

(2) *If $r$ is a solution of the congruence $f(x) \equiv 0 \pmod{p}$ such that $f'(r) \not\equiv 0 \pmod{p}$, then there is a unique integer $t$, $0 \leq t < p^k$, such that $t \equiv r \pmod{p}$ and $f(t) \equiv 0 \pmod{p^k}$.*

The second number-theoretic result we record is concerned with the number of ways certain definite quadratic forms such as $n^2+m^2$ in two integers $n, m$ can take the same value.

**Lemma 2.4.** *Let $q(n,m)$ be any of $n^2+m^2$, $n^2+nm+m^2$, or $n^2-nm+m^2$. Then, for every $\epsilon > 0$,*

$$r_q(N) := \#\{(n,m) \in \mathbb{Z}^2 : q(n,m) = N\} \ll_\epsilon N^\epsilon.$$

For $q_0(n,m) = n^2+m^2$, the estimate $r_{q_0}(N) \ll_\epsilon N^\epsilon$ follows from the famous theorem of Gauss for the number of representations of a positive integer $N$ as the sum of two squares [Rosen 1984, Theorem 14.13]: if $N$ has a canonical prime power factorization as $N = 2^m p_1^{e_1} \cdots p_s^{e_s} q_1^{f_1} \cdots q_t^{f_t}$, where primes $p_i$ are of the form $4k+1$ and primes $q_j$ are of the form $4k+3$, then

$$r_{q_0}(N) = 4(e_1+1)(e_2+1)\cdots(e_s+1)$$

if all $f_j$ are even, and $r_{q_0}(N) = 0$ otherwise. In particular, $r_{q_0}(N)$ is bounded by the number of divisors $\tau(N)$ as $r_{q_0}(N) \leq 4\tau(N)$; hence $r_{q_0}(N) \ll_\epsilon N^\epsilon$ by the standard divisor bound; see, for example, [Stopple 2003, Section 3.5; Iwaniec and Kowalski 2004, (12.82)].

Gauss' formula for $r_{q_0}(N)$ can be proved using the arithmetic of the ring of Gaussian integers $\mathbb{Z}[i]$. This is a Euclidean domain (relative to the usual norm), and hence a unique factorization domain, in which 2 is the sole ramified prime, rational primes of the form $4k+1$ split as the product of two distinct conjugate Gaussian primes, and rational primes of the form $4k+3$ remain as Gaussian primes [Rosen 1984, Theorem 14.12]. A similar argument could be made for $q_1(n,m) = n^2+nm+m^2$ and $q_2(n,m) = n^2-nm+m^2$ by using the arithmetic of the ring of Eisenstein integers $\mathbb{Z}[\omega]$, where $\omega = -\frac{1}{2} + i\frac{\sqrt{3}}{2}$ is a primitive cube

root of unity, and by distinguishing between primes of the form $6k+1$ and $6k+5$. In each of these cases, unique factorization allows for very pretty formulas for $r_q(N)$; however, this is ultimately not so important if all we need is the upper bound of Lemma 2.4. To make this clear, we provide a streamlined argument that applies in more general situations.

*Proof.* Note that, if $N = n^2 + m^2 = (n+mi)(n-mi)$, then $(n+mi) \mid N$ in the ring $\mathbb{Z}[i]$. Similarly, if $N = n^2 - nm + m^2 = (n+m\omega)(n+m\omega^2)$, then $(n+m\omega) \mid N$ in $\mathbb{Z}[\omega]$, and if $N = n^2 + nm + m^2 = (n-m\omega)(n-m\omega^2)$, then $(n-m\omega) \mid N$ in $\mathbb{Z}[\omega]$. Therefore, writing $F = \mathbb{Q}(i)$ if $q(n,m) = n^2 + m^2$ and $F = \mathbb{Q}(\omega)$ if $q(n,m) = n^2 \pm nm + m^2$, we have

$$r_q(N) \ll \tau_F(N).$$

Here, $\tau_F(N)$ denotes the number of ideal divisors of the ideal $(N) = N\mathcal{O}_F$ in the ring of integers $\mathcal{O}_F$ of $F$, and the absolute implied constant accounts for the finite group of units, which, in this case, are all roots of unity. Therefore the desired estimate follows from the divisor bound

$$\tau_F(\mathfrak{n}) \ll_\epsilon \mathfrak{N}\mathfrak{n}^\epsilon \tag{2-4}$$

in terms of the absolute ideal norm, which is valid in any number field $F$ (with a constant possibly depending on $F$).

The estimate (2-4) can be proved for any number field $F$ along the same lines as over $\mathbb{Q}$ [Stopple 2003, Section 3.5]. It is clear that

$$\tau_F(\mathfrak{p}^\alpha) = \alpha + 1 \le (\mathfrak{N}\mathfrak{p}^\alpha)^\epsilon = \mathfrak{N}\mathfrak{p}^{\epsilon\alpha}$$

for all prime powers $\mathfrak{p}^\alpha$ with $\alpha \ge 1$ and sufficiently large $\mathfrak{N}\mathfrak{p}$ (say, $\mathfrak{N}\mathfrak{p} \ge e^{1/\epsilon}$). A similar inequality holds, by allowing for a larger (but fixed once and for all for a given $F$) implied constant, for powers of the finitely many prime ideals with $\mathfrak{N}\mathfrak{p} < 2^{1/\epsilon}$. The estimate (2-4) follows by multiplicativity. □

## 3. The diagonal contribution

Writing the sum in (2-3) as the sum of terms with $n = m$ plus the sum of terms with $n \ne m$, we get

$$\frac{1}{|\mathcal{O}|}\sum_{\chi\in\mathcal{O}}\left|L\left(\tfrac{1}{2},\chi\right)\right|^2 = \frac{1}{|\mathcal{O}|}\sum_{\chi\in\mathcal{O}}\left(2\sum_{\substack{n\ge 1\\(n,p)=1}}\frac{1}{n}V\left(\frac{n^2}{q}\right)\right)$$
$$+\frac{1}{|\mathcal{O}|}\sum_{\chi\in\mathcal{O}}\left(2\sum_{\substack{nm\ge 1\\n\ne m}}\frac{\chi(n)\bar{\chi}(m)}{(nm)^{\frac{1}{2}}}V\left(\frac{nm}{q}\right)\right). \tag{3-1}$$

The first sum above is the "diagonal" and it forms the main term of Theorem 1.1. This is not surprising because there are no character values in the sum, so the sum

over characters on the outside cannot produce any cancellation. By [Khan et al. 2016, Section 3.3] we have

$$\frac{1}{|\mathcal{O}|}\sum_{\chi\in\mathcal{O}}\Bigg(2\sum_{\substack{n\geq 1\\(n,p)=1}}\frac{1}{n}V\bigg(\frac{n^2}{q}\bigg)\Bigg)=\frac{p-1}{p}(\log q+C)+O(q^{-1/2+\epsilon}).$$

We recall that this argument uses the integral representation (2-1) and contour shifting and is fully effective.

Now it remains to bound the off-diagonal sum of (3-1). This will be the dominant part of the error term in Theorem 1.1.

## 4. The off-diagonal contribution

Applying Lemma 2.1 and (2-2), we get

$$\frac{1}{|\mathcal{O}|}\sum_{\chi\in\mathcal{O}}\Bigg(2\sum_{\substack{nm\geq 1\\n\neq m}}\frac{\chi(n)\bar{\chi}(m)}{(nm)^{1/2}}V\bigg(\frac{nm}{q}\bigg)\Bigg)\ll\sum_{\substack{nm<q^{1+\epsilon}\\(nm,p)=1,n\neq m\\n^{p-1}\equiv m^{p-1}\bmod p^{k-1}}}\frac{1}{(nm)^{1/2}}+q^{-100}.$$

We will analyze this sum in dyadic intervals

$$N\leq n<2N,\quad M\leq m<2M,$$

where

$$NM<q^{1+\epsilon}. \tag{4-1}$$

Since there are at most $q^{\epsilon}$ such dyadic intervals, the task is reduced to bounding

$$\mathcal{S}_p=\mathcal{S}_p(N,M):=\frac{1}{(NM)^{1/2}}\sum_{\substack{N\leq n<2N\\M\leq m<2M\\(nm,p)=1,n\neq m\\n^{p-1}\equiv m^{p-1}\bmod p^{k-1}}}1;$$

for a proof of Theorem 1.1, we require the bound $\mathcal{S}_p\ll q^{-\lambda_p+\epsilon}$ in the range (4-1). Let us first note a "trivial" bound (this argument is from [Khan et al. 2016, Section 3.3]).

**Lemma 4.1.** *We have*

$$\mathcal{S}_p\ll\min\bigg\{\bigg(\frac{N}{M}\bigg)^{1/2},\bigg(\frac{M}{N}\bigg)^{1/2}\bigg\}+q^{-1/2+\epsilon}, \tag{4-2}$$

$$\mathcal{S}_p\ll\min\bigg\{\frac{q^{1/2+\epsilon}}{M},\frac{q^{1/2+\epsilon}}{N}\bigg\}+q^{-1/2+\epsilon}. \tag{4-3}$$

*Proof.* Suppose without loss of generality that $N\leq M$. For each of the $N$ choices of $n$ in the sum $\mathcal{S}_p$, the value of $m^{p-1}$ is uniquely determined modulo $p^{k-1}$, namely,

$m^{p-1} \equiv n^{p-1} \pmod{p^{k-1}}$. By Lemma 2.3(2) with $f(x) = x^{p-1} - n^{p-1}$, for every $m_0 \pmod p$, $(m_0, p) = 1$, there is a unique value of $m$ modulo $p^{k-1}$ such that $m \equiv m_0 \pmod p$ and $f(m) \equiv 0 \pmod{p^{k-1}}$. Therefore, once the value of $n$ in $\mathcal{S}_p$ has been fixed, there are at most $O(1)$ choices for the congruence class of $m \pmod{p^{k-1}}$, and thus there are at most $O(M/q + 1)$ choices for $m$ itself. So the sum $\mathcal{S}_p$ is bounded as

$$\mathcal{S}_p \ll \frac{1}{(NM)^{1/2}} \cdot N \cdot \left(\frac{M}{q} + 1\right).$$

This gives the bound (4-2) by using (4-1). The bound (4-3) follows from (4-2) by using (4-1) again. □

We can see that the bound of Lemma 4.1 is sufficient as long as the sizes of $N$ and $M$ are apart by a certain power of $q$. From this point onwards, our argument differs from that of [Khan et al. 2016].

**4.1.** ***The case*** $p = 3$**.** The sum we need to bound is

$$\mathcal{S}_3 = \frac{1}{(NM)^{1/2}} \sum_{\substack{N \le n < 2N \\ M \le m < 2M \\ (nm,3)=1, n \ne m \\ n^2 \equiv m^2 \bmod 3^{k-1}}} 1.$$

The congruence condition of $\mathcal{S}_3$ implies that $3^{k-1}$ divides $(n-m)(n+m)$. Since $(nm, 3) = 1$, we know that $n-m$ and $n+m$ are not both divisible by 3 (for if they were, their sum would be too and this would lead to a contradiction). This means that either $3^{k-1}$ divides $n-m$, or $3^{k-1}$ divides $n+m$. We also have the condition $n \ne m$. So we must have that at least one of $N$ and $M$ is at least as large as $3^{k-1}/4$, lest $n-m$ and $n+m$ be too small to satisfy the divisibility condition. Thus by (4-3) we get

$$\mathcal{S}_3 \ll q^{-1/2+\epsilon}.$$

**4.2.** ***The case*** $p = 5$**.** The sum we need to bound is

$$\mathcal{S}_5 = \frac{1}{(NM)^{1/2}} \sum_{\substack{N \le n < 2N \\ M \le m < 2M \\ (nm,5)=1, n \ne m \\ n^4 \equiv m^4 \bmod 5^{k-1}}} 1. \tag{4-4}$$

Suppose without loss of generality that $M \ge N$. The congruence condition of $\mathcal{S}_5$ implies that $5^{k-1}$ divides $(n^2 - m^2)(n^2 + m^2)$. Since $(nm, 5) = 1$, we know that $n^2 - m^2$ and $n^2 + m^2$ are not both divisible by 5 (for if they were, their sum would be too and this would lead to a contradiction). Thus $5^{k-1}$ divides either $n^2 - m^2$

or $n^2+m^2$. The subsum of $S_5$ consisting of terms satisfying $5^{k-1} \mid (n^2-m^2)$ is $O(q^{-1/2+\epsilon})$ by the argument given for $p=3$.

Consider the terms satisfying $5^{k-1} \mid (n^2+m^2)$. First note that we must have $M \gg q^{1/2}$ or else $n^2+m^2$ is too small to satisfy the divisibility. Now, writing

$$n^2+m^2 = 5^{k-1}h,$$

we see that $h \ll M^2/q$. By Lemma 2.4, for each choice of $h$, there are $O(q^\epsilon)$ choices for $n$ and $m$. So there are at most $q^\epsilon(M^2/q)$ summands satisfying $5^{k-1} \mid (n^2+m^2)$ in (4-4). We get

$$\mathcal{S}_5 \ll q^{-1/2+\epsilon} + \frac{1}{(NM)^{1/2}}\frac{M^2}{q^{1-\epsilon}} \ll q^{-1/2+\epsilon} + \left(\frac{M}{N}\right)^{1/2}\frac{M}{q^{1-\epsilon}}. \tag{4-5}$$

Now we consider two cases: when $N$ and $M$ are quite close and when they are not.

Suppose that $M/N < q^{1/3}$. Then by (4-1) we have $M^2 \ll (M/N)q^{1+\epsilon} \ll q^{4/3+\epsilon}$. So (4-5) becomes

$$\mathcal{S}_5 \ll q^{-1/6+\epsilon}. \tag{4-6}$$

Now suppose that $M/N \geq q^{1/3}$. Then by (4-2), we get the same bound (4-6).

**4.3.** ***The case*** $p=7$**.** The sum we need to bound is

$$\mathcal{S}_7 = \frac{1}{(NM)^{1/2}} \sum_{\substack{N\leq n<2N \\ M\leq m<2M \\ (nm,7)=1, n\neq m \\ n^6\equiv m^6 \bmod 7^{k-1}}} 1.$$

The congruence condition of $\mathcal{S}_7$ implies

$$7^{k-1} \mid (n^2-m^2)(n^2+nm+m^2)(n^2-nm+m^2).$$

Since $(nm,7)=1$, we get that 7 cannot divide more than one factor on the right-hand side. For example, if $7 \mid (n^2-m^2)$ then $n \equiv \pm m \bmod 7$. So if also $7 \mid (n^2 \pm nm+m^2)$, then $7 \mid (n^2 \pm n^2+n^2)$, which is impossible. On the other hand, if $7 \mid (n^2+nm+m^2)$ and $7 \mid (n^2-nm+m^2)$, then $7 \mid nm$, which is again impossible. So we have the cases $7^{k-1} \mid (n^2-m^2)$ or $7^{k-1} \mid (n^2 \pm nm+m^2)$. By the argument given for $p=3$, the subsum of $\mathcal{S}_7$ consisting of terms satisfying $7^{k-1} \mid (n^2-m^2)$ is $O(q^{-1/2+\epsilon})$.

For the cases when $7^{k-1} \mid (n^2 \pm nm+m^2)$, we proceed analogously to the case $p=5$. We must have $M \gg q^{1/2}$ or else $n^2 \pm nm+m^2$ is too small to be divisible by $7^{k-1}$, and in fact $n^2 \pm nm+m^2 = 7^{k-1}h$ for some $h \ll M^2/q$. By Lemma 2.4 (this time applied with the form $n^2 \pm nm+m^2$), the number of choices of $(n,m)$ is $O(q^\epsilon)$ for each choice of $h$ and thus at most $q^\epsilon(M^2/q)$ altogether. Therefore,

$$\mathcal{S}_7 \ll q^{-1/2+\epsilon} + \frac{1}{(NM)^{1/2}}\frac{M^2}{q^{1-\epsilon}} \ll q^{-1/2+\epsilon} + \left(\frac{M}{N}\right)^{1/2}\frac{M}{q^{1-\epsilon}}.$$

Using this bound and (4-1) when $M/N < q^{1/3}$ and (4-2) when $M/N \geq q^{1/3}$, we obtain

$$\mathcal{S}_7 \ll q^{-1/6+\epsilon}.$$

## 5. Effective estimates

In this section, we show how all the estimates of previous sections can be made fully effective for any desired choice of $\epsilon > 0$. We follow the exposition in Sections 2–4 and indicate explicit constants at each place. Since many of these computations are routine, we condense some of the details but provide all the essential steps.

**5.1.** ***Preliminaries.*** When estimating expressions involving $\Gamma(s)$, we use the following well-known facts valid for $\sigma = \operatorname{Re} s > 0$ and integers $N \geq 2$:

$$\begin{gathered}\Gamma(s) = \frac{1}{s}\Gamma(s+1), \quad \Gamma(s)\Gamma\big(s+\tfrac{1}{2}\big) = 2^{1-2s}\sqrt{\pi}\,\Gamma(2s), \quad |\Gamma(s)| \leq \Gamma(\sigma), \\ \big|\Gamma\big(\sigma+\tfrac{1}{4}\big)\Gamma\big(\sigma+\tfrac{5}{4}\big)\big| \leq \big|\Gamma(\sigma)\Gamma\big(\sigma+\tfrac{3}{2}\big)\big|, \quad |\Gamma(N)| \leq \tfrac{1}{4}e^2(N/e)^N.\end{gathered} \tag{5-1}$$

The first inequality in the second row follows from the convexity of $\log\Gamma(\sigma)$, and the second one follows by using integral comparison to estimate $\sum_{n<N}\log n$.

In Lemma 2.2, for $\kappa = 0$, we find by shifting contours to $\operatorname{Re} s = N \geq 3$ that

$$\begin{aligned}|V(x)| &\leq \frac{1}{2\pi\,\Gamma\big(\frac{1}{4}\big)^2}\Gamma\big(\tfrac{1}{2}N+\tfrac{1}{4}\big)\Gamma\big(\tfrac{1}{2}N+\tfrac{5}{4}\big)\int_{-\infty}^{\infty}\frac{\mathrm{d}t}{\big|\frac{1}{2}(N+it)+\frac{1}{4}\big|\,|N+it|}\cdot(\pi x)^{-N} \\ &\leq \frac{\sqrt{\pi}}{2\pi\,\Gamma\big(\frac{1}{4}\big)^2}\cdot(N+1)\Gamma(N)(2\pi)^{-N}\cdot\frac{8}{N}\cdot x^{-N} < \frac{3}{4}\Big(\frac{N}{2\pi e}\Big)^N\cdot x^{-N},\end{aligned}$$

where the integral is split into $|t| \leq N$ and $|t| > N$ and then estimated trivially. Similarly, by shifting to $\operatorname{Re} s = -\frac{1}{4}$,

$$|V(x)-1| \leq \frac{\pi^{1/4}\Gamma\big(\frac{1}{8}\big)\Gamma\big(\frac{7}{8}\big)}{2\pi\,\Gamma\big(\frac{1}{4}\big)^2}\int_{-\infty}^{\infty}\frac{\mathrm{d}t}{\big|\frac{1}{8}+\frac{1}{2}it\big|\big|-\frac{1}{4}+it\big|}\cdot x^{1/4} < 3x^{1/4},$$

where the integral is $\leq 16\sqrt{2}$ by splitting into $|t| \leq \frac{1}{2\sqrt{2}}$ and $|t| > \frac{1}{2\sqrt{2}}$ and estimating trivially. One similarly verifies that the same upper bounds hold for $\kappa = 1$. Using the first bound for $x \geq N/(2\pi e)$ and the second one for $x < N/(2\pi e)$, we obtain for $N \geq 3$

$$|V(x)| \leq \min\Big\{\frac{5N^{1/4}}{2}, \frac{3}{4}\Big(\frac{N}{2\pi e}\Big)^N\cdot x^{-N}\Big\}. \tag{5-2}$$

Next, we make (2-4) effective. Since the group of units in an imaginary quadratic number field such as $F = \mathbb{Q}(i)$ or $F = \mathbb{Q}(\omega)$ is a cyclic group of order at most 6, it is easy to see that $r_q(N) \leq 3\cdot\tau_F(N)$. For a prime ideal $\mathfrak{p}$ with $\mathfrak{N}\mathfrak{p} \geq e^{1/\epsilon}$, we simply have $\tau_F(\mathfrak{p}^\alpha) = 1+\alpha \leq (\mathfrak{N}\mathfrak{p}^\alpha)^\epsilon$. For the remaining primes, the function

$F(\alpha) = \mathfrak{N}\mathfrak{p}^{\alpha\epsilon}/(1+\alpha)$ has a minimum at $\alpha_0 = 1/(\epsilon \log \mathfrak{N}\mathfrak{p}) - 1 > 0$ with $F(\alpha_0) \geq e\epsilon \log \mathfrak{N}\mathfrak{p}/\mathfrak{N}\mathfrak{p}^{\epsilon}$. Therefore, for every integral ideal $\mathfrak{n} \subseteq \mathcal{O}_F$,

$$\frac{\tau_F(\mathfrak{n})}{\mathfrak{N}\mathfrak{n}^{\epsilon}} \leq \prod_{\mathfrak{N}\mathfrak{p} \leq e^{1/\epsilon}} \left( e\epsilon \frac{\log \mathfrak{N}\mathfrak{p}}{\mathfrak{N}\mathfrak{p}^{\epsilon}} \right)^{-1} \leq \frac{\epsilon^{-2\pi(e^{1/\epsilon})}}{\log 2},$$

where $\pi(x) = \#\{p \leq x\}$ is the classical prime-counting function, and the number of prime ideals $\mathfrak{p}$ with $\mathfrak{N}\mathfrak{p} \leq x$ is clearly $\leq 2\pi(x)$. Using the explicit estimate $\pi(x) \leq 2x/\log x$ [Stopple 2003, Section 5.2], we thus find that, for $\epsilon \leq \frac{1}{2}$,

$$\tau_F(\mathfrak{n}) \leq \frac{e^{4\epsilon |\log \epsilon| e^{1/\epsilon}}}{\log 2} \cdot \mathfrak{N}\mathfrak{n}^{\epsilon} \leq e^{(3/2)e^{1/\epsilon}} \cdot \mathfrak{N}\mathfrak{n}^{\epsilon}. \tag{5-3}$$

**5.2.** ***Diagonal terms.*** Proceeding to the evaluation of the diagonal contribution in Section 3, following [Khan et al. 2016, Section 3.3], we substitute the integral representation for $V(x)$ and exchange the order of summation and integration to find that the diagonal contribution equals, for $\kappa = 0$,

$$\frac{1}{2\pi i} \int_{(2)} \zeta_p(2s+1) \frac{\Gamma\left(\frac{1}{2}s + \frac{1}{4}\right)^2}{\Gamma\left(\frac{1}{4}\right)^2} \left(\frac{q}{\pi}\right)^s \frac{\mathrm{d}s}{s}.$$

We evaluate the integral by shifting to $\operatorname{Re} s = -\frac{1}{2} + \epsilon$, collecting the residue from the double pole at $s = 0$. We can use a simple estimate for $\zeta(s)$ with $0 < \sigma \leq \frac{1}{2}$ as

$$|(1 - 2^{1-s})\zeta(s)| = \left| \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n^s} \right| \leq \sum_{n=1}^{\infty} |s| (2n-1)^{-\sigma-1} \leq |s| \left(1 + \frac{1}{2\sigma}\right) \leq \frac{|s|}{\sigma},$$

by integral comparison. Further, $|1 - 2^{1-s}| \geq 2^{2/3} - 1 > \frac{1}{2}$ for $\sigma \leq \frac{1}{3}$ and $|1 - p^{-s}| \leq \min(|s| \log p, 2)$ for $p \geq 3$, so that the remainder from the contour at $\operatorname{Re} s = -\frac{1}{2} + \epsilon$, with $\epsilon \leq \frac{1}{6}$, is

$$\begin{aligned} &\leq \frac{1}{2\pi} \frac{2}{2\epsilon} \frac{1}{\Gamma\left(\frac{1}{4}\right)^2} \int_{-\infty}^{\infty} \frac{|2\epsilon + 2it| \min(|2\epsilon + 2it| \log p, 2)}{\left|\frac{1}{2}(\epsilon + it)\right|^2} \frac{\mathrm{d}t}{|\epsilon + it|} \cdot \left(\frac{q}{\pi}\right)^{-1/2+\epsilon} \\ &\leq \frac{8\sqrt{\pi}}{\pi \Gamma\left(\frac{1}{4}\right)^2 \epsilon} \cdot 2 \log p \cdot (2 + |\log \epsilon|) \cdot q^{-1/2+\epsilon} < \frac{3}{2} \log p \frac{|\log \epsilon|}{\epsilon} \cdot q^{-1/2+\epsilon}, \end{aligned} \tag{5-4}$$

by splitting the integral into $|t| \leq \epsilon$, $\epsilon < |t| \leq 1/\log p$, and $|t| > 1/\log p$ and estimating trivially. It is similarly verified that this explicit estimate for the error term in the evaluation of the diagonal contribution in Section 3 holds also when $\kappa = 1$.

**5.3.** ***Off-diagonal terms.*** We now come to the crux of the matter, the estimation of the off-diagonal terms in Section 4, in which $p \leq 7$ and we may assume that

$0 < \epsilon \le \lambda_p$. As a preliminary step, we find that the function $G(q) = q^\epsilon / \log q$ has a minimum at $q_0 = e^{1/\epsilon}$ with $G(q_0) = \epsilon e$, so that

$$\log q \le \frac{1}{\epsilon e} \cdot q^\epsilon.$$

For $x \ge 1$ and $N \ge 2$,

$$\sum_{m \le x} \frac{1}{m^{N+1/2}} \le \frac{1}{x^{N-1/2}} \left( \frac{1}{N - \frac{1}{2}} + \frac{1}{x} \right) \le \frac{2}{x^{N-1/2}}.$$

Using this, the contribution of the terms with $nm > q^{1+\epsilon}$ is estimated using (5-2) as

$$\begin{aligned} &2 \sum_{nm > q^{1+\epsilon}} \frac{1}{(nm)^{1/2}} V\left(\frac{nm}{q}\right) \\ &\quad \le 4 \frac{3}{4} \left(\frac{N}{2\pi e}\right)^N q^N \left( \sum_{n \le q^{1/2+\epsilon}} \frac{1}{n^{1/2+N} (q^{1+\epsilon}/n)^{N-1/2}} + \sum_{n > q^{1+\epsilon}} \frac{1}{n^{1/2+N}} \right) \\ &\quad \le 3 \left(\frac{N}{2\pi e}\right)^N \frac{q^{1/2}}{q^{(N-1)\epsilon}} (\log q^{1+\epsilon} + 2) < \frac{4}{\epsilon} \left(\frac{N}{2\pi e}\right)^N \frac{q^{1/2}}{q^{(N-2)\epsilon}} \end{aligned}$$

for $\epsilon \le \frac{1}{2}$. Taking $N \ge 1/\epsilon + 2$, the total contribution of terms with $nm > q^{1+\epsilon}$ is

$$\le \frac{4}{\epsilon (2\pi e \epsilon)^{1/\epsilon+2}} (1 + 3\epsilon)^{1/\epsilon+3} \cdot q^{-1/2} < \frac{1}{\epsilon^3 (2\pi e)^{1/\epsilon}} \cdot q^{-1/2}. \tag{5-5}$$

The terms with $nm < q^{1+\epsilon}$ can be split into at most

$$\frac{\log q^{1+\epsilon}}{\log 2} + 1 \le \frac{7}{6} \frac{1}{\log 2 \cdot e\epsilon} q^\epsilon + 1 < \left(\frac{1}{\epsilon}\right) q^\epsilon$$

dyadic ranges. Referring again to (5-2), the contribution of terms with $nm \le q^{1+\epsilon}$ is

$$\le \frac{5}{2\epsilon} \left( \frac{1}{\epsilon} + 3 \right)^{1/4} \max \mathcal{S}_p(N, M) < \frac{3}{\epsilon^{5/4}} \max \mathcal{S}_p(N, M), \tag{5-6}$$

where $NM < q^{1+\epsilon}$ and $\mathcal{S}_p(N, M)$ are as in Section 4.

Arguing as in the proof of Lemma 4.1, for every value of $n$ in $\mathcal{S}_p$, there are at most six choices for $m$ mod $p^{k-1}$ and thus the total number of choices for $m$ is at most $42(M/q + 1)$. From this, we get for $N \le M$,

$$\begin{aligned} \mathcal{S}_p &\le \frac{42}{(NM)^{1/2}} N \left( \frac{M}{q} + 1 \right) \\ &\le 42 \left( \left( \frac{N}{M} \right)^{1/2} + q^{-1/2+\epsilon} \right) \le 42 \left( \frac{q^{1/2+\epsilon/2}}{M} + q^{-1/2+\epsilon} \right). \end{aligned} \tag{5-7}$$

As in Section 4.1, the subsum of $\mathcal{S}_p(N, M)$ consisting of terms with $p^{k-1} \mid (n^2-m^2)$ is empty unless $M \geq q/28$, in which case their contribution is

$$\leq 42(28q^{-\epsilon/2}+1)q^{-1/2+\epsilon} < 1200 \cdot q^{-1/2+\epsilon}. \tag{5-8}$$

If $p = 3$, this is an upper bound on the full $\mathcal{S}_3$.

If $p = 5$, we must also consider the terms satisfying $5^{k-1} \mid (n^2+m^2)$. These occur only if $M \geq q^{1/2}/\sqrt{40}$ and $n^2+m^2 = 5^{k-1}h$ for some $1 \leq h \leq 40M^2/q$. For each $h$, we may bound $r_q(5^{k-1}h)$ by (5-3) and thus obtain

$$\mathcal{S}_5 \leq 1200 \cdot q^{-1/2+\epsilon} + 120e^{(3/2)e^{2/(5\epsilon)}} \frac{1}{(NM)^{1/2}} \frac{M^2}{q^{1-(5/2)\epsilon}}.$$

If $M/N < q^{1/3-2\epsilon}$, then $M^2/(NM)^{1/2} \leq (M/N)(NM)^{1/2} \leq q^{5/6-(3/2)\epsilon}$ and so

$$\mathcal{S}_5 \leq (1200q^{-1/3} + 3 \cdot 40e^{(3/2)e^{2/(5\epsilon)}})q^{-1/6+\epsilon} \leq 125e^{(3/2)e^{2/(5\epsilon)}} \cdot q^{-1/6+\epsilon}. \tag{5-9}$$

If, on the other hand, $M/N \geq q^{1/3+2\epsilon}$, we have $\mathcal{S}_5 \leq 70 \cdot q^{-1/6+\epsilon}$ by (5-7), so the above holds anyway. The same reasoning for $p = 7$ yields

$$\mathcal{S}_7 \leq (1200q^{-1/3} + 2 \cdot 3 \cdot 56e^{(3/2)e^{2/(5\epsilon)}})q^{-1/6+\epsilon} \leq 340e^{(3/2)e^{2/(5\epsilon)}} \cdot q^{-1/6+\epsilon}. \tag{5-10}$$

Combining (5-4)–(5-6) and (5-8)–(5-10), we obtain Theorem 1.1 in the effective form

$$\left| \frac{1}{|\mathcal{O}|} \sum_{\chi \in \mathcal{O}} \left|L\left(\tfrac{1}{2}, \chi\right)\right|^2 - \frac{p-1}{p}(\log q + C) \right| \leq c(\epsilon) q^{-\lambda_p+\epsilon}, \tag{5-11}$$

where, for $0 < \epsilon \leq \lambda_p$,

$$c(\epsilon) \leq \frac{3}{2}\log 7 \frac{|\log \epsilon|}{\epsilon} + \frac{1}{\epsilon^3(2\pi e\epsilon)^{1/\epsilon}} + \frac{3,600}{\epsilon^{5/4}} + \frac{1,020}{\epsilon^{5/4}} e^{(3/2)e^{2/(5\epsilon)}} < \frac{1,100}{\epsilon^{5/4}} e^{(3/2)e^{2/(5\epsilon)}}.$$

Indeed, it is seen directly that the function $f(x) = \frac{3}{2}e^{(2/5)x} - \left(x + \frac{7}{4}\right)\log x + 10$ is positive on $[2, 5]$ and on $[5, 8]$, and

$$f'(x) \geq \tfrac{3}{5}e^{16/5}(x-7) - \left(\log 8 + \tfrac{1}{8}x - 1\right) - \tfrac{39}{32} > 14x - 102 > 0$$

for $x \geq 8$, so that $f(x) > 0$ for all $x \geq 2$. Therefore,

$$(2\pi e)^{-1/\epsilon}\epsilon^{-(1/\epsilon+7/4)} < \frac{e^{10}}{(2\pi e)^2} \cdot e^{(3/2)e^{2/(5\epsilon)}} < 76e^{(3/2)e^{2/(5\epsilon)}},$$

which suffices to estimate the second summand; for the others it suffices to note that $\epsilon^{1/4}|\log \epsilon| \leq 4/e$ and $e^{(3/2)e^{4/5}} > 28$.

## Acknowledgements

We would like to thank the referee for constructive suggestions. In particular, the referee suggested that, since our estimates are effective, we demonstrate this aspect of our results by writing down the fully explicit version, and we now do so in the final section of the paper. This article grew out of Lei's thesis at Bryn Mawr College, jointly supervised by Khan and Milićević. Milićević was supported by the National Science Foundation, grant DMS-1503629.

## References

[Blomer and Milićević 2015] V. Blomer and D. Milićević, "The second moment of twisted modular $L$-functions", *Geom. Funct. Anal.* **25**:2 (2015), 453–516. MR Zbl

[Bui 2012] H. M. Bui, "Non-vanishing of Dirichlet $L$-functions at the central point", *Int. J. Number Theory* **8**:8 (2012), 1855–1881. MR Zbl

[Burgess 1963] D. A. Burgess, "On character sums and $L$-series, II", *Proc. London Math. Soc.* (3) **13** (1963), 524–536. MR Zbl

[Chinta 2002] G. Chinta, "Analytic ranks of elliptic curves over cyclotomic fields", *J. Reine Angew. Math.* **544** (2002), 13–24. MR Zbl

[Conrey and Iwaniec 2000] J. B. Conrey and H. Iwaniec, "The cubic moment of central values of automorphic $L$-functions", *Ann. of Math.* (2) **151**:3 (2000), 1175–1216. MR Zbl

[Davenport 2000] H. Davenport, *Multiplicative number theory*, 3rd ed., Graduate Texts in Mathematics **74**, Springer, 2000. MR Zbl

[Greenberg 1985] R. Greenberg, "On the critical values of Hecke $L$-functions for imaginary quadratic fields", *Invent. Math.* **79**:1 (1985), 79–94. MR Zbl

[Heath-Brown 1981] D. R. Heath-Brown, "The fourth power mean of Dirichlet's $L$-functions", *Analysis* **1**:1 (1981), 25–32. MR Zbl

[Iwaniec and Kowalski 2004] H. Iwaniec and E. Kowalski, *Analytic number theory*, American Mathematical Society Colloquium Publications **53**, American Mathematical Society, Providence, RI, 2004. MR Zbl

[Katz and Sarnak 1999] N. M. Katz and P. Sarnak, "Zeroes of zeta functions and symmetry", *Bull. Amer. Math. Soc.* (*N.S.*) **36**:1 (1999), 1–26. MR Zbl

[Khan and Ngo 2016] R. Khan and H. T. Ngo, "Nonvanishing of Dirichlet $L$-functions", *Algebra Number Theory* **10**:10 (2016), 2081–2091. MR Zbl

[Khan et al. 2016] R. Khan, D. Milićević, and H. T. Ngo, "Non-vanishing of Dirichlet $L$-functions in Galois orbits", *Int. Math. Res. Not.* **2016**:22 (2016), 6955–6978. MR

[Milićević 2016] D. Milićević, "Sub-Weyl subconvexity for Dirichlet $L$-functions to prime power moduli", *Compos. Math.* **152**:4 (2016), 825–875. MR

[Paley 1931] R. E. A. C. Paley, "On the k-analogues of some theorems in the theory of the Riemann $\zeta$-function", *Proc. London Math. Soc.* (2) **32**:4 (1931), 273–311. MR Zbl

[Ridout 1958] D. Ridout, "The $p$-adic generalization of the Thue–Siegel–Roth theorem", *Mathematika* **5** (1958), 40–48. MR Zbl

[Rohrlich 1984] D. E. Rohrlich, "On $L$-functions of elliptic curves and anticyclotomic towers", *Invent. Math.* **75**:3 (1984), 383–408. MR Zbl

[Rosen 1984] K. H. Rosen, *Elementary number theory and its applications*, Addison-Wesley, Reading, MA, 1984. MR Zbl

[Soundararajan 2000] K. Soundararajan, "Nonvanishing of quadratic Dirichlet $L$-functions at $s = \frac{1}{2}$", *Ann. of Math.* (2) **152**:2 (2000), 447–488. MR Zbl

[Soundararajan 2007] K. Soundararajan, "The fourth moment of Dirichlet $L$-functions", pp. 239–246 in *Analytic number theory*, edited by W. Duke and Y. Tschinkel, Clay Math. Proc. **7**, Amer. Math. Soc., Providence, RI, 2007. MR Zbl

[Stopple 2003] J. Stopple, *A primer of analytic number theory: from Pythagoras to Riemann*, Cambridge University Press, 2003. MR Zbl

[Young 2011] M. P. Young, "The fourth moment of Dirichlet $L$-functions", *Ann. of Math.* (2) **173**:1 (2011), 1–50. MR Zbl

[Zacharias 2017] R. Zacharias, "Simultaneous non-vanishing for Dirichlet $L$-functions", preprint, 2017. arXiv

Received: 2017-12-29 Revised: 2018-09-19 Accepted: 2018-10-12

rrkhan@olemiss.edu *Department of Mathematics, University of Mississippi, University, MS, United States*

rlei@brynmawr.edu *Department of Mathematics, Bryn Mawr College, Bryn Mawr, PA, United States*

dmilicevic@brynmawr.edu *Department of Mathematics, Bryn Mawr College, Bryn Mawr, PA, United States*

INVOLVE 12:3 (2019)
dx.doi.org/10.2140/involve.2019.12.491

# On the preservation of properties by piecewise affine maps of locally compact groups

Serina Camungol, Matthew Morison, Skylar Nicol and Ross Stokke

(Communicated by David Royal Larson)

As shown by Cohen (1960) and Ilie and Spronk (2005), for locally compact groups $G$ and $H$, there is a one-to-one correspondence between the completely bounded homomorphisms of their respective Fourier and Fourier–Stieltjes algebras $\varphi : A(G) \to B(H)$ and piecewise affine continuous maps $\alpha : Y \subseteq H \to G$. Using elementary arguments, we show that several (locally compact) group-theoretic properties, including amenability, are preserved by certain continuous piecewise affine maps. We discuss these results in relation to Fourier algebra homomorphisms.

Piecewise affine maps are, loosely speaking, finite unions of translations of subgroup homomorphisms. They seem to have been exclusively studied in connection with their applications to abstract harmonic analysis; see for example [Cohen 1960; Rudin 1962; Ilie 2004; Ilie and Spronk 2005; Ilie and Stokke 2008]. Our motivation in writing this paper has been to view piecewise affine maps as weak types of "generalized homomorphisms" and to study of them, in their own right, accordingly. Observe that most of our topologically imposed conditions are automatically satisfied by (discrete) groups and our results are also new in this situation.

Throughout this note, $G$ and $H$ are locally compact groups, and $\mathcal{P}$ will denote a property of locally compact groups. If $E$ is a coset of a closed subgroup $H_0$ of $H$, we will say that $E$ has $\mathcal{P}$ when $H_0$ has $\mathcal{P}$, and we define the *index* of $E$ in $H$ to be the index of $H_0$ in $H$. As noted in [Ilie 2004], a subset $E$ of $H$ is a coset of some subgroup of $H$ exactly when $EE^{-1}E = E$, and a map $\alpha : E \to G$ is called *affine* if for any $x, y, z \in E$, $\alpha(xy^{-1}z) = \alpha(x)\alpha(y)^{-1}\alpha(z)$. Thus, the affine maps are the natural morphisms of cosets and the affine image of a coset is also a coset.

*MSC2010:* primary 22D05, 43A22, 43A07, 43A30; secondary 20E99.
*Keywords:* locally compact group, piecewise affine map, amenability, Fourier algebra.
This note is based on undergraduate student research projects conducted under the supervision of Stokke. Nicol was supported by an NSERC Undergraduate Student Research Award; the other authors received financial support from Stokke's NSERC grant.

Note that for any $y_0 \in E$, $H_0 = y_0^{-1}E = E^{-1}E$ is a subgroup of $H$, and the map defined by $\beta(h) = \alpha(y_0)^{-1}\alpha(y_0 h)$ $(h \in H_0)$ is a homomorphism of $H_0$ into $G$ when $\alpha$ is an affine map; conversely, if $\beta : H_0 \to G$ is a homomorphism and $x_0 \in G$, then $\alpha(x) = x_0\beta(y_0^{-1}x)$ defines an affine map on $E$; see [Ilie 2004, Remark 2.2]. Thus, affine maps are exactly the translates of subgroup homomorphisms. A map $\alpha : E \to G$ is *antiaffine* if for any $x, y, z \in E$, $\alpha(xy^{-1}z) = \alpha(z)\alpha(y)^{-1}\alpha(x)$. Hence, the antiaffine image of a coset is also a coset and, as with the affine case, one can readily check that the antiaffine maps on $E$ are precisely the translates of subgroup antihomomorphisms on $H_0 = E^{-1}E$.

We let $\Omega(H)$ denote the ring of sets generated by the open cosets of $H$. Then every set in $\Omega(H)$ can be expressed as a finite union of disjoint sets in

$$\Omega_0(H) = \left\{ E_0 \setminus \left( \bigcup_1^m E_k \right) : E_0 \subseteq H \text{ an open coset}, \atop E_1, \ldots, E_m \text{ open subcosets of infinite index in } E_0 \right\}$$

[Cohen 1960; Ilie 2004]. If $Y = E_0 \setminus \left(\bigcup_1^m E_k\right) \in \Omega_0(H)$, Ilie showed that $\mathrm{Aff}(Y)$, the coset generated by $Y$, is exactly $E_0$ and that there is a finite subset $F$ of $E_0^{-1}E_0$ such that $E_0 = YF$ [Ilie 2004]. A map $\alpha : Y \to G$ is *piecewise affine* if

(†) there exist pairwise disjoint sets $Y_1, \ldots, Y_n \in \Omega_0(H)$ such that $Y = \bigcup_{i=1}^n Y_i$ and for each $i$, $\alpha|_{Y_i}$ has an affine extension $\alpha_i$ mapping $E_i = \mathrm{Aff}(Y_i)$ into $G$;

when each $\alpha_i$ is antiaffine, $\alpha$ is *piecewise antiaffine*, and when each $\alpha_i$ is affine *or* antiaffine, $\alpha$ is *mixed piecewise affine*.

**Notation.** Whenever we say that $\alpha : Y \subseteq H \to G$ is a (mixed) piecewise affine map, we shall use precisely the notation found in (†).

The continuous (mixed) piecewise affine maps can be viewed as the natural morphisms to consider on sets in the open coset ring $\Omega(H)$ of $H$. Equivalent definitions of piecewise affine maps on nonabelian groups are found in [Ilie 2004]. We note that if $\alpha$ is proper, open, closed or injective then so is $\alpha_i$ for each $i = 1, \ldots, n$ [Ilie 2004, Proposition 4.6], [Ilie and Stokke 2008, Lemma 3.3] (the same argument works for closed maps) and [Pham 2010, proof of Theorem 6.4]. As well, if $\alpha$ is continuous, then so is each $\alpha_i$ (and the converse also holds). This is almost certainly known but the authors were unable to locate the statement in the literature; note that continuity of the affine extensions seems to be implicitly assumed in the definition of a piecewise affine map in [Ilie 2004]. Nevertheless, this is easy to see: Let $F_i$ be a finite subset of $E_i^{-1}E_i$ such that $E_i = Y_i F_i$ [Ilie 2004, Lemma 4.5]. Let $x \in F_i$, say $x = u^{-1}v$ where $u, v \in E_i$. Then, for each $y \in Y_i x$, $\alpha_i(y) = \alpha(yx^{-1})\alpha_i(u)^{-1}\alpha_i(v)$, so $\alpha_i$ is continuous on $Y_i x$. By the pasting lemma, $\alpha_i$ is continuous on $Y_i F_i = E_i$.

Since the terminology of group extensions varies in the literature, we note that when $N$ is a closed normal subgroup of $G$, we will call $G$ an extension of $G/N$ by $N$, (whereas in [Palmer 2001], for example, $G$ is called an extension of $N$ by $G/N$).

## 1. Properties preserved by mixed piecewise affine maps

One typically begins studying a property $\mathcal{P}$ of locally compact groups by asking if $\mathcal{P}$ is preserved by closed subgroups, quotients and extensions. Phrased in terms of homomorphic images, $\mathcal{P}$ is preserved by closed subgroups if whenever there exists a continuous, injective homomorphism $\phi : H \to G$ such that $\phi$ is a homeomorphism onto $\phi(H)$ with its relative topology and $G$ has $\mathcal{P}$, then $H$ has $\mathcal{P}$. Additionally, $\mathcal{P}$ is preserved by closed quotients if whenever there exists a continuous, surjective homomorphism $\phi : H \to G$ such that $\phi$ is an open map and $H$ has $\mathcal{P}$, then $G$ has $\mathcal{P}$. More generally, given a continuous mixed piecewise affine map $\alpha : Y \subseteq H \to G$, the main purpose of this section is to address the following two questions:

(a) If $\alpha$ has dense image in $G$ and $H$ has $\mathcal{P}$, when does $G$ have $\mathcal{P}$?

(b) If $Y = H$ and $\alpha$ is injective (or proper) and $G$ has $\mathcal{P}$, when does $H$ have $\mathcal{P}$?

Since any homomorphism is a piecewise affine map, properties for which there is a positive answer to (a) must be preserved by quotients and properties for which there is a positive answer to (b) must be preserved by closed subgroups. As the following example shows, other restrictions on $\mathcal{P}$ must also be imposed.

**Example 1.** Suppose that $G$ contains a finite-index closed — and therefore open — normal subgroup $N$. Let $\beta : N \to G/N \times N$ be the continuous open homomorphism defined by $\beta(z) = (e_{G/N}, z)$. Let $F \subseteq G$ be a complete set of representatives of distinct cosets of $N$ and for each $x \in F$ define

$$\alpha_x : xN \to G/N \times N \quad \text{by} \quad \alpha_x(y) = (xN, e)\beta(x^{-1}y) = (xN, x^{-1}y).$$

Then $\alpha_x$ is an affine homeomorphism of $xN$ onto $\{xN\} \times N$, so $\alpha : G \to G/N \times N$, defined by putting $\alpha|_{xN} = \alpha_x$ $(x \in F)$, is a homeomorphic piecewise affine bijection. Observe that since the inverse of an affine bijection between cosets is also affine, $\alpha^{-1} : G/N \times N \to G$ is also a piecewise affine homeomorphism. Thus, if $\mathcal{P}$ is a property for which there is a positive answer to either question (a) or (b) above, $G/N \times N$ has $\mathcal{P}$ exactly when $G$ has $\mathcal{P}$ in this situation.

In particular, if $N \rtimes H$ is a semidirect product of a locally compact group $N$ and a finite group $H$, the identity map is a piecewise affine homeomorphism of $N \times H$ onto $N \rtimes H$. However, $N \rtimes H$ may fail to be a homomorphic image of $N \times H$, such as when $N$ and $H$ are chosen to be abelian groups with $N \rtimes H$ nonabelian. As a specific example, consider $G = \mathbb{R} \rtimes \mathbb{Z}_2$, where $\mathbb{Z}_2 = \{\pm 1\}$ acts on $\mathbb{R}$ via $(\pm 1)t = \pm t$. Then $G$ not nilpotent or $[\mathrm{FC}]^-$ (and fails to have any property implying either

of these properties) [Palmer 2001, Chapter 12], but since $\mathbb{R} \times \mathbb{Z}_2$ is abelian, it is both nilpotent and $[\mathrm{FC}]^-$. Hence, these are examples of properties $\mathcal{P}$ that are preserved by both quotients and closed subgroups, yet fail to provide a positive answer to either question (a) or (b), even when the piecewise affine maps involved are homeomorphisms with piecewise affine inverses.

**Definition 2.** We say that a property $\mathcal{P}$ of locally compact groups is *preserved by*

(a) *direct products with finite groups* if $H \times F$ has $\mathcal{P}$ whenever $H$ has $\mathcal{P}$ and $F$ is a finite group;

(b) *locally compact extensions of finite groups* if $G$ has $\mathcal{P}$ whenever it contains a closed (and open) normal subgroup $N$ such that $G/N$ is finite and $N$ has $\mathcal{P}$;

(c) *finite coset unions* if $G$ has $\mathcal{P}$ whenever it can be written as a finite union of closed cosets, each of which has $\mathcal{P}$;

(d) *$\mathcal{P}$-by-compact extensions* if $G$ has $\mathcal{P}$ whenever it contains a compact normal subgroup $K$ such that $G/K$ has $\mathcal{P}$.

The meaning of the statements "$\mathcal{P}$ is preserved by open (closed) subgroups", "$\mathcal{P}$ is preserved by dense-range continuous homomorphisms" and "$\mathcal{P}$ is preserved by dense-range continuous mixed piecewise affine maps" will be clear.

We remark that if $\mathcal{P}$ is preserved by direct products with finite groups and the trivial group has $\mathcal{P}$, then every finite group must have $\mathcal{P}$. Also, $\mathcal{P}$ satisfies condition (a) in Definition 2 whenever it satisfies condition (b), but not conversely: since $[\mathrm{FC}]^-$ is trivially closed under the formation of direct products and contains all finite groups, $\mathcal{P} = [\mathrm{FC}]^-$ satisfies (a), but $\mathbb{R} \rtimes \mathbb{Z}_2$ is not in $[\mathrm{FC}]^-$, so $[\mathrm{FC}]^-$ does not satisfy (b). Observe as well that $\mathcal{P}$ satisfies condition (b) in Definition 2 whenever it satisfies condition (c). In the proof of the following lemma, which establishes a partial converse to this last implication, we will use a theorem due to Neumann [1954] that says that a group cannot be expressed as a finite union of cosets of infinite index. An elegant, analytic proof of this theorem can be found in [Ilie and Spronk 2005]. Recall that open subgroups are always closed, and finite-index closed subgroups are always open.

**Lemma 3.** *If $\mathcal{P}$ is preserved by finite-index closed* (*equivalently open*) *normal subgroups and locally compact extensions of finite groups, then $\mathcal{P}$ is preserved by finite coset unions.*

*Proof.* Suppose that $G$ can be expressed as a finite union of closed cosets, each with property $\mathcal{P}$. By Neumann's theorem, $G$ contains a finite-index closed subgroup $M$ such that $M$ has $\mathcal{P}$. Then $N = \bigcap_{g \in G} gMg^{-1}$, the core of $M$ in $G$, is a finite-index closed normal subgroup of $G$ that is contained in $M$; see, e.g., [Isaacs 1994, Corollary 4.6]. Hence, $N$ has $\mathcal{P}$ and $G$ is an extension of the finite group $G/N$. Therefore, $G$ also has $\mathcal{P}$. □

If $\phi : H \to G$ is an antihomomorphism, then $\check{\phi}(x) := \phi(x^{-1})(= \phi(x)^{-1})$ is a homomorphism with the same range as $\phi$. Therefore if $\mathcal{P}$ is preserved by dense-range homomorphisms, it is also preserved by dense-range antihomomorphisms.

**Proposition 4.** *The following statements are equivalent*:

(i) *$\mathcal{P}$ is preserved by continuous dense-range homomorphisms* (*i.e.*, *quotients in the discrete case*), *locally compact extensions of finite groups and finite-index closed normal subgroups.*

(ii) *$\mathcal{P}$ is preserved by continuous dense-range mixed piecewise affine maps and products with finite groups.*

*Proof.* Assume that statement (i) holds, $H$ has $\mathcal{P}$, and $\alpha : Y \subseteq H \to G$ is a continuous mixed piecewise affine map with dense range in $G$. Employing the notation (†), each of the cosets $\overline{\alpha_i(E_i)}$ has $\mathcal{P}$ and $G = \bigcup_{i=1}^{n} \overline{\alpha_i(E_i)}$. By Lemma 3, $G$ has $\mathcal{P}$. Hence (ii) holds. Suppose, conversely, that statement (ii) holds, and let $N$ be a finite-index closed normal subgroup of $G$. Then, as shown in Example 1, there is a piecewise affine homeomorphic mapping of $G/N \times N$ onto $G$ with piecewise affine inverse. Hence, if $N$ has $\mathcal{P}$, then $G/N \times N$ has $\mathcal{P}$, and therefore $G$ has $\mathcal{P}$. If $G$ has $\mathcal{P}$, then $G/N \times N$ has $\mathcal{P}$, whence $N$, as a quotient of $G/N \times N$, has $\mathcal{P}$. □

**Remarks 5.** If we replace the assumption that $\mathcal{P}$ is preserved by continuous dense-range homomorphisms in statement (i) of Proposition 4 with the statement that $\mathcal{P}$ is preserved by continuous open (closed) epimorphisms — i.e., quotients in the case of open epimorphisms — then we can conclude that $\mathcal{P}$ is preserved by continuous open (closed) surjective mixed piecewise affine maps: when $\alpha$ is open (closed) in the above proof, so is each $\alpha_i$ and therefore $\overline{\alpha_i(E_i)} = \alpha(E_i)$. For discrete groups, each of these conditions is equivalent to the statement that $\mathcal{P}$ is preserved by quotients.

We say that $G$ is *virtually* $\mathcal{P}$ if $G$ contains a finite-index closed (equivalently, open) subgroup with property $\mathcal{P}$.

**Proposition 6.** *The following statements hold*:

(i) *Virtually $\mathcal{P}$ is preserved by locally compact extensions of finite groups.*

(ii) *If $\mathcal{P}$ is preserved by finite-index closed normal subgroups, then so is virtually $\mathcal{P}$.*

(iii) *If $\mathcal{P}$ is preserved by continuous dense-range homomorphisms, then so is virtually $\mathcal{P}$.*

(iv) *If $\mathcal{P}$ is preserved by finite-index closed normal subgroups and locally compact extensions of finite groups, then every virtually-$\mathcal{P}$ group has property $\mathcal{P}$.*

*Proof.* (i) This is obvious.

(ii) Let $M$ be a finite-index closed subgroup of $G$ with property $\mathcal{P}$, and let $N$ be a finite-index closed normal subgroup of $G$. Then $N \cap M$ is a finite-index closed

subgroup of $N$, since — by a standard (readily verified) fact — $|N : N \cap M| \leq |G : M| < \infty$. Moreover, $|M : N \cap M| \leq |G : N| < \infty$, so $N \cap M$ is a finite-index closed normal subgroup of $M$. Since $M$ has $\mathcal{P}$, so does $N \cap M$. Hence, $N$ is virtually $\mathcal{P}$.

(iii) Let $\phi : H \to G$ be a continuous dense-range homomorphism and suppose that $M$ is a closed subgroup of $H$ with finite index — say $H = \bigcup_{i=1}^{n} h_i M$ — with property $\mathcal{P}$. Then $\overline{\phi(M)}$ has $\mathcal{P}$ and

$$\bigcup_{i=1}^{n} \phi(h_i)\overline{\phi(M)} = \overline{\bigcup_{i=1}^{n} \phi(h_i M)} = \overline{\phi(H)} = G.$$

Hence, $G$ is virtually $\mathcal{P}$.

(iv) A virtually-$\mathcal{P}$ group is a finite union of cosets with $\mathcal{P}$, so this is an immediate consequence of Lemma 3. □

We note as well that if $\mathcal{P}$ is preserved by open (respectively closed) subgroups, then so is virtually $\mathcal{P}$. The following, which is an immediate consequence of Propositions 4 and 6, shows that virtually $\mathcal{P}$ is often preserved by mixed piecewise affine maps.

**Corollary 7.** *If $\mathcal{P}$ is preserved by continuous dense-range homomorphisms and finite-index closed normal subgroups, then virtually $\mathcal{P}$ is preserved by continuous dense-range mixed piecewise affine maps.*

We were unable to find a reference for the following lemma.

**Lemma 8.** *Let $\phi : H \to G$ be a continuous (anti-)homomorphism, $K = \ker \phi$, $\phi_K : H/K \to G : xK \mapsto \phi(x)$. Then $\phi$ is proper if and only if $K$ is compact and $\phi_K$ is proper.*

*Proof.* Suppose that $K$ is compact and $A$ is a compact subset of $H/K$. Choose a compact subset $L$ of $H$ such that $\pi(L) = A$, where $\pi : H \to H/K$ is the quotient map [Fell and Doran 1988, Proposition III.2.5]. Since $\pi^{-1}(A) = LK$, which is compact, $\pi$ is proper. Hence, if $K$ is compact and $\phi_K$ is proper, then $\phi = \phi_K \circ \pi$ is proper. Conversely, if $\phi$ is proper, then $K = \phi^{-1}(\{e_G\})$ is compact and given any compact subset $C$ of $G$, $\phi^{-1}(C) = \pi^{-1}(\phi_K^{-1}(C))$ is compact, whence $\phi_K^{-1}(C) = \pi(\phi^{-1}(C))$ is compact. Hence, $\phi_K$ is proper. □

**Lemma 9.** *A property $\mathcal{P}$ of locally compact groups is preserved by closed subgroups and $\mathcal{P}$-by-compact extensions if and only if*

(∗) *$H$ has $\mathcal{P}$ whenever there exists a proper continuous (anti-)homomorphism mapping $H$ into a locally compact group $G$ that has $\mathcal{P}$.*

*Proof.* Suppose that $\alpha : H \to G$ is a proper continuous (anti-)homomorphism where $G$ has $\mathcal{P}$ and $\mathcal{P}$ is preserved by closed subgroups and $\mathcal{P}$-by-compact extensions. Letting $K = \ker \alpha$, $K$ is compact and $\alpha_K$ is a continuous proper (anti-)isomorphism of $H/K$ onto its image by Lemma 8. Since proper maps are closed, $\alpha_K$ is, in fact, a topological (anti-)isomorphism of $H/K$ onto $\alpha_K(H/K)$. We can conclude that $H/K$ has $\mathcal{P}$, and therefore $H$ has $\mathcal{P}$. Conversely, if $\mathcal{P}$ satisfies $(*)$, then $\mathcal{P}$ is obviously preserved by closed subgroups and, since the quotient map of $G$ onto $G/K$ is proper when $K$ is compact, it is preserved by $\mathcal{P}$-by-compact extensions. □

Suppose that $H$ has $\mathcal{P}$ whenever there exists a continuous mixed piecewise affine proper mapping of $H$ into a locally compact group $G$ that has $\mathcal{P}$ and, further, that $\mathcal{P}$ is preserved by the formation of direct products with finite groups. If $N$, a closed finite-index normal subgroup of $G$, has $\mathcal{P}$, then $G/N \times N$ has $\mathcal{P}$ and, by Example 1, one can define a homeomorphic piecewise affine mapping of $G/N \times N$ onto $G$; hence $G$ has $\mathcal{P}$. This, together with Lemma 9, establishes "(ii) implies (i)" of the following proposition.

**Proposition 10.** *The following statements are equivalent*:

(i) *$\mathcal{P}$ is preserved by closed subgroups, $\mathcal{P}$-by-compact extensions, and locally compact extensions of finite groups.*

(ii) *$\mathcal{P}$ is preserved by the formation of direct products with finite groups, and $H$ has $\mathcal{P}$ whenever there exists a continuous mixed piecewise affine proper mapping of $H$ into a locally compact group $G$ that has $\mathcal{P}$.*

*Proof.* We only need to show that statement (i) implies the second condition found in statement (ii). To this end, let $\alpha : H \to G$ be a continuous mixed piecewise affine proper map. Using the notation (†), each $\alpha_i$ is a proper continuous affine, or antiaffine, mapping of $E_i$ into $G$ [Ilie 2004, Proposition 4.6]. Since each $\alpha_i$ can be obtained through translation of a continuous proper homomorphism, or antihomomorphism, on the subgroup $E_i^{-1}E_i$ of $H$, each coset $E_i$ has $\mathcal{P}$ by Lemma 9. Since $H$ is the union of the closed cosets $E_i$ $(i = 1, \ldots, n)$, Lemma 3 allows us to conclude that $H$ has $\mathcal{P}$. □

**Example 11.** Some examples of properties of locally compact groups that are preserved by open subgroups (and therefore by finite-index closed normal subgroups), continuous dense-range homomorphisms, and locally compact extensions of finite groups are amenability and compactness; within the class of discrete groups, torsion, local finiteness, polynomial growth and exponential growth have these hereditary properties. Thus, for each of these properties, virtually $\mathcal{P}$ and $\mathcal{P}$ are equivalent, and each is preserved by continuous dense-range mixed piecewise affine maps.

Examples of some properties that are preserved by open subgroups, quotients, and locally compact extensions of finite groups are those listed in the last paragraph

and the properties of being an [IN]-group or a [SIN]-group. These properties are all preserved by continuous, open mixed piecewise affine surjections.

Examples of some properties $\mathcal{P}$ that are preserved by closed subgroups, $\mathcal{P}$-by-compact extensions, and locally compact extensions of finite groups are amenability, compactness and the property of being an [IN]-group. For each of these properties, $H$ has $\mathcal{P}$ whenever there exists a continuous mixed piecewise affine proper mapping of $H$ into $G$ for some $G$ with $\mathcal{P}$. A reference for these assertions is [Palmer 2001, Chapter 12].

## 2. Remarks concerning Fourier algebra homomorphisms

With pointwise-defined operations and a particular norm that dominates the uniform norm, the Fourier–Stieltjes algebra $B(G)$ is a Banach algebra of continuous complex-valued functions on $G$ containing the Fourier algebra $A(G)$ as a closed ideal [Eymard 1964]. A long-standing open problem in abstract harmonic analysis asks for a description of every homomorphism mapping $A(G)$ into $B(H)$ and, as we have already noted, piecewise affine maps have primarily been studied in relation to this problem. A solution was obtained by Cohen [1960] in the abelian case, a solution that was generalized by Ilie and Spronk [2005] when $G$ is amenable and the homomorphism is completely bounded, and Pham [2010] when the homomorphism is norm decreasing.

Using the fact that $A(G)$ separates points and closed sets, i.e., $A(G)$ is a regular algebra of continuous functions on $G$, and the fact that the Gelfand spectrum of $A(G)$ — the set of nonzero multiplicative linear functionals on $A(G)$ — is exactly the set of point-evaluation maps $\delta_g(u) := u(g)$ $(g \in G,\ u \in A(G))$, one can see that for any homomorphism $\varphi : A(G) \to B(H)$ there is an open subset $Y$ of $H$ and a continuous map $\alpha : Y \to G$ such that $\varphi = j_\alpha$, where for $u \in A(G)$

$$j_\alpha(u) = \begin{cases} u \circ \alpha & \text{on } Y, \\ 0 & \text{on } H \backslash Y. \end{cases}$$

(For each $h \in H$, either $\delta_h \circ \varphi = 0$ or $\delta_h \circ \varphi$ belongs to the Gelfand spectrum of $A(G)$, whence $\delta_h \circ \varphi = \delta_{\alpha(h)}$ for some $\alpha(h) \in G$. Letting $Y = \{h \in H : \delta_h \circ \varphi \neq 0\}$, one obtains $\alpha : Y \to G$ such that $\varphi = j_\alpha$.) By [Ilie 2004, Proposition 3.9], which does not require that $G$ be amenable or that $\alpha$ be piecewise affine, $j_\alpha : A(G) \to B(H)$ maps $A(G)$ into $A(H)$ exactly when $\alpha$ is a proper map. An easy application of the regularity of $A(G)$ is that a map $\varphi = j_\alpha : A(G) \to B(H)$ is injective exactly when $\alpha : Y \subseteq H \to G$ has dense range. Observe as well that $Y = H$ exactly when $\delta_h \circ \varphi \neq 0$ for each $h \in H$. These facts are used below without comment.

As preduals of von Neumann algebras, $A(G)$ and $B(H)$ have operator space structures with respect to which they are completely contractive Banach algebras

[Effros and Ruan 2000], so it makes sense to speak of completely bounded homomorphisms $\varphi : A(G) \to B(H)$. If $\alpha : Y \subseteq H \to G$ is continuous and piecewise affine, Ilie and Spronk showed that $j_\alpha$ is a completely bounded homomorphism of $A(G)$ into $B(H)$ and, moreover, when $G$ is amenable, every completely bounded homomorphism $\varphi : A(G) \to B(H)$ equals $j_\alpha$ for some continuous piecewise affine map $\alpha : Y \subseteq H \to G$ [Ilie and Spronk 2005, Theorem 3.7]. Thus, the following statement is an immediate consequence of Proposition 4, Corollary 7 and Proposition 10.

**Proposition 12.** *Suppose that $G$ is amenable and there exists a completely bounded homomorphism $\varphi$ mapping $A(G)$ into $B(H)$:*

(i) *Suppose that $\mathcal{P}$ is preserved by continuous dense-range homomorphisms, locally compact extensions of finite groups, and finite-index closed normal subgroups. If $\varphi$ is injective and $H$ has $\mathcal{P}$, then so does $G$.*

(ii) *Suppose that $\mathcal{P}$ is preserved by continuous dense-range homomorphisms and finite-index closed normal subgroups. If $\varphi$ is injective and $H$ is virtually $\mathcal{P}$, then so is $G$.*

(iii) *Suppose that $\mathcal{P}$ is preserved by closed subgroups, $\mathcal{P}$-by-compact extensions and locally compact extensions of finite groups. Suppose further that $\varphi$ maps $A(G)$ into $A(H)$ and for each $h \in H$, $\delta_h \circ \varphi \neq 0$. If $G$ has $\mathcal{P}$, then so does $H$.*

Amenability of Banach algebras is not, in general, preserved by closed subalgebras, much less injective homomorphisms; for example, the semigroup algebra $\ell^1(\mathbb{N})$ is a nonamenable subalgebra of the (Connes) amenable Banach algebra $\ell^1(\mathbb{Z})$. However, since $A(H)$ is an amenable Banach algebra ($B(H)$ is a Connes amenable Banach algebra) exactly when $H$ is virtually abelian [Forrest and Runde 2005; Runde and Uygul 2015] and the property of being abelian is preserved by subgroups and continuous dense-range homomorphisms, the following is an immediate corollary of Proposition 12(ii).

**Corollary 13.** *Suppose that $G$ is amenable and $A(H)$ is amenable (equivalently, $B(H)$ is Connes amenable). If there exists an injective completely bounded homomorphism $\varphi$ mapping $A(G)$ into $A(H)$ or $B(H)$, then $A(G)$ is amenable.*

We remark that by applying the main result in [Pham 2010], we obtain the same conclusions in Proposition 12 and Corollary 13 if we drop the condition that $G$ is amenable and replace the assumption of the existence of a completely bounded homomorphism with that of a norm-decreasing homomorphism.

N. Spronk [2010, Conjecture 4.8] has conjectured that when $G$ is amenable, every homomorphism $\varphi : A(G) \to B(H)$ takes the form $\varphi = j_\alpha$ for some mixed piecewise affine map $\alpha : Y \subseteq H \to G$. If correct, then Propositions 4, 7 and 10 would imply that the statements of Proposition 12 and Corollary 13 hold without the assumption that

$\varphi$ is completely bounded. Thus, the results of this note suggest a possible method of testing the conjecture: for instance, an example of an amenable, but not virtually abelian, group $G$ and a virtually abelian group $H$ for which there exists an injective homomorphism $\varphi$ mapping $A(G)$ into $B(H)$ would disprove the conjecture. On the other hand, since the conjecture may well be correct, Proposition 12 suggests that when $G$ is amenable, every Fourier algebra homomorphism $A(G) \to B(H)$ preserves certain properties $\mathcal{P}$, as described in the proposition. Explicitly, we have the following question:

**Question 14.** Given a specific property $\mathcal{P}$ satisfying the conditions described in one of the statements in Proposition 12, does the corresponding statement of Proposition 12 hold if the homomorphism $\varphi$ is not assumed to be completely bounded? That is, can such a statement be established without necessarily verifying the Spronk conjecture?

Any positive answer would lend evidence in support of the conjecture (and a negative answer would disprove it). For example, since it is known that the property of being an amenable locally compact group satisfies all of the conditions considered in this note, Proposition 12(iii) suggests the following, which, as we now observe, is a consequence of [Kaniuth and Ülger 2010, Theorem 5.1]: this theorem states that a locally compact group $G$ is amenable if and only if $A(G)$ contains a bounded net $(e_i)_i$ converging pointwise on $G$ to 1 (i.e., $A(G)$ contains a "$\Delta$-weak bounded approximate identity").

**Proposition 15.** *Suppose there exists a homomorphism $\varphi : A(G) \to A(H)$ such that for each $h \in H$, $\delta_h \circ \varphi \neq 0$. If $G$ is amenable, then so is $H$.*

*Proof.* Since $G$ is amenable, $A(G)$ has a $\Delta$-weak bounded approximate identity $(e_i)_i$. As noted above, $\varphi = j_\alpha$ for some (continuous, proper) map $\alpha : H \to G$. As noted by Pham [2010], since $A(H)$ is semisimple, $\varphi$ is automatically bounded, so $\varphi(e_i)$ is a bounded net in $A(H)$ such that for each $h \in H$, $\varphi(e_i)(h) = e_i(\alpha(h)) \to 1$. Thus, $\varphi(e_i)$ is a $\Delta$-weak bounded approximate identity in $A(H)$, whence $H$ is amenable by [Kaniuth and Ülger 2010, Theorem 5.1]. □

We remark that when $G$ is amenable, $A(G)$ actually has a bounded approximate identity $(e_i)_i$ (and the converse holds) by Leptin's theorem, but it is not clear that $\varphi(e_i)$ in the proof of Proposition 15 is then a bounded approximate identity for $A(H)$. That is, more than Leptin's theorem was required to prove the above proposition. Observe that in establishing Proposition 15, we did not assume that amenability actually satisfies any of the hereditary properties described in Propositions 10 and 12(iii), because these hereditary properties are not employed in the proof of [Kaniuth and Ülger 2010, Theorem 5.1] (and the theory on which it depends). Since whenever $\alpha : H \to G$ is a proper continuous mixed piecewise affine map, $\varphi = j_\alpha$ is a

homomorphism of $A(G)$ into $A(H)$ such that, for each $h \in H$, $\delta_h \circ \varphi \neq 0$, we obtain — *independent of the hereditary properties of amenability* (and therefore independent of Proposition 10) — the following immediate corollary of Proposition 15.

**Corollary 16.** *If $G$ is amenable and there exists a proper continuous mixed piecewise affine map $\alpha$ of $H$ into $G$, then $H$ is amenable. In particular, closed subgroups of amenable locally compact groups are amenable.*

Thus, [Kaniuth and Ülger 2010, Theorem 5.1] and the basic fact that proper continuous group homomorphisms determine Fourier algebra homomorphisms yield a new proof that closed subgroups of locally compact groups are amenable. This seems interesting because although [Kaniuth and Ülger 2010, Theorem 5.1] is certainly not at all obvious, the standard proof of this fundamental hereditary property, which in the nondiscrete case involves the construction of a Bruhat function for $H$ on $G$ (e.g., see [Pier 1984, Section 13] or [Runde 2002, Section 1.2]), is also not at all obvious.

## References


[Cohen 1960] P. J. Cohen, "On homomorphisms of group algebras", *Amer. J. Math.* **82** (1960), 213–226. MR Zbl

[Effros and Ruan 2000] E. G. Effros and Z.-J. Ruan, *Operator spaces*, London Mathematical Society Monographs (N.S.) **23**, Clarendon Press/Oxford University Press, New York, 2000. MR Zbl

[Eymard 1964] P. Eymard, "L'algèbre de Fourier d'un groupe localement compact", *Bull. Soc. Math. France* **92** (1964), 181–236. MR Zbl

[Fell and Doran 1988] J. M. G. Fell and R. S. Doran, *Representations of $^*$-algebras, locally compact groups, and Banach $^*$-algebraic bundles, I: Basic representation theory of groups and algebras*, Pure and Applied Mathematics **125**, Academic Press, Boston, 1988. MR Zbl

[Forrest and Runde 2005] B. E. Forrest and V. Runde, "Amenability and weak amenability of the Fourier algebra", *Math. Z.* **250**:4 (2005), 731–744. MR Zbl

[Ilie 2004] M. Ilie, "On Fourier algebra homomorphisms", *J. Funct. Anal.* **213**:1 (2004), 88–110. MR Zbl

[Ilie and Spronk 2005] M. Ilie and N. Spronk, "Completely bounded homomorphisms of the Fourier algebras", *J. Funct. Anal.* **225**:2 (2005), 480–499. MR Zbl

[Ilie and Stokke 2008] M. Ilie and R. Stokke, "Weak*-continuous homomorphisms of Fourier–Stieltjes algebras", *Math. Proc. Cambridge Philos. Soc.* **145**:1 (2008), 107–120. MR Zbl

[Isaacs 1994] I. M. Isaacs, *Algebra: a graduate course*, Brooks/Cole, Pacific Grove, CA, 1994. MR

[Kaniuth and Ülger 2010] E. Kaniuth and A. Ülger, "The Bochner–Schoenberg–Eberlein property for commutative Banach algebras, especially Fourier and Fourier–Stieltjes algebras", *Trans. Amer. Math. Soc.* **362**:8 (2010), 4331–4356. MR Zbl

[Neumann 1954] B. H. Neumann, "Groups covered by permutable subsets", *J. London Math. Soc.* **29** (1954), 236–248. MR Zbl

[Palmer 2001] T. W. Palmer, *Banach algebras and the general theory of $*$-algebras, II*, Encyclopedia of Mathematics and its Applications **79**, Cambridge University Press, Cambridge, 2001. MR Zbl

[Pham 2010] H. L. Pham, "Contractive homomorphisms of the Fourier algebras", *Bull. Lond. Math. Soc.* **42**:5 (2010), 937–947. MR Zbl

[Pier 1984] J.-P. Pier, *Amenable locally compact groups*, John Wiley & Sons, New York, 1984. MR Zbl

[Rudin 1962] W. Rudin, *Fourier analysis on groups*, Interscience Tracts in Pure and Applied Mathematics **12**, Interscience, New York, 1962. MR Zbl

[Runde 2002] V. Runde, *Lectures on amenability*, Lecture Notes in Mathematics **1774**, Springer, 2002. MR Zbl

[Runde and Uygul 2015] V. Runde and F. Uygul, "Connes-amenability of Fourier–Stieltjes algebras", *Bull. Lond. Math. Soc.* **47**:4 (2015), 555–564. MR Zbl

[Spronk 2010] N. Spronk, "Amenability properties of Fourier algebras and Fourier–Stieltjes algebras: a survey", pp. 365–383 in *Banach algebras 2009*, edited by R. J. Loy et al., Banach Center Publ. **91**, Polish Acad. Sci. Inst. Math., Warsaw, 2010. MR Zbl

Received: 2018-02-27 Accepted: 2018-09-09

serina.camungol@concordia.ab.ca *Department of Mathematics and Statistics, University of Winnipeg, Winnipeg, MB, Canada*

mmorison@uwaterloo.ca *Department of Mathematics and Statistics, University of Winnipeg, Winnipeg, MB, Canada*

skylarnicol93@gmail.com *Department of Mathematics and Statistics, University of Winnipeg, Winnipeg, MB, Canada*

r.stokke@uwinnipeg.ca *Department of Mathematics and Statistics, University of Winnipeg, Winnipeg, MB, Canada*

INVOLVE 12:3 (2019)
dx.doi.org/10.2140/involve.2019.12.503

# Bin decompositions

Daniel Gotshall, Pamela E. Harris, Dawn Nelson,
Maria D. Vega and Cameron Voigt

(Communicated by Stephan Garcia)

It is well known that every positive integer can be expressed as a sum of nonconsecutive Fibonacci numbers provided the Fibonacci numbers satisfy $F_n = F_{n-1} + F_{n-2}$ for $n \geq 3$, $F_1 = 1$ and $F_2 = 2$. For any $n, m \in \mathbb{N}$ we create a sequence called the $(n, m)$-bin sequence with which we can define a notion of a legal decomposition for every positive integer. These sequences are not always positive linear recurrences, which have been studied in the literature, yet we prove, that like positive linear recurrences, these decompositions exist and are unique. Moreover, our main result proves that the distribution of the number of summands used in the $(n, m)$-bin legal decompositions displays Gaussian behavior.

## 1. Introduction

Edouard Zeckendorf [1972] proved that any positive integer can be uniquely decomposed as a sum of nonconsecutive Fibonacci numbers provided we use the recurrence $F_1 = 1$, $F_2 = 2$, and $F_n = F_{n-1} + F_{n-2}$ for $n \geq 3$. Since then numerous researchers have generalized Zeckendorf's theorem to other recurrence relations [Miller and Wang 2014; Catral et al. 2014; Demontigny et al. 2014a; 2014b; Koloğlu et al. 2011; Lengyel 2006]. Most work involved recurrence relations with positive leading terms, called positive linear recurrences (PLRs), until Catral, Ford, Harris, Miller, and Nelson [Catral et al. 2014; 2016; 2017] generalized these results to the $(s, b)$-Generacci sequences and the Fibonacci quilt sequence, which are defined by nonpositive linear recurrences, and Dorward, Ford, Fourakis, Harris, Miller, Palsson, and Paugh [Dorward et al. 2017a; 2017b] generalized them to the $m$-gonal sequences, which arise from a geometric construction via inscribed $m$-gons. The main results in these studies involved determining the uniqueness of the decompositions of nonnegative integers using the numbers in these new

*MSC2010:* 11B39, 65Q30, 60B10.

*Keywords:* Zeckendorf decompositions, bin decompositions, Gaussian behavior, integer decompositions.

Harris was supported by NSF award DMS-1620202.

sequences, determining whether the behavior arising from the mean number of summands in these decompositions is Gaussian, and other related results.

A way to interpret the creation of the $(s, b)$-Generacci sequences is to imagine an infinite number of bins each containing $b$ distinct positive integers. Given a number $\ell \in \mathbb{N}$, we decompose it as a sum of elements in the sequence such that the terms satisfy (1) no two numbers in the sequence used in the decomposition appear in the same bin, and (2) we do not use numbers in $s$ bins to the left and right of any bin containing a summand used in the decomposition of $\ell$. If such a decomposition of $\ell$ exists using the numbers in the sequence, we then say that $\ell$ has a legal decomposition. If every positive integer $\ell$ has a legal decomposition, then we call the sequence of numbers satisfying this property the $(s, b)$-Generacci sequence. Note that the $(1, 1)$-Generacci sequence gives rise to the Fibonacci sequence, as we have bins with only one integer and we cannot use any consecutive integers in any decomposition.

Motivated by the bin construction used in the $(s, b)$-Generacci sequences, we create the $(n, m)$-bin sequences. These sequences are defined by nonpositive linear recurrences and depend on the positive integer parameters $s, b$ for Generacci sequences and $n, m$ for bin sequences. The terms of an $(n, m)$-bin sequence $\{a_x\}_{x=0}^{\infty}$ can be pictured via

$$\underbrace{\underbrace{a_0, \ldots, a_{n-1}}_{n}, \underbrace{a_n, \ldots, a_{n+m-1}}_{m}}_{\mathcal{B}_0}, \ldots, \underbrace{\underbrace{a_{(n+m)k}, \ldots, a_{(n+m)k+n-1}}_{n}, \underbrace{a_{(n+m)k+n}, \ldots, a_{(n+m)k+n+m-1}}_{m}}_{\mathcal{B}_k}, \ldots. \quad (1)$$

Note that the first term in the sequence is indexed by 0. Notice also that there are $n$ terms in the first bin and $m$ terms in the next. The number of terms in each subsequent bin alternates between $n$ and $m$. We use the notation $\mathcal{B}_k$ to indicate a pair of bins of size $n$ and $m$, in that order. Given a term in the sequence, $a_x$, we can determine which $\mathcal{B}_k$ contains $a_x$ and whether $a_x$ is in the $n$- or $m$-sized bin by using the division algorithm to write $x = (n+m)k + i$. If $0 \le i \le n-1$ then $a_x$ is in the $n$-sized bin. If $n \le i \le m+n-1$ then $a_x$ is in the $m$-sized bin. For example, consider the (2,3)-bin sequence and term $a_{44}$. Since $44 = (2+3)8+4$, we know $a_{44} \in \mathcal{B}_8$ and since $i = 4 \ge 2 = n$, we know $a_{44}$ is the third term in the $m = 3$-sized bin.

Before defining how we construct the sequences, we need to establish the notion of a legal decomposition.

**Definition 1.1.** Let an increasing sequence of integers $\{a_i\}_{i=0}^{\infty}$, divided into bins of sizes $n$ and $m$ be given. For any $n, m \in \mathbb{N}$, an $(n, m)$-*bin legal decomposition* of an integer using summands from this sequence is a decomposition in which no two summands are from the same or adjacent bins.

As described in [Demontigny et al. 2014a], this notion of legal decompositions is an $f$-decomposition defined by the function $f : \mathbb{N}_0 \to \mathbb{N}_0$ with

$$f(j) = \begin{cases} m+i & \text{if } j \equiv i \bmod m+n \text{ and } 0 \le i \le n-1, \\ i & \text{if } j \equiv i \bmod m+n \text{ and } n \le i \le m+n-1. \end{cases} \tag{2}$$

In other words, if $a_j$ is a summand in an $(n, m)$-bin legal decomposition, then none of the previous $f(j)$ terms $(a_{j-f(j)}, a_{j-f(j)+1}, \ldots, a_{j-1})$ are in the decomposition. Consider the (2,3)-bin legal decompositions. Then $f : \mathbb{N}_0 \to \mathbb{N}_0$ is the periodic function

$$\{f(j)\} = \{3, 4, 2, 3, 4, 3, 4, 2, 3, 4, \ldots\}.$$

Note $f(44) = 4$, so if $a_{44}$ is a term in an $(n, m)$-bin legal decomposition, then $a_{40}, a_{41}, a_{42}, a_{43}$ are not in the decomposition. Notice that $a_{42}, a_{43}$ are other terms in the 3-bin (the bin of size 3) that contains $a_{44}$ and that $a_{40}, a_{41}$ are the two terms in the previous 2-bin (the bin of size 2).

Through an immediate application of Theorems 1.2 and 1.3 from [Demontigny et al. 2014a] we can establish that for any $n, m \in \mathbb{N}$, $(n, m)$-bin legal decompositions are unique and we get Proposition 1.2.

**Proposition 1.2.** *For each pair of $n, m \in \mathbb{N}$ there is a unique sequence such that every positive integer has a unique $(n, m)$-bin legal decomposition.*

With this result at hand, we can now formally define an $(n, m)$-bin sequence.

**Definition 1.3.** For each pair of $n, m \in \mathbb{N}$, an *$(n, m)$-bin sequence* is the unique sequence such that every positive integer has a unique $(n, m)$-bin legal decomposition.

Using this definition one can verify that the $(2, 3)$-bin sequence begins

$$\underbrace{1, 2}, \underbrace{3, 4, 5}, \underbrace{6, 9}, \underbrace{12, 18, 24}, \underbrace{30, 42}, \underbrace{54, 84, 114},$$
$$\underbrace{144, 198}, \underbrace{252, 396, 540}, \underbrace{684, 936}, \underbrace{1188, 1872, 2556}, \ldots$$

and that the $(2, 3)$-bin legal decomposition of 2018 is $2018 = 1872 + 144 + 2$. We also note that we can once again recover the Fibonacci sequence, which in this case is given by the $(1, 1)$-bin sequence.

In Section 2 we establish a recurrence for the $(n, m)$-bin sequences.

**Theorem 1.4.** *Assume $\{a_x\}_{x=0}^{\infty}$ is an $(n, m)$-bin sequence. Then for all $n, m \ge 1$ and $x \ge 2(m+n)$,*

$$a_x = (m+n+1)a_{x-(m+n)} - mna_{x-2(m+n)}. \tag{3}$$

We note that the recurrence above is sometimes a PLR and sometimes it is not. For example, as noted previously, the $(1, 1)$-bin legal decompositions are exactly the Zeckendorf decompositions, and use the Fibonacci numbers, which are defined

via a PLR. However, when $n = 2$ and $m = 1$ the recurrence above is not a PLR and we show this in the Appendix. This provides further motivation to study sequences that are more broadly defined and do not necessarily fall under (or out of) the PLR definition.

Our main result establishes that the number of summands used in $(n, m)$-bin legal decompositions of the natural numbers follows a Gaussian distribution.

**Theorem 1.5** (Gaussian behavior of summands). *Let the random variable $Y_k$ denote the number of summands in the (unique) $(n, m)$-bin legal decomposition of an integer chosen uniformly at random from $[0, a_{(n+m)k})$. Normalize $Y_k$ to $Y_k' = (Y_k - \mu_k)/\sigma_k$, where $\mu_k$ and $\sigma_k$ are the mean and variance of $Y_k$ respectively. Then*

$$\mu_k = Ck + O(1), \quad \sigma_k^2 = C'k + O(1) \tag{4}$$

*for some positive constants*

$$C = \frac{\sqrt{(1+m+n)^2 - 4mn} - 1}{\sqrt{(1+m+n)^2 - 4mn}}, \quad C' = \frac{(m+n)(1+m+n) - 4mn}{\sqrt{(1+m+n)^2 - 4mn}^3}.$$

*Moreover, $Y_k'$ converges in distribution to the standard normal distribution as $k \to \infty$.*

As we noted earlier, the (1, 1)-bin sequence is simply the Fibonacci sequence. In this case, the formulas for the mean and the variance given in (4) simplify to the known formulas obtained by Lekkerkerker [1952] and Kolǒglu et al. [2011]. Lekkerkerker computed that for $x \in [F_n, F_{n+1})$ the mean number of summands in a Zeckendorf decomposition is $n/(\phi^2 + 1) + O(1)$, where $\phi = \frac{1}{2}(1 + \sqrt{5})$. The result is the same when the interval is extended to $x \in [0, F_n)$. In [Koloğlu et al. 2011], the authors showed that for $x \in [F_n, F_{n+1})$ the variance of the number of summands in a Zeckendorf decomposition is $\phi n/(5(\phi + 2)) + O(1)$. Again the result is same when the interval is extended to $x \in [0, F_n)$.

**Corollary 1.6.** *Consider the (1, 1)-bin sequence. For $x \in [0, a_{2k})$ the mean and variance of the number of summands in a (1, 1)-bin legal decomposition are*

$$\mu_k = \frac{\sqrt{5} - 1}{\sqrt{5}} k + O(1) = \frac{1}{\phi^2 + 1} 2k + O(1),$$

$$\sigma_k^2 = \frac{2}{5\sqrt{5}} k + O(1) = \frac{\phi}{5(\phi + 2)} 2k + O(1).$$

The paper is organized as follows. Section 2 establishes needed recurrence relations and proves Theorem 1.4, Section 3 develops helpful generating functions, and Section 4 pulls these ideas together and contains the proof of Theorem 1.5. We end with some directions for future research.

## 2. Recurrence relations

In this section we establish recurrence relations for $(n, m)$-bin sequences. We will establish Theorem 1.4 via the following two technical results. Lemma 2.1 provides a family of recurrence relations. For example, (5) computes the first term in the $n$-bin, (6) computes the remaining terms in the $n$-bin and the first term in the $m$-bin, and (7) computes the remaining terms in the $m$-bin. In contrast, Theorem 1.4 provides a single recurrence relation that can be used to compute any term regardless of its position in the bins.

**Lemma 2.1.** *If $n, m \in \mathbb{N}$, then for $k \geq 1$*

$$a_{(m+n)(k+1)} = a_{(m+n)k+m+n-1} + a_{(m+n)k}, \tag{5}$$

$$a_{(m+n)(k+1)+i} = a_{(m+n)(k+1)+(i-1)} + a_{(m+n)k+n} \quad \text{for } 1 \leq i \leq n, \tag{6}$$

$$a_{(m+n)(k+1)+j} = a_{(m+n)(k+1)+j-1} + a_{(m+n)(k+1)} \quad \text{for } n+1 \leq j \leq m+n-1. \tag{7}$$

*Proof.* Using Theorems 1.2 and 1.3 in [Demontigny et al. 2014a], $a_x = a_{x-1} + a_{x-1-f(x-1)}$. If $x = (m+n)(k+1)$, then $x - 1 = (m+n)k + m + n - 1$ and $f((m+n)k+m+n-1) = m+n-1$. Hence (5), is immediate. The other equations follow from a similar argument. □

Lemma 2.2 interweaves the family of recurrence relations to show that if the single recurrence relation (of Theorem 1.4) is true for $x \equiv 0 \pmod{m+n}$, then it is true for all $x$.

**Lemma 2.2.** *Assume $n, m \geq 1$. If*

$$a_x = (m+n+1)a_{x-(m+n)} - mna_{x-2(m+n)} \tag{8}$$

*for $x \geq 2(m+n)$ and $x \equiv 0 \pmod{m+n}$, then (8) is true for all $x \geq 2(m+n)$.*

*Proof.* By hypothesis,

$$a_{(m+n)k} = (m+n+1)a_{(m+n)k-(m+n)} - mna_{(m+n)k-2(m+n)}.$$

In other words,

$$a_{(m+n)k} = (m+n+1)a_{(m+n)(k-1)} - mna_{(m+n)(k-2)}.$$

So applying (5), we have

$$\begin{aligned} a_{(m+n)(k-1)+m+n-1} + a_{(m+n)(k-1)} = {} & (m+n+1)[a_{(m+n)(k-2)+m+n-1} + a_{(m+n)(k-2)}] \\ & - mn[a_{(m+n)(k-3)+m+n-1} + a_{(m+n)(k-3)}]. \end{aligned}$$

Thus

$$\begin{aligned} a_{(m+n)(k-1)+m+n-1} - {} & [(m+n+1)a_{(m+n)(k-2)+m+n-1} - mna_{(m+n)(k-3)+m+n-1}] \\ & = -a_{(m+n)(k-1)} + [(m+n+1)a_{(m+n)(k-2)} - mna_{(m+n)(k-3)}]. \end{aligned}$$

By hypothesis, the right-hand side of this equation is 0. Hence so is the left side and thus (8) is true for $x \equiv m+n-1 \pmod{m+n}$.

Repeating a similar argument several more times shows that (8) is true for all $x$. □

It remains to prove that (8) is true for $x \equiv 0 \pmod{m+n}$. We do this in the following proof and thus establish Theorem 1.4.

*Proof of Theorem 1.4.* Assume $\{a_x\}_{x=0}^{\infty}$ is an $(n, m)$-bin sequence. As explained in Section 1, this sequence is an $f$-sequence defined by the function $f(j)$ given in (2). Note that the period of $f(j)$ is $m+n$ and $m+n \geq f(j)+1$ for all $j$.

By Theorem 1.5 in [Demontigny et al. 2014a], since $f(j)$ is periodic, we know that there is a single recurrence relation for our sequence, and the proof of that theorem gives us an algorithm for computing the single recurrence relation.

Consider the $m+n$ subsequences of $\{a_x\}_{x=0}^{\infty}$ given by terms whose indices are all in the same residue class mod $m+n$. We will begin by finding a recurrence relation for each subsequence

$$a_x = \sum_{i=1}^{m+n+1} c_i a_{x-(m+n)i}. \tag{9}$$

A priori, these relations may be different for each residue class, but Lemma 2.2 tells us that all relations are in fact the same. Thus we focus on the subsequence corresponding to the 0 residue class.

It remains to solve for the constants $c_i$ in (9). To solve for these constants we will use linear algebra techniques; in particular we use matrices and vectors to represent systems of equations. Each of the equations in Lemma 2.1 can be rewritten as vectors (the starred columns are those that are indexed by multiples of $m+n$, beginning with 0, and the columns marked with $\circ$ are indices congruent to $m$ modulo $m+n$):

$$\begin{array}{rcccccccccccccccc}
 & & \star & & & & \circ & & & & & & \star & & & & \circ \\
\vec{v}_0 & = & [1, & -1, & 0, & & & \ldots, & 0, & -1, & 0, & \ldots], & & & & & \\
\vec{v}_1 & = & [0, & 1, & -1, & 0, & & \ldots, & 0, & -1, & 0, & \ldots], & & & & & \\
 & & \vdots & & & & & & & & & & & & & & \\
\vec{v}_{m-1} & = & [0, & \ldots, & 0, & 1, & -1, & 0, & \ldots, & 0, & -1, & 0, & \ldots], & & & & \\
\vec{v}_m & = & [0, & \ldots, & & 0, & 1, & -1, & 0, & \ldots, & 0, & \ldots, & & & 0, & -1], & \\
 & & \vdots & & & & & & & & & & & & & & \\
\vec{v}_{m+n-1} & = & [0, & \ldots, & & & & & 0, & 1, & -1, & 0, & \ldots, & 0, & -1]. & &
\end{array}$$

Vector $\vec{v}_0$ corresponds to the recurrence relation in (5), $\vec{v}_1$ to $\vec{v}_{m-1}$ correspond to the recurrence relations in (7), and $\vec{v}_m$ to $\vec{v}_{m+n-1}$ correspond to the recurrence relations in (6). For all $\vec{v}_j$ the number of leading 0's is $j$ and the number of middle 0's is $f(m+n-j)-1$.

Define $T$ to be the transformation that shifts all coordinates to the right by $(m+n)$ places.

According to the algorithm in [Demontigny et al. 2014a] the goal is to zero out the coordinates that are not indexed by multiples of $m+n$ (the period). Note the first column is indexed by 0. Our first step in this process is to define $\vec{w}_1$, a linear combination of the $\vec{v}_j$. We have

$$\vec{w}_1 = \vec{v}_0 + \cdots + \vec{v}_{m+n-1} = [1,\ 0,\ \ldots,\ 0,\ -m-1,\ 0,\ \ldots,\ 0,\ -n,\ 0],$$

where there are $(m+n-1)$ 0's in the first set and $(m-1)$ 0's in the second set. We continue and use $T$ to define $\vec{w}_2$:

$$\vec{w}_2 = \vec{w}_1 + n\sum_{j=m}^{m+n-1} T\vec{v}_j = [1,\ 0,\ \ldots,\ 0,\ -m-1,\ 0,\ \ldots,\ 0,\ -n,\ 0,\ \ldots,\ 0,\ -n^2],$$

where there are $(m+n-1)$ 0's in the first and second sets and $(m-1)$ 0's in the last set.

Note that in $\vec{w}_0 = \vec{v}_0$, $\vec{w}_1$, and $\vec{w}_2$, the bad coordinates (the coordinates that are not 0 and not indexed by multiples of $m+n$) are given by

$$\vec{u}_0 = [-1, 0 \ldots, 0], \quad \vec{u}_1 = [0, \ldots, 0, -n], \quad \vec{u}_2 = [0, \ldots, 0, -n^2].$$

We simplify by removing the common strings of 0's:

$$\vec{u}_0 = [-1, 0], \quad \vec{u}_1 = [0, -n], \quad \vec{u}_2 = [0, -n^2].$$

There exists a nontrivial solution to $\sum_{j=0}^{2} \lambda_j \vec{u}_j = 0$, namely $\lambda_0 = 0$, $\lambda_1 = -n$, $\lambda_2 = 1$. Using these values, we can write a linear combination of the $\vec{w}_j$ in which we succeed in zeroing out the coordinates that are not multiples of $m+n$:

$$\sum_{j=0}^{2} \lambda_j T^{2-j} \vec{w}_j = [1,\ 0,\ \ldots,\ 0,\ -(m+n+1),\ 0,\ \ldots,\ 0,\ mn,\ 0,\ \ldots].$$

Thus (9) becomes

$$a_x = (m+n+1)a_{x-(m+n)} - mna_{x-2(m+n)}.$$

Note that a priori this is only the recurrence relation for the subsequence given by the terms whose indices are congruent to 0 $(\mathrm{mod}\, m+n)$. Fortunately, applying Lemma 2.2, we see that this recurrence relation is the single relation for the entire sequence. □

## 3. Counting summands with generating functions

In this section we provide generating functions for counting integers with a fixed number of summands in their $(n, m)$-bin legal decomposition. We continue to assume throughout that $\{a_x\}_{x=0}^{\infty}$ is an $(n, m)$-bin sequence.

Let $p_{k,c}$ denote the number of integers $z \in [0, a_{(n+m)k})$ whose legal decomposition contains exactly $c$ summands, where $c \geq 0$. Then by definition

$$p_{0,c} = \begin{cases} 1, & c = 0, \\ 0, & c > 0, \end{cases} \tag{10}$$

$$p_{1,c} = \begin{cases} 1, & c = 0, \\ n+m, & c = 1, \\ 0 & c > 1. \end{cases} \tag{11}$$

Also, for all $k \geq 0$, we have $p_{k,0} = 1$ and $p_{k,1} = k(n+m)$. Moreover, for all $c > k \geq 0$, we have $p_{k,c} = 0$. We also have the following recurrence relation for the values of $p_{k,c}$.

**Proposition 3.1.** *If $k \geq 2$ and $c \geq 0$, then*

$$p_{k,c} = p_{k-1,c} + (m+n)\, p_{k-1,c-1} - nm p_{k-2,c-2}. \tag{12}$$

*Proof.* The decomposition of an integer $z \in [0, a_{(n+m)k})$ either has a summand from the bin $\mathcal{B}_{k-1}$ or it doesn't. If it doesn't then the number of integers with $c$ summands is $p_{k-1,c}$.

If $z$ has a summand in the bin $\mathcal{B}_{k-1}$, then there are two possibilities: either the summand lies in the bin of size $m$ or it lies in the bin of size $n$. In what follows we need to recall that the first sub-bin of $\mathcal{B}_{k-1}$ has size $n$ and the second has size $m$. If the largest summand appearing in the decomposition of $z$ is in the sub-bin of size $m$ then there are $m$ ways to choose it, and since the next-largest legal summand is less than $a_{(n+m)(k-1)}$, there are $p_{k-1,c-1}$ ways to choose the remaining $c-1$ summands. Hence there are $mp_{k-1,c-1}$ integers with $c$ summands and with largest summand from the $m$ sub-bin of $\mathcal{B}_{k-1}$. On the other hand, if the largest summand in the decomposition of $z$ is in the sub-bin of size $n$, the quantity $np_{k-1,c-1}$ over-counts by $nmp_{k-2,c-2}$, because a decomposition with a summand from the sub-bin of size $n$ of $\mathcal{B}_{k-1}$ and a summand from the sub-bin of size $m$ of $\mathcal{B}_{k-2}$ does not give rise to an $(n, m)$-bin legal decomposition. Hence $p_{k,c} = p_{k-1,c} + (m+n)\, p_{k-1,c-1} - nm p_{k-2,c-2}$. □

**Proposition 3.2.** *Let $F(x, y) = \sum_{k\geq 0} \sum_{c\geq 0} p_{k,c} x^k y^c$ be the generating function of the $p_{k,c}$ arising from the $(n, m)$-bin legal decompositions. Then*

$$F(x, y) = \frac{1}{1 - x - (m+n)xy + mnx^2y^2}. \tag{13}$$

*Proof.* Noting that $p_{k,c} = 0$ if either $k < 0$ or $c < 0$, using explicit values of $p_{k,c}$ and the recurrence relation from Proposition 3.1, after some straightforward algebra we obtain

$$F(x, y) = xF(x, y) + (m+n)xyF(x, y) - mnx^2y^2F(x, y) + 1$$

from which (13) follows. □

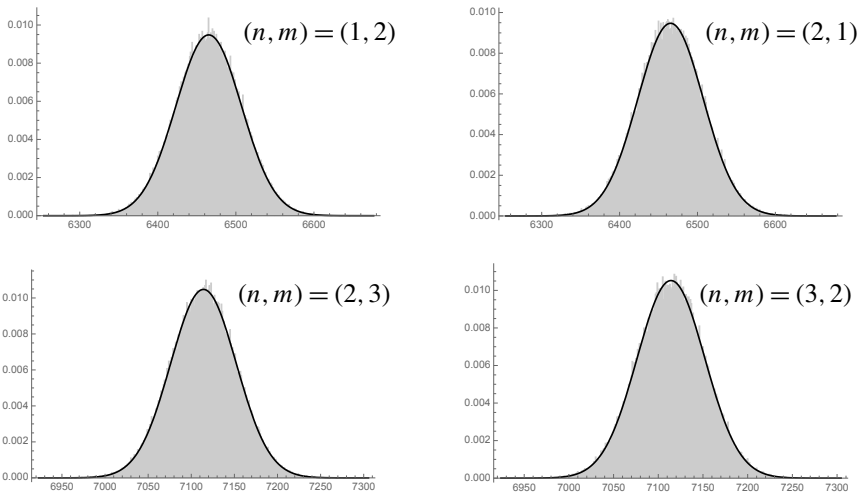


**Figure 1.** Distributions for the number of summands in the $(n, m)$-bin decomposition for a random sample of 100,000 integers from the intervals $[0, a_{10000(m+n)})$.

| $(n, m)$ | predicted mean | sample mean | predicted variance | sample variance |
|---|---|---|---|---|
| (1, 2) | 6464.466094 | 6465.205230 | 1767.766953 | 1770.751318 |
| (2, 1) | 6464.466094 | 6465.418910 | 1767.766953 | 1774.385128 |
| (2, 3) | 7113.248654 | 7114.140920 | 1443.375673 | 1450.656668 |
| (3, 2) | 7113.248654 | 7114.202700 | 1443.375673 | 1437.312966 |

**Table 1.** Predicted means and variances versus sample means and variances for simulations from Figure 1.

## 4. Gaussian behavior

To motivate the main result of this section, we point the reader to experimental observations. Taking samples of 100,000 integers from the intervals $[0, a_{10000(m+n)})$, in Figure 1 we provide a histogram for the distribution of the number of summands in the $(n, m)$-bin decomposition of these integers, when $(n, m) = (1, 2)$, $(2, 1)$, $(2, 3)$, $(3, 2)$. In these figures we also provide the Gaussian curve computed using each sample's mean and variance. Furthermore, Table 1 gives the values of the predicted means and variances as computed using Theorem 1.5, as well as the sample means and variances, for each of the samples considered.

From these observations one might speculate that for any pair of integers $n, m \in \mathbb{N}$ the distribution of the number of summands in the $(n, m)$-bin legal decompositions

of integers in the interval $[0, a_{(n+m)k})$ displays Gaussian behavior. This is in fact the statement of Theorem 1.5.

To prove Theorem 1.5 we first need the following technical results.

**Lemma 4.1.** *For all $m, n, y > 0$, the following inequalities hold*:

$$\gamma^2 > 1 + (m+n)y, \tag{14}$$

$$\gamma > 1, \tag{15}$$

$$1 + (m+n)y + \gamma > 1 + (m+n)y - \gamma > 0, \tag{16}$$

*where $\gamma = \sqrt{(1+(m+n)y)^2 - 4mny^2}$.*

*Proof.* To establish (14) and (15) we note that

$$\gamma^2 = (1+(m+n)y)^2 - 4mny^2 = 1 + 2(m+n)y + (m-n)^2y^2 > 1 + (m+n)y > 1.$$

The first inequality in (16) is clear, while the second is true because

$$(1+(m+n)y)^2 > (1+(m+n)y)^2 - 4mny^2 = \gamma^2 > 1.$$

Hence $1 + (m+n)y > \gamma$. □

**Proposition 4.2.** *Let $g_k(y) := \sum_{c=0}^{k} p_{k,c} y^c$ denote the coefficient of $x^k$ in $F(x, y)$. Then*

$$g_k(y) = \frac{1}{\gamma}\left[\left(\frac{2mny^2}{(1+(m+n)y) - \gamma}\right)^{k+1} - \left(\frac{2mny^2}{(1+(m+n)y) + \gamma}\right)^{k+1}\right],$$

*where again $\gamma = \sqrt{(1+(m+n)y)^2 - 4mny^2}$.*

*Proof.* From Proposition 3.2 we know that

$$F(x, y) = \frac{1}{1 - x - (m+n)xy + mnx^2y^2} = \frac{1}{mny^2}\left(x^2 - \frac{1+(m+n)y}{mny^2} + \frac{1}{mny^2}\right)^{-1}.$$

In order to expand $F(x, y)$ into a power series we will use partial fraction decomposition, but first we must factor

$$x^2 - \frac{1+(m+n)y}{mny^2} + \frac{1}{mny^2}$$

into two linear factors. Using the quadratic formula yields

$$x^2 - \frac{1+(m+n)y}{mny^2} + \frac{1}{mny^2} = (x - \lambda_1)(x - \lambda_2)$$

where

$$\lambda_1 = \lambda_1(y) = \frac{(1+(m+n)y) - \gamma}{2mny^2}, \tag{17}$$

$$\lambda_2 = \lambda_2(y) = \frac{(1+(m+n)y) + \gamma}{2mny^2}. \tag{18}$$

Since the discriminant is positive, by (15), we can use partial fraction decomposition

$$F(x,y)=\frac{1}{mny^2}\left(x^2-\frac{1+(m+n)y}{mny^2}+\frac{1}{mny^2}\right)^{-1}=\frac{1}{mny^2}\left(\frac{A_1}{x-\lambda_1}+\frac{A_2}{x-\lambda_2}\right).$$

Solving for $A_1, A_2$, we get

$$1=A_1(x-\lambda_2)+A_2(x-\lambda_1).$$

If $x=\lambda_1$, then $1=A_1(\lambda_1-\lambda_2)$. Hence $A_1=1/(\lambda_1-\lambda_2)$ and

$$\lambda_1-\lambda_2=\left(\frac{(1+(m+n)y)-\gamma}{2mny^2}\right)-\left(\frac{(1+(m+n)y)+\gamma}{2mny^2}\right)=-\frac{\gamma}{mny^2}.$$

Thus $A_1=-mny^2/\gamma$. Similarly, if $x=\lambda_2$, then $1=A_2(\lambda_1-\lambda_1)$. So $A_2=1/(\lambda_2-\lambda_1)=-A_1$.

Thus

$$\begin{aligned}F(x,y)&=\frac{1}{mny^2}\left(\frac{-A_1}{\lambda_1-x}-\frac{A_2}{\lambda_1-x}\right)\\&=\frac{1}{mny^2}\left(\frac{-A_1}{\lambda_1}\sum_{i=0}^{\infty}\left(\frac{x}{\lambda_1}\right)^i-\frac{A_2}{\lambda_2}\sum_{i=0}^{\infty}\left(\frac{x}{\lambda_2}\right)^i\right).\end{aligned}\tag{19}$$

If $g_k(y)$ denotes the coefficient of $x^k$ in $F(x,y)$, then using (19) we have

$$\begin{aligned}g_k(y)&=\frac{1}{mny^2}\left(\frac{-A_1}{\lambda_1}\left(\frac{1}{\lambda_1}\right)^k-\frac{A_2}{\lambda_2}\left(\frac{1}{\lambda_2}\right)^k\right)\\&=\frac{1}{\lambda_1\gamma}\left(\frac{2(mny^2)}{(1+(m+n)y)-\gamma}\right)^k+\frac{-1}{\lambda_2\gamma}\left(\frac{2(mny^2)}{(1+(m+n)y)+\gamma}\right)^k.\end{aligned}$$

□

To complete the proof of Theorem 1.5 we make use the following result from [Demontigny et al. 2014b].

**Theorem 4.3** [Demontigny et al. 2014b, Theorem 1.8]. *Let $\kappa$ be a fixed positive integer. For each $n$, let a discrete random variable $Y_n$ in $I_n=\{1,2,\dots,n\}$ have*

$$\operatorname{Prob}(Y_n=j)=\begin{cases}p_{j,n}/\sum_{j=1}^{n}p_{j,n} & \text{if } j\in I_n,\\ 0 & \text{otherwise}\end{cases}$$

*for some positive real numbers $p_{1,n}, p_{2,n}, \dots, p_{n,n}$. Let $g_n(y):=\sum_j p_{j,n}y^j$.*

*If $g_n$ has the form $g_n(y)=\sum_{i=1}^{\kappa}q_i(y)\alpha_i^n(y)$, where*

(i) *for each $i\in\{1,\dots,\kappa\}$, $q_i,\alpha_i:\mathbb{R}\to\mathbb{R}$ are three-times differentiable functions which do not depend on $n$;*

(ii) *there exists some small positive* $\epsilon$ *and some positive constant* $\lambda < 1$ *such that for all* $y \in I_\epsilon = [1-\epsilon, 1+\epsilon]$ *we have* $|\alpha_1(y)| > 1$ *and* $|\alpha_i(y)/\alpha_1(y)| < \lambda < 1$ *for all* $i = 2, \ldots, \kappa$;

*then the mean* $\mu_n$ *and variance* $\sigma_n^2$ *of* $Y_n$ *both grow linearly with* $n$. *Specifically*,

$$\mu_n = Cn + d + o(1), \quad \sigma_n^2 = C'n + d' + o(1),$$

*where*

$$C = \frac{\alpha_1'(1)}{\alpha_1(1)}, \quad C' = \frac{d}{dy}\left(\frac{y\alpha_1'(y)}{\alpha_1(y)}\right)\Big|_{y=1} = \frac{\alpha_1(1)[\alpha_1'(1)+\alpha_1''(1)] - \alpha_1'(1)^2}{\alpha_1(1)^2},$$
$$d = \frac{q_1'(1)}{q_1(1)}, \quad d' = \frac{d}{dy}\left(\frac{yq_1'(y)}{q_1(y)}\right)\Big|_{y=1} = \frac{q_1(1)[q_1'(1)+q_1''(1)] - q_1'(1)^2}{q_1(1)^2}.$$

*Moreover*, *if*

(iii) $\alpha_1'(1) \neq 0$ *and* $\frac{d}{dy}[y\alpha_1'(y)/\alpha_1(y)]|_{y=1} \neq 0$, *i.e.*, $C, C' > 0$,

*then as* $n \to \infty$, $Y_n$ *converges in distribution to a normal distribution.*

Throughout the following proof we will simplify some calculations with the substitutions

$$s = m + n, \quad p = mn, \quad \text{and} \quad \beta = \sqrt{(1+m+n)^2 - 4mn}.$$

*Proof of Theorem 1.5.* To prove Gaussian behavior we need only show that $g_k(y)$ satisfies the hypothesis of Theorem 4.3. Note that

$$g_k(y) = q_1(y)\alpha_1^k(y) + q_2(y)\alpha_2^k(y),$$

where

$$q_i(y) = \frac{(-1)^{i+1}2mny^2}{\left(1+(m+n)y+(-1)^i\sqrt{(1+(m+n)y)^2-4mny^2}\right)\sqrt{(1+(m+n)y)^2-4mny^2}}$$

and

$$\alpha_i(y) = \frac{2mny^2}{1+(m+n)y+(-1)^i\sqrt{(1+(m+n)y)^2-4mny^2}}.$$

Condition (i): For each $i = 1, 2$, the functions $q_i(y)$ and $\alpha_i(y)$ are three-times differentiable.

Condition (ii): Let $\epsilon$ be some small positive constant and assume $y \in I_\epsilon = [1-\epsilon, 1+\epsilon]$.

By (16), we see that $0 < \alpha_2(y) < \alpha_1(y)$. Thus for some positive constant $\lambda$, $|\alpha_2(y)/\alpha_1(y)| < \lambda < 1$. Next we show that $\alpha_1(y) > 1$. We begin by noting that

$py^2 > 0$ and $\sqrt{(1+sy)^2-4py^2} > 1$ by (15). Hence

$$\begin{aligned}
0 &< 4py^2\big(py^2+\sqrt{(1+sy)^2-4py^2}-1\big)\\
(1+sy)^2 &< 4py^2\big(py^2+\sqrt{(1+sy)^2-4py^2}-1\big)+(1+sy)^2\\
(1+sy)^2 &< 4p^2y^4+4py^2\sqrt{(1+sy)^2-4py^2}+(1+sy)^2-4py^2\\
(1+sy)^2 &< \big(2py^2+\sqrt{(1+sy)^2-4py^2}\big)^2\\
1+sy &< 2py^2+\sqrt{(1+sy)^2-4py^2}\\
1 &< \frac{2py^2}{1+sy-\sqrt{(1+sy)^2-4py^2}}.
\end{aligned}$$

Condition (iii): First we compute $C=\alpha_1'(1)/\alpha_1(1)$ and prove that it is not 0. Using

$$\alpha_1(y)=\frac{2py^2}{1+sy-\sqrt{(1+sy)^2-4py^2}}$$

we compute

$$\begin{aligned}
\alpha_1'(y)=&\frac{4py}{1+sy-\sqrt{(1+sy)^2-4py^2}}\\
&-\frac{2py^2\big[s-\frac{1}{2}((1+sy)^2-4py^2)^{-1/2}(2s(1+sy)-8py)\big]}{(1+sy-\sqrt{(1+sy)^2-4py^2})^2}.
\end{aligned}$$

Substituting $y=1$ and using a common denominator to add fractions, the numerator of $\alpha_1'(1)$ simplifies to

$$\begin{aligned}
4p(1+s-\beta)-2p\left(s-\frac{2s(1+s)-8p}{2\beta}\right)&=2p\left(2(1+s-\beta)-s+\frac{s(1+s)-4p}{\beta}\right)\\
&=\frac{2p}{\beta}(1+s-\beta)(\beta-1).
\end{aligned}$$

Hence

$$\begin{aligned}
C=\frac{\alpha_1'(1)}{\alpha_1(1)}&=\frac{2p(1+s-\beta)(\beta-1)}{\beta(1+s-\beta)^2}\cdot\frac{1+s-\beta}{2p}\\
&=\frac{\beta-1}{\beta}=\frac{\sqrt{(1+m+n)^2-4mn}-1}{\sqrt{(1+m+n)^2-4mn}}.
\end{aligned}$$

Note that this final value is positive (in particular not zero); see (15).

Second we compute

$$C'=\frac{\alpha_1'(1)-\alpha_1''(1)}{\alpha_1(1)}-\left(\frac{\alpha_1'(1)}{\alpha_1(1)}\right)^2$$

and prove that it is not 0. Note

$$\begin{aligned}\alpha_1''(1) &= \frac{4p\left(s+\frac{4p-s(1+s)}{\beta}\right)^2}{(1+s-\beta)^3} - \frac{8p\left(s+\frac{4p-s(1+s)}{\beta}\right)}{(1+s-\beta)^2} \\ &\qquad + \frac{4p}{1-s-\beta} - \frac{2p\left(\frac{(-4p+s(1+s))^2}{\beta^3}+\frac{4p-s^2}{\beta}\right)}{(1+s-\beta)^2} \\ &= \frac{4p}{1+s-\beta}\left(\frac{4p-s-s^2-\beta-4p+1+2s+s^2}{\beta(1+s-\beta)}\right)^2 - \frac{2p}{(1+s-\beta)^2}\frac{4p}{\beta^3} \\ &= \frac{4p}{(1+s-\beta)\beta^2} - \frac{8p^2}{(1+s-\beta)^2\beta^3}\end{aligned}$$

and using this we find that

$$\begin{aligned}\frac{\alpha_1'(1)-\alpha_1''(1)}{\alpha_1(1)} &= \left(\frac{2p(\beta-1)}{\beta(1+s-\beta)} + \frac{4p}{(1+s-\beta)\beta^2} - \frac{8p^2}{(1+s-\beta)^2\beta^3}\right)\frac{1+s-\beta}{2p} \\ &= \frac{\beta-1}{\beta} + \frac{\beta-1-s}{\beta^3}.\end{aligned}$$

Finally

$$\begin{aligned}C' = \frac{\alpha_1'(1)-\alpha_1''(1)}{\alpha_1(1)} - \left(\frac{\alpha_1'(1)}{\alpha_1(1)}\right)^2 &= \frac{\beta-1}{\beta} + \frac{\beta-1-s}{\beta^3} - \left(\frac{\beta-1}{\beta}\right)^2 && (20)\\ &= \frac{\beta^2-1-s}{\beta^3} && (21)\\ &= \frac{s(1+s)-4p}{\beta^3}. && (22)\end{aligned}$$

By considering (21) with (14) we see that $C' > 0$.

Therefore, by satisfying the conditions of Theorem 4.3, we have completed our proof. □

## 5. Directions for future research

In this paper we considered the construction of $(n, m)$-bin sequences. For $d \in \mathbb{Z}_+$, one natural extension is to consider $\boldsymbol{N} = (n_1, n_2, \ldots, n_d) \in \mathbb{Z}_+^d$ and define $\boldsymbol{N}$-bin sequences in an analogous way to that of $(n, m)$-bin sequences. One could then study the $\boldsymbol{N}$-bin decompositions of positive integers. Namely, do these decompositions exist and are they unique? What is the behavior of the mean number of summands used in the $\boldsymbol{N}$-bin legal decompositions; i.e., is it Gaussian?

Another further generalization would be to consider introducing a new parameter $s \in \mathbb{N}$ which accounts for the number of bins which must be skipped between summands used in a legal $\boldsymbol{N}$-bin decomposition. We call such decompositions the

$(s, \boldsymbol{N})$-bin with skip decompositions. Note that when $s = 1$ and $\boldsymbol{N} = (n, m)$, the $(s, \boldsymbol{N})$-bin with skip decompositions are exactly the $(n, m)$-bin decompositions and when $s \in \mathbb{Z}_+$ and $\boldsymbol{N} = b \in \mathbb{Z}_+$, the $(s, \boldsymbol{N})$-bin with skip decompositions are exactly the $(s, b)$-Generacci decompositions. Therefore the study of the $(s, \boldsymbol{N})$-bin with skip decompositions provides natural ways to generalize prior results in this area.

## Appendix: Negative coefficient in linear recurrence

**Proposition A.1.** *The* $(2, 3)$*-bin sequence is not a positive linear recurrence sequence* (*PLRS*).

*Proof.* By (3) the recurrence relation for the $(2, 3)$-bin sequence is

$$a_x = 4a_{x-3} - 2a_{x-6}.$$

This has characteristic equation $y^6 - 4y^3 + 2$. By Eisenstein's criterion the polynomial $y^6 - 4y^3 + 2$ is irreducible in $\mathbb{Q}[y]$ since there exists a prime $p = 2$ such that $p$ divides all nonleading coefficients of the polynomial, does not divide the leading coefficient, and whose square does not divide the constant term. Thus the polynomial $y^6 - 4y^3 + 2$ cannot be factored into the product of nonconstant polynomials with rational coefficients. Moreover, since this equation is irreducible in $\mathbb{Q}[y]$ our recurrence relation is minimal. By applying Lemma B.1 in [Demontigny et al. 2014a], it is enough to show that all multiples of the characteristic equation cannot have the form

$$y^{k+6} - \sum_{i=0}^{k+5} c_i y^i,$$

with all $c_i > 0$.

Consider the multiple of the characteristic equation (with $p_k \neq 0$)

$$\sum_{i=0}^{k+6} c_i y^i = \left(\sum_{j=0}^{k} p_j y^j\right)(y^6 - 4y^3 + 2) = \sum_{i=0}^{k+6} (p_{i-6} - 4p_{i-3} + 2p_i) y^i.$$

Thus $c_i = p_{i-6} - 4p_{i-3} + 2p_i$. Note that $p_i = 0$ when $i < 0$ and when $i > k$.

We will proceed by contradiction. Hence we assume $c_{k+6} > 0$, and $c_i \leq 0$ whenever $i < k + 6$. Let $t$ be the smallest nonnegative integer such that $p_t \neq 0$. Note that $0 \leq t \leq k$.

We claim that for all integers $j \geq 0$ with $t + 3j < k + 6$, we have $p_{t+3j} < p_{t+3j-3}$ and $p_{t+3j} < 0$. In other words the coefficients become increasingly negative. The proof of this claim is by induction.

Base case $n = 0$: By the definition of $t$, $c_t = p_{t-6} - 4p_{t-3} + 2p_t = 2p_t$. Hence $2p_t = c_t < 0$, because $p_t \neq 0$ and $t < k + 6$. Thus $p_t < 0 = p_{t-3}$.

Base case $n = 1$: We have

$$\begin{aligned} c_{t+3} = p_{t-3} - 4p_t + 2p_{t+3} &\leq 0 \\ 2p_{t+3} &\leq 4p_t \\ p_{t+3} &\leq 2p_t < p_t, \end{aligned}$$

where the last inequality is true because $p_t < 0$.

Inductive step: We have

$$c_{t+3j} = p_{t+3j-6} - 4p_{t+3j-3} + 2p_{t+3j} \leq 0 \tag{23}$$

$$2p_{t+3j} \leq 4p_{t+3j-3} - p_{t+3j-6} \tag{24}$$

$$2p_{t+3j} \leq 4p_{t+3j-3} - p_{t+3j-3} \tag{25}$$

$$p_{t+3j} \leq 1.5p_{t+3j-3} \tag{26}$$

$$p_{t+3j} \leq p_{t+3j-3}. \tag{27}$$

Step (23) is true because $t + 3j < k + 6$. Step (25) is true by the inductive assumption. Finally step (27) is true because $p_{t+3j-3} < 0$.

To establish our contradiction, choose $j^*$ such that $k < t + 3j^* < k + 6$. Thus we have

$$c_{t+3j^*} = p_{t+3j^*-6} - 4p_{t+3j^*-3} + 2p_{t+3j^*} \leq 0 \tag{28}$$

$$p_{t+3j^*-6} \leq 4p_{t+3j^*-3} \tag{29}$$

$$p_{t+3j^*-6} \leq p_{t+3j^*-3}. \tag{30}$$

Step (28) is true because $t + 3j^* < k + 6$. Step (29) is true because $p_i = 0$ when $i > k$. Step (30) is true because $p_{t+3j^*-3} < 0$. But this last line contradicts the claim we just proved above. □

## References

[Catral et al. 2014] M. Catral, P. Ford, P. Harris, S. J. Miller, and D. Nelson, "Generalizing Zeckendorf's theorem: the Kentucky sequence", *Fibonacci Quart.* **52**:5 (2014), 68–90. MR

[Catral et al. 2016] M. Catral, P. L. Ford, P. E. Harris, S. J. Miller, and D. Nelson, "Legal decomposition arising from non-positive linear recurrences", *Fibonacci Quart.* **54**:4 (2016), 348–365. MR

[Catral et al. 2017] M. Catral, P. L. Ford, P. E. Harris, S. J. Miller, D. Nelson, Z. Pan, and H. Xu, "New behavior in legal decompositions arising from non-positive linear recurrences", *Fibonacci Quart.* **55**:3 (2017), 252–275. MR

[Demontigny et al. 2014a] P. Demontigny, T. Do, A. Kulkarni, S. J. Miller, D. Moon, and U. Varma, "Generalizing Zeckendorf's Theorem to $f$-decompositions", *J. Number Theory* **141** (2014), 136–158. MR Zbl

[Demontigny et al. 2014b] P. Demontigny, T. Do, A. Kulkarni, S. J. Miller, and U. Varma, "A generalization of Fibonacci far-difference representations and Gaussian behavior", *Fibonacci Quart.* **52**:3 (2014), 247–273. MR Zbl

[Dorward et al. 2017a] R. Dorward, P. L. Ford, E. Fourakis, P. E. Harris, S. J. Miller, E. Palsson, and H. Paugh, "A generalization of Zeckendorf's theorem via circumscribed $m$-gons", *Involve* **10**:1 (2017), 125–150. MR Zbl

[Dorward et al. 2017b] R. Dorward, P. L. Ford, E. Fourakis, P. E. Harris, S. J. Miller, E. A. Palsson, and H. Paugh, "Individual gap measures from generalized Zeckendorf decompositions", *Unif. Distrib. Theory* **12**:1 (2017), 27–36. MR

[Kologlu et al. 2011] M. Kologlu, G. S. Kopp, S. J. Miller, and Y. Wang, "On the number of summands in Zeckendorf decompositions", *Fibonacci Quart.* **49**:2 (2011), 116–130. MR Zbl

[Lekkerkerker 1952] C. G. Lekkerkerker, "Voorstelling van natuurlijke getallen door een som van getallen van Fibonacci", *Simon Stevin* **29** (1952), 190–195. MR Zbl

[Lengyel 2006] T. Lengyel, "A counting based proof of the generalized Zeckendorf's theorem", *Fibboniacci Quart.* **44**:4 (2006), 324–325. MR Zbl

[Miller and Wang 2014] S. J. Miller and Y. Wang, "Gaussian behavior in generalized Zeckendorf decompositions", pp. 159–173 in *Combinatorial and additive number theory: CANT 2011 and 2012*, edited by M. B. Nathanson, Springer Proc. Math. Stat. **101**, Springer, 2014. MR Zbl

[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 Zbl

Received: 2018-04-18 Revised: 2018-07-10 Accepted: 2018-07-22

dgotshall16@saintpeters.edu *Department of Mathematics, Saint Peter's University, Jersey City, NJ, United States*

peh2@williams.edu *Department of Mathematics and Statistics, Williams College, Williamstown, MA, United States*

dnelson1@saintpeters.edu *Department of Mathematics, Saint Peter's University, Jersey City, NJ, United States*

maria.vega@usma.edu *Department of Mathematical Sciences, United States Military Academy, West Point, NY, United States*

cdv1218@gmail.com *Department of Mathematical Sciences, United States Military Academy, West Point, NY, United States*

INVOLVE 12:3 (2019)
dx.doi.org/10.2140/involve.2019.12.521

# Rigidity of Ulam sets and sequences

Joshua Hinman, Borys Kuca,
Alexander Schlesinger and Arseniy Sheydvasser

(Communicated by Kenneth S. Berenhaut)

We give a number of results about families of Ulam sequences and sets, further exploring recent work on rigidity phenomena. For Ulam sequences, using elementary methods we give an upper bound on the density and prove regularity for various families of sequences. For Ulam sets, we consider extensions of classification work done by Kravitz and Steinerberger.

## 1. Introduction and main results

***Introduction.*** Let $U(a, b)$ be the integer sequence that starts with two integers $0 < a < b$ and each subsequent term is the smallest integer that can be written as the sum of two distinct prior terms in exactly one way. Such sequences are known as Ulam sequences, in honor of Stanisław Ulam [1964], who first introduced the sequence $U(1, 2)$.

Considering the simplicity of the definition, surprisingly little is known about Ulam sequences, despite recent resurgence in interest — see [Gibbs 2015; Gibbs and McCranie 2017; Steinerberger 2017; Kravitz and Steinerberger 2017; Kuca 2018]. However, recent numerical evidence suggests that families of Ulam sequences have unexpected rigidity phenomena. In particular, in [Hinman et al. 2018], the authors make the following conjecture.

**Conjecture 1.1.** *There exist integer coefficients $m_i$, $p_i$, $k_i$, $r_i$ such that for all integers $n \geq 4$,*

$$U(1, n) = \bigsqcup_{i=1}^{\infty} [m_i n + p_i, k_i n + r_i].$$

*MSC2010:* primary 11B83, 11P70; secondary 05A16.
*Keywords:* Ulam sequence, additive number theory.

While this conjecture is at present open, the authors did prove that it holds for all terms up to $50{,}000n$ — that is, for all $n \geq 4$,

$$\begin{aligned} U(1,n) \cap [1, 50000n] = \{1\} \cup [n, 2n] \cup \{2n+2\} \cup \{4n\} \cup [4n+2, 5n-1] \\ \cup \{5n+1\} \cup [7n+3, 8n+1] \cup \{10n+2\} \cup \{11n+2\} \\ \cup \cdots \cup \{49991n+6949\} \cup \{49993n+6950\}. \end{aligned}$$

This suggests that while individual Ulam sequences may be difficult to deal with, we may be able to get substantially better results about families of sequences. In the present paper, we investigate various results related to the rigidity conjecture above.

***Summary of main results.*** We begin by revisiting the setting of Conjecture 1.1. By considering long runs of consecutive terms in an Ulam sequence $U(1, n)$, we prove the following elementary result.

**Theorem 1.2.** *Let $m_i$, $p_i$, $k_i$, $r_i$ be integer coefficients as in Conjecture 1.1. Then for all $i$, we have $k_i - m_i = 0$ or $1$, and $r_i \leq p_i$.*

Given a set of integers $K$, recall that its asymptotic density is defined as the constant

$$\delta(K) = \lim_{N \to \infty} \frac{\#(K \cap [1, N])}{N},$$

assuming that it exists. Using similar methodology as for Theorem 1.2, we also establish an upper bound on the asymptotic density for sequences $U(1, n)$.

**Theorem 1.3.** *The density of $U(1, n)$ is bounded above by $(n + 1)/(3n)$.*

It should be noted that this is likely not a tight upper bound — asymptotically,

$$\frac{n+1}{3n} \approx \frac{1}{3},$$

but numerical data for $n \geq 4$ suggests that the actual density is $\approx \frac{1}{6}$. Furthermore, while our methods provide an upper bound, they do not provide any lower bound on the density — unfortunately, this is not surprising, as no positive lower bound on the density of sequences $U(1, n)$ is known at this time.

In Section 3, we turn to a question first studied by Queneau[1] [1972]: when is the Ulam sequence regular — that is, when is the sequence of differences between consecutive terms periodic? It was proved by Finch [1991; 1992a; 1992b] that if an Ulam sequence contains finitely many even terms, then it is regular. It is

[1]Raymond Queneau is better known for his work as a French poet and novelist, but he was interested in the role of mathematics in literature, which led him to cofound the Oulipo in 1960 [Motte 1998], together with chemical engineer and mathematician François Le Lionnais.

conjectured that $U(a, b)$ with $a < b$ coprime contains finitely many even terms if and only if

(1) $a = 2,\ b \geq 5$,

(2) $a = 4$,

(3) $a = 5,\ b = 6$, or

(4) $a \geq 6$ and $a$ or $b$ is even.

Schmerl and Spiegel [1994] proved the $a = 2,\ b \geq 5$ case; Cassaigne and Finch [1995] proved the case where $a = 4,\ b \equiv 1 \bmod 4$. It is worthwhile to note that the proofs of these results use a limited form of rigidity similar to Conjecture 1.1; furthermore, if some generalization of that conjecture holds for sequences $U(a, b)$ with $a \neq 1$, this would seem to give a means of proving that certain families of Ulam sequences are all regular — if you can show that some Ulam sequence $U(a, b)$ with $b$ sufficiently large has only finitely many even terms, then this will be true of all subsequent Ulam sequences in that family. We prove a far more modest, but nevertheless interesting result that gives a semi-algorithm for determining whether an Ulam sequence is regular — unfortunately, it is only a semi-algorithm, as it is not ever guaranteed to halt. Using this, we were able to establish the following.

**Theorem 1.4.** *For integer pairs* $(a, b)$ *given below, the sequence of differences between consecutive terms of* $U(a, b)$ *is eventually periodic*:

$$\begin{array}{ccccccc}
(4,11), & (4,19), & (6,7), & (6,11), & (7,8), & (7,10), & (7,12),\\
(7,16), & (7,18), & (7,20), & (8,9), & (8,11), & (9,10), & (9,14),\\
(9,16), & (9,20), & (10,11), & (10,13), & (10,17), & (11,12), & (11,14),\\
(11,16), & (11,18), & (11,20), & (12,13), & (12,17), & (13,14). &
\end{array}$$

In another direction, we also consider "Ulam-like" behavior and rigidity in higher dimensions. Using the terminology of [Kravitz and Steinerberger 2017], we define Ulam sets as follows.

**Definition 1.5.** Let $|\cdot|$ be a norm on $\mathbb{Z}^n$ that increases monotonically in each coordinate. A $(k, n)$-*Ulam set* $U(v_1, v_2, \ldots, v_k)$ is a recursively defined set that contains $v_1, v_2, \ldots, v_n \in \mathbb{Z}^n_{\geq 0}$ and each subsequent vector is the vector of smallest norm that can be written as a sum of two distinct vectors in the set in exactly one way. We shall say $U(v_1, v_2, \ldots, v_k)$ is *nondegenerate* if $v_i \notin U(v_1, v_2, \ldots, v_{i-1}, v_{i+1}, \ldots, v_k)$ for every $1 \leq i \leq k$.

Two remarks are necessary here: first, it may appear that the definition of Ulam set depends on the choice of monotonically increasing norm $|\cdot|$. In fact, this is not so, as proved in [Kravitz and Steinerberger 2017]. Secondly, it may be unclear which vector is added if there is more than one of equal norm. However, by the above, this is irrelevant.

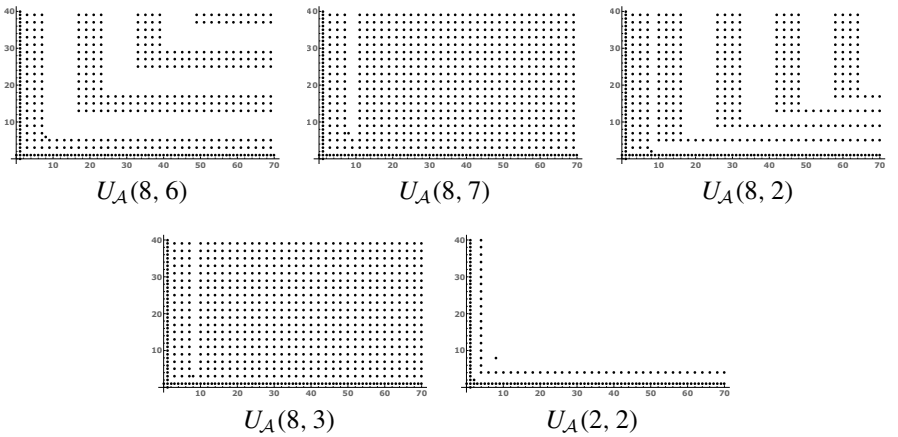


**Figure 1.** From left to right and top to bottom: sets $U_{\mathcal{A}}(v_1, v_2)$ of $L$, column-deleted, column-deleted $L$, shifted column-deleted, and exceptional type.

Contingent on some natural restrictions described in Section 4, we classify all $(3, 2)$-Ulam sets, showing that they necessarily belong to one of a finite number of different types, illustrated in Figure 1.

**Theorem 1.6.** *Let $\mathcal{U} = U((1, 0), (0, 1), (v_1, v_2))$ be a nondegenerate $(3, 2)$-Ulam set such that $v_1, v_2 \neq 0$. Then exactly one of the following is true of either $\mathcal{U}$ or its reflection about the line $y = x$:*

(1) $v_1, v_2 \in 2\mathbb{Z} \cap [4, \infty)$ *and $\mathcal{U}$ is of L type.*

(2) $v_1 \in 2\mathbb{Z}$, $v_2 \in (1 + 2\mathbb{Z}) \cap [4, \infty)$, *and $\mathcal{U}$ is of column-deleted type.*

(3) $v_1 \in 2\mathbb{Z} \cap [4, \infty)$, $v_2 = 2$, *and $\mathcal{U}$ is of column-deleted L type.*

(4) $v_1 \in 2\mathbb{Z}$, $v_2 = 3$, *and $\mathcal{U}$ is of shifted column-deleted type.*

(5) $v_1 = v_2 = 2$ *and $\mathcal{U}$ is of exceptional type.*

See Section 4 for definitions of the various types of Ulam sets. Note that this is in a sense a higher-dimensional version of rigidity — we are varying the Ulam sets in some parameter, and outside of some odd exceptional cases when the norm of the vector is small, the parity of the coordinates of the added vector wholly determine the structure of the set. We also show that there are restrictions for more general $(k, 2)$-Ulam sets — in particular, in Section 5 we show that there is always a parity restriction.

**Theorem 1.7.** *Let $\mathcal{U} = U((1, 0), (0, 1), v_1, v_2, \ldots, v_n)$ be a nondegenerate $(n+2, 2)$-Ulam set such that none of the $v_i$ lie on the coordinate axes. Then there exists $(w_1, w_2) \in \mathbb{Z}^2_{\geq 0}$ such that for all $(m, n) \in \mathcal{U}$, if $m \geq w_1$, $n \geq w_2$, then $m = w_1 \mod 2$, $n = w_2 \mod 2$.*

Finally, in Section 6, we demonstrate that if the added vectors are not too small, then the corresponding Ulam set must be periodic in following sense.

**Definition 1.8.** A $(k, 2)$-Ulam set $\mathcal{U}$ is *eventually* $(m, n)$*-periodic* if there exists $(m_0, n_0)$ such that for all $(a, b) \in \mathbb{Z}^2_{\geq 0}$ with $a \geq m_0$ and $b \geq n_0$ we have $(a, b) \in \mathcal{U}$ if and only if $(a+m, b+n) \in \mathcal{U}$. We call $(m, n)$ a *period* of $\mathcal{U}$.

**Theorem 1.9.** *All* $(4, 2)$*-Ulam sets* $\mathcal{U} = U_{\mathcal{A}}(v_1, v_2)$ *with* $v_i = (x_i, y_i)$ *such that* $x_i, y_i \geq 4$ *are eventually periodic.*

## 2. Consecutive terms in sequences $U(1, n)$

Our main goal for this section is to find bounds on the runs of consecutive terms in the Ulam sequences $U(1, n)$. As an example, we prove the following theorem.

**Theorem 2.1.** *Let* $n \geq 2$ *and let* $I$ *be a set of* $3n$ *consecutive positive integers greater than* $2n+2$*. Then* $|I \cap U(1, n)| \leq n+1$.

As an immediate corollary, this implies a bound on the density of $U(1, n)$.

**Corollary 2.2.**
$$\delta(U(1, n)) \leq \frac{n+1}{3n}.$$

*Proof.* Partition the first $k$ integers greater than $2n+2$ into runs of $3n$ consecutive integers. Each such partition contains at most $n+1$ terms of $U(1, n)$. The proportion of Ulam numbers less than or equal to $k$ is then no bigger than

$$\frac{(n+1)(k/(3n)+1)+2n+2}{k} = \frac{n+1}{3n}\left(1+\frac{1}{k}\right)+\frac{2n+2}{k}.$$

In the limit, we get the desired upper bound. □

We will give an improvement on this upper bound for the special case $U(1, 2)$ at the end of this section. Before we prove Theorem 2.1, we give a few useful lemmas, some of which are very useful in their own right.

**Lemma 2.3.** *Let* $n \geq 2$*. The first three intervals of* $U(1, n)$ *are* $\{1\}$, $[n, 2n]$, *and* $\{2n+2\}$.

*Proof.* Clearly, all elements of the form $n+i$ for $1 \leq i \leq n$ have the unique Ulam representation $n+i = (n+i-1)+1$. However, $2n+1 \notin U(1, n)$, because it has a second Ulam representation $n+(n+1)$. Finally, $2n+2 = n+(n+2)$, which is its only Ulam representation, and $2n+3 \notin U(1, n)$ since $2n+3 = (2n+2)+1 = n+(n+3)$. □

**Lemma 2.4.** *If* $a, a+k \in U(1, n)$ *for some* $1 \leq k \leq n$, *then* $[a+k+n, a+2n] \subset \mathbb{Z} \setminus U(1, n)$.

*Proof.* Every integer in this interval is of the form $a+k+n+i$ for $0 \le i \le n-k$; hence it has at least two Ulam representations: $(a+k)+(n+i)$ and $a+(n+k+i)$, where we have used the fact that $n+i, n+k+i \in [n, 2n]$, and hence are in the Ulam sequence by Lemma 2.3. □

As an immediate corollary of this lemma, we get a proof of Theorem 1.2.

*Proof of Theorem 1.2.* Suppose that $m_i n + p_i, m_i n + p_i + 1 \in U(1, n)$. Then $(m_i+1)n + (p_i+1) \notin U(1, n)$. Therefore, $k_i - m_i = 0$ or 1 and $r_i \le p_i$. □

**Lemma 2.5.** *Let* $1 \le k \le n$. *If* $[a, a+k] \subset U(1, n)$, *then* $[a+n+1, a+k+2n-1] \subset \mathbb{Z} \setminus U(1, n)$.

*Proof.* We have the partition

$$[a, a+k] = \bigcup_{i=0}^{k-1} [a+i, a+i+1],$$

and so it suffices to prove the claim with $k = 1$, which is an immediate corollary of Lemma 2.4. □

Lemma 2.5 shows that if there are long runs of consecutive elements in the Ulam sequence, then there must be a longer run of consecutive elements later on that do not belong to the Ulam sequence. With this observation in hand, we proceed to the proof of Theorem 2.1.

*Proof of Theorem 2.1.* If $I \cap U(1, n) = \varnothing$, we are done. Otherwise, let $a > 2n+2$ be the smallest element in $I \cap U(1, n)$. There are two cases: either $[a, a+n-1]$ contains at least two consecutive elements $u, u+1 \in U(1, n)$, or it does not. We consider these cases separately.

Case 1: Since we are given that $[a, a+n-1] \cap U(1, n)$ contains at least two consecutive elements, we can partition it into disjoint intervals

$$[a, a+n-1] \cap U(1, n) = \bigsqcup_{i=1}^{m} [a+k_i, a+l_i] = \bigsqcup_{j=1}^{t} \{a+c_j\}$$

such that $k_i \le l_i + 1 < k_{i+1}$, $c_j + 1 < c_{j+1}$, and for no $i$, $j$ is $c_j \in [k_i - 1, l_i + 1]$. By Lemma 2.5, $[a+n+k_i+1, a+l_i+2n-1] \subset \mathbb{Z} \setminus U(1, n)$ for $1 \le i \le m$. Note that since $k_m \le n-1$ and $l_1 \ge 1$, we have $a+n+k_m+1 \le a+l_1+2n-1$, and hence

$$\bigcup_{i=1}^{m} [a+n+k_i+1, a+l_i+2n-1] = [a+n+k_1+1, a+2n+l_m-1] \subset \mathbb{Z} \setminus U(1, n).$$

Therefore,

$$I \cap U(1, n) \subset ([a, a+n+k_1]) \cup ([a+2n+l_m, a+3n-1]).$$

However, we claim that

$$|[a+2n+l_m, a+3n-1]\cap U(1,n)|+|[a+l_m, a+n-1]\cap U(1,n)| \le n-l_m.$$

It suffices to prove this assuming that $[a+l_m, a+n-1]\cap U(1,n)\neq\varnothing$. Let $u_1, u_2, \ldots, u_s$ be the Ulam numbers in $[a+l_m, a+n-1]$. If $s=1$, then

$$a+2n+l_m = (a+l_m)+2n = u_1+(2n-(u_1-a-l_m)),$$

and as this gives two representations, it must be that $a+2n+l_m\notin U(1,n)$. If $s>1$, then for every $1\le i<j\le s$, by Lemma 2.4,

$$[u_j+n, u_i+2n]\subset\mathbb{Z}\setminus U(1,n).$$

Hence

$$[a+2n+l_m, u_{s-1}+2n]\subset\mathbb{Z}\setminus U(1,n).$$

Note that

$$|[a+2n+l_m, u_{s-1}+2n]|\ge s$$

unless $u_{s-1}=a+l_m+s-1$, which is to say that $[a+l_m, a+l_m+s-1]\subset U(1,n)$. But by the definition of $l_m$, it can only be that $a+l_m\in U(1,n)$ if $l_m=n-1$, which is not possible since we assumed that there are at least two Ulam numbers in $[a+l_m, a+n-1]$. As desired, we conclude that

$$|[a+l_m, a+n-1]\cap U(1,n)|+|[a+2n+l_m, a+3n-1]\cap U(1,n)|\le n-l_m,$$

and therefore

$$\begin{aligned}|I\cap U(1,n)| &\le |[a+l_m, a+n-1]\cap U(1,n)|+|[a+n, a+n+k_1]\cap U(1,n)|\\ &\qquad +|[a+2n+l_m, a+3n-1]\cap U(1,n)|\\ &\le n-l_m+k_1-1\\ &\le n-1.\end{aligned}$$

Case 2: In this case, we are given that

$$[a, a+n-1]\cap U(1,n)=\bigsqcup_{j=1}^{t}\{a+c_j\},$$

where $c_j+1<c_{j+1}$. This implies that for $k>j$,

$$k-j<c_k-c_j<n.$$

By Lemma 2.4, we have

$$[a+c_k+n, a+c_j+2n]\subset\mathbb{Z}\setminus U(1,n),$$

and consequently,

$$[a+c_2+n, a+c_{t-1}+2n]=\bigcup_{1\le i<j\le t}[a+c_k+n, a+c_j+2n]\subset\mathbb{Z}\setminus U(1,n).$$

Ergo,

$$\begin{aligned}|I \cap U(1,n)| &= |[a, a+n-1] \cap U(1,n)| \\ &\quad + |[a+n, a+c_2+n-1] \cap U(1,n)| \\ &\quad\quad + |[a+c_2+n, a+c_{t-1}+2n] \cap U(1,n)| \\ &\quad\quad\quad + |[a+c_{t-1}+2n+1, a+3n-1] \cap U(1,n)| \\ &\leq t + c_2 + n - c_{t-1} - 1 \\ &\leq n+1. \end{aligned}$$

□

For $n = 2$, Corollary 2.2 gives an upper bound of $\frac{1}{2}$ on the density. Using similar techniques to the proof of Theorem 2.1, we can improve this upper bound to $\frac{6}{17} \approx 0.353$.

**Theorem 2.6.** *The density of* $U(1,2)$ *is at most* $\frac{6}{17}$.

*Proof.* Let $a \in U(1,n)$ and define $I = [a, a+8]$, $J = [a, a+16]$. We claim that either $|I \cap U(1,2)| \leq 3$ or $|J \cap U(1,2)| \leq 6$. We make use of the fact that

$$1, 2, 3, 4, 6, 8, 11, 13, 16 \in U(1,2).$$

If $|I \cap U(1,2)| > 3$, then $I = \{a, a+2, a+5, a+7\}$. Otherwise, $I \cap U(1,2)$ contains a pair of elements $u, u+1$ such that $u+1 = a+2, a+3, a+4, a+6$, or $a+8$, which gives two representations; this is a contradiction.

In this case, $J \cap U(1,2) \subset \{a, a+2, a+5, a+7, a+12, a+14\}$ — otherwise, it contains an element with two representations. Consequently, $|J \cap U(1,2)| \leq 6$. This means we can now define two sequences $u_1, u_2, u_3, \ldots$, $L_1, L_2, L_3, \ldots$ recursively — let $u_1 = 1$ and $L_1 = 17$, and then define $u_{i+1}$ to be the smallest element of the Ulam sequence larger than $u_i + L_i$, and

$$L_{i+1} = \begin{cases} 17 & \text{if } |[u_{i+1}, u_{i+1}, +16] \cap U(1,2)| \leq 6, \\ 9 & \text{otherwise.} \end{cases}$$

We can then partition the positive integers into sets of the forms $[u_{i+1}, u_{i+1} + L_i]$ and $[u_{i+1} + L_i + 1, u_{i+2} - 1]$. The density of $U(1,2)$ in any of these sets is no more than $\frac{6}{17}$, and that implies that the density of $U(1,2)$ is bounded by $\frac{6}{17}$. □

## 3. Regular Ulam sequences

We now consider regular sequences. Let $1_{U(a,b)}$ be the indicator function of $U(a,b)$. Given a positive integer $l$ and a positive odd number $k$, define

$$S^l_{a,b}(k) = \{1_{U(a,b)}(k+2i)\}_{i=0}^{l-2}.$$

With this terminology, we can now easily state the main theorem we want to prove.

**Theorem 3.1.** *Let $0 < a < b$ be coprime integers. Let $l, p, q$ be positive integers such that $p < q$, $p, q$ are odd, $q \geq 2l$, $a < b < 2l - 2$, $S^l_{a,b}(p) = S^l_{a,b}(q)$, and*

$$U(a, b) \cap 2\mathbb{Z} \cap [2l, 3q - p] = \varnothing.$$

*Then*

$$U(a, b) \cap 2\mathbb{Z} \cap [2l, \infty) = \varnothing.$$

Theorem 3.1 provides a semi-algorithm for determining whether a sequence is regular — simply do a brute force search for triples $(l, p, q)$ satisfying the conditions of the theorem. If such a triple is found, then we conclude that $U(a, b)$ is regular. This gives us the following corollary.

**Corollary 3.2.** *For integer pairs $(a, b)$ given below, $U(a, b)$ is regular*:

| | | | | | | |
|---|---|---|---|---|---|---|
| (4, 11), | (4, 19), | (6, 7), | (6, 11), | (7, 8), | (7, 10), | (7, 12), |
| (7, 16), | (7, 18), | (7, 20), | (8, 9), | (8, 11), | (9, 10), | (9, 14), |
| (9, 16), | (9, 20), | (10, 11), | (10, 13), | (10, 17), | (11, 12), | (11, 14), |
| (11, 16), | (11, 18), | (11, 20), | (12, 13), | (12, 17), | (13, 14). | |

*Proof.* By direct computation, we find triples $(l, p, q)$ satisfying the conditions of Theorem 3.1:

| $(a, b)$ | $(l, p, q)$ |
|---|---|
| (4, 11) | (25, 107, 1425) |
| (4, 19) | (41, 14745, 17305) |
| (6, 7) | (57, 8537, 70987) |
| (6, 11) | (89, 1032425, 1033833) |
| (7, 8) | (71, 14331, 57089) |

| $(a, b)$ | $(l, p, q)$ |
|---|---|
| (7, 10) | (85, 95587, 102181) |
| (7, 12) | (99, 79423, 80991) |
| (7, 16) | (127, 46957, 47965) |
| (7, 18) | (141, 196513, 198753) |
| (7, 20) | (155, 50893, 52125) |

| $(a, b)$ | $(l, p, q)$ |
|---|---|
| (8, 9) | (91, 1037093, 1038533) |
| (8, 11) | (111, 2125501, 4308725) |
| (9, 10) | (109, 117117, 747935) |
| (9, 14) | (145, 558073, 560377) |
| (9, 16) | (163, 60093, 65277) |
| (9, 20) | (199, 219761, 222929) |
| (10, 11) | (133, 470303, 485615) |
| (10, 13) | (157, 5804601, 5807097) |
| (10, 17) | (205, 3919981, 3933037) |

| $(a, b)$ | $(l, p, q)$ |
|---|---|
| (11, 12) | (155, 140511, 142975) |
| (11, 14) | (177, 507965, 509373) |
| (11, 16) | (199, 394379, 400715) |
| (11, 18) | (221, 29995, 37035) |
| (11, 20) | (243, 46291, 54035) |
| (12, 13) | (183, 3329465, 3330921) |
| (12, 17) | (239, 3204117, 3211733) |
| (13, 14) | (209, 1421023, 1427679) |

□

To prove Theorem 3.1, we start with a useful lemma that establishes that if it is false, then there is a bijective correspondence between odd Ulam numbers in different intervals.

**Lemma 3.3.** *Let $l, a, b$ be positive integers, and $p < q$ be positive odd integers such that $q \geq 2l$, $a < b < 2l - 2$, $S^l_{a,b}(p) = S^l_{a,b}(q)$,*

$$U(a,b) \cap 2\mathbb{Z} \cap [2l, 3q - p] = \varnothing,$$
$$U(a,b) \cap 2\mathbb{Z} \cap [2l, \infty) \neq \varnothing.$$

*Let $\tilde{u}$ be the smallest even number in $U(a,b)$ greater than $3q - p$. Then there is a well-defined bijection*

$$U(a,b) \cap (1 + 2\mathbb{Z}) \cap [p, \tilde{u} + p - q - 1] \to U(a,b) \cap (1 + 2\mathbb{Z}) \cap [q, \tilde{u} - 1],$$
$$u \mapsto u + q - p.$$

*Proof.* We will show that there is a well-defined bijection

$$\phi_m : U(a,b) \cap (1 + 2\mathbb{Z}) \cap [p, p + 2m] \to U(a,b) \cap (1 + 2\mathbb{Z}) \cap [q, q + 2m],$$
$$u \mapsto u + q - p,$$

for all integers $0 \leq m \leq \frac{1}{2}(\tilde{u} - q - 1)$. We know that $S^l_{a,b}(p) = S^l_{a,b}(q)$; hence $p + 2m' \in U(a,b)$ if and only if $q + 2m' \in U(a,b)$ for all $0 \leq m' \leq l - 2$, which proves the claim for $m \leq l - 2$.

For all other $m$, we apply induction — that is, let $l - 2 < h \leq \frac{1}{2}(\tilde{u} - q - 1)$ such that $\phi_{h-1}$ is a bijection. We need to show that $\phi_h$ is bijection. This is equivalent to proving that $p + 2h \in U(a,b)$ if and only if $q + 2h \in U(a,b)$. Define sets

$$P = \{(u,v) \in U(a,b)^2 \mid u \equiv 0 \bmod 2,\ v \equiv 1 \bmod 2,\ u + v = p + 2h\},$$
$$Q = \{(u,v) \in U(a,b)^2 \mid u \equiv 0 \bmod 2,\ v \equiv 1 \bmod 2,\ u + v = q + 2h\},$$

which enumerate the number of representations of $p + 2h$ and $q + 2h$, respectively. If we can show that $|P| = |Q|$, then this will imply that $p + 2h \in U(a,b)$ if and only if $q + 2h \in U(a,b)$. However, we can construct a bijection between these two sets by

$$\psi : P \to Q,$$
$$(u,v) \mapsto (u, \phi_{h-1}(v)) = (u, v + q - p).$$

This is well-defined since $u + v = p + 2h$ implies $v \leq p + 2h - 1$. □

*Proof of Theorem 3.1.* We argue by contradiction. That is, suppose that there exist even Ulam numbers larger than $3q - p$. Let $\tilde{u}$ be the smallest such element. We know $\tilde{u} = u_1 + u_2$ for some $u_1 < u_2 \in U(a,b)$. Every even Ulam number less than $\tilde{u}$ is smaller than $2l$; hence one of $u_1, u_2$ is odd — otherwise, we have

$$u_1 + u_2 < 4l \leq 3q - p,$$

which is a contradiction. Since $\tilde{u}$ is even, we conclude that $u_1, u_2$ are both odd. Next, we show that $\tilde{u} - q + p$ has at least two representations as the sum of two

distinct elements of $U(a, b)$. Note that

$$\tilde{u} - q + p \geq (3q - p) - q + p = 2q > 2l,$$

and since $\tilde{u} - q + p$ is even, this implies it is not in $U(a, b)$. Consequently, it will suffice to prove that it has at least one representation. Note that

$$u_2 > \tfrac{1}{2}\tilde{u} > \tfrac{1}{2}(3q - p) > q,$$
$$u_2 \leq \tilde{u} - 1,$$

so by Lemma 3.3, since $u_2 \in U(a, b)$ it follows $u_2 + q - p \in U(a, b)$. Therefore, $\tilde{u} + q - p = u_1 + (u_2 + q - p)$ is a representation.

Write

$$\tilde{u} - q + p = v_1 + v_2 = v_1' + v_2',$$

where $v_1 < v_2$, $v_1' < v_2' \in U(a, b)$. Note that $v_2 > q$, since

$$v_2 > \tfrac{1}{2}(\tilde{u} - q + p) > \tfrac{1}{2}((3q - p) - q + p) > q.$$

Similarly, $v_2' > q$. From this it follows that $v_2, v_2' > 2l$, and we conclude that $v_2, v_2'$ must be odd. Finally, note that

$$p < q < v_2, v_2' \leq \tilde{u} + p - q - 1,$$

and therefore by Lemma 3.3, $v_2 + q - p$, $v_2' + q - p \in U(a, b)$, which is a contradiction since

$$\tilde{u} = v_1 + (v_2 + q - p) = v_1' + (v_2' + q - p). \qquad \square$$

## 4. Classification of (3, 2)-Ulam sets

Up until this point, we have only considered (2, 1)-Ulam sets; we now turn to the problem of classifying higher-dimensional Ulam sets. The classification problem for nondegenerate (2, 2)-Ulam sets was solved by Kravitz and Steinerberger [2017]. In particular, they showed that after a linear transformation, the Ulam set becomes $U((1, 0), (0, 1))$, illustrated in Figure 2. We shall denote this set by $\mathcal{A}$.

We shall consider (3, 2)-Ulam sets that are extensions of such Ulam sets — that is, we shall assume that two of the basis vectors are (1, 0) and (0, 1). For convenience, we define

$$U_{\mathcal{A}}(v_1, v_2) = U((1, 0), (0, 1), (v_1, v_2)),$$
$$W_{(v_1, v_2)} = \{(m, n) \in \mathbb{Z}^2_{\geq 0} \mid m < v_1 \text{ or } n < v_2\},$$
$$L_{(v_1, v_2)} = \{(m, n) \in \mathbb{Z}_{\geq v_1} \times \mathbb{Z}_{\geq v_2}\}.$$

Note that if $(a, b) \in L_{(v_1, v_2)}$, then any representations it has have to lie in the set $W_{(v_1, v_2)}$. We use this fact to our advantage to prove the following lemma.

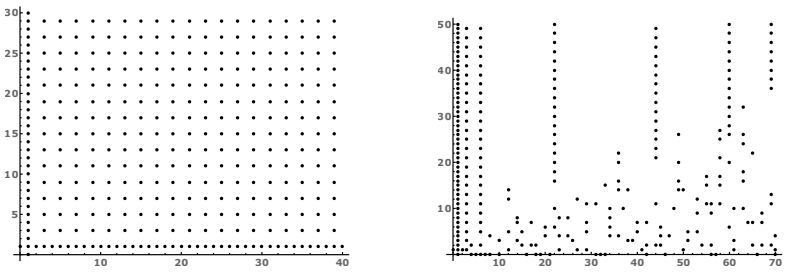


**Figure 2.** The (2, 2)-Ulam set $\mathcal{A}$ and the (3, 2)-Ulam set $U_{\mathcal{A}}(4, 0)$.

**Lemma 4.1.** *Let $\mathcal{U} = U_{\mathcal{A}}(v_1, v_2)$ be a nondegenerate* (3, 2)*-Ulam set with $v_1, v_2 \neq 0$. Then the following statements hold*:

(1) $v_1, v_2 > 1$ *and at least one of* $v_1, v_2$ *is even.*

(2) $\mathcal{A} \cap W_{(v_1,v_2)} = \mathcal{U} \cap W_{(v_1,v_2)}$.

(3) *Every point* $(m, n) \in \mathbb{Z}^2_{\geq 0}$ *has at least one representation.*

*Proof.* It was shown in [Kravitz and Steinerberger 2017] that

$$\mathcal{A} = \{(m, 1) \mid m \in \mathbb{Z}_{\geq 0}\} \cup \{(1, m) \mid m \in \mathbb{Z}_{\geq 0}\} \cup \{(2m + 1, 2n + 1) \mid m, n \in \mathbb{Z}_{\geq 0}\}.$$

For $\mathcal{U}$ to be nondegenerate, it must be that $(v_1, v_2) \notin \mathcal{A}$, and since $v_1, v_2 \neq 0$, this implies $v_1, v_2 > 1$ and at least one of $v_1, v_2$ is even.

All representations of points in $W_{(v_1,v_2)}$ are representations by elements in $\mathcal{U}$. It follows $\mathcal{A} \cap W_{(v_1,v_2)} = \mathcal{U} \cap W_{(v_1,v_2)}$. However, this implies

$$(m, n) = (m - 1, 1) + (1, n - 1)$$

is a representation of $(m, n)$. □

We shall call $(m, n) = (m - 1, 1) + (1, n - 1)$ the *standard representation* of $(m, n)$. By Lemma 4.1, proving that $(m, n) \notin U_{\mathcal{A}}(v_1, v_2)$ for $v_1, v_2 \neq 0$ is equivalent to proving that it has a nonstandard representation. This makes working with Ulam sets of this form much simpler. On the other hand, if one of $v_1, v_2$ is 0, then the set $U_{\mathcal{A}}(v_1, v_2)$ has a copy of a (2, 1)-Ulam set on either the $x$- or $y$-axis. An example of such a set is given in Figure 2. Some partial results about such sets are given in [Kravitz and Steinerberger 2017], but in general describing their structure is an open problem.

We now give five examples of possible structures of sets $U_{\mathcal{A}}(v_1, v_2)$ with $v_1, v_2 \neq 0$, which are derived from numerical observations. An illustration of each of these five types is provided in Figure 1.

**Definition 4.2.** Let $U \subset \mathbb{Z}^2_{\geq 0}$ and let $(v_1, v_2)$ be a vector in $U$. We say $U$ is of *L type for* $(v_1, v_2)$ if

$$U = \{(v_1, v_2)\} \cup \{(m, 1) \mid m \in \mathbb{Z}_{\geq 0}\} \cup \{(1, m) \mid m \in \mathbb{Z}_{\geq 0}\}$$
$$\cup \{(a+2mv_1, b+2mv_2) \mid a, b, m \geq 0,\ a, b \in 1+2\mathbb{Z},\ m \in \mathbb{Z},\ (a, b) \in W_{(v_1,v_2)}\}.$$

We say $U$ is of *column-deleted type for* $(v_1, v_2)$ if

$$U = \{(v_1, v_2)\} \cup \{(m, 1) \mid m \in \mathbb{Z}_{\geq 0}\} \cup \{(1, m) \mid m \in \mathbb{Z}_{\geq 0}\}$$
$$\cup \{(2m+1, 2n+1) \mid m, n \in \mathbb{Z}_{\geq 0},\ \text{if } 2m+1 = v_1+1 \text{ then } 2n+1 < v_2\}.$$

We say $U$ is of *column-deleted L type for* $(v_1, v_2)$ if

$$U = \{(v_1, v_2)\} \cup \{(m, 1) \mid m \in \mathbb{Z}_{\geq 0}\} \cup \{(1, m) \mid m \in \mathbb{Z}_{\geq 0}\}$$
$$\cup \{(a+(m+1)v_2+2, b+2m+5) \mid a, b, m \geq 0,\ a, b, m \in 2\mathbb{Z},\ a < m \text{ or } b = 0\}.$$

We say that $U$ is of *shifted column-deleted type for* $(v_1, v_2)$ if

$$U = \{(v_1, v_2)\} \cup \{(m, 1) \mid m \in \mathbb{Z}_{\geq 0}\} \cup \{(1, m) \mid m \in \mathbb{Z}_{\geq 0}\}$$
$$\cup \{(m, n) \mid m, n \geq 0,\ m < v_1,\ m, n \in 1+2\mathbb{Z}\}$$
$$\cup \{(m, n) \mid m, n \geq 0,\ m > v_1,\ m \in 2\mathbb{Z},\ n \in 1+2\mathbb{Z}\}.$$

We say $U$ is of *exceptional type* if

$$U = \{(v_1, v_2)\} \cup \{(8, 8)\} \cup \{(m, 1) \mid m \in \mathbb{Z}_{\geq 0}\} \cup \{(1, m) \mid m \in \mathbb{Z}_{\geq 0}\}$$
$$\cup \{(4, 2m+4) \mid m \in \mathbb{Z}_{\geq 0}\} \cup \{(2m+4, 4) \mid m \in \mathbb{Z}_{\geq 0}\}.$$

This list enumerates all the possibilities for sets $U_{\mathcal{A}}(v_1, v_2)$ if $v_1, v_2 \neq 0$.

**Theorem 4.3.** *Let $\mathcal{U} = U_{\mathcal{A}}(v_1, v_2)$ be a nondegenerate $(3, 2)$-Ulam set such that $v_1, v_2 \neq 0$. Then exactly one of the following is true of either $\mathcal{U}$ or its reflection about the line $y = x$:*

(1) $v_1, v_2 \in 2\mathbb{Z} \cap [4, \infty)$ *and $\mathcal{U}$ is of L type.*

(2) $v_1 \in 2\mathbb{Z},\ v_2 \in (1+2\mathbb{Z}) \cap [4, \infty)$, *and $\mathcal{U}$ is of column-deleted type.*

(3) $v_1 \in 2\mathbb{Z} \cap [4, \infty),\ v_2 = 2$, *and $\mathcal{U}$ is of column-deleted L type.*

(4) $v_1 \in 2\mathbb{Z},\ v_2 = 3$, *and $\mathcal{U}$ is of shifted column-deleted type.*

(5) $v_1 = v_2 = 2$ *and $\mathcal{U}$ is of exceptional type.*

*Proof.* By Lemma 4.1, the given list enumerates all possibilities for $v_1, v_2$, after accounting for a possible reflection around the line $y = x$. Furthermore, it is easy to check that $\mathcal{U} \cap W_{(v_1,v_2)}$ is of the specified type in each case — that is, it is equal to the intersection of a set $U$ of the desired type with $W_{(v_1,v_2)}$.

Consider the case $v_1, v_2 \in 2\mathbb{Z} \cap [4, \infty)$. We shall show that $\mathcal{U} \cap W_{(a,b)}$ is of $L$ type for all $a, b \geq 0$. Note that by Lemma 4.1,

$$\begin{aligned}\mathcal{A} \cap W_{(3,3)} &= \{(m, 1) \mid m \in \mathbb{Z}_{\geq 0}\} \cup \{(1, m) \mid m \in \mathbb{Z}_{\geq 0}\} \\ &\qquad \cup \{(3, 2m+1) \mid m \in \mathbb{Z}_{\geq 0}\} \cup \{(2m+1, 3) \mid m \in \mathbb{Z}_{\geq 0}\} \\ &= \mathcal{U} \cap W_{(3,3)}.\end{aligned}$$

It follows that if $(m, n) \in \mathcal{U}$ and $m, n > 1$, then $m, n \in 1 + 2\mathbb{Z}$. This is evident if $(m, n) \in W_{(3,3)}$ — otherwise, either $(m, n) = (k+3, 2l+2)$ or $(2l+2, k+3)$ for some $k, l \in \mathbb{Z}_{\geq 0}$, and we have nonstandard representations

$$\begin{aligned}(k+3, 2l+2) &= (3, 2l+1) + (k, 1), \\ (2l+2, k+3) &= (2l+1, 3) + (1, k).\end{aligned}$$

Furthermore, it must be that $\mathcal{U} \cap W_{(2v_1,2v_2)}$ is of $L$ type. To see this, it suffices to show that

$$\mathcal{U} \cap W_{(2v_1,2v_2)} \cap L_{(v_1,v_2)} = \{(v_1, v_2)\},$$

but as we know any point in this intersection must necessarily be of the form $(2m+1, 2n+1)$, we have a nonstandard representation

$$(2m+1, 2n+1) = (v_1, v_2) + (2m+1-v_1, 2n+1-v_2).$$

We now prove that $\mathcal{U} \cap W_{(2kv_1,2kv_2)}$ is of $L$ type by inducting on $k \in \mathbb{Z}$ — we have proved the base case $k = 1$, so it suffices to assume $\mathcal{U} \cap W_{(2mv_1,2mv_2)}$ is $L$ type for some $m \in \mathbb{Z}_{\geq 0}$ and prove that $\mathcal{U} \cap W_{(2(m+1)v_1,2(m+1)v_2)}$ is $L$ type. This amounts to proving that

$$\begin{aligned}&\mathcal{U} \cap W_{((2m+1)v_1,(2m+1)v_2)} \cap L_{(2mv_1,2mv_2)} \\ &\quad = W_{((2m+1)v_1,(2m+1)v_2)} \cap L_{(2mv_1,2mv_2)} \cap (1 + 2\mathbb{Z}_{\geq 0})^2 \mathcal{U} \\ &\qquad\qquad \cap W_{((2m+2)v_1,(2m+2)v_2)} \cap L_{((2m+1)v_1,(2m+1)v_2)} \\ &\quad = \varnothing.\end{aligned}$$

This is easily proven by noting that the former set cannot possibly have any nonstandard representations, whereas the latter set is nothing more than

$$(v_1, v_2) + \mathcal{U} \cap W_{((2m+1)v_1,(2m+1)v_2)} \cap L_{(2mv_1,2mv_2)}.$$

The other cases are similar. □

## 5. Parity restrictions on $(k, 2)$-Ulam sets

Let us now consider the more general case where multiple vectors are added to a $(2, 2)$-Ulam set, rather than just one. As in the previous section, we consider

nondegenerate Ulam sets containing $(1, 0)$, $(0, 1)$, and so we define

$$U_{\mathcal{A}}(v_1, v_2, \ldots, v_n) = U((1, 0), (0, 1), v_1, \ldots, v_n).$$

We shall show that the parity of any element in $U_{\mathcal{A}}(v_1, v_2, \ldots, v_n)$ is eventually fixed, as long as none of the $v_i$ lie on the coordinate axes.

**Theorem 5.1.** *Let $\mathcal{U} = U_{\mathcal{A}}(v_1, v_2, \ldots, v_n)$ be a nondegenerate $(n+2, 2)$-Ulam set such that none of the $v_i$ lie on the coordinate axes. Then there exists a $v$ such that for all $u \in \mathcal{U} \cap L_v$, we have $u = v \bmod 2$.*

To prove Theorem 5.1, we first note that if $\mathcal{U}$ contains a point $(u_1, u_2)$ such that $(u_1, u_2 + 2k) \in \mathcal{U}$ for all $k \in \mathbb{Z}_{\geq 0}$, then for all $(u_1', u_2') \in \mathcal{U} \cap L_{(u_1, u_2)}$, we have $u_2 = u_2' \bmod 2$. This is because if $u_2' \neq u_2 \bmod 2$,

$$(u_1', u_2') = (u_1, u_2' - 1) + (u_1' - u_1, 1)$$

gives a nonstandard representation. It shall therefore suffice to prove the existence of such a point. Toward this end, we give a useful lemma.

**Lemma 5.2.** *Let $\mathcal{U} = U_{\mathcal{A}}(v_1, v_2, \ldots, v_n)$ be a nondegenerate $(n+2, 2)$-Ulam set such that none of the $v_i$ lie on the coordinate axes. If there exists $m \in \mathbb{Z}_{>1}$ such that there are infinitely many points of the form $(m, n) \in \mathcal{U}$, then there exists a point $(u_1, u_2)$ such that $(u_1, u_2 + 2k) \in \mathcal{U}$ for all $k \in \mathbb{Z}_{\geq 0}$.*

*Proof.* Let $M \in \mathbb{Z}_{>1}$ be the smallest $m$ such that there are infinitely many points of the form $(m, n) \in \mathcal{U}$. Note that in fact $M > 2$, since every element $(2, n)$ has at least two representations. Therefore, we can define $N$ be the largest $n$ such that $(m, n) \in \mathcal{U}$, where $1 < m < M$.

Consider any point $(M, n) \in \mathbb{Z}_{\geq 0}^2$ with $n > 2N$. For any representation of $(M, n)$, at least one of the summands must have $x$-coordinate 1 or $M$ — otherwise, the $y$-coordinates are too small to add up to $n$. If this representation is

$$(M, n) = (1, n') + (M - 1, n - n'),$$

then it is nonstandard if and only if $n - n' \neq 1$. However, if $n - n' \neq 1$, then every point $(M, n'')$ with $n'' > n$ has a nonstandard representation, which is impossible.

On the other hand, the only other possible representation is $(M, n) = (M, n-1) + (0, 1)$, so we conclude that $(M, n) \in \mathcal{U}$ if and only if $(M, n-1) \notin \mathcal{U}$. Thus if we take

$$(u_1, u_2) = \begin{cases} (M, n) & \text{if } (M, n) \in \mathcal{U}, \\ (M, n+1) & \text{otherwise,} \end{cases}$$

it satisfies the desired conditions. □

This is sufficient to prove Theorem 5.1.

*Proof of Theorem 5.1.* We claim that there must exist some $m \in \mathbb{Z}_{>1}$ such that there are infinitely many points of the form $(m, n) \in \mathcal{U}$. Suppose otherwise — then there must exist some strictly increasing function $\phi : \mathbb{Z}_{>1} \to \mathbb{Z}_{>1}$ such that if $(m, n) \in \mathcal{U}$ and $m, n > 1$, then $n < \phi(m)$.

Let $m > 2$ and $n > 2\phi(m)$. Then if

$$(m, n) = (m_1, n_1) + (m_2, n_2)$$

is a representation of $(m, n)$, it must be the standard representation — otherwise, $n_1 + n_2 < 2\phi(m) < n$. But this implies $(m, n) \in \mathcal{U}$, which is a contradiction.

Consequently, we can apply Lemma 5.2. By our earlier remarks, we know there exists a point $(u_1, u_2) \in \mathcal{U}$ such that for all $(u_1', u_2') \in \mathcal{U} \cap W_{(u_1,u_2)}$, we have $u_2' \equiv u_2 \bmod 2$.

On the other hand, the reflection of $\mathcal{U}$ about the line $y = x$ is also an Ulam set, which we shall denote by $\mathcal{V}$. It is easy to check that $\mathcal{V}$ also satisfies the requirements of the theorem, and therefore must contain a point $(v_1, v_2)$ such that for all $(v_1', v_2') \in \mathcal{V} \cap W_{(v_1,v_2)}$, we have $v_2' \equiv v_2 \bmod 2$. However, this means that if we take

$$v = (\max\{u_1, v_2\}, \max\{u_2, v_1\}),$$

then for all $u \in \mathcal{U} \cap L_v$, we have $u = v \bmod 2$, as desired. □

## 6. Periodicity of Ulam sets

We close by considering the periodicity of Ulam sets $U_{\mathcal{A}}(v_1, v_2, \ldots, v_n)$, under the additional constraint that the added vectors are not too small — that is, all of their components are at least 4. With this restriction, such sets become far more manageable.

**Lemma 6.1.** *Let $\mathcal{U} := U_{\mathcal{A}}(v_1, v_2, \ldots, v_n)$ be a $(k, 2)$-Ulam set such that all $v_i = (x_i, y_i)$ have $x_i, y_i \geq 4$. Then $\mathcal{U} \subset \mathcal{A} \cup \{v_1, v_2, \ldots, v_n\}$.*

*Proof.* Since all the initial vectors have components greater than or equal to 4, all elements of $\mathcal{A}$ with at least one coordinate less than 4 are also in $\mathcal{U}$. In particular, $\mathcal{U}$ contains all vectors of the forms $(2n - 1, 3)$ and $(3, 2n - 1)$ for $n \geq 1$. Thus for $n \geq 2$, we have $(2n, m) = (2n - 1, 3) + (1, m - 3)$ is a representation of $(2n, m)$ as a sum of vectors in the sets — as this representation is not the standard one, we conclude that $(2n, m) \notin \mathcal{U}$. By symmetry, we also have that $(m, 2n) \notin \mathcal{U}$ for $n \geq 2$. Thus, all vectors with at least one coordinate less than 4 in $\mathcal{U}$ are the vectors in $\mathcal{A}$ with at least one coordinate less than 4, and all other vectors are in $\mathcal{U}$ only if their coordinates are both odd; hence they are in $\mathcal{U}$. This proves the claim. □

In fact, we can be far more precise in our characterization of this set.

**Lemma 6.2.** *Let $\mathcal{U} = U_{\mathcal{A}}(v_1, v_2, \ldots, v_n)$ be a $(k, 2)$-Ulam set such that the vectors $v_i$ all have both components greater than or equal to 4. Let $a, b \geq 1$ be odd integers such that $(a, b) \neq v_i$ for any $i$. Then $(a, b) \in \mathcal{U}$ if and only if $(a, b) - v_i \notin \mathcal{U}$ for all $i$, $1 \leq i \leq n$.*

*Proof.* If $(a, b) - v_i = u \in \mathcal{U}$, then clearly $u + v_i$ is a second representation of $(a, b)$ outside of the standard one, and so $(a, b) \notin \mathcal{U}$. On the other hand, if $(a, b) \notin \mathcal{U}$, then we know there must be some nonstandard representation of it. We know that $(a, b) \in \mathcal{A}$; hence at least one term in this representation must come from $\mathcal{U} \backslash \mathcal{A}$. Since $\mathcal{U} \subset \mathcal{A} \cup \{v_1, v_2, \ldots, v_n\}$, that means that one of the summands must be $v_i$ for some $i$, which is to say that $(a, b) - v_i \in \mathcal{U}$, as desired. □

Note that if $n = 1$, Lemma 6.2 tells us precisely that $\mathcal{U}$ is eventually periodic, which is consistent with the result of Section 4. On the other hand, based on numerical evidence, it is almost certainly not true that all $(k, 2)$-Ulam sets are eventually periodic. However, we are interested in whether one can build new eventually periodic sets from existing eventually periodic sets. As an example, we know from the results of Section 4 that adding an initial vector whose coordinates are at least 4 to $\mathcal{A}$ yields another set that is eventually periodic. This leads us to conjecture that adding an initial vector whose coordinates are sufficiently large to an eventually periodic set yields an eventually periodic set. We prove two theorems in this direction.

**Theorem 6.3.** *Let $\mathcal{U} = U_{\mathcal{A}}(v_1, v_2, \ldots, v_n)$ be a nondegenerate $(k, 2)$-Ulam set such that all vectors $v_i = (x_i, y_i)$ have $x_i, y_i \geq 4$ and even. Furthermore, suppose that there exist integers $m, n$ such for all $i$, there exists a $j$ such that $m - v_i = v_j$. Then $\mathcal{U}$ is eventually $(m, n)$-periodic, and for any other vector $v_{n+1} = (x_{n+1}, y_{n+1})$ with $x_{n+1}, y_{n+1} \geq 4$ such that at least one of $x_{n+1}, y_{n+1}$ is odd, $\mathcal{U}' := U_{\mathcal{A}}(v_1, v_2, \ldots, v_{n+1})$ is eventually $(m, n)$-periodic.*

*Proof.* Let $(a, b)$ be a vector such that $a, b > 0$ are both odd. We shall show that $(a, b) \in \mathcal{U}$ if and only if $(a, b) + (m, n) \in U$. Indeed, for all $i$, we have $(a, b) + (m, n) - v_i = (a, b) + v_j$ for some $j$. If $(a, b) \in \mathcal{U}$, this gives a nonstandard representation of $(a, b) + (m, n) - v_i$; hence it is not in $\mathcal{U}$, and so by Lemma 6.2, it follows that $(a, b) + (m, n) \in U$. On the other hand, if $(a, b) \notin \mathcal{U}$, then again by Lemma 6.2, we know that $(a, b) - v_i \in \mathcal{U}$ for some $i$, and it follows that $(a, b) + (m, n) - v_i \in \mathcal{U}$. But this implies $(a, b) + (m, n) \notin \mathcal{U}$. Since by Lemma 6.1 we know that all sufficiently large vectors in $\mathcal{U}$ have odd coordinates, we conclude that $\mathcal{U}$ is eventually $(m, n)$-periodic.

It remains to prove that $\mathcal{U}'$ is eventually periodic. Let $(a, b)$ be a vector such that $a, b$ are both odd, and $a > x_i$, $b > y_i$ for every $1 \leq i \leq n + 1$. If $(a, b) \in \mathcal{U}'$, then $(a, b) + (m, n) - v_i = (a, b) + v_j \notin \mathcal{U}'$ for every $1 \leq i \leq n$, so it suffices to prove that $(a, b) + (m, n) - v_{n+1} \notin \mathcal{U}'$ to conclude that $(a, b) + (m, n) \in \mathcal{U}'$. However,

the coordinates of $(a, b) + (m, n) - v_{n+1}$ are both integers greater than 1, and at least one of them is even. By Lemma 6.1, this implies $(a, b) + (m, n) - v_{n+1} \notin U'$, and so $(a, b) + (m, n) \in U'$. In the other direction, we know that if $(a, b) \notin \mathcal{U}'$, then $(a, b) - v_i \in U'$ for some $i$. If $i \neq n + 1$, the proof is the same as before. If $(a, b) - v_{n+1} \in U'$, then we note that $(a, b) + (m, n) \notin U'$ by parity considerations. We thus conclude that $\mathcal{U}'$ is eventually $(m, n)$-periodic. □

**Theorem 6.4.** *Let $\mathcal{U} = U_{\mathcal{A}}(v_1, v_2, \ldots, v_n)$ be a nondegenerate $(k, 2)$-Ulam set such that all vectors $v_i = (x_i, y_i)$ have $x_i, y_i \geq 4$ and at least one of $x_i, y_i$ is odd. Then $\mathcal{U}$ is eventually periodic, with period $(2, 2)$.*

*Proof.* Note that if $x_i, y_i$ are both odd, then $(x_i, y_i) \in \mathcal{A}$, which would contradict the fact that $\mathcal{U}$ is nondegenerate. Thus all vectors $v_i$ have one even component. We claim that if $x_n$ is even, then

$$\mathcal{U} = \{v_n\} \cup U_{\mathcal{A}}(v_1, \ldots, v_{n-1}) \setminus (\{(x_n+1, y_n+2l) \mid l \in \mathbb{Z}_{\geq 0}\} \cup \{v_i + v_j \mid 1 \leq i < j \leq n\}),$$

and if $y_n$ is even, then

$$\mathcal{U} = \{v_n\} \cup U_{\mathcal{A}}(v_1, \ldots, v_{n-1}) \setminus (\{(x_n+2l, y_n+1) \mid l \in \mathbb{Z}_{\geq 0}\} \cup \{v_i + v_j \mid 1 \leq i < j \leq n\}).$$

The base case follows from the results of Section 4. Now, note that if $v_n \in \mathcal{U}$, then certainly either $v_n + (1, 2l)$ or $v_n + (2l, 1)$ is a nonstandard representation, so correspondingly $(x_n + 1, y_n + 2l)$ or $(x_n + 2l, y_n + 1)$ is not in $\mathcal{U}$. Similarly, all vectors $v_i + v_j$ have at least two representations. It remains to prove that removing these vectors doesn't lead to removing representations of other points. This cannot be — all removed vectors have both coordinates odd, and $\mathcal{U}$ contains all vectors with positive odd coordinates, all of which have one standard representation that we know has not been removed. That this is true for all the sets $U_{\mathcal{A}}(v_1, \ldots, v_k)$ follows by induction. This concludes the proof, since it is clear that each of the sets $U_{\mathcal{A}}(v_1, \ldots, v_n)$ is eventually periodic, with period $(2, 2)$, by induction. □

These two results immediately imply Theorem 1.9.

*Proof of Theorem 1.9.* If both $v_1$ and $v_2$ have at least one odd coordinate, the result follows from Theorem 6.4. Otherwise, let $v_i, v_j$ be the vectors that have both coordinates even — here, $i, j$ need not be distinct. Then by Theorem 6.3, $\mathcal{U}$ is eventually $(v_i+v_j)$-periodic. □

**Acknowledgements**

The authors are indebted Stefan Steinerberger for introducing them to Ulam sets and their odd properties, as well as providing helpful advice along the way. They would also like to thank the organizers of SUMRY 2017, where the collaboration that led to this paper took place.

## References

[Cassaigne and Finch 1995] J. Cassaigne and S. R. Finch, "A class of 1-additive sequences and quadratic recurrences", *Experiment. Math.* **4**:1 (1995), 49–60. MR Zbl

[Finch 1991] S. R. Finch, "Conjectures about $s$-additive sequences", *Fibonacci Quart.* **29**:3 (1991), 209–214. MR Zbl

[Finch 1992a] S. R. Finch, "On the regularity of certain 1-additive sequences", *J. Combin. Theory Ser. A* **60**:1 (1992), 123–130. MR Zbl

[Finch 1992b] S. R. Finch, "Patterns in 1-additive sequences", *Experiment. Math.* **1**:1 (1992), 57–63. MR Zbl

[Gibbs 2015] P. Gibbs, "An efficient method for computing Ulam numbers", preprint, 2015, available at http://vixra.org/abs/1508.0085.

[Gibbs and McCranie 2017] P. Gibbs and J. McCranie, "The Ulam numbers up to one trillion", preprint, 2017, available at http://vixra.org/abs/1711.0134.

[Hinman et al. 2018] J. Hinman, B. Kuca, A. Schlesinger, and A. Sheydvasser, "The unreasonable rigidity of Ulam sequences", *J. Number Theory* (online publication July 2018).

[Kravitz and Steinerberger 2017] N. Kravitz and S. Steinerberger, "Ulam sequences and Ulam sets", preprint, 2017. arXiv

[Kuca 2018] B. Kuca, "Structures in additive sequences", preprint, 2018. arXiv

[Motte 1998] W. F. Motte, Jr. (editor), *Oulipo: a primer of potential literature*, Dalkey Archive, 1998.

[Queneau 1972] R. Queneau, "Sur les suites $s$-additives", *J. Combinatorial Theory Ser. A* **12** (1972), 31–71. MR Zbl

[Schmerl and Spiegel 1994] J. Schmerl and E. Spiegel, "The regularity of some 1-additive sequences", *J. Combin. Theory Ser. A* **66**:1 (1994), 172–175. MR Zbl

[Steinerberger 2017] S. Steinerberger, "A hidden signal in the Ulam sequence", *Exp. Math.* **26**:4 (2017), 460–467. MR Zbl

[Ulam 1964] S. Ulam, "Combinatorial analysis in infinite sets and some physical theories", *SIAM Rev.* **6** (1964), 343–355. MR Zbl

Received: 2018-07-10 Revised: 2018-07-31 Accepted: 2018-09-05

joshua.hinman@yale.edu *Department of Mathematics, Yale University, New Haven, CT, United States*

borys.kuca@postgrad.manchester.ac.uk *School of Mathematics, University of Manchester, Manchester, United Kingdom*

alexander.schlesinger@yale.edu *Department of Mathematics, Yale University, New Haven, CT, United States*

sheydvasser@gmail.com *Department of Mathematics, The Graduate Center, New York, NY, United States*

## Guidelines for Authors

Submissions in all mathematical areas are encouraged. All manuscripts accepted for publication in *Involve* are considered publishable in quality journals in their respective fields, and include a minimum of one-third student authorship. Submissions should include substantial faculty input; faculty co-authorship is strongly encouraged. Authors may submit manuscripts in PDF format on-line at the Submission page at the Involve website.

**Originality**. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

**Language**. Articles in *Involve* are usually in English, but articles written in other languages are welcome.

**Required items**. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

**Format**. Authors are encouraged to use LaTeX but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

**References**. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

**Figures**. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@msp.org with details about how your graphics were generated.

**White space**. Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

**Proofs**. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

# involve

2019 vol. 12 no. 3





1944-4176(2019)12:3;1-2