involve

a journal of mathematics

msp

2012 vol. 5, no. 2

involve

msp.berkeley.edu/involve

PRODUCTION

Silvio Levy, Scientific Editor   Sheila Newbery, Senior Production Editor   Cover design: © 2008 Alex Scorpan

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

The subscription price for 2012 is US $105/year for the electronic version, and $145/year (+$35 shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to Mathematical Sciences Publishers, Department of Mathematics, University of California, Berkeley, CA 94704-3840, USA.

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

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

PUBLISHED BY

mathematical sciences publishers

http://msp.org/

A NON-PROFIT CORPORATION

Typeset in LaTeX

INVOLVE 5:2 (2012)
dx.doi.org/10.2140/involve.2012.5.115

A Giambelli formula for the S^1-equivariant cohomology of type A Peterson varieties

Darius Bayegan and Megumi Harada

(Communicated by Ravi Vakil)

We prove a *Giambelli formula* for the Peterson Schubert classes in the S^1-equivariant cohomology ring of a type A Peterson variety. The proof uses the Monk formula for the equivariant structure constants for the Peterson Schubert classes derived by Harada and Tymoczko. In addition, we give proofs of two facts observed by H. Naruse: firstly, that some constants that appear in the multiplicative structure of the S^1-equivariant cohomology of Peterson varieties are Stirling numbers of the second kind, and secondly, that the Peterson Schubert classes satisfy a stability property in a sense analogous to the stability of the classical equivariant Schubert classes in the T-equivariant cohomology of the flag variety.

1. Introduction

The main result of this note is a formula of Giambelli type in the S^1-equivariant cohomology[1] of type A Peterson varieties. Specifically, we give an explicit formula that expresses an arbitrary *Peterson Schubert class* in terms of the degree-2 Peterson Schubert classes. We call this a "Giambelli formula" by analogy with the standard Giambelli formula in classical Schubert calculus [Fulton 1997], which expresses an arbitrary Schubert class in terms of degree-2 Schubert classes.

We briefly recall the setting of our results. *Peterson varieties* in type A can be defined as the following subvariety Y of $\mathcal{F}\ell ags(\mathbb{C}^n)$:

$$Y := \{V_\bullet = (0 \subseteq V_1 \subseteq V_2 \subseteq \cdots \subseteq V_{n-1} \subseteq V_n = \mathbb{C}^n) \mid NV_i \subseteq V_{i+1} \text{ for all } i = 1, \dots, n-1\}, \quad (1\text{-}1)$$

MSC2010: primary 14N15; secondary 55N91.

Keywords: Giambelli formula, Peterson variety, Schubert calculus, equivariant cohomology.

The second author is partially supported by an NSERC Discovery Grant, an NSERC University Faculty Award, and an Ontario Ministry of Research and Innovation Early Researcher Award.

[1] All our cohomology rings are with coefficients in $\mathbb{C}$.

where $N : \mathbb{C}^n \to \mathbb{C}^n$ denotes the principal nilpotent operator. These varieties have been much studied due to their relation to the quantum cohomology of the flag variety [Kostant 1996; Rietsch 2003]. Thus it is natural to study their topology, including the structure of their (equivariant) cohomology rings. We do so through Schubert calculus techniques. Our results extend techniques initiated and developed in [Harada and Tymoczko 2010; 2011], to which we refer the reader for further details and motivation.

There is a natural circle subgroup of $U(n, \mathbb{C})$, recalled in Section 2, that acts on Y. The inclusion of Y into $\mathcal{F}\ell ags(\mathbb{C}^n)$ induces a natural ring homomorphism

$$H^*_T(\mathcal{F}\ell ags(\mathbb{C}^n)) \to H^*_{S^1}(Y) \tag{1-2}$$

where T is the subgroup of diagonal matrices of $U(n, \mathbb{C})$ acting in the usual way on $\mathcal{F}\ell ags(\mathbb{C}^n)$. One of the main results of [Harada and Tymoczko 2011] is that a certain subset of the equivariant Schubert classes $\{\sigma_w\}_{w \in S_n}$ in $H^*_T(\mathcal{F}\ell ags(\mathbb{C}^n))$ maps under the projection (1-2) to a computationally convenient module basis of $H^*_{S^1}(Y)$. We refer to the images via (1-2) of $\{\sigma_w\}_{w \in S_n}$ in $H^*_{S^1}(Y)$ as *Peterson Schubert classes*. Theorem 6.12 of the same reference gives a manifestly positive *Monk formula* for the product of a degree-2 Peterson Schubert class with an arbitrary Peterson Schubert class, expressed as a $H^*_{S^1}(\mathrm{pt})$-linear combination of Peterson Schubert classes. This is an example of equivariant Schubert calculus in the realm of Hessenberg varieties (of which Peterson varieties are a special case), and we view the Giambelli formula (Theorem 3.2) as a further development of this theory. The Giambelli formula for Peterson varieties was also independently observed by H. Naruse.

Our Giambelli formula also allows us to simplify the presentation of the ring $H^*_{S^1}(Y)$ given in [Harada and Tymoczko 2011, Section 6]. This is because the previous presentation used as its generators all of the elements in the module basis given by Peterson Schubert classes, although the ring $H^*_{S^1}(Y)$ is multiplicatively generated by only the degree-2 Peterson Schubert classes. Details are explained starting on page 123 below, where we also give a concrete example in $n = 4$ to illustrate our results. We also formulate a conjecture (cf. Remark 3.12), suggested to us by the referee of this manuscript, that the ideal of defining relations is in fact generated by the quadratic relations only. If true, this would be a significant further simplification of the presentation of this ring and would lead to interesting further questions (both combinatorial and geometric).

In Sections 4 and 5, we present proofs of two facts concerning Peterson Schubert classes, which we learned from H. Naruse. The results are due to Naruse but the proofs given here are our own. We chose to include these results because they do not appear elsewhere in the literature. The first fact is that *Stirling numbers of the second kind* (see Section 4 for the definition) appear naturally in the product structure of $H^*_{S^1}(Y)$. The second is that the Peterson Schubert classes satisfy a *stability condition* with respect to the natural inclusions of Peterson varieties induced from the inclusions $\mathcal{F}\ell ags(\mathbb{C}^n) \hookrightarrow \mathcal{F}\ell ags(\mathbb{C}^{n+1})$.

2. Peterson varieties and S^1-fixed points

In this section we briefly recall the objects under study. For details we refer the reader to [Harada and Tymoczko 2011]. Since we work exclusively in Lie type A we henceforth omit it from our terminology.

By the *flag variety* we mean the homogeneous space $\mathrm{GL}(n, \mathbb{C})/B$, where B is the standard Borel subgroup of upper-triangular invertible matrices. The flag variety can also be identified with the space of nested subspaces in $\mathbb{C}^n$, that is,

$$\begin{aligned}\mathcal{F}\ell ags(\mathbb{C}^n) &:= \{V_\bullet = (\{0\} \subseteq V_1 \subseteq V_2 \subseteq \cdots \subseteq V_{n-1} \subseteq V_n = \mathbb{C}^n) \mid \dim_{\mathbb{C}}(V_i) = i\} \\ &\cong \mathrm{GL}(n, \mathbb{C})/B.\end{aligned}$$

Let N be the $n \times n$ principal nilpotent operator given with respect to the standard basis of $\mathbb{C}^n$ as the matrix with one $n \times n$ Jordan block of eigenvalue 0, that is,

$$N = \begin{bmatrix} 0 & 1 & 0 & & & \\ 0 & 0 & 1 & & & \\ 0 & 0 & 0 & & & \\ & & & \ddots & & \\ & & & & 0 & 1 \\ & & & & 0 & 0 \end{bmatrix}. \tag{2-1}$$

Fix n a positive integer. The main geometric object under study, the *Peterson variety* Y, is the subvariety of $\mathcal{F}\ell ags(\mathbb{C}^n)$ defined in (1-1) where N is the standard principal nilpotent in (2-1). The variety Y is a (singular) projective variety of complex dimension $n-1$.

We recall some facts from [Harada and Tymoczko 2011]. The following circle subgroup of $U(n, \mathbb{C})$ preserves Y:

$$S^1 = \left\{ \begin{bmatrix} t^n & 0 & \cdots & 0 \\ 0 & t^{n-1} & & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & & t \end{bmatrix} \;\middle|\; t \in \mathbb{C},\ \|t\| = 1 \right\} \subseteq T^n \subseteq U(n, \mathbb{C}). \tag{2-2}$$

Here T^n is the standard maximal torus of $U(n, \mathbb{C})$ consisting of diagonal unitary matrices. The S^1-fixed points of Y are isolated and are a subset of the T^n-fixed points of $\mathcal{F}\ell ags(\mathbb{C}^n)$. As is standard, we identify the T^n-fixed points in $\mathcal{F}\ell ags(\mathbb{C}^n)$ with the permutations S_n. In particular since Y^{S^1} is a subset of $\mathcal{F}\ell ags(\mathbb{C}^n)^{T^n}$, we think of the Peterson fixed points as permutations in S_n. There is a natural bijective correspondence between the Peterson fixed points Y^{S^1} and subsets of $\{1, 2, \dots, n-1\}$ which we now briefly recall. It is explained in [Harada and Tymoczko 2011, Section 2.3] that a permutation $w \in S_n$ is in Y^{S^1} precisely when the one-line notation of w^{-1} is of the form

$$w^{-1} = \underbrace{j_1\, j_1 - 1 \,\cdots\, 1}_{j_1 \text{ entries}}\; \underbrace{j_2\, j_2 - 1 \,\cdots\, j_1 + 1}_{j_2 - j_1 \text{ entries}} \cdots \underbrace{n\, n-1 \,\cdots\, j_m + 1}_{n - j_m \text{ entries}} \tag{2-3}$$

where $1 \leq j_1 < j_2 < \cdots < j_m < n$ is any sequence of strictly increasing integers. For example, for $n = 9$, $m = 2$ and $j_1 = 3$, $j_2 = 7$, then the permutation w^{-1} in (2-3) has one-line notation 321765498. Thus for each permutation $w \in S_n$ satisfying (2-3) we define

$$\mathscr{A} := \{i : w^{-1}(i) = w^{-1}(i+1) + 1 \text{ for } 1 \leq i \leq n-1\} \subseteq \{1, 2, \dots, n-1\}.$$

This gives a one-to-one correspondence between the power set of $\{1, 2, \dots, n-1\}$ and Y^{S^1}. We denote the Peterson fixed point corresponding to a subset $\mathscr{A} \subseteq \{1, 2, \dots, n-1\}$ by $w_{\mathscr{A}}$.

Example 2.1. Let $n = 5$ and suppose $\mathscr{A} = \{1, 2, 4\}$. Then the associated permutation is $w_{\mathscr{A}} = 32154$.

Indeed, for a fixed n, we can also easily enumerate all the Peterson fixed points by using this correspondence.

Example 2.2. Let $n = 4$. Then Y^{S^1} consists of $2^3 = 8$ elements in correspondence with the subsets of $\{1, 2, 3\}$, namely: $w_{\varnothing} = 1234$, $w_{\{1\}} = 2134$, $w_{\{2\}} = 1324$, $w_{\{3\}} = 1243$, $w_{\{1,2\}} = 3214$, $w_{\{2,3\}} = 1432$, $w_{\{1,3\}} = 2143$, $w_{\{1,2,3\}} = 4321$.

Given a choice of subset $\mathscr{A} \subseteq \{1, 2, \dots, n-1\}$, there is a natural decomposition of $\mathscr{A}$ as follows. We say that a set of consecutive integers

$$\{a, a+1, \dots, a+k\} \subseteq \mathscr{A}$$

is a *maximal consecutive* (*sub*)*string* of $\mathscr{A}$ if a and k are such that neither $a-1$ nor $a+k+1$ is in $\mathscr{A}$. For $a_1 := a$ and $a_2 := a_1 + k$, we denote the corresponding maximal consecutive substring by $[a_1, a_2]$. It is straightforward to see that any $\mathscr{A}$ uniquely decomposes into a disjoint union of maximal consecutive substrings

$$\mathscr{A} = [a_1, a_2] \cup [a_3, a_4] \cup \cdots \cup [a_{m-1}, a_m].$$

For instance, if $\mathscr{A} = \{1, 2, 3, 5, 6, 8\}$, then its decomposition into maximal consecutive substrings is $\{1, 2, 3\} \cup \{5, 6\} \cup \{8\} = [1, 3] \cup [5, 6] \cup [8, 8]$.

Suppose $\mathscr{A} = \{j_1 < j_2 < \cdots < j_m\}$. Finally we recall that we can associate to each $w_{\mathscr{A}}$ a permutation $v_{\mathscr{A}}$ by the recipe

$$w_{\mathscr{A}} \mapsto v_{\mathscr{A}} := s_{j_1} s_{j_2} \cdots s_{j_m} \tag{2-4}$$

where an s_i denotes the simple transposition $(i, i+1)$ in S_n.

3. The Giambelli formula for Peterson varieties

The Giambelli formula. In this section we prove the main result of this note, namely, a *Giambelli formula for Peterson varieties*.

As recalled above, the Peterson variety Y is an S^1-space for a subtorus S^1 of T^n and it can be checked that $Y^{S^1} = (\mathcal{F}\ell ags(\mathbb{C}^n))^{T^n} \cap Y$. There is a forgetful map from T^n-equivariant cohomology to S^1-equivariant cohomology obtained by the inclusion $S^1 \hookrightarrow T$, so there is a commutative diagram

$$\begin{tikzcd}
H^*_{T^n}(\mathcal{F}\ell ags(\mathbb{C}^n)) \arrow[r] \arrow[d] & H^*_{T^n}((\mathcal{F}\ell ags(\mathbb{C}^n))^{T^n}) \arrow[d] \\
H^*_{S^1}(\mathcal{F}\ell ags(\mathbb{C}^n)) \arrow[r] \arrow[d] & H^*_{S^1}((\mathcal{F}\ell ags(\mathbb{C}^n))^{T^n}) \arrow[d] \\
H^*_{S^1}(Y) \arrow[r] & H^*_{S^1}(Y^{S^1}).
\end{tikzcd} \tag{3-1}$$

The *equivariant Schubert classes* $\{\sigma_w\}$ in $H^*_{T^n}(\mathcal{F}\ell ags(\mathbb{C}^n))$ are well-known to form a $H^*_{T^n}(\mathrm{pt})$-module basis for $H^*_{T^n}(\mathcal{F}\ell ags(\mathbb{C}^n))$. We call the image of σ_w under the projection map $H^*_{T^n}(\mathcal{F}\ell ags(\mathbb{C}^n)) \to H^*_{S^1}(Y)$ the *Peterson Schubert class corresponding to w*. For the permutations $v_{\mathcal{A}}$ defined in (2-4), we denote by $p_{\mathcal{A}}$ the corresponding Peterson Schubert class, that is the image of $\sigma_{v_{\mathcal{A}}}$. (This is slightly different notation from that used in [Harada and Tymoczko 2011].) We denote by $p_{\mathcal{A}}(w) \in H^*_{S^1}(\mathrm{pt}) \cong \mathbb{C}[t]$ the restriction of the Peterson Schubert class $p_{\mathcal{A}}$ to the fixed point $w \in Y^{S^1}$.

One of the main results of [Harada and Tymoczko 2011] is that the set of 2^{n-1} Peterson Schubert classes $\{p_{\mathcal{A}}\}_{\mathcal{A} \subseteq \{1,2,\ldots,n-1\}}$ form a $H^*_{S^1}(\mathrm{pt})$-module basis for $H^*_{S^1}(Y)$ where $v_{\mathcal{A}}$ is defined in (2-4). (The fact that $H^*_{S^1}(Y)$ is a free module of rank 2^{n-1} over $H^*_{S^1}(\mathrm{pt})$ fits nicely with the result [Sommers and Tymoczko 2006, Theorem 10.2] that the Poincaré polynomial of Y is given by $(q^2+1)^{n-1}$.) It is also shown in [Harada and Tymoczko 2011] that the $n-1$ degree-2 classes $\{p_i := p_{s_i}\}_{i=1}^{n-1}$ form a multiplicative set of generators for $H^*_{S^1}(Y)$. These classes p_i are also (equivariant) Chern classes of certain line bundles over Y. Moreover, there is a *Monk formula* [Harada and Tymoczko 2011, Theorem 6.12] which expresses a product

$$p_i \, p_{\mathcal{A}}$$

for any $i \in \{1, 2, \ldots, n-1\}$ and any $\mathcal{A} \subseteq \{1, 2, \ldots, n-1\}$ as a $H^*_{S^1}(\mathrm{pt})$-linear combination of the additive module basis $\{p_{\mathcal{A}}\}$. Since the p_i multiplicatively generate the ring, this Monk formula completely determines the ring structure of $H^*_{S^1}(Y)$. Furthermore it is in principle possible to express any $p_{\mathscr{A}}$ in terms of the p_i. Our Giambelli formula is an explicit formula which achieves this (cf. for example [Fulton 1997] for the version in classical Schubert calculus).

We begin by recalling the Monk formula, for which we need some terminology. Fix $\mathscr{A} \subseteq \{1, 2, \dots, n-1\}$. We define $\mathscr{H}_{\mathscr{A}} : \mathscr{A} \to \mathscr{A}$ by

$$\mathscr{H}_{\mathscr{A}}(j) = \text{the maximal element in the maximal consecutive substring of } \mathscr{A} \text{ containing } j.$$

Similarly, we define $\mathscr{T}_{\mathscr{A}} : \mathscr{A} \to \mathscr{A}$ by

$$\mathscr{T}_{\mathscr{A}}(j) = \text{the minimal element in the maximal consecutive substring of } \mathscr{A} \text{ containing } j.$$

We say that the maps $\mathscr{H}_{\mathscr{A}}$ and $\mathscr{T}_{\mathscr{A}}$ give the "head" and "tail" of each maximal consecutive substring of $\mathscr{A}$. For an example see [Harada and Tymoczko 2011, Example 5.6]. We recall the following.

Theorem 3.1 (Monk formula for Peterson varieties [Harada and Tymoczko 2011, Theorem 6.12]). *Fix a positive integer n. Let Y be the Peterson variety in $\mathscr{F}\ell ags(\mathbb{C}^n)$ with the natural S^1-action defined by (2-2). For $\mathscr{A} \subseteq \{1, 2, \dots, n-1\}$, let $v_{\mathscr{A}} \in S_n$ be the permutation in (2-4), and let $p_{\mathscr{A}}$ be the corresponding Peterson Schubert class in $H^*_{S^1}(Y)$. Then*

$$p_i \cdot p_{\mathscr{A}} = p_i(w_{\mathscr{A}}) \cdot p_{\mathscr{A}} + \sum_{\mathscr{A} \subsetneq \mathscr{B} \text{ and } |\mathscr{B}|=|\mathscr{A}|+1} c^{\mathscr{B}}_{i,\mathscr{A}} \cdot p_{\mathscr{B}}, \tag{3-2}$$

where, for a subset $\mathscr{B} \subseteq \{1, 2, \dots, n-1\}$ which is a disjoint union $\mathscr{B} = \mathscr{A} \cup \{k\}$,

- *if $i \notin \mathscr{B}$ then $c^{\mathscr{B}}_{i,\mathscr{A}} = 0$,*
- *if $i \in \mathscr{B}$ and $i \notin [\mathscr{T}_{\mathscr{B}}(k), \mathscr{H}_{\mathscr{B}}(k)]$, then $c^{\mathscr{B}}_{i,\mathscr{A}} = 0$,*
- *if $k \leq i \leq \mathscr{H}_{\mathscr{B}}(k)$, then*
$$c^{\mathscr{B}}_{i,\mathscr{A}} = (\mathscr{H}_{\mathscr{B}}(k) - i + 1) \cdot \binom{\mathscr{H}_{\mathscr{B}}(k) - \mathscr{T}_{\mathscr{B}}(k) + 1}{k - \mathscr{T}_{\mathscr{B}}(k)}, \tag{3-3}$$
- *if $\mathscr{T}_{\mathscr{B}}(k) \leq i \leq k-1$,*
$$c^{\mathscr{B}}_{i,\mathscr{A}} = (i - \mathscr{T}_{\mathscr{B}}(k) + 1) \cdot \binom{\mathscr{H}_{\mathscr{B}}(k) - \mathscr{T}_{\mathscr{B}}(k) + 1}{k - \mathscr{T}_{\mathscr{B}}(k) + 1}. \tag{3-4}$$

We also recall that [Harada and Tymoczko 2011, Lemma 6.7] implies that if $\mathscr{B}, \mathscr{B}'$ are two disjoint subsets of $\{1, 2, \dots, n-1\}$ such that there is no i in $\mathscr{B}$ and j in $\mathscr{B}'$ with $|i-j| = 1$, then $p_{\mathscr{B} \cup \mathscr{B}'} = p_{\mathscr{B}} p_{\mathscr{B}'}$. It follows that for any $\mathscr{A}$ we have

$$p_{\mathscr{A}} = p_{[a_1,a_2]} \cdot p_{[a_3,a_4]} \cdots p_{[a_{m-1},a_m]} \tag{3-5}$$

where $\mathcal{A} = [a_1, a_2] \cup [a_3, a_4] \cup \cdots \cup [a_{m-1}, a_m]$ is the decomposition of $\mathcal{A}$ into maximal consecutive substrings. In particular, in order to give an expression for $p_{\mathcal{A}}$ in terms of the elements p_i, from (3-5) we see that it suffices to give a formula only for the special case in which $\mathcal{A}$ consists of a single maximal consecutive string.

We now state and prove our Giambelli formula.

Theorem 3.2. *Fix n a positive integer. Let Y be the Peterson variety in $\mathcal{F}\ell ags(\mathbb{C}^n)$ with the S^1-action defined by (2-2). Suppose $\mathcal{A} = \{a, a+1, a+2, \dots, a+k\}$ where $1 \leq a \leq n-1$ and $0 \leq k \leq n-1-a$. Let $v_{\mathcal{A}}$ be the permutation corresponding to $\mathcal{A}$ defined in (2-4) and let $p_{\mathcal{A}}$ be the associated Peterson Schubert class. Then*

$$p_{\mathcal{A}} = \frac{1}{(k+1)!} \prod_{j \in \mathcal{A}} p_j.$$

We use the following lemma.

Lemma 3.3. *Suppose $i \in \{1, 2, \dots, n-1\}$ and $\mathcal{A} \subseteq \{1, 2, \dots, n-1\}$. Suppose further that $i \notin \mathcal{A}$. Then the Monk relation*

$$p_i \cdot p_{\mathcal{A}} = p_i(w_{\mathcal{A}}) \cdot p_{\mathcal{A}} + \sum_{\mathcal{A} \subset \mathcal{B} \text{ and } |\mathcal{B}| = |\mathcal{A}|+1} c_{i,\mathcal{A}}^{\mathcal{B}} \cdot p_{\mathcal{B}}$$

simplifies to

$$p_i \cdot p_{\mathcal{A}} = c_{i,\mathcal{A}}^{\mathcal{A} \cup \{i\}} \cdot p_{\mathcal{A} \cup \{i\}}. \tag{3-6}$$

Proof. First observe that the Monk relation simplifies to

$$p_i \cdot p_{\mathcal{A}} = \sum_{\mathcal{A} \subset \mathcal{B} \text{ and } |\mathcal{B}| = |\mathcal{A}|+1} c_{i,\mathcal{A}}^{\mathcal{B}} \cdot p_{\mathcal{B}} \tag{3-7}$$

if $i \notin \mathcal{A}$, since in that case $p_i(w_{\mathcal{A}}) = 0$ by [Harada and Tymoczko 2011, Lemma 6.4]. Moreover, from Theorem 3.1 we also know that $c_{i,\mathcal{A}}^{\mathcal{B}} = 0$ if $i \notin \mathcal{B}$. Hence the summands appearing in (3-7) correspond to $\mathcal{B}$ satisfying $\mathcal{A} \subseteq \mathcal{B}$, $|\mathcal{B}| = |\mathcal{A}| + 1$, and $i \in \mathcal{B}$. On the other hand, since $i \notin \mathcal{A}$ by assumption, this means that there is only one nonzero summand in the right hand side of (3-7), namely, the term corresponding to $\mathcal{B} = \mathcal{A} \cup \{i\}$. Then (3-6) follows. □

Proof of Theorem 3.2. We proceed by induction on k. First consider the base case where $k = 0$. Then $A = \{a\}$, so $p_{v_{\mathcal{A}}} = p_a$. On the right hand side, we have $\frac{1}{(0+1)!} \prod_{j \in \mathcal{A}} p_j = p_a$. This verifies the base case.

By induction, suppose the claim holds for $k-1$. We now show that the claim holds for k. Consider $\mathcal{A}' := \{a, a+1, \dots, a+k-1\}$ and consider the product $p_{a+k} \cdot p_{\mathcal{A}'}$. From the Monk formula in Theorem 3.1 we know that

$$p_{a+k} \cdot p_{\mathcal{A}'} = p_{a+k}(w_{\mathcal{A}'}) \cdot p_{\mathcal{A}'} + \sum_{\substack{\mathcal{A}' \subseteq \mathcal{B} \\ |\mathcal{B}| = |\mathcal{A}'|+1}} c_{a+k,\mathcal{A}'}^{\mathcal{B}} \cdot p_{\mathcal{B}}. \tag{3-8}$$

On the other hand since by definition $a+k \notin \mathscr{A}'$, by Lemma 3.3 the equality (3-8) further simplifies to

$$p_{a+k} \cdot p_{\mathscr{A}'} = c^{\mathscr{A}}_{a+k,\mathscr{A}'} \cdot p_{\mathscr{A}}.$$

Moreover, since $\mathscr{A} = \mathscr{A}' \cup \{a+k\}$, we have $\mathscr{H}_{\mathscr{A}}(a+k) = a+k$ and $\mathscr{T}_{\mathscr{A}}(a+k) = a$. Hence, by Theorem 3.1,

$$\begin{aligned} c^{\mathscr{A}}_{a+k,\mathscr{A}'} &= (\mathscr{H}_{\mathscr{A}}(a+k) - (a+k) + 1) \binom{\mathscr{H}_{\mathscr{A}}(a+k) - \mathscr{T}_{\mathscr{A}}(a+k) + 1}{(a+k) - \mathscr{T}_{\mathscr{A}}(a+k)} \\ &= ((a+k) - (a+k) + 1) \binom{a+k-a+1}{(a+k) - a} \\ &= k+1. \end{aligned} \tag{3-9}$$

Therefore

$$p_{a+k} \cdot p_{\mathscr{A}'} = (k+1) \cdot p_{\mathscr{A}}.$$

By the inductive hypothesis we have for the set $\mathscr{A}' = \{a, a+1, \dots, a+k-1\}$

$$p_{\mathscr{A}'} = \frac{1}{k!} \prod_{j \in \mathscr{A}'} p_j.$$

Substituting into the above equation yields

$$p_{\mathscr{A}} = \frac{1}{(k+1)!} \prod_{j \in \mathscr{A}} p_j$$

as desired. This completes the proof. □

Remark 3.4. We thank the referee for the following observation. The formula in Theorem 3.2 suggests that the classes p_i behave like a normal crossings divisor (up to quotient singularities), with all other classes arising (up to rational coefficients) as intersections of the components. It would certainly be of interest to understand more precisely the underlying geometry which gives rise not only to the Giambelli relation in Theorem 3.2 but also to the original Monk formula [Harada and Tymoczko 2011, Theorem 6.12].

From Theorem 3.2 it immediately follows that for any subset

$$\mathscr{A} = [a_1, a_2] \cup [a_3, a_4] \cup \cdots \cup [a_{m-1}, a_m]$$

with its decomposition into maximal consecutive substrings, we have

$$p_{\mathscr{A}} = \frac{1}{(a_2 - a_1 + 1)!} \cdot \frac{1}{(a_4 - a_3 + 1)!} \cdots \frac{1}{(a_m - a_{m-1} + 1)!} \prod_{j \in \mathscr{A}} p_j. \tag{3-10}$$

For the purposes of the next section we introduce the notation

$$\sigma(\mathscr{A}) := \frac{1}{(a_2-a_1+1)!}\cdot\frac{1}{(a_4-a_3+1)!}\cdots\frac{1}{(a_m-a_{m-1}+1)!} \tag{3-11}$$

for the rational coefficient appearing in (3-10). The following is an immediate corollary of this discussion.

Corollary 3.5. *Let*

$$\mathscr{A} = [a_1, a_2] \cup [a_3, a_4] \cup \cdots \cup [a_{m-1}, a_m].$$

Then

$$p_{\mathscr{A}} = \sigma(\mathscr{A}) \prod_{j\in\mathscr{A}} p_j.$$

Simplification of the Monk relations. In this section we explain how to use the Giambelli formula to simplify the ring presentation of $H^*_{S^1}(Y)$ given in [Harada and Tymoczko 2011, Section 6]. Recall that the Peterson Schubert classes $\{p_{\mathscr{A}}\}$ form an additive module basis for $H^*_{S^1}(Y)$ and the degree-2 classes $\{p_i\}_{i=1}^{n-1}$ form a multiplicative basis, so the Monk relations give a presentation of the ring $H^*_{S^1}(Y)$ via generators and relations as follows.

Theorem 3.6 [Harada and Tymoczko 2011, Corollary 6.14]. *Fix n a positive integer. Let Y be the Peterson variety in $\mathscr{F}\ell ags(\mathbb{C}^n)$ with the S^1-action defined by* (2-2). *For $\mathscr{A} \subseteq \{1, 2, \dots, n-1\}$, let $v_{\mathscr{A}} \in S_n$ be the permutation given in* (2-4), *and let $p_{\mathscr{A}}$ be the corresponding Peterson Schubert class in $H^*_{S^1}(Y)$. Let $t \in H^*_{S^1}(\mathrm{pt}) \cong \mathbb{C}[t]$ denote both the generator of $H^*_{S^1}(\mathrm{pt})$ and its image $t \in H^*_{S^1}(Y)$. Then the S^1-equivariant cohomology $H^*_{S^1}(Y)$ is given by*

$$H^*_{S^1}(Y) \cong \mathbb{C}[t, \{p_{\mathscr{A}}\}_{\mathscr{A}\subseteq\{1,2,\dots,n-1\}}]/\mathscr{J}$$

where $\mathscr{J}$ is the ideal generated by the relations (3-2).

In order to state the main result of this section we introduce some notation. For i with $1 \le i \le n-1$ and $\mathscr{A} \subseteq \{1, 2, \dots, n-1\}$ define

$$m_{i,\mathscr{A}} := p_i \cdot p_{\mathscr{A}} - p_i(w_{\mathscr{A}}) \cdot p_{\mathscr{A}} - \sum_{\substack{\mathscr{A}\subseteq\mathscr{B} \\ |\mathscr{B}|=|\mathscr{A}|+1}} c^{\mathscr{B}}_{i,\mathscr{A}} \cdot p_{\mathscr{B}},$$

thought of as an element in $\mathbb{C}[t, \{p_{\mathscr{A}}\}_{\mathscr{A}\subseteq\{1,2,\dots,n-1\}}]$, where the $c^{\mathscr{B}}_{i,\mathscr{A}} \in \mathbb{C}[t]$ are the coefficients computed in Theorem 3.1. Motivated by the Giambelli formula we also define the following elements in $\mathbb{C}[t, p_1, p_2, \dots, p_{n-1}]$:

$$q_{i,\mathscr{A}} := p_i\cdot\sigma(\mathscr{A})\cdot\Bigl(\prod_{j\in\mathscr{A}} p_j\Bigr) - p_i(w_{\mathscr{A}})\cdot\sigma(\mathscr{A})\cdot\Bigl(\prod_{j\in\mathscr{A}} p_j\Bigr) - \sum_{\substack{\mathscr{A}\subseteq\mathscr{B} \\ |\mathscr{B}|=|\mathscr{A}|+1}} c^{\mathscr{B}}_{i,\mathscr{A}}\cdot\sigma(\mathscr{B})\Bigl(\prod_{k\in\mathscr{B}} p_k\Bigr),$$

where $\sigma(\mathscr{A}) \in \mathbb{Q}$ is the constant defined in (3-11).

Example 3.7. Let $n = 4$ and $i = 1$ and $\mathscr{A} = \{1, 2\}$. Consider

$$m_{1,\{1,2\}} = p_1 p_{v_{\{1,2\}}} - 2t\, p_{v_{\{1,2\}}} + p_{v_{\{1,2,3\}}}.$$

Expanding in terms of the Giambelli formula, we obtain

$$q_{1,\{1,2\}} = \tfrac{1}{2} p_1^2 p_2 - 2t \cdot \left(\tfrac{1}{2} p_1 p_2\right) + \tfrac{1}{6} p_1 p_2 p_3 = t\, p_1 p_2 + \tfrac{1}{6} p_1 p_2 p_3.$$

The main theorem of this section gives a ring presentation of $H^*_{S^1}(Y)$ using fewer generators and fewer relations than that in Theorem 3.6. More specifically let $\mathscr{K}$ denote the ideal in $\mathbb{C}[t, p_1, \dots, p_{n-1}]$ generated by the $q_{i,\mathscr{A}}$ for which $i \notin \mathscr{A}$, that is,

$$\begin{aligned} \mathscr{K} := \big\langle q_{i,\mathscr{A}} \bigm| 1 \le i \le n-1, \mathscr{A} \subseteq \{1, 2, \dots, n-1\}, i \notin \mathscr{A} \big\rangle \\ \subseteq \mathbb{C}[t, p_1, \dots, p_{n-1}]. \end{aligned} \tag{3-12}$$

Theorem 3.8. *Fix n a positive integer. Let Y be the Peterson variety in $\mathscr{F}\ell ags(\mathbb{C}^n)$ equipped with the action of the S^1 in (2-2). Then the S^1-equivariant cohomology $H^*_{S^1}(Y)$ is isomorphic to the ring*

$$\mathbb{C}[t, p_1, p_2, \dots, p_{n-1}]/\mathscr{K}$$

where $\mathscr{K}$ is the ideal in (3-12).

To prove the theorem we need the following lemma.

Lemma 3.9. *Let $i \in \{1, 2, \dots, n-1\}$ and $\mathscr{A} \subseteq \{1, 2, \dots, n-1\}$. Suppose $i \notin \mathscr{A}$. Then $q_{i,\mathscr{A}} = 0$ in $\mathbb{C}[t, p_1, p_2, \dots, p_{n-1}]$.*

Proof. Since $i \notin \mathscr{A}$ by assumption, Lemma 3.3 implies that

$$m_{i,\mathscr{A}} = p_i \cdot p_{\mathscr{A}} - p_i(w_{\mathscr{A}}) \cdot p_{\mathscr{A}} - \sum_{\substack{\mathscr{A} \subseteq \mathscr{B} \\ |\mathscr{B}| = |\mathscr{A}|+1}} c^{\mathscr{B}}_{i,\mathscr{A}} \cdot p_{\mathscr{B}}$$

simplifies to

$$m_{i,\mathscr{A}} = p_i \cdot p_{\mathscr{A}} - c^{\mathscr{A} \cup \{i\}}_{i,\mathscr{A}} \cdot p_{\mathscr{A} \cup \{i\}}. \tag{3-13}$$

Thus in order to compute the corresponding $q_{i,\mathscr{A}}$ it remains to compute $c^{\mathscr{A} \cup \{i\}}_{i,\mathscr{A}}$ and apply the Giambelli formula.

Let $\mathscr{A} = [a_1, a_2] \cup [a_3, a_4] \cup \cdots \cup [a_{m-1}, a_m]$ be the decomposition of $\mathscr{A}$ into maximal consecutive substrings. Consider the decomposition of $\mathscr{A} \cup \{i\}$ into maximal consecutive substrings. There are several cases to consider:

(1) The singleton set $\{i\}$ is a maximal consecutive substring of $\mathscr{A} \cup \{i\}$, that is, $i - 1 \notin \mathscr{A}$ and $i + 1 \notin \mathscr{A}$.

(2) The inclusion of i extends a maximal consecutive substring to its right by 1 element, that is, there exists a maximal consecutive string $[a_\ell, a_{\ell+1}] \subseteq \mathscr{A}$ such that $i = a_{\ell+1} + 1$ and that $[a_\ell, i]$ is a maximal consecutive substring of $\mathscr{A} \cup \{i\}$.

(3) The inclusion of i extends a maximal consecutive substring to its left by 1 element, that is, there exists a maximal consecutive string $[a_\ell, a_{\ell+1}] \subseteq \mathscr{A}$ such that $i = a_\ell - 1$ and that $[i, a_{\ell+1}]$ is a maximal consecutive substring of $\mathscr{A} \cup \{i\}$.

(4) The inclusion of i glues together two maximal consecutive substrings of $\mathscr{A}$, that is, there exist two maximal consecutive substrings $[a_\ell, a_{\ell+1}]$, $[a_{\ell+2}, a_{\ell+3}]$ such that $i = a_{\ell+1} + 1 = a_{\ell+2} - 1$ and hence $[a_\ell, a_{\ell+3}] = [a_\ell, a_{\ell+1}] \cup \{i\} \cup [a_{\ell+2}, a_{\ell+3}]$ is a maximal consecutive substring of $\mathscr{A} \cup \{i\}$.

We consider each case separately.

Case (1): Suppose $\{i\}$ is a maximal consecutive substring in $\mathscr{A} \cup \{i\}$. In this case, the coefficient $c_{i,\mathscr{A}}^{\mathscr{A}\cup\{i\}}$ is 1. Hence we have

$$m_{i,\mathscr{A}} = p_i \, p_{\mathscr{A}} - p_{v_{\mathscr{A}\cup\{i\}}}.$$

Since $\{i\}$ is a maximal consecutive substring in $\mathscr{A} \cup \{i\}$, we have $\sigma(\mathscr{A}) = \sigma(\mathscr{A} \cup \{i\})$. We conclude that

$$q_{i,\mathscr{A}} = p_i \cdot \left(\sigma(\mathscr{A}) \cdot \left(\prod_{j \in \mathscr{A}} p_j \right) \right) - \sigma(\mathscr{A} \cup \{i\}) \cdot \left(\prod_{j \in \mathscr{A}\cup\{i\}} p_j \right) = 0,$$

as desired.

Cases (2) and (3) are very similar, so we only present the argument for Case (2). Suppose i extends a maximal consecutive substring $[a_\ell, a_{\ell+1}]$ of $\mathscr{A}$ to its right. Then

$$m_{i,\mathscr{A}} = p_i \cdot p_{\mathscr{A}} - (i - a_\ell + 1) p_{v_{\mathscr{A}\cup\{i\}}},$$

since $k = i = \mathscr{H}_{\mathscr{B}}(i)$ and $\mathscr{T}_{\mathscr{B}}(i) = a_\ell$ so $c_{i,\mathscr{A}}^{\mathscr{A}\cup\{i\}} = i - a_\ell + 1$. We compute

$$\begin{aligned} q_{i,\mathscr{A}} = p_i &\left(\left(\prod_{\substack{1 \le s \le m-1 \\ s \text{ odd}}} \frac{1}{(a_{s+1} - a_s + 1)!} \right) \cdot \left(\prod_{j \in \mathscr{A}} p_j \right) \right) \\ &- (i - a_\ell + 1) \cdot \left(\prod_{\substack{1 \le s \le m-1 \\ s \text{ odd}, \, s \neq \ell}} \frac{1}{(a_{s+1} - a_s + 1)!} \right) \cdot \left(\frac{1}{(i - a_\ell + 1)!} \right) \cdot \left(\prod_{j \in \mathscr{A}\cup\{i\}} p_j \right), \end{aligned}$$

where one of the factors in the product in the second expression has changed because the maximal consecutive string $[a_\ell, a_{\ell+1}]$ has been extended in $\mathscr{A} \cup \{i\}$. Since

$$(i - a_\ell + 1) \left(\frac{1}{(i - a_\ell + 1)!} \right) = \frac{1}{(a_{\ell+1} - a_\ell + 1)!}$$

by assumption on i, we conclude $q_{i,\mathscr{A}} = 0$ as desired.

Case (4). Here the inclusion of i glues together two maximal consecutive substrings $[a_\ell, a_{\ell+1}]$, $[a_{\ell+2}, a_{\ell+3}]$ in $\mathscr{A}$. We then have $k = i$, $\mathscr{H}_{\mathscr{B}}(i) = a_{\ell+3}$, $\mathscr{T}_{\mathscr{B}}(i) = a_\ell$. Hence the coefficient $c_{i,\mathscr{A}}^{\mathscr{A}\cup\{i\}}$ is

$$c_{i,\mathscr{A}}^{\mathscr{A}\cup\{i\}} = (a_{\ell+3} - i + 1)\binom{a_{\ell+3} - a_\ell + 1}{i - a_\ell} = \frac{(a_{\ell+3} - a_\ell + 1)!}{(i - a_\ell)!\,(a_{\ell+3} - i)!}.$$

The expansion of $p_i \cdot p_{\mathscr{A}}$ is the same as in the previous cases. The term corresponding to $c_{i,\mathscr{A}}^{\mathscr{A}\cup\{i\}} \cdot p_{\mathscr{A}\cup\{i\}}$ is

$$\frac{(a_{\ell+3}-a_\ell+1)!}{(i-a_\ell)!\,(a_{\ell+3}-i)!} \cdot \Biggl(\prod_{\substack{1\le s\le m-1\\ s \text{ odd and } s\neq \ell,\ell+2}} \frac{1}{(a_{s+1}-a_s+1)!}\Biggr)\cdot\Biggl(\frac{1}{(a_{\ell+3}-a_\ell+1)!}\Biggr)\cdot\Biggl(\prod_{j\in\mathscr{A}\cup\{i\}} p_j\Biggr).$$

Since by assumption on i we have $i = a_{\ell+1} + 1 = a_{\ell+2} - 1$, we obtain the simplification

$$\frac{(a_{\ell+3} - a_\ell + 1)!}{(i - a_\ell)!\,(a_{\ell+3} - i)!}\Biggl(\frac{1}{(a_{\ell+3} - a_\ell + 1)!}\Biggr) = \frac{1}{(i - a_\ell)!\,(a_{\ell+3} - i)!} = \frac{1}{(a_{\ell+1} - a_\ell + 1)!} \cdot \frac{1}{(a_{\ell+3} - a_{\ell+2} + 1)!}$$

from which it follows that $q_{i,\mathscr{A}} = 0$ also in this case. The result follows. □

Example 3.10. Let $n = 5$, $i = 4$ and let $\mathscr{A} = \{1, 2\}$. Consider

$$m_{4,\{1,2\}} = p_4 \cdot p_{v_{\{1,2\}}} - c_{4,\{1,2\}}^{\{1,2,4\}} \cdot p_{v_{\{1,2,4\}}}.$$

From (3-3) it follows that $c_{4,\{1,2\}}^{\{1,2,4\}} = 1$. The corresponding $q_{4,\{1,2\}}$ can be computed to be

$$q_{4,\{1,2\}} = p_4\left(\tfrac{1}{2!}p_1p_2\right) - \left(\tfrac{1}{2!}p_1p_2\right)p_4 = 0.$$

Proof of Theorem 3.8. By Theorem 3.6 we know that

$$H^*_{S^1}(Y) \cong \mathbb{C}[t, \{p_{\mathscr{A}}\}_{\mathscr{A}\subseteq\{1,2,\dots,n-1\}}]/\mathscr{J},$$

where $\mathscr{J}$ is the ideal generated by the relations (3-2) so we wish to prove

$$\mathbb{C}[t, p_1, \dots, p_{n-1}]/\mathscr{K} \cong \mathbb{C}[t, \{p_{\mathscr{A}}\}_{\mathscr{A}\subseteq\{1,2,\dots,n-1\}}]/\mathscr{J}.$$

The content of the Giambelli formula (Theorem 3.2) is that the expressions

$$p_{\mathscr{A}} - \sigma(\mathscr{A})\prod_{j\in\mathscr{A}} p_j$$

are elements of $\mathcal{J}$. Hence

$$\begin{aligned}\mathcal{J} &= \big\langle m_{i,\mathcal{A}} \;\big|\; 1 \le i \le n-1, \mathcal{A} \subseteq \{1,2,\dots,n-1\}\big\rangle \\ &\qquad + \Big\langle p_{\mathcal{A}} - \sigma(\mathcal{A}) \prod_{j\in\mathcal{A}} p_j \;\Big|\; 1 \le i \le n-1, \mathcal{A} \subseteq \{1,2,\dots,n-1\}\Big\rangle \\ &= \big\langle q_{i,\mathcal{A}} \;\big|\; 1 \le i \le n-1, \mathcal{A} \subseteq \{1,2,\dots,n-1\}\big\rangle \\ &\qquad + \Big\langle p_{\mathcal{A}} - \sigma(\mathcal{A}) \prod_{j\in\mathcal{A}} p_j \;\Big|\; 1 \le i \le n-1, \mathcal{A} \subseteq \{1,2,\dots,n-1\}\Big\rangle.\end{aligned}$$

We therefore have

$$\frac{\mathbb{C}[t, \{p_{\mathcal{A}}\}_{\mathcal{A}\subseteq\{1,2,\dots,n-1\}}]}{\mathcal{J}} \cong \frac{\mathbb{C}[t, p_1, \dots, p_{n-1}]}{\big\langle q_{i,\mathcal{A}} \;\big|\; 1 \le i \le n-1, \mathcal{A} \subseteq \{1,2,\dots,n-1\}\big\rangle};$$

but since $q_{i,\mathcal{A}} = 0$ if $i \notin \mathcal{A}$ by Lemma 3.9 we conclude that

$$\begin{aligned}&\big\langle q_{i,\mathcal{A}} \;\big|\; 1 \le i \le n-1, \mathcal{A} \subseteq \{1,2,\dots,n-1\}\big\rangle \\ &\qquad = \big\langle q_{i,\mathcal{A}} \;\big|\; 1 \le i \le n-1, \mathcal{A} \subseteq \{1,2,\dots,n-1\} \text{ and } i \notin \mathcal{A}\big\rangle,\end{aligned}$$

from which the result follows. □

Example 3.11. Let $n = 4$ and Y the Peterson variety in $\mathcal{F}\ell ags(\mathbb{C}^4)$. The degree-2 multiplicative generators are p_1, p_2, and p_3. Then the statement of Theorem 3.8 yields a presentation of the equivariant cohomology ring of Y as

$$H^*_{S^1}(Y) \cong \mathbb{C}[t, p_1, p_2, p_3]/\mathcal{K},$$

where $\mathcal{K}$ is the ideal generated by the following 12 elements:

$$\begin{gathered}2p_1^2 - 2t\,p_1 - p_1p_2, \\ 2p_2^2 - 2t\,p_2 - p_1p_2 - p_2p_3, \\ 2p_3^2 - 2t\,p_3 - p_2p_3, \\ 3p_1^2p_2 - 6t\,p_1p_2 - p_1p_2p_3, \\ 3p_1p_2^2 - 6t\,p_1p_2 - 2p_1p_2p_3, \\ 2p_1^2p_3 - 2t\,p_1p_3 - p_1p_2p_3, \\ 2p_1p_3^2 - 2t\,p_1p_3 - p_1p_2p_3, \\ 3p_2^2p_3 - 6t\,p_2p_3 - 2p_1p_2p_3, \\ 3p_2p_3^2 - 6t\,p_2p_3 - p_1p_2p_3, \\ p_1^2p_2p_3 - 3t\,p_1p_2p_3, \\ p_1p_2^2p_3 - 4t\,p_1p_2p_3, \\ p_1p_2p_3^2 - 3t\,p_1p_2p_3.\end{gathered}$$

This list is not minimal: for instance, one can immediately see the sixth and seventh expressions in this list are multiples of the first and third ones, so evidently they are unnecessary for defining the ideal $\mathcal{K}$. In fact, more is true: a Macaulay 2 computation shows that the ideal $\mathcal{K}$ is in fact generated by just the *quadratic* relations, that is, the first three elements in the above list. (We thank the referee for pointing this out.) Note that the original presentation given in Theorem 3.6 uses 8 generators and 24 relations, so this discussion shows that our presentation indeed gives a simplification of the description of the ring.

Remark 3.12. We thank the referee for the following comment. Based on our Giambelli formula, Theorem 3.8, and the example of $n = 4$ discussed above, it seems natural to conjecture that for any value of n, the corresponding ideal $\mathcal{K}$ is generated by just the quadratic relations. Using Macaulay 2, we have verified that the conjecture holds for a range of small values of n, but we were unable to give a proof for the general case. If the conjecture is true, then it would be a very significant simplification of the presentation of this ring and would lead to many interesting geometric and combinatorial questions.

4. Stirling numbers of the second kind

In this section we prove that Stirling numbers of the second kind appear in the multiplicative structure of the ring $H^*_{S^1}(Y)$. We learned this result from H. Naruse and do not claim originality, though the proof given is our own. The *Stirling number of the second kind*, which we denote $S(n, k)$, counts the number of ways to partition a set of n elements into k nonempty subsets (see, e.g., [Knuth 1975, Section 1.2.6]). For example, $S(3, 2)$ is the number of ways to put balls labeled 1, 2, and 3 into two identical boxes such that each box contains at least one ball. It is then easily seen that $S(3, 2) = 3$.

Theorem 4.1. *Fix a positive integern. Let Y be the Peterson variety in $\mathcal{F}\ell ags(\mathbb{C}^n)$ equipped with the action of the S^1 in (2-2). For $\mathcal{A} \subseteq \{1, 2, \dots, n-1\}$, let $v_{\mathcal{A}}$, $p_{\mathcal{A}}$ be as in Theorem 3.6. The following equality holds in $H^*_{S^1}(Y)$ for any k with $1 \le k \le n-1$:*

$$p_1^k = \sum_{j=1}^{k} S(k, j) t^{k-j} p_{v_{[1,j]}}. \tag{4-1}$$

Proof. We proceed by induction on k. Consider the base case $k = 1$. Then (4-1) becomes the equality

$$p_1 = S(1, 1) p_1.$$

Here $S(1, 1)$ is the number of ways to put 1 ball into 1 box, so $S(1, 1) = 1$ and the claim follows.

Now assume that (4-1) holds for k. We need to show that it also holds for $k+1$, that is,

$$p_1^{k+1} = \sum_{j=1}^{k+1} S(k+1, j) t^{k+1-j} \, p_{v_{[1,j]}}.$$

By the inductive hypothesis this is equivalent to showing that

$$\sum_{i=1}^{k} S(k, i) t^{k-i} p_1 p_{v_{[1,i]}} = \sum_{j=1}^{k+1} S(k+1, j) t^{k+1-j} \, p_{v_{[1,j]}}. \tag{4-2}$$

We now expand the left-hand side using the Monk formula. For each i it can be computed that

$$p_1 \, p_{v_{[1,i]}} = it \, p_{v_{[1,i]}} + p_{v_{[1,i+1]}}$$

where we have used [Harada and Tymoczko 2011, Lemma 6.4] to compute $p_1(w_{[1,i]})$. Therefore

$$\begin{aligned}
&\sum_{i=1}^{k} S(k,i) t^{k-i} p_1 p_{v_{[1,i]}} \\
&= \sum_{i=1}^{k} S(k,i) t^{k-i} (it \, p_{v_{[1,i]}} + p_{v_{[1,i+1]}}) \\
&= \sum_{i=1}^{k} i \, S(k,i) t^{k+1-i} p_{v_{[1,i]}} + \sum_{i=1}^{k} S(k,i) t^{k-i} p_{v_{[1,i+1]}} \\
&= S(k,1) t^k p_1 + \sum_{i=2}^{k} i \, S(k,i) t^{k+1-i} p_{v_{[1,i]}} + \sum_{i=1}^{k-1} S(k,i) t^{k-i} p_{v_{[1,i+1]}} + S(k,k) p_{v_{[1,k+1]}} \\
&= S(k,1) t^k p_1 + \sum_{i=2}^{k} i \, S(k,i) t^{k+1-i} p_{v_{[1,i]}} + \sum_{i=2}^{k} S(k,i-1) t^{k+1-i} p_{v_{[1,i]}} + S(k,k) p_{v_{[1,k+1]}} \\
&= S(k+1,1) t^k p_1 + \sum_{i=2}^{k} (i \, S(k,i) + S(k,i-1)) t^{k+1-i} p_{v_{[1,i]}} + S(k+1,k+1) p_{v_{[1,k+1]}} \\
&= S(k+1,1) t^k p_1 + \sum_{i=2}^{k} S(k+1,j) t^{k+1-i} p_{v_{[1,i]}} + S(k+1,k+1) p_{v_{[1,k+1]}} \\
&= \sum_{j=1}^{k+1} S(k+1,j) t^{k+1-j} p_{v_{[1,j]}}
\end{aligned}$$

where we have used the recurrence relation

$$S(k+1, j) = j S(k, j) + S(k, j-1)$$

for Stirling numbers and the fact that

$$S(k,1)=S(k,k)=S(k+1,1)=S(k+1,k+1)=1$$

for any k. The result follows. □

5. Stability of Peterson Schubert classes

We now observe that the Peterson Schubert classes $\{p_{\mathscr{A}}\}$ for the Peterson varieties satisfy a stability property for varying n, similar to that satisfied by the classical equivariant Schubert classes. This is an observation we learned from H. Naruse; we do not claim originality. For this section only, for a fixed positive integer n we denote by Y_n the Peterson variety in $\mathscr{F}\ell ags(\mathbb{C}^n)$.

Let $X_{w,n} \subseteq \mathscr{F}\ell ags(\mathbb{C}^n)$ denote the *Schubert variety* corresponding to $w \in S_n$ in $\mathscr{F}\ell ags(\mathbb{C}^n)$. By the standard inclusion of groups $S_n \hookrightarrow S_{n+1}$, we may also consider w to be an element in S_{n+1}. Furthermore there is a natural T^n-equivariant inclusion $\iota_n : \mathscr{F}\ell ags(\mathbb{C}^n) \hookrightarrow \mathscr{F}\ell ags(\mathbb{C}^{n+1})$ induced by the inclusion of the coordinate subspace $\mathbb{C}^n$ into $\mathbb{C}^{n+1}$. Then with respect to ι_n the Schubert variety $X_{w,n}$ maps isomorphically onto the corresponding Schubert variety $X_{w,n+1}$. Since the equivariant Schubert classes are cohomology classes corresponding to the Schubert varieties, this implies that for any $w \in S_n$ there exists an infinite sequence of Schubert classes $\{\sigma_{w,m}\}_{m=n}^{\infty}$ which lift the classes $\sigma_{w,n} \in H^*_{T^n}(\mathscr{F}\ell ags(\mathbb{C}^n))$, that is,

$$\begin{tikzcd}
\cdots \arrow[r] & H^*_{T^n}(\mathscr{F}\ell ags(\mathbb{C}^{n+2})) \arrow[r] & H^*_{T^n}(\mathscr{F}\ell ags(\mathbb{C}^{n+1})) \arrow[r] & H^*_{T^n}(\mathscr{F}\ell ags(\mathbb{C}^{n})) \\
\cdots \arrow[r, mapsto] & \sigma_{w,n+2} \arrow[r, mapsto] & \sigma_{w,n+1} \arrow[r, mapsto] & \sigma_{w,n}
\end{tikzcd} \tag{5-1}$$

and furthermore for any $v \in S_n$ and any $m \geq n$, the restriction $\sigma_{w,m}(v)$ is equal to $\sigma_{w,n}(v)$. The theorem below asserts that a similar statement holds for Peterson Schubert classes. Observe that the inclusion $\iota_n : \mathscr{F}\ell ags(\mathbb{C}^n) \hookrightarrow \mathscr{F}\ell ags(\mathbb{C}^{n+1})$ mentioned above also induces a natural inclusion $j_n : Y_n \hookrightarrow Y_{n+1}$ since the principal nilpotent operator on $\mathbb{C}^{n+1}$ preserves the coordinate subspace $\mathbb{C}^n$. Moreover, since the central circle subgroup of $U(n,\mathbb{C})$ acts trivially on $\mathscr{F}\ell ags(\mathbb{C}^n)$ for any n, the inclusion j_n is equivariant with respect to the S^1-actions on Y_n and Y_{n+1} given by the two circle subgroups defined by (2-2) in $U(n,\mathbb{C})$ and $U(n+1,\mathbb{C})$ respectively. Thus there is a pullback homomorphism $j_n^* : H^*_{S^1}(Y_{n+1}) \to H^*_{S^1}(Y_n)$ analogous to the map $\iota_n : H^*_{T^n}(\mathscr{F}\ell ags(\mathbb{C}^{n+1})) \to H^*_{T^n}(\mathscr{F}\ell ags(\mathbb{C}^n))$ above.

Theorem 5.1. *Fix a positive integer n. Let Y_n denote the Peterson variety in $\mathscr{F}\ell ags(\mathbb{C}^n)$ equipped with the natural S^1-action defined by* (2-2)*. For $w \in S_n$ let $p_{w,n} \in H^*_{S^1}(Y_n)$ denote the Peterson Schubert class corresponding to w. Then the natural inclusions $j_m : Y_m \hookrightarrow Y_{m+1}$ for $m \geq n$ induce a sequence of homomorphisms $j_m^* : H^*_{S^1}(Y_{m+1}) \to H^*_{S^1}(Y_m)$ such that $j_m^*(p_{w,m+1}) = p_{w,m}$, that is,* *there exists a infinite sequence of Peterson Schubert classes $\{p_{w,m}\}_{m=n}^{\infty}$ that lift $p_{w,n} \in H^*_{T^n}(\mathcal{F}\ell ags(\mathbb{C}^n))$:*

$$\begin{tikzcd} \cdots \arrow[r] & H^*_{S^1}(Y_{n+2}) \arrow[r] & H^*_{S^1}(Y_{n+1}) \arrow[r] & H^*_{S^1}(Y_n) \\ \cdots \arrow[r, mapsto] & p_{w,n+2} \arrow[r, mapsto] & p_{w,n+1} \arrow[r, mapsto] & p_{w,n} \end{tikzcd} \tag{5-2}$$

Moreover, for any $v \in Y_n^{S^1}$ and any $m \geq n$, the restriction $p_{w,m}(v)$ equals $p_{w,n}(v)$.

Proof. By naturality and the definition of Peterson Schubert classes $p_{w,n} \in H^*_{S^1}(Y_n)$ as the images of $\sigma_{w,n}$, it is immediate that (5-1) can be expanded to a commutative diagram

$$\begin{tikzcd} \cdots \arrow[r] & H^*_{T^n}(\mathcal{F}\ell ags(\mathbb{C}^{n+2})) \arrow[r, "\iota^*_{n+1}"] \arrow[d] & H^*_{T^n}(\mathcal{F}\ell ags(\mathbb{C}^{n+1})) \arrow[r, "\iota^*_n"] \arrow[d] & H^*_{T^n}(\mathcal{F}\ell ags(\mathbb{C}^{n})) \arrow[d] \\ \cdots \arrow[r] & H^*_{S^1}(Y_{n+2}) \arrow[r, "j^*_{n+1}"'] & H^*_{S^1}(Y_{n+1}) \arrow[r, "j^*_n"'] & H^*_{S^1}(Y_n) \end{tikzcd} \tag{5-3}$$

where the vertical arrows are the projection maps obtained by the composition of $H^*_{T^n}(\mathcal{F}\ell ags(\mathbb{C}^m)) \to H^*_{S^1}(\mathcal{F}\ell ags(\mathbb{C}^m))$ with $H^*_{S^1}(\mathcal{F}\ell ags(\mathbb{C}^m)) \to H^*_{S^1}(Y_m)$, for $m = n+2, n+1, n$. In particular, for any $w \in S_n$ and $m \geq n$, the vertical maps send $\sigma_{w,m}$ to $p_{w,m}$. The result follows. □

Acknowledgements

We are grateful to Hiroshi Naruse for communicating to us his observations on the stability of Peterson Schubert classes and the presence of Stirling numbers of the second kind. We thank Julianna Tymoczko for support of and interest in this project and Alex Yong for useful conversations. We also thank an anonymous referee for a careful reading of the manuscript and many helpful suggestions; in particular, the referee suggested the conjecture regarding quadratic relations recorded in Remark 3.12, as well as the question which we record in Remark 3.4.

References

[Fulton 1997] W. Fulton, *Young tableaux: with applications to representation theory and geometry*, London Mathematical Society Student Texts **35**, Cambridge University Press, Cambridge, 1997. MR 99f:05119 Zbl 0878.14034

[Harada and Tymoczko 2010] M. Harada and J. Tymoczko, "Poset pinball, GKM-compatible subspaces, and Hessenberg varieties", preprint, 2010. arXiv 1007.2750

[Harada and Tymoczko 2011] M. Harada and J. Tymoczko, "A positive Monk formula in the S^1-equivariant cohomology of type A Peterson varieties", *Proc. Lond. Math. Soc.* (3) **103**:1 (2011), 40–72. MR 2012f:14108 Zbl 1219.14065 arXiv 0908.3517

[Knuth 1975] D. E. Knuth, *The art of computer programming, 1: Fundamental algorithms*, 2nd ed., Addison-Wesley, Reading, MA, 1975. MR 51 #14624 Zbl 0895.68055

[Kostant 1996] B. Kostant, "Flag manifold quantum cohomology, the Toda lattice, and the representation with highest weight ρ", *Selecta Math.* (*N.S.*) **2**:1 (1996), 43–91. MR 97e:17029 Zbl 0868.14024

[Rietsch 2003] K. Rietsch, "Totally positive Toeplitz matrices and quantum cohomology of partial flag varieties", *J. Amer. Math. Soc.* **16**:2 (2003), 363–392. MR 2004d:14081 Zbl 1057.14065

[Sommers and Tymoczko 2006] E. Sommers and J. Tymoczko, "Exponents for B-stable ideals", *Trans. Amer. Math. Soc.* **358**:8 (2006), 3493–3509. MR 2007a:17016 Zbl 1105.20036

Received: 2011-01-02 Revised: 2011-09-12 Accepted: 2011-10-13

dbayegan@gmail.com

Department of Pure Mathematics and Mathematical Statistics, Centre for Mathematical Sciences,
University of Cambridge, Wilberforce Road,
Cambridge CB30WA, United Kingdom

megumi.harada@math.mcmaster.ca

Department of Mathematics and Statistics,
McMaster University, 1280 Main Street, West,
Hamilton, Ontario L8S4K1, Canada

INVOLVE 5:2 (2012)
dx.doi.org/10.2140/involve.2012.5.133

Weak Allee effect, grazing, and S-shaped bifurcation curves

Emily Poole, Bonnie Roberson and Brittany Stephenson

(Communicated by John Baxley)

We study a one-dimensional reaction-diffusion model arising in population dynamics where the growth rate is a weak Allee type. In particular, we consider the effects of grazing on the steady states and discuss the complete evolution of the bifurcation curve of positive solutions as the grazing parameter varies. We obtain our results via the quadrature method and Mathematica computations. We establish that the bifurcation curve is S-shaped for certain ranges of the grazing parameter. We also prove this occurrence of an S-shaped bifurcation curve analytically.

1. Introduction

For a given population, a linear relationship between the population's size and its per capita growth rate is often assumed. This correlation is known as logistic type growth and reflects that as a given population grows, its per capita growth rate declines linearly. However, it has been observed that for small population densities, the per capita growth rate increases rather than declines. Logistic growth cannot account for this initial increase, and an alternate model, dubbed the Allee effect [Allee 1938] must be invoked.

The general idea behind the Allee effect is that for small population densities, a variety of factors (such as a shortage of mates or predator saturation) result in an initial increase in the per capita growth rate. There are two types of Allee phenomena: strong Allee effect, whose per capita growth rate begins negative, and weak Allee effect, whose per capita growth rate is initially positive; these are usually modeled in the literature by quadratic functions of the population size. Thus, the

MSC2010: 34B15.

Keywords: ordinary differential equations, nonlinear boundary value problems, grazing, Allee effect, population dynamics.

A significant portion of this research work was completed during the 2010 NSF REU Program at the Center for Computational Sciences at Mississippi State University. Research was also supported by National Science Foundation Award DMS 0852032.

mathematical analysis of such models is considerably more challenging since the per capita growth rates are neither linear nor always nondecreasing.

When considering the long-term stability of a given population, it is insightful to study other factors affecting the population. By including an additional term that accounts for these natural phenomena, such as grazing, more accurate models can be obtained. Grazing is a type of predation in which an herbivore feeds from plant life. It is also similar to natural predation found in fish populations. The grazing term used in previous models, $cu^2/(1+u^2)$ (see [Van Nes and Scheffer 2005]), is known as the rate of grazing. Since the grazing population is constant, the term converges to c at high vegetation density levels. The effects of grazing have previously been studied with logistic growth, as in [Lee et al. 2011]. In the latter paper it was shown that, for certain ranges of the parameters involved (including c), the bifurcation curve of positive steady states is S-shaped.

Our primary motivation is to analyze the consequences of grazing on a weak Allee effect problem and on a strong Allee effect, in order to determine its effect on the steady state solutions.

Hence, we examine the structure of positive solutions of the steady state equations obtained from the reaction diffusion model,

$$u_t = \frac{1}{\lambda} u_{xx} + u\tilde{f}(u) - \frac{cu^2}{1+u^2} \quad \text{in } (0, 1),$$

with Dirichlet boundary conditions, namely

$$\begin{aligned} &-u'' = \lambda\left[u\tilde{f}(u) - \frac{cu^2}{1+u^2}\right] = \lambda f(u) \quad \text{in } (0, 1),\\ &u(0) = 0, \qquad u(1) = 0, \end{aligned}$$

where u is the population density, $\tilde{f}(u)$ is the per capita growth rate, $1/\lambda$ is the diffusion coefficient where $\lambda > 0$ is a constant, and $c \geq 0$ is also a constant.

Previous studies have analyzed positive solutions to Allee effect problems, both strong and weak (see [Shi and Shivaji 2006], for example), but to our knowledge no information is known about the combination of grazing with Allee effect. In this paper, we will analyze how the addition of a grazing term in combination with a weak Allee and also in combination with a strong Allee type affects the steady states in the one-dimensional problem. Our analysis is completed via the quadrature method [Brown et al. 1981; Laetsch 1970], which we will discuss in Section 2. In Sections 3–5, we provide a detailed analysis on the case when $\tilde{f}$ is weak Allee, i.e.,

$$\tilde{f}(u) = (u+1)(b-u), \quad b > 1.$$

In Section 3, we present some necessary analysis of the zeros of our nonlinearity, f, and in Section 4, we provide the complete evolution of the bifurcation curve of positive solutions via Mathematica computations. In particular, we obtain S-shaped bifurcation curves for certain ranges of parameters bifurcating from the nontrivial branch of solutions. We note here that for all parameter values $u \equiv 0$ is a solution of (1-1). In Section 5, we provide various analytical results, including a proof of the occurrence of such an S-shaped bifurcation curve. In Section 6, we study the case when $\tilde{f}$ represents a logistic growth rate, that is,

$$\tilde{f}(u) = (1 - bu)$$

and provide the evolution of the bifurcation curve as c varies. Next, Section 7 provides the evolution of the bifurcation curve for the case when $\tilde{f}$ is of strong Allee type. That is,

$$\tilde{f}(u) = (u-1)(b-u), \quad b > 1. \tag{1-1}$$

Unlike the weak Allee case and the logistic case, for the strong Allee case, we notice that the variation of c had little effect on the general structure of the bifurcation curve. In particular, no S-shaped bifurcation curve occurred for any parameter values.

Finally, in Section 8, we conclude the paper by considering the biological implications arising from our results. In particular, the ranges of conditional and unconditional persistence in terms of the diffusion coefficient as c varies will also be discussed. Interestingly, our analysis proves that for weak Allee effect growth models, when grazing is large, there exist no ranges of the diffusion coefficient for which conditional persistence exists.

2. Quadrature method

In this section, we recall results via the quadrature method developed by Laetsch [1970] and Brown, Ibrahim, and Shivaji [Brown et al. 1981] to analyze positive solutions to the boundary value problem

$$\begin{aligned} -u''(x) &= \lambda f(u(x)), \quad x \in (0,1), \\ u(0) &= 0, \\ u(1) &= 0, \end{aligned} \tag{2-1}$$

where $f : [0, \infty) \to (0, \infty)$ is a C^1 function and λ is a nonnegative parameter.

Lemma 2.1 [Laetsch 1970]. *Let u be a positive solution to (2-1) with $\|u\|_\infty = u(\frac{1}{2}) = \rho > 0$. Such a solution exists if and only if*

$$G(\rho) := \int_0^\rho \frac{dt}{\sqrt{F(\rho) - F(t)}} = \sqrt{\frac{\lambda}{2}}, \tag{2-2}$$

where

$$F(u) = \int_0^u f(s)\,ds.$$

Proof. ($\Rightarrow$) Since (2-1) is an autonomous differential equation, if u is a positive solution to (2-1) such that $u'(x_0)=0$ for some $x_0 \in (0, 1)$, then both $v(x) := u(x_0+x)$ and $w(x) := u(x_0 - x)$ satisfy the initial value problem

$$\begin{gathered} -z''(x) = \lambda f(z(x)), \quad x \in [0, d), \\ z(0) = u(x_0), \quad z'(0) = 0, \end{gathered}$$

where $d = \min\{x_0, 1 - x_0\}$. By Picard's existence and uniqueness theorem, we can infer that $u(x_0 + x) \equiv u(x_0 - x)$ for all $x \in [0, d)$. Thus, solutions of (2-1) must be symmetric around $x = \frac{1}{2}$, at which point u attains its maximum $\rho := u(\frac{1}{2})$. Multiplying the differential equation in (2-1) by $u'(x)$ gives

$$-\left(\frac{[u'(x)]^2}{2}\right)' = \lambda[F(u(x))]', \tag{2-3}$$

where $F(s) = \int_0^s f(z)\,dz$.

Integrating both sides, we derive

$$\frac{u'(x)}{\sqrt{F(\rho) - F(u(x))}} = \sqrt{2\lambda}, \quad x \in \left[0, \tfrac{1}{2}\right). \tag{2-4}$$

Integrating again, we obtain

$$\int_0^{u(x)} \frac{dt}{\sqrt{F(\rho) - F(t)}} = \sqrt{2\lambda}x, \quad x \in \left[0, \tfrac{1}{2}\right]. \tag{2-5}$$

Substituting $x = \frac{1}{2}$ into (2-5) and using $u(\frac{1}{2}) = \rho$, we now have

$$G(\rho) := \int_0^{\rho} \frac{dt}{\sqrt{F(\rho) - F(t)}} = \sqrt{\frac{\lambda}{2}}. \tag{2-6}$$

Thus, if u is a solution of (2-1) with $\|u\|_\infty = \rho$, then ρ must satisfy the equation $G(\rho) = \sqrt{\lambda/2}$.

($\Leftarrow$) Now, if we have such a value for ρ, we can define our solution u through the equation

$$\int_0^{u(x)} \frac{dt}{\sqrt{F(\rho) - F(t)}} = \sqrt{2\lambda}x, \quad x \in \left[0, \tfrac{1}{2}\right].$$

By the implicit function theorem, u is differentiable; therefore,

$$u'(x) = \sqrt{2\lambda[F(\rho) - F(u(x))]}.$$

Differentiating again gives us

$$-u''(x) = \lambda f(u(x)).$$

Also, it is easy to see that $u(0) = 0$. Finally, defining $u(x)$ to be a symmetric solution, we have that u is a positive solution to (2-1) with $\|u\|_\infty = \rho$ if and only if $\sqrt{\lambda/2} = G(\rho)$. □

Remark 2.1. For values of ρ satisfying the following two conditions, the improper integral in (2-2) will be well-defined and convergent:

$$\begin{aligned} &f(\rho) > 0, \\ &F(\rho) > F(s) \quad \text{for all } s \in [0, \rho). \end{aligned}$$

The following lemma is taken from [Brown et al. 1981], and the proof makes critical use of Lebesgue's dominated convergence theorem to prove the existence of the integral in (2-7).

Lemma 2.2. *$G(\rho)$ is continuous and differentiable on the set*

$$\boldsymbol{S} := \{\rho > 0 \mid f(\rho) > 0 \textit{ and } F(\rho) - F(s) > 0 \textit{ for all } s \in [0, \rho)\};$$

moreover,

$$G'(\rho) = \int_0^1 \frac{H(\rho) - H(\rho v)}{[F(\rho) - F(\rho v)]^{\frac{3}{2}}} dv, \tag{2-7}$$

where

$$H(s) := F(s) - \frac{s}{2} f(s). \tag{2-8}$$

3. Preliminaries

We consider the following reaction term, which combines weak Allee effect and grazing:

$$f(u) = u(u+1)(b-u) - \frac{cu^2}{1+u^2} = \frac{u(u+1)(b-u)(1+u^2) - cu^2}{1+u^2},$$

where $b > 1$ and $c \geq 0$. The numerator of $f(u)$ is a fifth degree polynomial. Study of the roots of $f(u)$ reveals the existence of one negative root and one root at $u = 0$, regardless of the values chosen for b and c. On the contrary, the three remaining roots are dependent on the value of the constant c. These three roots fluctuate between real and imaginary values as c changes. Let σ represent the smallest positive root of $f(u)$ in all cases, and let σ_0 and σ_1 denote the two remaining roots. We must also note that a special case occurs for small values of b: for $b \in (1, b_0)$, (for some $b_0 > 1$), there is always exactly one positive real root of $f(u)$, σ. In

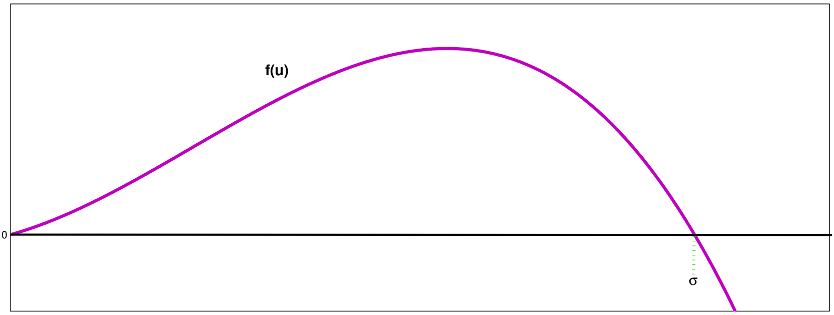


Figure 1. Graph of $f(u)$ with root at σ.

order for $G(\rho)$ to be defined, this variance demands that further analysis of $f(u)$ be completed case-wise (see Remark 2.1).

Remark 3.1. Based on our computations and aid from Mathematica, we conjecture that $b_0 \approx 2.852$.

If $b \in (b_0, \infty)$, the characteristic shape of $f(u)$ varies as the value of c changes. There exists $c_0 > 0$ so that for $c \in (0, c_0)$, $f(u)$ has only one real root, σ. In this case, $f(u)$ resembles Figure 1.

Correspondingly, in this case, $F(u)$ will take the form exemplified in Figure 2. As you can see, $G(\rho)$ will be well-defined for $\rho \in (0, \sigma)$. As c increases, the shape of $f(u)$ changes. There exists $c_1 > c_0$ so that for $c \in (c_0, c_1)$, $f(u)$ has exactly 3 real positive roots, (σ, σ_0, and σ_1). The shape of $f(u)$ is illustrated in Figure 3.

There exists $\hat{c}_1 < c_1$ such that for $c \in (c_0, \hat{c}_1)$, the graph of $F(u)$ resembles Figure 4. We let $\gamma \in (\sigma_0, \sigma_1)$, so that $F(\gamma) = F(\sigma)$. Recall from Remark 2.1 that, to guarantee that $G(\rho)$ is well-defined, we need $f(\rho) > 0$ as well as $F(\rho) > F(u)$ whenever $0 \leq u < \rho$. Thus, in this case, $G(\rho)$ will be viable only for $\rho \in (0, \sigma)$ and $\rho \in (\gamma, \sigma_1)$. (The boxed region in Figure 4 has been magnified in Figure 5.)

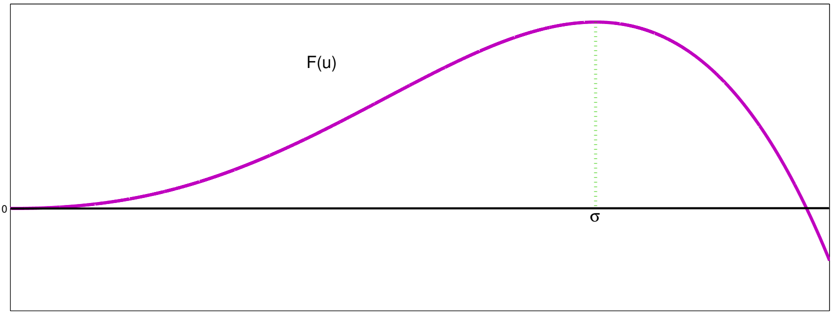


Figure 2. Graph of $F(u)$ for $c \in (0, c_0)$.

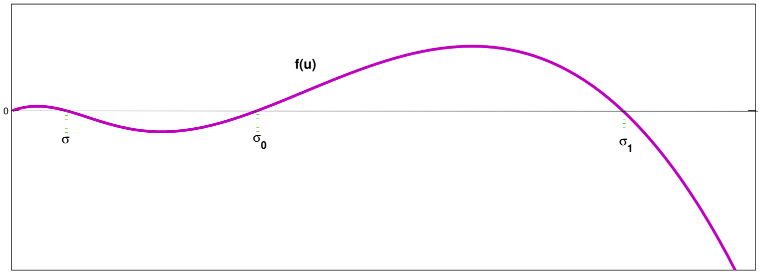


Figure 3. Graph of $f(u)$ with roots at σ, σ_0, and σ_1.

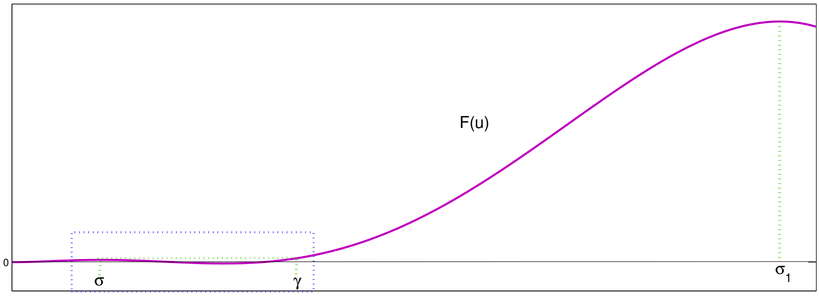


Figure 4. Graph of $F(u)$ for $c \in (c_0, \hat{c}_1)$.

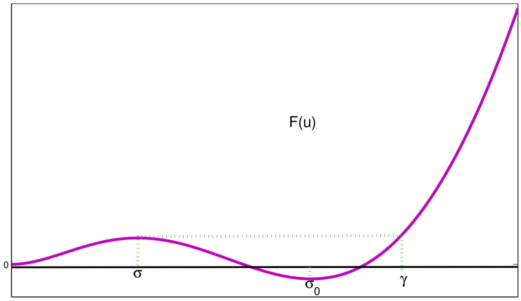


Figure 5. Magnified picture of the boxed region in Figure 4.

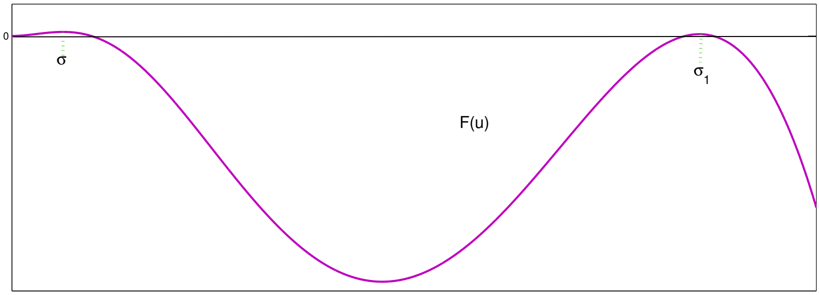


Figure 6. Graph of $F(u)$ for $c \in (\hat{c}_1, c_1)$.

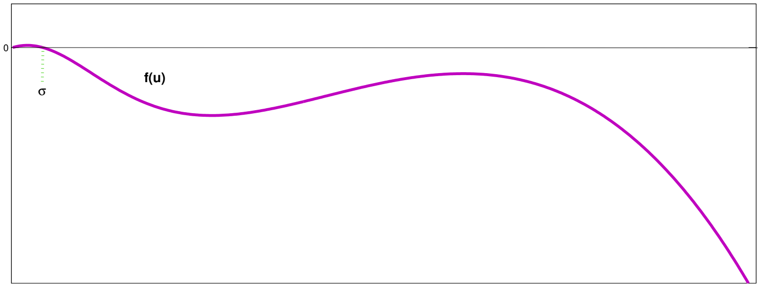


Figure 7. Graph of $f(u)$ with root at σ.

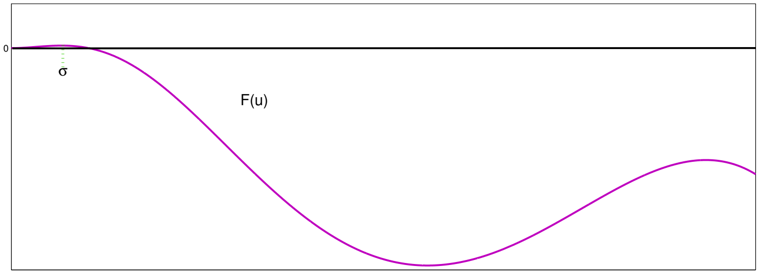


Figure 8. $F(u)$ for $c > c_1$.

The graph of $F(u)$ for $c \in (\hat{c}_1, c_1)$ is illustrated in Figure 6. Clearly, $G(\rho)$ in this instance is only well-defined for $\rho \in (0, \sigma)$. When c exceeds c_1, $f(u)$ is pulled downward and once again has only one real positive root. We will denote this root as σ while σ_0 and σ_1 are imaginary in this case. This is portrayed in Figure 7. Again, $G(\rho)$ is well-defined only for $\rho \in (0, \sigma)$. For $c > c_1$, $F(u)$ will take the form illustrated in Figure 8.

For each of these cases, the structure of positive solutions for (1-1) changes; thus, distinct bifurcation diagrams are obtained, as we now explain.

4. Computational results

In this section, we present the bifurcation diagrams for the weak Allee effect. Recalling Lemma 2.1, we obtained these results via Mathematica by plotting (2-2) for a fixed b-value over a range of c-values.

If $b \in (b_0, \infty)$, then there exist $c_0^*, c_1^*, c_2^* > 0$ such that:

1. If $c \in [0, c_0^*)$, there exist $\lambda_0 > 0$ and $\Lambda = \pi^2/f'(0)$ such that (2-1) has
   - no positive solution for $\lambda \in (0, \lambda_0)$,
   - exactly 1 positive solution for $\lambda = \lambda_0$,
   - exactly 2 positive solutions for $\lambda \in (\lambda_0, \Lambda)$, and
   - exactly 1 positive solution for $\lambda \in [\Lambda, \infty)$.

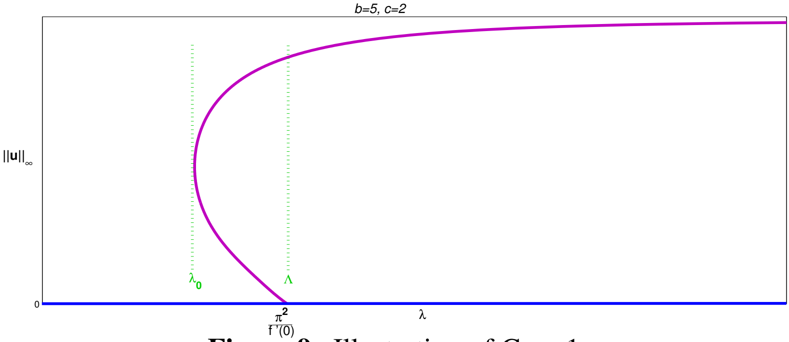


Figure 9. Illustration of Case 1.

(See illustration in Figure 9.)

2. If $c \in [c_0^*, c_1^*)$, there exist $\lambda_0, \lambda_1, \lambda_1, \lambda_2, \Lambda > 0$ such that (2-1) has
   - no positive solution for $\lambda \in (0, \lambda_0)$,
   - exactly 1 positive solution for $\lambda = \lambda_0$,
   - exactly 2 positive solutions for $\lambda \in (\lambda_0, \lambda_1)$,
   - exactly 3 positive solutions for $\lambda = \lambda_1$,
   - exactly 4 positive solutions for $\lambda \in (\lambda_1, \lambda_2)$,
   - exactly 3 positive solutions for $\lambda = \lambda_2$,
   - exactly 2 positive solutions for $\lambda \in (\lambda_2, \Lambda)$, and
   - exactly 1 positive solution for $\lambda \in [\Lambda, \infty)$.

   (See illustration in Figure 10.)

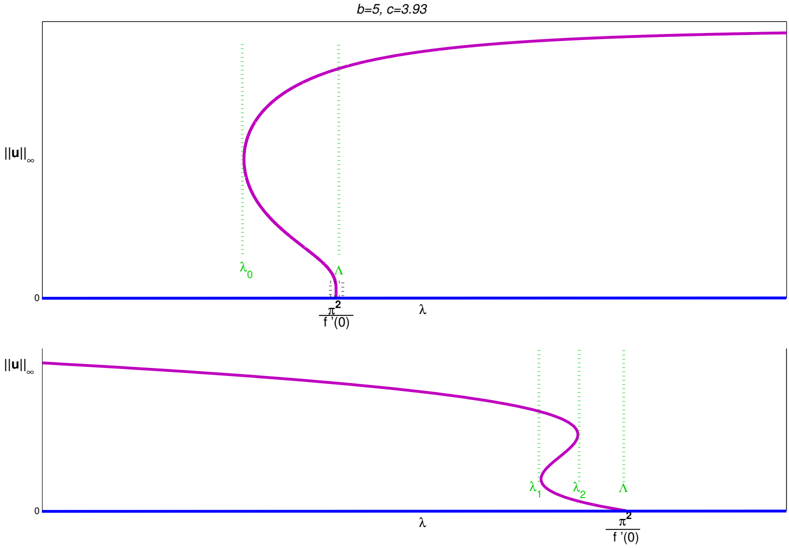


Figure 10. Illustration of Case 2. The bottom diagram shows the contents of the small dotted box under magnification.

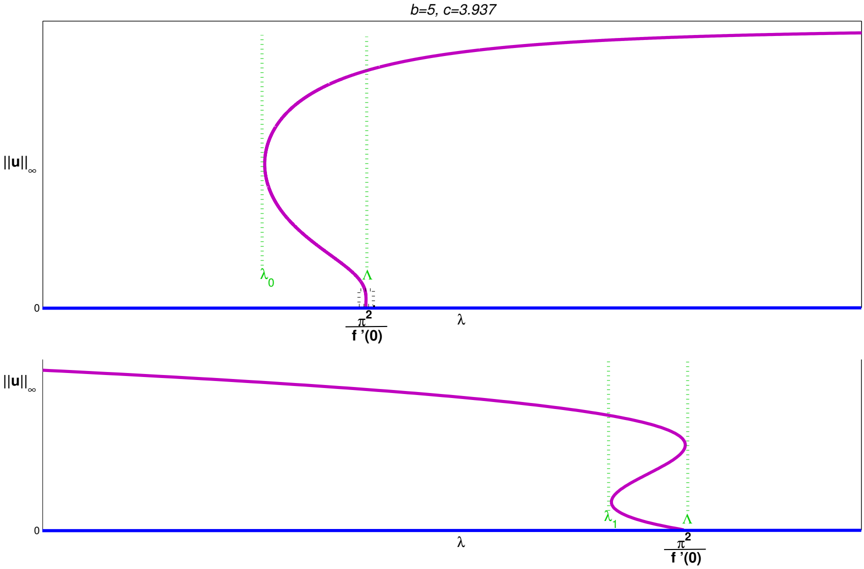


Figure 11. Illustration of Case 3, including magnified detail.

3. If $c = c_1^*$, there exist $\lambda_0, \lambda_1, \Lambda > 0$ such that (2-1) has
   - no positive solution for $\lambda \in (0, \lambda_0)$,
   - exactly 1 positive solution for $\lambda = \lambda_0$,
   - exactly 2 positive solutions for $\lambda \in (\lambda_0, \lambda_1)$,
   - exactly 3 positive solutions for $\lambda = \lambda_1$,
   - exactly 4 positive solutions for $\lambda \in (\lambda_1, \Lambda)$,
   - exactly 2 positive solutions for $\lambda = \Lambda$, and
   - exactly 1 positive solution for $\lambda \in (\Lambda, \infty)$.

   (See illustration in Figure 11.)

4. If $c = (c_1^*, c_2^* = b - 1)$, there exist $\lambda_0, \lambda_1, \lambda_2, \Lambda > 0$ such that (2-1) has
   - no positive solution for $\lambda \in (0, \lambda_0)$,
   - exactly 1 positive solutions for $\lambda = \lambda_0$,
   - exactly 2 positive solutions for $\lambda \in (\lambda_0, \lambda_1)$,
   - exactly 3 positive solutions for $\lambda = \lambda_1$,
   - exactly 4 positive solutions for $\lambda \in (\lambda_1, \Lambda)$,
   - exactly 3 positive solutions for $\lambda \in [\Lambda, \lambda_2)$,
   - exactly 2 positive solutions for $\lambda = \lambda_2$, and
   - exactly 1 positive solution for $\lambda \in (\lambda_2, \infty)$.

   (See illustration in Figure 12.)

5. If $c = c_2^* = b - 1$, there exist $\lambda_0, \Lambda > 0$ such that (2-1) has
   - no positive solution for $\lambda \in (0, \lambda_0)$,
   - exactly 1 positive solutions for $\lambda = \lambda_0$,

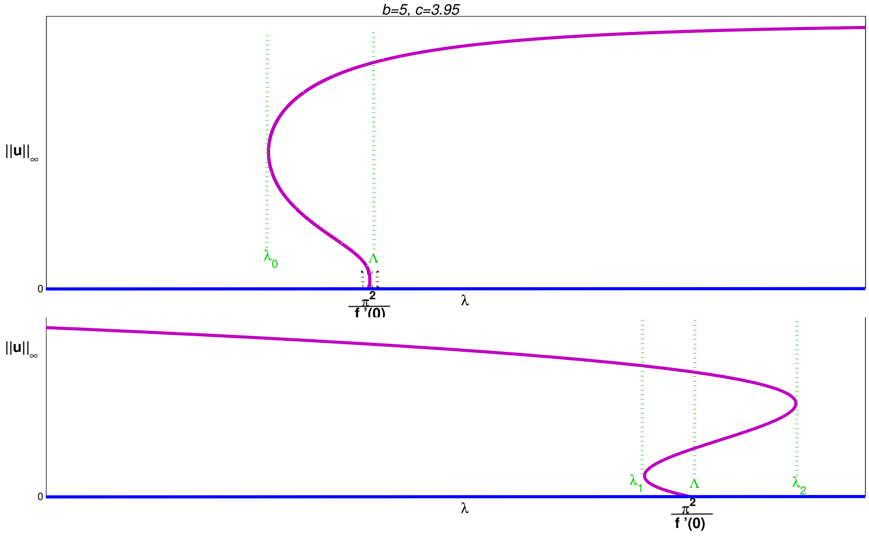


Figure 12. Illustration of Case 4, including magnified detail.

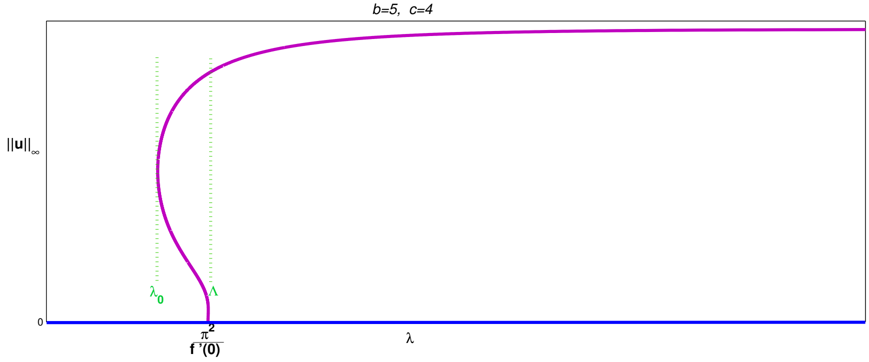


Figure 13. Illustration of Case 5.

- exactly 2 positive solutions for $\lambda \in (\lambda_0, \Lambda)$, and
- exactly 1 positive solutions for $\lambda \in [\Lambda, \infty)$.

(See illustration in Figure 13.)

6. If $c \in (c_2^* = b - 1, c_3^*)$, there exist $\lambda_0, \Lambda, \lambda_2 > 0$ such that (2-1) has
   - no positive solution for $\lambda \in (0, \lambda_0)$,
   - exactly 1 positive solution for $\lambda = \lambda_0$,
   - exactly 2 positive solutions for $\lambda \in (\lambda_0, \Lambda]$,
   - exactly 3 positive solutions for $\lambda \in (\Lambda, \lambda_2)$,
   - exactly 2 positive solutions for $\lambda = \lambda_2$, and
   - exactly 1 positive solution for $\lambda \in (\lambda_2, \infty)$.

   (See illustration in Figure 14.)

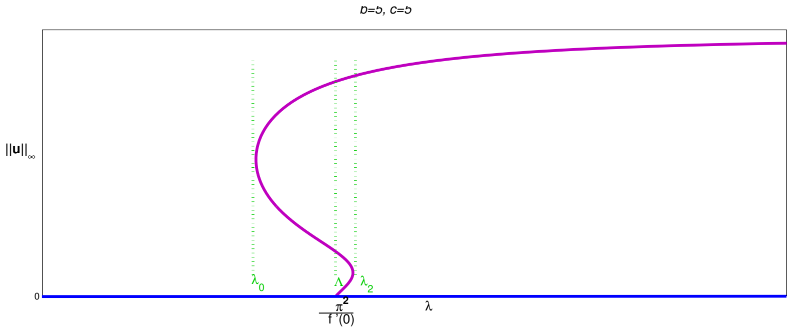


Figure 14. Illustration of Case 6.

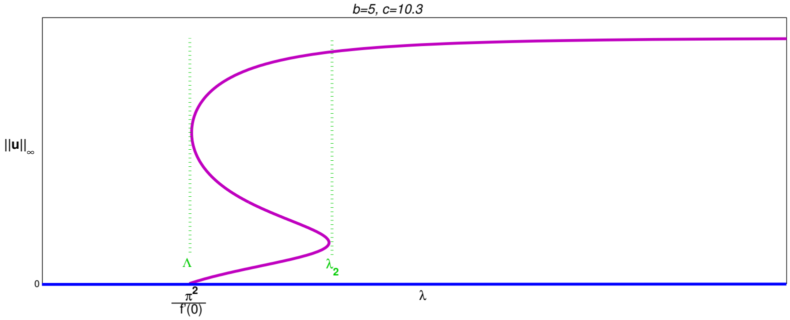


Figure 15. Illustration of Case 7.

7. If $c = c_3^*$, there exist $\Lambda, \lambda_2 > 0$ such that (2-1) has
   - no positive solution for $\lambda \in (0, \Lambda)$,
   - exactly 1 positive solution for $\lambda = \Lambda$,
   - exactly 3 positive solutions for $\lambda \in (\Lambda, \lambda_2)$,
   - exactly 2 positive solutions for $\lambda = \lambda_2$, and
   - exactly 1 positive solution for $\lambda \in (\lambda_2, \infty)$.

   (See illustration in Figure 15.)

8. If $c \in (c_3^*, c_4^*)$, there exist $\Lambda, \lambda_0, \lambda_2 > 0$ such that (2-1) has

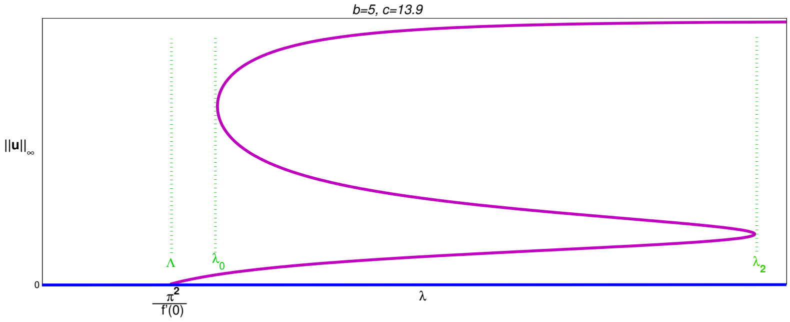


Figure 16. Illustration of Case 8.

- no positive solution for $\lambda \in (0, \Lambda]$,
- exactly 1 positive solution for $\lambda \in (\Lambda, \lambda_0)$,
- exactly 2 positive solutions for $\lambda = \lambda_0$,
- exactly 3 positive solutions for $\lambda \in (\lambda_0, \lambda_2)$,
- exactly 2 positive solutions for $\lambda = \lambda_2$, and
- exactly 1 positive solution for $\lambda \in (\lambda_2, \infty)$.

(See illustration in Figure 16.)

9. If $c \in [c_4^*, c_5^*)$, there exist $\Lambda, \lambda_0 > 0$ such that (2-1) has

- no positive solution for $\lambda \in (0, \Lambda]$,
- exactly 1 positive solution for $\lambda \in (\Lambda, \lambda_0)$,
- exactly 2 positive solutions for $\lambda = \lambda_0$, and
- exactly 3 positive solutions for $\lambda \in (\lambda_0, \infty)$.

(See illustration in Figure 17.)

10. If $c \in [c_5^*, \infty)$, there exists $\Lambda > 0$ such that (1-1) has

- no positive solution for $\lambda \in (0, \Lambda]$, and
- exactly 1 positive solution for $\lambda \in (\Lambda, \infty)$.

(See illustration in Figure 18.)

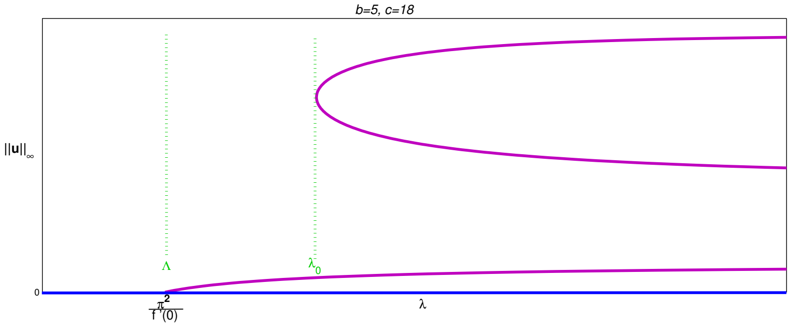


Figure 17. Illustration of Case 9.

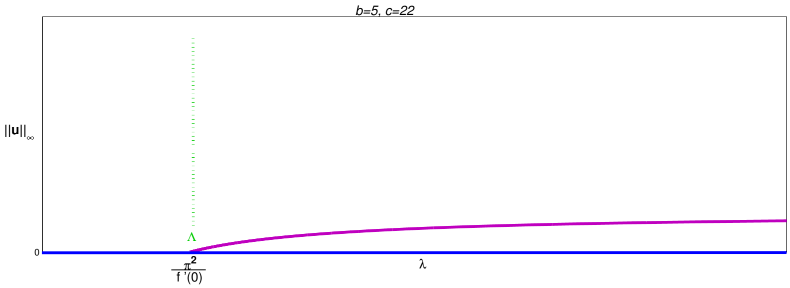


Figure 18. Illustration of Case 10.

5. Analytical results

In this section, we provide analytical proofs of several results that help detail the global behavior of bifurcation curves and further corroborate our computational results presented in the previous section. First we state two results on the behavior of $G(\rho)$ when $\rho \to 0^+$ and when $\rho \to \sigma^-$, where σ is the smallest positive root of $f(u)$. The proofs of these results are provided in the Appendix. One may also refer to [Laetsch 1970], where such results were discussed.

Lemma 5.1. $\lim_{\rho\to 0^+} G(\rho) = \pi/\sqrt{2b}$.

Lemma 5.2. $\lim_{\rho\to \sigma^-} G(\rho) = \infty$.

Next we establish a precise value of c such that for c smaller than this value, the bifurcation curve bifurcates to the left near $(\pi/\sqrt{2b}, 0)$ while for c greater than this value, the bifurcation curve bifurcates to the right near $(\pi/\sqrt{2b}, 0)$. Namely, we establish the following:

Theorem 5.1. *Let* $c_2^* = b - 1$. *If* $0 \le c < c_2^*$, *then* $G'(\rho) < 0$ *on some interval* $(0, \rho_1)$ *and if* $c > c_2^*$, *then* $G'(\rho) > 0$ *on some interval* $(0, \rho_1)$.

Proof. Recall

$$G'(\rho) = \int_0^1 \frac{H(\rho) - H(\rho v)}{[F(\rho) - F(\rho v)]^{\frac{3}{2}}} dv \tag{5-1}$$

where

$$H(s) = F(s) - \frac{s}{2} f(s). \tag{5-2}$$

Note that $H(0) = 0$. Thus, showing $H'(s) > 0$ for some $0 < s < s_0$ with $s_0 \approx 0$ implies $G'(\rho) > 0$ on some interval $(0, \rho_1)$. We have

$$H'(s) = \tfrac{1}{2}[f(s) - sf'(s)].$$

$H'(0) = 0$ also. Therefore, we differentiate again.

$$H''(s) = -\frac{s}{2} f''(s).$$

Clearly, the sign of $H''(s)$ is dependent only on the sign of $f''(s)$. We know

$$f(s) = s(s+1)(b-s) - c\frac{s^2}{1+s^2} \tag{5-3}$$

$$= -s^3 + (b-1)s^2 + bs - \frac{cs^2}{1+s^2}. \tag{5-4}$$

By taking the first derivative and simplifying, we get

$$f'(s) = -3s^2 + 2(b-1)s + b - c\frac{2s}{(1+s^2)^2}.$$

Then, taking the second derivative and simplifying, we obtain

$$f''(s) = -6s + 2(b-1) - c\frac{2-6s^2}{(1+s^2)^3}.$$

Evaluating $f''(s)$ when $s = 0$ gives

$$f''(0) = 2(b-1) - 2c. \tag{5-5}$$

By analysis of (5-5), we see that $c < (b-1) \Rightarrow f''(0) > 0 \Rightarrow H''(s) < 0 \Rightarrow H'(s) < 0$ for $s \in (0, s_0)$ for some $s_0 > 0 \Rightarrow G'(\rho) < 0$ for $\rho \approx 0$. Conversely, we have $c > (b-1) \Rightarrow f''(0) < 0 \Rightarrow H''(s) > 0 \Rightarrow H'(s) > 0$ for $s \in (0, s_1)$ for some $s_1 > 0 \Rightarrow G'(\rho) > 0$ for $\rho \approx 0$. □

Now we establish our main result of this section, the occurrence of an S-shaped bifurcation curve.

Theorem 5.2. *Let $b > 4$ and $c \in (b-1, \frac{3}{2}b - 3)$. Then the bifurcation curve for (1-1) is guaranteed to be at least S-shaped.* (See Figure 19.)

Proof. The proof is divided into three steps. In Step 1, we establish that if $b > 2$ and $c \in (\max\{0, b-5\}, \frac{3}{2}b - 3)$, then $2 \in (0, \sigma)$, for which we must recall that σ is the smallest positive root of $f(u)$. In Step 2, we prove that if $b > 4$ and $c \in (\max\{0, b-5\}, \frac{3}{2}b - 3)$, then $H(2) < 0$. In Step 3, we prove that the bifurcation curve is at least S-shaped.

Step 1. Consider the functions

$$f(u) = u(u+1)(b-u) - \frac{cu^2}{u^2+1}$$

and

$$k(u) = u(u+1)(b-u) - cu^2.$$

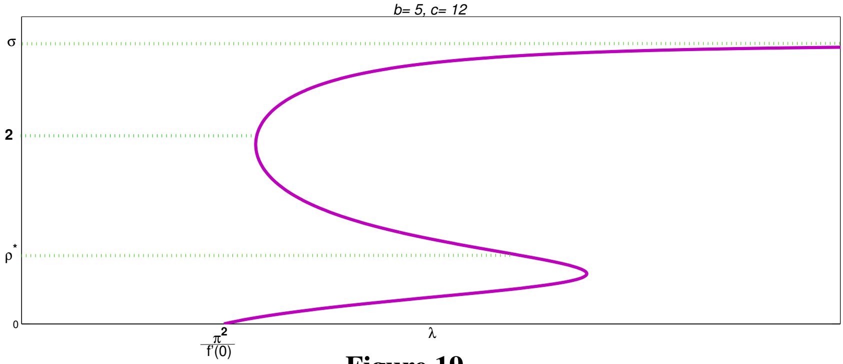


Figure 19

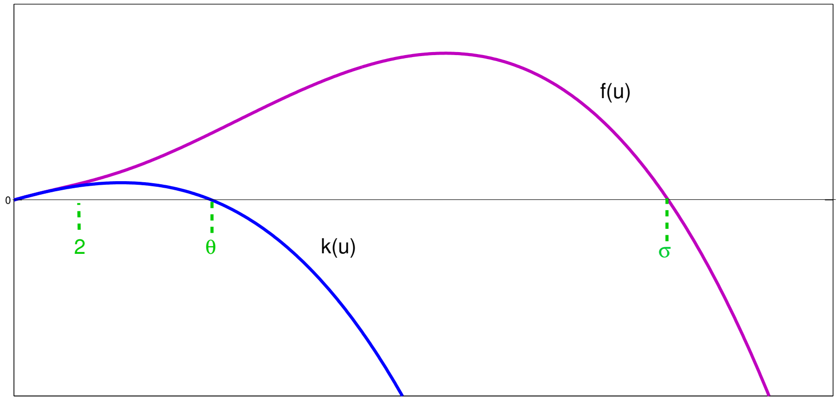


Figure 20

As shown in Figure 20, it is clear $f(u) \geq k(u)$. Any positive root of $k(u)$, θ, will occur before any positive root of $f(u)$, σ. Thus, if $r \in (0, \theta)$, then $r \in (0, \sigma)$.

Therefore, it suffices to show that $2 \in (0, \theta)$. By solving $k(u) = 0$, we obtain

$$\theta = \frac{(b-1-c)+\sqrt{4b+(b-1-c)^2}}{2}. \tag{5-6}$$

We want $\theta > 2$, so

$$\frac{(b-1-c)+\sqrt{4b+(b-1-c)^2}}{2} > 2 \tag{5-7}$$

Simplifying, we obtain

$$\sqrt{4b+(b-1-c)^2} > (5+c-b) \tag{5-8}$$

If $5+c-b \leq 0$, then it is clear the above inequality holds true. If $5+c-b > 0$, then $c > b-5$ and squaring and solving (5-8) gives

$$c < \tfrac{3}{2}b - 3. \tag{5-9}$$

Thus, if $b > 2$ and $c \in (\max\{0, b-5\}, \frac{3}{2}b-3)$, then $2 \in (0, \theta)$; hence, $2 \in (0, \sigma)$.

Step 2. Recall that

$$H(s) = F(s) - \frac{s}{2}f(s) = \frac{s^4}{4} - (b-1)\frac{s^3}{6} + c\left[\frac{s^3}{2(1+s^2)} - s + \arctan s\right].$$

Then,

$$H(2) = 4 - (b-1)\tfrac{4}{3} + c[\tfrac{4}{5} - 2 + \arctan 2] \leq \frac{16-4b}{3} + c(-\tfrac{6}{5}+1) < 0$$

for all $b > 4$.

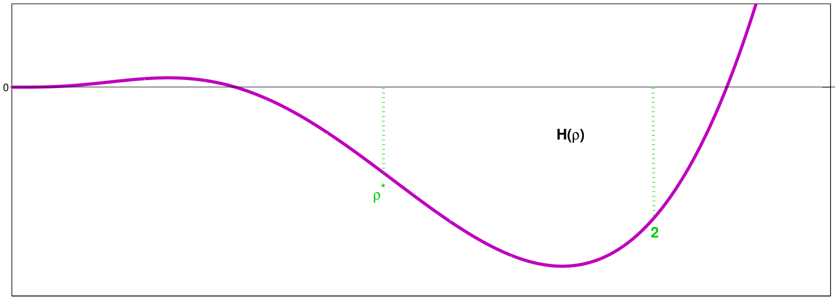


Figure 21

Step 3. Let $b > 4$ and $c \in (b-1, \frac{3}{2}b - 3)$. Recall from Lemma 2.2 that

$$G'(\rho) = \int_0^1 \frac{H(\rho) - H(\rho v)}{[F(\rho) - F(\rho v)]^{\frac{3}{2}}} dv. \tag{5-10}$$

Then from Theorem 5.1, we conclude that $G'(\rho)$ begins positive. From Steps 1 and 2, we know $2 \in (0, \sigma)$ and $H(2) < 0$. Hence there exists $\rho^* \in (0, 2)$, (say, the first zero of $H(\rho)$), such that $G'(\rho^*) < 0$. (See Figure 16.) Also, $\lim_{\rho \to \sigma^-} G(\rho) = \infty$ by Lemma 5.2. Therefore, the graph of $G(\rho)$ must be at least S-shaped. □

Next we study the bifurcation diagram for a larger range of c values. First, we consider the case when the bifurcation curve is split, providing various numbers of positive solutions for different ranges of λ:

Theorem 5.3. *There exist $c_0, \hat{c}_1$ $(< c_1)$ such that for $c \in (c_0, \hat{c}_1)$, there exists a $\lambda_1 > (\pi/\sqrt{2b})^2$ such that (1-1) has at least one positive solution in $((\pi/\sqrt{2b})^2, \lambda_1)$, at least two positive solutions for $\lambda = \lambda_1$, and at least three positive solutions for $\lambda > \lambda_1$.*

Proof. By Lemma 5.1, $\lim_{\rho \to 0^+} G(\rho) = \pi/\sqrt{2b}$. Recall from Section 3 that for $c_0 < c < c_1$, $f(s)$ has the appearance shown in Figure 22.

Furthermore, recall from Section 3 that $G(\rho)$ is well-defined for $\rho \in (0, \sigma)$ and $\rho \in (\gamma, \sigma_1)$, and that $F(\gamma) = F(\sigma)$ (see the graph of $F(u)$ for $c \in (c_0, \hat{c}_1)$ in Figure 4, page 139). From Lemma 5.2, we know $\lim_{\rho \to \sigma^-} G(\rho) = \infty$. A similar argument can be applied to show that $\lim_{\rho \to \sigma_1^-} G(\rho) = \infty$. The proof of Theorem 5.3 is complete if $\lim_{\rho \to \gamma^+} G(\rho) = +\infty$. Such a result was proved in [Brown and Budin 1979], which we will recall now. First recall that $F(\gamma) = F(\sigma)$. We let $A = \max\{|f'(s)|; s \in [0, \sigma_1]\}$. Then we can note that $|f(s)| \le A|s - \sigma|$ for all $s \in [0, \sigma_1]$. Next we let $B = \max\{|f(s)|; 0 \le s \le \sigma_1\}$. Now, if $\sigma_1 > \rho > \gamma$ and $0 \le s < \rho$, then

$$F(\rho) - F(s) = F(\rho) - F(\gamma) + F(\sigma) - F(s).$$

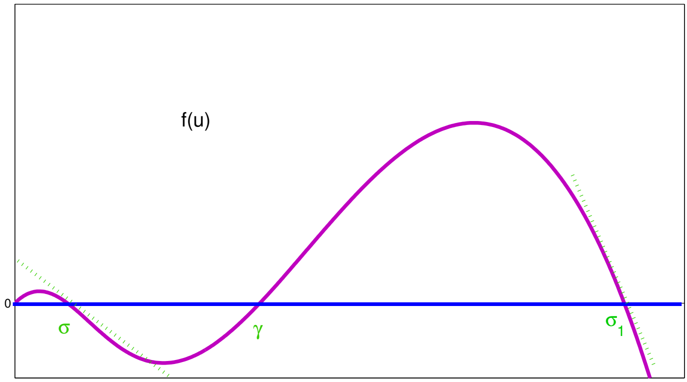


Figure 22

By the mean value theorem, we can then write this as

$$F(\rho)-F(s)=(\rho-\gamma)f(\xi)+(\sigma-s)f(\eta),$$

where $\xi\in(\gamma,\rho)$ and η lies between σ and s. Hence,

$$F(\rho)-F(s)\le B(\rho-\gamma)+A(\sigma-s)^2.$$

Recall that

$$G(\rho)=\sqrt{2}\int_0^\rho [F(\rho)-F(s)]^{-\frac{1}{2}}ds.$$

By substitution, we can write

$$G(\rho)\ge\int_0^\gamma \sqrt{2}\left[B(\rho-\gamma)+A(\sigma-s)^2\right]^{-\frac{1}{2}}ds.$$

Let $H_\rho(s)=\sqrt{2}\left[B(\rho-\gamma)+A(\sigma-s)^2\right]^{-\frac{1}{2}}$. Since H_ρ is nondecreasing as $\rho\to\gamma^+$, we can apply the monotonic convergence theorem, which gives us

$$\begin{aligned}\lim_{\rho\to\gamma^+}G(\rho)\ge\lim_{\rho\to\gamma^+}\int_0^\gamma H_\rho(s)\,ds&=\int_0^\gamma\sqrt{2}A^{-\frac{1}{2}}|\sigma-s|^{-1}ds\\&=\sqrt{2}A^{-\frac{1}{2}}\int_0^\sigma(\sigma-s)^{-1}ds+\int_\sigma^\gamma(s-\sigma)^{-1}ds.\end{aligned}$$

Both these integrals diverge to $+\infty$, so $\lim_{\rho\to\gamma^+}G(\rho)=+\infty$. □

Thus, we obtain the bifurcation diagram represented in Figure 23 (see next page). Next we establish a result for large values of c:

Theorem 5.4. *There exists a $\tilde{c}$ such that if $c>\tilde{c}$, then (1-1) has a unique positive solution for all $\lambda>(\pi/\sqrt{2b})^2$.*

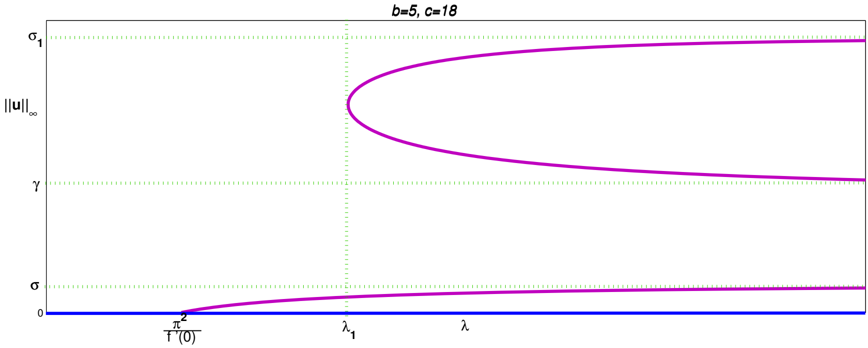


Figure 23

Proof. From Section 3 we know that for $c > \hat{c}_1$, $G(\rho)$ is only defined for $\rho \in (0, \sigma)$. With $f_1(u) = u(u+1)(b-u)$ and $f_2(u) = cu^2/(1+u^2)$, the graph of $f_1 - f_2$ is illustrated in Figure 24. Recall the equality

$$G'(\rho) = \int_0^1 \frac{H(\rho) - H(\rho v)}{[F(\rho) - F(\rho v)]^{\frac{3}{2}}} dv$$

with

$$H(s) = F(s) - \tfrac{1}{2} s f(s), \quad H'(s) = \tfrac{1}{2}[f(s) - s f'(s)], \quad H''(s) = -\tfrac{1}{2} s f''(s).$$

We wish to show that $f''(s) < 0$ for $0 < s < \sigma$. This will alternatively imply that $H''(s) > 0$ for $0 < s < \sigma$, noting once again that $H(0) = H'(0) = 0$. Therefore, showing $H'(s) > 0$ for $0 < s < \sigma$ implies $G'(\rho) > 0$ for $0 < \rho < \sigma$, as shown in the bifurcation diagram in Figure 25.

We begin with the analysis of $f''(s)$.

$$f''(s) = -6s + 2(b-1) - c\frac{2 - 6s^2}{(1+s^2)^3} \tag{5-11}$$

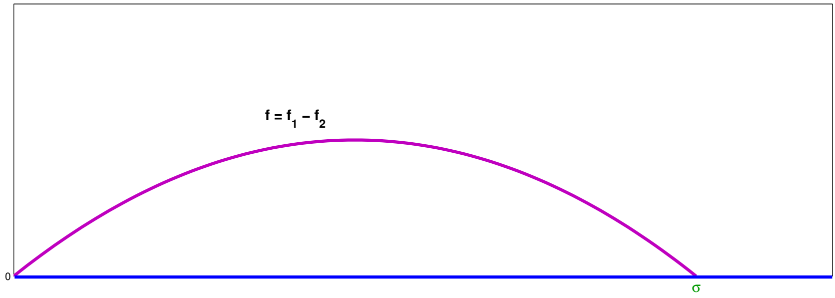


Figure 24. $f(u) = f_1 - f_2$.

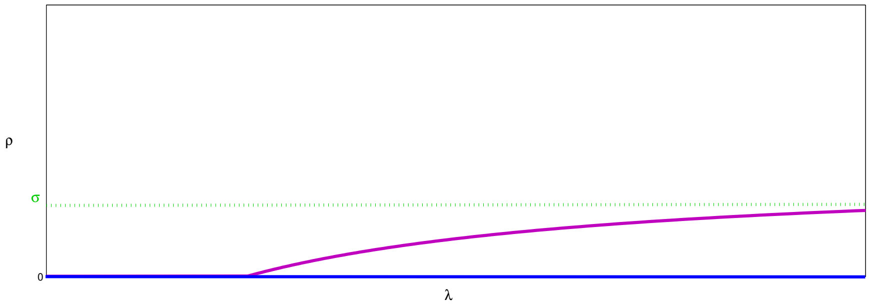


Figure 25

We can then bound this function by a larger one, so we choose

$$f''(s) \leq 2(b-1) - c\frac{2-6s^2}{(1+s^2)^3}. \tag{5-12}$$

Let

$$B(s) = \frac{2-6s^2}{(1+s^2)^3}.$$

Note that $B(s) > 0$ on $0 \leq s < \frac{1}{\sqrt{3}}$. For $c \gg 1$, we can assume that $\sigma < \frac{1}{2\sqrt{3}}$. Hence, for $c \gg 1$, there exists $\delta > 0$ such that $B(s) \geq \delta$ for all $0 \leq s \leq \sigma$. Thus,

$$f''(s) \leq 2(b-1) - c\delta \quad \text{for all } 0 \leq s \leq \sigma. \tag{5-13}$$

Therefore, for $c \gg 1$, $f''(s) < 0$ for $0 < s < \sigma$. Hence, we know then that $H''(s) > 0$ for $0 < s < \sigma$ and $G'(\rho) > 0$ for $0 < \rho < \sigma$. □

The corresponding bifurcation diagram is illustrated in Figure 25.

6. Computational results for logistic growth

In [Lee et al. 2011], the effect of grazing on a logistic growth rate was studied. Several bifurcation diagrams were provided, but a complete bifurcation evolution for the one-dimensional case as c varies was not provided. It is useful to compare these computational results to those of the weak and strong Allee effect. The combination of grazing with a logistic growth rate can be illustrated by the following equation:

$$\hat{f}(u) = u(1-bu) - c\frac{u^2}{1+u^2}; \; b > 0, c \geq 0. \tag{6-1}$$

We obtain our evolution results via the quadrature method and Mathematica computations. The following figures illustrate this evolution for a fixed b as c increases.

If $b \in (0, b_0)$, then there exist $\hat{c}_0, \hat{c}_1, \hat{c}_2, \hat{c}_3, \hat{c}_4 > 0$ such that:

1. If $c \in [0, \hat{c}_0)$, there exists a $\Lambda > 0$ such that (6-1) has

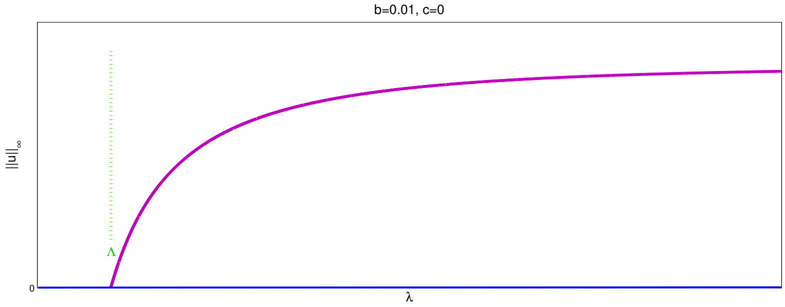


Figure 26. Illustration of Case 1.

- no positive solution for $\lambda \in (0, \Lambda)$, and
- exactly 1 positive solution for $\lambda \in [\Lambda, \infty)$.

(See illustration in Figure 26.)

2. If $c \in (\hat{c}_0, \hat{c}_1)$, there exist $\Lambda, \lambda_0, \lambda_1 > 0$ such that (6-1) has
   - no positive solution for $\lambda \in (0, \Lambda)$,
   - exactly 1 positive solution for $\lambda \in [\Lambda, \lambda_0)$,
   - exactly 2 positive solutions for $\lambda = \lambda_0$,
   - exactly 3 positive solutions for $\lambda \in (\lambda_0, \lambda_1]$, and
   - exactly 1 positive solution for $\lambda \in (\lambda_1, \infty)$.

   (See illustration in Figure 27.)

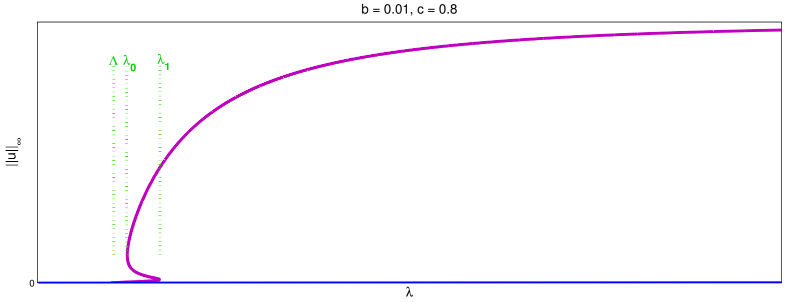


Figure 27. Illustration of Case 2.

3. If $c \in (\hat{c}_2, \hat{c}_3)$, there exist $\Lambda, \lambda_0 > 0$ such that (6-1) has
   - no positive solution for $\lambda \in (0, \Lambda)$,
   - exactly 1 positive solution for $\lambda \in (\Lambda, \lambda_0)$,
   - exactly 2 positive solutions for $\lambda = \lambda_0$, and
   - exactly 3 positive solutions for $\lambda \in (\lambda_0, \infty]$.

   (See illustrations in Figure 28.)

4. If $c \in (\hat{c}_3, \hat{c}_4)$, there exists a $\Lambda > 0$, such that (6-1) has
   - no positive solution for $\lambda \in (0, \Lambda)$, and

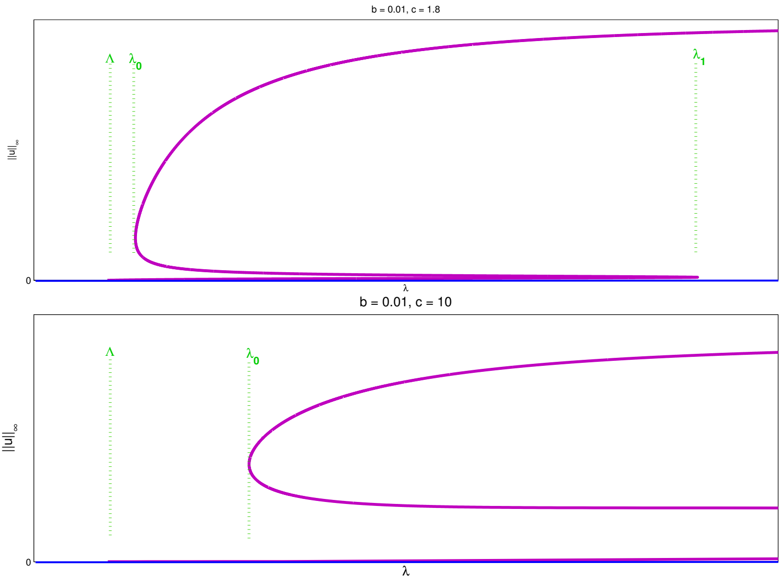


Figure 28. Illustrations of Case 3.

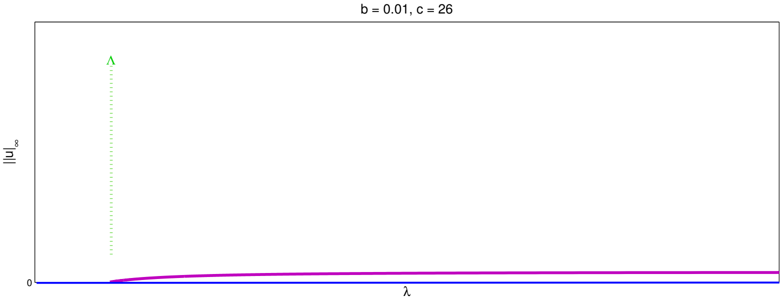


Figure 29. Illustration of Case 4.

- exactly 1 positive solution for $\lambda \in [\Lambda, \infty)$.

(See illustration in Figure 29.)

7. Computational results for strong Allee effect

This section describes the case in which a grazing term is combined with a strong Allee growth rate. Thus, our reaction term is

$$\bar{f}(u) = u(u-1)(b-u) - \frac{cu^2}{1+u^2}, \qquad b > 1,\ c \geq 0.$$

As in Section 4, we obtain our results via the quadrature method detailed in Section 2 and apply Mathematica to complete our computations. In the top left part of Figure 30, we present the bifurcation curve with no grazing term ($c = 0$). The

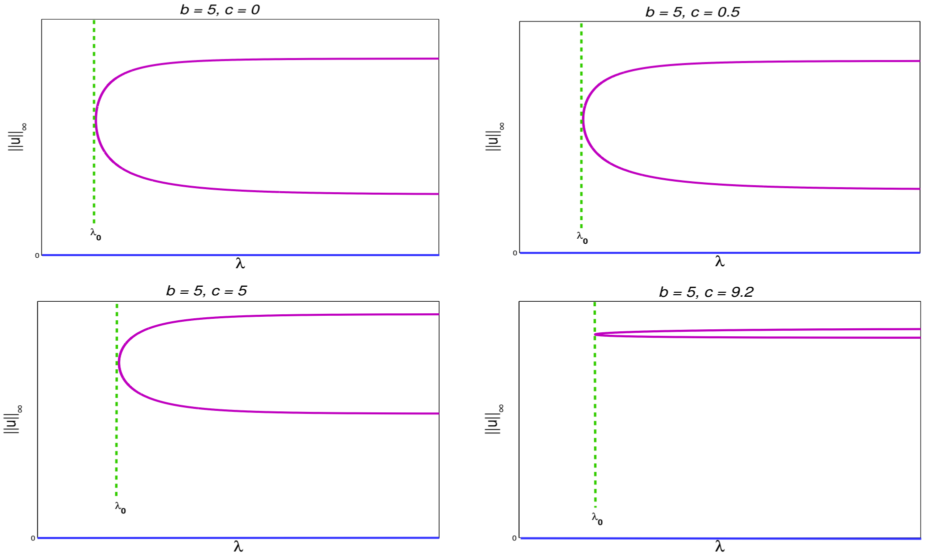


Figure 30. Evolution of the bifurcation curve as c increases.

resulting bifurcation curve evolution for increasing c is briefly exemplified in the remaining three parts of the figure.

After careful computational analysis and application of values for ranges of $b > 1$ and $c \geq 0$, we observe that the grazing term ultimately has little effect on the overall structure of the resulting bifurcation curve. However, we must note that for large values of c the grazing term will overcome the strong Allee effect and there will no longer exist any steady states. Thus, in this case the population will die out.

8. Biological implications

Analysis of the steady states of (1-1) provides valuable information on the long-term survival of a population. Given an initial population size and grazing rate, the bifurcation diagrams we have included provide ranges of λ for which the population persists. Depending on the range of λ, the persistence is either conditional or unconditional.

When λ is small, the diffusion coefficient is large enough to cause a population to die out despite its initial size. This is clearly illustrated through the bifurcation diagrams for the range $\lambda \in (0, \lambda_0)$. Using Figure 31 as an example, it is clear that whether a population has an initial size of k or of l, it will still die out.

Between λ_1 and λ_2, we have unconditional persistence. That is, a population with an initial size of m will decrease until achieving stability at the bottom branch. However, between λ_2 and λ_3, the stability of the steady states results in conditional persistence. In this range, the top and bottom branches are stable solutions while the

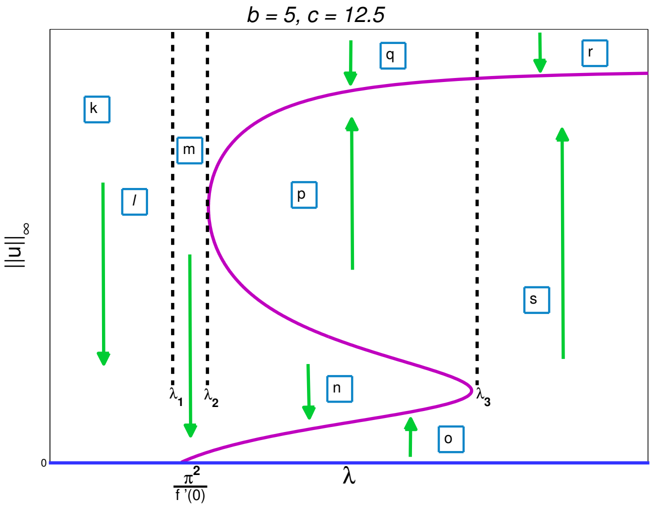


Figure 31

middle branch is unstable. Thus, a population's persistence is dictated by its initial population size. For example, a population beginning with a size of n will decline until reaching stability at the bottom branch; whereas, a population beginning with a size of o will increase until reaching stability at the bottom branch. Furthermore, for an initial size of p, the population will grow until obtaining stability at the top branch while an initial size of q will diminish until obtaining stability at the top branch.

For $\lambda > \lambda_3$, the population will unconditionally persist. With an initial size of r, a population will decrease until stabilizing at the top branch while an initial size of s will increase until stabilizing at the top branch.

9. Appendix: Proofs of Lemma 5.1 and Lemma 5.2

Proof of Lemma 5.1. (See also [Laetsch 1970].)

$$\begin{aligned}\lim_{\rho\to 0^+} G(\rho) &= \lim_{\rho\to 0^+}\int_0^\rho \frac{dz}{\sqrt{F(\rho)-F(z)}} = \lim_{\rho\to 0^+}\int_0^1 \frac{\rho\,dv}{\sqrt{F(\rho)-F(\rho v)}}\\ &= \lim_{\rho\to 0^+}\int_0^1 \frac{dv}{\sqrt{\frac{F(\rho)-F(\rho v)}{\rho^2}}}\end{aligned} \tag{9-1}$$

By Lebesgue's dominated convergence theorem, the limit can be moved inside the integral. Thus, we evaluate

$$\lim_{\rho\to 0^+}\frac{F(\rho)-F(\rho v)}{\rho^2}.$$

After applying L'Hospital's rule twice, we obtain

$$\lim_{\rho\to 0^+} \frac{f'(\rho)-v^2 f'(\rho v)}{2} = \frac{f'(0)}{2}(1-v^2). \tag{9-2}$$

Combining (9-1) and (9-2), we have

$$\lim_{\rho\to 0^+} G(\rho) = \sqrt{\frac{2}{f'(0)}}\int_0^1 \frac{dv}{\sqrt{1-v^2}} = \frac{\pi}{\sqrt{2f'(0)}} = \frac{\pi}{\sqrt{2b}}. \qquad \square$$

Proof of Lemma 5.2. (See also [Laetsch 1970].) Let $N > 0$ be large enough such that $f(u) \le N(\sigma - u)$ for all $0 \le u \le \sigma$. By the mean value theorem,

$$F(\rho)-F(s) = F'(\theta)(\rho-s) = f(\theta)(\rho-s) \le N(\sigma-\theta)(\rho-s) \le N(\sigma-s)^2.$$

Then $\sqrt{F(\rho)-F(s)} \le \sqrt{N}(\sigma-s)$, or, with $n = 1/\sqrt{N}$,

$$\frac{1}{\sqrt{F(\rho)-F(s)}} \ge \frac{n}{\sigma - s}.$$

Integrating both sides gives

$$\int_0^\rho \frac{ds}{\sqrt{F(\rho)-F(s)}} \ge n\int_0^\rho \frac{ds}{\sigma-s} G(\rho) \ge -n\ln(\sigma-\rho)+n\ln(\sigma).$$

As $\rho\to\sigma^-$, the right side of the inequality approaches ∞. Thus, $G(\rho)\to\infty$ as $\rho\to\sigma^-$. $\square$

Acknowledgements

We would like to extend a special thanks to our advisor and mentor Dr. Ratnasingham Shivaji for his guidance throughout this research. We would also like to thank our graduate mentors Jerome Goddard II and Dagny Grillis for their assistance.

References

[Allee 1938] W. C. Allee, *The social life of animals*, Norton, New York, 1938.

[Brown and Budin 1979] K. J. Brown and H. Budin, "On the existence of positive solutions for a class of semilinear elliptic boundary value problems", *SIAM J. Math. Anal.* **10**:5 (1979), 875–883. MR 82k:35043 Zbl 0414.35029

[Brown et al. 1981] K. J. Brown, M. M. A. Ibrahim, and R. Shivaji, "*S*-shaped bifurcation curves", *Nonlinear Anal.* **5**:5 (1981), 475–486. MR 82h:35007 Zbl 0458.35036

[Laetsch 1970] T. Laetsch, "The number of solutions of a nonlinear two point boundary value problem", *Indiana Univ. Math. J.* **20** (1970), 1–13. MR 42 #4815 Zbl 0215.14602

[Lee et al. 2011] E. Lee, S. Sasi, and R. Shivaji, "S-shaped bifurcation curves in ecosystems", *J. Math. Anal. Appl.* **381**:2 (2011), 732–741. MR 2012e:92080 Zbl 1221.35421

[Shi and Shivaji 2006] J. Shi and R. Shivaji, "Persistence in reaction diffusion models with weak Allee effect", *J. Math. Biol.* **52**:6 (2006), 807–829. MR 2007g:92070 Zbl 1110.92055

[Van Nes and Scheffer 2005] E. H. Van Nes and M. Scheffer, "Implications of spatial heterogeneity for catastrophic regime shifts in ecosystems", *Ecology* **86**:7 (2005), 1797–1807.

Received: 2011-02-08 Revised: 2011-09-26 Accepted: 2011-09-27

ekpoole@uark.edu *Department of Mathematics and Statistics,*
University of Arkansas, Fayatteville, AR 72701, United States

bjr76@msstate.edu *Department of Mathematics and Statistics, Center for*
Computational Sciences, Mississippi State University,
Mississippi State, MS 39762, United States

bcs173@msstate.edu *Department of Mathematics and Statistics, Center for*
Computational Sciences, Mississippi State University,
Mississippi State, MS 39762, United States

mathematical sciences publishers **msp**

INVOLVE 5:2 (2012)
dx.doi.org/10.2140/involve.2012.5.159

A BMO theorem for ϵ-distorted diffeomorphisms on $\mathbb{R}^D$ and an application to comparing manifolds of speech and sound

Charles Fefferman, Steven B. Damelin and William Glover

(Communicated by Kenneth S. Berenhaut)

This paper deals with a BMO theorem for ϵ-distorted diffeomorphisms on $\mathbb{R}^D$ and an application comparing manifolds of speech and sound.

1. Introduction

From the very beginning of time, mathematicians have been intrigued by the fascinating connections which exist between music, speech and mathematics. Indeed, these connections were already in some subtle form in the writings of Gauss. The aim of this paper is to study estimates in measure for diffeomorphisms $\mathbb{R}^D$ to $\mathbb{R}^D$, $D \geq 2$ of small distortion and provide an application to comparing music and speech manifolds.

This paper originated from discussions where Glover, an undergraduate student of Damelin and a passionate practitioner of music (particularly the piano), introduced Damelin to the beautiful world of beats, movements, scales, measures and time signatures. A fruitful and inspiring collaboration ensued, enriched by wonderful contributions from Fefferman.

2. Preliminaries

Fix a dimension $D \geq 2$. We work in $\mathbb{R}^D$. We write $B(x, r)$ to denote the open ball in $\mathbb{R}^D$ with centre x and radius r. We write A to denote Euclidean motions on $\mathbb{R}^D$. A Euclidean motion may be orientation-preserving or orientation reversing. We write c, C, C', etc. to denote constants depending on the dimension D. These expressions

MSC2010: 00A65, 65D19, 97M80.

Keywords: measure, BMO, diffeomorphism, small distortion, manifold, music, sound, noise, speech, alignment.

Damelin thanks Princeton University and Wayne Country Day School for financial support. He also thanks Prof Charles Fefferman for wonderfully inspiring collaborations and his generosity.

need not denote the same constant in different occurrences. For a $D \times D$ matrix, $M = (M_{ij})$, we write $|M|$ to denote the Hilbert–Schmidt norm

$$|M| = \Big(\sum_{ij} |M_{ij}|^2\Big)^{1/2}.$$

Note that if M is real and symmetric and if

$$(1-\lambda)I \le M \le (1+\lambda)I$$

as matrices, where $0 < \lambda < 1$, then

$$|M - I| \le C\lambda. \tag{2-1}$$

This follows from working in an orthonormal basis for which M is diagonal. One way to understand the formulas above is to think of λ as being close to zero. See also (2-6) below.

A function $f : \mathbb{R}^D \to \mathbb{R}$ is said to be BMO (Bounded mean oscillation)if there is a constant $K \ge 0$ such that, for every ball $B \subset \mathbb{R}^D$, there exists a real number H_B such that

$$\frac{1}{\operatorname{vol} B}\int_B |f(x) - H_B|\, dx \le K. \tag{2-2}$$

The least such K is denoted by $\|f\|_{\mathrm{BMO}}$.

In harmonic analysis, a function of bounded mean oscillation, also known as a BMO function, is a real-valued function whose mean oscillation is bounded (finite). The space of functions of bounded mean oscillation (BMO), is a function space that, in some precise sense, plays the same role in the theory of Hardy spaces, that the space of essentially bounded functions plays in the theory of Lp-spaces: it is also called a John–Nirenberg space, after Fritz John and Louis Nirenberg who introduced and studied it for the first time [John 1961; John and Nirenberg 1961].

The John–Nirenberg inequality asserts the following: Let $f \in$ BMO and let $B \subset \mathbb{R}^D$ be a ball. Then there exists a real number H_B such that

$$\operatorname{vol}\{x \in B : |f(x) - H_B| > C\lambda\|f\|_{\mathrm{BMO}}\} \le \exp(-\lambda)\operatorname{vol} B, \quad \lambda \ge 1. \tag{2-3}$$

As a corollary of the John–Nirenberg inequality, we have

$$\left(\frac{1}{\operatorname{vol} B}\int_B |f(x) - H_B|^4 dx\right)^{1/4} \le C\lambda\|f\|_{\mathrm{BMO}}. \tag{2-4}$$

There is nothing special about the 4th power in the above; it will be needed later.

The definition of BMO, the notion of the BMO norm, the John–Nirenburg inequality (2-3) and its corollary (2-4) carry through to the case of functions f on $\mathbb{R}^D$ which take their values in the space of $D \times D$ matrices. Indeed, we take

H_B in (2-2)–(2-4) to be a $D \times D$ matrix for such f. The matrix valued norms of (2-3)–(2-4) follow easily from the scalar case.

We will need some potential theory. If f is a smooth function of compact support in $\mathbb{R}^D$, then we can write $\Delta^{-1} f$ to denote the convolution of f with the Newtonian potential. Thus, $\Delta^{-1} f$ is smooth and $\Delta(\Delta^{-1} f) = f$ on $\mathbb{R}^D$.

We will use the estimate:

$$\left\| \frac{\partial}{\partial x_i} \Delta^{-1} \frac{\partial}{\partial x_j} f \right\|_{L^2(\mathbb{R}^D)} \leq C \| f \|_{L^2(\mathbb{R}^D)}, \quad i, j = 1, \dots, D, \tag{2-5}$$

valid for any smooth function f with compact support. Estimate (2-5) follows by applying the Fourier transform.

We will work with a positive number ε. We always assume that $\varepsilon \leq \min(1, C)$. An ε-distorted diffeomorphism of $\mathbb{R}^D$ is a one to one and onto diffeomorphism $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ such as

$$(1-\varepsilon)I \leq (\Phi'(x))^T (\Phi'(x)) \leq (1+\varepsilon)I$$

as matrices. Thanks to (2-1), such Φ satisfy

$$\left| (\Phi'(x))^T (\Phi'(x)) - I \right| \leq C\varepsilon. \tag{2-6}$$

We end this section with the following inequality from [Fefferman and Damelin ≥ 2012]:

Approximation Lemma. *Let $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ be an ε-distorted diffeomorphism. Then, there exists an Euclidean motion A such that*

$$|\Phi(x) - A(x)| \leq C\varepsilon \tag{2-7}$$

for all $x \in B(0, 10)$.

3. An overdetermined system

We will need to study the following elemetary overdetermined system of partial differential equations.

$$\frac{\partial \Omega_i}{\partial x_j} + \frac{\partial \Omega_j}{\partial x_i} = f_{ij}, \quad i, j = 1, \dots, D, \tag{3-1}$$

on $\mathbb{R}^D$. Here, Ω_i and f_{ij} are C^∞ functions on $\mathbb{R}^D$. A result concerning (3-1) we need is:

PDE Theorem. *Let $\Omega_1, \dots, \Omega_D$ and f_{ij}, for $i, j = 1, \dots, D$, be smooth functions on $\mathbb{R}^D$. Assume that (3-1) holds and suppose that*

$$\| f_{ij} \|_{L^2(B(0,4))} \leq 1. \tag{3-2}$$

Then, there exist real numbers Δ_{ij}, for $i, j = 1, \dots, D$, such that

$$\Delta_{ij} + \Delta_{ji} = 0 \quad \textit{for all } i, j \tag{3-3}$$

and

$$\left\| \frac{\partial \Omega_i}{\partial x_j} - \Delta_{ij} \right\|_{L^2(B(0,1))} \le C. \tag{3-4}$$

Proof. From (3-1), we see at once that

$$\frac{\partial \Omega_i}{\partial x_i} = \tfrac{1}{2} f_{ii}$$

for each i. Now, by differentiating (3-1) with respect to x_j and then summing on j, we see that

$$\Delta \Omega_i + \frac{1}{2} \frac{\partial}{\partial x_i} \left(\sum_j f_{jj} \right) = \sum_j \frac{\partial f_{ij}}{\partial x_j}$$

for each i. Therefore, we may write

$$\Delta \Omega_i = \sum_j \frac{\partial}{\partial x_j} g_{ij}$$

for smooth functions g_{ij} with

$$\|g_{ij}\|_{L^2(B(0,4)} \le C.$$

This holds for each i. Let χ be a C^∞ cutoff function on $\mathbb{R}^D$ equal to 1 on $B(0, 2)$ vanishing outside $B(0, 4)$ and satisfying $0 \le \chi \le 1$ everywhere. Now let

$$\Omega_i^{\text{err}} = \Delta^{-1} \sum_j \frac{\partial}{\partial x_j} \left(\chi g_{ji} \right)$$

and let

$$\Omega_i^* = \Omega_i - \Omega_i^{\text{err}}.$$

Then,

$$\Omega_i = \Omega_i^* + \Omega_i^{\text{err}} \tag{3-5}$$

each i. The function

$$\Omega_i^* \tag{3-6}$$

is harmonic on $B(0, 2)$ and

$$\left\| \nabla \Omega_i^{\text{err}} \right\|_{L^2(B(0,2))} \le C \tag{3-7}$$

thanks to (2-5). By (3-1), (3-2), (3-5), (3-7), we can write

$$\frac{\partial \Omega_i^*}{\partial x_j} + \frac{\partial \Omega_j^*}{\partial x_i} = f_{ij}^*, \quad i, j = 1, \dots, D, \tag{3-8}$$

on $B(0, 2)$ and with

$$\| f_{ij}^* \|_{L^2(B(0,2))} \leq C. \tag{3-9}$$

From (3-6) and (3-8), we see that each f_{ij}^* is a harmonic function on $B(0, 2)$. Consequently, (3-9) implies

$$sup_{B(0,1)} |\nabla f_{ij}^*| \leq C. \tag{3-10}$$

From (3-8), we have for each i, j, k,

$$\frac{\partial^2 \Omega_i^*}{\partial x_j \partial x_k} + \frac{\partial^2 \Omega_k^*}{\partial x_i \partial x_j} = \frac{\partial f_{ik}^*}{\partial x_j}, \quad \frac{\partial^2 \Omega_i^*}{\partial x_j \partial x_k} + \frac{\partial^2 \Omega_j^*}{\partial x_i \partial x_k} = \frac{\partial f_{ij}^*}{\partial x_k}, \tag{3-11}$$

$$\frac{\partial^2 \Omega_j^*}{\partial x_i \partial x_k} + \frac{\partial^2 \Omega_k^*}{\partial x_i \partial x_j} = \frac{\partial f_{jk}^*}{\partial x_i}. \tag{3-12}$$

Now adding the first two equations above and subtracting the last, we obtain:

$$2\frac{\partial^2 \Omega_i^*}{\partial x_j \partial x_k} = \frac{\partial f_{ik}^*}{\partial x_j} + \frac{\partial f_{ij}^*}{\partial x_k} - \frac{\partial f_{jk}^*}{\partial x_i} \tag{3-13}$$

on $B(0, 1)$. Now from (3-10) and (3-13), we obtain the estimate

$$\left| \frac{\partial^2 \Omega_i^*}{\partial x_j \partial x_k} \right| \leq C \tag{3-14}$$

on $B(0, 1)$ for each i, j, k. Now for each i, j, let

$$\Delta_{ij}^* = \frac{\partial \Omega_i^*}{\partial x_j}(0). \tag{3-15}$$

By (3-14), we have

$$\left| \frac{\partial \Omega_i^*}{\partial x_j} - \Delta_{ij}^* \right| \leq C \tag{3-16}$$

on $B(0, 1)$ for each i, j. Recalling (3-5) and (3-7), we see that (3-16) implies that

$$\left\| \frac{\partial \Omega_i}{\partial x_j} - \Delta_{ij}^* \right\|_{L^2(B(0,1))} \leq C. \tag{3-17}$$

Unfortunately, the Δ_{ij}^* need not satisfy (3-3). However, (3-1), (3-2) and (3-17) imply the estimate

$$|\Delta_{ij}^* + \Delta_{ji}^*| \leq C$$

for each i, j. Hence, there exist real numbers Δ_{ij}, $(i, j = 1, \dots, D)$ such that

$$\Delta_{ij} + \Delta_{ji} = 0 \tag{3-18}$$

and

$$|\Delta^*_{ij} - \Delta_{ij}| \le C \tag{3-19}$$

for each i, j. From (3-17) and (3-19), we see that

$$\left\| \frac{\partial \Omega_i}{\partial x_j} - \Delta_{ij} \right\|_{L^2(B(0,1))} \le C \tag{3-20}$$

for each i and j.

Thus (3-18) and (3-20) are the desired conclusions of the theorem. □

4. A BMO theorem

BMO Theorem 1. *Let $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ be an ε diffeomorphism and let $B \subset \mathbb{R}^D$ be a ball. Then, there exists $T \in O(D)$ such that*

$$\frac{1}{\text{vol}\, B} \int_B |\Phi'(x) - T| \, dx \le C\varepsilon^{1/2}. \tag{4-1}$$

Proof. Estimate (4-1) is preserved by translations and dilations. Hence we may assume that

$$B = B(0, 1). \tag{4-2}$$

Now we know that there exists an Euclidean motion $A : \mathbb{R}^D \to \mathbb{R}^D$ such that

$$|\Phi(x) - A(x)| \le C\varepsilon \tag{4-3}$$

for $x \in B_{(0,10)}$. Our desired conclusion (4-1) holds for Φ if and only if it holds for $A^{-1} o \Phi$ (with a different T). Hence, without loss of generality, we may assume that $A = I$. Thus, (4-3) becomes

$$|\Phi(x) - x| \le C\varepsilon, \quad x \in B(0, 10). \tag{4-4}$$

We set up some notation: We write the diffeomorphism Φ in coordinates by setting:

$$\Phi(x_1, \dots, x_D) = (y_1, \dots, y_D) \tag{4-5}$$

where for each i, $1 \le i \le D$,

$$y_i = \psi_i(x_1, \dots, x_D). \tag{4-6}$$

First claim: For each $i = 1, \dots, D$,

$$\int_{B(0,1)} \left| \frac{\partial \psi_i(x)}{\partial x_i} - 1 \right| \le C\varepsilon. \tag{4-7}$$

For this, for fixed $(x_2, \dots, x_D) \in B'$, we apply (4-4) to the points $x^+ = (1, \dots, x_D)$ and $x^- = (1, \dots, x_D)$. We have

$$\left|\psi_1(x^+) - 1\right| \le C\varepsilon$$

and

$$\left|\psi_1(x^{-1}) + 1\right| \le C\varepsilon.$$

Consequently,

$$\int_{-1}^{1} \frac{\partial \psi_1}{\partial x_1}(x_1, \dots, x_D) dx_1 \ge 2 - C\varepsilon. \tag{4-8}$$

On the other hand, since,

$$\left(\psi'(x)\right)^T \left(\psi'(x)\right) \le (1+\varepsilon) I,$$

we have for each $i = 1, \dots, D$ the inequality

$$\left(\frac{\partial \psi_i}{\partial x_i}\right)^2 \le 1 + \varepsilon.$$

Therefore,

$$\left|\frac{\partial \psi_i}{\partial x_i}\right| - 1 \le \sqrt{1+\varepsilon} - 1 \le \varepsilon. \tag{4-9}$$

Set

$$I^+ = \left\{x_1 \in [-1, 1] : \frac{\partial \psi_1}{\partial x_1}(x_1, \dots, x_D) - 1 \le 0\right\},$$

$$I^{-1} = \left\{x_1 \in [-1, 1] : \frac{\partial \psi_1}{\partial x_1}(x_1, \dots, x_D) - 1 \ge 0\right\},$$

$$\Delta^+ = \int_{I^+} \left(\frac{\partial \psi_1}{\partial x_1}(x_1, \dots, x_D) - 1\right) dx_1,$$

$$\Delta^- = \int_{I^-} \left(\frac{\partial \psi_1}{\partial x_1}(x_1, \dots, x_D) - 1\right) dx_1.$$

The inequality (4-8) implies that $-\Delta^{-1} \le C\varepsilon + \Delta^+$. The inequality (4-9) implies that

$$\frac{\partial \psi_1}{\partial x_1} - 1 \le C\varepsilon.$$

Integrating the last inequality over I^+, we obtain $\Delta^+ \le C\varepsilon$. Consequently,

$$\int_{-1}^{1} \left|\frac{\partial \psi_1}{\partial x_1}(x_1, \dots, x_D) - 1\right| dx_1 = \Delta^+ - \Delta^- \le C\varepsilon. \tag{4-10}$$

Integrating this last equation over $(x_2, \dots, x_D) \in B'$ and noting that $B(0, 1) \subset [-1, 1] \times B'$, we conclude that

$$\int_{B(0,1)} \left| \frac{\partial \psi_1}{\partial x_1}(x_1, \dots, x_D) - 1 \right| dx \leq C\varepsilon.$$

Similarly, for each $i = 1, \dots, D$, we obtain (4-7).

Second claim: For each $i, j = 1, \dots, D,\ i \neq j$, we have

$$\int_{B(0,1)} \left| \frac{\partial \psi_i(x)}{\partial x_j} \right| dx \leq C\sqrt{\varepsilon}. \tag{4-11}$$

Since

$$(1-\varepsilon)I \leq (\Phi'(x))^T(\Phi'(x)) \leq (1+\varepsilon)I,$$

we have

$$\sum_{i,j=1}^{D} \left(\frac{\partial \psi_i}{\partial x_j} \right)^2 \leq (1 + C\varepsilon)D. \tag{4-12}$$

Therefore,

$$\sum_{i \neq j} \left(\frac{\partial \psi_i}{\partial x_j} \right)^2 \leq C\varepsilon + \sum_{i=1}^{D} \left(1 - \frac{\partial \psi_i}{\partial x_i} \right) \left(1 + \frac{\partial \psi_i}{\partial x_i} \right).$$

Using (4-9) for i, we have $|\partial \psi_i / \partial x_i| + 1 \leq C$. Therefore,

$$\sum_{i \neq j} \left(\frac{\partial \psi_i}{\partial x_j} \right)^2 \leq C\varepsilon + C \left| \frac{\partial \psi_i}{\partial x_i} - 1 \right|.$$

Now integrating the last inequality over the unit ball and using (4-7), we find that

$$\int_{B(0,1)} \sum_{i \neq j} \left(\frac{\partial \psi_i}{\partial x_j} \right)^2 dx \leq C\varepsilon + \int_{B(0,1)} \left| \frac{\partial \psi_i}{\partial x_i} - 1 \right| dx \leq C\varepsilon. \tag{4-13}$$

Consequently, by the Cauchy–Schwarz inequality, we have

$$\int_{B(0,1)} \sum_{i \neq j} \left| \frac{\partial \psi_i}{\partial x_j} \right| dx \leq C\sqrt{\varepsilon}.$$

Third claim:

$$\int_{B(0,1)} \left| \frac{\partial \psi_i}{\partial x_i} \right| dx \leq C\sqrt{\varepsilon} \tag{4-14}$$

Since,

$$\int_{B(0,1)} \left(\frac{\partial \psi_i}{\partial x_i} - 1\right)^2 dx \le \int_{B(0,1)} \left|\frac{\partial \psi_i}{\partial x_i} - 1\right| \left|\frac{\partial \psi_i}{\partial x_i} + 1\right| dx,$$

using (4-7) and $|\partial \psi_i / \partial x_i| \le 1 + C\varepsilon$, we obtain

$$\int_{B(0,1)} \left(\frac{\partial \psi_i}{\partial x_i}\right)^2 dx \le C\varepsilon.$$

Thus, an application of Cauchy–Schwarz, yields (4-14).

Final claim: By the Hilbert–Schmidt definition, we have

$$\begin{aligned}\int_{B(0,1)} |\Psi'(x) - I|\, dx &= \int_{B(0,1)} \left(\sum_{i,j=1}^{D} \left(\frac{\partial \psi_i}{\partial x_j} - \delta_{ij}\right)^2\right)^{1/2} \\ &\le \int_{B(0,1)} \sum_{i,j=1}^{D} \left|\frac{\partial \psi_i}{\partial x_j} - \delta_{ij}\right| dx.\end{aligned}$$

The estimate (4-11) combined with (4-14) yields:

$$\int_{B(0,1)} \left|\Phi'(x) - I\right| dx \le C\varepsilon^{1/2}.$$

Thus we have proved (4-1) with $T = I$. The proof of the BMO Theorem 1 is complete. □

Corollary. *Let $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ be an ε-distorted diffeomorphism. For each, ball $B \subset \mathbb{R}^D$, there exists $T_B \in O(D)$, such that*

$$\left(\frac{1}{\operatorname{vol} B} \int_B \left|\Phi'(x) - T\right|^4 dx\right)^{1/4} \le C\varepsilon^{1/2}.$$

Proof. The proof follows from that of BMO Theorem 1 just proved and the John Nirenberg inequality. (See (2-4).) □

5. A refined BMO theorem

BMO Theorem 2. *Let $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ be an ε diffeomorphism and let $B \in \mathbb{R}^D$ be a ball. Then, there exists $T \in O(D)$ such that*

$$\frac{1}{\operatorname{vol} B} \int_B \left|\Phi'(x) - T\right| dx \le C\varepsilon. \tag{5-1}$$

Proof. We may assume without loss of generality that

$$B = B(0, 1). \tag{5-2}$$

We know that there exists $T_B^* \in O(D)$ such that

$$\left(\int_B |\Phi'(x) - T_B^*|^4 dx \right)^{1/4} \le C\varepsilon^{1/2}.$$

Our desired conclusion holds for Φ if and only if it holds for $(T_B^*)^{-1} o \Phi$. Hence without loss of generality, we may assume that $T_B^* = I$. Thus we have

$$\left(\int_B |\Phi'(x) - I)|^4 dx \right)^{1/4} \le C\varepsilon^{1/2}. \tag{5-3}$$

Let

$$\Omega(x) = \big(\Omega_1(x), \Omega_2(x), \dots, \Omega_D(x)\big) = \Phi(x) - x, \quad x \in \mathbb{R}^D. \tag{5-4}$$

Thus (5-3) asserts that

$$\left(\int_{B(0,1)} |\nabla \Omega(x)|^4 \, dx \right)^{1/4} \le C\varepsilon^{1/2}. \tag{5-5}$$

We know that

$$\left| (\Phi'(x))^T \Phi'(x) - I \right| \le C\varepsilon, \quad x \in \mathbb{R}^D. \tag{5-6}$$

In coordinates, $\Phi'(x)$ is the matrix $\left(\delta_{ij} + \dfrac{\partial \Omega_i(x)}{\partial x_j} \right)$; hence $\Phi'(x)^T \Phi'(x)$ is the matrix whose ij-th entry is

$$\delta_{ij} + \frac{\partial \Omega_j(x)}{\partial x_i} + \frac{\partial \Omega_i(x)}{\partial x_j} + \sum_l \frac{\partial \Omega_l(x)}{\partial x_i} \frac{\partial \Omega_l(x)}{\partial x_j}.$$

Thus (5-6) says that

$$\left| \frac{\partial \Omega_j}{\partial x_i} + \frac{\partial \Omega_i}{\partial x_j} + \sum_l \frac{\partial \Omega_l}{\partial x_i} \frac{\partial \Omega_l}{\partial x_j} \right| \le C\varepsilon \tag{5-7}$$

on $\mathbb{R}^D$, $i, j = 1, \dots, D$. Thus, we have from (5-5), (5-7) and the Cauchy–Schwarz inequality the estimate

$$\left\| \frac{\partial \Omega_i}{\partial x_j} + \frac{\partial \Omega_j}{\partial x_i} \right\|_{L^2(B(0,10))} \le C\varepsilon.$$

By the PDE Theorem, there exists, for each i, j, an antisymmetric matrix $S = (S)_{ij}$, such that

$$\left\| \frac{\partial \Omega_i}{\partial x_j} - S \right\|_{L^2(B(0,1))} \le C\varepsilon. \tag{5-8}$$

Recalling (5-4), this is equivalent to

$$\left\| \Phi' - (I + S) \right\|_{L^2(B(0,1))} \le C\varepsilon. \tag{5-9}$$

Note that (5-5) and (5-8) show that

$$|S| \le C\varepsilon^{1/2}$$

and thus,

$$|\exp(S) - (I + S)| \le C\varepsilon.$$

Hence, (5-9) implies via Cauchy–Schwarz.

$$\int_{B(0,1)} \left|\Phi'(x) - \exp(S)(x)\right| dx \le C\varepsilon^{1/2}. \tag{5-10}$$

This implies the result because S is antisymmetric, which means that $\exp S \in O(D)$. □

6. A BMO theorem for diffeomorphisms of small distortion

Theorem. *Let $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ be an ε distorted diffeomorphism. Let $B \subset \mathbb{R}^D$ be a ball. Then, there exists $T_B \in O(D)$ such that for every $\lambda \ge 1$,*

$$\operatorname{vol}\left\{x \in B : |\Phi'(x) - T_B| > C\lambda\varepsilon\right\} \le \exp(-\lambda)\operatorname{vol}(B). \tag{6-1}$$

Moreover, the result (6-1) is sharp in the sense of small volume if one takes a slow twist defined as follows: For $x \in \mathbb{R}^D$, let S_x be the block-diagonal matrix

$$\begin{pmatrix} D_1(x) & 0 & 0 & 0 & 0 & 0 \\ 0 & D_2(x) & 0 & 0 & 0 & 0 \\ 0 & 0 & \cdot & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdot & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdot & 0 \\ 0 & 0 & 0 & 0 & 0 & D_r(x) \end{pmatrix}$$

where, for each i, either $D_i(x)$ is the 1×1 identity matrix or else

$$D_i(x) = \begin{pmatrix} \cos f_i(|x|) & \sin f_i(|x|) \\ -\sin f_i(|x|) & \cos f_i(|x|) \end{pmatrix}$$

for a function f_i of one variable.

Now define for each $x \in \mathbb{R}^D$, $\Phi(x) = \Theta^T S_x(\Theta x)$ where Θ is any fixed matrix in $SO(D)$. One checks that Φ is ε-distorted, provided for each i, $t|f_i'(t)| < c\varepsilon$ for all $t \in [0, \infty)$.

Proof. The theorem follows from BMO Theorem 2 and the Nirenberg inequality. The sharpness can be easily checked. □

7. On the approximate and exact alignment of data in Euclidean space, speech and music manifolds

Approximate and exact alignment of data. A classical problem in geometry goes as follows. Suppose we are given two sets of D-dimensional data, that is, sets of points in Euclidean D-space, where $D \geq 1$. The data sets are indexed by the same set, and we know that pairwise distances between corresponding points are equal in the two data sets. In other words, the sets are isometric. Can this correspondence be extended to an isometry of the ambient Euclidean space?

In this form the question is not terribly interesting; the answer has long known to be yes (see [Wells and Williams 1975], for example). But a related question is actually fundamental in data analysis: here the known points are samples from larger, unknown sets — say, manifolds in $\mathbb{R}^D$ — and we seek to know what can be said about the manifolds themselves. A typical example might be a face recognition problem, where all we have is multiple finite images of people's faces from various views.

An added complication is that in general we are not given exact distances. We have noise and so we need to demand that instead of the pairwise distances being equal, they should be close in some reasonable metric. Some results on almost isometries in Euclidean spaces can be found in [John 1961; Alestalo et al. 2003].

In [Fefferman and Damelin $\geq$ 2012], the following two theorems are established which tell us about how to handle manifold identification when the point set function values given are not exactly equal but are close.

Theorem. *Given $\varepsilon > 0$ and $k \geq 1$, there exists $\delta > 0$ such that the following holds. Let $y_1, \dots, y_k$ and $z_1, \dots, z_k$ be points in $\mathbb{R}^D$. Suppose*

$$(1+\delta)^{-1} \leq \frac{|z_i - z_j|}{|y_i - y_j|} \leq 1+\delta, \quad i \neq j.$$

Then, there exists a Euclidean motion $\Phi_0 : x \to Tx + x_0$ such that

$$|z_i - \Phi_0(y_i)| \leq \varepsilon \operatorname{diam}\{y_1, \dots, y_k\}$$

for each i. If $k \leq D$, then we can take Φ_0 to be a proper Euclidean motion on $\mathbb{R}^D$.

Theorem. *Let $\varepsilon > 0$, $D \geq 1$ and $1 \leq k \leq D$. Then there exists $\delta > 0$ such that the following holds: Let $E := y_1, \dots, y_k$ and $E' := z_1, \dots z_k$ be distinct points in $\mathbb{R}^D$. Suppose that*

$$(1+\delta)^{-1} \leq \frac{|z_i - z_j|}{|y_i - y_j|} \leq (1+\delta), \quad 1 \leq i, j \leq k,\ i \neq j.$$

Then there exists a diffeomorphism $\Psi : \mathbb{R}^D \to \mathbb{R}^D$ with

$$(1+\varepsilon)^{-1} \le \frac{|\Psi(x)-\Psi(y)|}{|x-y|} \le (1+\varepsilon), \quad x, y \in \mathbb{R}^D,\ x \ne y$$

satisfying

$$\Psi(y_i) = z_i, \quad 1 \le i \le k.$$

The theorem above shows that any $1+\delta$ bilipchitz mapping Φ of $1 \le k \le D$ points from $\mathbb{R}^D$ to $\mathbb{R}^D$ may be extended to a $1+\varepsilon$ bilipchitz diffeomorphism of $\mathbb{R}^D$ to $\mathbb{R}^D$.

Given the two theorems above, we now need to ask ourselves. Can we take, in any particular data application, an ε-distorted map and replace it by a Euclidean motion or visa versa. Clearly this is very important since the theorems themselves provide in the once case a Euclidean motion and in the other a diffeomorphism of small distortion. We understand that our main BMO theorems tell us that at least in measure, diffeomorphisms of small distortion are very close to Euclidean motions motions.

Speech and music manifolds. Recently (see [Damelin and Miller 2012] and the references cited therein) there has been much interest in geometrically motivated dimensionality reduction algorithms. The reason for this is that these algorithms exploit low dimensional manifold structure in certain natural datasets to reduce dimensionality while preserving categorical content. In [Jansen and Niyogi 2006], the authors motivated the existence of low dimensional manifold structure to voice and speech sounds. As an immediate application of our results from this paper and from [Fefferman and Damelin $\ge$ 2012], we are now able to answer the following question related to speech and music manifolds. Suppose that we are given two collections of data functions in time which arise from vocal tract functions used in speech and music production. These manfolds exist; see the results of [Jansen and Niyogi 2006]. Suppose that all we know is that the functions are the same within a small δ distrortion. Then what can one say about the manifolds themselves. For example, can one identify different musical instruments or people/animals via speech using Euclidean motions or diffeomorphisms of ε distortion? What can one say about the differences in measure between the Euclidean motions or diffeomorphisms themselves? The theorems proved in this paper and in [Fefferman and Damelin $\ge$ 2012] provide a fascinating insight into these very interesting questions.

References

[Alestalo et al. 2003] P. Alestalo, D. A. Trotsenko, and J. Väisälä, "The linear extension property of bi-Lipschitz mappings", *Sibirsk. Mat. Zh.* **44**:6 (2003), 1226–1238. In Russian; translated as Linear bilipschitz extension property in *Sib. Math. J.* **44**:6 (2003), 959–968. MR 2004k:30042 Zbl 1063.30019

[Damelin and Miller 2012] S. B. Damelin and W. Miller, Jr., *The mathematics of signal processing*, Cambridge University Press, Cambridge, 2012. MR 2883645 Zbl 06002972

[Fefferman and Damelin $\geq$ 2012] C. Fefferman and S. B. Damelin, "On the approximate and exact alignment of δ distorted data in Euclidean space I", manuscript.

[Jansen and Niyogi 2006] A. Jansen and P. Niyogi, "Intrinsic Fourier analysis on the manifold of speech sounds", in *Acoustics, speech, and signal processing: proceedings of ICASSP* (Toulouse, 2006), IEEE, Piscataway, NJ, 2006.

[John 1961] F. John, "Rotation and strain", *Comm. Pure Appl. Math.* **14** (1961), 391–413. MR 25 #1672 Zbl 0102.17404

[John and Nirenberg 1961] F. John and L. Nirenberg, "On functions of bounded mean oscillation", *Comm. Pure Appl. Math.* **14** (1961), 415–426. MR 24 #A1348 Zbl 0102.04302

[Wells and Williams 1975] J. H. Wells and L. R. Williams, *Embeddings and extensions in analysis*, Ergebnisse der Mathematik und ihrer Grenzgebiete **84**, Springer, New York, 1975. MR 57 #1092 Zbl 0324.46034

Received: 2011-04-01 Revised: 2011-10-06 Accepted: 2011-10-08

cf@math.princeton.edu
Department of Mathematics, Princeton University,
Princeton, NJ 08544, United States

steve.damelin@gmail.com
Department of Mathematics, Wayne Country Day School,
Goldsboro, NC 27530, United States

Department of Mathematics and Physics, Wake Technical
Community College, Raleigh, NC 27603, United States

will25655@yahoo.com
Department of Music, Albany State University,
Albany, GA 31701, United States

INVOLVE 5:2 (2012)
dx.doi.org/10.2140/involve.2012.5.173

Modular magic sudoku

John Lorch and Ellen Weld

(Communicated by Kenneth S. Berenhaut)

A modular magic sudoku solution is a solution to a sudoku puzzle with symbols in {0, 1, . . . , 8} such that rows, columns, and diagonals of each subsquare add to 0 mod 9. We count these sudoku solutions by using the action of a suitable symmetry group and we also describe maximal mutually orthogonal families.

1. Introduction

1A. ***Terminology and goals.*** Upon completing a newspaper sudoku puzzle one obtains a *sudoku solution* of order nine, namely, a nine-by-nine array in which all of the symbols {0, 1, . . . , 8} occupy each row, column, and subsquare. For example, both

$$\begin{array}{|ccc|ccc|ccc|}\hline 7&2&3&1&8&5&4&6&0\\ 4&0&5&3&2&6&1&8&7\\ 6&1&8&4&0&7&2&3&5\\ \hline 1&7&0&6&3&2&5&4&8\\ 5&4&6&8&1&0&7&2&3\\ 8&3&2&7&5&4&0&1&6\\ \hline 2&6&4&0&7&8&3&5&1\\ 3&5&7&2&6&1&8&0&4\\ 0&8&1&5&4&3&6&7&2\\ \hline \end{array} \quad \text{and} \quad \begin{array}{|ccc|ccc|ccc|}\hline 1&8&0&7&5&6&4&2&3\\ 2&3&4&8&0&1&5&6&7\\ 6&7&5&3&4&2&0&1&8\\ \hline 8&4&6&5&1&3&2&7&0\\ 7&0&2&4&6&8&1&3&5\\ 3&5&1&0&2&7&6&8&4\\ \hline 5&1&3&2&7&0&8&4&6\\ 4&6&8&1&3&5&7&0&2\\ 0&2&7&6&8&4&3&5&1\\ \hline \end{array} \tag{1-1}$$

are sudoku solutions. The righthand array in (1-1) is a *modular magic sudoku solution*: in addition to satisfying the ordinary sudoku conditions, the rows, columns, and diagonals of each subsquare add to 0 mod 9. These subsquares are called *modular magic squares*. Plain magic squares of order 3 can't be cobbled into sudoku solutions but modular magic squares can.

One of our goals is to count the modular magic sudoku solutions. In Sections 2 and 3 we discuss properties and relabelings of modular magic squares; in Section 4

MSC2010: 05B15.
Keywords: sudoku, magic square, orthogonality, Latin square.

we introduce a natural symmetry group G acting on the set X of modular magic sudoku solutions and determine its structure. These ideas, coupled with a G-invariant property possessed by certain elements of X, are used to show that there are exactly two G-orbits on X (Theorem 4.3) and that there are 32256 modular magic sudoku solutions (Theorem 4.4).

Two sudoku solutions are *orthogonal* if upon superimposition there is no repetition in the resulting ordered pairs. The set of ordered pairs formed by superimposing the righthand sudoku solution in (1-1) and the solution x_2' given in Section 5 is

10	88	01	73	52	64	46	25	37
24	33	42	87	06	15	51	60	78
65	77	56	38	41	20	02	14	83
86	44	68	50	17	32	23	71	05
72	00	27	45	63	81	18	36	54
31	55	13	04	28	76	67	82	40
53	11	35	26	74	08	80	47	62
48	66	84	12	30	57	75	03	21
07	22	70	61	85	43	34	58	16

.

One can check directly that the two solutions are orthogonal; each is called an *orthogonal mate* of the other. On the other hand, the lefthand sudoku solution in (1-1) is not orthogonal to *any* sudoku solution (or to any Latin square, for that matter). A collection of sudoku solutions is said to be *mutually orthogonal* if every pair of distinct members is orthogonal.

Another of our goals is to investigate the orthogonality of modular magic sudoku solutions. In Section 5 we show that every modular magic sudoku solution possesses an orthogonal modular magic sudoku mate and that each such pair forms a largest possible family of mutually orthogonal modular magic sudoku solutions (Theorem 5.1).

1B. ***Background: Latin squares, orthogonality, and sudoku.*** A *Latin square* of order n is an $n \times n$ array with n symbols such that every symbol appears in each row and column. Latin squares have been of mathematical interest for hundreds of years, at first in their own right (for example, Euler's 36 officers problem; see [Euler 1923; Ball and Coxeter 1987]) and then in concert with other mathematical structures when it was discovered in the early 20th century that Latin squares are intimately connected with statistical design, coding theory, finite geometry, and graph theory. (See [Colbourn and Dinitz 1996; Dénes and Keedwell 1974; Roberts 1984] for more information.) A classical theorem illustrating some of these connections, largely due to Bose [1938], is:

Theorem 1.1. *Let m be an integer with $m \geq 2$. The following are equivalent*:

(a) *There is a collection of $m-1$ mutually orthogonal Latin squares of order m.*

(b) *There is a finite projective plane of order m.*

(c) *There exists a symmetric balanced incomplete block design with the type $(m^2+m+1, m+1, 1)$.*

The theorem indicates that counting Latin squares is of fundamental importance. The exact number of Latin squares of order nine (approximately 5.52×10^{27}; see [Bammel and Rothstein 1975]) wasn't known until 1975, and the exact number for orders twelve and larger is currently unknown. Regarding families of mutually orthogonal Latin squares, it has long been known that there are at most $n-1$ mutually orthogonal Latin squares of order n and that this bound is achieved when n is a prime power. However, for nonprime power orders larger than six, the largest size of a family of mutually orthogonal Latin squares is unknown. This open problem has been proposed by Mullen [1995] as a candidate for the "next Fermat problem."

Sudoku solutions, being special types of Latin squares, inherit both the legacy and the problems associated with Latin squares. In [Felgenhauer and Jarvis 2006] and [Jarvis and Russell 2006], using computer-aided arguments, it has been shown that there are 6670903752021072936960 distinct sudoku solutions of order nine and 5472730538 orbits under the action of a natural symmetry group (consisting of rotations, relabelings, *et cetera*), respectively. Moving on to orthogonal families of sudoku solutions, it is known that there are at most $n(n-1)$ mutually orthogonal sudoku solutions of order n^2; this bound is achieved when n is a prime power. More generally it has recently been shown (for example, [Bailey et al. 2008]) that if $p_1^{k_1} \dots p_s^{k_s}$ is the prime factorization of n and $q = \min\{p_i^{k_i}\}$, then there is a family of $q(q-1)$ mutually orthogonal sudoku solutions of order n^2. As in the case of Latin squares, the maximum size of a family of mutually orthogonal sudoku solutions is unknown in general. Given the difficulty of these counting problems, it is desirable to understand tractable settings such as modular magic sudoku thoroughly so that they can be used as a testing ground for new counting methods.

1C. ***Miscellaneous remarks.*** In addition to modular magic sudoku, both *magidoku* and *quasimagic sudoku* (each described in [Forbes 2007] and certain of the latter painstakingly counted in [Jones et al. 2011]) are types of sudoku solutions characterized by additional sum conditions on the subsquares. Also, our modular magic squares are equivalent (in order three) to the pseudomagic, modular magic squares considered by Evans [1996], provided that one adds a diagonal condition to Evans' definition.

2. Properties of modular magic squares

Before investigating modular magic sudoku, we establish a few properties of modular magic squares. For example, all of the modular magic squares presented thus far have the entries $\{0, 3, 6\}$ on a diagonal; this is not coincidental. Throughout we let $U = \{1, 2, 4, 5, 7, 8\}$ and $D = \{0, 3, 6\}$ be subsets of $\{0, 1, \dots, 8\}$, and we let the *remainder square* associated to a modular magic square consist of remainders $\bmod 3$ of the original entries. We often identify $\{0, 1, \dots, 8\}$ with the ring $\mathbb{Z}_9$.

Lemma 2.1. *A remainder square associated to a modular magic square must be a Latin square.*

Proof. Given a modular magic square, we make the following observations about its remainder square:

(a) Each of the symbols $\{0, 1, 2\}$ must appear exactly three times in the remainder square.

(b) The rows, columns, and diagonals of the remainder square must each add to $0 \bmod 3$ or else the rows, columns, and diagonals of the original modular magic square won't sum to $0 \bmod 9$.

(c) No row or column can consist of the same symbol.

Item (a) must hold because there are exactly three numbers in $\mathbb{Z}_9$ possessing each of the three possible remainders $\bmod 3$. Item (b) must hold or else the rows, columns, and diagonals of the original modular magic square won't sum to $0 \bmod 9$. Regarding item (c), rows or columns of 1s or 2s in the remainder square lead to sums of the form $7+4+1$ and $8+5+2$, respectively, in the original modular magic square; neither is equal to 0 in $\mathbb{Z}_9$. In view of items (a) and (b), a row or column of 0s in the remainder square implies a row or column of 1s, which is not allowed.

These observations imply that the remainder square is Latin: item (a) says that we have an order-three grid with three symbols each appearing three times. Further, if there is repetition of symbols in a given row or column then item (b) forces that row or column to consist of all the same symbol, thus violating item (c). □

Proposition 2.2. *In any modular magic square the elements of D must lie on a diagonal.*

Proof. We first show that the central entry of a given modular magic square must lie in D. Suppose otherwise that $\alpha \in U$ occupies the central location. Since α is not a zero divisor in $\mathbb{Z}_9$, it follows that $-(2^{-1}\alpha)$ is distinct from α. Therefore, α, $-(2^{-1}\alpha)$, and a third element of $\mathbb{Z}_9$ must form a row, column, or diagonal of the square. But the zero sum condition forces this third element to be $-(2^{-1}\alpha)$, contradicting the uniqueness of symbols in a modular magic square.

Then, given a modular magic square, the fact that an element of D lies in the center together with Lemma 2.1 indicate that the associated remainder square must be Latin with a 0 in the center. This means that all the 0s in the remainder square must occupy one of the diagonals, and so the elements of D must lie on this same diagonal in the original modular magic square. □

Finally, we observe that a modular magic square is uniquely determined by a choice of diagonal type (either "main" or "off"), elements of D to occupy this diagonal, and one element of U occupying a location away from the chosen diagonal. All of the remaining entries of the square can be filled in via the zero sum condition. This gives $2 \times 6 \times 6 = 72$ modular magic squares.

3. Modular magic relabelings

Ultimately we will use a group generated by certain grid symmetries and relabelings to count modular magic sudoku solutions. As opposed to ordinary sudoku, we cannot allow all relabelings because not every relabeling preserves modular magic squares. For example, the relabeling that swaps 0 and 1 and leaves everything else fixed is not allowable, as when

$$\boxed{\begin{matrix}4&8&6\\2&0&7\\3&1&5\end{matrix}} \quad \text{becomes} \quad \boxed{\begin{matrix}4&8&6\\2&1&7\\3&0&5\end{matrix}}\,.$$

Our purpose here is to describe the collection of *modular magic relabelings*, namely, those bijections of $\mathbb{Z}_9$ onto itself that preserve modular magic squares. We begin by making a few observations:

Lemma 3.1. *Let S denote the group of magic relabelings.*

(a) *Members of S become permutations of D when restricted to D.*

(b) *Given a permutation μ of $\{0, 3, 6\}$ and $\lambda \in U$, there is at most one $\sigma \in S$ with $\sigma|_D = \mu$ and $\sigma(\lambda) = 1$.*

(c) $|S| \leq 36$.

Proof. Part (a) must hold or else the action of such a relabeling on a modular magic square can produce a square having a member of U in its central location, contradicting Proposition 2.2. For part (b), more than one such σ would imply the existence of multiple modular magic squares possessing the data described immediately after Proposition 2.2, again a contradiction. Part (c) follows from part (b): we have $|S| \leq |S_3| \times |U| = 36$. □

Let's produce some magic relabelings. Given $k \in U$ and $l \in D$, consider the mapping $\mu_{k,l} : \mathbb{Z}_9 \to \mathbb{Z}_9$ defined by $\mu_{k,l}(n) = kn + l$. Let $H = \{\mu_{k,l} \mid k \in U, l \in D\}$, and it is not too difficult to see that H is an order-18 subgroup of S. In addition to H there are rather less obvious magic relabelings. For example, consider the mapping $\rho : \mathbb{Z}_9 \to \mathbb{Z}_9$ defined[1] by

$$\rho(n) = \begin{cases} 2n^{-1} & \text{if } n \in U, \\ n & \text{if } n \in D. \end{cases}$$

To see that ρ preserves the magic sum property, if $m, n \in U$ and $a \in D$ with $m + n + a = 0$ (that is, $\{m, n, a\}$ make a typical row/column/diagonal triple), then

$$\begin{aligned} \rho(m) + \rho(n) + \rho(a) &= 2m^{-1} + 2n^{-1} + a \\ &= (mn)^{-1}(2m + 2n + mna) \\ &= (mn)^{-1}\big((2m + 2n + mna) + (m + n + a)\big) \\ &= (mn)^{-1}\big(3(m + n) + (mn + 1)a\big) \\ &= (mn)^{-1}(0 + 0) = 0, \end{aligned}$$

where we've used the facts that $m + n + a = 0$ in $\mathbb{Z}_9$ implies $m + n \equiv 0 \bmod 3$ and that $mn \equiv 2 \bmod 3$ for all $m, n \in U$. All told, these relabelings generate S:

Proposition 3.2. *The group S of modular magic relabelings is generated by the set $\{\mu_{k,l}, \rho \mid k \in U, l \in D\}$ and is isomorphic to $(S_3 \times \mathbb{Z}_3) \rtimes \mathbb{Z}_2$.*

Proof. Using the fact that $\mu_{k,l} \circ \rho = \rho \circ \mu_{k^{-1},l}$, which we verify at the end of the proof, we know $H \rtimes \mathbb{Z}_2$ is a subgroup of S and so $|S| \geq 36$. But Lemma 3.1 says $|S| \leq 36$, so we conclude that $|S| = 36$ and that $S = \langle \mu_{k,l}, \rho \rangle \cong H \rtimes \mathbb{Z}_2$. Regarding H, observe that $|\mu_{1,6}| = 3$, $|\mu_{8,0}| = 2$, and $\mu_{1,6} \circ \mu_{8,0} = \mu_{8,0} \circ \mu_{1,6}^{-1}$. Therefore, these two elements generate a copy of S_3 within H. Likewise, $\mu_{4,0}$ generates a copy of $\mathbb{Z}_3$ in H that commutes with and has trivial intersection with the previously mentioned copy of S_3. Therefore, the direct product of these groups is an order-18 subgroup of H; this subgroup must be the entirety of H because $|H| = 18$. We conclude that $H \cong S_3 \times \mathbb{Z}_3$ and that $S \cong (S_3 \times \mathbb{Z}_3) \rtimes \mathbb{Z}_2$.

Finally, we verify that $\mu_{k,l} \circ \rho = \rho \circ \mu_{k^{-1},l}$. Note that

$$\mu_{k,l} \circ \rho(n) = \begin{cases} kn + l & \text{if } n \in D, \\ 2kn^{-1} + l & \text{if } n \in U, \end{cases}$$

while

$$\rho \circ \mu_{k^{-1},l}(n) = \begin{cases} k^{-1}n + l & \text{if } n \in D, \\ 2(k^{-1}n + l)^{-1} & \text{if } n \in U. \end{cases}$$

If $n \in D$, we require $kn + l = k^{-1}n + l$ in $\mathbb{Z}_9$, or equivalently $(k - k^{-1})n = 0$ in $\mathbb{Z}_9$. The latter statement is an immediate consequence of the facts that k and k^{-1}

[1] As a product of cycles $\rho = (12)(45)(78)$.

have the same remainder mod 3 and $3|n$. If on the other hand $n \in U$, we require $(k^{-1}n+l)^{-1} = kn^{-1} + 2^{-1}l$ in $\mathbb{Z}_9$. This follows from

$$\begin{aligned}(k^{-1}n+l)(kn^{-1}+2^{-1}l) &= 1 + l\big(kn^{-1} + 2^{-1}(kn^{-1})^{-1}\big) + 2^{-1}l^2 \\ &= 1 + l(0) + 0 = 1,\end{aligned}$$

where for the latter equation $kn^{-1} + 2^{-1}(kn^{-1})^{-1} \equiv 0 \bmod 3$ and $l^2 = 0$ when $l \in D$. □

Since $|S| = 36$, all of the relabelings in part (b) of Lemma 3.1 are achieved:

Corollary 3.3. *Given $\lambda \in U$ and μ a permutation of D, there exists $\sigma \in S$ with $\sigma|_D = \mu$ and $\sigma(\lambda) = 1$.*

4. Counting modular magic sudoku solutions

We use a symmetry group G, called the *modular magic sudoku group*, to assist us in the task of counting modular magic sudoku solutions. We first describe the generators of this group and its action on the set X of modular magic sudoku solutions, then we count the number of G-orbits in X (this gives the number of "essentially different" modular magic sudoku solutions), and finally we count the total number of modular magic sudoku solutions.

4A. ***The modular magic sudoku group.*** The modular magic sudoku group G should consist of all reasonable grid transformations and relabelings that send one modular magic sudoku solution to another. We declare this group to have the generators:

- Modular magic relabelings. (Here a single relabeling is applied to each subsquare. Modular magic relabelings are discussed in Section 3 above.)
- Permutations of *large rows*. (A large row is a row of subsquares.)
- Swaps of the outer two rows within a given large row.
- Permutations of *large columns*. (A large column is a column of subsquares.)
- Swaps of the outer two columns within a large column.
- Transpose.

The first set of generators forms the group S of modular magic relabelings, whose structure we've already discussed in Section 3. The remaining generators form a group H of grid transformations (including rotations), and we have $G = H \times S$ because H, S commute and have trivial intersection.

We determine the structure of H. Observe that $H = [H_r \times H_c] \rtimes T$, where H_r denotes the subgroup of H generated by the large row and row permutations described, H_c is the analogous subgroup generated by column permutations, and

T is the two-element group generated by the transpose. The direct product arises because the groups H_r and H_c have trivial intersection and commute, while the semidirect product comes about as a result of the fact that $th_r = h_c t$ whenever t is transpose, $h_r \in H_r$, and $h_c \in H$, where h_c is obtained from h_r by simply replacing the word "row" by "column" in any generators used to produce h_r. Now H_r and H_c clearly have the same structure, so all that remains is to describe the structure of H_r, which we address in the following paragraphs.

Positions of rows within our sudoku grid can be labeled (a, b) where $a, b \in \mathbb{Z}_3$, a denotes the large row, and b denotes the row within a large row, with 1 denoting top, 0 denoting middle, and 2 denoting bottom for both large rows and rows within large rows. (This ordering of rows seems unnatural at the moment, but will suit our purpose.) The set of permutations of large rows is isomorphic to S_3, regarded as bijections of $\mathbb{Z}_3$ onto itself, with $\sigma(a, b) = \big(\sigma(a), b\big)$ for $\sigma \in S_3$. On the other hand, the set of swaps of outer rows within a given large row is isomorphic to $(\mathbb{Z}_3^*)^3 \cong \mathbb{Z}_2^3$ where if $s = (s_0, s_1, s_2) \in (\mathbb{Z}_3^*)^3$ then $s(a, b) = (a, s_a b)$.[2] (Here $s_a b$ is computed by multiplication in $\mathbb{Z}_3$.) Observe that S_3 acts on $\mathbb{Z}_2^3$ via $\sigma.s = (s_{\sigma^{-1}(0)}, s_{\sigma^{-1}(1)}, s_{\sigma^{-1}(2)})$ and that each such σ determines an automorphism $\phi_\sigma : \mathbb{Z}_2^3 \to \mathbb{Z}_2^3$. Further, if $\sigma \in S_3$ and $s \in \mathbb{Z}_2^3$ we have the commutation relation

$$\sigma s = (\sigma.s)\sigma \tag{4-1}$$

because

$$\begin{aligned}\sigma s(a, b) = \sigma(a, s_a b) = \big(\sigma(a), s_a b\big) &= \big(\sigma(a), s_{\sigma^{-1}(\sigma(a))} b\big)\\ &= \big(\sigma(a), (\sigma.s)_{\sigma(a)} b\big) = (\sigma.s)(\sigma(a), b) = (\sigma.s)\sigma(a, b).\end{aligned}$$

An example may help: According to our labeling, the $(0, 1)$ row is the top row within the middle large row. Further, if $s = (2, 2, 1) \in (Z_3^*)^2$ and $\sigma = (021) \in S_3$, then $\sigma.s = (2, 1, 2)$. Applying these to the $(0, 1)$ row we have

$$\sigma s(0, 1) = \sigma(0, s_0 \times 1) = \sigma(0, 2 \times 1) = \sigma(0, 2) = \big(\sigma(0), 2\big) = (2, 2). \tag{4-2}$$

Therefore, the outcome of $\sigma s(0, 1)$ is the bottom row within the bottom large row. Likewise, we have

$$(\sigma.s)\sigma(0, 1) = \sigma.s(2, 1) = \big(2, (\sigma.s)_2 \times 1\big) = (2, 2 \times 1) = (2, 2), \tag{4-3}$$

with the equality of (4-2) and (4-3) as required by (4-1).

Returning to the structure of H_r, the commutation relation (4-1) implies that $H_r \cong S_3 \ltimes \mathbb{Z}_2^3$ via

$$(f, \sigma)(g, \tau) = \big(f(\sigma.g), \sigma\tau\big),$$

[2] The simple action of $s \in (\mathbb{Z}_3^*)^3$ on a row (a, b) is facilitated by the strange ordering of rows given above.

where $f, g \in (\mathbb{Z}_3^*)^3 \cong \mathbb{Z}_2^3$ and $\tau, \sigma \in S_3$. Note here that S_3 is acting on multiple copies of $\mathbb{Z}_2$ (three copies), where S_3 acts to permute the copies of $\mathbb{Z}_2$ among themselves. Semidirect products of this type are known as *wreath products*: we denote $\mathbb{Z}_2^3 \rtimes S_3$ by $\mathbb{Z}_2 \text{ wr } S_3$. Summarizing the discussion above we have:

Proposition 4.1. *The modular magic sudoku group G is isomorphic to $S \times H$, where $S \cong (S_3 \times \mathbb{Z}_3) \rtimes \mathbb{Z}_2$ and $H \cong \big[(\mathbb{Z}_2 \, {}_{wr} \, S_3) \times (\mathbb{Z}_2 \, {}_{wr} \, S_3)\big] \rtimes \mathbb{Z}_2$. The order of this group is $|S| \times |H| = 36 \times (48 \times 48 \times 2) = 165888$.*

4B. *Orbits of the modular magic sudoku group.* We set about counting the G-orbits on X. Begin by declaring a modular magic sudoku solution to be in *proper form* if it has this aspect, as described by Lemma 4.2:

$$\begin{array}{|ccc|ccc|ccc|}
\hline
1 & 8 & 0 & & & 6 & & & 3 \\
2 & 3 & 4 & & 0 & & & 6 & \\
6 & 7 & 5 & 3 & & & 0 & & \\
\hline
 & & 6 & & & & & & \\
 & 0 & & & & & & & \\
3 & & & & & & & & \\
\hline
 & & 3 & & & & & & \\
 & 6 & & & & & & & \\
0 & & & & & & & & \\
\hline
\end{array} \;.$$

Lemma 4.2. *Every modular magic sudoku solution is in the same G-orbit as some proper form modular magic sudoku solution.*

Proof. Beginning with a modular magic sudoku solution, we apply the these group elements to produce something in proper form:

(a) Permute the large columns so that there is a 3 in the center of the upper left subsquare.

(b) Perform an outer row swap in the top large row and/or outer column swap in the left large column to place 0 in the upper right location of the upper left subsquare.

(c) Swap the middle and right large columns to place 0 in the center location of the upper middle subsquare.

(d) Make outer column swaps in the rightmost two large columns to make the $\{0, 3, 6\}$-diagonals go from lower left to upper right in the top rightmost two subsquares (rightmost two subsquares in the top large row).

(e) Swap the middle and bottom large rows so that 0 lies in the center location of the middle left subsquare (and 6 lies in the bottom left subsquare).

(f) Make outer row swaps in the bottommost two large rows to make the $\{0, 3, 6\}$-diagonals go from lower left to upper right in the leftmost bottom two subsquares (bottom two subsquares in the leftmost large column).

(g) Via Corollary 3.3 apply a modular magic relabeling to the resulting modular magic sudoku solution that fixes D and sends the upper leftmost symbol to 1. □

To count the number of proper form modular magic sudoku solutions, and thereby to determine an upper bound on the number of G-orbits, we first observe that in any proper form solution, such as

$$\begin{array}{|ccc|ccc|ccc|}
\hline
1 & 8 & 0 & a_1 & & 6 & a_2 & & 3 \\
2 & 3 & 4 & & 0 & & & 6 & \\
6 & 7 & 5 & 3 & & & 0 & & \\
\hline
a_3 & & 6 & & & & & & \\
& 0 & & & & & & & \\
3 & & & & & & & & \\
\hline
a_4 & & 3 & & & & & & \\
& 6 & & & & & & & \\
0 & & & & & & & & \\
\hline
\end{array}, \qquad (4\text{-}4)$$

each of the symbols a_1, a_2, a_3, a_4 shown in (4-4) has no more than two possible values. For example, in order for the first row to satisfy the Latin row condition, we know that a_1 and $-(a_1+6)$ cannot be 1 or 8, so $a_1 \in \{5, 7\}$. Further, one can check that values for a_1, a_2, a_3, a_4 either uniquely determine a proper form solution or lead to a contradiction of sudoku conditions. This implies that there are at most sixteen proper form solutions. A case-by-case check shows that seven of these sixteen are valid modular magic sudoku solutions and further that each of these seven is readily obtainable via G from one of the two proper form solutions:

$$x_1 = \begin{array}{|ccc|ccc|ccc|}
\hline
1 & 8 & 0 & 7 & 5 & 6 & 4 & 2 & 3 \\
2 & 3 & 4 & 8 & 0 & 1 & 5 & 6 & 7 \\
6 & 7 & 5 & 3 & 4 & 2 & 0 & 1 & 8 \\
\hline
7 & 5 & 6 & 4 & 2 & 3 & 1 & 8 & 0 \\
8 & 0 & 1 & 5 & 6 & 7 & 2 & 3 & 4 \\
3 & 4 & 2 & 0 & 1 & 8 & 6 & 7 & 5 \\
\hline
4 & 2 & 3 & 1 & 8 & 0 & 7 & 5 & 6 \\
5 & 6 & 7 & 2 & 3 & 4 & 8 & 0 & 1 \\
0 & 1 & 8 & 6 & 7 & 5 & 3 & 4 & 2 \\
\hline
\end{array} \quad \text{and} \quad x_2 = \begin{array}{|ccc|ccc|ccc|}
\hline
1 & 8 & 0 & 7 & 5 & 6 & 4 & 2 & 3 \\
2 & 3 & 4 & 8 & 0 & 1 & 5 & 6 & 7 \\
6 & 7 & 5 & 3 & 4 & 2 & 0 & 1 & 8 \\
\hline
8 & 4 & 6 & 5 & 1 & 3 & 2 & 7 & 0 \\
7 & 0 & 2 & 4 & 6 & 8 & 1 & 3 & 5 \\
3 & 5 & 1 & 0 & 2 & 7 & 6 & 8 & 4 \\
\hline
5 & 1 & 3 & 2 & 7 & 0 & 8 & 4 & 6 \\
4 & 6 & 8 & 1 & 3 & 5 & 7 & 0 & 2 \\
0 & 2 & 7 & 6 & 8 & 4 & 3 & 5 & 1 \\
\hline
\end{array}. \qquad (4\text{-}5)$$

This leads to the summary result:

Theorem 4.3. *There are exactly two G-orbits orbits in X. The modular magic sudoku solutions x_1 and x_2 can be taken as base points for these orbits.*

Proof. Our discussion up to this point indicates that there are at most two G-orbits. To finish we show that x_1 and x_2 from (4-5) must lie in different orbits. Recall that a *transversal* of a Latin square is a collection of locations that meets every row, column, and symbol exactly once. The property of possessing a transversal consisting of the diagonals of exactly three subsquares is a property that is invariant under the action of G: no modular magic sudoku group generator takes a central subsquare location to a noncentral subsquare location. We see that x_1 possesses such a transversal (the main diagonal) while x_2 does not. It follows that x_1 and x_2 must be in different G-orbits. □

4C. ***The total number of modular magic sudoku solutions.*** Let x_1 and x_2 be as in (4-5) and let G_{x_1} and G_{x_2} be the corresponding stabilizer subgroups of G (that is, $G_{x_i} = \{g \in G \mid g.x_i = x_i\}$). We introduce the notation:

- Large rows, and rows within large rows, are labeled 0, 1, and 2 from top to bottom. The same goes for columns, labeled left to right.[3]
- If σ is a permutation of $\{0, 1, 2\}$ then $\sigma_r, \sigma_c \in G$ denote the corresponding permutations of large rows and large columns, respectively, determined by σ.
- Let $s \in G$ denote the grid permutation that swaps the outer rows of every large row and the outer columns of every large column.
- Let $t \in G$ denote transpose.

We describe the structure of G_{x_1}. First observe that s is the only possible nontrivial combination of outer row/column swaps because any other nontrivial combination of these swaps when applied to x_1 yields a modular magic sudoku solution with some $\{0, 3, 6\}$ subsquare diagonal of the wrong type. This means that the possible generators of G_{x_1} have been reduced to relabelings, permutations of large rows/columns, s, and t. If $g \in G_{x_1}$ has no contribution from s or t, then the large row/column permutations must be *even*, or else $g.x_1$ is not in proper form. Likewise, if there is contribution from s or t (possibly both), then the large row/column permutations must be odd. This allows us to further narrow the possible generators for G_{x_1}, and, upon checking the possibilities, we find that all of the "allowable" large row/column permutations (in the sense of the previous two

[3] This is different from the ordering presented in Section 4A.

sentences) actually appear in elements of G_{x_1}. We therefore have

$$\begin{aligned} G_{x_1} &= \langle \mu_{1,6}(012)_c,\, \mu_{1,6}(012)_r,\, \rho\mu_{5,6}(12)_r(12)_c t,\, \mu_{8,6}(12)_r(12)_c s\rangle \\ &\cong (\mathbb{Z}_3 \times \mathbb{Z}_3) \rtimes (\mathbb{Z}_2 \times \mathbb{Z}_2). \end{aligned}$$

A similar analysis can be applied to G_{x_2}, which has the same "allowable" large row/column permutations, but here fewer of them actually work. Upon checking we have

$$G_{x_2} = \langle \mu_{1,6}(012)_c,\, \mu_{8,6}(12)_r(12)_c s\rangle \cong S_3,$$

a subgroup of G_{x_1}.

Theorem 4.4. *There are* 32256 *modular magic sudoku solutions.*

Proof. Let $G.x_i$ denote the G-orbit in X through x_i. From the discussion immediately above we have

$$|G.x_i| = \frac{|G|}{|G_{x_1}|} = \frac{165888}{36} = 4608 \quad \text{while} \quad |G.x_2| = \frac{|G|}{|G_{x_2}|} = \frac{165888}{6} = 27648.$$

The total number of modular magic sudoku solutions is $|G.x_1| + |G.x_2| = 32256$. □

5. Orthogonality of modular magic sudoku solutions

Here we investigate the orthogonality of modular magic sudoku solutions. We begin by observing that the solutions x_1' and x_2' given in (5-1) are modular magic and are orthogonal to the solutions x_1 and x_2 given in (4-5), respectively.

$$x_1' = \begin{array}{|ccc|ccc|ccc|} \hline 0&8&1&3&2&4&6&5&7 \\ 4&3&2&7&6&5&1&0&8 \\ 5&7&6&8&1&0&2&4&3 \\ \hline 6&5&7&0&8&1&3&2&4 \\ 1&0&8&4&3&2&7&6&5 \\ 2&4&3&5&7&6&8&1&0 \\ \hline 3&2&4&6&5&7&0&8&1 \\ 7&6&5&1&0&8&4&3&2 \\ 8&1&0&2&4&3&5&7&6 \\ \hline \end{array} \quad \text{and} \quad x_2' = \begin{array}{|ccc|ccc|ccc|} \hline 0&8&1&3&2&4&6&5&7 \\ 4&3&2&7&6&5&1&0&8 \\ 5&7&6&8&1&0&2&4&3 \\ \hline 6&4&8&0&7&2&3&1&5 \\ 2&0&7&5&3&1&8&6&4 \\ 1&5&3&4&8&6&7&2&0 \\ \hline 3&1&5&6&4&8&0&7&2 \\ 8&6&4&2&0&7&5&3&1 \\ 7&2&0&1&5&3&4&8&6 \\ \hline \end{array}. \qquad (5\text{-}1)$$

The selection of x_1' and x_2' is not entirely random. For example, one can see that the $\{0, 3, 6\}$ subsquare diagonals for x_j and x_j' with $j \in \{1, 2\}$ must be of opposite types and that by applying a relabeling (Corollary 3.3) the upper left subsquare of

x'_j can be chosen to be

$$\boxed{\begin{array}{ccc}0&8&1\\4&3&2\\5&7&6\end{array}}\,.$$

Due to the fact that orthogonality is preserved under relabelings and grid symmetries, Theorem 4.3 implies that every modular magic sudoku solution possesses an orthogonal modular magic sudoku mate.

If M is a modular magic sudoku solution let $C(M)$ denote the Latin square of order 3 with symbols in D containing the subsquare centers of M. We note that if two modular magic sudoku solutions M_1 and M_2 are orthogonal then so are $C(M_1)$ and $C(M_2)$. Since two is the maximal size of a family of mutually orthogonal Latin squares of order 3, this observation implies that the maximal size of a family of mutually orthogonal modular magic sudoku solutions is at most two. Summarizing, we have:

Theorem 5.1. *Every modular magic sudoku solution has an orthogonal modular magic sudoku mate*; *such a pair forms a largest possible family of mutually orthogonal modular magic sudoku solutions.*

References

[Bailey et al. 2008] R. A. Bailey, P. J. Cameron, and R. Connelly, "Sudoku, gerechte designs, resolutions, affine space, spreads, reguli, and Hamming codes", *Amer. Math. Monthly* **115**:5 (2008), 383–404. MR 2408485 Zbl 1149.05010

[Ball and Coxeter 1987] W. W. R. Ball and H. S. M. Coxeter, *Mathematical recreations and essays*, 13th ed., Dover, New York, 1987. MR 88m:00013 Zbl 0029.19701

[Bammel and Rothstein 1975] S. E. Bammel and J. Rothstein, "The number of 9 × 9 Latin squares", *Discrete Math.* **11**:1 (1975), 93–95. MR 51 #7882 Zbl 0304.05007

[Bose 1938] R. C. Bose, "On the application of properties of Galois fields to the problem of construction of hyper-Græco-Latin squares", *Sankhyā* **3** (1938), 323–338.

[Colbourn and Dinitz 1996] C. J. Colbourn and J. H. Dinitz (editors), *The CRC handbook of combinatorial designs*, CRC, Boca Raton, FL, 1996. MR 97a:05001 Zbl 0836.00010

[Dénes and Keedwell 1974] J. Dénes and A. D. Keedwell, *Latin squares and their applications*, Academic Press, New York, 1974. MR 50 #4338 Zbl 0283.05014

[Euler 1923] L. Euler, "Recherches sur une nouvelle espèce de quarrés magiques", pp. 291–392 in *Opera omnia*, Series 1, Volume 7, Teubner, Leipzig and Berlin, 1923.

[Evans 1996] A. B. Evans, "Magic rectangles and modular magic rectangles", *J. Statist. Plann. Inference* **51**:2 (1996), 171–180. MR 97b:05035 Zbl 0851.05031

[Felgenhauer and Jarvis 2006] B. Felgenhauer and F. Jarvis, "Mathematics of sudoku I", *Math. Spectrum* **39**:1 (2006), 15–22.

[Forbes 2007] A. D. Forbes, "Quasi-magic sudoku puzzles", *M500* **215** (2007), 1–11.

[Jarvis and Russell 2006] F. Jarvis and E. Russell, "Mathematics of sudoku II", *Math. Spectrum* **39**:2 (2006), 54–58.

[Jones et al. 2011] S. K. Jones, S. Perkins, and P. A. Roach, "Properties, isomorphisms and enumeration of 2-quasi-magic sudoku grids", *Discrete Math.* **311**:13 (2011), 1098–1110. MR 2012b:05054 Zbl 1226.05064

[Mullen 1995] G. L. Mullen, "A candidate for the 'next Fermat problem'", *Math. Intelligencer* **17**:3 (1995), 18–22. MR 97a:05040 Zbl 0845.05020

[Roberts 1984] F. S. Roberts, *Applied combinatorics*, Prentice Hall, Englewood Cliffs, NJ, 1984. MR 85h:05001 Zbl 0547.05001

Received: 2011-05-02 Revised: 2011-08-24 Accepted: 2011-09-17

jlorch@bsu.edu *Department of Mathematical Sciences, Ball State University, Muncie, IN 47306, United States*

elweld@bsu.edu *Department of Mathematical Sciences, Ball State University, Muncie, IN 47306, United States*

INVOLVE 5:2 (2012)
dx.doi.org/10.2140/involve.2012.5.187

Distribution of the exponents of primitive circulant matrices in the first four boxes of $\mathbb{Z}_n$

Maria Isabel Bueno, Kuan-Ying Fang,
Samantha Fuller and Susana Furtado

(Communicated by Joseph Gallian)

We consider the problem of describing the possible exponents of n-by-n boolean primitive circulant matrices. It is well known that this set is a subset of $[1, n-1]$ and not all integers in $[1, n-1]$ are attainable exponents. In the literature, some attention has been paid to the gaps in the set of exponents. The first three gaps have been proven, that is, the integers in the intervals $[\frac{n}{2}+1, n-2]$, $[\frac{n}{3}+2, \frac{n}{2}-2]$ and $[\frac{n}{4}+3, \frac{n}{3}-2]$ are not attainable exponents. Here we study the distribution of exponents in between those gaps by giving the exact exponents attained there by primitive circulant matrices. We also study the distribution of exponents in between the third gap and our conjectured fourth gap. It is interesting to point out that the exponents attained in between the $(i-1)$-th and the i-th gap depend on the value of n mod i.

1. Introduction

A boolean matrix is a matrix over the binary boolean algebra $\{0, 1\}$. An n-by-n boolean matrix C is said to be circulant if each row of C (except the first one) is obtained from the preceding row by shifting the elements cyclically 1 column to the right. In other words, the entries of a circulant matrix $C = (c_{ij})$ are related in the manner: $c_{i+1,j} = c_{i,j-1}$, where $0 \leq i \leq n-2$, $0 \leq j \leq n-1$, and the subscripts are computed modulo n. The first row of C is called the generating vector. Here and throughout we number the rows and columns of an n-by-n matrix from 0 to $n-1$.

The set of all n-by-n boolean circulant matrices forms a multiplicative commutative semigroup C_n with $|C_n| = 2^n$ [Davis 1979; Lancaster 1969]. This semigroup

MSC2010: 05C25, 05C50, 11P70.

Keywords: exponent, primitive circulant matrix, basis of a cyclic group, order, box.

Bueno was supported by Dirección General de Investigación (Ministerio de Ciencia y Tecnología) of Spain under grant MTM2009-09281 and NSF grant DMS-0852065. Fang and Fuller were supported by NSF Grant DMS-0852065 for the REU program at the University of California, Santa Barbara. Furtado's contribution was done within the activities of Centro de Estruturas Lineares e Combinatórias da Universidade de Lisboa.

was thoroughly investigated by K. K.-H. Butler and J. R. Krabill [1974] and by S. Schwarz [1974].

An n-by-n boolean matrix C is said to be primitive if there exists a positive integer k such that $C^k = J$, where J is the n-by-n matrix whose entries are all ones and the product is computed in the algebra $\{0, 1\}$. The smallest such k is called the exponent of C, and we denote it by $\exp C$. Let us denote $E_n = \{\exp C : C \in \mathcal{C}_n,\ C \text{ is primitive}\}$.

In [Bueno et al. 2009] we stated the following question: Given a positive integer n, what is the set E_n?

The previous question can easily be restated in terms of circulant graphs or bases for finite cyclic groups, as we explain next.

Let C be a boolean primitive circulant matrix and let S be the set of positions corresponding to the nonzero entries in the generating vector of C (where the columns are counted starting with zero). C is the adjacency matrix of the circulant digraph $Cay(\mathbb{Z}_n, S)$. The vertex set of this graph is $\mathbb{Z}_n$ and there is an arc from u to $u+a \pmod n$ for every $u \in \mathbb{Z}_n$ and every $a \in S$. A digraph D is called primitive if there exists a positive integer k such that for each ordered pair a, b of vertices there is a directed walk from a to b of length k in D. The smallest such integer k is called the exponent of the primitive digraph D. Thus, a circulant digraph G is primitive if and only if its adjacency matrix is. Moreover, if they are primitive, they have the same exponent. Therefore, finding the set E_n is equivalent to finding the possible exponents of circulant digraphs of order n.

Let n be a positive integer and let S be a nonempty subset of the additive group $\mathbb{Z}_n$. For a positive integer k we denote by kS the set given by

$$kS = \{s_1 + \cdots + s_k \bmod n : s_i \in S\} \subset \mathbb{Z}_n.$$

The set kS is called the k-fold sumset of S.

The set S is said to be a basis for $\mathbb{Z}_n$ if there exists a positive integer k such that $kS = \mathbb{Z}_n$. The smallest such k is called the order of S, denoted by order(S). It is well known [Butler and Krabill 1974; Schwarz 1974] that the set $S = \{s_0, s_1, \ldots, s_r\} \subset \mathbb{Z}_n$ is a basis if and only if $\gcd(s_1 - s_0, \ldots, s_r - s_0, n) = 1$. In [Bueno et al. 2009] we proved that, given a matrix C in $\mathcal{C}_n$, if S is the set of positions corresponding to the nonzero entries in the generating vector of C, then C is primitive if and only if S is a basis for $\mathbb{Z}_n$. Moreover, if C is primitive, then $\exp(C) = \text{order}(S)$. Therefore, finding the set E_n is equivalent to finding the possible orders of bases for the cyclic group $\mathbb{Z}_n$. This question is quite interesting by itself. We note that all the results in this paper will be given in terms of bases for $\mathbb{Z}_n$, as the techniques we can use following this approach result more convenient.

The problem we study in this paper has applications in different areas. In particular, circulant matrices appear as transition matrices in Markov processes

[Chou et al. 2008]. Also, the problem stated in terms of bases for $\mathbb{Z}_n$ has applications in coding theory and quantum information [Klopsch and Lev 2009].

In the literature, the problem of computing all possible exponents attained by circulant primitive matrices or, equivalently, by circulant digraphs, has been considered. In particular, the following results were obtained. Here and throughout, $[a, b]$ denotes the set of positive integers in the real interval $[a, b]$. If $a > b$ then $[a, b] = \varnothing$.

Lemma 1 [Huang 1990; Wang and Meng 1997]. *If C is a primitive circulant matrix, then its exponent is either $n-1$, $\lfloor \frac{n}{2} \rfloor$, $\lfloor \frac{n}{2} \rfloor - 1$ or does not exceed $\lfloor \frac{n}{3} \rfloor + 1$. Moreover, $\exp C = n-1$ if and only if the number of nonzero entries in the generating vector of C is exactly 2.*

Lemma 2 [Dukes et al. 2010]. *For every $n \geq 3$, the sets $\left[\lfloor \frac{n}{4} \rfloor + 3, \lfloor \frac{n}{3} \rfloor - 2\right]$ and E_n are disjoint.*

All these results can be immediately translated into results about the possible orders of bases for a finite cyclic group.

Note that the only primitive matrix in C_2 is J_2, so $E_2 = \{1\}$. From now on, we assume that $n \geq 3$. In [Bueno et al. 2009] we presented a conjecture concerning the possible exponents attained by n-by-n boolean primitive circulant matrices which we restate here in a more precise way. We start with a definition.

Definition. Let j be a positive integer. We call the j^{th} box of $\mathbb{Z}_n$, and denote it by B_j, the set of positive integers

$$\left[\left\lfloor \frac{n}{j} \right\rfloor - 1, \left\lfloor \frac{n}{j} \right\rfloor + j - 2\right].$$

Conjecture 3. *If $C \in C_n$ is primitive, then*

$$\exp C \in \left[1, \lfloor \sqrt{n} \rfloor\right] \cup \bigcup_{j=1}^{\lfloor \sqrt{n} \rfloor} B_j .$$

In [Dukes et al. 2010], it was proven that if $C \in C_n$ is primitive and its exponent is greater than k for some positive integer k, then there exists d_k such that the exponent of C is within d_k of n/l for some integer $l \in [1, k]$. Notice that the result we present in Conjecture 3 produces gaps in the set of exponents which are larger than the ones encountered in [Dukes et al. 2010]. In fact, we have shown that the gaps in our conjecture should be maximal [Bueno and Furtado 2010]. We say that a gap A in E_n is maximal if $A' \cap E_n \neq \varnothing$ for any interval of integers $A' \subset [1, n-1]$, with A strictly contained in A'. In [Bueno and Furtado 2010], we proved that for each positive integer j, there is an integer n, such that $B_{j,n}$ is a maximal gap in

E_n. However, as stated in [Dukes et al. 2010], we remain far from a complete characterization of the possible exponents of $n \times n$ primitive circulant matrices.

Lemmas 1 and 2 above show the gaps between the first and second box, between the second and third box, and between the third and fourth box when these boxes do not overlap. Here we present the distribution of orders of bases within the first three boxes by showing what orders are attained and which ones are not. The results for the first and second box were already known [Huang 1990; Wang and Meng 1997] and we include them for completeness. We also study the order of bases in the fourth box by giving orders that are attained and we conjecture that those are, in fact, the exact orders in that box. In addition, we also prove that all integers in $[1, \lfloor\sqrt{n}\rfloor]$ are attained by bases of $\mathbb{Z}_n$.

This paper is organized as follows. In Section 2 we state our main results and prove them in Section 4. In Section 3 we state and prove several auxiliary results concerning the order of bases for $\mathbb{Z}_n$, which will be used to prove our main theorems. The order of several bases for $\mathbb{Z}_n$ with cardinality at most 4 that are relevant to our proofs is studied in the Appendix.

2. Main results

In this section, we give the exact orders attained by bases for $\mathbb{Z}_n$ in the first three boxes of $\mathbb{Z}_n$. We also give orders attained in the fourth box. Notice that the results for the first and second box were already known [Huang 1990; Wang and Meng 1997] but we include them for completeness. Finally, we state that all integers up to $\lfloor\sqrt{n}\rfloor$ are in E_n.

The result for the first box is an immediate consequence of Lemma 1.

Theorem 4 [Huang 1990]. *For all n,*

$$B_1 \subseteq E_n.$$

Concerning the second box, we have the following result obtained in [Huang 1990; Wang and Meng 1997]. In Section 4.1 we include a proof of it using the techniques for bases.

Theorem 5 [Huang 1990; Wang and Meng 1997]. *Let $n \geq 17$ be a positive integer.*

- *If n is even, then $B_2 \subseteq E_n$.*
- *If n is odd, then $B_2 \cap E_n = \lfloor \frac{n}{2} \rfloor$.*

The next two theorems are our main results and will be proven in Section 4. In our first result we assume a lower bound n_0 for n, which is the smallest value of n for which the theorem holds for all $n > n_0$. The possible orders in E_n, with $n < n_0$, appear in Tables 1 and 2. We observe that, for any n for which the box under study does not overlap with adjacent boxes, the theorem holds. We also notice that, though we have a lower bound for n in our results, when $n \equiv 0 \bmod j$, $j = 3, 4$, B_j is a subset of E_n, for all n.

Theorem 6. *Let $n \geq 45$ be a positive integer.*

- *If $n \equiv 0 \pmod 3$, then $B_3 \subseteq E_n$.*
- *If $n \equiv 1 \pmod 3$, then $B_3 \cap E_n = \left\{\left\lfloor \frac{n}{3} \right\rfloor + 1, \left\lfloor \frac{n}{3} \right\rfloor\right\}$.*
- *If $n \equiv 2 \pmod 3$, then $B_3 \cap E_n = \left\{\left\lfloor \frac{n}{3} \right\rfloor + 1\right\}$.*

Theorem 7. *Let $n \geq 16$ be a positive integer.*

- *If $n \equiv 0 \pmod 4$, then $B_4 \subseteq E_n$.*
- *If $n \equiv 1 \pmod 4$, then $\left\{\left\lfloor \frac{n}{4} \right\rfloor + 2, \left\lfloor \frac{n}{4} \right\rfloor + 1, \left\lfloor \frac{n}{4} \right\rfloor\right\} \subseteq E_n$.*
- *If $n \equiv 2 \pmod 4$ or $n \equiv 3 \pmod 4$, then $\left\{\left\lfloor \frac{n}{4} \right\rfloor + 2, \left\lfloor \frac{n}{4} \right\rfloor + 1\right\} \subseteq E_n$.*

Though we do not prove it, we conjecture that $\left\lfloor \frac{n}{4} \right\rfloor - 1 \notin E_n$ when $n \equiv 1 \pmod 4$ and $\left\lfloor \frac{n}{4} \right\rfloor - 1, \left\lfloor \frac{n}{4} \right\rfloor \notin E_n$ when $n \equiv 2, 3 \pmod 4$.

In Tables 1 and 2 we give the exact orders attained by bases for $\mathbb{Z}_n$ with $n = 2, 3, 4, \ldots, 104$. As the numerical experiments show, for each n there is a number of

n	E_n	n	E_n	n	E_n
2	1	23	$1\ldots8, 11, 22$	44	$1\ldots13, 15, 21, 22, 43$
3	$1, 2$	24	$1\ldots9, 11, 12, 23$	45	$1\ldots16, 22, 44$
4	$1, 2, 3$	25	$1\ldots9, 12, 24$	46	$1\ldots13, 15, 16, 22, 23, 45$
5	$1, 2, 4$	26	$1\ldots9, 12, 13, 25$	47	$1\ldots13, 16, 23, 46$
6	$1, 2, 3, 5$	27	$1\ldots10, 13, 26$	48	$1\ldots17, 23, 24, 47$
7	$1, 2, 3, 6$	28	$1\ldots10, 13, 14, 27$	49	$1\ldots14, 16, 17, 24, 48$
8	$1\ldots4, 7$	29	$1\ldots10, 14, 28$	50	$1\ldots14, 17, 24, 25, 49$
9	$1\ldots4, 8$	30	$1\ldots11, 14, 15, 29$	51	$1\ldots14, 16, 17, 18, 25, 50$
10	$1\ldots5, 9$	31	$1\ldots11, 15, 30$	52	$1\ldots15, 17, 18, 25, 26, 51$
11	$1\ldots5, 10$	32	$1\ldots11, 15, 16, 31$	53	$1\ldots15, 18, 26, 52$
12	$1\ldots6, 11$	33	$1\ldots12, 16, 32$	54	$1\ldots15, 17, 18, 19, 26, 27, 53$
13	$1\ldots6, 12$	34	$1\ldots12, 16, 17, 33$	55	$1\ldots15, 18, 19, 27, 54$
14	$1\ldots7, 13$	35	$1\ldots10, 12, 17, 34$	56	$1\ldots16, 19, 27, 28, 55$
15	$1\ldots7, 14$	36	$1\ldots13, 17, 18, 35$	57	$1\ldots16, 18, 19, 20, 28, 56$
16	$1\ldots8, 15$	37	$1\ldots13, 18, 36$	58	$1\ldots16, 19, 20, 28, 29, 57$
17	$1\ldots6, 8, 16$	38	$1\ldots11, 13, 18, 19, 37$	59	$1\ldots16, 20, 29, 58$
18	$1\ldots9, 17$	39	$1\ldots14, 19, 38$	60	$1\ldots17, 19, 20, 21, 29, 30, 59$
19	$1\ldots7, 9, 18$	40	$1\ldots14, 19, 20, 39$	61	$1\ldots17, 20, 21, 30, 60$
20	$1\ldots7, 9, 10, 19$	41	$1\ldots12, 14, 20, 40$	62	$1\ldots17, 21, 30, 31, 61$
21	$1\ldots8, 10, 20$	42	$1\ldots14, 15, 20, 21, 41$	63	$1\ldots17, 20, 21, 22, 31, 62$
22	$1\ldots8, 10, 11, 21$	43	$1\ldots12, 14, 15, 21, 42$	64	$1\ldots18, 21, 22, 31, 32, 63$

Table 1. Orders of bases for $\mathbb{Z}_n$.

n	E_n
65	1...14, 16, 17, 18, 22, 32, 64
66	1...18, 21, 22, 23, 32, 33, 65
67	1...18, 22, 23, 33, 66
68	1...19, 23, 33, 34, 67
69	1...19, 22, 23, 24, 34, 68
70	1...15, 17, 18, 19, 23, 24, 34, 35, 69
71	1...19, 24, 35, 70
72	1...20, 23, 24, 25, 35, 36, 71
73	1...20, 24, 25, 36, 72
74	1...20, 25, 36, 37, 73
75	1...16, 18, 19, 20, 24, 25, 26, 37, 74
76	1...21, 25, 26, 37, 38, 75
77	1...16, 18, 19, 20, 21, 26, 38, 76
78	1...21, 25, 26, 27, 38, 39, 77
79	1...18, 20, 21, 26, 27, 39, 78
80	1...17, 19, 20, 21, 22, 27, 39, 40, 79
81	1...22, 26, 27, 28, 40, 80
82	1...17, 19, 20, 21, 22, 27, 28, 40, 41, 81
83	1...19, 21, 22, 28, 41, 82
84	1...23, 27, 28, 29, 41, 42, 83
85	1...18, 20, 21, 22, 23, 28, 29, 42, 84
86	1...18, 20, 21, 22, 23, 29, 42, 43, 85
87	1...18, 20, 22, 23, 28, 29, 30, 43, 86
88	1...24, 29, 30, 43, 44, 87
89	1...20, 22, 23, 24, 30, 44, 88
90	1...19, 21, 22, 23, 24, 29, 30, 31, 44, 45, 89
91	1...21, 23, 24, 30, 31, 45, 90
92	1...19, 21, 22, 23, 24, 25, 31, 45, 46, 91
93	1...21, 23, 24, 25, 30, 31, 32, 46, 92
94	1...21, 23, 24, 25, 31, 32, 46, 47, 93
95	1...20, 22, 24, 25, 32, 47, 94
96	1...26, 31, 32, 33, 47, 48, 95
97	1...18, 20, 22, 24, 25, 26, 32, 33, 48, 96
98	1...22, 24, 25, 26, 33, 48, 49, 97
99	1...22, 25, 26, 32, 33, 34, 49, 98
100	1...21, 23, 24, 25, 26, 27, 33, 34, 49, 50, 99
101	1...23, 25, 26, 27, 34, 50, 100
102	1...21, 23, 25, 26, 27, 33, 34, 35, 50, 51, 101
103	1...19, 21, 22, 23, 26, 27, 34, 35, 51, 102
104	1...19, 21, 22, 23, 25, 26, 27, 28, 35, 51, 52, 103

Table 2. Orders of bases for $\mathbb{Z}_n$.

consecutive orders that can be attained by bases of $\mathbb{Z}_n$. Though we prove Theorem 8, according to our numerical experiments, we conjecture that at least all consecutive integers up to $2\sqrt{n}-2$ are attainable orders.

Theorem 8. *Let n be a positive integer. Then* $[1, \lfloor\sqrt{n}\rfloor] \subseteq E_n$.

Though this result is cited in [Dukes et al. 2010], it seems that the paper where its proof is said to be is not available.

3. Order of bases for $\mathbb{Z}_n$

Computing the order of bases for $\mathbb{Z}_n$ is, in general, a challenging task. In this section we introduce some results relative to the order of bases of $\mathbb{Z}_n$ that will be helpful when proving our main results.

To start with, let us notice that the order of a basis S is invariant under shifts and multiplication by a unit of $\mathbb{Z}_n$, that is, for $a \in \mathbb{Z}_n$ and b a unit of $\mathbb{Z}_n$

$$\operatorname{order}(S) = \operatorname{order}(S+a), \quad \text{and} \quad \operatorname{order}(S) = \operatorname{order}(b * S) \tag{1}$$

where $b*S = \{bs \bmod n : s \in S\}$. In particular, this result implies that the set of orders attained by bases of $\mathbb{Z}_n$ is the same as the set of orders attained by bases of $\mathbb{Z}_n$ containing 0.

We now state some known results about the order of a basis for $\mathbb{Z}_n$. The following lemma gives an upper bound on the cardinality of a basis when a lower bound on its order is known.

Lemma 9 [Klopsch and Lev 2009]. *Let $n \in \mathbb{N}$ and $\rho \in [2, n-1]$. Let S be a basis for $\mathbb{Z}_n$ such that* $\operatorname{order}(S) \geq \rho$. *Then*

$$|S| \leq \max\left\{\frac{n}{d}\left(\left\lfloor\frac{d-2}{\rho-1}\right\rfloor+1\right) : d|n, d \geq \rho+1\right\}.$$

In particular, for each fixed $k \in \mathbb{N}$, if $\operatorname{order}(S) \geq \frac{n}{k}$ *and $n \gg 0$, then $|S| \leq 2k$.*

The next lemma gives an upper and a lower bound on the order of some bases for $\mathbb{Z}_n$ with cardinality 3.

Lemma 10 [Bueno and Furtado 2010]. *Let $2 \leq b \leq n-1$. Then*

$$\left\lfloor\tfrac{n}{b}\right\rfloor \leq \operatorname{order}(\{0, 1, b\}) \leq \left\lfloor\tfrac{n}{b}\right\rfloor+b-2.$$

We now give the exact order of some particular bases for $\mathbb{Z}_n$ that will be needed later. The next lemma shows, in particular, that the largest element of the j-th box, $j \leq \sqrt{n}$, belongs to E_n for all n.

Lemma 11 [Bueno et al. 2009]. *For $j \in \{1, 2, \ldots, \lfloor\sqrt{n}\rfloor\}$,*

$$\operatorname{order}(\{0, 1, j\}) = \left\lfloor\tfrac{n}{j}\right\rfloor+j-2.$$

Lemma 12 [Bueno and Furtado 2010]. *Let $2 \leq j \leq \sqrt{n}$ be a positive integer. Then*

$$\operatorname{order}\left(\left\{0, 1, \left\lfloor\tfrac{n}{j}\right\rfloor+1\right\}\right) = \left\lfloor\tfrac{n}{j}\right\rfloor+j-2.$$

Lemma 13 [Bueno et al. 2009]. *Let $2 \leq r \leq n-1$ and $t = n-r\left\lfloor\frac{n}{r}\right\rfloor$. Then*

$$\operatorname{order}(\{0, 1, 2, \ldots, r-1, r\}) = \begin{cases}\left\lfloor\frac{n}{r}\right\rfloor & \text{if } t \leq 1,\\ \left\lfloor\frac{n}{r}\right\rfloor+1 & \text{if } t > 1.\end{cases}$$

Lemma 14. *Let $2 \leq r \leq n-2$. Then*

$$\operatorname{order}(\{0, 1, 2, \ldots, r-1, r+1\}) = \left\lfloor\frac{n}{r+1}\right\rfloor+1.$$

Proof. Let $S = \{0, 1, 2, \ldots, r-1, r+1\}$. It can be shown by induction on k that, for $k \geq 1$, $kS = [0, \ldots, k(r+1)-2] \cup \{k(r+1)\}$. Thus, $\operatorname{order}(S) = k$ if and only if k is the minimum integer such that $k(r+1)-2 \geq n-1$, which implies the result. □

Lemma 15. *Suppose that m is a divisor of n and let $1 \leq q < m \leq n$. Then*

$$\operatorname{order}\left(\bigcup_{i=0}^{q}(i+\langle m\rangle)\right) = \left\lceil \frac{m-1}{q} \right\rceil.$$

Proof. Let S be the basis in the statement. Note that $kS = \bigcup_{i=0}^{kq}(i+\langle m\rangle)$ for all $k \geq 1$. Therefore, the order of S equals the minimum k such that $kq \geq m-1$ and the result follows. □

As a consequence of the previous result, we obtain that, if j is a divisor of n, the smallest element of the j-th box is an element of E_n, since

$$\operatorname{order}\left(\langle n/j\rangle \cup (1+\langle n/j\rangle)\right) = n/j - 1.$$

Using canonical projections we can bound the order of some bases in a convenient way. Given $\mathbb{Z}_n$ and a proper divisor m of n, we denote by ϕ the canonical quotient map $\phi : \mathbb{Z}_n \to \mathbb{Z}_{n/m}$. We denote by $\operatorname{order}_n(S)$ the order of the basis S as a subset of $\mathbb{Z}_n$. The next result is well known. For that reason, we include it without proof.

Lemma 16. *Let m be a proper divisor of n. If S is a basis for $\mathbb{Z}_n$ that contains zero and an element of order m, then $\phi(S)$ is a basis for $\mathbb{Z}_{n/m}$ and*

$$\operatorname{order}_{n/m}(\phi(S)) \leq \operatorname{order}_n(S) \leq \operatorname{order}_{n/m}(\phi(S)) + m - 1.$$

The next corollaries are immediate consequences of the previous lemma and Lemma 1.

Corollary 1 [Huang 1990]. *Suppose m is a proper divisor of n and S is a basis for $\mathbb{Z}_n$ that contains zero and an element of order m. Then* $\operatorname{order}(S) \leq (n/m) + m - 2$.

Corollary 2. *Let S be a basis for $\mathbb{Z}_n$ and assume that S contains zero and an element of order* 2. *Then* $\operatorname{order}(S) \leq \left\lfloor \frac{n}{4} \right\rfloor + 1$ *or* $\operatorname{order}(S) \geq \left\lfloor \frac{n}{2} \right\rfloor - 1$.

Corollary 3. *Let S be a basis for $\mathbb{Z}_n$ and assume that S contains zero and an element of order* 3. *Then* $\operatorname{order}(S) \leq \left\lfloor \frac{n}{6} \right\rfloor + 2$ *or* $\operatorname{order}(S) \geq \left\lfloor \frac{n}{3} \right\rfloor - 1$.

The next technical lemma allows us to prove Corollary 4, which is a key result in the proof of our main theorems.

Lemma 17. *Let $j \geq 2$ be an integer and assume that*

$$b \in I_j = \left[\left\lfloor \frac{n}{j+1} \right\rfloor + 2, \left\lfloor \frac{n}{j} \right\rfloor - 1\right].$$

Then

$$\operatorname{order}(\{0, 1, b\}) \leq \left\lfloor \frac{n}{j+2} \right\rfloor + j.$$

Proof. Let $S = \{0, 1, b\}$. First we observe that $j+1 < (j+1)b - n < b$. We divide the proof into three cases.

Case 1: Assume b is even and $(j+1)b-n=b/2$. This implies that $(2j+1)b/2=n$ and, therefore, b is not a divisor of n. Since $(2j+1)b=2n$, then b is an element of $\mathbb{Z}_n$ of order $2j+1$. Then

$$\operatorname{order}(S)\le \frac{n}{2j+1}+2j-1\le \left\lfloor \frac{n}{j+2}\right\rfloor+j.$$

The inequality on the left follows from Corollary 1 while the right inequality follows after a few computations. Thus,

$$\left\lfloor \frac{n}{j+2}\right\rfloor+j=\left\lfloor \frac{(2j+1)(j+2+k)}{j+2}\right\rfloor+j\ge 3j+1+k=\frac{n}{2j+1}+2j-1.$$

Case 2: Assume $(j+1)b-n<b/2$. Let $k=j+1$ and $p=(j+1)b-n$. Clearly, $[0,k]\cup\{p\}\cup[b,b+k-1]\subseteq kS$. It can be shown by induction on q that

$$\bigcup_{i=0}^{q}[ip,ip+(q-i)k]\cup[b,b+qk-1]\subset qkS \tag{2}$$

and

$$\bigcup_{i=0}^{q-1}[ip+(k-1)b,ip+(k-1)b+(q-(i+1))k]\subset(qk-1)S. \tag{3}$$

Now assume that q is the largest integer such that $qp<b$, that is, $q=\lfloor b/p\rfloor$ and let $l=\max\{b-pq,\ p-k\}$. Note that $q\ge 2$. Also, the gaps between consecutive intervals in the unions in (2) and (3) have at most $l-1$ elements. Thus, we have

$$[0,b+j]\cup[jb,\ jb+(q-1)p+l]\subseteq(qk+l-1)S.$$

Moreover, $[0,\ jb+(q-1)p+l+j-1]\subseteq(qk+l-1+(j-1))S$. Since $n-jb=b-p$, we get that $(q-1)p+l+j\ge n-jb$ is equivalent to $l+j\ge b-qp$, which is true because of the definition of l. This implies

$$\operatorname{order}(S)\le qk+\max\{b-pq,\ p-k\}+j-2. \tag{4}$$

Let $b=pq+r$, $0\le r<p$ and $q_1=\lfloor rk/p\rfloor$. It is easy to show that

$$\max\{b-pq,\ p-k\}\le q_1+p-k \tag{5}$$

which implies

$$\operatorname{order}(S)\le\left\lfloor\frac{bk}{p}\right\rfloor+p-k+j-2. \tag{6}$$

Taking into account (6), to complete the proof it is sufficient to show that

$$\left\lfloor\frac{bk}{p}\right\rfloor+p-k+j-2\le\left\lfloor\frac{n}{j+2}\right\rfloor+j. \tag{7}$$

Let g be the function given by

$$g(b)=\frac{bk}{p}+p-3=\frac{n}{p}+p-2.$$

To see that (7) holds it is enough to note that $g(b)\le\frac{n}{j+2}+j$, or, equivalently,

$$b\in\left[\frac{n+j+2}{j+1},\frac{n+\frac{n}{j+2}}{j+1}\right].$$

Case 3: Assume $(j+1)b-n>b/2$. Note that $j=\lfloor\frac{n}{b}\rfloor$. Let $n=jb+r_3$, $0\le r_3<b$. Thus, $(j+1)b-n=b-r_3$. Clearly, $[0,\,j+1]\cup\{b-r_3\}\cup[b,\,b+j]\cup[jb,\,jb+1]\subseteq(j+1)S$. It can be shown by induction on j that

$$[0,\,qj+1]\,\cup\,\bigcup_{i=0}^{q-1}[b-(q-i)r_3,\,b-(q-i)r_3+ij]\cup[b,\,b+qj]\subset(qj+1)S \tag{8}$$

Denote by q the largest integer such that $qj+2\le b-qr_3$, that is, $q=\left\lfloor\frac{b-2}{j+r_3}\right\rfloor$. An argument similar to Case 2 implies that

$$\operatorname{order}(S)\le qj+\max\{r_3,\,b-q(j+r_3)-1\}+j-1. \tag{9}$$

Let $l=\max\{r_3,\,b-q(j+r_3)-1\}$. Now we show that

$$qj+l+j-1\le\left\lfloor\frac{j(b-1)}{j+r_3}\right\rfloor+j+r_3-1\le\left\lfloor\frac{n}{j+2}\right\rfloor+j. \tag{10}$$

To see the first inequality in (10), it is enough to note that, by definition of q, $q(j+r_3)<b-1$ and $b-1\le(q+1)(j+r_3)$. To see the second inequality in (10), let h be the function given by

$$h(b)=\frac{j(b-1)}{j+r_3}+j+r_3-1=\frac{n}{j+n-jb}+j+n-jb-2.$$

Then we see that

$$h(b)\le\frac{n}{j+2}+(j+2)-2 \text{ if and only if } j+n-jb\in\left[j+2,\,\frac{n}{j+2}\right].$$

Moreover, for $j+n-jb=\lfloor\frac{n}{j+2}\rfloor+1$, we get $\lfloor h(b)\rfloor=\lfloor\frac{n}{j+2}\rfloor+j$, since by [Bueno and Furtado 2010, Theorem 5.7], and taking into account that $j<\sqrt{n}$,

$$\left\lfloor\frac{n}{\lfloor\frac{n}{j+2}\rfloor+1}\right\rfloor=j+1.$$

Therefore, if $j+n-jb \in \left[j+2,\, \frac{n}{j+2}+1\right]$, or equivalently, if

$$b \in \left[\frac{n+j-1-\frac{n}{j+2}}{j},\, \frac{n-2}{j}\right], \tag{11}$$

then the second inequality in (10) holds. We finish the proof by showing that any b satisfying our assumptions is such that (11) holds. Note that, as $(j+1)b-n > \frac{b}{2}$, we have $2n/(2j+1) < b \le \left\lfloor \frac{n}{j} \right\rfloor - 1$. Thus, because $j \ge 2$, it follows that $b \le \frac{n}{j} - 1 \le \frac{n-2}{j}$. First we note that if $|I_j| \ge 2$, then

$$\frac{n+j-1-\frac{n}{j+2}}{j} \le \frac{2n}{2j+1} < b. \tag{12}$$

If $|I_j| = 1$, then $b = \lfloor n/j \rfloor - 1$. If (12) holds, we are done. Otherwise, it can be proven that

$$\frac{n+j-1-\frac{n}{j+2}}{j} \le b = \left\lfloor \frac{n}{j} \right\rfloor - 1. \qquad \square$$

From the previous lemma we obtain the next corollary, which includes some results presented in [Dukes et al. 2010] without proof.

Corollary 4. *Let* $n \ge 16$. *Suppose that* $2 \le b \le \left\lfloor \frac{n}{2} \right\rfloor + 1$.

(i) *If either* $b \notin \left\{2,\, 3,\, \left\lfloor \frac{n}{3} \right\rfloor,\, \left\lfloor \frac{n}{3} \right\rfloor + 1,\, \left\lfloor \frac{n}{2} \right\rfloor,\, \left\lfloor \frac{n}{2} \right\rfloor + 1\right\}$, *or* $b = \left\lfloor \frac{n}{3} \right\rfloor$ *and* $n \ne 0 \bmod 3$, *then* $\text{order}(\{0,1,b\}) \le \left\lfloor \frac{n}{4} \right\rfloor + 2$.

(ii) *If either* $b \in \left\{3,\, \left\lfloor \frac{n}{3} \right\rfloor + 1\right\}$, *or* $b = \left\lfloor \frac{n}{3} \right\rfloor$ *and* $n \equiv 0 \bmod 3$, *or* $b = \left\lfloor \frac{n}{2} \right\rfloor$ *with* n *odd*, *then* $\text{order}(\{0,1,b\}) = \left\lfloor \frac{n}{3} \right\rfloor + 1$.

(iii) *If either* $b \in \left\{2,\, \left\lfloor \frac{n}{2} \right\rfloor + 1\right\}$, *or* $b = \left\lfloor \frac{n}{2} \right\rfloor$ *and* n *is even*, *then* $\text{order}(\{0,1,b\}) = \left\lfloor \frac{n}{2} \right\rfloor$.

Proof. By Lemma 17, if $b \in \left[\left\lfloor \frac{n}{4} \right\rfloor + 2,\, \left\lfloor \frac{n}{3} \right\rfloor - 1\right] \cup \left[\left\lfloor \frac{n}{3} \right\rfloor + 2,\, \left\lfloor \frac{n}{2} \right\rfloor - 1\right]$, the order of $\{0,1,b\}$ is at most $\left\lfloor \frac{n}{4} \right\rfloor + 2$. By Lemma 10, if $4 \le b \le n/4$, then $\text{order}(\{0,1,b\}) \le \left\lfloor \frac{n}{4} \right\rfloor + 2$. By Lemma 12, $\text{order}\left\{0,\, 1,\, \left\lfloor \frac{n}{4} \right\rfloor + 1\right\} = \left\lfloor \frac{n}{4} \right\rfloor + 2$. If $b = \left\lfloor \frac{n}{3} \right\rfloor$ and $n \ne 0 \bmod 3$ then

$$\text{order}(\{0,1,b\}) = \begin{cases} \text{order}(1+3*\{0,1,b\}) = \text{order}(\{0,1,4\}) & \text{if } n \equiv 1 \bmod 3, \\ \text{order}(2+3*\{0,1,b\}) = \text{order}(\{0,2,5\}) & \text{if } n \equiv 2 \bmod 3, \end{cases}$$

and the result follows from Lemmas 19 and 20. Thus, (i) follows. If $b \in \left\{3,\, \left\lfloor \frac{n}{3} \right\rfloor + 1\right\}$ the result follows from Lemmas 12 and 14. If n is odd, then

$$\text{order}\left(\left\{0,\, 1,\, \left\lfloor \frac{n}{2} \right\rfloor\right\}\right) = \text{order}(1+2*\{0,1,b\}) = \{0,1,3\}$$

and the result follows from Lemma 14. If $n \equiv 0 \bmod 3$ and $b = n/3$, then, for $k \ge 1$,

$$kS = [0,k] \cup [n/3,\, n/3+k-1] \cup [2n/3,\, 2n/3+k-2]$$

(in $\mathbb{Z}$). The order of S is the smallest positive integer k such that $k-2+2n/3 \geq n-1$, that is, $k = 1+n/3$, which completes the proof of (ii). To prove (iii), note that, if n is even and $b = n/2$, then, for $k \geq 1$,

$$kS = [0, k] \cup [n/2, n/2+k-1]$$

(in $\mathbb{Z}$). Thus, the order of S is the smallest positive integer k such that $k-1+n/2 \geq n-1$, that is, $\mathrm{order}(S) = n/2$. If $b \in \left\{2, \left\lfloor \frac{n}{2} \right\rfloor + 1\right\}$, the result follows from Lemmas 12 and 13. □

4. Proofs of the main results

In this section we prove Theorems 5, 6, 7, and 8. To prove the first three results, we initially show that certain orders in each box are attained by giving examples of bases with such orders. Then, regarding the first two theorems, we prove that the remaining orders are not attained.

4.1. ***Proof of Theorem 5.*** In the next table, we give examples of bases attaining the orders in the second box according to Theorem 5. The results follow from Lemmas 13 and 15.

Second Box for $\mathbb{Z}_n$		
$n \equiv 0 \bmod 2$	$n \equiv 1 \bmod 2$	Order(S)
$S = \langle n/2 \rangle \cup (1+\langle n/2 \rangle)$	—	$\left\lfloor \frac{n}{2} \right\rfloor - 1$
$S = \{0, 1, 2\}$	$S = \{0, 1, 2\}$	$\left\lfloor \frac{n}{2} \right\rfloor$

We now assume that $n \geq 17$ and n is odd, and show that there is no basis $S \subseteq \mathbb{Z}_n$ such that $\mathrm{order}(S) = \left\lfloor \frac{n}{2} \right\rfloor - 1$.

Assume that $S \subset \mathbb{Z}_n$ is a basis such that $\mathrm{order}(S) = \left\lfloor \frac{n}{2} \right\rfloor - 1$. By Lemma 9, $|S| \leq 3$. Note that, by definition of basis, $|S| \geq 2$ and, by Lemma 1, $|S| \neq 2$ if $\mathrm{order}(S) \neq n-1$. Thus $|S| = 3$. Suppose $S = \{0, a, b\}$ where $a, b \in \mathbb{Z}_n$. If a had order $m \neq n$, then $3 \leq m < \left\lfloor \frac{n}{2} \right\rfloor$, since n is odd. By Corollary 1, this would imply that $\mathrm{order}(S) \leq m + n/m - 2 < \left\lfloor \frac{n}{2} \right\rfloor - 1$, as $n \geq 17$. Therefore, a must have order n. Then S has the same order as $a^{-1}S = \{0, 1, c\}$ for some $c \in \mathbb{Z}_n$. If $c > \left\lfloor \frac{n}{2} \right\rfloor + 1$, then S has the same order as $1 - a^{-1}S = \{0, 1, d\}$ with $d \leq \left\lfloor \frac{n}{2} \right\rfloor + 1$. Thus, we can assume that $c \leq \left\lfloor \frac{n}{2} \right\rfloor + 1$. Now using Corollary 4, we get $\mathrm{order}(S) \neq \left\lfloor \frac{n}{2} \right\rfloor - 1$, a contradiction.

4.2. ***Proof of Theorem 6.*** The next table gives examples of bases attaining the conjectured orders in the third box according to Theorem 6. The results follow from Lemmas 13–15.

Third Box for $\mathbb{Z}_n$			
$n \equiv 0 \bmod 3$	$n \equiv 1 \bmod 3$	$n \equiv 2 \bmod 3$	Order(S)
$S = \langle n/3 \rangle \cup (1 + \langle n/3 \rangle)$	—	—	$\lfloor \frac{n}{3} \rfloor - 1$
$S = \{0, 1, 2, 3\}$	$S = \{0, 1, 2, 3\}$	—	$\lfloor \frac{n}{3} \rfloor$
$S = \{0, 1, 3\}$	$S = \{0, 1, 3\}$	$S = \{0, 1, 3\}$	$\lfloor \frac{n}{3} \rfloor + 1$

The fact that, for $n \geq 45$, order$(S) \neq \lfloor \frac{n}{3} \rfloor - 1$, if $n \equiv 1 \bmod 3$, and order$(S) \notin \{\lfloor \frac{n}{3} \rfloor - 1, \lfloor \frac{n}{3} \rfloor\}$, if $n \equiv 2 \bmod 3$, follows from Lemma 18. Just note that, if order$(S) \in \{\lfloor \frac{n}{3} \rfloor - 1, \lfloor \frac{n}{3} \rfloor\}$, then, by Lemma 9, $|S| \leq 4$.

The statement of the next lemma is stronger than what is needed to prove Theorem 6. However, the techniques we developed before allowed us to get this result, which in turn is useful in the proof of Corollary 5.

Lemma 18. *Let $n \geq 45$ and suppose that 3 is not a divisor of n. Let S be a basis for $\mathbb{Z}_n$. If $|S| \leq 4$, then* order$(S) \leq \lfloor \frac{n}{4} \rfloor + 2$ *or* order$(S) \geq \lfloor \frac{n}{3} \rfloor$. *Moreover, if* order$(S) = \lfloor \frac{n}{3} \rfloor$, *then $n \equiv 1 \bmod 3$.*

Proof. Without loss of generality, assume $0 \in S$. Suppose that $n \not\equiv 0 \bmod 3$. Since S is a basis, $|S| > 1$. If $|S| = 2$, then order$(S) = n - 1 > \lfloor \frac{n}{3} \rfloor$. Suppose that $|S| = 3$ or $|S| = 4$. If S has an element whose order is not $1, 2, n/2$ nor n, then, by Corollary 1, the result follows since there can't be an element of order 3 or $n/3$. Thus, this element has order ≥ 4 or $\leq n/4$. Suppose that the order of the elements in S is $1, 2, n/2$, or n, where 2 and $n/2$ only occur when n is even. If S has an element of order 2, then the result follows from Corollary 2. If S does not contain an element of order 2, then necessarily it contains an element of order n. Moreover, by (1), if S has an element of order n, the basis S has the same order as some basis of the form $\{0, 1, a, b\}$. If $|S| = 3$, then we can assume that $S = \{0, 1, a\}$, with $1 < a \leq \lfloor \frac{n}{2} \rfloor + 1$. In this case, the result follows from Corollary 4. If $|S| = 4$, assume that $S = \{0, 1, a, b\}$ with $a \leq \lfloor \frac{n}{2} \rfloor + 1$. Since for $S' \subset S$, order$(S) \leq$ order(S'), we have

$$\text{order}(\{0, 1, a, b\}) \leq \min\{\text{order}(\{0, 1, a\}), \text{order}(\{0, 1, b\})\}. \tag{13}$$

Let

$$A_1 = \{2, 3, \lfloor \tfrac{n}{3} \rfloor + 1, \lfloor \tfrac{n}{2} \rfloor, \lfloor \tfrac{n}{2} \rfloor + 1\},$$
$$A_2 = \{2, 3, \lfloor \tfrac{n}{3} \rfloor + 1, \lfloor \tfrac{n}{2} \rfloor, \lfloor \tfrac{n}{2} \rfloor + 1, 1 - \lfloor \tfrac{n}{2} \rfloor, -\lfloor \tfrac{n}{3} \rfloor, -2, -1\}.$$

Note that $-\lfloor \frac{n}{2} \rfloor \in A_2$. Also, $1 - \lfloor \frac{n}{2} \rfloor \equiv \lfloor \frac{n}{2} \rfloor + 1 \bmod n$, for n even. If $a \notin A_1$ or $b \notin A_2$ then, by Corollary 4 and taking into account (13),

$$\text{order}(\{0, 1, a, b\}) \leq \min\{\text{order}(\{0, 1, a\}), \text{order}(\{0, 1, b\})\} \leq \lfloor \tfrac{n}{4} \rfloor + 2.$$

Recall that order($\{0, 1, 1-b\}$) = order($\{0, 1, b\}$). If $a \in A_1$ and $b \in A_2$, the 25 and 26 → result follows from Lemmas 22–26. □

The following result was presented in [Dukes et al. 2010]. However, the authors leave most of the details of the proof to the reader and we do not see clearly that the result follows from their proof. For that reason and for completeness we are including it in this paper.

Corollary 5. *Let S be a basis for $\mathbb{Z}_n$. Then*

$$\operatorname{order}(S) \notin \left[\left\lfloor \tfrac{n}{4} \right\rfloor + 3, \left\lfloor \tfrac{n}{3} \right\rfloor - 2\right].$$

Proof. Note that, for $n < 45$, the interval in the statement is empty. Assume that $n \geq 45$. Without loss of generality, suppose that $0 \in S$. If $S \subset \mathbb{Z}_n$ is a basis such that $\left\lfloor \frac{n}{4} \right\rfloor + 3 \leq \operatorname{order}(S)$, by Lemma 9, $|S| \leq 6$. Assume that $n \not\equiv 0 \bmod 3$. If $|S| = 5$ or $|S| = 6$, by [Bueno et al. 2009, Theorem 3.7], $\operatorname{order}(S) \leq \left\lfloor \frac{n}{4} \right\rfloor + 1$. If $|S| \leq 4$, by Lemma 18, $\operatorname{order}(S) \leq \left\lfloor \frac{n}{4} \right\rfloor + 2$ or $\operatorname{order}(S) \geq \left\lfloor \frac{n}{3} \right\rfloor$.

Now assume that $n \equiv 0 \bmod 3$. If $|S| = 3$, the result follows from Corollary 4. Suppose that $|S| \in \{4, 5, 6\}$. If $\left\lfloor \frac{n}{4} \right\rfloor + 3 \leq \operatorname{order}(S)$, by Corollary 1, the order of the elements in S must be $1, 2, 3, n/2, n/3$, or n. First note that S contains, or has the same order as a basis which contains, an element of order 2, 3 or n. In fact, if $|S| = 4$ and S does not have an element of order 2, 3 or n, then S has an element of order $n/2$ and an element of order $n/3$. Hence, $\{0, 2a, 3b\} \subseteq S$ for some $a, b \in \mathbb{Z}_n$. Since S is a basis, $3b - 2a$ is not an element of order $n/2$ nor $n/3$ as, otherwise, 6 would divide $2a$ or $3b$ and all elements of S would be multiples of 2 or multiples of 3. Thus, S has the same order as $S - 2a$, which has an element of order 2, 3 or n. A similar argument can be applied if $|S| = 5$ or $|S| = 6$. Thus, assume that S contains an element of order 2, 3 or n. If S contains an element of order 2 or 3, the result follows from Corollaries 2 and 3. Now suppose that S contains an element of order n and no elements of order 2 and 3. If either $n/3 + 1 \in S$ or n is even and $n/2 + 1 \in S$, then S can be transformed into a basis with the same order containing zero and an element of order 2 or 3 and we reduce the problem to the previous case. Let

$$\begin{aligned} A_1 &= \left\{2, 3, \left\lfloor \tfrac{n}{2} \right\rfloor, \left\lfloor \tfrac{n}{2} \right\rfloor + 1\right\}, \\ A_2 &= \left\{2, 3, \left\lfloor \tfrac{n}{2} \right\rfloor, \left\lfloor \tfrac{n}{2} \right\rfloor + 1, 1 - \left\lfloor \tfrac{n}{2} \right\rfloor, -2, -1\right\}. \end{aligned}$$

Assume that $S = \{0, 1, a, b, c, d\}$, with $a \leq \left\lfloor \frac{n}{2} \right\rfloor + 1$ and $b = c = d$ if $|S| = 4$, and $c = d$ if $|S| = 5$. Note that if $S' \subset S$ then $\operatorname{order}(S) \leq \operatorname{order}(S')$. If $a \notin A_1$ or, b,c, or $d \notin A_2$ the result follows from Corollary 4. Suppose that $a \in A_1$, $b, c, d \in A_2$ and if a, b, c or $d \in \left\{\left\lfloor \frac{n}{2} \right\rfloor, \left\lfloor \frac{n}{2} \right\rfloor + 1, 1 - \left\lfloor \frac{n}{2} \right\rfloor\right\}$ then n is odd. If $|S| = 4$, the result follows from Lemmas 22–26. If $|S| = 5$ or $|S| = 6$ the result follows from the Remark on page 204 by noting that S has a subset of cardinality 4 containing 0 and 1 which is not one of the exceptional bases and, therefore, $\operatorname{order}(S) \leq \left\lfloor \frac{n}{4} \right\rfloor + 2$. □

4.3. ***Proof of Theorem 7.*** The next table gives examples of bases attaining the orders in the fourth box of $\mathbb{Z}_n$ claimed in Theorem 7. The results follow from Lemmas 12–15.

Fourth Box for $\mathbb{Z}_n$				
$n \equiv 0 \bmod 4$	$n \equiv 1 \bmod 4$	$n \equiv 2 \bmod 4$	$n \equiv 3 \bmod 4$	Order(S)
$\langle n/4\rangle \cup (1+\langle n/4\rangle)$	—	—	—	$\lfloor \frac{n}{4} \rfloor - 1$
$\{0, 1, 2, 3, 4\}$	$\{0, 1, 2, 3, 4\}$	$\bigcup_{i=0}^{2}(i+\langle n/2\rangle)$	—	$\lfloor \frac{n}{4} \rfloor$
$\{0, 1, 2, 4\}$	$\{0, 1, 2, 4\}$	$\{0, 1, 2, 4\}$	$\{0,1,2,4\}$	$\lfloor \frac{n}{4} \rfloor + 1$
$\{0, 1, \frac{n}{4}+1\}$	$\{0, 1, \lfloor \frac{n}{4} \rfloor + 1\}$	$\{0, 1, \lfloor \frac{n}{4} \rfloor + 1\}$	$\{0, 1, \lfloor \frac{n}{4} \rfloor + 1\}$	$\lfloor \frac{n}{4} \rfloor + 2$

4.4. ***Proof of Theorem 8.*** If $n \leq 4$, the result follows from Table 1. Assume $n \geq 5$. Notice that $\mathbb{Z}_n$ is always a basis for $\mathbb{Z}_n$, which implies that $1 \in E_n$. Consider the set $S = \{0, 1, 2, \dots, r-1, r+1\}$ with $2 \leq r \leq n-2$. By Lemma 14, order$(S) = \lceil \frac{n+1}{r+1} \rceil$. For all $r \geq \sqrt{n} - 1$

$$\frac{n+1}{r+1} - \frac{n+1}{r+2} = \frac{n+1}{(r+1)(r+2)} = \frac{n+1}{r^2+3r+2} \leq \frac{n+1}{n+\sqrt{n}} < 1.$$

It can be easily seen that, for positive real numbers a and b, $\lceil a \rceil - \lceil b \rceil \leq \lceil a - b \rceil$. Thus, $\lceil \frac{n+1}{r+1} \rceil - \lceil \frac{n+1}{r+2} \rceil \leq 1$ for all $r \geq \sqrt{n} - 1$, which implies that all integers from 2 to

$$\left\lceil \frac{n+1}{\lceil \sqrt{n} \rceil - 1 + 1} \right\rceil$$

are attained orders. But $\left\lceil \frac{n+1}{\lceil \sqrt{n} \rceil} \right\rceil \geq \left\lceil \frac{n}{\lceil \sqrt{n} \rceil} \right\rceil \geq \lfloor \sqrt{n} \rfloor$ and the result follows.

Appendix: Gallery of bases and their orders

Here we provide the order of some particular bases that are necessary to prove the main results in this paper. We do not include all the proofs since many of them are similar.

Lemma 19. *For $n \geq 6$,* order$(\{0, 1, 4\}) = \lfloor \frac{n}{4} \rfloor + 2$.

Proof. Let $S = \{0, 1, 4\}$. It can be shown by induction on k that in $\mathbb{Z}$, for all $k \geq 2$,

$$kS = [0, 4k-6] \cup [4k-4, 4k-3] \cup \{4k\}.$$

Let $q = \lfloor \frac{n}{4} \rfloor$. Then

$$(q+1)S = [0, 4q-2] \cup [4q, 4q+1] \cup \{4q+4\}$$

and $[0, 4q+2] \subseteq (q+2)S$. Note that $4q+4 \neq 4q-1 \pmod{n}$, since $n \geq 6$. Thus, $(q+1)S \neq \mathbb{Z}_n \pmod{n}$. On the other hand, $4q+2 \geq n-1$. The result follows. □

Lemma 20. *For* $n \geq 6$, $\operatorname{order}(\{0, 2, 5\}) \leq \lfloor \frac{n}{5} \rfloor + 3$.

Lemma 21. *For* $n \geq 4$, $\operatorname{order}(\{0, 2, 3, 4\}) = \lfloor \frac{n}{4} \rfloor + 1$.

Bases of the form {0, 1, 2, *a*}

Lemma 22. *Let* $n \geq 21$. *Let* $a \in \{3, \lfloor \frac{n}{3} \rfloor + 1, \lfloor \frac{n}{2} \rfloor, \lfloor \frac{n}{2} \rfloor + 1, 1 - \lfloor \frac{n}{2} \rfloor, -\lfloor \frac{n}{3} \rfloor, -2, -1\}$ *and* $S = \{0, 1, 2, a\}$. *Then* $\operatorname{order}(S) \leq \lfloor \frac{n}{4} \rfloor + 2$ *or* $\operatorname{order}(S) \geq \lfloor \frac{n}{3} \rfloor - 1$. *Moreover, if* $n \equiv 1 \bmod 3$, *then* $\operatorname{order}(S) \neq \lfloor \frac{n}{3} \rfloor - 1$ *and if* $n \equiv 2 \bmod 3$, *then* $\operatorname{order}(S) \notin \{\lfloor \frac{n}{3} \rfloor - 1, \lfloor \frac{n}{3} \rfloor\}$.

Proof. Case 1: If $a \in \{3, -1\}$, then the basis S has the same order as $\{0, 1, 2, 3\}$ and the result follows by Lemma 13.

Case 2: If $a = -2$, then S has the same order as $2 + S = \{0, 2, 3, 4\}$ and the result follows from Lemma 21.

Case 3: Suppose that $a \in \{\lfloor \frac{n}{2} \rfloor, \lfloor \frac{n}{2} \rfloor + 1, 1 - \lfloor \frac{n}{2} \rfloor\}$. Assume n is even. Note that $1 - \lfloor \frac{n}{2} \rfloor = \lfloor \frac{n}{2} \rfloor + 1$. In this case, S contains an element of order 2 or it has the same order as a basis containing 0 and an element of order 2. Thus, the result follows from Corollary 2. Assume n is odd. Then

$$\operatorname{order}\left(\{0, 1, 2, \lfloor \tfrac{n}{2} \rfloor\}\right) = \operatorname{order}(\{0, 1, 3, 5\}) \leq \operatorname{order}(\{0, 1, 5\}),$$
$$\operatorname{order}\left(\{0, 1, 2, \lfloor \tfrac{n}{2} \rfloor + 1\}\right) = \operatorname{order}(\{0, 1, 2, 4\}) \leq \operatorname{order}(\{0, 1, 4\}).$$

In both cases, $\operatorname{order}(S) \leq \lfloor \frac{n}{4} \rfloor + 2$ by Corollary 4. Also,

$$\operatorname{order}\left(\{0, 1, 2, \lfloor \tfrac{n}{2} \rfloor + 2\}\right) = \operatorname{order}(\{0, 2, 3, 4\}) \leq \lfloor \tfrac{n}{4} \rfloor + 2$$

by Lemma 21. Note that $1 - \lfloor \frac{n}{2} \rfloor = \lfloor \frac{n}{2} \rfloor + 2$.

Case 4: Suppose that $a \in \{-\lfloor \frac{n}{3} \rfloor, \lfloor \frac{n}{3} \rfloor + 1\}$. If $n \equiv 0 \bmod 3$, then S contains an element of order 3 or it has the same order as a basis containing 0 and an element of order 3. Thus, the result follows from Corollary 3. Let $n \equiv 1 \bmod 3$. If $a = -\lfloor \frac{n}{3} \rfloor$, then $3 * S = \{0, 1, 3, 6\}$ and

$$\operatorname{order}(S) = \operatorname{order}(3 * S) \leq \operatorname{order}(\{0, 1, 6\});$$

if $a = \lfloor \frac{n}{3} \rfloor + 1$, then $3 * S - 2 = \{0, 1, 4, -2\}$ and

$$\operatorname{order}(S) = \operatorname{order}(3 * S - 2) \leq \operatorname{order}(\{0, 1, 4\}).$$

In both cases, $\operatorname{order}(S) \leq \lfloor \frac{n}{4} \rfloor + 2$ by Corollary 4. If $n \equiv 2 \bmod 3$, then

$$\operatorname{order}\left(\{0, 1, 2, -\lfloor \tfrac{n}{3} \rfloor\}\right) = \operatorname{order}(\{0, 1, 4, -2\}),$$
$$\operatorname{order}\left(\{0, 1, 2, \lfloor \tfrac{n}{3} \rfloor + 1\}\right) = \operatorname{order}(\{0, 1, 3, 6\}),$$

and the result follows as before. □

Bases of the form $\{0, 1, 3, a\}$

Lemma 23. *Let $n \geq 30$. Let $a \in \left\{\left\lfloor\frac{n}{3}\right\rfloor+1, \left\lfloor\frac{n}{2}\right\rfloor, \left\lfloor\frac{n}{2}\right\rfloor+1, 1-\left\lfloor\frac{n}{2}\right\rfloor, -\left\lfloor\frac{n}{3}\right\rfloor, -2, -1\right\}$ and $S = \{0, 1, 3, a\}$. Then $\operatorname{order}(S) \leq \left\lfloor\frac{n}{4}\right\rfloor+2$ or $\operatorname{order}(S) \geq \left\lfloor\frac{n}{3}\right\rfloor-1$. Moreover, if $n \equiv 1 \bmod 3$, then $\operatorname{order}(S) \neq \left\lfloor\frac{n}{3}\right\rfloor-1$ and if $n \equiv 2 \bmod 3$, then $\operatorname{order}(S) \notin \left\{\left\lfloor\frac{n}{3}\right\rfloor-1, \left\lfloor\frac{n}{3}\right\rfloor\right\}$.*

Bases of the form $\left\{0, 1, \left\lfloor\frac{n}{3}\right\rfloor+1, a\right\}$

Lemma 24. *Let $n \geq 30$. Let $a \in \left\{\left\lfloor\frac{n}{2}\right\rfloor, \left\lfloor\frac{n}{2}\right\rfloor+1, 1-\left\lfloor\frac{n}{2}\right\rfloor, -\left\lfloor\frac{n}{3}\right\rfloor, -2, -1\right\}$ and $S = \left\{0, 1, \left\lfloor\frac{n}{3}\right\rfloor+1, a\right\}$. Then $\operatorname{order}(S) \leq \left\lfloor\frac{n}{4}\right\rfloor+2$ or $\operatorname{order}(S) \geq \left\lfloor\frac{n}{3}\right\rfloor-1$. Moreover, if $n \equiv 1 \bmod 3$, then $\operatorname{order}(S) \neq \left\lfloor\frac{n}{3}\right\rfloor-1$ and if $n \equiv 2 \bmod 3$, then $\operatorname{order}(S) \notin \left\{\left\lfloor\frac{n}{3}\right\rfloor-1, \left\lfloor\frac{n}{3}\right\rfloor\right\}$.*

Bases of the form $\left\{0, 1, \left\lfloor\frac{n}{2}\right\rfloor, a\right\}$

Lemma 25. *Let $n \geq 22$. Let $a \in \left\{\left\lfloor\frac{n}{2}\right\rfloor+1, 1-\left\lfloor\frac{n}{2}\right\rfloor, -\left\lfloor\frac{n}{3}\right\rfloor, -2, -1\right\}$ and $S = \left\{0, 1, \left\lfloor\frac{n}{2}\right\rfloor, a\right\}$. Then $\operatorname{order}(S) \leq \left\lfloor\frac{n}{4}\right\rfloor+2$ or $\operatorname{order}(S) \geq \left\lfloor\frac{n}{3}\right\rfloor-1$. Moreover, if $n \equiv 1 \bmod 3$, then $\operatorname{order}(S) \neq \left\lfloor\frac{n}{3}\right\rfloor-1$ and if $n \equiv 2 \bmod 3$, then $\operatorname{order}(S) \notin \left\{\left\lfloor\frac{n}{3}\right\rfloor-1, \left\lfloor\frac{n}{3}\right\rfloor\right\}$.*

Proof. If n is even, then S contains an element of order 2 and the result follows from Corollary 2. Now suppose that n is odd. Note that $2*S+1 = \{0, 1, 3, 2a+1\}$.

For $a = -1$, $\operatorname{order}(S) = \operatorname{order}(\{0, 1, 3, -1\}) = \operatorname{order}(\{0, 1, 2, 4\}) \leq \left\lfloor\frac{n}{4}\right\rfloor+2$, by Corollary 4.

For $a = -\left\lfloor\frac{n}{3}\right\rfloor$ and $n \equiv 0 \bmod 3$, S contains an element of order 3 and the result follows from Corollary 4.

For $a = \left\lfloor\frac{n}{2}\right\rfloor+1$, $\operatorname{order}(S) = \operatorname{order}(2*S+1) = \{0, 1, 2, 3\}$ and the result follows from Lemma 13.

Now suppose that a does not satisfy the previous cases. We have $\operatorname{order}(S) = \operatorname{order}(\{0, 1, 3, b\})$, with $b \in \left\{4, \left\lfloor\frac{n}{3}\right\rfloor+t+1, -3\right\}$, where $0 < t = n-3\left\lfloor\frac{n}{3}\right\rfloor \leq 2$. Thus, $\operatorname{order}(S) \leq \operatorname{order}(\{0, 1, b\}) \leq \left\lfloor\frac{n}{4}\right\rfloor+2$ by Corollary 4. □

Bases of the form $\left\{0, 1, \left\lfloor\frac{n}{2}\right\rfloor+1, a\right\}$

Lemma 26. *Let $n \geq 21$, $a \in \left\{1-\left\lfloor\frac{n}{2}\right\rfloor, -\left\lfloor\frac{n}{3}\right\rfloor, -2, -1\right\}$ and $S = \left\{0, 1, \left\lfloor\frac{n}{2}\right\rfloor+1, a\right\}$. Then $\operatorname{order}(S) \leq \left\lfloor\frac{n}{4}\right\rfloor+2$ or $\operatorname{order}(S) \geq \left\lfloor\frac{n}{3}\right\rfloor-1$. Moreover, if $n \equiv 1 \bmod 3$, then $\operatorname{order}(S) \neq \left\lfloor\frac{n}{3}\right\rfloor-1$ and if $n \equiv 2 \bmod 3$, then $\operatorname{order}(S) \notin \left\{\left\lfloor\frac{n}{3}\right\rfloor-1, \left\lfloor\frac{n}{3}\right\rfloor\right\}$.*

Proof. If n is even, then S has the same order as $S-1$, which contains 0 and an element of order 2. Thus, the result follows from Corollary 2. Now suppose that n is odd. Then $\operatorname{order}(S) = \operatorname{order}(\{0, 1, 2, 2a\})$.

If $a = 1-\left\lfloor\frac{n}{2}\right\rfloor = \left\lfloor\frac{n}{2}\right\rfloor+2$, then $2a = 3$ and the result follows from Lemma 13.

If $a = -2$, then $2a = -4$ and, by Corollary 4,

$$\text{order}(S) \le \text{order}(\{0, 1, -4\}) = \text{order}(\{0, 1, 5\}) \le \left\lfloor \tfrac{n}{4} \right\rfloor + 2.$$

If $a = -1$, then $2a = -2$ and, by Lemma 21, $\text{order}(S) = \text{order}(\{0, 2, 3, 4\}) \le \left\lfloor \frac{n}{4} \right\rfloor + 2$.

Suppose that $a = -\left\lfloor \frac{n}{3} \right\rfloor$. If $n \equiv 0 \bmod 3$, then S contains 0 and an element of order 3 and the result follows from Corollary 3. If $n \equiv 1 \bmod 3$, then $S = \left\{0, 1, 2, \left\lfloor \frac{n}{3} \right\rfloor + 1\right\}$ and the result follows from Lemma 22. If $n \equiv 2 \bmod 3$, then, by Corollary 4,

$$\text{order}(S) = \text{order}\Bigl(\Bigl\{0, 1, 2, \left\lfloor \tfrac{n}{3} \right\rfloor + 2\Bigr\}\Bigr) \le \text{order}\Bigl(\Bigl\{0, 1, \left\lfloor \tfrac{n}{3} \right\rfloor + 2\Bigr\}\Bigr) \le \left\lfloor \tfrac{n}{4} \right\rfloor + 2. \qquad \square$$

Remark. Suppose that $S = \{0, 1, a, b\}$, with $a \in \left\{2, 3, \left\lfloor \frac{n}{2} \right\rfloor, \left\lfloor \frac{n}{2} \right\rfloor + 1\right\}$ and $b \in \left\{2, 3, \left\lfloor \frac{n}{2} \right\rfloor, \left\lfloor \frac{n}{2} \right\rfloor + 1, 1 - \left\lfloor \frac{n}{2} \right\rfloor, -2, -1\right\}$, where n is odd if a or b belong to the set $\left\{\left\lfloor \frac{n}{2} \right\rfloor, \left\lfloor \frac{n}{2} \right\rfloor + 1, -\left\lfloor \frac{n}{2} \right\rfloor\right\}$. From the proofs of Lemmas 22–26, we get that $\text{order}(S) \le \left\lfloor \frac{n}{4} \right\rfloor + 2$ if S is not one of the next exceptional bases:

$$\{0, 1, 2, 3\}, \quad \{0, 1, 2, -1\}, \quad \left\{0, 1, \left\lfloor \tfrac{n}{2} \right\rfloor, \left\lfloor \tfrac{n}{2} \right\rfloor + 1\right\}, \quad \left\{0, 1, \left\lfloor \tfrac{n}{2} \right\rfloor + 1, 1 - \left\lfloor \tfrac{n}{2} \right\rfloor\right\}.$$

Note that all of them have the same order as $\{0, 1, 2, 3\}$.

Acknowledgements

We would like to thank the anonymous referee whose comments and suggestions helped us improve the presentation of our paper.

References

[Bueno and Furtado 2010] M. I. Bueno and S. Furtado, "On the gaps in the set of exponents of Boolean primitive circulant matrices", *Electron. J. Linear Algebra* **20** (2010), 640–660. MR 2011g:15054 Zbl 05850040

[Bueno et al. 2009] M. I. Bueno, S. Furtado, and N. Sherer, "Maximum exponent of Boolean circulant matrices with constant number of nonzero entries in their generating vector", *Electron. J. Combin.* **16**:1 (2009), Research Paper 66. MR 2010m:11121 Zbl 1165.05329

[Butler and Krabill 1974] K. K.-H. Butler and J. R. Krabill, "Circulant Boolean relation matrices", *Czechoslovak Math. J.* **24**:2 (1974), 247–251. MR 50 #515 Zbl 0329.20049

[Chou et al. 2008] W.-S. Chou, B.-S. Du, and P. J.-S. Shiue, "A note on circulant transition matrices in Markov chains", *Linear Algebra Appl.* **429**:7 (2008), 1699–1704. MR 2010a:15077 Zbl 1148.60045

[Davis 1979] P. J. Davis, *Circulant matrices*, Wiley, New York, 1979. MR 81a:15003 Zbl 0418.15017

[Dukes et al. 2010] P. Dukes, P. Hegarty, and S. Herke, "On the possible orders of a basis for a finite cyclic group", *Electron. J. Combin.* **17**:1 (2010), Research Paper 79. MR 2011e:11014 Zbl 1201.11017

[Huang 1990] D. D. Huang, "On circulant Boolean matrices", *Linear Algebra Appl.* **136** (1990), 107–117. MR 91m:15024 Zbl 0701.15010

[Klopsch and Lev 2009] B. Klopsch and V. F. Lev, "Generating abelian groups by addition only", *Forum Math.* **21**:1 (2009), 23–41. MR 2010c:20068 Zbl 1172.20038

[Lancaster 1969] P. Lancaster, *Theory of matrices*, Academic Press, New York, 1969. MR 39 #6885 Zbl 0186.05301

[Schwarz 1974] Š. Schwarz, "Circulant Boolean relation matrices", *Czechoslovak Math. J.* **24**:2 (1974), 252–253. MR 50 #516 Zbl 0315.15011

[Wang and Meng 1997] J.-Z. Wang and J.-X. Meng, "The exponent of the primitive Cayley digraphs on finite abelian groups", *Discrete Appl. Math.* **80**:2-3 (1997), 177–191. MR 99h:05058 Zbl 0897.05045

Received: 2011-06-10 Revised: 2011-09-21 Accepted: 2011-09-22

mbueno@math.ucsb.edu
Mathematics Department and College of Creative Studies,
University of California, Santa Barbara,
Santa Barbara, CA 93106, United States

kuanyingfang2011@u.northwestern.edu
Department of Mathematics, Northwestern University,
Evanston, IL 60208, United States

saf5132@psu.edu
Department of Mathematics, Pennsylvania State University,
University Park, PA 16802, United States

sbf@fep.up.pt
Faculdade de Economia do Porto, Universidade do Porto,
Rua Doutor Roberto Frias, 4200-464 Porto, Portugal

INVOLVE 5:2 (2012)
dx.doi.org/10.2140/involve.2012.5.207

Commutation classes of double wiring diagrams

Patrick Dukes and Joe Rusinko

(Communicated by Ravi Vakil)

We describe a new method for computing the graph of commutation classes of double wiring diagrams. Using these methods we compute the graph for five strings or less which allows us to confirm a positivity conjecture of Fomin and Zelevinsky when $n \leq 4$.

1. Introduction

In the theory of cluster algebras, the term Laurent phenomenon describes the mysterious instances in which recursively defined rational functions simplify to Laurent polynomials [Fomin and Zelevinsky 2002]. In many instances of the Laurent phenomenon, it is conjectured that the coefficients of the resulting Laurent polynomials are all positive.

One of the first examples of the Laurent phenomenon was found by Fomin and Zelevinsky [2000] when studying the relationships among minors of an $n \times n$ matrix with real coefficients. In this work they showed that every minor of a matrix was positive if and only if a particular subset of minors known as *chamber minors* was positive. These chamber minors were indexed by regions of a combinatorial object known as a double wiring diagram. Further, the relationships among minors were described in terms of a graph that describes the relationships among classes of double wiring diagrams.

As an example of the Laurent phenomenon, Fomin and Zelevinsky proved that every minor of an $n \times n$ matrix can be written as a Laurent polynomial in the chamber minors. They stated the following conjecture which remains open:

Conjecture 1.1 [Fomin and Zelevinsky 2000]. *For any n-string double wiring diagram w, every minor of an $n \times n$ matrix can be written as a Laurent polynomial with nonnegative coefficients in terms of the chamber minors of w.*

In this paper we confirm the conjecture for $n \leq 4$. To do so, we develop a new method of computing the graph of relationships among classes of double wiring

MSC2010: primary 05E15; secondary 15B48.

Keywords: wiring diagram, matrix positivity, Laurent phenomenon.

diagrams. We then use an original program for computing this graph based on this method, along with the algebra software Fermat [Lewis 2007] to compute the aforementioned Laurent polynomials.

In Section 2 of this paper we define *commutation classes of double wiring diagrams* and a graph which displays the relationships among these classes. In Section 3 we describe a new quiver representation of the commutation classes. This representation greatly simplifies the computation of the associated graph which we describe in Section 4. Finally, we return in Section 5 to the Laurent phenomenon and use our new computations to confirm the positivity conjecture for $n \leq 4$.

2. Double wiring diagrams

Fomin and Zelevinsky [2000] define an *n-stringed double wiring diagram* as two sets of n piecewise linear lines (black and gray) such that each line intersects every other line of the same color exactly once. We number gray lines from 1 to n with 1 on the top left and n on the bottom left. The black lines are labeled in the reverse order. In addition, for each chamber of the double wiring diagram we define the *chamber label* to be a pair of subsets (g, b) where g (respectively b) is the subset of $\{1, 2, \ldots, n\}$ identifying the gray (respectively black) strings that pass below the chamber. See Figure 1 for an example of a double wiring diagram with the chamber labels.

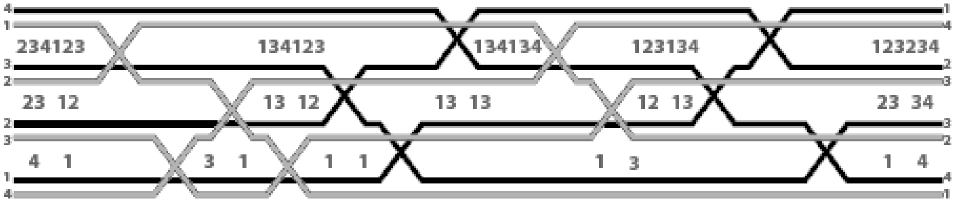


Figure 1. A four string double wiring diagram with chamber labels.

It is possible that slightly different wiring diagrams yield the same collection of chamber labels. Following Fomin and Zelevinsky, we consider two wiring diagrams that share the same collection of chamber labels *isotopic*. For single wiring diagrams, such collections of diagrams are called commutation classes, which have been studied in [Bédard 1999; Carter and Marsh 2000].

Definition 2.1. A *commutation class of double wiring diagrams* is the collection of all double wiring diagrams that share the same collection of chamber labels.

Any two commutation classes of double wiring diagrams can be linked by a sequence of the braid moves pictured in Figure 2 [Fomin and Zelevinsky 2000]. Note that in each exchange only one chamber label changes. In a braid move from wiring diagram w to wiring diagram w', we call the chamber label of w that changes under the braid move, the *center* of the braid move. A braid move is *centered* at a chamber label if that label changes under the braid move.

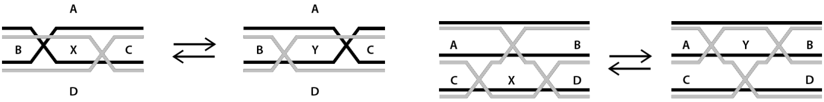


Figure 2. Braid moves. Left: 2-move; right: 3-move.

Since any two commutation classes of wiring diagrams can be connected by a sequence of braid moves, it is natural to construct a graph describing these relationships.

Definition 2.2. The graph of commutation classes of wiring diagrams, Φ_n, has a unique vertex for every commutation class of double wiring diagram with n strings. Two vertices are connected by an edge if their wiring diagrams differ by a single braid move.

Fomin and Zelevinsky [2000] prove that Φ_n is a finite connected graph and compute Φ_3. In this paper we present a method for computing Φ_n and use it to construct Φ_4 and Φ_5. We use these calculations to verify a positivity conjecture of Fomin and Zelevinsky when $n \leq 4$.

3. Using quivers to compute Φ_n

We have found that it is easier to compute Φ_n from the relationships among chamber labels than through the graphical structure of the wiring diagrams. This avoids the difficulty of keeping track of which wiring diagrams are in the same commutation class.

We introduce a quiver that describes the relevant relationships among the chamber labels of the double wiring diagram. This quiver is similar to a dual graph. The dual graph itself, however, is not an adequate data structure, as double wiring diagrams that are in the same commutation class may have differing dual graph structures.

Definition 3.1. For any double wiring diagram ω, define the quiver $Q(w)$ with vertices corresponding to chamber labels and an arrow from (g, b) to (g', b') if $g' = g \cup \{g_j\}$ and $b' = b \cup \{b_k\}$ for $g_j, b_k \in \{1, 2, \dots, n\}$. We label the arrows of the quiver with the pair of numbers (g_j, b_k). We refer to g_j (respectively b_k) as the gray (respectively black) labels of the arrow.

Figure 3 shows $Q(w)$ for the wiring diagram pictured in Figure 1. To keep track of the geometric relationship between the wiring diagram and the quiver we define the height and position of a vertex of a quiver, which roughly describe the location in the double wiring diagram of the corresponding chamber label.

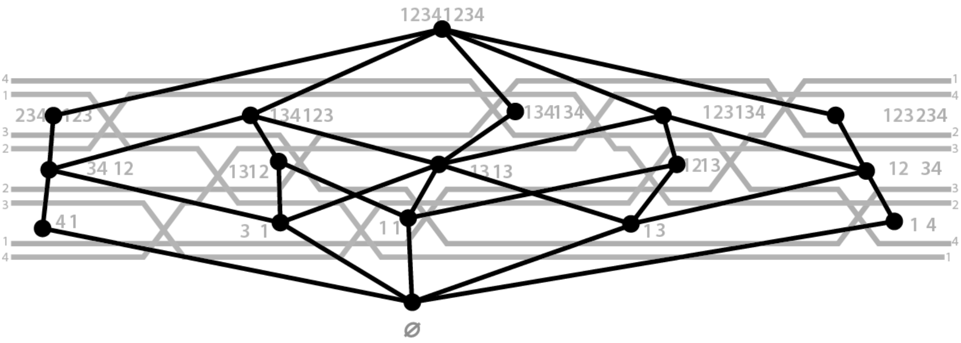


Figure 3. Quiver diagram.

Definition 3.2. Let v be a vertex of $Q(w)$ corresponding to the chamber label (g, b). We define the *height* of v to be the cardinality of g; $h(v) = |g|$. We define the position of v denoted $p(v) = \sum_{x \in b} x - \sum_{y \in g} y$.

Notice that the height increases the higher up one moves in the diagram while the position increases from left to right.

In order to construct Φ_n one needs to be able to identify the edges that are incident to a given vertex. This information is local in nature so we introduce language that allows us to discuss pieces of $Q(w)$.

Definition 3.3. A *subquiver* of $Q(w)$ is any subset of the vertices of $Q(w)$ together with a (possibly empty) set of arrows whose corresponding vertices are in the subset.

The following definitions provide the language needed to discuss the subquivers that are fundamental to the identifying braid moves.

Definition 3.4. A subquiver S of $Q(w)$ is *complete* if it contains every arrow of Q that connects two vertices of S.

Definition 3.5. A subquiver S of $Q(w)$ is *full* if, given that (g_1, b_1) and (g_2, b_2) are elements of S with height h, S contains the vertices corresponding to all chamber labels with height h and position between the positions of (g_1, b_1) and (g_2, b_2).

Notice that complete full subquivers completely determine a portion of a wiring diagram without missing arrows or vertices.

Using the language of complete full subquivers we can describe all of the edges that are incident to a vertex of Φ_n. Recall, each edge of Φ_n corresponds to a particular braid move centered at a particular chamber of the double wiring diagram.

Theorem 3.6. *There exists a* 3*-move centered at label* (g, b) *if and only if* $Q(w)$ *contains a complete, full subquiver of one of the two types shown in* Figure 4.

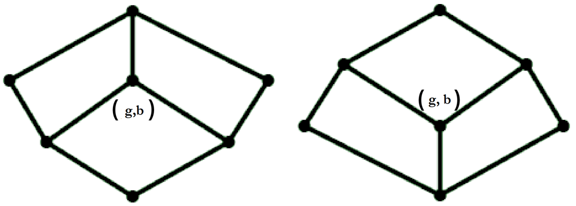


Figure 4. Subquivers for 3-move.

Proof. Assume a 3-move exists. Then there must be a region of the wiring diagram isomorphic to Figure 2 (right). Constructing the subquiver from this diagram yields Figure 4.

For the other direction, assume $Q(w)$ has a compete full subquiver isomorphic to Figure 4 (left). We examine the possible gray labels for this subquiver. Since the bottom vertex is connected to the top by a path of length three, we know that only three distinct edge labels may appear in this subquiver. We label the leftmost path from the bottom to the top that passes through (g, b), x, y, z as pictured in Figure 5.

For each four-cycle in Figure 5 only two distinct edge labels may be used since the bottom and top vertices are connected by a path of length two. This limits the potential labelings to those in Figure 6. The case of picture d) in the figure cannot exist because strings z and y are exchanged twice, which contradicts the definition of a double wiring diagram.

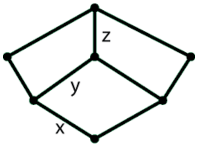


Figure 5. Grey labels for subquiver.

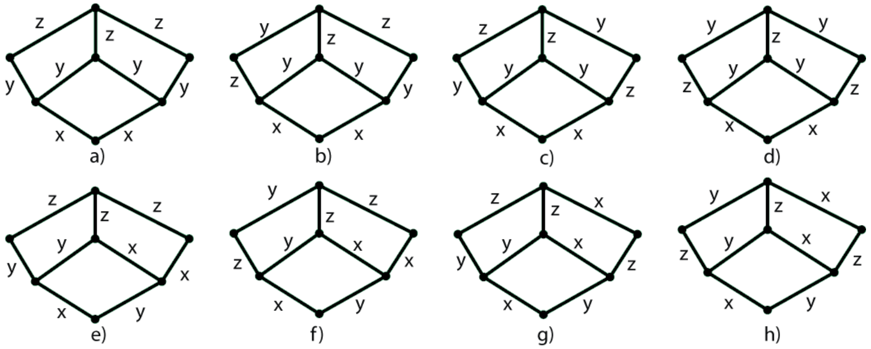


Figure 6. Possible labeled subquivers.

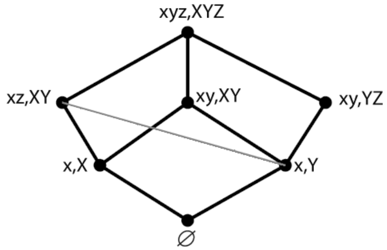


Figure 7. Invalid quiver labeling.

Repeat this argument for the black strings and label those cases A through H. We now determine which gray and black cases can be paired together. Since the labels must be distinct, the only potential pairs are (a, H), (b, G) and (c, F), and their opposites (h, A), (g, B) and (f, C).

If we draw a subquiver with the labels in the case (b, G), as in Figure 7, we recover an extra arrow, which contradicts the hypothesis that the subquiver was complete. The pairs (c, F), (g, B) and (f, C) are symmetric to (b, G), so they are also eliminated. This leaves only (a, H) and (h, A) as possible labelings.

By symmetry of the labelings we may assume the edge labels are of type (h, A). Since this subquiver is full, there are no missing vertices. This means that changes in chamber labels of the same cardinality indicate a unique braid crossing as pictured in Figure 8 (left). No other crossings may occur in this region because the quiver is complete. Therefore, the strings must connect without creating any other crossings. This yields the 3-move pictured in Figure 8 (right). The proof for Figure 4 (right) follows the same argument with reflected labels. □

Theorem 3.7. *There exists a* 2*-move centered at label* (g, b) *if and only if* $Q(w)$ *contains the full subquiver shown in* Figure 9.

Proof. Assume a 2-move exists. Then there must be a region of the wiring diagram isomorphic to Figure 2 (left). Constructing the quiver from this diagram yields the subquiver in Figure 9.

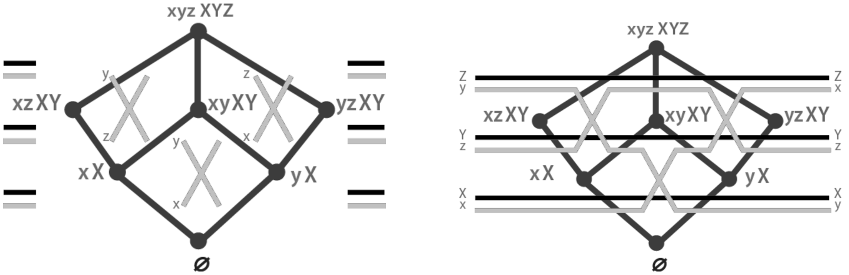


Figure 8. Reconstructed 3-move.

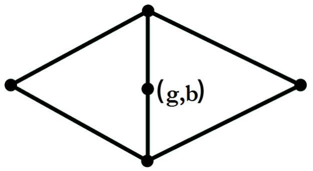


Figure 9. Subquiver for 2-move.

Now assume $Q(w)$ contains the full subquiver in Figure 9. We examine the possible gray labels for the subquiver. Label the arrows to and from (g, b) as x and y. Since there is a path from the bottom vertex to the top vertex of length two, all arrows in the subquiver must be labeled x or y. Figure 10 shows the possible labelings. The case corresponding to picture d) can be eliminated because it would require strings x and y to be exchanged twice.

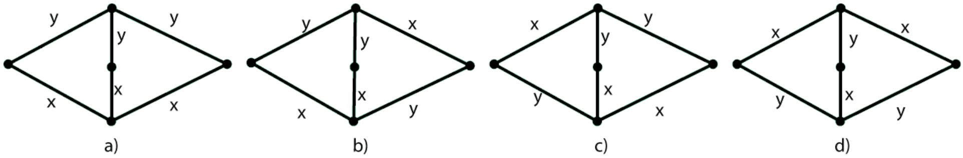


Figure 10. Quiver labelings.

We construct a similar pattern of possibilities for the black strings by labeling the arrows with X and Y. We need to determine which gray and black cases can be paired together. Since all of the labelings are distinct, the only potential pairs of cases are (b, C) and (c, B); see Figure 11.

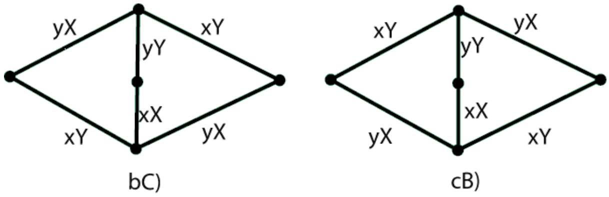


Figure 11. Potential quiver labelings.

As the labelings are symmetric, we can assume without loss of generality that the diagram has edge labels of type (b, C). Since this subquiver is full there are no missing vertices. This means that changes in chamber labels of the same height indicate a unique braid crossing as pictured in Figure 12 (left). Since no other crossings may occur in this region we connect the strings without creating any other crossings. Doing so yields the 2-move pictured in Figure 12 (right). □

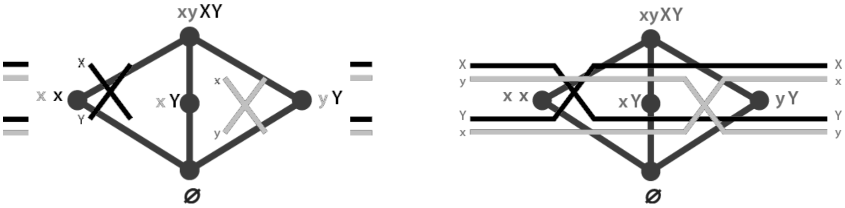


Figure 12. Reconstructed 2-move.

4. Describing Φ_n

Theorems 3.6 and 3.7 allow for the computation of the graph Φ_n for any n using the following algorithm:

(1) Choose an n-stringed double wiring diagram w with quiver $Q(w)$.

(2) Using Theorems 3.6 and 3.7 find and connect all vertices incident to the vertex corresponding to w.

(3) Repeat the previous step with the new set of vertices.

(4) Repeat this process until no new vertices can be added.

Since Φ_n is finite and connected this process will terminate and compute the entire graph. This process has been implemented in C++ using algorithms available at [Dukes 2011].

The smallest graph Φ_2 consists of two vertices connected by an edge. The graph of Φ_3 first appeared in [Fomin and Zelevinsky 2000]. Figure 13 shows a new

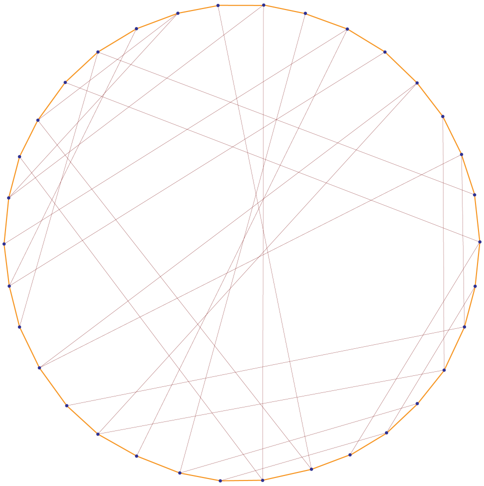


Figure 13. Φ_3 with Hamiltonian cycle highlighted.

	Φ_2	Φ_3	Φ_4	Φ_5
Vertices	2	34	4894	5520372
Edges	1	120	33300	60930112

Table 1. Φ_n edge and vertex data.

	1	2	3	4	5	6	7	8	9
Φ_2	2								
Φ_3			16	18					
Φ_4				2	522	1362	1754	1054	200

Table 2. Number of vertices of given degree for Φ_2 through Φ_4.

	6	7	8	9	10	11
Φ_5	84	28584	198596	632028	1165732	1402756
	12	13	14	15	16	
Φ_5	1165888	651188	227520	44452	3544	

Table 3. Number of vertices of given degree for Φ_5.

representation of Φ_3 which indicates the presence of a Hamiltonian path. Tables 1, 2 and 3 summarize information about Φ_n for $n \leq 5$.

5. Total positivity conjecture

With these explicit computations of Φ_n in hand, we return to the positivity conjecture of Fomin and Zelevinsky described in the introduction. We briefly review the setup of their work here. See [Fomin and Zelevinsky 2000] for a more complete description.

Definition 5.1. An $n \times n$ matrix M with entries in $\mathbb{R}$ is called *totally positive* if all minors of M are positive.

Definition 5.2 (Fomin and Zelevinsky). Let w be an n-stringed double wiring diagram. For each chamber label (b, g) of w we define the minor $\Delta_{g,b}$ to be the determinant of the matrix with rows of M corresponding to g and columns of M corresponding to b. We call the collection of all such minors the *chamber minors of* w.

Fomin and Zelevinsky [2000] proved that for any commutation class of double wiring diagrams w, a matrix M is totally positive if and only if all of its chamber minors are positive. In addition they conjectured that every minor could be written as a Laurent polynomial in the chamber minors with nonnegative coefficients.

Using our computations from Section 3, we can confirm this conjecture for $n \le 4$.

Theorem 5.3. *For $n \le 4$ and any n-stringed double wiring diagram w, every minor of an $n \times n$ matrix can be written as a Laurent polynomial with nonnegative coefficients in terms of the chamber minors of w.*

Proof. Fomin and Zelevinsky [2000] show that if w and w' are linked by a braid move as pictured in Figure 2, then their chamber minors satisfy the equation

$$AD + BC = XY. \tag{1}$$

Using the program Fermat [Lewis 2007] and a C++ program written by the first author, we verify Theorem 5.3 using the following algorithm:

(1) For each vertex $v \in \Phi_n$ and minor Δ, find a path from v to a vertex v' such that Δ is a chamber minor of v'. This is possible since Φ_n is connected and every minor appears as the chamber minor for some double wiring diagram.

(2) At each edge of this path use Fermat to compute the new minor as a Laurent polynomial in terms of the previous minors using Equation (1). The Laurent theorem [Fomin and Zelevinsky 1999] guarantees the result will be a Laurent polynomial in the chamber minors of v. Repeat the process until Δ is written as a Laurent polynomial in terms of the chamber minors of v.

(3) Verify that the corresponding Laurent polynomial has all positive coefficients.

The relevant code and data files can be found in [Dukes 2011]. □

Example 5.4. In this example we demonstrate that $\Delta_{14,12}$ can be written as a Laurent polynomial in the chamber minors of w as in the wiring diagram in Figure 1 with nonnegative coefficients. First, the diagrams in Figure 14 determine a path in Φ_4 from w to a vertex corresponding to a diagram with $\Delta_{14,12}$ as a chamber minor.

Each exchange along this path introduces a new chamber minor. Using (1), we compute the new chamber minor as a Laurent polynomial in terms of the chamber minors of w. The results of the Fermat computations of these Laurent polynomials are listed below.

$$\begin{aligned}
\Delta_{34,13} &= \frac{\Delta_{34,12}\Delta_{13,13} + \Delta_{134,123}\Delta_{3,1}}{\Delta_{13,12}} \\
&= \Delta_{134,123}\Delta_{13,12}^{-1}\Delta_{3,1} + \Delta_{34,12}\Delta_{13,12}^{-1}\Delta_{13,13},
\end{aligned}$$

$$\begin{aligned}
\Delta_{14,13} &= \frac{\Delta_{1,1}\Delta_{34,13} + \Delta_{4,1}\Delta_{13,13}}{\Delta_{3,1}} \\
&= \Delta_{134,123}\Delta_{13,12}^{-1} + \Delta_{34,12}\Delta_{13,12}^{-1}\Delta_{13,13}\Delta_{3,1}^{-1}\Delta_{1,1} + \Delta_{13,13}\Delta_{4,1}\Delta_{3,1}^{-1},
\end{aligned}$$

$$\Delta_{14,12} = \frac{\Delta_{14,13}\Delta_{34,12} + \Delta_{134,123}\Delta_{4,1}}{\Delta_{34,13}} = \Delta_{34,12}\Delta_{3,1}^{-1}\Delta_{1,1} + \Delta_{13,12}\Delta_{4,1}\Delta_{3,1}^{-1}.$$

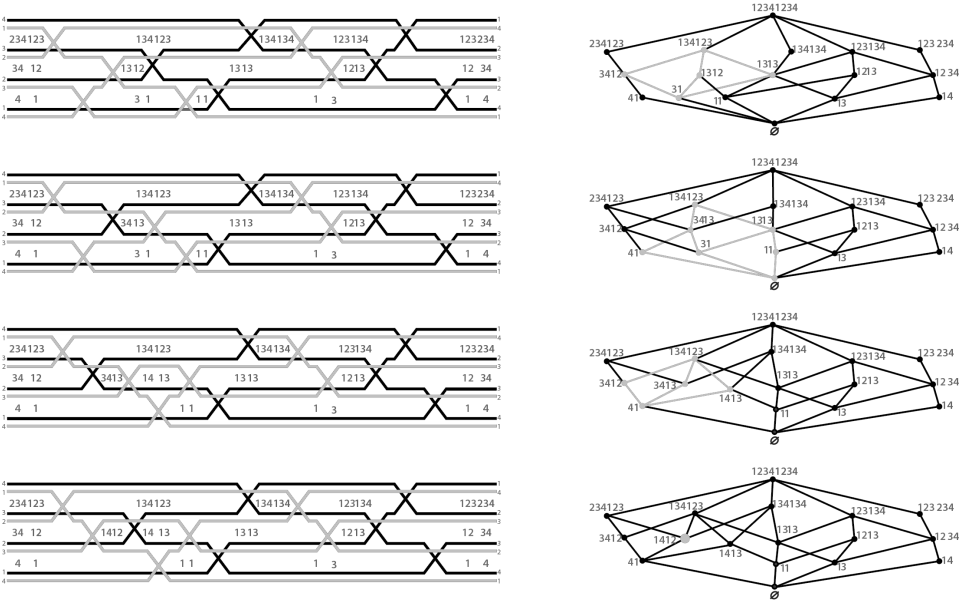


Figure 14. Path to a vertex with $\Delta_{14,12}$ as a chamber minor.

It suffices to observe that the coefficients in the expression of $\Delta_{14,12}$ in terms of the chamber minors of w, are all positive.

Remark 5.5. The example above is not indicative of the complexity of the computations. In Φ_4 the Laurent polynomials in the solution frequently had over 100 terms.

Although we were able to compute Φ_5 we were unable to confirm the conjecture for $n = 5$ because of number of computations required. There are $34 \times 14 = 476$ pairs of vertices and chamber minors in Φ_3, and $62 \times 4{,}894 = 303{,}420$ such combinations in Φ_4. To confirm the conjecture with brute force for $n = 5$ would require $242 \times 5{,}520{,}372 = 1{,}335{,}930{,}024$ computations each involving extremely large Laurent polynomials.

References

[Bédard 1999] R. Bédard, "On commutation classes of reduced words in Weyl groups", *European J. Combin.* **20**:6 (1999), 483–505. MR 2001i:05160 Zbl 0934.05126

[Carter and Marsh 2000] R. Carter and R. Marsh, "Regions of linearity, Lusztig cones, and canonical basis elements for the quantized enveloping algebra of type A_4", *J. Algebra* **234**:2 (2000), 545–603. MR 2001k:17019 Zbl 0999.17024

[Dukes 2011] P. Dukes, Double wiring diagram files, 2011, http://www.patrickdukes.com/double-wiring-diagram-files.html.

[Fomin and Zelevinsky 1999] S. Fomin and A. Zelevinsky, "Double Bruhat cells and total positivity", *J. Amer. Math. Soc.* **12**:2 (1999), 335–380. MR 2001f:20097 Zbl 0913.22011

[Fomin and Zelevinsky 2000] S. Fomin and A. Zelevinsky, "Total positivity: tests and parametrizations", *Math. Intelligencer* **22**:1 (2000), 23–33. MR 2001b:15030 Zbl 1052.15500

[Fomin and Zelevinsky 2002] S. Fomin and A. Zelevinsky, "The Laurent phenomenon", *Adv. in Appl. Math.* **28**:2 (2002), 119–144. MR 2002m:05013 Zbl 1012.05012

[Lewis 2007] R. Lewis, "Fermat: a computer algebra system for polynomial and matrix computation", 2007, http://home.bway.net/lewis.

Received: 2011-08-30 Revised: 2011-09-03 Accepted: 2011-09-14

pdukes3@gmail.com *School of Computing, Clemson University,*
Clemson, SC 29634, United States

rusinkoj@winthrop.edu *Department of Mathematics, Winthrop University,*
Rock Hill, SC 29733, United States

INVOLVE 5:2 (2012)
dx.doi.org/10.2140/involve.2012.5.219

A two-step conditionally bounded numerical integrator to approximate some traveling-wave solutions of a diffusion-reaction equation

Siegfried Macías and Jorge E. Macías-Díaz

(Communicated by Emil Minchev)

We develop a finite-difference scheme to approximate the bounded solutions of the classical Fisher–Kolmogorov–Petrovsky–Piskunov equation from population dynamics, in which the nonlinear reaction term assumes a generalized logistic form. Historically, the existence of wave-front solutions for this model is a well-known fact; more generally, the existence of solutions of this equation which are bounded between 0 and 1 at all time, is likewise known, whence the need to develop numerical methods that guarantee the positivity and the boundedness of such solutions follows necessarily. The method is implicit, relatively easy to implement, and is capable of preserving the positivity and the boundedness of the new approximations under a simple parameter constraint. The proof of the most important properties of the scheme is carried out with the help of the theory of M-matrices. Finally, the technique is tested against some traveling-wave solutions of the model under investigation; the results evince the fact that the method performs well in the cases considered.

1. Introduction

R. A. Fisher [1937] and A. Kolmogorov, I. Petrovsky and N. Piskunov [Kolmogorov et al. 1937] were the first to investigate the advance wave of mutant genes which are advantageous to some populations distributed on linear habitats. The model that they investigated is known as the Fisher–Kolmogorov–Petrovsky–Piskunov equation, the Fisher–KPP equation, or simply Fisher's equation, and it is one of the simplest diffusive equations with nonlinear reaction. This parabolic partial differential equation is a useful model in the description of the process of epidermal wound healing [Sherratt and Murray 1990], in the theory of the electrodynamics of

MSC2010: primary 35K15, 65M06; secondary 35K55, 35K57.
Keywords: Fisher's equation, bounded/positive solutions, finite-difference scheme, boundedness/positivity preservation, inverse-positive matrices.

semiconductors [Wallace 1984], in the investigation of excitons [Rashba and Sturge 1982], and as a model for neutron flux in nuclear reactor kinetics [Kastenberg and Chambré 1968].

The Fisher–KPP equation, like many other equations in mathematical physics, is well-known to possess traveling-wave solutions [Wang 1988]. The wave fronts connect the two stationary solutions, 0 and 1 in the equation's nondimensionalized version, via a monotone solution bounded within (0, 1) at all times. The existence of other bounded solutions for this model, apart from traveling waves, is also a standard result in the specialized literature [Wazwaz and Gorguis 2004]. This and the fact that the Fisher–KPP equation is a model for which there is no analytic solution for every admissible set of initial conditions justify interest in the design of numerical techniques preserving the boundedness of the solutions.

The design of numerical methods that preserve several physical or mathematical properties of the phenomena that they describe is a fruitful avenue of research in scientific computation. Thus, from the physical point of view, several methods have been proposed to approximate the solution and the energy dynamics of conservative [Furihata 2001] and dissipative [Furihata 1999] systems. From the mathematical point of view, the preservation of conditions such as symmetry, monotonicity, positivity and boundedness is sometimes a highly desirable characteristic in a numerical integrator. In fact, several numerical methods have been designed with these conditions in mind, particularly in those cases when the variable of interest is measured in an absolute scale. In these situations, the conditions of positivity and boundedness of solutions, which are typical in the study of some traveling waves, arise as constraints in the meaningfulness of the numerical results.

In this article we develop a finite-difference scheme to approximate bounded positive solutions to the Fisher–KPP equation, and test our method against known traveling-wave solutions. The main properties of our technique are consequences of the theory of M-matrices [Fujimoto and Ranade 2004], which are nonsingular, square matrices with the property that their inverses have only positive entries.

This work is organized as follows: In Section 2, we introduce the quantitative model under investigation (namely, the Fisher–KPP equation from population dynamics), and a family of traveling-wave solutions used in the sequel as comparison paradigms. Section 3 presents the numerical method employed to approximate solutions of the problem under investigation. There we prove our main result, which gives parameter conditions under which the method is able to preserve positivity and boundedness of the solutions of the Fisher–KPP model. Section 4 presents numerical evidence that the method is capable of preserving the properties mentioned above when the conditions of our main result are satisfied. We make some concluding remarks in Section 5.

2. The Fisher–KPP equation

Let p be a positive integer. Let $\mathbb{R}^+$ represent the set of nonnegative numbers, and let $I = [a, b]$ be a closed and bounded interval of $\mathbb{R}$. Let u be a real function defined on $I \times \mathbb{R}^+$ which, for practical purposes, is supposed to be twice differentiable in the interior of its domain. In this work, we approximate traveling-wave solutions of the classical Fisher–KPP equation, which, in nondimensional form, is the nonlinear, parabolic partial differential equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + u f(u), \tag{1}$$

where the function $f : \mathbb{R} \to \mathbb{R}$ has the generalized logistic form

$$f(u) = 1 - u^p. \tag{2}$$

As mentioned in the Introduction, this equation was first studied in the context of the dynamics of populations in a one-dimensional, unbounded habitat. (In the original studies, the exponent p was equal to 1.) For every real constant C, the functions

$$u(x, t) = \left\{ \frac{1}{2} \tanh\left[-\frac{p}{2\sqrt{2p+4}} \left(x - \frac{p+4}{\sqrt{2p+4}} t \right) + \frac{C}{2} \right] + \frac{1}{2} \right\}^{2/p} \tag{3}$$

are traveling wave solutions to (1), bounded in the interval (0, 1), and connecting the two constant solutions $u = 0$ and $u = 1$ (see [Wang 1988]). These solutions will be employed for comparison purposes in Section 4.

3. Numerical method

For the discretization, we consider a uniform partition $a = x_0 < x_1 < \cdots < x_N = b$ of the interval I and a uniform partition $0 = t_0 < t_1 < \cdots < t_M = T$ of the time interval $[0, T]$ over which we will compute approximate solutions of (1). We let u_n^k represent the approximation to the exact value of $u(x_n, t_k)$. For convenience, let $\Delta x = (b - a)/N$ and $\Delta t = T/M$, and consider the standard linear operators

$$\delta_t u_n^k = \frac{u_n^{k+1} - u_n^k}{\Delta t}, \tag{4}$$

defined for every $n \in \{0, 1, \ldots, N\}$ and every $k \in \{0, 1, \ldots, M - 1\}$, and

$$\delta_x^2 u_n^k = \frac{u_{n+1}^k - 2u_n^k + u_{n-1}^k}{(\Delta x)^2}, \tag{5}$$

defined for every $n \in \{1, 2, \ldots, N - 1\}$ and every $k \in \{0, 1, \ldots, M\}$. Let $n \in \{1, 2, \ldots, N - 1\}$ and $k \in \{0, 1, \ldots, M - 1\}$. With this notation at hand, we approximate the exact solution of u at (x_n, t_k) through the nonlinear difference equation

$$\delta_t u_n^k = \delta_x^2 u_n^{k+1} + u_n^{k+1} f(u_n^k). \tag{6}$$

Clearly, in order to approximate solutions of (1) using the numerical method (6), appropriate initial and boundary conditions must be imposed in both the continuous and the discrete scenarios. In the present work, we will consider an initial profile of the form $u(x, 0) = \phi(x)$ for every $x \in I$, a condition that translates to the discrete scene into the constraint $u_n^0 = \phi(x_n)$, for $n \in \{0, 1, \dots, N\}$. Similarly, we will consider boundary conditions of the form $u(a, t) = g(t)$ and $u(b, t) = h(t)$ for every $t \in [0, T]$, which translate, respectively, as $u_0^k = g(t_k)$ and $u_N^k = h(t_k)$, for every $k \in \{0, 1, \dots, M\}$. With these conventions, the finite-difference method (6) may be rewritten in vector form as the equation

$$A_k \boldsymbol{u}^{k+1} = \boldsymbol{v}^k \quad \text{for } k \in \{0, 1, \dots, M-1\}, \tag{7}$$

where $\boldsymbol{v}^k$ is the $(N+1)$-dimensional real vector

$$\boldsymbol{v}^k = \left(g(t_{k+1}), u_1^k, \dots, u_{N-1}^k, h(t_{k+1})\right)^t, \tag{8}$$

for $k \in \{0, 1, \dots, M\}$, and A is the matrix of size $(N+1) \times (N+1)$ given by

$$A_k = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -R & a_1^k & -R & 0 & \cdots & 0 & 0 & 0 \\ 0 & -R & a_2^k & -R & \cdots & 0 & 0 & 0 \\ 0 & 0 & -R & a_3^k & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & a_{N-2}^k & -R & 0 \\ 0 & 0 & 0 & 0 & \cdots & -R & a_{N-1}^k & -R \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{pmatrix}. \tag{9}$$

Here,

$$R = \frac{\Delta t}{(\Delta x)^2}, \tag{10}$$

$$a_n^k = 1 + 2R - f(u_n^k)\Delta t \quad \text{for } n \in \{1, 2, \dots, N-1\}. \tag{11}$$

The forward-difference stencil of our method is depicted in Figure 1. The method is clearly implicit and, after appropriate boundary conditions are specified at the endpoints of I, it only requires of an initial profile $\boldsymbol{u}^0$ in order to compute the subsequent approximations. Note also that, if f were a constant function, the matrix A_k would be a constant matrix A, and the approximation at time k would be given by $A^k \boldsymbol{u}^k = \boldsymbol{u}^0$.

We now establish conditions under which the finite-difference method (6) pre-

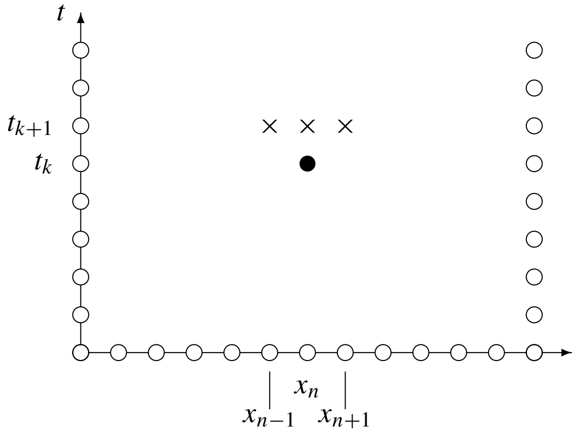


Figure 1. Forward-difference stencil of the finite-difference scheme (6). The black circle represents the known approximation to the exact solutions at the time t_k, and the crosses denote the unknown, new approximations at the time t_{k+1}.

serves the boundedness and the positivity of the solutions of (1), and it makes use of the nonsingularity properties of M-matrices [Fujimoto and Ranade 2004].

Proposition 1. *Let* $k \in \{0, 1, \dots, M-1\}$, *let* p *be equal to* 1, *and suppose that all the components of* $\boldsymbol{v}^k$ *are numbers in* $(0, 1)$. *If* $\Delta t < 1$ *then the components of* $\boldsymbol{u}^{k+1}$ *in* (7) *are all likewise bounded in* $(0, 1)$.

Proof. Clearly, A_k has nonpositive, off-diagonal entries. Moreover, if $f(u_n^k)\Delta t < 1$ for every $n \in \{1, 2, \dots, N-1\}$, then A_k is a strictly diagonally dominant matrix with positive diagonal entries (notice that such condition holds if $0 < u_n^k < 1$ for every $n \in \{1, 2, \dots, N-1\}$ and $\Delta t < 1$) and, as a consequence, A_k is an M-matrix, that is, a nonsingular matrix whose inverse only has positive entries. Together with (7), this implies that $\boldsymbol{u}^{k+1}$ is a vector with positive entries. Next, we establish the boundedness from above of the components of $\boldsymbol{u}^{k+1}$. Let $\boldsymbol{e}$ be the $(N+1)$-dimensional vector all of whose components are equal to 1, and let $\boldsymbol{w}^{k+1} = \boldsymbol{e} - \boldsymbol{u}^{k+1}$. A simple substitution in (7) gives us the equation

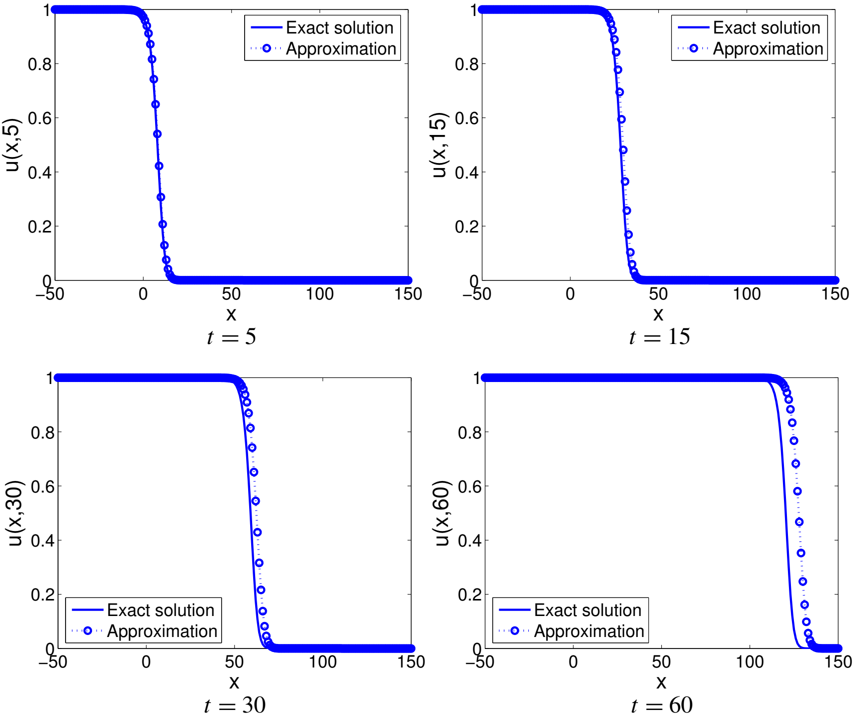
$$A_k \boldsymbol{w}^{k+1} = A_k \boldsymbol{e} - \boldsymbol{v}^k. \tag{12}$$

The first and last components of the right-hand side of (12) are, respectively, $1-g(t_k)$ and $1-h(t_k)$, which are positive, while for every $n \in \{1, 2, \dots, N-1\}$, the $(n+1)$-st component is given by the expression $(1-\Delta t)(1-u_n^k)$, which is also a positive number. As in the first part of this proof, it follows that the components of $\boldsymbol{w}^{k+1}$ are all positive numbers or, equivalently, that the components of $\boldsymbol{u}^{k+1}$ are all less than 1. □

We stress that (4) is a first-order accurate approximation of the partial derivative of u with respect to t at (x_n, t_k), and that (5) is an approximation of the second order to the value of the partial derivative of u with respect to x^2 at the same point. Under these circumstances, the linearized version of the finite-difference scheme (6) is consistent of order $\Delta t + (\Delta x)^2$ with the linearized version of (1) at (x_n, t_{k+1}).

4. Numerical results

To illustrate the validity of the our method and its computational implementation, we ran two numerical experiments, choosing the initial conditions so the exact solution is known, namely, the function (3). We set $C = 1$ and $p = 1$, and let the spatial domain be $I = [-50, 150]$, imposing at the endpoints Dirichlet conditions provided by the exact solution evaluated at -50 and 150.



$t = 5$ $t = 15$

$t = 30$ $t = 60$

Figure 2. Analytical solution (solid line) and the corresponding approximation (dotted line) versus the spatial variable x at four different times, of a system governed by (1) with $p = 1$. The initial profile is that given by (3) at $t = 0$ with $C = 1$, and the boundary conditions are provided by (3) at the endpoints of $[-50, 150]$ at any time. Numerically, the method (6) employed $\Delta x = 1$ and $\Delta t = 0.05$, and the times considered were $t = 5, 15, 30, 60$.

In the first run, we use the finite-difference method (6) with $\Delta x = 1$ and $\Delta t = 0.05$, so that the parameter constraint in Proposition 1 for the boundedness of the method be satisfied. Under these conditions, Figure 2 compares the exact solutions with the corresponding numerical approximations provided by our technique at four different times, namely $t = 5$, 15, 30 and 60. The results show that the computational solution remains bounded within (0, 1), as expected. Additionally, there exists a good agreement between both solutions at small times; the difference between the exact solutions and the numerical approximations is more pronounced at the times $t = 30$ and 60.

In the second run, we change only the parameter values $\Delta x = 0.5$ and $\Delta t = 0.005$. The numerical results are presented in Figure 3, and one immediately notices a better agreement between the analytical solutions and the computational approximations

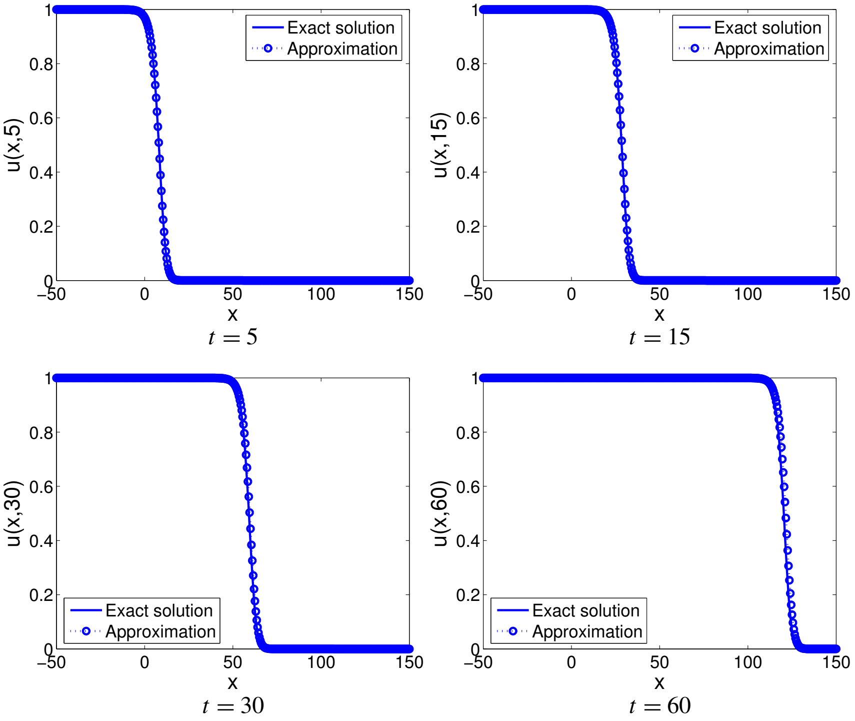


Figure 3. Analytical solution (solid line) and the corresponding approximation (dotted line) versus the spatial variable x at four different times, of a system governed by (1) with $p = 1$. The initial profile is that given by (3) at $t = 0$ with $C = 1$, and the boundary conditions are provided by (3) at the endpoints of $[-50, 150]$ at any time. Numerically, the method (6) employed $\Delta x = 0.5$ and $\Delta t = 0.005$, and the times considered were $t = 5$, 15, 30, 60.

to the problem under consideration, even for larger values of time. We also see that the numerical approximations, like the exact solutions, remain bounded within $(0, 1)$. This is in agreement with Proposition 1.

5. Conclusions

We have presented a numerical method to approximate bounded solutions of the classical Fisher–KPP equation from population dynamics. The proposed finite-difference scheme is a nonstandard method in the way that the reaction term is approximated, and it may be conveniently expressed in vector form in terms of the multiplication by a tridiagonal matrix which, under certain circumstances, is actually an M-matrix. In this way, new approximations may be written as the product of the previous approximation by the inverse of the M-matrix. Some simple and direct calculations show that the new approximations are bounded between 0 and 1 under suitable conditions on the computational parameters.

The method was implemented and tested against known exact solutions of the classical Fisher–KPP equation on a bounded spatial domain. The results show that the method performs well when approximating the analytical solutions considered. Moreover, one notices that the method preserves the boundedness and the positivity of the solutions considered when the parameter conditions derived in the work are satisfied.

References

[Fisher 1937] R. A. Fisher, "The wave of advance of advantageous genes", *Ann. Eugenics* **7** (1937), 355–369.

[Fujimoto and Ranade 2004] T. Fujimoto and R. R. Ranade, "Two characterizations of inverse-positive matrices: the Hawkins–Simon condition and the Le Chatelier–Braun principle", *Electron. J. Linear Algebra* **11** (2004), 59–65. MR 2005m:15053 Zbl 1069.15020

[Furihata 1999] D. Furihata, "Finite difference schemes for $\partial u/\partial t = (\partial/\partial x)^{\alpha}\delta G/\delta u$ that inherit energy conservation or dissipation property", *J. Comput. Phys.* **156**:1 (1999), 181–205. MR 2000j:65076 Zbl 0945.65103

[Furihata 2001] D. Furihata, "Finite-difference schemes for nonlinear wave equation that inherit energy conservation property", *J. Comput. Appl. Math.* **134**:1-2 (2001), 37–57. MR 2002g:65096 Zbl 0989.65099

[Kastenberg and Chambré 1968] W. E. Kastenberg and P. L. Chambré, "On the stability of nonlinear space-dependent reactor kinetics", *Nucl. Sci. Eng.* **31** (1968), 67–79.

[Kolmogorov et al. 1937] A. Kolmogoroff, I. Petrovsky, and N. Piscounoff, "Étude de l'équation de la diffusion avec croissance de la quantité de matière et son application a un problème biologique", *Bull. Univ. Moscou, Ser. Internat. A* **1**:6 (1937), 1–25. Zbl 0018.32106

[Rashba and Sturge 1982] E. I. Rashba and M. D. Sturge (editors), *Excitons*, North-Holland, Amsterdam, 1982.

[Sherratt and Murray 1990] J. A. Sherratt and J. D. Murray, "Models of epidermal wound healing", *Proc. R. Soc. Lond. B* **241** (1990), 29–36.

[Wallace 1984] P. R. Wallace, *Mathematical analysis of physical problems*, Dover Publications, New York, 1984. MR 86c:00006 Zbl 1092.00501

[Wang 1988] X. Y. Wang, "Exact and explicit solitary wave solutions for the generalised Fisher equation", *Phys. Lett. A* **131**:4-5 (1988), 277–279. MR 89h:35320

[Wazwaz and Gorguis 2004] A.-M. Wazwaz and A. Gorguis, "An analytic study of Fisher's equation by using Adomian decomposition method", *Appl. Math. Comput.* **154**:3 (2004), 609–620. MR 2005c:35152 Zbl 1054.65107

Received: 2011-11-28 Accepted: 2012-05-12

sieg_macias@hotmail.com

Universidad Autónoma de Aguascalientes,
Avenida Universidad 940, Ciudad Universitaria,
Aguascalientes, Aguascalientes 20131, Mexico

jemacias@correo.uaa.mx

Departamento de Matemáticas y Física,
Universidad Autónoma de Aguascalientes,
Avenida Universidad 940, Ciudad Universitaria,
Aguascalientes, Aguascalientes 20131, Mexico

INVOLVE 5:2 (2012)
dx.doi.org/10.2140/involve.2012.5.229

The average order of elements in the multiplicative group of a finite field

Yilan Hu and Carl Pomerance

(Communicated by Kenneth S. Berenhaut)

We consider the average multiplicative order of a nonzero element in a finite field and compute the mean of this statistic for all finite fields of a given degree over their prime fields.

1. Introduction

For a cyclic group of order n, let $\alpha(n)$ denote the average order of an element. For each $d \mid n$, there are exactly $\varphi(d)$ elements of order d in the group (where φ is Euler's function), so

$$\alpha(n) = \frac{1}{n}\sum_{d\mid n} d\varphi(d).$$

It is known [von zur Gathen 2004] that

$$\frac{1}{x}\sum_{n\le x}\alpha(n) = \frac{3\zeta(3)}{\pi^2}x + O\left((\log x)^{2/3}(\log\log x)^{4/3}\right).$$

We are interested here in obtaining an analogous result where n runs over the orders of the multiplicative groups of finite fields. Let p denote a prime number. We know that up to isomorphism, for each positive integer k, there is a unique finite field of p^k elements. The multiplicative group for this field is cyclic of size $p^k - 1$. We are concerned with the average order of an element in this cyclic group as p varies. We show the following results.

Theorem 1. *For each positive integer k there is a positive constant K_k such that the following holds. For each number $A > 0$, each number $x \ge 2$, and each positive*

MSC2010: 11B05, 11B75.

Keywords: average multiplicative order, finite field.

This paper is based on Hu's 2010 Dartmouth honors thesis, written under the direction of Pomerance. Both authors gratefully acknowledge Florian Luca, who suggested the problem. Pomerance was supported in part by NSF grant DMS-1001180.

integer k with $k \le (\log x)/(2\log\log x)$, we have

$$\frac{1}{\pi(x)}\sum_{p\le x}\frac{\alpha(p^k-1)}{p^k-1} = K_k + O_A\left(\frac{1}{\log^A x}\right).$$

This theorem in the case $k=1$ appears in [Luca 2005]. Using Theorem 1 and a partial summation argument we are able to show the following consequence.

Corollary 2. *For all numbers $A>0$, $x\ge 2$, and for any positive integer $k\le (\log x)/(2\log\log x)$, we have*

$$\frac{1}{\pi(x)}\sum_{p\le x}\alpha(p^k-1) = K_k\frac{\operatorname{li}(x^{k+1})}{\operatorname{li}(x)} + O_A\left(\frac{x^k}{\log^A x}\right),$$

where K_k is the constant from Theorem 1 and $\operatorname{li}(x) := \int_2^x \mathrm{d}t/\log t$.

Since $\operatorname{li}(x^{k+1})/\operatorname{li}(x) \sim x^k/(k+1)$ as $x\to\infty$, Corollary 2 implies that

$$\frac{1}{\pi(x)}\sum_{p\le x}\alpha(p^k-1) \sim \frac{K_k}{k+1}x^k, \quad \text{as } x\to\infty.$$

We identify the constants K_k as follows. Let $N_k(n)$ denote the number of solutions to the congruence $s^k\equiv 1 \pmod n$.

Proposition 3. *For each prime p and positive integer k let*

$$S_k(p) = \sum_{j=1}^{\infty}\frac{N_k(p^j)}{p^{3j-1}}.$$

Then $S_k(p)<1$ and

$$K_k := \prod_p (1-S_k(p))$$

is a real number with $0<K_k<1$.

2. Preliminary results

In this section we prove Proposition 3 and we also prove a lemma concerning the function $N_k(n)$.

Proof of Proposition 3. We clearly have $N_k(n)\le\varphi(n)$ for every n, since $N_k(n)$ counts the number of elements in the group $(\mathbb{Z}/n\mathbb{Z})^*$ with order dividing k and there are $\varphi(n)$ elements in all in this group. Thus, we have

$$S_k(p) \le \sum_{j=1}^{\infty}\frac{\varphi(p^j)}{p^{3j-1}} = \left(1-\frac{1}{p}\right)\sum_{j=1}^{\infty}\frac{p}{p^{2j}} = \left(1-\frac{1}{p}\right)\frac{p}{p^2-1} = \frac{1}{p+1}.$$

This proves the first assertion, but it is not sufficient for the second assertion. For p an odd prime, the group $(\mathbb{Z}/p^j Z)^*$ is cyclic so that the number of elements in this group of order dividing k is

$$N_k(p^j) = \gcd(k, \varphi(p^j)). \tag{1}$$

The same holds for $p^j = 2$ or 4, or if $p = 2$ and k is odd. Suppose now that $p = 2$, $j \geq 3$, and k is even. Since $(\mathbb{Z}/2^j\mathbb{Z})^*$ is the direct product of a cyclic group of order 2 and a cyclic group of order 2^{j-2}, we have

$$N_k(2^j) = 2 \cdot \gcd(k, 2^{j-2}) = \gcd(2k, \varphi(2^j)). \tag{2}$$

Thus, we always have $N_k(p^j) \leq 2k$, and so

$$S_k(p) \leq \sum_{j=1}^{\infty} \frac{2k}{p^{3j-1}} = \frac{2kp}{p^3 - 1}.$$

In particular, we have $S_k(p) = O_k(1/p^2)$, which with our first assertion implies that the product for K_k converges to a positive real number that is less than 1. This completes the proof. □

Lemma 4. *For every positive integer k and each real number $x \geq 1$ we have*

$$\sum_{n \leq x} \frac{N_k(n)}{n} \leq 2(1 + \log x)^k.$$

Proof. Let $\omega(n)$ denote the number of distinct primes that divide n and let $\tau_k(n)$ denote the number of ordered factorizations of n into k positive integral factors. Since $k^{\omega(n)}$ is the number of ordered factorizations of n into k pairwise coprime factors, we have $k^{\omega(n)} \leq \tau_k(n)$ for all n. Further, from (1), (2) and the fact that $N_k(n)$ is multiplicative in the variable n, we have $N_k(n) \leq 2k^{\omega(n)}$, so that $N_k(n) \leq 2\tau_k(n)$. Thus, it suffices to show that

$$\sum_{n \leq x} \frac{\tau_k(n)}{n} \leq (1 + \log x)^k. \tag{3}$$

We prove (3) by induction on k. It holds for $k = 1$ since $\tau_1(n) = 1$ for all n, so that

$$\sum_{n \leq x} \frac{N_1(n)}{n} = \sum_{n \leq x} \frac{1}{n} \leq 1 + \int_1^x \frac{\mathrm{d}t}{t} = 1 + \log x.$$

Assume now that $k \geq 1$ and that (3) holds for k. Since

$$\tau_{k+1}(n) = \sum_{d \mid n} \tau_k(n),$$

we have

$$\sum_{n\le x}\frac{\tau_{k+1}(n)}{n}=\sum_{n\le x}\frac{1}{n}\sum_{d\mid n}\tau_k(d)=\sum_{d\le x}\frac{\tau_k(d)}{d}\sum_{m\le x/d}\frac{1}{m}$$
$$\le\sum_{d\le x}\frac{\tau_k(d)}{d}(1+\log x)\le(1+\log x)^{k+1},$$

by the induction hypothesis. This completes the proof. □

Corollary 5. *For k a positive integer and y a positive real with $k\le 1+\log y$, we have*

$$\sum_{n>y}\frac{N_k(n)}{n^2}\le 2(k+1)\frac{(1+\log y)^k}{y}.$$

Proof. By partial summation, Lemma 4, and integration by parts, we have

$$\sum_{n>y}\frac{N_k(n)}{n^2}=\int_y^\infty\frac{1}{t^2}\sum_{y<n\le t}\frac{N_k(n)}{n}\,\mathrm{d}t\le 2\int_y^\infty\frac{(1+\log t)^k}{t^2}\,\mathrm{d}t$$
$$=\frac{2}{y}\bigl((1+\log y)^k+k(1+\log y)^{k-1}+k(k-1)(1+\log y)^{k-2}+\dots+k!\bigr)$$
$$\le 2(k+1)\frac{(1+\log y)^k}{y},$$

using $k\le 1+\log y$. This completes the proof. □

3. The main theorem

Proof of Theorem 1. The function

$$\frac{\alpha(m)}{m}=\frac{1}{m^2}\sum_{n\mid m}n\varphi(n)$$

is multiplicative and so by Möbius inversion, we may write

$$\frac{\alpha(m)}{m}=\sum_{n\mid m}\gamma(n),$$

where γ is a multiplicative function. It is easy to compute that

$$\gamma(p^j)=-\frac{p-1}{p^{2j}}\tag{4}$$

for every prime p and positive integer j. If $\mathrm{rad}(n)$ denotes the largest squarefree divisor of n, we thus have

$$\gamma(n)=(-1)^{\omega(n)}\frac{\varphi(\mathrm{rad}(n))}{n^2}\tag{5}$$

for each positive integer n. Note that (4), (5) are also in [Luca 2005].

For n a positive integer, label the $N_k(n)$ roots to the congruence $s^k \equiv 1 \pmod{n}$ as $s_{k,1}, s_{k,2}, \dots, s_{k,N_k(n)}$. We have

$$\sum_{p\le x}\frac{\alpha(p^k-1)}{p^k-1}=\sum_{p\le x}\sum_{n\mid p^k-1}\gamma(n)=\sum_{n\le x^k-1}\gamma(n)\sum_{\substack{p\le x\\ n\mid p^k-1}}1$$
$$=\sum_{n\le x^k-1}\gamma(n)\sum_{i=1}^{N_k(n)}\pi(x;n,s_{k,i}),$$

where $\pi(x;q,a)$ denotes the number of primes $p\le x$ with $p\equiv a \pmod{q}$.

If q is not too large in comparison to x and if a is coprime to q, we expect $\pi(x;q,a)$ to be approximately $\pi(x)/\varphi(q)$. With this thought in mind, let $E_{q,a}(x)$ be defined by the equation

$$\pi(x;q,a)=\frac{1}{\varphi(q)}\pi(x)+E_{q,a}(x).$$

Further, let $y=x^{1/2}/\log^{A+4}x$, where A is as in the statement of Theorem 1. From the above, we thus have

$$\begin{aligned}&\sum_{p\le x}\frac{\alpha(p^k-1)}{p^k-1}\\ &=\sum_{n\le x^k-1}\gamma(n)\sum_{i=1}^{N_k(n)}\pi(x;n,s_{k,i})\\ &=\sum_{n\le y}\frac{\gamma(n)N_k(n)}{\varphi(n)}\pi(x)+\sum_{n\le y}\gamma(n)\sum_{i=1}^{N_k(n)}E_{n,s_{k_i}}(x)+\sum_{y<n\le x^k-1}\gamma(n)\sum_{i=1}^{N_k(n)}\pi(x;n,s_{k,i})\\ &=: T_1+T_2+T_3,\quad\text{say.}\end{aligned}$$

We further refine the main term T_1 as

$$T_1=\pi(x)\sum_{n=1}^{\infty}\frac{\gamma(n)N_k(n)}{\varphi(n)}-\pi(x)\sum_{n>y}\frac{\gamma(n)N_k(n)}{\varphi(n)}.$$

The first sum here has an Euler product as

$$\sum_{n=1}^{\infty}\frac{\gamma(n)N_k(n)}{\varphi(n)}=\prod_p\left(1+\sum_{j=1}^{\infty}\frac{\gamma(p^j)N_k(p^j)}{\varphi(p^j)}\right)=\prod_p\left(1-\sum_{j=1}^{\infty}\frac{N_k(p^j)}{p^{3j-1}}\right)=K_k,$$

where we used (4). For the second sum in the expression for T_1, we have by (5) and Corollary 5,

$$\left|\sum_{n>y}\frac{\gamma(n)N_k(n)}{\varphi(n)}\right| \le \sum_{n>y}\frac{N_k(n)}{n^2} \le 2(k+1)\frac{(1+\log y)^k}{y}.$$

Here we have used $k \le (\log x)/(2\log\log x)$ and $y = x^{1/2}/\log^{A+4} x$, so that $k \le 1+\log y$ for all sufficiently large x depending on the choice of A. Further, with these choices for k, y we have $(1+\log y)^k < x^{1/2}$ for x sufficiently large, so that

$$\pi(x)\left|\sum_{n>y}\frac{\gamma(n)N_k(n)}{\varphi(n)}\right| \le \pi(x)\frac{2(k+1)(1+\log y)^k}{y} \le \frac{\pi(x)}{\exp\dfrac{\log x}{3\log\log x}}$$

for all sufficiently large values of x depending on A. Thus,

$$T_1 = K_k\pi(x) + O_A(\pi(x)/\log^A x).$$

It remains to show that both T_2 and T_3 are $O_A(\pi(x)/\log^A x)$. Using the elementary estimate $\pi(x;q,a) \le 1 + x/q$, we have

$$|T_3| \le \sum_{y<n\le x^k-1}|\gamma(n)|N_k(n)\left(1+\frac{x}{n}\right) \le \sum_{y<n\le x^k-1}\frac{N_k(n)}{n} + x\sum_{y<n\le x^k-1}\frac{N_k(n)}{n^2},$$

by (5). We have seen that the second sum here is negligible, and the first sum is bounded by $2(1+k\log x)^k$ using Lemma 4. This last expression is smaller than

$$\left(\frac{\log^2 x}{\log\log x}\right)^k \le \frac{x}{\exp\dfrac{\log x\log\log\log x}{2\log\log x}} = O_A\left(\frac{\pi(x)}{\log^A x}\right)$$

for any fixed choice of A.

To estimate T_2, note that

$$\begin{aligned}|T_2| &\le \sum_{n\le y}|\gamma(n)|N_k(n)\max_{(a,n)=1}\left|\pi(x;n,a)-\frac{1}{\varphi(n)}\pi(x)\right|\\ &\le \sum_{n\le y}\max_{(a,n)=1}\left|\pi(x;n,a)-\frac{1}{\varphi(n)}\pi(x)\right|,\end{aligned}$$

since $|\gamma(n)| \le \varphi(n)/n^2 \le 1/n$ and $N_k(n) \le \varphi(n) \le n$. Thus, by the Bombieri–Vinogradov theorem (see [Davenport 2000, Chapter 28]) we have

$$|T_2| = O_A(\pi(x)/\log^A x),$$

by our choice of y. These estimates conclude our proof of Theorem 1. □

4. Proof of Corollary 2 and more on the constants K_k

Proof of Corollary 2. By partial summation, we have

$$\begin{aligned}\sum_{p\le x}\alpha(p^k-1) &= \sum_{p\le x}\frac{\alpha(p^k-1)}{p^k-1}(p^k-1)\\ &= (x^k-1)\sum_{p\le x}\frac{\alpha(p^k-1)}{p^k-1} - \int_2^x kt^{k-1}\sum_{p\le t}\frac{\alpha(p^k-1)}{p^k-1}\,\mathrm{d}t.\end{aligned}$$

Thus, by Theorem 1, the prime number theorem, and integration by parts, we have

$$\begin{aligned}\sum_{p\le x}\alpha(p^k-1) &= (x^k-1)K_k\pi(x) - \int_2^x kt^{k-1}K_k\pi(t)\,\mathrm{d}t + O\left(\frac{\pi(x)x^k}{\log^A x}\right)\\ &= (x^k-1)K_k\mathrm{li}(x) - \int_2^x kt^{k-1}K_k\mathrm{li}(t)\,\mathrm{d}t + O\left(\frac{\pi(x)x^k}{\log^A x}\right)\\ &= \int_2^x K_k\frac{t^k}{\log t}\,\mathrm{d}t + O\left(\frac{\pi(x)x^k}{\log^A x}\right).\end{aligned}$$

This last integral is $K_k\mathrm{li}(x^{k+1}) - K_k\mathrm{li}(2^{k+1})$, so the corollary now follows via one additional call to the prime number theorem. □

We now examine the constants K_k for $k \le 4$. Since $N_1(p^j) = 1$ for all p^j, we have

$$K_1 = \prod_p\left(1-\sum_{j\ge1}\frac{p}{p^{3j}}\right) = \prod_p\left(1-\frac{p}{p^3-1}\right) = 0.5759599689\ldots.$$

(This constant is also worked out in [Luca 2005].) For K_2 we note that $N_2(p^j) = 2$ for all prime powers p^j except that $N_2(2) = 1$ and $N_2(2^j) = 4$ for $j \ge 3$. Thus,

$$\sum_{j\ge1}\frac{N_2(2^j)}{2^{3j-1}} = \frac{1}{4}+\frac{2}{32}+\frac{1}{56} = \frac{37}{112},$$

and so

$$K_2 = \frac{75}{112}\prod_{p>2}\left(1-\frac{2p}{p^3-1}\right) = 0.4269891575\ldots.$$

For K_3, we have $N_3(p^j) = 3$ for $p \equiv 1 \pmod 3$ and for $p = 3$ and $j \ge 2$. Otherwise, $N_3(p^j) = 1$. Thus,

$$K_3 = \frac{205}{234}\prod_{p\equiv1 \pmod 3}\left(1-\frac{3p}{p^3-1}\right)\prod_{p\equiv2 \pmod 3}\left(1-\frac{p}{p^3-1}\right) = 0.6393087751\ldots.$$

For K_4, we have $N_4(p^j) = 4$ for $p \equiv 1 \pmod 4$, $N_4(p^j) = 2$ for $p \equiv 3 \pmod 4$, $N_4(2) = 1$, $N_4(2^2) = 2$, $N_4(2^3) = 4$, and $N_4(2^j) = 8$ for $j \geq 4$. Thus,

$$K_4 = \frac{299}{448} \prod_{p\equiv 1 \pmod 4} \left(1 - \frac{4p}{p^3-1}\right) \prod_{p\equiv 3 \pmod 4} \left(1 - \frac{2p}{p^3-1}\right) = 0.3775394971\ldots.$$

These calculations were done with the aid of Mathematica. With a little effort other constants K_k may be computed, but if k has many divisors, the calculation gets more tedious.

We close with the observation that there is an infinite sequence of numbers k on which $K_k \to 0$. In particular, if $k = k_m$ is the least common multiple of all numbers up to m, then $N_k(p) = p - 1$ for every prime $p \leq m+1$, so that

$$K_k < \prod_p \left(1 - \frac{N_k(p)}{p^2}\right) < \prod_{p\leq m+1} \left(1 - \frac{p-1}{p^2}\right).$$

Since $\sum (p-1)/p^2 = +\infty$, it follows that as $m \to \infty$, $K_{k_m} \to 0$. Using the theorem of Mertens, we in fact have $\liminf K_k \log\log k < +\infty$.

References

[Davenport 2000] H. Davenport, *Multiplicative number theory*, 3rd ed., Graduate Texts in Mathematics **74**, Springer, New York, 2000. MR 2001f:11001 Zbl 1002.11001

[Luca 2005] F. Luca, "Some mean values related to average multiplicative orders of elements in finite fields", *Ramanujan J.* **9**:1-2 (2005), 33–44. MR 2006i:11111 Zbl 1155.11344

[von zur Gathen 2004] J. von zur Gathen, A. Knopfmacher, F. Luca, L. G. Lucht, and I. E. Shparlinski, "Average order in cyclic groups", *J. Théor. Nombres Bordeaux* **16**:1 (2004), 107–123. MR 2006d:11111 Zbl 1079.11003

Received: 2011-12-07 Revised: 2012-01-11 Accepted: 2012-01-13

yilan.hu.10@alum.dartmouth.org *55 Maple Hill Road, Thetford Center, VT 05075,*
United States

carl.pomerance@dartmouth.edu *Mathematics Department, Kemeny Hall,*
Dartmouth College, Hanover, NH 03755, United States
www.math.dartmouth.edu/~carlp

Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the Submission page at the Involve website.

Originality. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles in *Involve* are usually in English, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

Format. Authors are encouraged to use LaTeX but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@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

2012 vol. 5 no. 2