involve

a journal of mathematics

msp

2013 vol. 6, no. 1

involve

msp.org/involve

EDITORS

MANAGING EDITOR

Kenneth S. Berenhaut, Wake Forest University, USA, berenhks@wfu.edu

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2013 is US $105/year for the electronic version, and $145/year (+$35, if 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 MSP.

Involve (ISSN 1944-4184 electronic, 1944-4176 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840, is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

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

PUBLISHED BY
mathematical sciences publishers
nonprofit scientific publishing

http://msp.org/

INVOLVE 6:1 (2013)
dx.doi.org/10.2140/involve.2013.6.1

Refined inertias of tree sign patterns of orders 2 and 3

D. D. Olesky, Michael F. Rempel and P. van den Driessche

(Communicated by Charles R. Johnson)

Sign patterns are matrices with only the sign of each entry specified. The refined inertia of a matrix categorizes the eigenvalues as positive, negative, zero or nonzero imaginary, and the refined inertia of a sign pattern is the set of all refined inertias allowed by real matrices with that sign pattern. The complete sets of allowed refined inertias for all tree sign patterns of orders 2 and 3 (up to equivalence and negation) are determined.

1. Introduction

The *inertia* of an $n \times n$ real matrix A, denoted by $\mathrm{i}(A)$, is the triple (n_+, n_-, n_0), where n_+, n_- and n_0 are the numbers of eigenvalues with positive, negative and zero real part, respectively. Note that $n_+ + n_- + n_0 = n$. The *refined inertia* of A, $\mathrm{ri}(A)$, is the 4-tuple $(n_+, n_-, n_z, 2n_p)$ with n_+ and n_- as above, n_z the number of zero eigenvalues and n_p the number of nonzero imaginary complex conjugate pairs of eigenvalues; see [Kim et al. 2009]. The refined inertia of A distinguishes between zero and nonzero imaginary eigenvalues, which is important for linear dynamical systems.

A *sign pattern* $\mathcal{A} = [\alpha_{ij}]$ of order n is an $n \times n$ matrix with entries in $\{+, -, 0\}$. A real matrix A is a *realization* of $\mathcal{A}$ if the signs of the entries in A correspond to the entries in $\mathcal{A}$. The *sign pattern class* of $\mathcal{A}$ is $Q(\mathcal{A}) = \{A \mid A \text{ is a realization of } \mathcal{A}\}$. A sign pattern $\mathcal{B} = [\beta_{ij}]$ is a *superpattern* of $\mathcal{A}$ if $\beta_{ij} = \alpha_{ij}$ for all $\alpha_{ij} \in \{+, -\}$. The inertia of a sign pattern $\mathcal{A}$ is $\mathrm{i}(\mathcal{A}) = \{\mathrm{i}(A) \mid A \in Q(\mathcal{A})\}$ and the refined inertia of $\mathcal{A}$ is $\mathrm{ri}(\mathcal{A}) = \{\mathrm{ri}(A) \mid A \in Q(\mathcal{A})\}$. A sign pattern $\mathcal{A}$ *allows* a refined inertia $(n_+, n_-, n_z, 2n_p)$ if there exists some $A \in Q(\mathcal{A})$ having this refined inertia. See, for example, [Catral et al. 2009; Johnson and Summers 1989] for related allow problems on sign patterns.

Sign patterns have applications in areas where dynamical systems arise (see, for example, [Logofet 1993]), but characterizing sign patterns that have a particular

MSC2010: 15A18, 15B35.
Keywords: sign pattern, inertia, refined inertia, allow problems, characteristic equation, eigenvalues.
This research was supported in part by NSERC Discovery grants.

property can be challenging since each nonzero entry is free to take on any value in one half of the real line. Section 2 introduces more definitions and concepts required for our analysis of sign patterns. Sections 3 and 4 identify the refined inertias allowed by all tree sign patterns of orders 2 and 3, respectively, and these are listed in Appendices A and B.

2. Fundamentals

Given an $n \times n$ sign pattern $\mathcal{A} = [\alpha_{ij}]$, the *transpose* of $\mathcal{A}$ is $\mathcal{A}^T = [\alpha_{ji}]$. A *permutation similarity transformation* is $\mathcal{A} \mapsto P\mathcal{A}P^T$ where P is an $n \times n$ permutation matrix. A *signature similarity transformation* is $\mathcal{A} \mapsto D\mathcal{A}D^{-1}$ where $D = D^{-1}$ is an $n \times n$ diagonal matrix with each diagonal entry equal to ± 1. The refined inertia of a sign pattern $\mathcal{A}$ is preserved by each of these three transformations, which define equivalence classes of sign patterns. Two sign patterns $\mathcal{A}$ and $\mathcal{B}$ are *equivalent*, and therefore in the same equivalence class, if $\mathcal{B}$ can be derived from $\mathcal{A}$ by some sequence of the three transformations. One other important transformation is *negation*. Since $\mathrm{ri}(-\mathcal{A}) = \{(n_-, n_+, n_z, 2n_p) \mid (n_+, n_-, n_z, 2n_p) \in \mathrm{ri}(\mathcal{A})\}$, the refined inertia (and inertia) of $-\mathcal{A}$ is easily obtained from that of $\mathcal{A}$.

An $n \times n$ sign pattern $\mathcal{A}$ is a *spectrally arbitrary* pattern (SAP) if, given any set of n complex numbers closed under complex conjugation, there exists a realization $A \in Q(\mathcal{A})$ having these n numbers as its eigenvalues [Drew et al. 2000], and $\mathcal{A}$ is a *refined inertially arbitrary* pattern (rIAP) if, given any 4-tuple $(n_+, n_-, n_z, 2n_p)$ with $n_+ + n_- + n_z + 2n_p = n$, there exists a realization of $\mathcal{A}$ having this 4-tuple as its refined inertia [Kim et al. 2009]. An inertially arbitrary pattern (IAP) is defined similarly [Drew et al. 2000]. The properties of being an IAP, rIAP and SAP are invariant with respect to equivalence and negation. Any SAP is obviously an rIAP, and any rIAP is also an IAP. Conversely, for $n = 2$ and 3, $\mathcal{A}$ is a SAP if it is an rIAP [Kim et al. 2009], but it is not known if this holds for larger n.

A *tree sign pattern* $\mathcal{A} = [\alpha_{ij}]$ is a sign pattern that is combinatorially symmetric (i.e., $\alpha_{ij} \neq 0$ whenever $\alpha_{ji} \neq 0$) and has $\alpha_{i_1i_2}\alpha_{i_2i_3} \cdots \alpha_{i_ki_1} = 0$ for all $k \geq 3$. As in [Johnson and Summers 1989], associated with an $n \times n$ tree sign pattern $\mathcal{A}$ is a signed tree graph with n vertices labeled $1, 2, \dots, n$ and an edge between vertices i and $j \neq i$ if and only if α_{ij} is not 0; the sign on the edge is the sign of the product $\alpha_{ij}\alpha_{ji}$. In addition, if $\alpha_{ii} \neq 0$ the graph has a loop at vertex i, signed as the sign of α_{ii}. The negation of the sign pattern changes the signs of the loops of the graph, but not the signs of its edges. Every graph in the appendices uniquely represents those sign patterns that are the same up to equivalence, with one such sign pattern shown for each class. For $n = 2$ every irreducible sign pattern is a tree sign pattern. For $n = 3$, a tree sign pattern is equivalent to an irreducible tridiagonal sign pattern. We consider only irreducible sign patterns because the refined inertia of a reducible sign pattern is the sumset of the refined inertias of its irreducible components.

3. Sign patterns of order 2

For completeness, the refined inertias and graphs of all irreducible sign patterns of order 2 up to equivalence and negation are given in Appendix A. These can be determined simply by considering the trace and determinant of a real matrix with each sign pattern. The trace of a matrix is equal to the sum of its eigenvalues, and its determinant is equal to the product of its eigenvalues. For the 2×2 case the possible signs of the trace and determinant provide complete information on the refined inertia. For example, for an order 2 sign pattern, if the trace must be positive and the determinant must be negative, the only allowed refined inertia is $(1, 1, 0, 0)$, with the positive eigenvalue larger in magnitude than the negative one. Note that only one sign pattern of order 2 is an rIAP.

4. Tree sign patterns of order 3

Since there are a total of 3^9 sign patterns of order 3, a computer program was written to identify the set of all irreducible order 3 sign patterns up to equivalence and negation. This set was determined by examining in turn every possible 3×3 irreducible sign pattern and checking for equivalence or negation with each sign pattern already in the set. The program identified 187 sign patterns, of which 34 were tree sign patterns. The equivalence classes corresponding to the 34 tree sign patterns are each represented by a graph in Appendix B, along with their refined inertias, a representative sign pattern and its associated characteristic equation, and references to techniques used for finding the refined inertias.

We now present our techniques and methods for finding the exact set of refined inertias allowed by each sign pattern. For every sign pattern, each refined inertia was either proved to not be allowed or shown to be allowed either by a proof or a numerical realization, respectively.

A tree sign pattern can be represented by a tridiagonal matrix with (real) variables for each nonzero entry. The generic 3×3 tridiagonal matrix is

$$\begin{bmatrix} \pm e & \pm a & 0 \\ \pm b & \pm f & \pm c \\ 0 & \pm d & \pm g \end{bmatrix}. \tag{4-1}$$

Here a, b, c, $d > 0$ for an irreducible tree sign pattern and e, f, $g \geq 0$. For the moment we let $a' = \pm a$, and similarly for the other variables, but when working with specific sign patterns, we always use strictly positive variables.

We can normalize the matrix in (4-1) to reduce the number of unknowns by up to three, setting them to ± 1. If $e \neq 0$ we set $e = 1$ by multiplying the matrix by $1/e$, and if $e = 0$ and $f \neq 0$ we multiply by $1/f$ to set $f = 1$. By Lemma 2.3 in [Britz et al. 2004] we can also set $a = c = 1$, and since each 3×3 tree sign pattern has a', $c' > 0$ up to equivalence, $a' = c' = 1$. Thus the characteristic polynomial can be simplified to

$$x^3 - (e' + f' + g')x^2 + (e'f' + e'g' + f'g' - b' - d')x + e'd' + b'g' - e'f'g', \quad (4\text{-}2)$$

with one of e', f' equal to 1 or -1, or $e' = f' = g' = 0$.

In terms of the characteristic polynomial, the trace is the negative of the coefficient of x^2 and the determinant is the negative of the constant term. Each refined inertia corresponds to a unique product of three linear factors, i.e., to a unique factorization of a monic cubic polynomial. The 13 possible refined inertias for 3×3 matrices are listed in Table 1 with their factorizations. After expanding each factorization, the coefficients of the resulting polynomial can be compared directly to the coefficients of the characteristic polynomial in (4-2). If there is no solution to the resulting set of equations, then the corresponding refined inertia is not allowed by the sign pattern with the given characteristic polynomial.

The real parts of α, β are positive and $\gamma > 0$. Note that for refined inertias (a)–(d), (g) and (i) in Table 1, the unknowns α and β may be complex conjugate pairs. However, since $\alpha\beta$ and $\alpha + \beta$ are always real and positive, when used in these combinations the fact that α and β are possibly complex can be ignored.

The following list summarizes the techniques that we used to determine the refined inertias. Each sign pattern in Appendix B references the techniques used for that sign pattern.

T1. The rIAPs (equivalently SAPs for $n = 3$) were found by determining which of the 34 tree sign patterns are equivalent to one of the two 3×3 SAPs that are

	refined inertia	factorization	characteristic polynomial
(a)	(3,0,0,0)	$(x-\alpha)(x-\beta)(x-\gamma)$	$x^3 - (\alpha+\beta+\gamma)x^2 + (\alpha\beta + (\alpha+\beta)\gamma)x - \alpha\beta\gamma$
(b)	(2,1,0,0)	$(x-\alpha)(x-\beta)(x+\gamma)$	$x^3 - (\alpha+\beta-\gamma)x^2 + (\alpha\beta - (\alpha+\beta)\gamma)x + \alpha\beta\gamma$
(c)	(1,2,0,0)	$(x+\alpha)(x+\beta)(x-\gamma)$	$x^3 + (\alpha+\beta-\gamma)x^2 + (\alpha\beta - (\alpha+\beta)\gamma)x - \alpha\beta\gamma$
(d)	(0,3,0,0)	$(x+\alpha)(x+\beta)(x+\gamma)$	$x^3 + (\alpha+\beta+\gamma)x^2 + (\alpha\beta + (\alpha+\beta)\gamma)x + \alpha\beta\gamma$
(e)	(1,0,0,2)	$(x-\alpha)(x^2+\beta)$	$x^3 - \alpha x^2 + \beta x - \alpha\beta$
(f)	(0,1,0,2)	$(x+\alpha)(x^2+\beta)$	$x^3 + \alpha x^2 + \beta x + \alpha\beta$
(g)	(2,0,1,0)	$x(x-\alpha)(x-\beta)$	$x^3 - (\alpha+\beta)x^2 + \alpha\beta x$
(h)	(1,1,1,0)	$x(x+\alpha)(x-\beta)$	$x^3 + (\alpha-\beta)x^2 - \alpha\beta x$
(i)	(0,2,1,0)	$x(x+\alpha)(x+\beta)$	$x^3 + (\alpha+\beta)x^2 + \alpha\beta x$
(j)	(0,0,1,2)	$x(x^2+\alpha)$	$x^3 + \alpha x$
(k)	(1,0,2,0)	$x^2(x-\alpha)$	$x^3 - \alpha x^2$
(l)	(0,1,2,0)	$x^2(x+\alpha)$	$x^3 + \alpha x^2$
(m)	(0,0,3,0)	x^3	x^3

Table 1. All refined inertias for matrices of order 3.

trees, namely $\mathcal{T}_3$ and $\mathcal{U}_3$ in [Britz et al. 2004] (see also [Cavers and Vander Meulen 2005]), or are equivalent to a superpattern of one of them.

T2. If $e = g = 0$, then the determinant must be zero and the trace is ± 1 or 0. The characteristic polynomial factors into a zero root and a quadratic, and the possible refined inertias are easily determined.

T3. In order to find a realization of a given sign pattern that has a refined inertia with $n_z = n_p = 0$, a random matrix with that sign pattern was generated in MATLAB, and its eigenvalues were computed. This ad hoc technique was used for many sign patterns to find an example of refined inertias (a)–(d) in Table 1 that each allows, although it does not show the nonexistence of those that are not allowed.

T4. If a tridiagonal sign pattern $\mathcal{A}$ is symmetric, then any $A \in Q(\mathcal{A})$ is diagonally similar to a symmetric matrix, so its eigenvalues are real. Thus the refined inertias (e)–(f) and (j) in Table 1 are not allowed for such sign patterns.

T5. If the determinant must be positive or must be negative, then the sign pattern does not allow (g)–(m) in Table 1, as well as either (b), (d), (f) if positive, or (a), (c), (e) if negative.

T6. If the trace must be positive, then the sign pattern does not allow (d), (f), (i), (j), (l) and (m) in Table 1, and if negative it does not allow (a), (e), (g), (j), (k) and (m).

T7. If the sign pattern is such that the coefficient of the x term in (4-2) is negative, then only refined inertias (b), (c) and (h) in Table 1 can be allowed.

Note. The next four techniques are algebraic, involving the characteristic polynomial (4-2) and equations in Table 1. There are many possible ways of proceeding; however, we found the following techniques to be the most straightforward.

T8. To show that one of (e) or (f) in Table 1 is not allowed, equate the coefficients of the characteristic polynomial (4-2) to the coefficients of the polynomial corresponding to the refined inertia being considered. If this leads to a contradiction, then that refined inertia is not allowed.

For example, consider sign pattern (4e) in Appendix B. Equating its characteristic polynomial to the polynomial associated with (0, 1, 0, 2) ((f) in Table 1) gives

$$1 - f = \alpha \implies \alpha < 1, \quad d - f - b = \beta \Leftrightarrow d = \beta + b + f, \quad d = \alpha\beta.$$

A contradiction is immediate since $0 < \alpha < 1$ implies $d < \beta$ from the last equation, but the second equation implies $d > \beta$. Therefore this sign pattern does not allow refined inertia (0, 1, 0, 2).

A more complicated argument shows that sign pattern (5d) in Appendix B does not allow refined inertia (1, 0, 0, 2). In this case

$$1 + f + g = \alpha \implies \alpha > g \text{ and } \alpha > 1,$$

while

$$f+g+fg+b+d=\beta \implies \beta > fg+d+b$$

and $d+bg+fg=\alpha\beta$. From the inequality for β it follows that $\alpha\beta > \alpha(fg+d+b) = \alpha(fg+d)+\alpha b$ and from the inequalities for α it follows that $\alpha(fg+d)+\alpha b > fg+d+gb=\alpha\beta$, which is a contradiction. This method can also be used to show that (a) or (d) in Table 1 are not allowed, but it is easier to invoke continuity when appropriate (see T12).

T9. This technique is for eliminating or identifying allowed refined inertias with at least one eigenvalue equal to zero (refined inertias (g)–(m) in Table 1) when the determinant is not necessarily zero (hence differing from T2). The objective is to determine inequalities between unknowns in the characteristic equation that require the coefficient of x to have a certain sign.

If the coefficient of x must be positive, only (g), (i) and/or (j) in Table 1 can be allowed, while if it must be negative, only (h) is allowed. This argument can also be viewed in terms of the discriminant of the quadratic that arises after factoring out the zero root from the characteristic polynomial.

As a simple example, consider sign pattern (10a) in Appendix B with the coefficient of x equal to $f+g+fg-b-d$. Here we are interested in the case $d+bg-fg=0$ (i.e., at least one zero eigenvalue). This equation gives

$$fg=d+bg \implies fg>d \text{ and } f>b.$$

These imply that $f-b+fg-d+g>0$. Since the trace $1+f+g$ must be positive, (2, 0, 1, 0) is the only refined inertia with a zero eigenvalue that sign pattern (10a) allows, eliminating (h)–(m) in Table 1.

A more complex use of this technique is needed for sign pattern (11d) in Appendix B, with the coefficient of x equal to $g-f-fg+b+d$. Again we set the determinant to zero: $fg-d-bg=0$. Considering the case with trace not positive ($f \geq 1+g$) gives $f>1$, $f>g$, so:

- If $g \leq 1$, then

  $$fg=bg+d \implies fg \geq bg+dg \implies f \geq b+d$$

  and $fg>g$. Therefore the coefficient of x is negative.
- If $g>1$, then

  $$fg=bg+d \implies fg>b+d$$

  and $f>g$. Therefore the coefficient of x is negative.

Thus, when the trace is negative or zero and the determinant is zero, only (1, 1, 1, 0) is allowed, eliminating (i)–(j) and (l)–(m) in Table 1.

For certain sign patterns, this technique can also be used to show that when the determinant and trace are taken to be negative, then the coefficient of x is also negative, and therefore (d) and (f) in Table 1 are not allowed. For sign pattern (11d) the argument follows as in the above example, except that the equality $fg = bg + d$ is replaced by the inequality $fg > bg + d$.

Note. The next two techniques are for finding realizations with certain refined inertias when continuity cannot be obviously invoked. This could be done simply by trial and error, but it is easier to do some algebra first.

T10. The first technique is for finding realizations with a zero eigenvalue. First fix the trace to be either positive or negative, and then find inequalities that ensure that the coefficient of x is positive, negative and/or zero.

As an example, consider sign pattern (11c) in Appendix B with the coefficient of x equal to $f - g - fg - b + d$. For this sign pattern $fg - d - bg = 0$ ensures a zero eigenvalue while $1 + f > g$ implies a positive trace.

To find a realization with refined inertia (k) in Table 1, set

$$f + d = g + b + fg = g + b + d + bg \implies f - g = b(g + 1).$$

This reduces the number of unknowns by one, and further note that a solution to the last equation ensures that the trace is positive; therefore that condition can be ignored. Also, a solution implies $f > b$, and since $d = g(f - b)$, d is positive for all solutions. An obvious solution is then $b = 1$, $g = 1$, $f = 3$.

Similarly, to show that (g) in Table 1 is allowed, a realization is required with positive trace, positive coefficient of x and zero determinant, which imply that $f - g > b(g + 1)$. Thus, in the above solution, increase the value of f. Similarly, to show that (h) in Table 1 is allowed, increase b so that $f - g < b(g + 1)$ while maintaining $b < f$.

T11. In order to determine a realization with refined inertia (e) or (f) in Table 1, begin as in T8, but now instead of trying to reach a contradiction, the goal is to find a reduced set of expressions that will ensure that all unknowns are positive, and therefore show that the refined inertia exists.

For example consider sign pattern (11c) as in the previous example. Equating the coefficients of the characteristic polynomial with those of (e) in Table 1 gives

$$1 + f - g = \alpha \implies f - g = \alpha - 1,$$
$$f - g - fg - b + d = \beta, \quad d + bg - fg = \alpha\beta.$$

From the last equation, choosing $f > b$ ensures $d > 0$. Combining all three equations to eliminate d and f gives

$$b(1 + g) = \alpha\beta + \alpha - \beta - 1 = (\alpha - 1)(\beta + 1).$$

Therefore, by choosing $g > \beta$, necessarily $b < \alpha - 1 < \alpha - 1 + g = f$, which gives two simple expressions, namely $\alpha > 1$ and $g > \beta > 0$, that ensure all variables remain positive. A solution is, for example, $\alpha = 2$, $\beta = 1$, $g = 3$ giving $f = 4$, $b = 0.5$, $d = 12.5$.

T12. Since the eigenvalues are continuous functions of the matrix entries, continuity can require that a sign pattern allow or not allow a particular refined inertia. This can often be used after the set of possible refined inertias is narrowed down.

For an example of the first case, if a sign pattern has been shown to allow $(3, 0, 0, 0)$ and $(1, 2, 0, 0)$ and the determinant is nonzero, then the sign pattern must also allow $(1, 0, 0, 2)$.

For an example of the second case, if a sign pattern allows $(3, 0, 0, 0)$ but no others except possibly $(2, 1, 0, 0)$, as in Appendix B(5), by continuity $(2, 1, 0, 0)$ is also not possible since an eigenvalue would have to cross the imaginary axis, and therefore $(2, 0, 1, 0)$ would also have to be allowed.

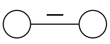
For each tree sign pattern of order 3 (up to equivalence and negation), the above techniques determine the set of all allowed refined inertias. Appendix B contains a list of sign patterns arranged according to these sets, and also includes the graph corresponding to each equivalence class. This list suggests some open questions. Given a list of refined inertias, what classes of sign patterns allow exactly these refined inertias, and which lists have at least one such sign pattern?

Appendix A. Refined inertias of all 2 x 2 irreducible sign patterns up to equivalence and negation

(1) $\mathrm{ri}(\mathscr{A}) = \{(0, 0, 0, 2)\}$

(1a)

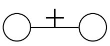
$\begin{bmatrix} 0 & + \\ - & 0 \end{bmatrix}$



(2) $\mathrm{ri}(\mathscr{A}) = \{(1, 1, 0, 0)\}$

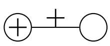
(2a)

$\begin{bmatrix} 0 & + \\ + & 0 \end{bmatrix}$



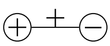
(2b)

$\begin{bmatrix} + & + \\ + & 0 \end{bmatrix}$



(2c)

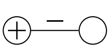
$\begin{bmatrix} + & + \\ + & - \end{bmatrix}$



(3) $\mathrm{ri}(\mathscr{A}) = \{(2, 0, 0, 0)\}$

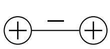
(3a)

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



(3b)

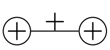
$\begin{bmatrix} + & + \\ - & + \end{bmatrix}$



(4) $\mathrm{ri}(\mathscr{A}) = \{(2, 0, 0, 0), (1, 1, 0, 0), (1, 0, 1, 0)\}$

(4a)

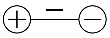
$\begin{bmatrix} + & + \\ + & + \end{bmatrix}$



(5) rIAP (allows all 7 refined inertias)

(5a)

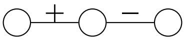
$$\begin{bmatrix} + & + \\ - & - \end{bmatrix}$$



Appendix B. Refined inertias of all 3 × 3 tree sign patterns up to equivalence and negation

(1) $\mathrm{ri}(\mathscr{A}) = \{(1, 1, 1, 0), (0, 0, 3, 0), (0, 0, 1, 2)\}$

(1a) $\begin{bmatrix} 0 & + & 0 \\ + & 0 & + \\ 0 & - & 0 \end{bmatrix}$

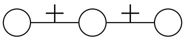


$x^3 - (b - d)x$

Technique: T2

(2) $\mathrm{ri}(\mathscr{A}) = \{(1, 1, 1, 0)\}$

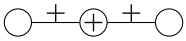
(2a) $\begin{bmatrix} 0 & + & 0 \\ + & 0 & + \\ 0 & + & 0 \end{bmatrix}$



$x^3 - (b + d)x$

Technique: T2

(2b) $\begin{bmatrix} 0 & + & 0 \\ + & + & + \\ 0 & + & 0 \end{bmatrix}$

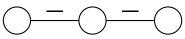


$x^3 - x^2 - (b + d)x$

Technique: T2

(3) $\mathrm{ri}(\mathscr{A}) = \{(0, 0, 1, 2)\}$

(3a) $\begin{bmatrix} 0 & + & 0 \\ - & 0 & + \\ 0 & - & 0 \end{bmatrix}$

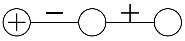


$x^3 + (b + d)x$

Technique: T2

(4) $\mathrm{ri}(\mathscr{A}) = \{(2, 1, 0, 0)\}$

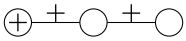
(4a) $\begin{bmatrix} + & + & 0 \\ - & 0 & + \\ 0 & + & 0 \end{bmatrix}$



$x^3 - x^2 + (b - d)x + d$

Techniques: T3, T5, T6

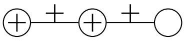
(4b) $\begin{bmatrix} + & + & 0 \\ + & 0 & + \\ 0 & + & 0 \end{bmatrix}$



$x^3 - x^2 - (b + d)x + d$

Techniques: T5, T7

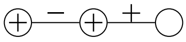
(4c) $\begin{bmatrix} + & + & 0 \\ + & + & + \\ 0 & + & 0 \end{bmatrix}$



$x^3 - (1 + f)x^2 + (f - b - d)x + d$

Techniques: T3, T5, T6

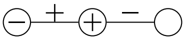
(4d) $\begin{bmatrix} + & + & 0 \\ - & + & + \\ 0 & + & 0 \end{bmatrix}$



$x^3 - (1 + f)x^2 + (f + b - d)x + d$

Techniques: T3, T5, T6

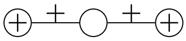
(4e) $\begin{bmatrix} - & + & 0 \\ + & + & + \\ 0 & - & 0 \end{bmatrix}$



$x^3 + (1 - f)x^2 - (f + b - d)x + d$

Techniques: T3, T5, T8, T12

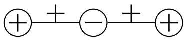
(4f) $\begin{bmatrix} + & + & 0 \\ + & 0 & + \\ 0 & + & + \end{bmatrix}$



$x^3-(1+g)x^2+(g-b-d)x+d+bg$

Techniques: T3, T5, T6

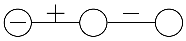
(4g) $\begin{bmatrix} + & + & 0 \\ + & - & + \\ 0 & + & + \end{bmatrix}$



$x^3-(1-f+g)x^2-(f-g+fg+b+d)x+d+bg+fg$

Techniques: T3, T4, T5, T12

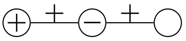
(4h) $\begin{bmatrix} - & + & 0 \\ + & 0 & + \\ 0 & - & 0 \end{bmatrix}$



$x^3+x^2-(b-d)x+d$

Techniques: T3, T5, T8, T12

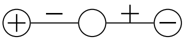
(4i) $\begin{bmatrix} + & + & 0 \\ + & - & + \\ 0 & + & 0 \end{bmatrix}$



$x^3-(1-f)x^2-(f+b+d)x+d$

Techniques: T5, T7

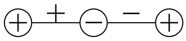
(4j) $\begin{bmatrix} + & + & 0 \\ - & 0 & + \\ 0 & + & - \end{bmatrix}$



$x^3-(1-g)x^2-(g-b+d)x+d+bg$

Techniques: T3, T5, T8, T12

(4k) $\begin{bmatrix} + & + & 0 \\ - & + & + \\ 0 & + & - \end{bmatrix}$

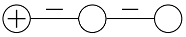


$x^3-(1+f-g)x^2+(f-g-fg+b-d)x+d+bg+fg$

Techniques: T3, T5, T8, T12

(5) $\mathrm{ri}(\mathscr{A})=\{(3,0,0,0)\}$

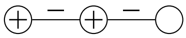
(5a) $\begin{bmatrix} + & + & 0 \\ - & 0 & + \\ 0 & - & 0 \end{bmatrix}$



$x^3-x^2+(b+d)x-d$

Techniques: T3, T5, T8, T12

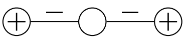
(5b) $\begin{bmatrix} + & + & 0 \\ - & + & + \\ 0 & - & 0 \end{bmatrix}$



$x^3-(1+f)x^2+(f+b+d)x-d$

Techniques: T3, T5, T8, T12

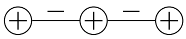
(5c) $\begin{bmatrix} + & + & 0 \\ - & 0 & + \\ 0 & - & + \end{bmatrix}$



$x^3-(1+g)x^2+(g+b+d)x-d-bg$

Techniques: T3, T5, T8, T12

(5d) $\begin{bmatrix} + & + & 0 \\ - & + & + \\ 0 & - & + \end{bmatrix}$

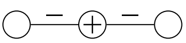


$x^3-(1+f+g)x^2+(f+g+fg+b+d)x-d-bg-fg$

Techniques: T3, T5, T8, T12

(6) $\mathrm{ri}(\mathscr{A})=\{(2,0,1,0)\}$

(6a) $\begin{bmatrix} 0 & + & 0 \\ - & + & + \\ 0 & - & 0 \end{bmatrix}$

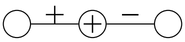


$x^3-x^2+(b+d)x$

Technique: T2

(7) $\mathrm{ri}(\mathscr{A})=\{(1,0,2,0),\,(1,1,1,0),\,(2,0,1,0)\}$

(7a) $\begin{bmatrix} 0 & + & 0 \\ + & + & + \\ 0 & - & 0 \end{bmatrix}$

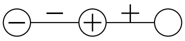


$x^3 - x^2 - (b-d)x$

Technique: T2

(8) $\mathrm{ri}(\mathscr{A}) = \{(3,0,0,0), (1,2,0,0), (1,0,0,2)\}$

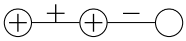
(8a) $\begin{bmatrix} - & + & 0 \\ - & + & + \\ 0 & + & 0 \end{bmatrix}$



$x^3 + (1-f)x^2 - (f-b+d)x - d$

Techniques: T3, T5, T12

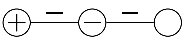
(8b) $\begin{bmatrix} + & + & 0 \\ + & + & + \\ 0 & - & 0 \end{bmatrix}$



$x^3 - (1+f)x^2 + (f-b+d)x - d$

Techniques: T3, T5, T12

(8c) $\begin{bmatrix} + & + & 0 \\ - & - & + \\ 0 & - & 0 \end{bmatrix}$

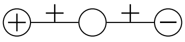


$x^3 - (1-f)x^2 - (f-b-d)x - d$

Techniques: T3, T5, T12

(9) $\mathrm{ri}(\mathscr{A}) = \{(2,1,0,0), (1,2,0,0), (1,1,1,0)\}$

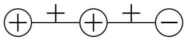
(9a) $\begin{bmatrix} + & + & 0 \\ + & 0 & + \\ 0 & + & - \end{bmatrix}$



$x^3 - (1-g)x^2 - (g+b+d)x + d - bg$

Techniques: T3, T7, T12

(9b) $\begin{bmatrix} + & + & 0 \\ + & + & + \\ 0 & + & - \end{bmatrix}$

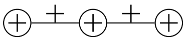


$x^3 - (1+f-g)x^2 + (f-g-fg-b-d)x + d - bg + fg$

Techniques: T3, T4, T9, T12

(10) $\mathrm{ri}(\mathscr{A}) = \{(3,0,0,0), (2,1,0,0), (2,0,1,0)\}$

(10a) $\begin{bmatrix} + & + & 0 \\ + & + & + \\ 0 & + & + \end{bmatrix}$

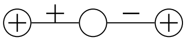


$x^3 - (1+f+g)x^2 + (f+g+fg-b-d)x + d + bg - fg$

Techniques: T3, T4, T5, T9, T12

(11) $\mathrm{ri}(\mathscr{A}) = \{(3,0,0,0), (1,2,0,0), (2,1,0,0), (2,0,1,0),$
$(1,0,2,0), (1,1,1,0), (1,0,0,2)\}$

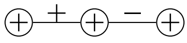
(11a) $\begin{bmatrix} + & + & 0 \\ + & 0 & + \\ 0 & - & + \end{bmatrix}$



$x^3 - (1+g)x^2 + (g-b+d)x - d + bg$

Techniques: T3, T6, T10, T11

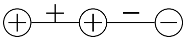
(11b) $\begin{bmatrix} + & + & 0 \\ + & + & + \\ 0 & - & + \end{bmatrix}$



$x^3 - (1+f+g)x^2 + (f+g+fg-b+d)x - d + bg - fg$

Techniques: T3, T6, T10, T11

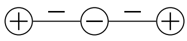
(11c) $\begin{bmatrix} + & + & 0 \\ + & + & + \\ 0 & - & - \end{bmatrix}$



$x^3 - (1+f-g)x^2 + (f-g-fg-b+d)x - d - bg + fg$

Techniques: T3, T9, T10, T11

(11d) $\begin{bmatrix} + & + & 0 \\ - & - & + \\ 0 & - & + \end{bmatrix}$

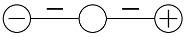


$x^3 - (1-f+g)x^2 - (f-g+fg-b-d)x - d - bg + fg$

Techniques: T3, T9, T10, T11

(12) rIAP (allows all 13 refined inertias)

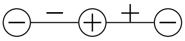
(12a) $\begin{bmatrix} - & + & 0\\ - & 0 & +\\ 0 & - & + \end{bmatrix}$



$x^3+(1-g)x^2-(g-b-d)x+d-bg$

Technique: T1

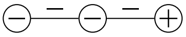
(12b) $\begin{bmatrix} - & + & 0\\ - & + & +\\ 0 & + & - \end{bmatrix}$



$x^3+(1-f+g)x^2-(f-g+fg-b+d)x-d+bg-fg$

Technique: T1

(12c) $\begin{bmatrix} - & + & 0\\ - & - & +\\ 0 & - & + \end{bmatrix}$



$x^3+(1+f-g)x^2+(f-g-fg+b+d)x+d-bg-fg$

Technique: T1

Acknowledgements

We thank Charles R. Johnson for his valuable comments.

References

[Britz et al. 2004] T. Britz, J. J. McDonald, D. D. Olesky, and P. van den Driessche, "Minimal spectrally arbitrary sign patterns", *SIAM J. Matrix Anal. Appl.* **26**:1 (2004), 257–271. MR 2005h:15030 Zbl 1082.15016

[Catral et al. 2009] M. Catral, D. D. Olesky, and P. van den Driessche, "Allow problems concerning spectral properties of sign pattern matrices: a survey", *Linear Algebra Appl.* **430**:11-12 (2009), 3080–3094. MR 2010i:15066 Zbl 1165.15009

[Cavers and Vander Meulen 2005] M. S. Cavers and K. N. Vander Meulen, "Spectrally and inertially arbitrary sign patterns", *Linear Algebra Appl.* **394** (2005), 53–72. MR 2005f:15008 Zbl 1065.15009

[Drew et al. 2000] J. H. Drew, C. R. Johnson, D. D. Olesky, and P. van den Driessche, "Spectrally arbitrary patterns", *Linear Algebra Appl.* **308**:1-3 (2000), 121–137. MR 2001c:15012 Zbl 0957.15012

[Johnson and Summers 1989] C. R. Johnson and T. A. Summers, "The potentially stable tree sign patterns for dimensions less than five", *Linear Algebra Appl.* **126** (1989), 1–13. MR 91f:15043 Zbl 0723.05047

[Kim et al. 2009] I.-J. Kim, D. D. Olesky, B. L. Shader, P. van den Driessche, H. van der Holst, and K. N. Vander Meulen, "Generating potentially nilpotent full sign patterns", *Electron. J. Linear Algebra* **18** (2009), 162–175. MR 2010f:15045 Zbl 1181.15039

[Logofet 1993] D. Logofet, *Matrices and graphs: stability problems in mathematical ecology*, CRC Press, 1993.

Received: 2011-08-04 Revised: 2012-04-25 Accepted: 2012-06-16

dolesky@cs.uvic.ca
Department of Computer Science, University of Victoria,
Victoria, British Columbia V8W 3P6, Canada

rempelm@uvic.ca
Department of Mathematics and Statistics, University
of Victoria, Victoria, British Columbia V8W 3R4, Canada

pvdd@math.uvic.ca
Department of Mathematics and Statistics, University
of Victoria, Victoria, British Columbia V8W 3R4, Canada
www.math.uvic.ca/faculty/pvdd/

mathematical sciences publishers msp

INVOLVE 6:1 (2013)
dx.doi.org/10.2140/involve.2013.6.13

The group of primitive almost pythagorean triples

Nikolai A. Krylov and Lindsay M. Kulzer

(Communicated by Scott Chapman)

We consider the triples of integer numbers that are solutions of the equation $x^2+qy^2=z^2$, where q is a fixed, square-free arbitrary positive integer. The set of equivalence classes of these triples forms an abelian group under the operation coming from complex multiplication. We investigate the algebraic structure of this group and describe all generators for each $q \in \{2, 3, 5, 6\}$. We also show that if the group has a generator with the third coordinate being a power of 2, such generator is unique up to multiplication by ± 1.

1. Introduction and the group of PPTs

The set of pythagorean triples has various interesting structures. One of such structures is induced by a binary operation introduced by Taussky [1970]. Recall that a pythagorean triple (PT from now on) is an ordered triple (a, b, c) of natural numbers satisfying the identity $a^2+b^2=c^2$, and given two such triples (a_1, b_1, c_1) and (a_2, b_2, c_2) we can produce another one using

$$A := a_1a_2 + b_1b_2, \quad B := |a_1b_2 - a_2b_1|, \quad C := c_1c_2. \tag{1}$$

The natural relation $(a, b, c) \simeq (\lambda a, \lambda b, \lambda c)$ for all $\lambda \in \mathbb{N}$, called projectivization, is an equivalence relation on this set. The operation mentioned above induces an abelian group structure on the set of equivalence classes of PTs where the identity element is the class of $(1, 0, 1)$. When a, b and c have no common prime divisors, the triple (a, b, c) is called *primitive*. It's easy to see that every equivalence class contains exactly one primitive pythagorean triple. Thus the set of all primitive pythagorean triples (PPTs from now on) forms an abelian group under the operation given in (1). The algebraic structure of this group, denoted by **P**, was investigated by Eckert [1984], who proved that the group of PPTs is a free abelian group generated by all primitive triples (a, b, c), where $a > b$ and c is a prime number of the linear form $c = 4n+1$. Every pythagorean triple (a, b, c) naturally gives a point on the unit circle with rational coordinates $(a/c, b/c)$ and the equivalence class of PTs

MSC2010: 11D09, 20K20.
Keywords: pythagorean triples, infinitely generated commutative groups.

corresponds to a unique point on the circle. Operation (1) on the pythagorean triples corresponds to the "angle addition" of rational points on S^1 and thus the group of PPTs is identified with the subgroup of all rational points on S^1. Analysis of this group was done by Tan [1996] and his Theorem 1 (p. 167) is equivalent to the proposition on page 25 of [Eckert 1984].

It is not hard to see that the composition law (1) naturally extends to the solutions of the Diophantine equation

$$X^2 + q \cdot Y^2 = Z^2 \tag{2}$$

where q is a fixed, square-free arbitrary positive integer. Via projectivization, we obtain a well defined binary operation on the set of equivalence classes of solutions to (2), and the set of such classes forms an abelian group as well. For some special values of q, including all $q \in \{2, 3, 5, 6, 7, 15\}$, such a group has been considered by Baldisserri [1999]. However, it seems that the generators $(3, 1, 4)$ for $q = 7$, and $(1, 1, 4)$ for $q = 15$, are missing in [Baldisserri 1999].

With the above in mind, we will consider in this paper the set of triples we call *almost pythagorean triples*, which are solutions to the equation (2). As in the case of PTs, each equivalence class here contains exactly one *primitive* almost pythagorean triple and therefore the set of equivalence classes is the set of *primitive almost pythagorean triples* (PAPTs).

In the next two sections we give a complete description of this group for $q \in \{2, 3, 5, 6\}$, similar to the one given in [Eckert 1984]. We also prove that for all $q \neq 3$ the group of PAPTs is free abelian of infinite rank. In the last section we will discuss solutions (a, b, c) where c is even. Please note that some of the results we prove here have been obtained earlier by Baldisserri; however, our proof of existence of elements of finite order is different from the one given in [Baldisserri 1999]. We also explain that if $(a, b, 2^k)$ is a nontrivial solution of (2) with $q \neq 3$, the set of all such solutions makes an infinite cyclic subgroup of the group of PAPTs. When $q = 7$ and $q = 15$ such a subgroup is missing in Theorem 2 of [Baldisserri 1999].

2. Group of PAPTs

Let T_q denote the set of all integer triples $(a, b, c) \in \mathbb{Z} \times \mathbb{Z} \times \mathbb{N}$ such that $a^2 + q \cdot b^2 = c^2$. We introduce the following relation on T_q: two triples (a, b, c) and (A, B, C) are equivalent if there exist $m, n \in \mathbb{Z} \setminus \{0\}$ such that $m(a, b, c) = n(A, B, C)$, where $m(a, b, c) = (ma, mb, |mc|)$. It is a straightforward check that this is an equivalence relation (also known as *projectivization*). We will denote the equivalence class of (a, b, c) by $[a, b, c]$. Note that $[a, b, c] = [-a, -b, c]$, but $[a, b, c] \neq [-a, b, c]$. We will denote the set of these equivalence classes by $\mathcal{P}_q$. Now we define a binary operation on $\mathcal{P}_q$ that generalizes the one on the set of PPTs defined by (1).

Definition 1. For two arbitrary classes $[a, b, c]$, $[A, B, C] \in \mathcal{P}_q$, define their sum by the formula

$$[a, b, c] + [A, B, C] := [aA - qbB, aB + bA, cC].$$

It is a routine check that this definition is independent of a particular choice of a triple and thus the binary operation is well defined. Here are two examples:

If $q = 7$,

$$[3, 1, 4] + [3, 1, 4] + [3, 1, 4] = [3, 1, 4] + [2, 6, 16] = [-36, 20, 64] = [-9, 5, 16].$$

If $q = 14$,

$$[5, 2, 9] + [13, 2, 15] = [9, 36, 135] = [1, 4, 15].$$

Since $[a, b, c] + [1, 0, 1] = [a, b, c]$ and $[a, b, c] + [-a, b, c] = [-a^2 - qb^2, 0, c^2] = [c^2, 0, c^2]$, and the operation is associative (this check is left for the reader), we obtain the following result (compare [Baldisserri 1999, Section 2] or [Weintraub 2008, Section 4.1]):

Theorem 1. *$(\mathcal{P}_q, +)$ is an abelian group. The identity element is $[1, 0, 1]$ and the inverse of $[a, b, c]$ is $[a, -b, c] = [-a, b, c]$.*

The purpose of this paper is to see what the algebraic structure of $(\mathcal{P}_q, +)$ is, and how it depends on q. From now on we will denote this group simply by $\mathcal{P}_q$. Please note that every equivalence class $[a, b, c] \in \mathcal{P}_q$ can be represented uniquely by a primitive triple $(\alpha, \beta, \gamma) \in T_q$, where $\alpha > 0$. In particular, this gives us freedom to refer to primitive triples to describe elements of the group.

Remark 1. The group $\mathcal{P}_q$ is a natural generalization of the group $\mathbf{P}$ of PPTs. However, $\mathcal{P}_1$ is not isomorphic to $\mathbf{P}$. The key point here is that the triple $(0, 1, 1) \notin T_q$, when $q > 1$, and the inverse of $[a, b, c]$ is $[a, -b, c] = [-a, b, c]$. In particular, it forces the consideration of triples with a and b being all integers and not only positive ones. As a result, the triples $(1, 0, 1)$ and $(0, 1, 1)$ are not equivalent in T_1. In order for the binary operation on the set of PPTs to be well defined, the triple $(0, 1, 1)$ must be equivalent to the identity triple $(1, 0, 1)$ [Eckert 1984, (5), p. 23]. The relation between our group $\mathcal{P}_1$ and the group $\mathbf{P}$ of PPTs is given by the following direct sum decomposition:

$$\mathcal{P}_1 \cong \mathbf{P} \oplus \mathbb{Z}/2\mathbb{Z},$$

where the 2-torsion subgroup $\mathbb{Z}/2\mathbb{Z}$ is generated by the element $[0, 1, 1]$. To prove this, one uses the map $f : \mathbf{P} \oplus \mathbb{Z}/2\mathbb{Z} \longrightarrow \mathcal{P}_1$ defined by

$$f\big((a, b, c), n\big) := \begin{cases} [a, b, c] + [1, 0, 1] = [a, b, c] & \text{if } n = 0, \\ [a, b, c] + [0, 1, 1] = [-b, a, c] & \text{if } n = 1. \end{cases}$$

It's easy to see that this f is an isomorphism.

Remark 2. The group $\mathcal{P}_q$ also has a geometric interpretation: consider the set $\mathcal{P}(\mathbb{Q})$ of all points $(X, Y) \in \mathbb{Q} \times \mathbb{Q}$ that belong to the conic $X^2 + qY^2 = 1$. Let $N = (1, 0)$ and take any two A, $B \in \mathcal{P}(\mathbb{Q})$. Draw the line through N parallel to the line (AB); then its second point of intersection with the conic $X^2 + qY^2 = 1$ will be $A + B$ (see [Lemmermeyer 2003b, Section 2.2; 2003a, Section 1] for the details). Via such geometric point of view, Lemmermeyer draws a close analogy between the groups $\mathcal{P}(\mathbb{Z})$ of integral points on the conics in the affine plane and the groups $E(\mathbb{Q})$ of rational points on elliptic curves in the projective plane. One of the key characteristics of $\mathcal{P}(\mathbb{Z})$ and $E(\mathbb{Q})$ is that both of the groups are finitely generated. Note however that if $q > 0$, the curve $X^2 + qY^2 = 1$ has only two integer points, $(\pm 1, 0)$. One could consider the solutions of $X^2 + qY^2 = 1$ over a finite field $\mathbb{F}_q$ or over the p-adic numbers $\mathbb{Z}_p$. In each of these cases the group of all solutions is also finitely generated and we refer the reader to [Lemmermeyer 2003a, Section 4.2] for the exact formulas. In the present paper we investigate the group structure of all rational points on the conic $X^2 + qY^2 = 1$ when $q \geq 2$ and such group is never finitely generated, as we explain below.

3. Algebraic structure of $\mathcal{P}_q$

The classical enumeration of primitive pythagorean triples in the form

$$(a, b, c) = (u^2 - v^2, 2uv, u^2 + v^2) \text{ or } \left(\frac{u^2 - v^2}{2}, uv, \frac{u^2 + v^2}{2}\right)$$

is a useful component in understanding the group structure on the set of PPTs. We assume here that integers u and v have no common prime divisors; otherwise (a, b, c) won't be primitive. One could use the Diophantus chord method (see, for example, [Stillwell 2003, Section 1.7]) to derive such enumeration of all PPTs. This method can be generalized to enumerate all solutions to (2) for all square-free $q > 1$. In particular, if a primitive triple $(a, b, c) \in T_q$, then there exists a pair (u, v) of integers with no common prime divisors, such that

$$(a, b, c) = \big(\pm(u^2 - qv^2), 2uv, u^2 + qv^2\big) \text{ or } \left(\pm\frac{u^2 - qv^2}{2}, uv, \frac{u^2 + qv^2}{2}\right).$$

We can use this enumeration right away to prove that if c is prime, and $(a,b,c) \in T_q$, then such a pair of integers (a, b) is essentially unique. Here is the precise statement.

Claim 1. *If c is prime and*

$$x^2 + qy^2 = c^2 = a^2 + qb^2, \quad \textit{where } abxy \neq 0,$$

then $(x, y) = (h_1 a, h_2 b)$, where $h_i = \pm 1$.

Proof. We apply Lemma 5.48 from Section 5.5 of [Weintraub 2008]. When $2c = u^2 + qv^2$, the proof needs an additional argument explaining why not just β/α_0 but also $\beta/(2\alpha_0)$ will be in the ring of integers. It can be easily done considering separate cases of even and odd q and using the fact that if q is odd, then u and v used in the enumeration are both odd, and if q is even, then u will be even and v will be odd. We leave details to the reader. □

We will use these results when we discuss generators of $\mathcal{P}_q$ below, but first we will find for which $q > 1$ the group $\mathcal{P}_q$ will have elements of finite order.

3.1. *Torsion in $\mathcal{P}_q$.*

We follow Eckert's geometric argument [1984, p. 24] to understand the torsion of $\mathcal{P}_q$.

Lemma 1. *If $q = 2$ or $q > 3$, then $\mathcal{P}_q$ is torsion-free. $\mathcal{P}_3 \cong \mathcal{F}_3 \oplus \mathbb{Z}/3\mathbb{Z}$, where $\mathcal{F}_3$ is a free abelian group.*

Proof. Let us assume that $q \geq 2$, and suppose the triple (a, b, c) is a solution of (2); that is, we can identify point $(a/c, \sqrt{q} \cdot b/c)$ with $e^{i\alpha}$ on the unit circle $\mathbf{U}$. Then a circle S^1_r with radius $r = \alpha/(2\pi)$ is made to roll inside $\mathbf{U}$ in the counterclockwise direction. The radius r is chosen this way so that the length of the circle S^1_r equals the length of the smaller arc of $\mathbf{U}$ between the points $e^{i\alpha}$ and $e^0 = (1, 0)$. Let us denote the point $(1, 0)$ by P and assume that this point moves inside the unit disk when S^1_r rolls inside $\mathbf{U}$. When $1 = kr$ for some positive integer k, this point P traces out a curve known as a hypocycloid. In this case the point P will mark off $k - 1$ distinct points on $\mathbf{U}$ and will return to its initial position $(1, 0)$ so the hypocycloid will have exactly k cusps. If P doesn't return to $(1, 0)$ after the first revolution around the origin, it might come back to $(1, 0)$ after, say, n such revolutions. In that case $n \cdot 2\pi = m \cdot \alpha$ for some $m \in \mathbb{N}$. Thus, α is a rational multiple of π, or, to be more precise,

$$\alpha = \pi \cdot \frac{2n}{m}.$$

Due to Corollary 3.12 of [Niven 1956, Chapter 3, Section 5], in such a case the only possible rational values of $\cos(\alpha)$ are $0, \pm 1/2, \pm 1$. Since $\cos(\alpha) = a/c$, where $a \neq 0$, we see that $\mathcal{P}_q$ might have a torsion only if $a/c = \pm 1/2$ or $a/c = \pm 1$. In the latter case we must have $q \cdot b^2 = 0$, which implies that the element $[a, b, c]$ is the identity of $\mathcal{P}_q$. Suppose now $a/c = \pm 1/2$. Then $qb^2 = 3a^2$ and if $3 \neq q$ we will have a prime $t \neq 3$ dividing q. We can assume without loss of generality that $\gcd(a, b) = 1$; hence we obtain $t \mid a$ and therefore $t^2 \mid qb^2$. Since q is square-free, we must have $t \mid b^2$, which contradicts that $\gcd(a, b) = 1$. Therefore if $q = 2$ or $q > 3$, $\mathcal{P}_q$ is torsion-free. Suppose now $q = 3$. Then we obtain $a = \pm b$ and we can multiply $[a, b, c]$ by -1, if needed, to conclude that $[a, b, c] = [1, 1, 2]$ or $[a, b, c] = [1, -1, 2]$. We have $\langle [a, b, c] \rangle \cong \mathbb{Z}/3\mathbb{Z}$ in both these cases. This implies that $\mathcal{P}_3/(\mathbb{Z}/3\mathbb{Z})$ is free abelian and hence $\mathcal{P}_3 \cong \mathcal{F}_3 \oplus \mathbb{Z}/3\mathbb{Z}$. □

Remark 3. There is a another way to obtain this lemma via a different approach to the group $\mathcal{P}_q$, $q > 0$. The authors are very thankful to Wladyslaw Narkiewicz who explained this alternative viewpoint to us (compare also with [Baldisserri 1999]). Consider an imaginary quadratic field $\mathbb{Q}(\sqrt{-q})$ and the multiplicative subgroup of nonzero elements whose norm is a square of a rational number. Let us denote this subgroup by $\mathcal{A}_q$. Obviously $\mathbb{Q}^* \subset \mathcal{A}_q$ ($\mathbb{Q}^*$ denotes the group of nonzero rational numbers). It is easy to see that $\mathcal{P}_q \cong \mathcal{A}_q/\mathbb{Q}^*$, and it follows from Theorem A of [Schenkman 1964] that $\mathcal{A}_q$ is a direct product of cyclic groups. Hence the same holds for $\mathcal{P}_q$. If $q = 1$ or $q = 3$ the group $\mathcal{A}_q$ will have elements of finite order since the field $\mathbb{Q}(\sqrt{-q})$ has units different from ± 1. These units will generate in $\mathcal{P}_q$ the torsion factors $\mathbb{Z}/2\mathbb{Z}$ or $\mathbb{Z}/3\mathbb{Z}$, when $q = 1$ or $q = 3$, respectively.

3.2. ***On generators of $\mathcal{P}_q$ when $q \leq 6$.*** In this subsection we assume that $2 \leq q \leq 6$ and will describe the generators of $\mathcal{P}_q$ similar to the way it was done in the proposition on pages 25 and 26 of [Eckert 1984]. We will use $\mathcal{F}_q$ to denote the free subgroup of $\mathcal{P}_q$. As follows from Section 3.1 above, $\mathcal{F}_q = \mathcal{P}_q$ for $q \neq 3$, and $\mathcal{P}_3 \cong \mathcal{F}_3 \oplus (\mathbb{Z}/3\mathbb{Z})$.

The key point in Eckert's description of the generators of the group of primitive pythagorean triples is the fact that a prime p can be a hypotenuse in a pythagorean triangle if and only if $p \equiv 1 \pmod 4$. Our next lemma generalizes this fact to the cases of primitive triples from T_q, with $q \in \{2, 3, 5, 6\}$.

Lemma 2. *If $(a, b, c) \in T_2$ is primitive and p is a prime divisor of c, then there exist $u, v \in \mathbb{Z}$ such that $p = u^2 + 2v^2$. If $(a, b, c) \in T_3$ is primitive and p is a prime divisor of c, then either $p = 2$ or there exist $u, v \in \mathbb{Z}$ such that $p = u^2 + 3v^2$. If $(a, b, c) \in T_q$ is primitive where $q = 5$ or $q = 6$, and p is a prime divisor of c, then there exist $u, v \in \mathbb{Z}$ such that $p = u^2 + qv^2$ or $2p = u^2 + qv^2$.*

Proof. Consider $(a, b, c) \in T_q$. Since $a^2 + qb^2 = c^2$ where $q \in \{2, 3, 5, 6\}$, it follows from the generalized Diophantus chord method that there exist $s, t \in \mathbb{Z}$ such that $c = s^2 + qt^2$ or $2c = s^2 + qt^2$. Suppose $c = p_1^{n_1} \cdots p_k^{n_k}$ is the prime decomposition of c.

<u>Case 1</u>: $q = 2$. We want to show that each prime p_i dividing c can be written in the form $p_i = u^2 + 2v^2$ for some $u, v \in \mathbb{Z}$ (note that if q is even, $p_i \neq 2$). It is well known that a prime p can be written in the form

$$p = u^2 + 2v^2 \iff p = 8n + 1 \text{ or } p = 8n + 3 \quad \text{for some integer } n$$

(see [Stillwell 2003, Chapter 9] or [Cox 1989, Chapter 1]). Thus it's enough to show that if a prime $p \mid c$ then $p = 8n + 1$ or $p = 8n + 3$. Since $p \mid c$, and $c = s^2 + qt^2$ or $2c = s^2 + qt^2$, we see that there exists $m \in \mathbb{Z}$ such that $pm = s^2 + 2t^2$ and hence $-2t^2 \equiv s^2 \pmod p$; that is, the Legendre symbol $\left(\frac{-2t^2}{p}\right)$ equals 1. Using basic properties of the Legendre symbol, this implies that $\left(\frac{-2}{p}\right) = 1$. But $\left(\frac{-2}{p}\right) = 1$ if and only if $p = 8n + 1$ or $p = 8n + 3$ as follows from the supplements to quadratic reciprocity law. This finishes the case with $q = 2$.

Case 2: Suppose now that $q = 3$. Then $(1, 1, 2) \in T_3$ gives an example when c is divisible by prime $p = 2$. Note also that prime $p = 2$ is of the form $2p = u^2 + 3v^2$. Assuming from now on that prime p dividing c is odd, we want to show that there exist u, $v \in \mathbb{Z}$ such that $p = u^2 + 3v^2$, which is true if and only if there exists $n \in \mathbb{Z}$ such that $p = 3n + 1$ (again, see [Stillwell 2003] or [Cox 1989]). Hence, in our case, it suffices to show that if $p \mid c$ then there exists $n \in \mathbb{Z}$ such that $p = 3n + 1$. As in Case 1, there exists $m \in \mathbb{Z}$ such that $pm = s^2 + 3t^2$ for some s, $t \in \mathbb{Z}$. Therefore, we have that the Legendre symbol $\left(\frac{-3}{p}\right) = 1$, which holds if and only if $p = 3n + 1$. One can prove this using the quadratic reciprocity law (e.g., [Stillwell 2003, Section 6.8]).

Case 3: Suppose now that $q = 5$. Note that in this case c must be odd. Indeed, if c were even, $x^2 + 5y^2$ would be divisible by 4, but on the other hand, since both of x and y must be odd when q is odd and c is even, we see that $x^2 + 5y^2 \not\equiv 0 \pmod 4$. Since $p \mid c$ then again there exists $m \in \mathbb{Z}$ such that $pm = s^2 + 5t^2$ for some s, $t \in \mathbb{Z}$; that is, $\left(\frac{-5}{p}\right) = 1$. It is true that for any integer n and odd prime p not dividing n the Legendre symbol $\left(\frac{-n}{p}\right) = 1$ if and only if p is represented by a primitive form $ax^2 + bxy + cy^2$ of discriminant $-4n$ such that a, b, and c are relatively prime [Cox 1989, Corollary 2.6]. Following an algorithm in Section 2.A of [Cox 1989] to show that every primitive quadratic form is equivalent to a reduced form one can show that the only two primitive reduced forms of discriminant $-4 \cdot 5 = -20$ are $x^2 + 5y^2$ and $2x^2 + 2xy + 3y^2$. Through a simple calculation it's easy to see that a prime p is of the form

$$p = 2x^2 + 2xy + 3y^2 \iff 2p = x^2 + 5y^2.$$

This finishes the third case.

Case 4: Lastly, let's consider the case when $q = 6$. Once again since $p \neq 2$ and $p \mid c$ then $\left(\frac{-6}{p}\right) = 1$. Using the same corollary used in Case 3, we see that p must be represented by a primitive quadratic form of discriminant $-4 \cdot 6 = -24$. Also, following the same algorithm used in Case 3 to determine such primitive reduced forms, we find that there are only two: $x^2 + 6y^2$ and $2x^2 + 3y^2$. Through a simple calculation it can be determined that a prime p is of the form

$$p = 2x^2 + 3y^2 \iff 2p = x^2 + 6y^2.$$

Thus, the lemma is proven. □

Remark 4. One could write prime divisors from this lemma in a linear form if needed. It is a famous problem of classical number theory which primes can be expressed in the form x^2+ny^2. The reader will find a complete solution of this problem in [Cox 1989]. For example, if p is prime, then for some $n \in \mathbb{Z}$ we have

$$p = \begin{cases} 20n+1, \\ 20n+3, \\ 20n+7, \\ 20n+9, \end{cases}$$

if and only if $p = x^2+5y^2$ or $p = 2x^2+2xy+3y^2$. We refer the reader for the details to [Cox 1989, Chapter 1].

Now we are ready to describe all generators of $\mathcal{P}_q$, where $q \in \{2, 3, 5, 6\}$. Our proof is similar to the proof given by Eckert [1984], where he decomposes the hypotenuse of a right triangle into the product of primes and after that peels off one prime at a time, together with the corresponding sides of the right triangle. His description of prime $p \equiv 1 \pmod 4$ is equivalent to the statement that p can be written in the form $p = u^2 + v^2$, for some integers u and v, which is the case of Fermat's two square theorem. In the theorem below we also use quadratic forms for the primes.

Theorem 2. *Let us fix $q \in \{2, 3, 5, 6\}$. Then $\mathcal{P}_q$ is generated by the set of all triples $(a, b, p) \in T_q$ where $a > 0$, and p is prime such that there exist $u, v \in \mathbb{Z}$ with $p = u^2+qv^2$, or $2p = u^2+qv^2$.*

Proof. Take arbitrary $[r, s, d] \in \mathcal{P}_q$ and let us assume that $(r, s, d) \in T_q$ will be the corresponding primitive triple with $r > 0$. Let $d = p_1^{n_1} \cdots p_k^{n_k}$ be the prime decomposition of d. It is clear from what we've said above that d will be odd when $[r, s, d] \in \mathcal{F}_q$, and d will be even only if $q = 3$ and $[r, s, d] \notin \mathcal{F}_3$. Our goal is to show that

$$[r, s, d] = \sum_{i=1}^{k} n_i \cdot [a_i, b_i, p_i],$$

where

$$a_i > 0, \quad n_i \cdot [a_i, b_i, p_i] := \underbrace{[a_i, b_i, p_i] + \cdots + [a_i, b_i, p_i]}_{n_i \text{ times}},$$

and p_i is either of the form u^2+qv^2 or of the form $(u^2+qv^2)/2$. We deduce from Lemma 2 that each prime $p_i \mid d$ can be written in one of these two forms. Hence, for all p_i, there exist $a_i, b_i \in \mathbb{Z}$ such that $a_i^2+qb_i^2 = p_i^2$. Indeed, if we have $2p = u^2+qv^2$, then

$$4p^2 = (u^2-qv^2)^2 + 4q(uv)^2$$

and since u^2+qv^2 is even, u^2-qv^2 will be even as well, and therefore we could write $\alpha^2+q\beta^2=p^2$, where $\alpha=(u^2-qv^2)/2$ and $\beta=uv$. Thus $[a_i, b_i, p_i]\in\mathcal{P}_q$. Since $\mathcal{P}_q$ is a group, the equations

$$[r,s,d]=\begin{cases}[X_1, Y_1, D_1]+[a_k, b_k, p_k],\\ [X_2, Y_2, D_2]+[-a_k, b_k, p_k]\end{cases}$$

always have a solution with $(X_i, Y_i, D_i)\in\mathbb{Z}\times\mathbb{Z}\times\mathbb{N}$. The key observation now is that only one of the triples (X_i, Y_i, D_i) will be equivalent to a primitive triple (x, y, d_1), with $d_1<d$. Indeed, we have $[r,s,d]=[X,Y,D]\pm[a,b,p]$ or

$$[X,Y,D]=[r,s,d]\pm[-a,b,p]=\begin{cases}[-ra-qsb, rb-sa, dp],\\ [ra-qsb, rb+sa, dp].\end{cases}$$

Since $p\mid d$, we have $dp\equiv 0 \pmod{p^2}$ and hence it is enough to show that either $ra+qsb\equiv rb-sa\equiv 0 \pmod{p^2}$, or $ra-qsb\equiv rb+sa\equiv 0 \pmod{p^2}$ (see lemma on page 24 of [Eckert 1984]). From the identity

$$(sa-rb)(sa+rb)=s^2a^2-r^2b^2=s^2(a^2+qb^2)-b^2(r^2+qs^2)\equiv 0 \pmod{p^2},$$

we deduce that either p divides each of $sa-rb$ and $sa+rb$, or p^2 divides exactly one of these two terms. In the first case $p\mid 2sa$, which is impossible if p is odd, since then either $a^2>p^2$ or (r,s,d) won't be primitive. If we assume $p=2$, then, as we explained in Lemma 2, $q=3$ and therefore $(a,b,p)=(1,1,2)$ so $(ra-qsb, rb+sa, dp)=(r-3s, r+s, 2d)$. But $r+s\equiv r-3s \pmod 4$ and if $4\mid r+s$ we can write $(ra-qsb, rb+sa, 2d)=4\big((r-3s)/4, (r+s)/4, d_1\big)$, where $d_1=d/2$. If $r+s\equiv 2 \pmod 4$, we will divide each element of the other triple by 4.

Thus we can assume from now on that p is an odd prime and that either $p^2\mid sa-rb$ or $p^2\mid sa+rb$. Let us assume without loss of generality that $sa-rb=kp^2$ for some $k\in\mathbb{Z}$. Since the triple $(-ra-qsb, rb-sa, dp)$ is a solution of (2), and the last two elements are divisible by p^2, it is obvious that the first element must be divisible by p^2 too, that is, that $ra+qsb=tp^2$. This implies that

$$[X,Y,D]=[-ra-qsb, rb-sa, dp]=[-t, -k, d_1],$$

where $d_1=d/p<d$, which we wanted to show. The other case is solved similarly. Note that only one of the two triples will have all three elements divisible by 4, which means that only $[a,b,p]$ or $[-a,b,p]$ can be subtracted from the original element $[r,s,d]$ in such a way that the result will be in the required form.

Thus we can "peel off" the triple $[a_k, b_k, p_k]$ from the original one $[r,s,d]$ ending up with the element $[x, y, d_1]$, where now $d_1<d$. Note that we can always assume that $a_k>0$ by using either $[a_k, b_k, p_k]$ or $[-a_k, -b_k, p_k]$. Then simply keep peeling off until all prime divisors of d give the required presentation of the element $[r, s, d]$ as a linear combination of the generators $[a_i, b_i, p_i]$. □

Remark 5. Since these primes are the generators of $\mathcal{P}_q$ when $q \in \{2, 3, 5, 6\}$ and each prime (with exception $p = 2$ when $q = 3$) generates an infinite cyclic subgroup, it is obvious that $\mathcal{P}_q$ contains an infinite number of elements. The same holds for $\mathcal{P}_q$ when $q \geq 7$. This can be shown through properties of Pell's equation $c^2 - qb^2 = 1$ where q is a square-free positive integer different from 1. This equation can be rewritten as $c^2 = 1^2 + qb^2$, which is in fact Equation (2) with specific solutions $(1, b, c)$. It is a classical fact of number theory that this equation always has a nontrivial solution and, in result, has infinitely many solutions (see [Weintraub 2008, Section 4.2] or [Stillwell 2003, Section 5.9]).

Note that it is not obvious that Pell's equation has a nontrivial solution for arbitrary q. For example, the smallest solution of the equation

$$1^2 + 61b^2 = c^2 \quad \text{is} \quad b = 226{,}153{,}980, \quad c = 1{,}766{,}319{,}049.$$

Let us observe that the equation $a^2 + 61b^2 = c^2$, where a is allowed to be any integer, has many solutions with "smaller" integer triples. Three examples are [3, 16, 125], [6, 7, 55], and [10, 9, 71].

3.3. ***On generators of $\mathcal{P}_q$ when $q \geq 7$ and the triples $(a, b, 2^k)$.*** It is interesting to see how the method of peeling off breaks down in specific cases of q for $q \geq 7$. Here are some examples of PAPTs $(a, b, c) \in T_q$, where c is divisible by a prime p but there exist no nontrivial pair $r, s \in \mathbb{Z}$ such that $(r, s, p) \in T_q$.

The primitive triple (9, 1, 10) $\in T_{19}$ is a solution, where 10 is divisible by primes 2 and 5; however, it is impossible to find nonzero $a, b \in \mathbb{Z}$ such that $a^2 + 19b^2 = 5^2$.

The primitive triple (3, 1, 4) $\in T_7$ is a solution, where 4 is divisible by prime 2, however, it is impossible to solve $a^2 + 7b^2 = 2^2$ in integers. In T_{15} the primitive triple (1, 1, 4) is a problematic solution for the same reason.

Baldisserri [1999, Observation 2, p. 304] mentions that if a nontrivial and primitive (a, b, c) solves (2), then c can be even only when $q \equiv 3 \pmod 4$. Moreover, if $q \equiv 3 \pmod 8$, we must have $c = 2 \cdot \text{odd}$, but if $q \equiv 7 \pmod 8$ we can have c divisible by any power of 2. Indeed, as we just mentioned above, the triple (3, 1, 4) solves (2) with $q = 7$, and clearly can not be presented as a sum of two "smaller" triples. Since $\mathcal{P}_7$ is free, we see that (3, 1, 4) must generate a copy of $\mathbb{Z}$ inside $\mathcal{P}_7$, and one can easily check that we have

$$2\cdot[3,1,4] = \pm[1,3,2^3], \quad 3\cdot[3,1,4] = \pm[9,5,2^4], \quad 4\cdot[3,1,4] = \pm[31,3,2^5], \quad \ldots.$$

The same holds for the triple (1, 1, 4) $\in T_{15}$ but somehow these two generators of $\mathcal{P}_7$ and $\mathcal{P}_{15}$ are not mentioned in Theorem 2 of [Baldisserri 1999].

Can we have more than one such generator for a fixed q? In other words, how many nonintersecting $\mathbb{Z}$-subgroups of $\mathcal{P}_q$ can exist, provided that each subgroup is generated by a triple where c is a power of 2? The following theorem shows that there can be only one such generator (for the definition of *irreducible solution* we refer the reader to [Baldisserri 1999, p. 304], but basically it means that this solution is a generator of the group of PAPTs).

Theorem 3. *Fix q as above and assume that the triple $(a, b, 2^k)$ is an irreducible solution of* (2). *If $(x, y, 2^r) \in T_q$ and $r \geq k$, then there exists $n \in \mathbb{Z}$ such that*

$$[x, y, 2^r] = n \cdot [a, b, 2^k].$$

Proof. Our idea of the proof is to show that given such a triple $(x, y, 2^r) \in T_q$ with $r \geq k$, we can always peel off (i.e., add or subtract) one copy of $(a, b, 2^k)$ so the resulting primitive triple will have the third coordinate less than or equal to 2^{r-1}. Thus we consider

$$[S, T, V] := [x, y, 2^r] \pm [a, b, 2^k] = \begin{cases} [xa - qyb, ay + xb, 2^{r+k}], \\ [xa + qyb, ay - xb, 2^{r+k}]. \end{cases}$$

Since a, b, x and y are all odd, either $ay + xb$ or $ay - xb$ must be divisible by 4. Let's assume that $4 \mid ay - xb$ and hence we can write $ay - xb = 2^d \cdot R$, where $d \geq 2$. Clearly, it's enough to prove that $d \geq k + 1$. We prove it by induction; that is, we will show that if $d \leq k$, then R must be even.

Since $S = xa + qyb$ we could write

$$\begin{pmatrix} 2^d \cdot R \\ S \end{pmatrix} = \begin{pmatrix} -b & a \\ a & qb \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} \quad \text{and hence} \quad \begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{2^{2k}} \cdot \begin{pmatrix} qb & -a \\ -a & -b \end{pmatrix} \cdot \begin{pmatrix} 2^d \cdot R \\ S \end{pmatrix},$$

which gives $bS = -2^{2k} y - a 2^d R$. Since $(bS, bT, bV) \in T_q$, we can also write

$$(2^{2k} y + a 2^d R)^2 + qb^2 \cdot (2^d R)^2 = b^2 \cdot 2^{2r+2k}.$$

This last identity is equivalent to the following one (after using $a^2 + qb^2 = 2^{2k}$ and dividing all terms by 2^{2k}):

$$2^{2k} y^2 + 2^{d+1} ayR + 2^{2d} R^2 = b^2 2^{2r}.$$

Furthermore, we can cancel 2^{d+1} as well, because $1 < d \leq k \leq r$, and then we will obtain that

$$ayR = b^2 2^{2r-d-1} - 2^{d-1} R^2 - 2^{2k-d-1} y^2 = \text{even},$$

which finishes the proof since a and y are odd. □

Remark 6. Please note that if a primitive triple $(a, b, 2 \cdot d) \in T_q$ for $q \equiv 7 \pmod 8$, it is easy to show that d must be even (compare with Observation 2 of [Baldisserri 1999], where λ must be at least 2). When $q \in \{7, 15\}$, we obtain the generators $(3, 1, 4)$ and $(1, 1, 4)$, respectively. However, if, for example $q = 23$, the primitive solution (a, b, c) where c is the smallest power of 2 is $(7, 3, 16)$ but $(11, 1, 12)$ also belongs to $\mathcal{P}_{23}$.

References

[Baldisserri 1999] N. Baldisserri, "The group of primitive quasi-Pythagorean triples", *Rend. Circ. Mat. Palermo* (2) **48**:2 (1999), 299–308. MR 2000g:11025 Zbl 0938.11014

[Cox 1989] D. A. Cox, *Primes of the form $x^2 + ny^2$: Fermat, class field theory and complex multiplication*, John Wiley & Sons, New York, 1989. MR 90m:11016 Zbl 0701.11001

[Eckert 1984] E. J. Eckert, "The group of primitive Pythagorean triangles", *Math. Mag.* **57**:1 (1984), 22–27. MR 85a:51017 Zbl 0534.10010

[Lemmermeyer 2003a] F. Lemmermeyer, "Conics: a poor man's elliptic curves", preprint, 2003. arXiv 0311306

[Lemmermeyer 2003b] F. Lemmermeyer, "Higher descent on Pell conics, III: The first 2-descent", preprint, 2003. arXiv 0311310

[Niven 1956] I. Niven, *Irrational numbers*, Carus Mathematical Monographs **11**, Mathematical Association of America, 1956. MR 18,195c Zbl 0070.27101

[Schenkman 1964] E. Schenkman, "On the multiplicative group of a field", *Arch. Math.* (*Basel*) **15** (1964), 282–285. MR 30 #89 Zbl 0126.06801

[Stillwell 2003] J. Stillwell, *Elements of number theory*, Springer, New York, 2003. MR 2004j:11001 Zbl 1112.11002

[Tan 1996] L. Tan, "The group of rational points on the unit circle", *Math. Mag.* **69**:3 (1996), 163–171. MR 1394795 Zbl 1044.11584

[Taussky 1970] O. Taussky, "Sums of squares", *Amer. Math. Monthly* **77** (1970), 805–830. MR 42 #3020 Zbl 0208.05202

[Weintraub 2008] S. H. Weintraub, *Factorization: unique and otherwise*, Canadian Mathematical Society, Ottawa, ON, 2008. MR 2009b:11193 Zbl 1162.11003

Received: 2011-08-11 Revised: 2012-03-29 Accepted: 2012-04-29

nkrylov@siena.edu
*Department of Mathematics, Siena College,
515 Loudon Road, Loudonville, NY 12211, United States*

lindsaykulzer@gmail.com
*Department of Mathematics, Siena College,
515 Loudon Road, Loudonville, NY 12211, United States*

INVOLVE 6:1 (2013)
dx.doi.org/10.2140/involve.2013.6.25

Properties of generalized derangement graphs

Hannah Jackson, Kathryn Nyman and Les Reid

(Communicated by Ann Trenk)

A permutation on n elements is called a *k-derangement* ($k \leq n$) if no k-element subset is mapped to itself. One can form the *k-derangement graph* on the set of all permutations on n elements by connecting two permutations σ and τ if $\sigma\tau^{-1}$ is a k-derangement. We characterize when such a graph is connected or Eulerian. For n an odd prime power, we determine the independence, clique and chromatic numbers of the 2-derangement graph.

1. Introduction

Permutations which leave no element fixed, known as derangements, were first considered in [de Montmort 1708] and have been extensively studied since. A derangement graph is a graph whose vertices are the elements of the symmetric group S_n and whose edges connect two permutations that differ by a derangement. Derangement graphs have been shown to be connected (for $n > 3$) and Hamiltonian, and their independence number, clique number, and chromatic number have been calculated [Renteln 2007].

Here we consider the generalization of derangements known as k-derangements, which are those permutations in S_n that do not fix any k-element subset of the set being permuted. A k-derangement graph is defined in an analogous manner to a derangement graph. We examine some of the graph-theoretical properties of k-derangement graphs.

2. Preliminaries

Let S_n be the group of permutations on the set $\{1, 2, \ldots, n\}$. A permutation $\sigma \in S_n$ maps any k-element subsetof $\{1, \ldots, n\}$ to a k-element subset of $\{1, \ldots, n\}$; in the usual notation,

$$\sigma(\{a_1, \ldots, a_k\}) = \{\sigma(a_1), \ldots, \sigma(a_k)\}.$$

If $\{a_1, \ldots, a_k\} = \{\sigma(a_1), \ldots, \sigma(a_k)\}$ (as sets, that is, without regard to order), we

MSC2010: primary 05C69, 05A05; secondary 05C45.

Keywords: derangements, Eulerian, chromatic number, maximal clique, Cayley graph, independent set.

say that σ fixes the unordered k-tuple $\{a_1, \dots, a_k\}$. ("Unordered k-tuple" is another name for a k-element set.)

If σ does not map *any* of the $\binom{n}{k}$ possible unordered k-tuples to itself, we say that σ is a *k-derangement*. For example, with $n = 4$, the cyclic permutation $\sigma = (1234)$ is a 2-derangement, because (taking $k = 2$) we have

$$\begin{aligned}
(1234)(\{1,2\}) &= \{(1234)(1), (1234)(2)\} = \{2,3\},\\
(1234)(\{1,3\}) &= \{(1234)(1), (1234)(3)\} = \{2,4\},\\
(1234)(\{1,4\}) &= \{(1234)(1), (1234)(4)\} = \{2,1\} = \{1,2\},\\
(1234)(\{2,3\}) &= \{(1234)(2), (1234)(3)\} = \{3,4\},\\
(1234)(\{2,4\}) &= \{(1234)(2), (1234)(4)\} = \{3,1\} = \{1,3\},\\
(1234)(\{3,4\}) &= \{(1234)(3), (1234)(4)\} = \{4,1\} = \{1,4\}.
\end{aligned}$$

This extends the ordinary notion of a derangement, defined as a permutation $\sigma \in S_n$ such that $\sigma(x) \neq x$ for all $x \in \{1, \dots, n\}$.

The set of k-derangements in S_n is denoted by $\mathcal{D}_{k,n}$, and its cardinality $|\mathcal{D}_{k,n}|$ — the number of k-derangements in S_n — is denoted by $D_k(n)$. As we have seen, (1234) is in $\mathcal{D}_{2,4}$. Specifically,

$$\begin{aligned}
\mathcal{D}_{2,4} = \big\{&(1234), (1243), (1324), (1342), (1423), (1432), (123)(4), (124)(3),\\
&(132)(4), (134)(2), (142)(3), (143)(2), (234)(1), (243)(1)\big\},
\end{aligned}$$

and thus $D_2(4) = 14$. The sequence $D_2(n)$ appears as A137482 in the *On-Line Encyclopedia of Integer Sequences*; see [Henshaw 2008]. The number $D_1(n)$ is also simply called the derangement number.

The cycle structure of a permutation σ, denoted by C_σ, is the multiset of the lengths of the cycles in its cycle decomposition (e.g., $C_{(12)(3)(45)} = \{2, 2, 1\}$). Note that the cycle structure of $\sigma \in S_n$ is a partition of n. Given a partition r of n, let P_r be the set of all permutations in S_n whose cycle structure is r. For example (as usual, excluding singletons in our notation) $P_{\{2,1,1\}} = \{(12), (13), (14), (23), (24), (34)\}$.

We first note that if the cycle structure of a permutation σ contains a multiset which partitions k, then σ is not a k-derangement. For example, (12)(34) is a 3-derangement in S_4, but (12)(3)(4) is not, because it fixes the set $\{1, 2, 3\}$, for example. And we see that $\{2, 1\} \subseteq C_{(12)(3)(4)} = \{2, 1, 1\}$ is a partition of 3. Thus we observe that the cycle structure of a permutation determines whether or not it is a k-derangement, and we have the following.

Proposition 1. *A permutation $\sigma \in S_n$ is a k-derangement if and only if the cycle decomposition of σ does not contain a set of cycles whose lengths partition k.*

Proof. If $\{q, r, \dots, s\}$ is a partition of k, and $(a_1 \cdots a_q)(b_1 \cdots b_r) \cdots (c_1 \cdots c_s)$ are cycles of σ, then, for $x = \{a_1, \dots, a_q, b_1, \dots, b_r, c_1, \dots, c_s\}$, $\sigma(x) = x$. Conversely,

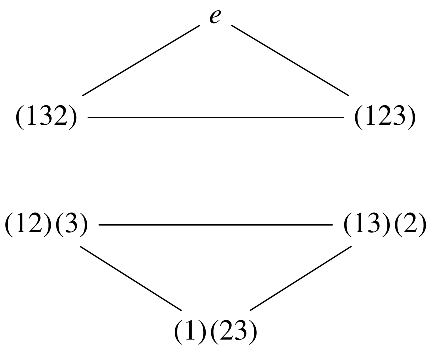


Figure 1. The 2-derangement graph on 6 vertices, $\Gamma_{2,3}$.

if σ has no set of cycles whose lengths partition k, then, given any k-element subset x of $\{1, \ldots, n\}$, there is a cycle in σ which contains at least one element in x and contains some element not in x. Hence σ sends an element in x to an element not in x and so $\sigma(x) \neq x$. □

Let $CD_{k,n}$ be the set of cycle structures corresponding to k-derangements in S_n; for example, $CD_{2,4} = \{\{4\}, \{3, 1\}\}$. Since a cycle structure C_σ is in $CD_{k,n}$ if and only if it is in $CD_{n-k,n}$, we have $\mathfrak{D}_{k,n} = \mathfrak{D}_{n-k,n}$.

Let G be a group, and let S be a subset of G that is closed under taking inverses. The *Cayley graph* $\Gamma(G, S)$ is the graph whose vertices are the elements of G such that an edge connects two vertices $u, v \in G$ if $su = v$ for some $s \in S$. A *k-derangement graph* is a Cayley graph defined by $\Gamma_{k,n} := \Gamma(S_n, \mathfrak{D}_{k,n})$. (Note that $\mathfrak{D}_{k,n}$ is symmetric, as the inverse of a k-derangement is a k-derangement, and thus satisfies the requirements for a Cayley graph.) It is worth noting that $\Gamma_{k,n}$ is, by construction, $D_k(n)$-regular, and that, since $\mathfrak{D}_{k,n} = \mathfrak{D}_{(n-k),n}$, $\Gamma_{k,n} = \Gamma_{(n-k),n}$. Figure 1 illustrates the 2-derangement graph on 6 vertices, $\Gamma_{2,3}$.

It is possible to consider k-derangements in S_n for any positive k and n. However, if $k = n$, there will be no k-derangements in S_n, since every partition in S_n will have a cycle structure such that the cycle lengths partition k. As such, $\Gamma_{k,n}$ will be the empty (edgeless) graph on n vertices. If $k > n$, then every permutation in S_n is a k-derangement vacuously, and thus $\Gamma_{k,n}$ will be the complete graph on $|S_n|$ vertices. As neither of these cases is particularly interesting, henceforth we will only consider k-derangements where $k < n$.

3. Properties of derangement graphs

Figure 1 shows that $\Gamma_{2,3}$ is not a connected graph, and, since $\Gamma_{2,3} = \Gamma_{1,3}$, we see that $\Gamma_{k,3}$ is disconnected for all $k < n$. But this is an exception rather than the rule, as the following theorem demonstrates.

Theorem 2. *For $n > 3$ and $k < n$, $\Gamma_{k,n}$ is connected.*

Proof. Every permutation in S_n can be written as the product of adjacent transpositions $(h\ (h{+}1))$. These, in turn, can be expressed as products of two k-derangements, so long as $n > 3$, as we will demonstrate. As a result, for $n > 3$, the elements of $\mathscr{D}_{k,n}$ generate S_n, which means that every vertex of $\Gamma_{k,n}$ can be reached by a path from the identity.

We show that the permutation $(1\ 2)$ can be written as the product of two k-derangements and then note that, since it is the form and not the individual labels that are important, any adjacent transposition can be written as the product of two k-derangements. We consider two cases: $k = 1$ and $k \geq 2$.

Case 1: If $k = 1$, then $(1\,2) = (1\,2\,\cdots\,n)^2 \cdot (n\ (n{-}1)\,\cdots\,1)^2(1\,2)$. We claim that $(1\,2\,\cdots\,n)^2$ and $(n\ (n{-}1)\,\cdots\,1)^2(1\,2)$ are each 1-derangements in S_n for all $n > 3$. If n is even, then $(1\,2\,\cdots\,n)^2 = (1\,3\,\cdots\,(n{-}3)\,(n{-}1))(2\,4\,\cdots\,(n{-}2)\,n)$, which is a 1-derangement in S_n for all n. Additionally,

$$(n\ (n{-}1)\ \cdots\ 1)^2(1\,2) = (1\ n\ (n{-}2)\ (n{-}4)\ \cdots\ 2\ (n{-}1)\ (n{-}3)\ \cdots\ 3),$$

which is also a 1-derangement in S_n for any n.

On the other hand, if n is odd, then

$$(1\,2\ \cdots\ n)^2 = (1\,3\ \cdots\ (n{-}2)\ n\ 2\,4\ \cdots\ (n{-}3)\ (n{-}1)),$$

which is a 1-derangement in S_n for all n. And

$$\begin{aligned}(n\ (n{-}1)\ \cdots\ 1)^2(1\,2) &= (n\ (n{-}2)\ (n{-}4)\ \cdots\ 3\ 1\ (n{-}1)\ (n{-}3)\ \cdots\ 4\ 2)(1\,2)\\ &= (1\ n\ (n{-}2)\ (n{-}4)\ \cdots\ 3)(2\ (n{-}1)\ (n{-}3)\ \cdots\ 4),\end{aligned}$$

which is a 1-derangement in S_n so long as $n > 3$. (If $n = 3$, $(312)(12) = (13)(2)$, which is not a 1-derangement.)

Thus, for $n > 3$, we have shown that $(1\ 2)$ can be written as the product of two 1-derangements, and, by extension, every adjacent transposition can be written as the product of two 1-derangements.

Case 2: For $k \geq 2$, $(1\,2) = (1\,2\,\cdots\,n)^{-1}(1\,3\,4\,\cdots\,n)$. We know $(1\,2\,\cdots\,n)^{-1}$ is a k-derangement for all k since the inverse of a k-derangement is a k-derangement. And, by the cycle structure, we see that $(1\ 3\ 4\ \cdots\ n) = (1\ 3\ 4\ \cdots\ n)(2)$ is a k-derangement for all k, except $k = 1$ and $k = (n{-}1)$ (however, since $\Gamma_{1,n} = \Gamma_{(n-1),n}$, Case 1 addresses $(n{-}1)$-derangements as well as 1-derangements).

So we have shown that, for $k \geq 2$, $(1\ 2)$ can be written as the product of two k-derangements, and again, by extension, we can write any adjacent transposition as the product of two k-derangements. Thus every vertex is connected by a path to the identity, and $\Gamma_{k,n}$ is connected. □

It is worth noting that Theorem 2 holds for $n = 2$ as well. Since we are only interested in k-derangements in S_n such that $k < n$, when $n = 2$, k must equal 1, and so $\Gamma_{1,2}$ is the connected graph on two vertices.

Next, we give a characterization in terms of n and k for when a derangement graph is Eulerian. We will require the following result.

Lemma 3. *If a cycle structure includes a cycle of length greater than* 2, *then there are an even number of permutations with that cycle structure.*

Proof. Consider P_r, the set of permutations with a given cycle structure, r. We can pair each $\sigma \in P_r$ with its inverse $\sigma^{-1} \in P_r$, and, so long as $\sigma \neq \sigma^{-1}$ for any $\sigma \in P_r$, $|P_r|$ will be even. Suppose there exists a $\sigma \in P_r$ such that $\sigma = \sigma^{-1}$. Then $\sigma^2 = e$, and so the order of σ is at most 2. The order of a permutation is the least common multiple of the orders of the elements of its cycle structure, so σ must not include a cycle of length greater than 2. This is a contradiction; thus $|P_r|$ is even. □

Theorem 4. *For $n > 3$ and $k < n$, $\Gamma_{k,n}$ is Eulerian if and only if k is even or k and n are both odd.*

Proof. A graph is Eulerian if and only if it is connected and each vertex has an even degree. In light of Theorem 2 and the previously noted fact that $\Gamma_{k,n}$ is $D_k(n)$-regular, in order to ascertain if $\Gamma_{k,n}$ is Eulerian, we must determine whether $D_k(n)$ is even or odd.

If k is even, we claim that $D_k(n)$ is the sum of even numbers. Any cycle structure composed entirely of 2- or 1-cycles will partition an even k, and thus any permutation which is in $\mathcal{D}_{k,n}$ for an even k will contain a cycle of length 3 or greater in its cycle decomposition. Now, $\mathcal{D}_{k,n} = P_{r_1} \dot{\cup} P_{r_2} \dot{\cup} \cdots \dot{\cup} P_{r_m}$ (disjoint union) such that no r_i partitions k, and, by Lemma 3, $|P_{r_i}|$ is even for all $i \in \{1, \ldots, m\}$. Thus, when k is even, $D_k(n)$ is even.

If k and n are both odd, again we see that every permutation in $\mathcal{D}_{k,n}$ will contain a cycle of length 3 or greater in its cycle decomposition, since an odd k can be partitioned by a set of cycles of lengths 1 or 2 if there is at least one 1-cycle. Furthermore, since n is odd, there are no permutations whose cycle structure is composed only of length-2 cycles. Thus, $D_k(n)$ is even.

Finally, we show that, if k is odd and n is even, then $\Gamma_{k,n}$ is not Eulerian. In this case, $P_{\{2,2,\ldots,2\}}$ is in $CD_{k,n}$. By choosing pairs of elements for the cycles and dividing by the number of ways to order the cycles, we see that the number of permutations in $P_{\{2,2,\ldots,2\}}$ is given by

$$\begin{aligned}\frac{\binom{n}{2}\binom{n-2}{2}\cdots\binom{2}{2}}{\left(\frac{n}{2}\right)!} &= \frac{n(n-1)(n-2)\cdots(3)(2)(1)}{\left(2\cdot\frac{n}{2}\right)\left(2\cdot\left(\frac{n}{2}-1\right)\right)\cdots(6)(4)(2)} \\ &= \frac{n(n-1)(n-2)\cdots(3)(2)(1)}{n(n-2)\cdots(6)(4)(2)} = (n-1)(n-3)\cdots(5)(3)(1).\end{aligned}$$

Since n is even, the product $(n-1)(n-3)\cdots(5)(3)(1)$ is odd. Every other k-derangement in S_n will contain a cycle with length greater than 2, since any combination of 1-cycles or 1- and 2-cycles will partition k. So $D_k(n)$ is the sum of one odd number and even numbers, and so is odd. □

4. Chromatic, independence and clique numbers for $k = 2$ and n an odd prime power

For the majority of this section, we will think of permutations in terms of the result of their application to the ordering $\{1, 2, 3, \dots, n\}$. Thus, $\{2, 3, 1, 4, 5\}$ represents the permutation which has moved 2 to the first position, 3 to the second, 1 to the third, and left 4 and 5 fixed; that is, the permutation $(132)(4)(5)$ in cycle notation, or the inverse of the permutation $\binom{12345}{23145}$ in two line notation.

We note that in order for vu^{-1} (or, equivalently, $v^{-1}u$) to be a k-derangement, it is necessary and sufficient that no unordered k-tuple of elements be sent to the same unordered k-tuple of positions by both u and v. For example, the permutations $u = \{2, 3, 1, 4, 5\}$ and $v = \{4, 1, 3, 5, 2\}$ both send the pair $\{1, 3\}$ to the second and third positions. Thus $(vu^{-1})(\{2, 3\}) = \{2, 3\}$, and so vu^{-1} is not a 2-derangement and there is no edge between u and v in the 2-derangement graph. More formally, suppose u and v both send the k-tuple $M' = \{a'_1, a'_2, \dots, a'_k\}$ to positions $M = \{a_1, a_2, \dots, a_k\}$. Then, $(vu^{-1})(M) = v(M') = M$. Thus, vu^{-1} is not a k-derangement.

On the other hand, if u and v send no k-tuple to the same positions we claim vu^{-1} is a k-derangement. Consider an arbitrary k-tuple, $M = \{a_1, a_2, \dots, a_k\}$, and suppose u maps the k-tuple $M' = \{a'_1, a'_2, \dots, a'_k\}$ to the positions given in M. Then $(vu^{-1})(M) = v(M') \neq M$ since v cannot send the k-tuple M' to the same positions as u does. Thus, vu^{-1} is a k-derangement.

In Theorem 6, we find the clique number of the 2-derangement graph, $\omega(\Gamma_{2,n})$, for n an odd prime power, by constructing a clique of maximal size. Before establishing this clique number, we note an upper bound on the clique number of a general k-derangement graph.

Lemma 5. *For $k < n$, $\omega(\Gamma_{k,n}) \leq \binom{n}{k}$.*

Proof. The clique number of the k-derangement graph, $\omega(\Gamma_{k,n})$, cannot be greater than $\binom{n}{k}$, since there are only $\binom{n}{k}$ subsets of size k and hence at most $\binom{n}{k}$ different unordered k-tuples of positions for an arbitrary k-tuple of elements to be sent under a permutation. □

Theorem 6. *If n is an odd prime power, then $\omega(\Gamma_{2,n}) = \binom{n}{2}$.*

Proof. We will explicitly construct a clique with $\binom{n}{2}$ elements. Let $n = p^r$, with p a prime greater than 2, and let $\mathbb{F}_{p^r}$ denote the field with p^r elements. Rather than letting S_n act on $\{1, \ldots, n\}$, we will let it act on $\mathbb{F}_{p^r}$ and construct $\Gamma_{2,n}$ accordingly. Let $v = (x_1, \ldots, x_n)$ be an ordered n-tuple whose entries are the elements of $\mathbb{F}_{p^r}$ in some order. Given any function $\phi : \mathbb{F}_{p^r} \to \mathbb{F}_{p^r}$, we define $\phi(v) = (\phi(x_1), \ldots, \phi(x_n))$. Partition the nonzero elements of $\mathbb{F}_{p^r}$ by pairing each element with its (additive) inverse, and let T be a set obtained by choosing exactly one element from each pair, giving $|T| = (p^r - 1)/2$.

Define $f_{s,\alpha}(x) = sx + \alpha$, and consider the set $X = \{f_{s,\alpha}(v) \mid s \in T \text{ and } \alpha \in \mathbb{F}_{p^r}\}$. Since $s \neq 0$, $f_{s,\alpha}$ is a bijection and $f_{s,\alpha}(v)$ is a permutation of the elements of $\mathbb{F}_{p^r}$. We claim that X is a clique in $\Gamma_{2,n}$. Suppose not; that is, suppose there are $s, t \in T$ and $\alpha, \beta \in \mathbb{F}_{p^r}$, $(s, \alpha) \neq (t, \beta)$, such that $f_{s,\alpha}(v)$ is not a 2-derangement of $f_{s,\beta}(v)$. In that case there exist $x, y \in \mathbb{F}_{p^r}$, $x \neq y$, such that either $f_{s,\alpha}(x) = f_{t,\beta}(x)$ and $f_{s,\alpha}(y) = f_{t,\beta}(y)$ or $f_{s,\alpha}(x) = f_{t,\beta}(y)$ and $f_{s,\alpha}(y) = f_{t,\beta}(x)$. In the first case, subtracting the two equations and rewriting yields $(s-t)(x-y) = 0$. If $s = t$, then $\alpha = \beta$, giving a contradiction. If $s \neq t$, then $x = y$ and again we have a contradiction. In the second case, subtracting and rewriting yields $(s+t)(x-y) = 0$ and, since $s + t \neq 0$ for $s, t \in T$, $x = y$ and this also give a contradiction. Thus, X is a clique of size $p^r(p^r - 1)/2 = \binom{n}{2}$. □

The next example illustrates the construction when $n = 7$.

Example 7. We build a clique of size $\binom{7}{2}$ in the derangement graph $\Gamma_{2,7}$ consisting of $\frac{7-1}{2}$ blocks, each of which contains 7 permutations. We let $v = (1, 2, 3, 4, 5, 6, 7)$ (writing 7 instead of 0) and take $T = \{1, 4, 5\}$. Then

$$f_{1,0}(v) = (1, 2, 3, 4, 5, 6, 7), \quad f_{4,0}(v) = (4, 1, 5, 2, 6, 3, 7),$$
$$f_{5,0}(v) = (5, 3, 1, 6, 4, 2, 7).$$

Increasing α from 0 cyclically permutes the 7-tuples. Block 1 consists of the arrangements $\{f_{1,\alpha}(v) \mid \alpha \in \mathbb{F}_7\}$, that is, the arrangement $(1, 2, 3, 4, 5, 6, 7)$ and the remaining 6 rotations of this arrangement (e.g., $(2, 3, 4, 5, 6, 7, 1)$, $(3, 4, 5, 6, 7, 1, 2)$, etc.). Block 2 consists of the arrangement $f_{4,0}(v)$ along with all of its rotations. Finally, block 3 consists of $f_{5,0}(v)$ and its rotations. To see that these permutations form a clique, consider, for example, the pair $\{1, 2\}$. These elements are one position apart in block 1, two positions apart in block 2 and three positions apart in block 3 (counting the shortest distance between them either forwards or backwards). So the pair $\{1, 2\}$ cannot occupy the same positions in two permutations which appear in different blocks. Furthermore, within a block, the rotations insure that the pair never occupies the same positions.

Remark 8. Cliques achieving the upper bound of Lemma 5 are known as *sharply k-homogeneous sets* of permutations. A corollary in [Nomura 1985] shows that, for $2k \leq n$, the existence of such a k-homogeneous set implies $n + 1 \equiv 0 \mod k$. Thus Theorem 6 cannot be extended to even n, and we have the following.

Corollary 9. *For n even and $n \geq 4$, $\omega(\Gamma_{2,n}) < \binom{n}{2}$.*

A computer search confirms that $\omega(\Gamma_{2,4}) = 5 < \binom{4}{2}$.

Next we turn to the independence number $\alpha(\Gamma_{k,n})$ and the chromatic number $\chi(\Gamma_{k,n})$ of the k-derangement graph. We will require the following lemma which has been adapted from Frankl and Deza's lemma [1977] and applied to k-tuples of elements.

Lemma 10. *For $k < n$, $\alpha(\Gamma_{k,n})\omega(\Gamma_{k,n}) \leq n!$.*

Proof. Let $\mathcal{P}$ be a set of permutations in S_n, every pair of which has at least one unordered k-tuple of elements in the same unordered k-tuple of positions. That is, for any u, $v \in \mathcal{P}$, there exists a set $M = \{a_1, \dots, a_k\} \subseteq \{1, \dots, n\}$ such that $(v^{-1}u)(M) = M$. Note that $\mathcal{P}$ is an independent set in the k-derangement graph. Let $\mathcal{Q}$ be a set of permutations in S_n such that each pair of permutations has no k-tuple of elements in the same positions; that is, $\mathcal{Q}$ is a clique in the k-derangement graph. We claim that products of the form PQ with $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$ give distinct permutations of n. Suppose, for the sake of contradiction, that $P_1Q_1 = P_2Q_2$ for P_1, $P_2 \in \mathcal{P}$ and Q_1, $Q_2 \in \mathcal{Q}$ with $P_1 \neq P_2$ and $Q_1 \neq Q_2$. This implies that $P_1^{-1}P_2 = Q_1Q_2^{-1}$. Now, since P_1 and P_2 are in $\mathcal{P}$, there is a k-tuple of elements $M = \{a_1, \dots, a_k\}$ such that $(P_1^{-1}P_2)(M) = M$. However, this implies $(Q_1Q_2^{-1})(M) = M$. But we know that the permutations in $\mathcal{Q}$ agree on no k-tuples, and so we must have $Q_1 = Q_2$ and, hence, $P_1 = P_2$. Finally, since each product gives a unique permutation of n, there can be no more than $n!$ such products. □

Theorem 11. *For $k < n$, $\alpha(\Gamma_{k,n}) \geq k!(n-k)!$ and $\chi(\Gamma_{k,n}) \leq \binom{n}{k}$.*

Proof. Consider H, the set of all permutations in S_n that send $\{1, 2, \dots, k\}$ to itself (and hence $\{k+1, \dots, n\}$ to itself). It is clear that H is a subgroup of S_n isomorphic to $S_k \times S_{n-k}$ and that $|H| = k!(n-k)!$. Since the unordered k-tuple $\{1, 2, \dots, k\}$ is fixed, none of these are k-derangements of each other, so H is an independent set and $\alpha(\Gamma_{k,n}) \geq k!(n-k)!$.

The cosets of H partition S_n, and each forms an independent set, since $\tau_1, \tau_2 \in \sigma H$ implies that $\tau_1^{-1}\tau_2 \in H$ is not a k-derangement and hence the vertices associated to τ_1 and τ_2 are not connected by an edge. Giving each of the $\frac{n!}{k!(n-k)!} = \binom{n}{k}$ cosets a different color results in a valid coloring of $\Gamma_{k,n}$, so $\chi(\Gamma_{k,n}) \leq \binom{n}{k}$. □

Corollary 12. *For n an odd prime power, $\alpha(\Gamma_{2,n}) = 2(n-2)!$ and $\chi(\Gamma_{2,n}) = \binom{n}{2}$.*

Proof. By Lemma 10 and Theorem 6, we have $\binom{n}{2} \cdot \alpha(\Gamma_{2,n}) \leq n!$. Thus

$$\alpha(\Gamma_{2,n}) \leq n! \cdot \frac{2(n-2)!}{n!} = 2(n-2)!$$

and Theorem 11 gives the reverse inequality. For any graph G, $\chi(G) \geq \omega(G)$, so, by Theorem 6, $\chi(\Gamma_{2,n}) \geq \binom{n}{2}$ and again Theorem 11 gives the reverse inequality. □

5. Further questions

In the last section, we showed that the clique number of the 2-derangement graph is equal to $\binom{n}{2}$ when n is an odd prime power and strictly less than that if n is even (and at least 4). The clique construction of Theorem 6 fails to work when n is odd and not a prime power since there is no field of that cardinality. We believe that in this case the clique number is strictly smaller than $\binom{n}{2}$. For arbitrary k, we have some faint hope that the bounds given in Theorem 11 for $\alpha(\Gamma_{k,n})$ and $\chi(\Gamma_{k,n})$ are actually equalities, but the situation for $\omega(\Gamma_{k,n})$ remains unclear.

In another direction, the numerical evidence is overwhelming that the derangement graphs are Hamiltonian. We hope to explore these and other questions in future work.

Acknowledgements

We would like to acknowledge Sam Tencer's contribution to this investigation. We are also grateful to an anonymous referee for spotting the connection between cliques attaining the bound of Lemma 5 and sharply k-homogeneous sets of permutations.

The first author worked on this topic with the third author during an REU at Missouri State University in the summer of 2009 (award #0552573).

References

[Frankl and Deza 1977] P. Frankl and M. Deza, "On the maximum number of permutations with given maximal or minimal distance", *J. Combinatorial Theory Ser. A* **22**:3 (1977), 352–360. MR 55 #12534 Zbl 0352.05003

[Henshaw 2008] J. Henshaw, "A137482: Number of permutations of n objects such that no two-element subset is preserved", entry A137482 in *The On-Line Encyclopedia of Integer Sequences* (http://oeis.org), 2008.

[de Montmort 1708] P. R. de Montmort, *Essay d'analyse sur les jeux de hazard*, 1st ed., Jacque Quillau, Paris, 1708.

[Nomura 1985] K. Nomura, "On t-homogeneous permutation sets", *Arch. Math.* (*Basel*) **44**:6 (1985), 485–487. MR 87e:20003 Zbl 0547.20003

[Renteln 2007] P. Renteln, "On the spectrum of the derangement graph", *Electron. J. Combin.* **14**:1 (2007), Research Paper 82, 17. MR 2008j:05217 Zbl 1183.05047

Received: 2011-09-14 Revised: 2012-05-22 Accepted: 2012-07-13

hljackso@syr.edu
Mathematics Department, Syracuse University, 215 Carnegie, Syracuse, NY 13244, United States

knyman@willamette.edu
Mathematics Department, Willamette University, 900 State Street, Salem, OR 97301, United States

lesreid@missouristate.edu
Mathematics Department, Missouri State University, 901 South National Avenue, Springfield, MO 65897, United States

INVOLVE 6:1 (2013)
dx.doi.org/10.2140/involve.2013.6.35

Rook polynomials in three and higher dimensions

Feryal Alayont and Nicholas Krzywonos

(Communicated by Jim Haglund)

The rook polynomial of a board counts the number of ways of placing nonattacking rooks on the board. In this paper, we describe how the properties of the two-dimensional rook polynomials generalize to the rook polynomials of "boards" in three and higher dimensions. We also define families of three-dimensional boards which generalize the two-dimensional triangle boards and the boards representing the problème des rencontres. The rook coefficients of these three-dimensional boards are shown to be related to famous number sequences such as the central factorial numbers, the number of Latin rectangles and the Genocchi numbers.

Introduction

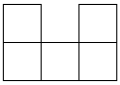
The theory of rook polynomials provides a way of counting permutations with restricted positions. This theory was developed in [Kaplansky and Riordan 1946] and has been researched and studied quite extensively since then. Two rather comprehensive resources on it are [Riordan 1958] and [Stanley 1997]. In this paper, we generalize these properties and theorems of the two-dimensional rook polynomials to higher dimensions, which was partially done for the three-dimensional case in [Zindle 2007]. A Maple program to calculate the rook numbers of a given three-dimensional board using this generalization is included in the Appendix. In Section 1 we review the two-dimensional rook polynomials and their properties, including a discussion of famous families of boards, namely, the boards corresponding to the problème des rencontres, and the triangle boards. The results provided in this review most of the time form the basis of the proofs for the three- and higher-dimensional cases. In Section 2 we discuss the generalization of the rook polynomials to three and higher dimensions, starting with a discussion of the three-dimensional boards and how rooks attack in three dimensions. We provide the generalizations of the properties and theorems of two-dimensional rook polynomials to three and higher dimensions as well as the three-dimensional counterparts of the boards

MSC2010: primary 05A05, 05A10; secondary 11B73.
Keywords: rook polynomial, three dimensions, central factorial numbers, Genocchi numbers, problème des rencontres.

corresponding to the problème des rencontres and the triangle boards. In Section 2.3 we introduce another family of three-dimensional boards related to the triangle boards. This family is named Genocchi boards due to its connection to the Genocchi numbers.

1. Overview of the rook theory in two dimensions

Given a natural number m, let $[m]$ denote the set $\{1, 2, \dots, m\}$. In two dimensions, we define a *board* B with m rows and n columns to be a subset of $[m] \times [n]$. We call such a board an $m \times n$ board if m and n are the smallest such natural numbers. Each of the elements in the board is referred to as a *cell* of the board. The set $[m] \times [n]$ is called the *full* $m \times n$ board. An example of how we visualize a board is this:



Numbering the rows from top to bottom and columns from left to right, the above picture corresponds to the 2×3 board $B = \{(1, 1), (1, 3), (2, 1), (2, 2), (2, 3)\}$. We sometimes highlight the cells missing from the board by shading them in gray.

The *rook polynomial* $R_B(x) = r_0(B) + r_1(B)x + \cdots + r_k(B)x^k + \cdots$ of a board B represents the number of ways that one can place various numbers of nonattacking rooks on B; i.e., no two rooks can lie in the same column or row. More specifically, $r_k(B)$ is equal to the number of ways of placing k nonattacking rooks on B. For any board, $r_0(B) = 1$ and $r_1(B)$ is equal to the number of cells in B. For the above example, $r_2(B) = 4$ as there are four different ways to place 2 nonattacking rooks on the board. It is not possible to place 3 or more rooks on this board. Hence the rook polynomial of this board is $R_B(x) = 1 + 5x + 4x^2$. In general, the number of nonattacking rooks placed on an $m \times n$ board cannot exceed n or m, and hence the rook polynomial, as indicated by its name, is a polynomial of degree less than or equal to $\min\{m, n\}$. Note that the rook polynomial of a board is invariant under permuting the rows and columns of the board.

Theorem. *The number of ways of placing k nonattacking rooks, with $0 \le k \le \min\{m, n\}$, on the full $m \times n$ board is equal to $\binom{m}{k}\binom{n}{k}k!$.*

Proof. First choose k of the m rows and k of the n columns on which the rooks will be placed. This can be done in $\binom{m}{k}\binom{n}{k}$ ways. Once we have selected the rows and columns, we place a rook in each column and row. For the first row, there are k columns to choose from. Once the first rook is placed, for the second row, there are $k - 1$ choices left. Continuing in this way, we find that there are $k!$ ways to place the k rooks on the chosen rows and columns. Hence we have $\binom{m}{k}\binom{n}{k}k!$ ways to place k nonattacking rooks on a full $m \times n$ board. □

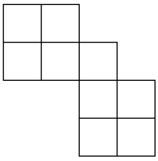
We define two boards to be *disjoint* if the boards do not share any rows or columns. If a board is composed of two disjoint subboards, the rook polynomial of the board can be calculated in terms of the rook polynomials of the subboards.

Theorem (disjoint board decomposition). *Let A and B be boards that share no rows or columns. Then the rook polynomial of the board $A \cup B$ consisting of the union of the cells in A and B is $R_{A\cup B}(x) = r_A(x) \times r_B(x)$.*

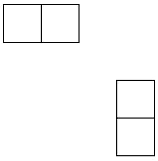
Proof. Let $R_A(x) = \sum_{k=0}^{\infty} r_k(A)x^k$ and $R_B(x) = \sum_{k=0}^{\infty} r_k(B)x^k$ be the rook polynomials of A and B. Consider the number of ways to place k rooks on $A \cup B$. We can place k rooks on A and 0 rooks on B, in $r_k(A)r_0(B)$ ways, or place $k-1$ rooks on A and 1 rook on B, in $r_{k-1}(A)r_1(B)$ ways, and so on. Hence, the number of ways to place k rooks on $A \cup B$ is $\sum_{i=0}^{k} r_{k-i}(A)r_i(B)$, which is the coefficient of x^k in $r_A(x) \times r_B(x)$. Therefore $R_{A\cup B}(x) = R_A(x) \times R_B(x)$. □

The rook polynomial of a board which can be decomposed into two disjoint subboards, possibly after permuting rows and/or columns, can thus be calculated efficiently via this theorem.

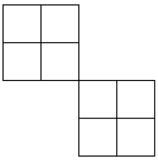
Similarly, *cell decomposition* is another method of expressing the rook polynomial of a board in terms of smaller boards. Consider the board B shown below.



This board cannot be decomposed into two disjoint boards even if we permute the rows and columns. The cell decomposition method breaks the rook placements down into cases: when there is a rook in a specific position, say cell $(2, 3)$, and when there is no rook in that position. If there is a rook on cell $(2, 3)$, we cannot have another rook in row two or column three. By deleting row two and column three we create a new board B' on which the rest of the rooks can be placed. For the case that no rook is placed on $(2, 3)$, we create a board B'' by deleting the cell $(2, 3)$.



B'



B''

In order to find $r_k(B)$, the number of ways of placing k rooks on B, we add $r_{k-1}(B')$ and $r_k(B'')$ using the two cases. In terms of the rook polynomial, this implies that $R_B(x) = xR_{B'}(x) + R_{B''}(x)$. For this specific example, the disjoint board decomposition can be used to compute the rook polynomials of B' and B'', making them much easier to compute than that of the board B.

The same idea of considering cases in regards to a specific cell as described above proves the more general cell decomposition.

Theorem (cell decomposition). *Let B be a board, B' be the board obtained by removing the row and column corresponding to a cell from B, and B'' be the board obtained by deleting the same cell from B. Then $R_B(x) = xR_{B'}(x) + R_{B''}(x)$.*

Another property of rook polynomials relates the rook polynomial of a board to that of the board consisting of the missing cells. Given an $m \times n$ board B, we define the *complement* of B, denoted by $\bar{B}$, to consist of all cells missing from B so that the disjoint union of B and $\bar{B}$ is the full $m \times n$ board. In other words $\bar{B} = [m] \times [n] \backslash B$. Sometimes we clearly indicate with respect to which board the complement is taken by saying that the complement is calculated inside $[m] \times [n]$.

Theorem (complementary board theorem). *Let $\bar{B}$ be the complement of B inside $[m] \times [n]$ and $R_B(x) = \sum r_i(B)x^i$ the rook polynomial of B. Then the number of ways to place k nonattacking rooks on $\bar{B}$ is*

$$r_k(\bar{B}) = \sum_{i=0}^{k} (-1)^i \binom{m-i}{k-i} \binom{n-i}{k-i} (k-i)!\, r_i(B), \tag{1}$$

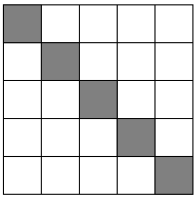
taking r_i to be 0 for i greater than the degree of $R_B(x)$.

Proof. In order to find the number of ways to place k nonattacking rooks on $\bar{B}$, we consider all the placements of k nonattacking rooks on the full $m \times n$ board and remove those where one or more rooks are placed on B using the inclusion-exclusion principle. We temporarily number the k rooks in our counting process, which means we will be counting $k!\, r_k(\bar{B})$. The total number of ways to place k numbered rooks on a full $m \times n$ board is $\binom{m}{k}\binom{n}{k}k!^2$, the additional $k!$ factor coming from the numbering of the rooks. Let A_i denote the set of placements of the rooks where the i-th rook is on the board B. We have to remove these placements from the set of all placements. There are $r_1(B)$ ways to place the i-th rook on B and $\binom{m-1}{k-1}\binom{n-1}{k-1}(k-1)!^2$ ways to place the rest in the other rows and columns. Hence there are $r_1(B)\binom{m-1}{k-1}\binom{n-1}{k-1}(k-1)!^2$ elements in A_i and there are k A_i's. Similarly, there are $r_2(B)2!\binom{m-2}{k-2}\binom{n-2}{k-2}(k-2)!^2$ elements in $A_i \cap A_j$ for any $i \neq j$ and there are $\binom{k}{2}$ of these double intersections. There are $r_3(B)3!\binom{m-3}{k-3}\binom{n-3}{k-3}(k-3)!^2$ elements in $\binom{k}{3}$ triple intersections $A_i \cap A_j \cap A_\ell$, and so on. Hence, using the inclusion-exclusion principle, the number of ways to place k numbered rooks on $\bar{B}$ is

$$\sum_{i=0}^{k} (-1)^i \binom{k}{i} i! \binom{m-i}{k-i} \binom{n-i}{k-i} (k-i)!^2 r_i(B).$$

Dividing this by $k!$ we arrive at (1). □

1.1. ***Problème des rencontres.*** We now consider the family of boards which correspond to the famous problème des rencontres, or equivalently to derangements. An example of such a problem is as follows. Suppose that five people enter a restaurant, each with his or her own hat. As they leave, they each take a hat, but not necessarily their own. We want to find the number of ways that everyone can leave the restaurant without their own hats, ignoring the order in which they leave. The boards which correspond to the problème des rencontres are $m \times m$ boards with the cells along the main diagonal removed. For the hat problem, $m = 5$ and we obtain the following board B, where we have highlighted the missing cells with gray.



The number of ways to place 5 rooks on B corresponds with the number of permutations of five elements where no element is in its original position. Such a permutation is equivalent to matching each owner with someone else's hat.

Instead of B, consider its complement. The complement $\bar{B}$ consists of 5 disjoint boards, each of which is a single cell. The rook polynomial of each cell is $1 + x$. Hence, using the disjoint board decomposition, we find that the rook polynomial of $\bar{B}$ is

$$(1+x)^5 = 1 + 5x + 10x^2 + 10x^3 + 5x^4 + x^5.$$

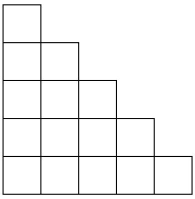
Now, using the theorem on rook polynomials of complementary boards, we find that the number of ways to place 5 rooks on B is equal to

$$\binom{5}{5}\binom{5}{5}5!\cdot 1 - \binom{4}{4}\binom{4}{4}4!\cdot 5 + \binom{3}{3}\binom{3}{3}3!\cdot 10 - \binom{2}{2}\binom{2}{2}2!\cdot 10 \\ + \binom{1}{1}\binom{1}{1}1!\cdot 5 - \binom{0}{0}\binom{0}{0}0!\cdot 1 = 44.$$

In general, r_k of the rook polynomial of an $m \times m$ problème des rencontres board is

$$\sum_{i=0}^{k}(-1)^i \binom{m-i}{k-i}^2 (k-i)!\binom{m}{i}.$$

1.2. ***Triangle boards.*** We next consider the family of two-dimensional boards called the triangle boards. A *triangle board* of size m consists of the cells of the form (i, j) where $j \le i$ and $1 \le i \le m$. The triangle board of size 5 is shown at the top of the next page.



The rook numbers of this family correspond with the Stirling numbers of the second kind. Recall that the Stirling numbers of the second kind, $S(n, k)$, count the number of ways to partition a set of size n into k nonempty sets, and can be defined recursively by

$$S(n, k) = S(n-1, k-1) + kS(n-1, k)$$

with $S(n, 1) = 1$ and $S(n, n) = 1$.

Theorem. *The number of ways to place k nonattacking rooks on a triangle board of size m is equal to* $S(m+1, m+1-k)$, *where* $0 \leq k \leq m$.

Proof. We will prove this by induction on m and using the recursive definition of the Stirling numbers.

The rook polynomial of the triangle board of size 1 is equal to $1 + x$, which corresponds to $S(2, 1) = 1$ and $S(2, 2) = 1$.

Assume now the theorem is true for some m, i.e., that the number of ways of placing k rooks on a size m triangle board is equal to $S(m+1, m+1-k)$ for $0 \leq k \leq m$. We will show that the number of ways of placing k rooks on an $(m+1) \times (m+1)$ board is equal to $S((m+1)+1, (m+1)+1-k) = S(m+2, m+2-k)$ for $0 \leq k \leq m+1$.

For $k = m+1$, there is only one way to place k nonattacking rooks on a size $m+1$ triangle board: by placing all rooks on the diagonal. This corresponds to $S(m+2, m+2-k) = S(m+2, 1) = 1$. For $k = 0$, placing k rooks on the board can be done in only one way, which corresponds to $S(m+2, m+2) = 1$. Therefore the rook numbers and the Stirling numbers agree for $k = 0$ and $k = m+1$.

We now show that these numbers agree for $0 < k < m+1$. When finding the number of ways to place k rooks on the size $m+1$ triangle board, we consider two cases. The first is when all k rooks are placed on the top m rows, forming a size m triangle board. There are $S(m+1, m+1-k)$ ways to do so by our inductive hypothesis. The second case is when one rook lies in the bottom row. In this case, $k-1$ rooks must be placed on the top m rows, which can be done in $S(m+1, m+1-(k-1))$ ways. We then have $m+1-(k-1)$ cells available in the last row to place our last rook, resulting in $(m+2-k)S(m+1, m+2-k)$ ways to place k rooks on the board with one rook in the last row. So there are a total of

$$S(m+1, m+1-k) + (m+2-k)S(m+1, m+2-k)$$

ways to place k rooks on a size $m+1$ triangle board. Using the recursive definition of the Stirling numbers, this sum corresponds to $S(m+2, m+2-k)$.

Therefore, by induction, the k-th rook number for any size m triangle board is $S(m+1, m+1-k)$. □

2. Rook polynomials in three and higher dimensions

The theory of rook polynomials in two dimensions as described above can be generalized to three and higher dimensions. The theory for three dimensions is introduced in [Zindle 2007] and the theory we describe in this paper is a more generalized version of Zindle's theory.

In three dimensions, our boards will be subsets of $[m] \times [n] \times [p]$. We refer to such a board as an $m \times n \times p$ board. More generally, a board in d dimensions is a subset of $[m_1] \times [m_2] \times \cdots \times [m_d]$. A *full board* is again a board that is the whole set $[m_1] \times [m_2] \times \cdots \times [m_d]$. In three and higher dimensions, a *cell* again refers to an element of the board. In particular, in three dimensions, a cell is a 3-tuple (i, j, k) with $1 \le i \le m$, $1 \le j \le n$, $1 \le k \le p$.

In two dimensions, rows correspond to cells with the same first coordinate, and columns correspond to cells with the same second coordinate. We extend this idea to three dimensions to introduce new groupings of cells. All cells with the same third coordinate are said to lie in the same *layer* and we number the layers from top to bottom. All cells with the same first coordinate are said to be in the same *slab* and all cells with the same second coordinate in the same *wall*. We still have rows and columns within a layer. We also have *towers* which correspond to the cells with the same first and second coordinate. In four and higher dimensions, we use *layer* to represent cells along any hyperplane formed by fixing a coordinate.

We generalize the rook theory to three dimensions so that a rook in three dimensions will attack along walls, slabs and layers. In higher dimensions, rooks attack along hyperplanes corresponding to cells with one fixed coordinate. In three dimensions, when we place a rook in a cell, we can no longer place another rook in the same wall, slab, or layer. In higher dimensions, a rook placed in a cell means we cannot place another rook in the fixed coordinate hyperplanes that this cell belongs to. In other words, if a rook is placed in cell $(i_1, i_2, \dots, i_d)$, then a rook may not be placed in any other cell sharing a coordinate with this cell. With this generalization, the rook polynomial of a board is invariant under permuting the layers of the board.

In another generalization of rook polynomials to three and higher dimensions, the rooks attack along lines instead of attacking over hyperplanes. For example, in three dimensions, a rook placed in cell (i, j, k) prohibits another rook from being placed in cells $(i, j, \cdot)$, $(i, \cdot, k)$ and $(\cdot, j, k)$. This approach has possible applications as well; however, we will not pursue this generalization in this paper.

Our first theorem on the generalized rook theory deals with a three-dimensional board obtained from a two-dimensional board extended in the z-direction. In other words, if A is a two-dimensional board, the three-dimensional extension of A with p layers consists of elements of the form (i, j, k) where $(i, j) \in A$ and $1 \leq k \leq p$. It is natural that there is a relation between the rook polynomials of the two boards.

Theorem. *Let A be an $m \times n$ board and B be a three-dimensional extension of A with p layers. Then, for* $0 \leq k \leq \min\{m, n, p\}$,

$$r_k(B) = \frac{p!}{(p-k)!} r_k(A).$$

Proof. Given a three-dimensional rook placement on the board B, consider the projection onto the board A. Since each rook can attack along either coordinate, when projected onto A no two rooks occupy the same cell in A and we get a placement of k rooks on A. There are $r_k(A)$ such placements. Given such a placement, we must distribute the k rooks among p layers. This is equivalent to k permutations of p numbers, which corresponds with $p!/(p-k)!$. So we have $r_k(A)p!/(p-k)!$ ways to place k rooks on B. Also note that for $k > \min\{m, n, p\}$, $r_k(B) = 0$ since k rooks cannot fit into the board. □

As a corollary of this theorem, we can obtain the rook numbers of the full three-dimensional boards, which are extensions of the full two-dimensional boards. However, we provide a proof similar to the two-dimensional case below which gives the idea of the proof of the general higher-dimensional theorem.

Theorem. *There are $\binom{m}{k}\binom{n}{k}\binom{p}{k}(k!)^2$ ways to place k nonattacking rooks on the full $m \times n \times p$ board for* $0 \leq k \leq \min\{m, n, p\}$.

Proof. Since we are placing k rooks on m slabs, n walls, and p layers, we have $\binom{m}{k}\binom{n}{k}\binom{p}{k}$ ways to choose the k slabs, walls, layers to place the rooks on. Since we have k rooks and k layers, there will be exactly one rook on each layer. For the first layer, we have k walls and k slabs from which we can choose to place the rook. After placing the first rook, on the second layer we will have $k-1$ slabs and $k-1$ walls as options. Continuing this way, we find that we have

$$k \cdot k \cdot (k-1) \cdot (k-1) \cdot (k-2) \cdot (k-2) \cdot \cdots \cdot 2 \cdot 2 \cdot 1 \cdot 1 = (k!)^2$$

ways to place the rooks on the chosen walls, slabs and layers. So there are $\binom{m}{k}\binom{n}{k}\binom{p}{k}(k!)^2$ ways to place k nonattacking rooks on the full $m \times n \times p$ board. □

The theorem for the most general case is:

Theorem. *There are $\binom{m_1}{k}\binom{m_2}{k} \cdots \binom{m_d}{k}(k!)^{d-1}$ ways to place k nonattacking rooks, with* $0 \leq k \leq \min_i m_i$, *on a full $m_1 \times m_2 \times \cdots \times m_d$ board in d dimensions.*

The decomposition theorems of the two-dimensional case also generalize naturally to three and higher dimensions. We define two boards in three dimensions to be *disjoint* if the boards do not share any walls, slabs or layers. In four and higher dimensions, the boards are *disjoint* if they do not share any layers. We then have the following disjoint board decomposition in the general case.

Theorem (disjoint board decomposition). *Let A and B be two boards in three or higher dimensions that share no layers. Then the rook polynomial of the board $A \cup B$ consisting of the union of the cells in A and B is $R_{A\cup B}(x) = R_A(x) \times R_B(x)$.*

The disjoint board theorem allows easy calculation of rook polynomials of a board which can be decomposed into disjoint subboards, possibly after permuting layers.

The cell decomposition method from the two-dimensional case generalizes to three and higher dimensions as follows with the proof being a slight modification of the proof in Section 1.

Theorem (cell decomposition). *Let B be a board, B' be the board obtained by removing the layers that correspond to a cell from B, and B'' be the board obtained by removing the same cell from B. Then $R_B(x) = xR_{B'}(x) + R_{B''}(x)$.*

The theorem on complementary boards generalizes to three and higher dimensions, with slight modification:

Theorem (complementary board theorem). *Let $\bar{B}$ be the complement of B inside $[m_1] \times [m_2] \times \cdots \times [m_d]$ and*

$$R_B(x) = \sum_i r_i(B)x^i$$

the rook polynomial of B. Then the number of ways to place $0 \le k \le \min_i m_i$ nonattacking rooks on B is

$$r_k(\bar{B}) = \sum_{i=0}^{k} (-1)^i \binom{m_1 - i}{k-i}\binom{m_2 - i}{k-i} \cdots \binom{m_d - i}{k-i}(k-i)!^{d-1} r_i(B). \quad (2)$$

Proof. The proof proceeds as in the two-dimensional case. We number the rooks and let A_i be the set of placements of the rooks where the i-th rook is on B. There are

$$\binom{m_1}{k}\binom{m_2}{k} \cdots \binom{m_d}{k} k!^d$$

ways to place k numbered rooks on the full board. There are

$$r_1(B)\binom{m_1 - 1}{k-1}\binom{m_2 - 1}{k-1} \cdots \binom{m_3 - 1}{k-1}(k-1)!^d$$

elements in A_i and there are k A_i's. Similarly, there are

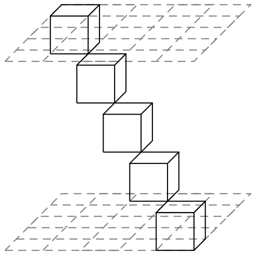
$$r_2(B)2!\binom{m_1-2}{k-2}\binom{m_2-2}{k-2}\cdots\binom{m_d-2}{k-2}(k-2)!^d$$

elements in $A_i \cap A_j$ for any $i \neq j$ and there are $\binom{k}{2}$ of these double intersections, and so on. Hence, using the inclusion-exclusion principle, the number of ways to place k numbered rooks on $\bar{B}$ is

$$\sum_{i=0}^{k}(-1)^i\binom{k}{i}i!\binom{m_1-i}{k-i}\binom{m_2-i}{k-i}\cdots\binom{m_d-i}{k-i}(k-i)!^d r_i(B).$$

Dividing this by $k!$ we arrive at (2). □

2.1. ***Problème des rencontres in three dimensions.*** Recall the problème des rencontres from earlier. The problème des rencontres dealt with a board with restrictions along the main diagonal. When creating a three-dimensional version of the problème des rencontres board, we will again place restrictions along the diagonal. In two dimensions we explained the problème des rencontres by considering five people leaving a restaurant without their hats. For this type of problem to make sense in three dimensions we will have to alter the scenario. We will once again consider five people entering a restaurant and introduce another dimension to the story. Let these people each have a hat and coat. We are now interested in the number of ways that the five people can leave the restaurant without both of their items. Let B be an $5 \times 5 \times 5$ board with elements (i, i, i) for $i = 1, \ldots, 5$ removed; we will consider placing 5 rooks on B. A visual representation of B is shown on the right.



For this board we let each layer represent a person, and walls and slabs represent coats and hats, respectively. The missing cells correspond with no person leaving with both their hat and coat. We refer to this board as the problème des rencontres board of the first kind. To find the rook numbers of this board, notice that the 5 missing cells form disjoint boards. The rook polynomial for each cell is $1 + x$. Hence, using the cell decomposition, we get

$$(1+x)^5 = 1 + 5x + 10x^2 + 10x^3 + 5x^4 + x^5$$

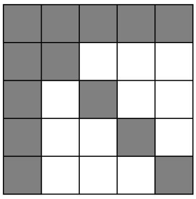
as the rook polynomial for the missing cells. Using the complementary board theorem, we then find that the number of ways to place 5 rooks on B is

$$\binom{5}{5}\binom{5}{5}\binom{5}{5}(5!)^2 - \binom{4}{4}\binom{4}{4}\binom{4}{4}(4!)^2\cdot 5 + \binom{3}{3}\binom{3}{3}\binom{3}{3}(3!)^2\cdot 10 \\ - \binom{2}{2}\binom{2}{2}\binom{2}{2}(2!)^2\cdot 10 + \binom{1}{1}\binom{1}{1}\binom{1}{1}(1!)^2\cdot 5 - \binom{0}{0}\binom{0}{0}\binom{0}{0}(0!)^2\cdot 1 = 11844.$$

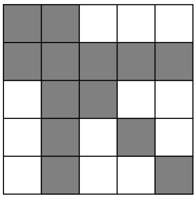
More generally, the number of ways that we can place k rooks on an $m \times m \times m$ problème des rencontres board of this kind is

$$\sum_{i=0}^{k}(-1)^i \binom{m-i}{k-i}^3 (k-i)!^2 \binom{k}{j}.$$

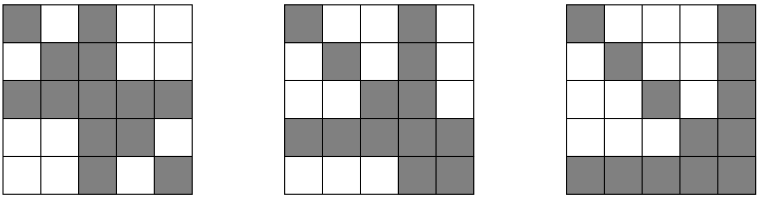
Another generalization of the problème des rencontres is to remove the rows, columns, and towers that pass through a diagonal cell, i.e., to remove cells of the form $(i, i, \cdot)$, $(i, \cdot, i)$ and $(\cdot, i, i)$. This second generalization corresponds to finding the number of ways that the five people can leave the restaurant without their coats, hats, or any proper pairing of a coat and hat. This means that each person must leave the restaurant with a hat that is not his or hers, and a coat that belongs neither to that person nor to the owner of the hat. The rook board for this problem is a bit more difficult to visualize so we will first discuss how to construct it. We again let each layer of the board represent a person, and the walls and slabs represent coats and hats, respectively. For layer 1, corresponding to the first person, we remove $(\ell, 1, 1)$ and $(1, \ell, 1)$ for $1 \le \ell \le 5$. This removes the column and row corresponding to the first person not leaving with their own coat or hat. We will also remove all cells along the main diagonal, meaning cells of the form $(\ell, \ell, 1)$ for $1 \le \ell \le 5$. This corresponds with person one not leaving with another person's coat and hat. The first layer of the board will then appear as follows:



For the second layer we remove $(\ell, 2, 2)$ and $(2, \ell, 2)$ for $1 \le \ell \le 5$. This will remove the row and column associated with the second person leaving with his or her own coat or hat. We will also remove $(\ell, \ell, 2)$ for $1 \le \ell \le 5$. This corresponds with the second person not leaving with another person's coat and hat. This layer will appear as follows:



Continuing this method for the final three layers we get:



The problème des rencontres board of the second kind of any size m is constructed in a similar fashion.

We use a Maple program to compute the rook polynomials of this type board of various sizes. The program is included in the Appendix. The rook numbers of boards of size from 3 up to 7 are given in Table 1.

Notice from the table that the rook numbers for $k = m$ correspond to the number of $3 \times m$ Latin rectangles. In fact, the correspondence between these rook placements and the Latin rectangles is very natural.

Theorem. *The number of ways to place m rooks on the size m problème des rencontres board of the second kind is equal to the number of* $3 \times m$ *Latin rectangles in which the first row is in order.*

Proof. A $3 \times m$ Latin rectangle consists of three rows, each of which is a permutation of the numbers in $[m]$ and where in each of the m columns no number is repeated. Given such a rectangle, each column can be represented by an ordered triple (r_1, r_2, r_3) in which no two entries are the same. These are exactly the cells missing from the problème des rencontres board of the second kind. We then take these m ordered triples and place rooks in the corresponding cells of this board. Because each number appears in each row of the Latin rectangle exactly once, we have exactly one rook per slab, wall, and layer. Therefore, the rooks are nonattacking. This shows that any $3 \times m$ Latin rectangle corresponds with a valid placement of m rooks on the size m problème des rencontres board of the second kind.

Now consider an arbitrary placement of m rooks on the size m problème des rencontres board of the second kind. Since there are m rooks, there is a rook in each slab. We read the positions of the rooks starting with the rook in the first slab, and record these into the columns of a $3 \times m$ array. In this way, the first row is

$m\backslash k$	0	1	2	3	4	5	6	7
3	1	6	6	2				
4	1	24	132	176	24			
5	1	60	960	4580	5040	552		
6	1	120	4260	52960	213000	206592	21280	
7	1	210	14070	368830	3762360	13109712	11404960	1073160

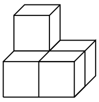
Table 1. Rook numbers of boards of size 3–7.

arranged from 1 to m in increasing order, and, as explained above, each row is a permutation of $[m]$, and no two entries in each column are the same. This also shows that the correspondence is one-to-one. □

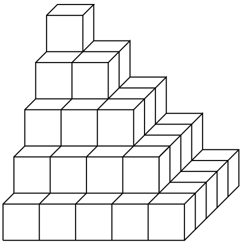
This second kind of the problème des rencontres board can be generalized to dimensions higher than three as follows: a size m problème des rencontres board in d dimensions is a subset of the set $[m]^d$ where the cells with at least two equal coordinates are removed. With this generalization, using a method similar to the proof of the above theorem, we obtain the following theorem:

Theorem. *The number of ways to place m rooks on a size m problème des rencontres board in d dimensions is equal to the number of $d \times m$ Latin rectangles in which the first row is in order.*

2.2. ***Triangle boards in three dimensions.*** In two dimensions the triangle board of size m contains the cells of the form (i, j) with $j \leq i$ and $1 \leq i \leq m$. This board has the property that there is only one way to place m rooks on a size m triangle board. Another property of the triangle board is that removing both the row and column corresponding to a diagonal cell of a size m triangle board results in a size $m-1$ triangle board. We want to replicate these aspects of a triangle board in three dimensions, and this is how the three-dimensional triangle board evolved. In three dimensions a size 1 triangle board is simply one cell. The size 2 triangle board is obtained by placing a 2×2 layer below the size 1 triangle board as follows:



We build the larger triangle boards recursively in a similar way, by adding an $(m+1) \times (m+1)$ layer at the bottom of a size m triangle board. The cells included in the size m triangle are (i, j, k) with $1 \leq i, j \leq k$ and $1 \leq k \leq m$. With this definition, there is only one way to place m rooks on a size m triangle board. Additionally, removing the wall, slab and layer including a diagonal cell of a size m triangle board results in a size $m-1$ triangle board. The size 5 triangle board is depicted below.



The rook numbers of the triangle boards up to size 8 are calculated using Maple and are shown in the table on the next page. The numbers turn out to be the *central factorial numbers* defined recursively by

$$T(n,k)=T(n-1,k-1)+k^2T(n-1,k),$$

with $T(n,1)=1$ and $T(n,n)=1$; see Table 2.

Theorem. *The number of ways to place k rooks on a size m triangle board in three dimensions is equal to* $T(m+1,m+1-k)$, *where* $0\le k\le m$.

Proof. We will prove this theorem by induction on m.

For the base case, $m=1$, the rook polynomial is $1+x$ and the corresponding central factorial numbers are $T(2,2)=T(2,1)=1$. Hence the result is true for $m=1$.

Assume now the theorem is true for some m, i.e., that the number of ways of placing k rooks on a size m triangle board is equal to $T(m+1,m+1-k)$ for $0\le k\le m$. We will show that the number of ways of placing k rooks on an $(m+1)\times(m+1)$ board is equal to $T((m+1)+1,\ (m+1)+1-k)=T(m+2,m+2-k)$ for $0\le k\le m+1$.

We know that there is only one way to place no rooks, which corresponds to $T(m+2,m+2)=1$. We also know that there is only one way to place the maximum number of rooks, $m+1$ rooks, which corresponds to

$$T(m+2,m+2-(m+1))=T(m+2,1)=1.$$

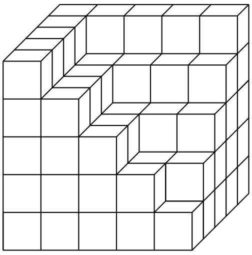
Now let $0<k<m+1$. Similar to the two-dimensional case, we consider two cases, when all rooks are on the top m layers and when one of the rooks is on the bottom layer. The top m layers form a triangle board of size m and hence the number of ways to place k rooks on the top m layers is $T(m+1,m+1-k)$. If one

$n\backslash k$	0	1	2	3	4	5	6	7	8
0	1								
1	1	1							
2	1	5	1						
3	1	21	14	1					
4	1	85	147	30	1				
5	1	341	1408	627	55	1			
6	1	1365	13013	11440	2002	91	1		
7	1	5461	118482	196053	61490	5278	140	1	
8	1	21845	1071799	3255330	1733303	251498	12138	204	1

Table 2. Sequence A008957 in [Sloane 2009], triangle of central factorial numbers.

rook is on the bottom layer, the rest of the rooks will be on the top m layers, which can be done in $T(m+1, m+1-(k-1))$ ways. Once these $k-1$ rooks are placed, the corresponding $k-1$ rows and columns in the bottom layer are restricted for the last rook, leaving $(m+1-(k-1))^2$ cells available for that rook. Hence, there are a total of $(m+2-k)^2 T(m+1, m+2-k)$ ways to have k rooks on the board with one being on the bottom layer. Adding the results from the two cases, we obtain $T(m+1, m+1-k)+(m+2-k)^2 T(m+1, m+2-k)$ ways of placing the k rooks on the size $m+1$ triangle board. By the recursive definition of the central factorial numbers, this sum corresponds to $T(m+2, m+2-k)$, proving the theorem by induction. □

2.3. ***Genocchi board.*** Another possible three-dimensional generalization of the triangle boards is obtained by generalizing the following property of the two-dimensional triangle boards. The number of cells in each row of a two-dimensional triangle board is equal to the row number. We generalize this property by letting the number of cells in a tower over a fixed row and column be equal to the maximum of the row and column numbers. In terms of the coordinates, the cells in the size m three-dimensional triangle board are of the form (i, j, k) with $1 \leq k \leq \max\{i, j\}$ and $1 \leq i, j \leq m$. The rook numbers of these boards are related to the Genocchi numbers; hence we call this family the *Genocchi boards*. Below is the depiction of the size 5 Genocchi board turned upside down and rotated for clarity. From the picture, we can see that the complement of the size m Genocchi board inside the $m \times m \times m$ cube is the size $m-1$ triangle board.



Using Maple, we generated rook numbers for various Genocchi boards. We found that the number of ways to place m rooks on a board of size m corresponds with the unsigned $(m+1)$-th Genocchi number; see Table 3.

m	0	1	2	3	4	5	6	7
r_m	1	1	3	17	155	2073	38227	929569
G_{m+1}	−1	1	−3	17	−155	2073	−38227	929569

Table 3. Sequence A001469 in [Sloane 2009], Genocchi numbers (of first kind).

Theorem. *The number of ways to place m nonattacking rooks on a size m Genocchi board is the unsigned $(m+1)$-th Genocchi number.*

Proof. Recall that the complement of a size m Genocchi board in an $m \times m \times m$ cube is a size $m-1$ triangle board. Hence, using the theorem of complementary boards, we can calculate the number of ways to place m rooks on a size m Genocchi board in terms of the rook numbers of the triangle board. Recall that r_k for a size m triangle board is $T(m+1, m+1-k)$ and that the number of ways to place k rooks on the complement of a three-dimensional board B in the $m \times m \times m$ cube is

$$\sum_{i=0}^{k} \binom{m-i}{k-i}^3 (k-i)!^2 r_i(B).$$

Using these two formulas, we find that the number of ways to place $k = m$ rooks on a size m Genocchi board is

$$\sum_{i=0}^{m-1} \binom{m-i}{m-i}^3 (m-i)!^2 T(m, m-i).$$

We omitted the term corresponding to $i = m$ in the summation because $r_m(B) = 0$ for the triangle board of size $m-1$. This last summation can be rewritten via a change of variables $j = m - i$ as

$$(-1)^{m+1} \sum_{j=1}^{m} (-1)^{j+1} j!^2 T(m, j),$$

which is shown to equal $(-1)^{m+1} G_{m+1}$ in [Dumont 1974]; thus the number of ways to place m rooks on a size m Genocchi board is the unsigned $(m+1)$-th Genocchi number. □

Appendix

```
Rook:=proc(A,m,n,p,B,k,rem)
local C,i,j,h,g,l,count,v;
count:=0;
if k=1 then
 for i from 1 to m do
  for j from 1 to n do
   for g from 1 to p do
    if ’not’(’in’([i,j,g],B)) then
     if add(add(A[i,a1,a2],a1=1..n),a2=1..p)=0 then
      if add(add(A[b1,j,b2],b1=1..m),b2=1..p)=0 then
       if add(add(A[c1,c2,g],c1=1..m),c2=1..n)=0 then
        count:=count+1
       end if
      end if
     end if
    end if
   end do
  end do
 end do
else
C:=Array(1..m,1..n,1..p);
 for i from 1 to m do
  for j from 1 to n do
   for g from 1 to p do
    if ’not’(’in’([i,j,g],B)) then
     if add(add(A[i,a1,a2],a1=1..n),a2=1..p)=0 then
      if add(add(A[b1,j,b2],b1=1..m),b2=1..p)=0 then
       if add(add(A[c1,c2,g],c1=1..m),c2=1..n)=0 then
        for h from 1 to m do
         for l from 1 to n do
          for v from 1 to p do
           C[h,l,v]:=A[h,l,v]
          end do
         end do
        end do
        C[i,j,g]:=1;
        count:=count+Rook(C,m,n,p,B,k-1);
        C[i,j,g]:=0
       end if
      end if
     end if
    end if
   end do
  end do
 end do
end if
count:=count/k
end proc:
```

References

[Dumont 1974] D. Dumont, “Interprétations combinatoires des nombres de Genocchi”, *Duke Math. J.* **41** (1974), 305–318. MR 49 #2412 Zbl 0297.05004

[Kaplansky and Riordan 1946] I. Kaplansky and J. Riordan, “The problem of the rooks and its applications”, *Duke Math. J.* **13**:2 (1946), 259–268. MR 7,508d Zbl 0060.02903

[Riordan 1958] J. Riordan, *An introduction to combinatorial analysis*, Wiley, New York, 1958. MR 20 #3077 Zbl 0078.00805

[Sloane 2009] N. J. A. Sloane, "The on-line encyclopedia of integer sequences", 2009, available at http://oeis.org/.

[Stanley 1997] R. P. Stanley, *Enumerative combinatorics, I*, Wadsworth & Brooks/Cole, Monterey, CA, 1997. MR 98a:05001 Zbl 0608.05001

[Zindle 2007] B. Zindle, *Rook polynomials for chessboards of two and three dimensions*, Master's thesis, Rochester Institute of Technology, 2007, available at https://ritdml.rit.edu/bitstream/handle/1850/5968/BZindle_Thesis-2007.pdf.

Received: 2011-09-15 Accepted: 2012-05-30

alayontf@gvsu.edu *Department of Mathematics, Grand Valley State University,*
1 Campus Drive, Allendale, MI 49401, United States

krzywonos123@gmail.com *Department of Mathematics, Grand Valley State University,*
1 Campus Drive, Allendale, MI 49401, United States

mathematical sciences publishers msp

INVOLVE 6:1 (2013)
dx.doi.org/10.2140/involve.2013.6.53

New confidence intervals for the AR(1) parameter

Ferebee Tunno and Ashton Erwin

(Communicated by Robert B. Lund)

This paper presents a new way to construct confidence intervals for the unknown parameter in a first-order autoregressive, or AR(1), time series. Typically, one might construct such an interval by centering it around the ordinary least-squares estimator, but this new method instead centers the interval around a linear combination of a weighted least-squares estimator and the sample autocorrelation function at lag one. When the sample size is small and the parameter has magnitude closer to zero than one, this new approach tends to result in a slightly thinner interval with at least as much coverage.

1. Introduction

Consider the causal stationary AR(1) time series given by

$$X_t = \phi X_{t-1} + \epsilon_t, \quad t = 0, \pm 1, \pm 2, \dots, \tag{1-1}$$

where $|\phi| < 1$, $E(X_t) = 0$ and $\{\epsilon_t\} \overset{\text{iid}}{\sim} N(0, \sigma^2)$. We seek a new way to construct confidence intervals for the unknown parameter ϕ.

If $X_1, X_2, \dots, X_n$ are sample observations from this process, then a point estimate for ϕ is found by calculating

$$\tilde{\phi}_p = \frac{\sum_{t=2}^{n} S_{t-1}|X_{t-1}|^p X_t}{\sum_{t=2}^{n} |X_{t-1}|^{p+1}},$$

where $p \in \{0, 1, 2, \dots\}$ and S_t is the sign function, defined by

$$S_t = \begin{cases} 1 & \text{if } X_t > 0, \\ 0 & \text{if } X_t = 0, \\ -1 & \text{if } X_t < 0. \end{cases}$$

The estimator $\tilde{\phi}_p$ can be thought of as a weighted least-squares estimator with form

$$\frac{\sum_{t=2}^{n} W_{t-1} X_{t-1} X_t}{\sum_{t=2}^{n} W_{t-1} X_{t-1}^2}$$

MSC2010: 60G10, 62F12, 62F99, 62M10.

Keywords: confidence interval, autoregressive parameter, weighted least squares, linear combination.

and weight $W_t = |X_t|^{p-1}$. Note that, when $p = 1$, we get the ordinary (unweighted) least-squares estimator (OLSE), and when $p = 0$, we get what has come to be called the Cauchy estimator:

$$\tilde{\phi}_1 = \frac{\sum_{t=2}^{n} X_{t-1}X_t}{\sum_{t=2}^{n} X_{t-1}^2} \quad \text{(OLSE)}, \qquad \tilde{\phi}_0 = \frac{\sum_{t=2}^{n} S_{t-1}X_t}{\sum_{t=2}^{n} |X_{t-1}|} \quad \text{(Cauchy)}.$$

The OLSE has been studied since the time of Gauss and its optimal properties for linear models are well known. The eponymously named Cauchy estimator dates back to about the same time and is sometimes used as a surrogate for the OLSE. Traditionally, confidence intervals for ϕ have been centered around the OLSE, although So and Shin [1999] and Phillips, Park and Chang [Phillips et al. 2004] showed that the Cauchy estimator has certain advantages over the OLSE when dealing with a unit root autoregression. Gallagher and Tunno [2008] constructed a confidence interval for ϕ centered around a linear combination of both estimators.

Another point estimate for ϕ comes from the sample autocorrelation function of $\{X_t\}$ at lag one, given by

$$\hat{\rho}(1) = \frac{\sum_{t=2}^{n} X_{t-1}X_t}{\sum_{t=1}^{n} X_t^2}.$$

The autocovariance function of $\{X_t\}$ at lag h for an AR(1) series is given by $\gamma(h) = \operatorname{Cov}(X_t, X_{t+h}) = \phi^{|h|}\sigma^2/(1-\phi^2)$, which makes the true lag-one autocorrelation function equal to

$$\rho(1) = \frac{\gamma(1)}{\gamma(0)} = \frac{\phi\sigma^2/(1-\phi^2)}{\sigma^2/(1-\phi^2)} = \phi.$$

Observe that the structure of $\hat{\rho}(1)$ is similar to that of the OLSE. In fact, for an AR(1) series, the Yule–Walker, maximum likelihood, and least-squares estimators for ϕ are all approximately the same [Shumway and Stoffer 2006, Section 3.6]. Note also that, in general, if $\{X_t\}$ is not mean-zero, we would subtract $\overline{X}$ from each observation when calculating things like $\hat{\rho}(h)$ and $\tilde{\phi}_p$.

To get a feel for how $\tilde{\phi}_0$, $\tilde{\phi}_1$ and $\hat{\rho}(1)$ behave relative to one another, Figure 1 shows their empirical bias and mean squared error (MSE) when $\phi \in (-1, 1)$ and $n = 50$. The Cauchy estimator has the lowest absolute bias, and $\hat{\rho}(1)$ has the smallest MSE for parameter values (roughly) between -0.5 and 0.5, while the OLSE has the smallest MSE elsewhere. Other simulations not shown here reveal that the MSE and absolute bias of $\tilde{\phi}_p$ keep growing as p gets larger.

The goal of this paper is to construct a confidence interval for ϕ centered around a linear combination of an arbitrary weighted least-squares estimator and the sample autocorrelation function at lag one. That is, the center will take the form

$$a_1\tilde{\phi}_p + a_2\hat{\rho}(1), \tag{1-2}$$

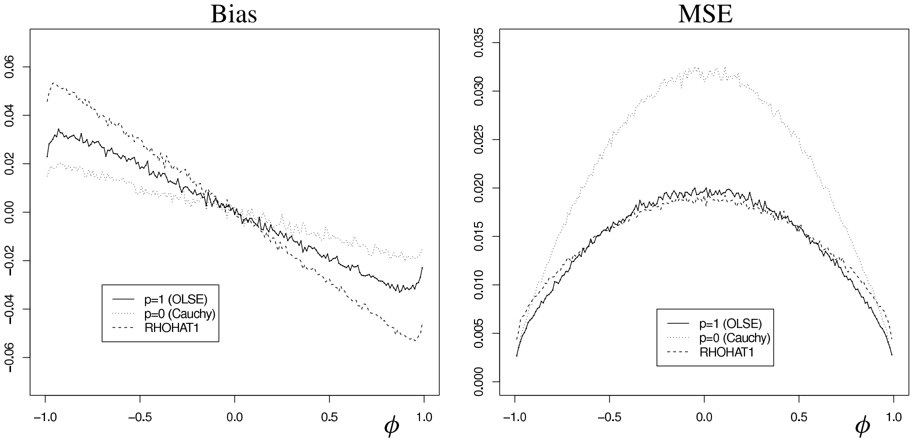


Figure 1. Empirical bias (left) and mean squared error (right) of $\tilde{\phi}_0$, $\tilde{\phi}_1$ and $\hat{\rho}(1)$ for $\phi \in (-1, 1)$; 10,000 simulations were run for each parameter value, with distribution $N(0, 1)$ and $n = 50$.

where $a_1 + a_2 = 1$ and $p \neq 1$. We first, however, need to take a brief look at how intervals centered around a single estimator behave in order to find a proper target for our new interval to outperform.

Theorem 2.1 from [Gallagher and Tunno 2008] states that for the AR(1) series given in (1-1), we have

$$\sqrt{n}(\tilde{\phi}_p - \phi) \xrightarrow{D} N\left(0, \frac{\sigma^2 E(X_t^{2p})}{(E|X_t|^{p+1})^2}\right) \tag{1-3}$$

for all p such that $E(|X_t|^r) < \infty$, where $r = \max(2p, p+1)$. Since the error terms in our series are normal, the X_t's have finite moments of all orders. Thus, this theorem can be used to create confidence intervals for ϕ centered at $\tilde{\phi}_p$ for any choice of p.

Specifically, if $X_1, X_2, \dots, X_n$ are sample observations from (1-1), then an approximate $(1-\alpha) \times 100\%$ confidence interval for ϕ has endpoints

$$\tilde{\phi}_p \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\tilde{\phi}_p)},$$

where

$$n\widehat{\mathrm{Var}}(\tilde{\phi}_p) = \frac{\sigma^2 n^{-1} \sum_{t=2}^{n} X_{t-1}^{2p}}{\left(n^{-1} \sum_{t=2}^{n} |X_{t-1}|^{p+1}\right)^2} \xrightarrow{P} \frac{\sigma^2 E(X_{t-1}^{2p})}{\left(E|X_{t-1}|^{p+1}\right)^2}$$

and $z_{\alpha/2}$ is the standard normal critical value with area $\alpha/2$ to its right.

Similarly, we can create confidence intervals for ϕ centered at $\hat{\rho}(1)$. If we think of $\hat{\rho}(1)$ as being nearly the equivalent of the OLSE, then an approximate $(1-\alpha) \times 100\%$ confidence interval for ϕ has endpoints

$$\hat{\rho}(1) \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{\rho}(1))},$$

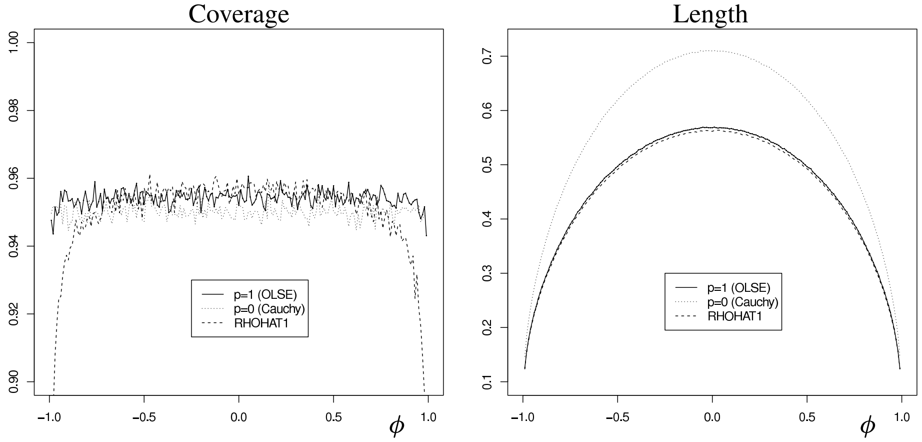


Figure 2. Empirical coverage capability (left) and length (right) of 95% confidence intervals for ϕ centered at $\tilde{\phi}_0$, $\tilde{\phi}_1$ and $\hat{\rho}(1)$ for $\phi \in (-1, 1)$; 10,000 simulations were run for each parameter value, with distribution $N(0, 1)$ and $n = 50$.

where

$$n\widehat{\mathrm{Var}}(\hat{\rho}(1)) = \frac{n\sigma^2}{\sum_{t=1}^{n} X_t^2} \xrightarrow{P} \frac{\sigma^2}{E(X_t^2)}.$$

Figure 2 shows the empirical coverage capability and length of 95% confidence intervals for ϕ centered at $\tilde{\phi}_0$, $\tilde{\phi}_1$ and $\hat{\rho}(1)$ when $\phi \in (-1, 1)$ and $n = 50$. The thinnest intervals occur when $\hat{\rho}(1)$ is used, although not by much. The OLSE also has the best overall coverage, except (roughly) for $|\phi| \leq 0.5$, which is where $\hat{\rho}(1)$ once again outperforms the OLSE. Other simulations not shown here reveal that the length of intervals centered at $\tilde{\phi}_p$ keeps growing as p gets larger, while coverage capability starts to break down for $|\phi|$ near 1.

In this paper, we will aim to construct intervals with center (1-2) that outperform those centered at the OLSE. The next section shows the details of this construction, while Section 3 presents some simulations. Section 4 closes the paper with an application and some remarks.

2. Interval construction

Suppose for the moment that we wish to construct a confidence interval for ϕ centered at a linear combination of two weighted least-squares estimators. That is, instead of (1-2), the center would take the form

$$a_1\tilde{\phi}_p + a_2\tilde{\phi}_q, \tag{2-1}$$

where $a_1 + a_2 = 1$ and $p \neq q$. Minimizing the variance of this quantity is equivalent to minimizing the length of the corresponding interval and occurs when

$$a_1 = \frac{\operatorname{Var}(\tilde{\phi}_q) - \operatorname{Cov}(\tilde{\phi}_p, \tilde{\phi}_q)}{\operatorname{Var}(\tilde{\phi}_p - \tilde{\phi}_q)}. \tag{2-2}$$

Theorem 2.1. *Let $a_1 + a_2 = 1$. If a_1 is given by* (2-2), *then* $\operatorname{Var}(a_1\tilde{\phi}_p + a_2\tilde{\phi}_q)$ *is minimized and has upper bound* $\operatorname{Var}(\tilde{\phi}_q)$.

Proof. Let

$$\begin{aligned} f(a_1) &= \operatorname{Var}\big(a_1\tilde{\phi}_p + (1-a_1)\tilde{\phi}_q\big) \\ &= a_1^2 \operatorname{Var}(\tilde{\phi}_p) + (1-a_1)^2 \operatorname{Var}(\tilde{\phi}_q) + 2a_1(1-a_1)\operatorname{Cov}(\tilde{\phi}_p, \tilde{\phi}_q) \\ &= a_1^2 \operatorname{Var}(\tilde{\phi}_p - \tilde{\phi}_q) + 2a_1\big(\operatorname{Cov}(\tilde{\phi}_p, \tilde{\phi}_q) - \operatorname{Var}(\tilde{\phi}_q)\big) + \operatorname{Var}(\tilde{\phi}_q). \end{aligned}$$

Then $f'(a_1) = 2a_1 \operatorname{Var}(\tilde{\phi}_p - \tilde{\phi}_q) + 2\big(\operatorname{Cov}(\tilde{\phi}_p, \tilde{\phi}_q) - \operatorname{Var}(\tilde{\phi}_q)\big) = 0$

$$\Rightarrow \quad a_1 = \frac{\operatorname{Var}(\tilde{\phi}_q) - \operatorname{Cov}(\tilde{\phi}_p, \tilde{\phi}_q)}{\operatorname{Var}(\tilde{\phi}_p - \tilde{\phi}_q)}.$$

Since $f''(a_1) = 2\operatorname{Var}(\tilde{\phi}_p - \tilde{\phi}_q) > 0$, then this critical value minimizes f. Note that this means

$$a_2 = 1 - a_1 = \frac{\operatorname{Var}(\tilde{\phi}_p) - \operatorname{Cov}(\tilde{\phi}_p, \tilde{\phi}_q)}{\operatorname{Var}(\tilde{\phi}_p - \tilde{\phi}_q)},$$

where the choices of p and q determine the ranges of a_1 and a_2. Specifically, we have

$$\begin{aligned} \operatorname{Var}(\tilde{\phi}_p) > \operatorname{Var}(\tilde{\phi}_q) &\iff a_1 < 0.5 \text{ and } a_2 > 0.5, \\ \operatorname{Var}(\tilde{\phi}_q) > \operatorname{Var}(\tilde{\phi}_p) &\iff a_1 > 0.5 \text{ and } a_2 < 0.5, \\ \operatorname{Var}(\tilde{\phi}_p) = \operatorname{Var}(\tilde{\phi}_q) &\iff a_1 = a_2 = 0.5. \end{aligned}$$

Finally, since the critical value found above minimizes f, we have $f(a_1) \le f(0)$, which is equivalent to saying

$$\operatorname{Var}\big(a_1\tilde{\phi}_p + (1-a_1)\tilde{\phi}_q\big) \le \operatorname{Var}(\tilde{\phi}_q),$$

where the inequality is strict for $a_1 \neq 0$. □

We would like for the variance of $a_1\tilde{\phi}_p + a_2\tilde{\phi}_q$ to be less than or equal to that of the OLSE. Setting $q = 1$ makes this happen since Theorem 2.1 tells us that

$$\operatorname{Var}(a_1\tilde{\phi}_p + a_2\tilde{\phi}_1) \le \operatorname{Var}(\tilde{\phi}_1).$$

It turns out, however, that the window where these two variances are distinguishable may be brief since a_1 goes to zero as the sample size increases. This in turn causes $a_1\tilde{\phi}_p + a_2\tilde{\phi}_1$ to be asymptotically normal.

Theorem 2.2. *Let* $a_1 + a_2 = 1$. *If* a_1 *is given by* (2-2) *with* $q = 1$, *then*

$$\sqrt{n}\big(a_1\tilde{\phi}_p + a_2\tilde{\phi}_1 - \phi\big) \xrightarrow{D} N\left(0, \frac{\sigma^2}{E(X_t^2)}\right).$$

Proof. First, we note that

$$n\,\mathrm{Cov}(\tilde{\phi}_p, \tilde{\phi}_q) \xrightarrow{P} \frac{\sigma^2 E|X_t|^{p+q}}{E|X_t|^{p+1}E|X_t|^{q+1}}.$$

Then

$$\begin{aligned} a_1 &= \frac{\mathrm{Var}(\tilde{\phi}_1) - \mathrm{Cov}(\tilde{\phi}_p, \tilde{\phi}_1)}{\mathrm{Var}(\tilde{\phi}_p - \tilde{\phi}_1)} \\ &= \frac{n\,\mathrm{Var}(\tilde{\phi}_1) - n\,\mathrm{Cov}(\tilde{\phi}_p, \tilde{\phi}_1)}{n\,\mathrm{Var}(\tilde{\phi}_p) + n\,\mathrm{Var}(\tilde{\phi}_1) - 2n\,\mathrm{Cov}(\tilde{\phi}_p, \tilde{\phi}_1)} \\ &\xrightarrow{P} \frac{\dfrac{\sigma^2}{E|X_t|^2} - \dfrac{\sigma^2}{E|X_t|^2}}{\dfrac{\sigma^2 E|X_t|^{2p}}{(E|X_t|^{p+1})^2} + \dfrac{\sigma^2}{E|X_t|^2} - \dfrac{2\sigma^2}{E|X_t|^2}} = \frac{0}{\dfrac{\sigma^2 E|X_t|^{2p}}{(E|X_t|^{p+1})^2} - \dfrac{\sigma^2}{E|X_t|^2}} \\ &\qquad =: R. \end{aligned}$$

The denominator of R is strictly positive since

$$\operatorname*{plim}_{n\to\infty} n\,\mathrm{Var}(\tilde{\phi}_p) > \operatorname*{plim}_{n\to\infty} n\,\mathrm{Var}(\tilde{\phi}_1) \quad \text{for } p \neq 1.$$

Thus, $R = 0$.

Since $a_1 \xrightarrow{P} 0$, we obtain $a_2 \xrightarrow{P} 1$. Hence, $a_1\tilde{\phi}_p + a_2\tilde{\phi}_1$ and $\tilde{\phi}_1$ have the same asymptotic distribution. By (1-3), we have

$$\sqrt{n}(\tilde{\phi}_1 - \phi) \xrightarrow{D} N\left(0, \frac{\sigma^2}{E(X_t^2)}\right).$$

Thus,

$$\sqrt{n}(a_1\tilde{\phi}_p + a_2\tilde{\phi}_1 - \phi) \xrightarrow{D} N\left(0, \frac{\sigma^2}{E(X_t^2)}\right)$$

as well. □

If $X_1, X_2, \ldots, X_n$ are sample observations from (1-1), then an approximate $(1-\alpha) \times 100\%$ confidence interval for ϕ centered at $a_1\tilde{\phi}_p + a_2\tilde{\phi}_1$ has endpoints

$$a_1\tilde{\phi}_p + a_2\tilde{\phi}_1 \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(a_1\tilde{\phi}_p + a_2\tilde{\phi}_1)}.$$

Letting

$$\hat{\sigma}_i^2 = \frac{\sigma^2 \sum_{t=2}^{n} |X_{t-1}|^{2i}}{\left(\sum_{t=2}^{n} |X_{t-1}|^{i+1}\right)^2}, \quad \hat{\sigma}_{ij} = \frac{\sigma^2 \sum_{t=2}^{n} |X_{t-1}|^{i+j}}{\sum_{t=2}^{n} |X_{t-1}|^{i+1} \sum_{t=2}^{n} |X_{t-1}|^{j+1}},$$

$$\hat{a}_1 = \frac{\hat{\sigma}_1^2 - \hat{\sigma}_{p1}}{\hat{\sigma}_p^2 + \hat{\sigma}_1^2 - 2\hat{\sigma}_{p1}}, \quad \hat{a}_2 = \frac{\hat{\sigma}_p^2 - \hat{\sigma}_{p1}}{\hat{\sigma}_p^2 + \hat{\sigma}_1^2 - 2\hat{\sigma}_{p1}},$$

we then have

$$\begin{aligned} n\widehat{\mathrm{Var}}(a_1\tilde{\phi}_p + a_2\tilde{\phi}_1) &= n\big(\hat{a}_1^2\widehat{\mathrm{Var}}(\tilde{\phi}_p) + \hat{a}_2^2\widehat{\mathrm{Var}}(\tilde{\phi}_1) + 2\hat{a}_1^2\hat{a}_2^2\widehat{\mathrm{Cov}}(\tilde{\phi}_p, \tilde{\phi}_1)\big) \\ &= \frac{n(\hat{\sigma}_1^2\hat{\sigma}_p^2 - \hat{\sigma}_{p1}^2)}{\hat{\sigma}_1^2 + \hat{\sigma}_p^2 - 2\hat{\sigma}_{p1}} \xrightarrow{P} \frac{\sigma^2}{E(X_t^2)}. \end{aligned}$$

However, observe that $\hat{\sigma}_{p1} = \hat{\sigma}_1^2$ which implies that $\widehat{\mathrm{Var}}(a_1\tilde{\phi}_p + a_2\tilde{\phi}_1) = \widehat{\mathrm{Var}}(\tilde{\phi}_1)$. Thus, our choice of asymptotic estimators when $q = 1$ has the unintended consequence of causing our interval to be equivalent to that of the OLSE.

Herein lies the motive to go with center (1-2) in lieu of center (2-1). By replacing $\tilde{\phi}_1$ with $\hat{\rho}(1)$, we avoid this asymptotic equivalence, while preserving some of the desirable properties associated with (2-1). In the upcoming simulations, we also replace $\hat{\sigma}_1^2 = \widehat{\mathrm{Var}}(\tilde{\phi}_1)$ with

$$\widehat{\mathrm{Var}}(\hat{\rho}(1)) = \frac{\sigma^2}{\sum_{t=1}^{n} X_t^2},$$

but retain $\hat{\sigma}_{p1} = \widehat{\mathrm{Cov}}(\tilde{\phi}_p, \tilde{\phi}_1)$.

3. Simulations

We now look at the length and coverage capability of 95% confidence intervals for ϕ centered at the OLSE and $a_1\tilde{\phi}_p + a_2\hat{\rho}(1)$ for various $p \neq 1$. Each figure reflects 10,000 simulation runs of $n = 50$ independent observations with distribution $N(0, 1)$.

In Figure 3, top, we see that the $a_1\tilde{\phi}_0 + a_2\hat{\rho}(1)$ interval has at least as much coverage as the OLSE interval when (roughly) $|\phi| \leq 0.5$. The $a_1\tilde{\phi}_0 + a_2\hat{\rho}(1)$ interval is also slightly shorter over this same region. In Figure 3, bottom, p has increased to 2, but the coverage of the $a_1\tilde{\phi}_2 + a_2\hat{\rho}(1)$ interval has degenerated with no meaningful difference in interval lengths.

In Figure 4, p has increased to 3 (top two graphs) and 4 (middle row of graphs). The $a_1\tilde{\phi}_3 + a_2\hat{\rho}(1)$ and $a_1\tilde{\phi}_4 + a_2\hat{\rho}(1)$ intervals both return back to the performance level of the $a_1\tilde{\phi}_0 + a_2\hat{\rho}(1)$ interval, with (roughly) $|\phi| \leq 0.5$ again being the domain of interest.

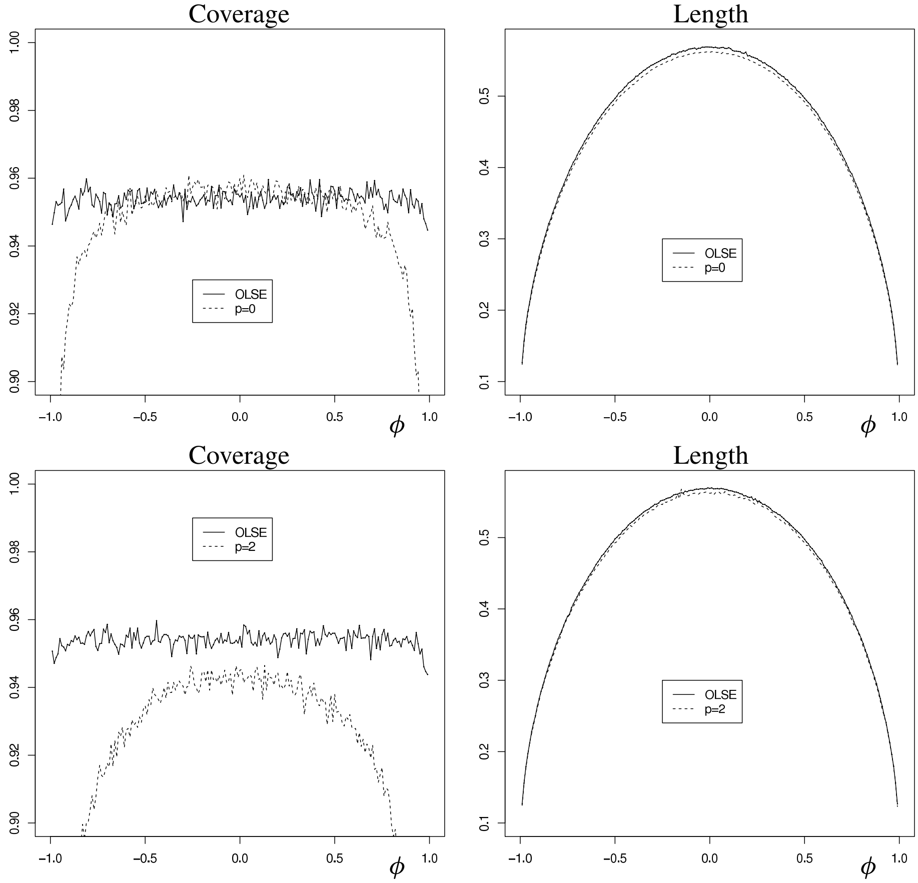


Figure 3. Top row: Empirical coverage capability (left) and length (right) of 95% confidence intervals for ϕ centered at the OLSE and $a_1\tilde{\phi}_0 + a_2\hat{\rho}(1)$ for $\phi \in (-1, 1)$. Bottom row: Same information, for $a_1\tilde{\phi}_2 + a_2\hat{\rho}(1)$.

In Figure 4, bottom, we create intervals whose centers are simply unweighted averages of the OLSE and the sample correlation coefficient. That is, the intervals' endpoints take the form

$$0.5\tilde{\phi}_1 + 0.5\hat{\rho}(1) \pm 1.96\sqrt{\widehat{\mathrm{Var}}\big(0.5\tilde{\phi}_1 + 0.5\hat{\rho}(1)\big)}.$$

Using the fact that $2\widehat{\mathrm{Cov}}(\tilde{\phi}_1, \hat{\rho}(1)) \approx \widehat{\mathrm{Var}}(\tilde{\phi}_1) + \widehat{\mathrm{Var}}(\hat{\rho}(1))$, this is approximately

$$0.5\tilde{\phi}_1 + 0.5\hat{\rho}(1) \pm 1.96\sqrt{0.5\big(\widehat{\mathrm{Var}}(\tilde{\phi}_1) + \widehat{\mathrm{Var}}(\hat{\rho}(1))\big)}.$$

There is no significant difference between the $0.5\tilde{\phi}_1 + 0.5\hat{\rho}(1)$ and OLSE intervals.

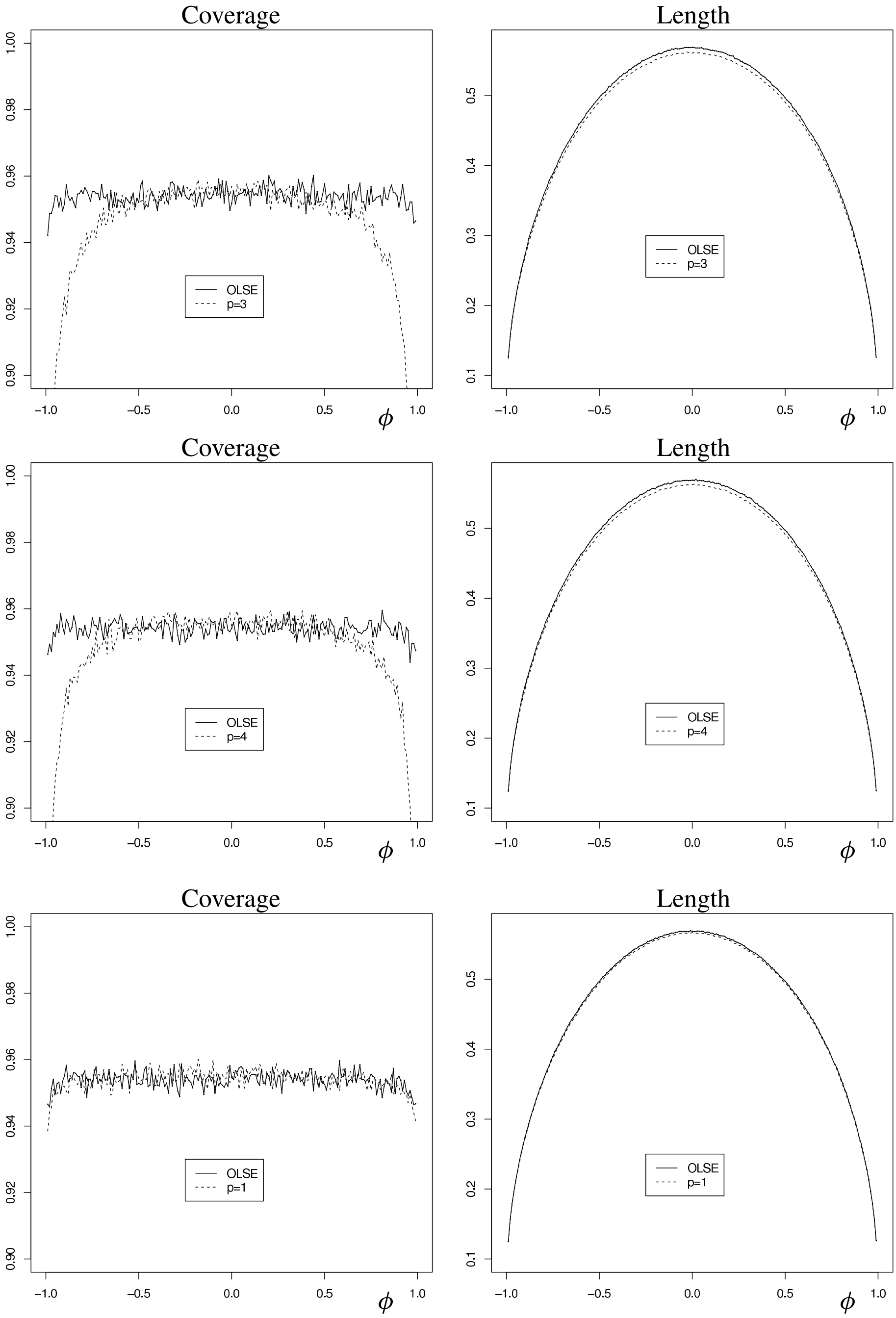


Figure 4. Top row: Empirical coverage capability (left) and length (right) of 95% confidence intervals for ϕ centered at the OLSE and $a_1\tilde{\phi}_3 + a_2\hat{\rho}(1)$ for $\phi \in (-1, 1)$. Middle row: Same, for $a_1\tilde{\phi}_4 + a_2\hat{\rho}(1)$. Bottom row: Same, for $0.5\tilde{\phi}_1 + 0.5\hat{\rho}(1)$.

4. Closing remarks

The performance of the confidence interval centered at $a_1\tilde{\phi}_p + a_2\hat{\rho}(1)$ presented in this paper is modest, but not unimportant. For parameter values (roughly) between -0.5 and 0.5, its coverage tends to be at least as good as that of the OLSE interval while having a slightly smaller margin of error. This interval also does not require a large sample size, which can be good for certain practical purposes.

For example, consider the daily stock prices for Exxon Mobil Corporation during the fall quarter of 2011 (i.e., September 23 to December 21). A reasonable model for this time series is an ARIMA(1, 1, 0), where $\{X_t\} \sim \text{ARIMA}(p, 1, q)$ implies $\{X_t - X_{t-1}\} \sim \text{ARMA}(p, q)$. Thus, if X_t stands for the price at time t and $Y_t = X_t - X_{t-1}$, it follows that $\{Y_t\} \sim \text{AR}(1)$ with estimated model $Y_t = -0.0444Y_{t-1} + \epsilon_t$. Both the $\{X_t\}$ and $\{Y_t\}$ processes are shown in Figure 5.

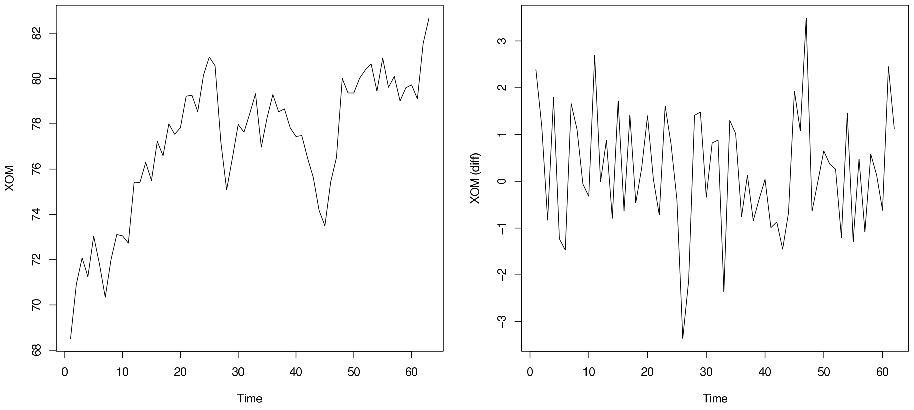


Figure 5. The original (left) and differenced (right) stock prices for Exxon Mobil (XOM) from 9/23/11 to 12/21/11. The sample sizes are 63 and 62, respectively.

If we supplement this model with 95% confidence intervals for ϕ, we get the following:

Center	Interval	Length
$\tilde{\phi}_1$	$(-0.2089871, 0.1719117)$	0.3808988
$\hat{\rho}(1)$	$(-0.2076522, 0.1710108)$	0.3786630
$0.5\tilde{\phi}_1 + 0.5\hat{\rho}(1)$	$(-0.2083205, 0.1714621)$	0.3797825
$a_1\tilde{\phi}_0 + a_2\hat{\rho}(1)$	$(-0.2036185, 0.1749948)$	0.3786133
$a_1\tilde{\phi}_2 + a_2\hat{\rho}(1)$	$(-0.2127095, 0.1658015)$	0.3785110
$a_1\tilde{\phi}_3 + a_2\hat{\rho}(1)$	$(-0.2100079, 0.1686098)$	0.3786177
$a_1\tilde{\phi}_4 + a_2\hat{\rho}(1)$	$(-0.2092005, 0.1694378)$	0.3786383

All seven intervals contain the point estimate $\hat{\phi} = -0.0444$, but the last four are slightly thinner than the first three.

One extension of the research presented in this paper would be to create confidence intervals centered around a linear combination of an arbitrary number of weighted least-squares estimators. For example, it can be shown that the variance of $a_1\tilde{\phi}_p + a_2\tilde{\phi}_q + a_3\tilde{\phi}_r$ is minimized when

$$a_1 = \frac{\sigma^*_{qr}(\sigma^2_r - \sigma_{pr}) + M(\sigma^2_r - \sigma_{qr})}{\sigma^*_{pr}\sigma^*_{qr} - M^2}, \quad a_2 = \frac{\sigma^*_{pr}(\sigma^2_r - \sigma_{qr}) + M(\sigma^2_r - \sigma_{pr})}{\sigma^*_{pr}\sigma^*_{qr} - M^2},$$

and $a_3 = 1 - a_1 - a_2$, where $\sigma^2_i = \operatorname{Var}(\tilde{\phi}_i)$, $\sigma_{ij} = \operatorname{Cov}(\tilde{\phi}_i, \tilde{\phi}_j)$, $\sigma^*_{ij} = \operatorname{Var}(\tilde{\phi}_i - \tilde{\phi}_j)$, and $M = \sigma_{pr} + \sigma_{qr} - \sigma_{pq} - \sigma^2_r$. However, once the number of estimators in the center goes beyond two, the work required to construct and analyze the interval may outweigh any benefits it would bestow.

Another (less tedious) extension would be to find a new sequence $\{a_{1,n}\}$ that converges to zero while yielding a linear combination of estimators with smaller MSE than the OLSE. This new combination would still have the same distributional limit as the OLSE and could then serve as the center for another competitive interval for ϕ. Specifically, if we simply set the standard error equal to the square root of the asymptotic variance of the OLSE, the resulting interval should have length equal to that of the OLSE, but with better coverage capability.

References

[Gallagher and Tunno 2008] C. Gallagher and F. Tunno, "A small sample confidence interval for autoregressive parameters", *J. Statist. Plann. Inference* **138**:12 (2008), 3858–3868. MR 2009m:62257 Zbl 1146.62065

[Phillips et al. 2004] P. C. B. Phillips, J. Y. Park, and Y. Chang, "Nonlinear instrumental variable estimation of an autoregression", *J. Econometrics* **118**:1-2 (2004), 219–246. MR 2004j:62186 Zbl 1033.62085

[Shumway and Stoffer 2006] R. H. Shumway and D. S. Stoffer, *Time series analysis and its applications*, 2nd ed., Springer, New York, 2006. MR 2007a:62003 Zbl 1096.62088

[So and Shin 1999] B. S. So and D. W. Shin, "Cauchy estimators for autoregressive processes with applications to unit root tests and confidence intervals", *Econometric Theory* **15**:2 (1999), 165–176. MR 2000f:62228 Zbl 0985.62072

Received: 2011-12-31 Revised: 2012-04-26 Accepted: 2012-05-10

ftunno@astate.edu
Department of Mathematics and Statistics,
Arkansas State University, P.O. Box 70,
State University, AK 72467, United States

ashton.erwin@smail.astate.edu
Department of Mathematics and Statistics,
Arkansas State University, P.O. Box 70,
State University, AK 72467, United States

INVOLVE 6:1 (2013)
dx.doi.org/10.2140/involve.2013.6.65

Knots in the canonical book representation of complete graphs

Dana Rowland and Andrea Politano

(Communicated by Joel Foisy)

We describe which knots can be obtained as cycles in the canonical book representation of the complete graph K_n, and we conjecture that the canonical book representation of K_n attains the least possible number of knotted cycles for any embedding of K_n. The canonical book representation of K_n contains a Hamiltonian cycle that is a composite knot if and only if $n \geq 12$. When p and q are relatively prime, the (p, q) torus knot is a Hamiltonian cycle in the canonical book representation of K_{2p+q}. For each knotted Hamiltonian cycle α in the canonical book representation of K_n, there are at least $2^k\binom{n+k}{k}$ Hamiltonian cycles that are ambient isotopic to α in the canonical book representation of K_{n+k}. Finally, we list the number and type of all nontrivial knots that occur as cycles in the canonical book representation of K_n for $n \leq 11$.

1. Introduction

In K_n, the *complete graph* on n vertices, every pair of distinct vertices is joined by an edge. An embedding or spatial representation of K_n is a particular way of joining the n vertices in three-dimensional space. Conway and Gordon [1983] proved that every spatial representation of K_6 contains at least one pair of linked triangles and every spatial representation of K_7 contains at least one knotted Hamiltonian cycle. They included examples of embeddings of K_6 and K_7 that were minimally linked or knotted — their embedding of K_6 contained exactly one pair of linked triangles and their embedding of K_7 contained exactly one knotted Hamiltonian cycle.

Otsuki [1996] introduced a family of spatial representations of K_n that generalized these examples of Conway and Gordon. Otsuki's spatial representation of K_n is an example of a book representation. Projections of book representations prevent complicated interactions between edges. In particular,

MSC2010: 05C10, 57M15, 57M25.
Keywords: spatial graph, intrinsically knotted, canonical book.
The authors gratefully acknowledge funding for this project received from the Merrimack College Paul E. Murray Fellowship.

- no edge crosses itself;
- a pair of edges cross at most once;
- if edge e_1 crosses over edge e_2 and edge e_2 crosses over edge e_3, then edge e_1 crosses over edge e_3.

Because book representations minimize the entanglement among the edges in a graph, they are good candidates for minimizing the linking and knotting in an embedding of a graph.

Otsuki called his family of embeddings the *canonical book representations* of K_n, which in this paper we denote by $\widetilde{K}_n$. He showed that any subcollection of m vertices of $\widetilde{K}_n$ induces a subgraph that is ambient isotopic to $\widetilde{K}_m$. Note this implies that, for $n \geq 6$, $\widetilde{K}_n$ contains exactly $\binom{n}{6}$ linked triangles, all of which are ambient isotopic to the Hopf link, and for $n \geq 7$, $\widetilde{K}_n$ contains exactly $\binom{n}{7}$ knotted 7-cycles, all of which are trefoil knots. Since Conway and Gordon's theorem implies that any spatial embedding of K_n contains at least $\binom{n}{6}$ linked triangles and at least $\binom{n}{7}$ nontrivially knotted 7-cycles, a canonical book representation is minimally linked and knotted in this sense.

In addition, Fleming and Mellor [2009] proved that a canonical book representation of K_n attains $14\binom{n}{7}$ triangle-square links, and showed this is the minimum possible for any embedding of K_n. They also conjectured that for any graph G there is some book representation that realizes the minimal number of nontrivial links possible in an embedding of G.

Similarly, the canonical book representation $\widetilde{K}_n$ is a candidate for the minimal number of knotted cycles in an embedding.

In this paper, we focus on which knots arise as knotted Hamiltonian cycles in the canonical book representation for $n > 7$. In Section 2 we review the definitions of book representations and Otsuki's canonical book representation, and show how knotted cycles in $\widetilde{K}_n$ are related to knotted cycles in $\widetilde{K}_{n+1}$. In Section 3 we show that $\widetilde{K}_n$ contains a (p, q) torus knot (or link) when $n \geq q+2p$. In Section 4 we examine composite knots in the canonical book representation, and in Section 5 we give a listing of all the knots that appear as cycles in $\widetilde{K}_n$ for $8 \leq n \leq 11$ and we conjecture about the ways in which $\widetilde{K}_n$ may achieve the minimal possible knotting complexity.

2. The canonical book representation of K_n

In this section, we review the right canonical book representation, as defined in [Otsuki 1996]. (In the right canonical book representation, the knotted 7-cycles are right-handed trefoil knots. The left canonical book representation is the mirror image of the one presented here.)

Definition 1. A *k-book* is a subset of $\mathbb{R}^3$ consisting of a line L and distinct half-planes $S_1, S_2, \dots, S_k$ with boundary L. The line L forms the spine of the book and the half-planes S_i form the pages, or sheets. We denote a k-book by B_k. Let G be a graph, and let $f : G \to B_k \subset \mathbb{R}^3$ be a tame embedding of G. We say that the spatial representation $f(G)$ is a k-book representation of G if:

(1) each vertex of $f(G)$ is on the line L;

(2) each edge of $f(G)$ is contained in exactly one sheet S_i.

If $\widetilde{G}$ is a k-book representation of G, then $\widetilde{G}$ can be deformed by an ambient isotopy so that the vertices lie on a circle C and the edges are chords on k internally disjoint topological disks, all of which have C as their boundary. For the remainder of this paper, we will treat the sheets S_i for $1 \le i \le k$ as topological disks. In a projection of the embedding onto the plane containing C, we assume that the sheets are labeled so that sheet S_i is "above" sheet S_j if $i < j$. A k-book embedding is determined, up to ambient isotopy, by specifying which edges are in which sheet.

The *sheet-number* of a graph G is the smallest possible k for which G has a k-book representation.

For K_n, the sheet-number is $\lceil n/2 \rceil$, the smallest integer greater than or equal to $n/2$; see [Bernhart and Kainen 1979] or [Kobayashi 1992] for proofs. Otsuki's right canonical book representation, $\widetilde{K}_n$, provides an example of a minimal-sheet book embedding of K_n [Otsuki 1996].

To describe $\widetilde{K}_n$, it suffices to list which sheet contains each edge. Label the n vertices with the integers 1 through n.

Case 1: The number of vertices is even. When $n = 2m$, there are m sheets and each of the sheets $S_1, S_2, \dots, S_m$ contains $2m-1$ edges. Sheet S_i contains the edges joining vertex i to vertex $i+j$, for $1 \le j \le m$, and the edges joining vertex $i+m$ to $(i+m+j) \bmod 2m$, for $1 \le j \le m-1$. See Figure 1.

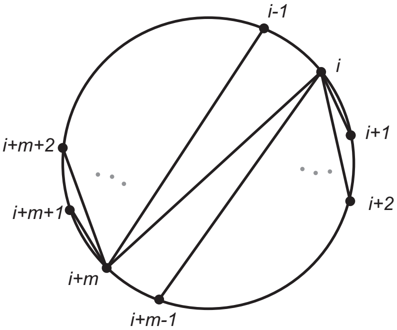


Figure 1. Sheet S_i in the canonical book representation of K_{2m}.

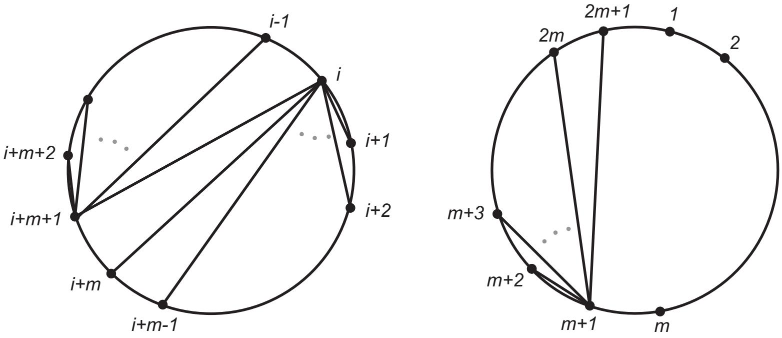


Figure 2. Sheets S_i $(i \leq m)$ and S_{m+1} in the canonical book representation of K_{2m+1}.

Alternatively, if we are given an edge joining vertex i to vertex j, we can determine which sheet the edge is in:

Lemma 2. *Let $n = 2m$ and let (i, j), with $i < j$, be the edge joining vertices i and j in the projection of $\widetilde{K}_n$. Then we can determine which sheet contains (i, j):*

- *If $i \leq m$ and $j - i \leq m$, then (i, j) is in S_i.*
- *If $i \leq m$ and $j - i \geq m+1$, then (i, j) is in S_{j-m}.*
- *If $i \geq m+1$, then (i, j) is in S_{i-m}.*

Furthermore, suppose the edge (i, j) crosses the edge (k, l) in the projection. We may assume without loss of generality that $1 \leq i < k < j < l \leq 2m$. Then edge (k, l) is on top of edge (i, j) if and only if $i \leq m$ and $k \geq m+1$.

Case 2: The number of vertices is odd. When $n = 2m+1$, the sheets S_1, S_2, $\ldots$, S_m each contain $2m$ edges and sheet S_{m+1} is a "half-sheet" containing m edges. For each $1 \leq i \leq m$, sheet S_i contains the edges joining vertex i to vertex $i+j$, and the edges joining vertex $i+m+1$ to $(i+m+j+1) \bmod (2m+1)$, for $1 \leq j \leq m$. Sheet S_{m+1} contains the edges joining vertex $m+1$ to vertex $m+1+j$, for $1 \leq j \leq m$. See Figure 2.

If we are given an edge joining vertex i to vertex j, we can determine which sheet the edge is in:

Lemma 3. *Let $n = 2m+1$, and let (i, j), with $i < j$, be the edge joining vertices i and j in the projection of $\widetilde{K}_n$. Then we can determine which sheet contains (i, j).*

- *If $i \leq m+1$ and $j - i \leq m+1$, then the edge is in S_i.*
- *If $i \leq m+1$ and $j - i \geq m+2$, then the edge is in S_{j-m-1}.*
- *If $i \geq m+2$, then the edge is in S_{i-m-1}.*

Furthermore, suppose the edge (i, j) *crosses the edge* (k, l) *in the projection. We may assume without loss of generality that* $1 \leq i < k < j < l \leq 2m+1$. *Then edge* (k, l) *is on top of edge* (i, j) *if and only if* $i \leq m+1$ *and* $k \geq m+2$.

Otsuki [1996] proved that the canonical book representation has the property that any subgraph induced by a subcollection of vertices is ambient isotopic to a canonical book representation. In particular, we have the following:

Proposition 4. *Let* $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)$ *be an n-cycle through the vertices* $1, \dots, n$ *in* $\widetilde{K}_N$. *Then the n-cycle* $(\alpha_1, \alpha_2, \dots, \alpha_n)$ *through the vertices* $1, \dots, n$ *in* $\widetilde{K}_{N+1}$ *is ambient isotopic to* α.

Proof. Let (i, j) and (k, l) be edges of the cycle α in $\widetilde{K}_N$, labeled so that $i < j$, $k < l$, and $i < k$. Two edges cross in the projection of $\widetilde{K}_N$ if and only if they cross in the projection of $\widetilde{K}_{N+1}$, which occurs if and only if $i < k < j < l$.

First, consider the case $N = 2m+1$. We can use Lemmas 2 and 3 to verify that:

(1) If $i < k \leq m+1$ or if $m+2 \leq i < k$, then (i, j) crosses over (k, l) in both $\widetilde{K}_N$ and $\widetilde{K}_{N+1}$.

(2) If $i \leq m+1$ and $k \geq m+2$, then (k, l) crosses over (i, j) in both $\widetilde{K}_N$ and $\widetilde{K}_{N+1}$.

Since there are no crossing changes between edges, the cycle $(\alpha_1, \alpha_2, \dots, \alpha_n)$ represents the same knot in both $\widetilde{K}_{2m+1}$ and $\widetilde{K}_{2m+2}$.

Now suppose that $N = 2m$. Using Lemmas 2 and 3 we observe that:

(1) If $i < k \leq m$ or if $m+2 \leq i < k$, then (i, j) crosses over (k, l) in both $\widetilde{K}_N$ and $\widetilde{K}_{N+1}$.

(2) If $i \leq m$ and $k \geq m+2$, then (k, l) crosses over (i, j) in both $\widetilde{K}_N$ and $\widetilde{K}_{N+1}$.

(3) If $i = m+1$ or if $k = m+1$ then a crossing change occurs between edges (i, j) and (k, l) when moving from $\widetilde{K}_N$ to $\widetilde{K}_{N+1}$.

Notice that if $i = m+1$ (or $k = m+1$), then (i, j) (or, respectively, (k, l)) is in the top sheet in $\widetilde{K}_N$ and the bottom sheet in $\widetilde{K}_{N+1}$. An edge in the bottom sheet of $\widetilde{K}_{N+1}$ is under *all* other edges and can be moved by an ambient isotopy so that it lies over all other edges. Thus, the only crossing changes that occur do not change the knot type, and the cycle $(\alpha_1, \alpha_2, \dots, \alpha_n)$ represents the same knot in both $\widetilde{K}_{2m}$ and $\widetilde{K}_{2m+1}$. □

We also know that, if a Hamiltonian cycle with a certain knot type appears in $\widetilde{K}_n$, then $\widetilde{K}_N$ must contain a Hamiltonian cycle with the same knot type for any $N > n$. The following theorem indicates one way to find such a cycle:

Theorem 5. *Let* $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)$ *denote an n-cycle through all the vertices* $1, 2, \dots, n+1$ *except* $i+1$ *in* $\widetilde{K}_{n+1}$. *Suppose that* $\alpha_k = i$ *and* $\alpha_{k+1} = j$. *Then the Hamiltonian cycle* $(\alpha_1, \dots, \alpha_k, i+1, \alpha_{k+1}, \dots, \alpha_n)$ *is ambient isotopic to* α.

Proof. It suffices to check that in $\widetilde{K}_n$ any edge $(i,\, j)$ is at most one sheet apart from the edge $(i+1,\, j)$. This will guarantee that the edge $(i,\, j)$ can be moved to the path $(i,\, i+1,\, j)$ by an ambient isotopy, since if the edge $(i+1,\, j)$ is one sheet level above or one below the edge $(i,\, j)$ then the path $(i,\, i+1,\, j)$ crosses the same edges as the edge $(i,\, j)$ and in the same manner. In other words, no edge can pass through the triangle formed by the cycle $(i,\, i+1,\, j)$. Note that the top and bottom sheets can also be considered consecutive, since an edge on the bottom sheet can be deformed by ambient isotopy to be on top of all the sheets, and vice versa.

We will verify that edges $(i,\, j)$ and $(i+1,\, j)$ are at most one sheet apart when $i < j$. The proof for when $i > j$ is similar, and is left to the reader. There are six cases to check.

Case 1: *$i \geq m+1$ and there are an even number of vertices.* Refer to Lemma 2. The edge $(i,\, j)$ is in S_{i-m}. The edge $(i+1,\, j)$ is in S_{i+1-m}. Therefore, the edges are in consecutive sheets.

Case 2: *$i \geq m+2$ and there are an odd number of vertices.* Refer to Lemma 3. The edge $(i,\, j)$ is in S_{i-m-1}. The edge $(i+1,\, j)$ is in $S_{i+1-m-1}$ which equals S_{i-m}. Therefore, the edges are in consecutive sheets.

Case 3: *$i \leq m$, $j-i \leq m$ and there are an even number of vertices.* Refer to Lemma 2. The edge $(i,\, j)$ is in S_i. There are two possibilities for the sheet level of the edge $(i+1,\, j)$. First, if $i < m$, then $i+1 \leq m$, and

$$j-(i+1) = j-i-1 \leq m-1 \leq m.$$

Therefore, edge $(i+1,\, j)$ is in S_{i+1}. Second, if $i = m$, then $i+1 \geq m+1$ so edge $(i+1,\, j)$ is in $S_{i+1-m} = S_{m+1-m} = S_1$. This would not change the knot type because the edge $(i,\, j)$ was in the very last sheet and this edge is in the very first sheet. In both cases, the edges are in consecutive sheets.

Case 4: *$i \leq m+1$, $j-i \leq m+1$ and there are an odd number of vertices.* Refer to Lemma 3. The edge $(i,\, j)$ is in S_i. Again, there are two possibilities for the sheet level of edge $(i+1,\, j)$. First, if $i < m+1$, then $i+1 \leq m+1$, and so $j-(i+1) = j-i-1 \leq m \leq m+1$ meaning this edge is found in S_{i+1}. This is one sheet level below the original edge. Second, if $i = m+1$, then $i+1 \geq m+2$. The edge $(i+1,\, j)$ is therefore in $S_{i+1-m-1} = S_{m+1+1-m-1} = S_1$, the very first sheet. As in case 3, this means that the knot type remains unchanged.

Case 5: *$i \leq m$, $j-i \geq m+1$ and there are an even number of vertices.* Refer to Lemma 2. The edge $(i,\, j)$ is in S_{j-m}. Note that, if $i = m$, then $j \geq 2m+1$, which is impossible, since there are only $2m$ vertices. That leaves two possibilities for the sheet level of edge $(i+1,\, j)$. First, if $i < m$ and $j > i+m+1$, then $i+1 \leq m$ and $j-(i+1) \geq m+1$. This forces the edge to be in S_{j-m}, and so both edges are in the same sheet. Second, if $i < m$ and $j = i+m+1$, then $i+1 \leq m$ and $j-(i+1) = m$. This means the edge is in S_{i+1}, which is equivalent to S_{j-m} because $j = i+m+1$. Again, both edges are in the same sheet.

Case 6: $i \leq m+1$, $j-i \geq m+2$ *and there are an odd number of vertices.* Refer to Lemma 3. The edge (i, j) is in S_{j-m-1}. Note that, if $i = m+1$, then $j \geq 2m+3$, which is impossible, so we can assume $i < m+1$. There are two possibilities for the sheet level of edge $(i+1, j)$. First, if $i < m+1$ and $j > i+m+2$, then $i+1 \leq m+1$ and $j-(i+1) \geq m+2$. This means the edge is in S_{j-m-1}. Second, if $i < m+1$ and $j = i+m+2$, then $i+1 \leq m+1$ and $j-(i+1) = m+1$. Again, the edge is in $S_{i+1} = S_{j-m-1}$. Both edges are in the same sheet. □

Corollary 6. *Suppose α is a Hamiltonian cycle in $\widetilde{K}_n$ with the property that no edge of α joins consecutively labeled vertices. Let $N = n+k$ for $k \geq 0$. Then $\widetilde{K}_N$ contains at least $2^k\binom{N}{k}$ Hamiltonian cycles that are ambient isotopic to α.*

Proof. The subgraph induced by any n vertices of $\widetilde{K}_N$ is ambient isotopic to $\widetilde{K}_n$, so there are at least $\binom{N}{n}$ n-cycles in $\widetilde{K}_N$ that are ambient isotopic to α. These cycles share the property that no edge joins consecutive vertices. Let $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ be such an n-cycle. Choose the smallest integer j such that $j = \alpha_i$ for some $1 \leq i \leq n$ but $j+1 \neq \alpha_l$ for any $1 \leq l \leq n$. (Note: if $j = N$, we interpret $j+1$ as 1.) By the proof of Theorem 5, the cycles $(\alpha_1, \ldots, \alpha_{i-1}, j+1, \alpha_i, \ldots, \alpha_n)$ and $(\alpha_1, \ldots, \alpha_i, j+1, \alpha_{i+1}, \ldots, \alpha_n)$ are both ambient isotopic to α. Repeat this step until all vertices in $\widetilde{K}_N$ are used. This gives 2^k ways to extend each n-cycle, which produces $2^k\binom{N}{k}$ distinct Hamiltonian cycles that are ambient isotopic to α, as claimed. □

This immediately implies that there are at least $2^{N-7}\binom{N}{7}$ Hamiltonian cycles that are trefoil knots in $\widetilde{K}_N$ when $N \geq 7$. This bound is not sharp, however, as shown in the table in Section 5.

3. Torus knots in the canonical book embedding

Recall that a *torus link* is a knot or link that can be embedded on the standard (unknotted) torus in $\mathbb{R}^3$. A (p, q) torus link can be deformed so that it crosses every meridian (a closed curve that bounds a topological disk that is "inside" the torus) of the torus p times and every longitude (a closed curve that bounds a topological disk that is "outside" the torus) of the torus q times. When p and q are relatively prime, the link is a knot. See [Adams 1994, Section 5.1] for a general description of (p, q) torus knots and links.

A (p, q) torus knot can also be described as the closure of a braid on p strands, with braid word $(\sigma_1\sigma_2 \cdots \sigma_{p-1})^q$. Recall that σ_i denotes that the i-th strand of the braid crosses over the $(i+1)$-st strand of the braid, and equivalent braid words can

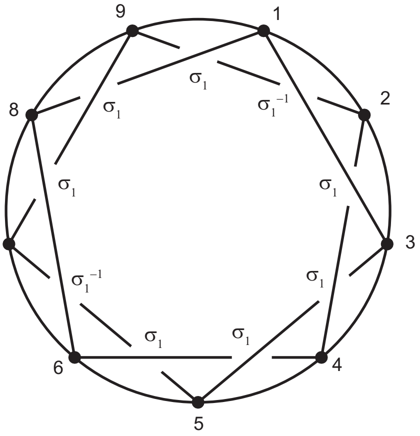


Figure 3. The cycle $(1,3,5,7,9,2,4,6,8)$ in $\widetilde{K}_9$ is the knot 5_1. It can be described by the braid word $\sigma_1^4\sigma_1^{-1}\sigma_1^3\sigma_1^{-1} = \sigma_1^5$.

be obtained using the braid relations $\sigma_i\sigma_{i+1}\sigma_i = \sigma_{i+1}\sigma_i\sigma_{i+1}$ and $\sigma_i\sigma_j = \sigma_j\sigma_i$ when $|i-j| \geq 2$. See [Adams 1994, Section 5.4] or [Birman 1975] for references on braids.

Consider the Hamiltonian cycle $(1,3,5,\dots,2m+1,2,4,\dots,2m)$ in $\widetilde{K}_{2m+1}$. This cycle forms the closure of a 2-strand braid with $2m+1$ crossings. See Figure 3. For each $i \leq 2m-2$, we know from Lemma 3 that the edge $(i, i+2)$ crosses over the edge $(i+1, i+3)$ except when $i = m+1$. Edge $(2m-1, 2m+1)$ crosses over edge $(2m, 1)$, edge $(2m, 1)$ crosses over edge $(2m+1, 2)$, and edge $(2m+1, 2)$ crosses under edge $(1, 3)$. The resulting braid word is

$$\sigma^m\sigma^{-1}\sigma^{(2m-2)-(m+1)}\sigma\sigma\sigma^{-1} = \sigma^{2m-3}.$$

Therefore, we see that $\widetilde{K}_{2m+1}$ contains a $(2, 2m-3)$ torus knot. When q is odd, $\widetilde{K}_n$ contains a $(2, q)$ torus knot as one of its Hamiltonian cycles for all $n \geq q+4$. (Note: when q is even, the same argument shows that $\widetilde{K}_n$ contains a $(2, q)$ torus link.)

Suppose $n > 6$ is not a multiple of 3, and consider the Hamiltonian cycle $(1,4,7,\dots)$ in $\widetilde{K}_n$. This cycle forms the closure of a 3-strand braid with word $\prod_{i=1}^n(\sigma_1^{\delta_1(i)}\sigma_2^{\delta_2(i)})$ where $\delta_1(i) = 1$ if the edge $(i, i+3)$ is over the edge $(i+1, i+4)$ and -1 otherwise, and $\delta_2(i) = 1$ if the edge $(i, i+3)$ is over the edge $(i+2, i+5)$ and -1 otherwise. (The vertex labels are to be taken modulo n.)

Suppose $n = 2m$ is even. (The case for when n is odd is similar, and omitted.) Then Lemma 2 implies that $\delta_1(i) = -1$ if and only if i is m or n, and $\delta_2(i) = -1$ if and only if i is one of $m-1, m, n-1$ or n. The braid word becomes

$$(\sigma_1\sigma_2)^{m-2}\sigma_1\sigma_2^{-1}\sigma_1^{-1}\sigma_2^{-1}(\sigma_1\sigma_2)^{m-2}\sigma_1\sigma_2^{-1}\sigma_1^{-1}\sigma_2^{-1}.$$

Since the braid relations imply that

$$\sigma_2^{-1}\sigma_1^{-1}\sigma_2^{-1} = \sigma_1^{-1}\sigma_2^{-1}\sigma_1^{-1},$$

we see that $\sigma_1\sigma_2^{-1}\sigma_1^{-1}\sigma_2^{-1}\sigma_1\sigma_2$ is the identity. Therefore the braid word can be reduced to $(\sigma_1\sigma_2)^{n-6}$. This shows that $\widetilde{K}_n$ contains a $(3, n-6)$ torus knot. For any $n \geq q+6$, the spatial representation $\widetilde{K}_n$ contains a $(3, q)$ torus knot (or link, if q is a multiple of 3).

An extension of this argument leads to the following theorem:

Theorem 7. *Let p, q, and n be positive integers such that $p \leq q$ and $n \geq q+2p$. Then the canonical book representation of K_n contains a (p, q) torus knot (or link).*

Proof. By Theorem 5, it suffices to prove this theorem when $n = 2p+q$. Consider the knot or link in $\widetilde{K}_{2p+q}$ consisting of all edges of the form $(i, i+p)$ for $1 \leq i \leq n$, where the vertex labels are taken modulo n. This knot or link can be described as a braid on p strands with braid word

$$w = \prod_{i=1}^{n} \big[\sigma_1^{\delta_1(i)}\sigma_2^{\delta_2(i)} \cdots \sigma_{p-1}^{\delta_{p-1}(i)}\big],$$

where

$$\delta_j(i) = \begin{cases} 1 & \text{if edge } (i, i+p) \text{ is over edge } (i+j, i+j+p), \\ -1 & \text{otherwise.} \end{cases}$$

We will use Lemma 2 to prove the case when n is even. The case for n odd is left to the reader.

Suppose $n = 2m$. By Lemma 2,

$$\delta_j(i) = \begin{cases} 1 & \text{if } 1 \leq i \leq m-j, \\ -1 & \text{if } m-j+1 \leq i \leq m, \\ 1 & \text{if } m+1 \leq i \leq n-j, \\ -1 & \text{if } n-j+1 \leq i \leq n, \end{cases}$$

and this allows us to express w as

$$w = \big[(\sigma_1\sigma_2\cdots\sigma_{p-1})^{m-p+1}(\sigma_1\sigma_2\cdots\sigma_{p-2})\sigma_{p-1}^{-1} \\ \cdot (\sigma_1\sigma_2\cdots\sigma_{p-3})\sigma_{p-2}^{-1}\sigma_{p-1}^{-1}\cdots\sigma_1^{-1}\sigma_2^{-1}\cdots\sigma_{p-1}^{-1}\big]^2.$$

Next, observe that

$$(\sigma_1\sigma_2\cdots\sigma_{p-1})(\sigma_1\sigma_2\cdots\sigma_{p-2})\sigma_{p-1}^{-1}(\sigma_1\sigma_2\cdots\sigma_{p-3})\sigma_{p-2}^{-1}\sigma_{p-1}^{-1}\cdots\sigma_1^{-1}\sigma_2^{-1}\cdots\sigma_{p-1}^{-1}$$

is equivalent to the identity. For example, when $p = 4$, we can use the braid relations

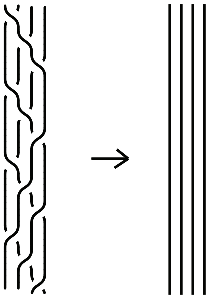


Figure 4. The braid word $(\sigma_1\sigma_2\sigma_3)(\sigma_1\sigma_2\sigma_3^{-1})(\sigma_1\sigma_2^{-1}\sigma_3^{-1})\cdot(\sigma_1^{-1}\sigma_2^{-1}\sigma_3^{-1})$ is equivalent to the identity.

to obtain

$$\begin{aligned}(\sigma_1\sigma_2\sigma_3)&(\sigma_1\sigma_2\sigma_3^{-1})(\sigma_1\sigma_2^{-1}\sigma_3^{-1})(\sigma_1^{-1}\sigma_2^{-1}\sigma_3^{-1})\\ &=(\sigma_1\sigma_2\sigma_1)(\sigma_3\sigma_2\sigma_3^{-1})(\sigma_1\sigma_2^{-1}\sigma_1^{-1})(\sigma_3^{-1}\sigma_2^{-1}\sigma_3^{-1})\\ &=(\sigma_2\sigma_1\sigma_2)(\sigma_2^{-1}\sigma_3\sigma_2)(\sigma_2^{-1}\sigma_1^{-1}\sigma_2)(\sigma_2^{-1}\sigma_3^{-1}\sigma_2^{-1})\\ &=\sigma_2\sigma_1\sigma_3\sigma_1^{-1}\sigma_3^{-1}\sigma_2^{-1}\\ &=\sigma_2\sigma_3\sigma_1\sigma_1^{-1}\sigma_3^{-1}\sigma_2^{-1}\\ &=1.\end{aligned}$$

(See Figure 4.) This implies that the braid word simplifies to

$$w=(\sigma_1\sigma_2\cdots\sigma_{p-1})^{2m-2p}=(\sigma_1\sigma_2\cdots\sigma_{p-1})^q,$$

so we obtain a (p,q) torus link as claimed. □

4. Composite knots in the canonical book embedding

In this section we prove that the canonical book embedding of K_n contains a composite knot for all $n\geq 12$. We also show that, if we choose any two knotted Hamiltonian cycles contained in $\widetilde{K}_p$ and $\widetilde{K}_q$, respectively, their composite will be a Hamiltonian cycle in $\widetilde{K}_{p+q+1}$.

Theorem 8. *Let $n\geq 14$. Then the cycle*

$$(1,3,5,7,9,11,13,8,10,12,14,15,16,\ldots,n,2,4,6)$$

in the canonical book representation of K_n is the composite of two trefoils.

Proof. We can find a composite knot in $\widetilde{K}_{14}$ by first finding trefoils in two disjoint subgraphs. The first subgraph is induced by vertices 1 through 7. The second subgraph is induced by vertices 8 through 14.

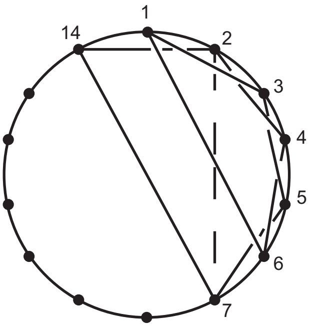


Figure 5. Cycle $(1, 3, 5, 7, 14, 2, 4, 6)$ is ambient isotopic to cycle $(1, 3, 5, 7, 2, 4, 6)$ since the edge $(2, 7)$, shown as a dashed line, can be replaced by the path $(2, 14, 7)$.

Any set of seven vertices of $\widetilde{K}_{14}$ induces a graph that is ambient isotopic to the canonical book representation of K_7. In $\widetilde{K}_7$ there is exactly one trefoil knot. Therefore, there is exactly one trefoil in each subgraph of $\widetilde{K}_{14}$ induced by seven vertices. The first subgraph has a trefoil in the cycle $(1, 3, 5, 7, 2, 4, 6)$. Notice that this cycle is ambient isotopic to the cycle $(1, 3, 5, 7, 14, 2, 4, 6)$ in $\widetilde{K}_{14}$. See Figure 5.

These cycles are ambient isotopic because the only edges of the cycles which intersect the path $(2, 14, 7)$ and the edge $(2, 7)$ are edges $(1, 3)$ and $(1, 6)$. Both of these edges lie in S_1 meaning that any path or edge that crosses those two edges will fall in a lower sheet. This means that the edge $(2, 7)$ can be replaced with the path $(2, 14, 7)$ without changing the knot type.

The second subgraph (induced by vertices 8 through 14) has a trefoil in the cycle

$$(8, 10, 12, 14, 9, 11, 13).$$

Notice that this cycle is ambient isotopic to the cycle

$$(8, 10, 12, 14, 7, 9, 11, 13).$$

See Figure 6.

These cycles are ambient isotopic because both the path $(9, 7, 14)$ and the edge $(9, 14)$ cross edges $(8, 10)$ and $(8, 13)$, which are both in S_1. Therefore, any edge or path that crosses these two edges will still remain under them, meaning that the path $(9, 7, 14)$ can be replaced with the edge $(9, 14)$ without affecting the knot type.

Place the two cycles on $\widetilde{K}_{14}$. When these two cycles are layered they share the edge $(7, 14)$. Removing edge $(7, 14)$ (which is the shared edge that has no crossings), will create the composite of the two trefoils. See Figure 7. The cycle

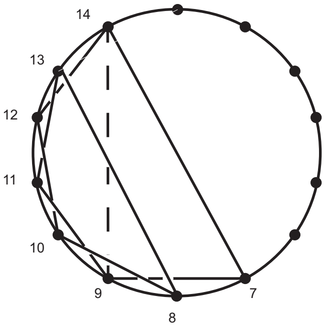


Figure 6. Cycle (8, 10, 12, 14, 7, 9, 11, 13) is ambient isotopic to cycle (8, 10, 12, 14, 9, 11, 13) since the edge (9, 14), shown as a dashed line, can be replaced by the path (9, 7, 14).

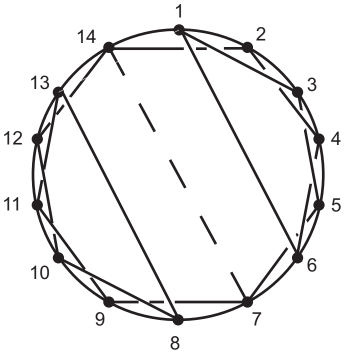


Figure 7. Composite knot in K_{14}. The dashed line is the edge removed from both factor knots to form the composite.

with the composite knot in $\widetilde{K}_{14}$ is

$$(1, 3, 5, 7, 9, 11, 13, 8, 10, 12, 14, 2, 4, 6).$$

For $n > 14$, the fact that the cycle

$$(1, 3, 5, 7, 9, 11, 13, 8, 10, 12, 14, 15, 16, \ldots, n, 2, 4, 6)$$

in the canonical book representation of K_n is the composite of two trefoils follows immediately from Theorem 5. □

We can improve this result by finding a composite knot in $\widetilde{K}_{13}$. Consider two subgraphs of $\widetilde{K}_{13}$. Let the first subgraph of $\widetilde{K}_{13}$ be induced by vertices 1 through 7, and let the second subgraph be induced by vertices 7 through 13. Refer to Figure 8.

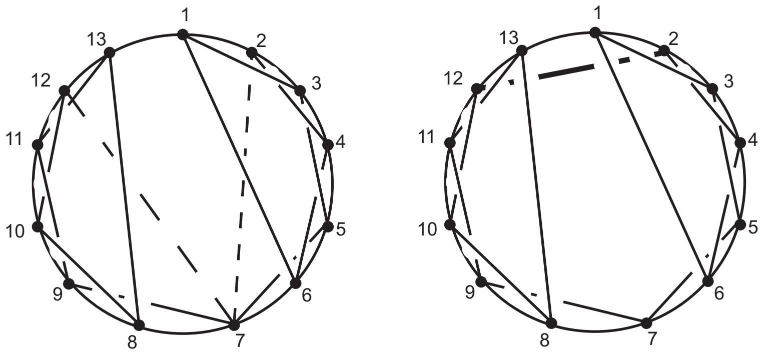


Figure 8. Composite knot in K_{13}. On the left, the two cycles are layered with the dashed lines representing the edges to be removed. On the right is the composite formed by adding the bold edge $(2, 7)$.

Since each subgraph is ambient isotopic to $\widetilde{K}_7$, each subgraph contains exactly one trefoil knot. The first subgraph has a trefoil knot in the cycle

$$(1, 3, 5, 7, 2, 4, 6).$$

The second subgraph has a trefoil in the cycle

$$(7, 9, 11, 13, 8, 10, 12).$$

Place these two cycles together in $\widetilde{K}_{13}$. Notice that 4 edges meet at vertex 7. Connect the knots by joining edges $(5, 7)$ and $(7, 9)$ and replacing the path $(2, 7, 12)$ with the edge $(2, 12)$. This results in the cycle

$$(1, 3, 5, 7, 9, 11, 13, 8, 10, 12, 2, 4, 6).$$

Note that edge $(2, 12)$ crosses edges $(1, 3)$, $(1, 6)$, $(8, 13)$ and $(11, 13)$. Edge $(2, 12)$ is in sheet five, edges $(1, 3)$, $(1, 6)$, and $(8, 13)$ are in sheet one, and lastly, edge $(11, 13)$ is in sheet four. This means that edge $(2, 12)$ crosses completely under all edges. Since edges $(2, 7)$ and $(7, 12)$ also cross under all the edges that edge $(2, 12)$ crosses, replacing the path $(2, 7, 12)$ by the edge $(2, 12)$ forms a composite of the two trefoil knots in $\widetilde{K}_{13}$.

The smallest $\widetilde{K}_n$ that a composite knot can be found in is $\widetilde{K}_{12}$; refer to Figure 9. To find this composite, once again we consider two subgraphs of $\widetilde{K}_{12}$ where the first subgraph is induced by the first 7 vertices and the second subgraph is induced by the last 7 vertices in the embedding of K_{12}.

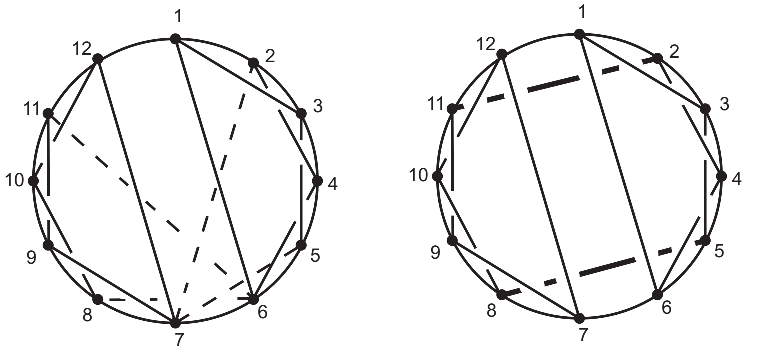


Figure 9. Composite knot in K_{12}. On the left, two knotted trefoils are shown. The dashed edges are the ones that will be replaced. On the right is a cycle which is the composite of the two trefoils.

Each subgraph contains exactly one Hamiltonian cycle that is a trefoil knot. The first subgraph has a trefoil in the cycle

$$(1, 3, 5, 7, 2, 4, 6).$$

The second subgraph has a trefoil in the cycle

$$(6, 8, 10, 12, 7, 9, 11).$$

Place these two cycles with the trefoil knots on K_{12}. Notice that there are 4 edges that meet at vertex 6 and vertex 7. Removing the paths (8, 6, 11) and (2, 7, 5) and adding edges (2, 11) and (5, 8) forms a composite knot. This cycle is

$$(1, 3, 5, 8, 10, 12, 7, 9, 11, 2, 4, 6).$$

Up to now we have shown how to find composites of trefoil knots. A similar method can be used to find other composite knots.

Theorem 9. *Let α be a Hamiltonian cycle in the canonical book representation of K_p and let β be a Hamiltonian cycle in the canonical book representation of K_q. Then $\alpha \# \beta$ is a Hamiltonian cycle in the canonical book representation of K_{p+q+1}.*

Proof. Without loss of generality, we assume that $p \leq q$. Consider the subgraph of $\widetilde{K}_{p+q+1}$ induced by vertices 1 through p, and let $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_p)$. Because we are dealing with a book representation, we know that there exists some edge (α_i, α_{i+1}) that is in a lower sheet than all other edges in the cycle. Change the orientation of the cycle if necessary so that $\alpha_i < \alpha_{i+1}$. Edges $(\alpha_i, p+q+1)$ and $(\alpha_{i+1}, p+q)$ are also in lower sheets than any of the edges of α, so the cycle $\tilde{\alpha} = (\alpha_1, \ldots, \alpha_i, p+q+1, p+q, \alpha_{i+1}, \alpha_{i+2}, \ldots, \alpha_p)$ is ambient isotopic to α.

n	knotted Hamiltonian cycles			out of
7	1 3_1 knot			1
8	21 3_1 knots			29
9	342 3_1 knots	9 4_1 knots	1 5_1 knot	577
10	5090 3_1 knots	245 4_1 knots	50 5_1 knots	9991
	20 5_2 knots	1 8_{19} knot		
11	74855 3_1 knots	5335 4_1 knots	1375 5_1 knots	165102
	836 5_2 knots	11 6_1 knots	11 6_2 knots	
	1 7_1 knot	56 8_{19} knot	1 10_{124} knot	

Table 1. Knotted Hamiltonian cycles and total number of knotted cycles (rightmost column) in the canonical book embedding of K_n, for $n \leq 11$.

Similarly, we can find a cycle $(\beta_1, \beta_2, \dots, \beta_q)$ that is ambient isotopic to β using vertices $p+1$ through $p+q$. We know such a cycle exists, because the subgraph induced by any q vertices is ambient isotopic to the canonical book representation of K_q. Suppose that $\beta_j = p+q$, and that the cycle is oriented so that $\beta_{j-1} < \beta_{j+1}$. Using the same argument used in the proof of Theorem 5, we can extend β to an ambient isotopic cycle $\tilde{\beta} = (\beta_1, \beta_2, \dots, \beta_{j-1}, p+q+1, p+q, \beta_{j+1}, \dots, \beta_q)$ that contains the edge $(p+q, p+q+1)$.

The cycles $\tilde{\alpha}$ and $\tilde{\beta}$ meet along the edge $(p+q, p+q+1)$. The only crossing between disjoint edges in the two cycles is a single crossing between the edges $(\alpha_{i+1}, p+q)$ and $(\beta_{j-1}, p+q+1)$. Since this crossing can be eliminated by flipping one of the components $\tilde{\alpha}$ or $\tilde{\beta}$, the cycle

$$(\alpha_1, \dots, \alpha_i, p+q+1, \beta_{j-1}, \beta_{j-2}, \dots, \beta_1, \beta_q, \beta_{q-1}, \dots, \beta_{j+1}, p+q, \alpha_{i+1}, \dots, \alpha_p)$$

is ambient isotopic to the composite knot $\alpha \# \beta$. □

5. Conclusion

In Table 1, we have identified the knotted Hamiltonian cycles and the total number of knotted cycles in the canonical book embedding of K_n for $7 \leq n \leq 11$. These values were obtained using a computer program that identifies knots from their Dowker–Thistlethwaite code [Toth and Walton 2010].

The values in column 3 of Table 1 are a consequence of the following:

Proposition 10. *Let $f(n)$ be the number of knotted Hamiltonian cycles in the canonical book representation of K_n. Then the total number of knotted cycles in $\widetilde{K}_n$*

is

$$\sum_{j=7}^{n} \binom{n}{j} f(j).$$

Proof. The proof follows immediately from Otsuki's result that any subset of vertices induces a subgraph that is ambient isotopic to the canonical book representation. □

Hirano [2010] proved that all spatial embeddings of K_8 must have at least 3 knotted Hamiltonian cycles; however, no known example achieves that bound. Abrams and Mellor [2010, Proposition 28] proved that the minimum number of knotted cycles in K_8 must be between 15 and 29. We conjecture the following:

Conjecture 11. *The canonical book representation of K_n contains the fewest total number of knotted cycles possible in any embedding of K_n.*

Conjecture 12. *The canonical book representation of K_n contains the fewest number of knotted Hamiltonian cycles possible in any embedding of K_n.*

Note that Conjecture 12 implies Conjecture 11. Both conjectures are true for $n \leq 7$.

Acknowledgments

We thank David Toth and Michael Walton for helpful conversations and for their assistance writing code to count the number of knotted Hamiltonian cycles in an embedding.

References

[Abrams and Mellor 2010] L. Abrams and B. Mellor, "Counting links and knots in complete graphs", preprint, 2010. arXiv 1008.1085

[Adams 1994] C. C. Adams, *The knot book*, W. H. Freeman, New York, 1994. MR 94m:57007 Zbl 0840.57001

[Bernhart and Kainen 1979] F. Bernhart and P. C. Kainen, "The book thickness of a graph", *J. Combin. Theory Ser. B* **27**:3 (1979), 320–331. MR 81f:05065 Zbl 0427.05028

[Birman 1975] J. S. Birman, *Braids, links, and mapping class groups*, Annals of Mathematics Studies **82**, Princeton University Press, 1975. MR 51 #11477 Zbl 0305.57013

[Conway and Gordon 1983] J. H. Conway and C. M. Gordon, "Knots and links in spatial graphs", *J. Graph Theory* **7**:4 (1983), 445–453. MR 85d:57002 Zbl 0524.05028

[Fleming and Mellor 2009] T. Fleming and B. Mellor, "Counting links in complete graphs", *Osaka J. Math.* **46**:1 (2009), 173–201. MR 2010j:05109 Zbl 1163.05008

[Hirano 2010] Y. Hirano, "Improved lower bound for the number of knotted Hamiltonian cycles in spatial embeddings of complete graphs", *J. Knot Theory Ramifications* **19**:5 (2010), 705–708. MR 2011g:57002 Zbl 1191.57003

[Kobayashi 1992] K. Kobayashi, "Standard spatial graph", *Hokkaido Math. J.* **21**:1 (1992), 117–140. MR 93e:57010 Zbl 0765.57003

[Otsuki 1996] T. Otsuki, "Knots and links in certain spatial complete graphs", *J. Combin. Theory Ser. B* **68**:1 (1996), 23–35. MR 97g:57013 Zbl 0858.05038

[Toth and Walton 2010] D. Toth and M. Walton, "Personal communication", 2010.

Received: 2012-01-18 Accepted: 2012-08-02

rowlandd@merrimack.edu *Department of Mathematics, Merrimack College,*
315 Turnpike Street, North Andover, MA 01845, United States

politanoa@merrimack.edu *Merrimack College, 315 Turnpike Street,*
North Andover, MA 01845, United States

INVOLVE 6:1 (2013)
dx.doi.org/10.2140/involve.2013.6.83

On closed modular colorings of rooted trees

Bryan Phinezy and Ping Zhang

(Communicated by Ann Trenk)

Two vertices u and v in a nontrivial connected graph G are twins if u and v have the same neighbors in $V(G) - \{u, v\}$. If u and v are adjacent, they are referred to as true twins, while if u and v are nonadjacent, they are false twins. For a positive integer k, let $c : V(G) \to \mathbb{Z}_k$ be a vertex coloring where adjacent vertices may be assigned the same color. The coloring c induces another vertex coloring $c' : V(G) \to \mathbb{Z}_k$ defined by $c'(v) = \sum_{u \in N[v]} c(u)$ for each $v \in V(G)$, where $N[v]$ is the closed neighborhood of v. Then c is called a closed modular k-coloring if $c'(u) \neq c'(v)$ in $\mathbb{Z}_k$ for all pairs u, v of adjacent vertices that are not true twins. The minimum k for which G has a closed modular k-coloring is the closed modular chromatic number $\overline{\text{mc}}(G)$ of G. A rooted tree T of order at least 3 is even if every vertex of T has an even number of children, while T is odd if every vertex of T has an odd number of children. It is shown that $\overline{\text{mc}}(T) = 2$ for each even rooted tree and $\overline{\text{mc}}(T) \leq 3$ if T is an odd rooted tree having no vertex with exactly one child. Exact values $\overline{\text{mc}}(T)$ are determined for several classes of odd rooted trees T.

1. Introduction

A weighting (or edge labeling with positive integers) of a connected graph G was introduced in [Chartrand et al. 1988] for the purpose of producing a weighted graph whose degrees (obtained by adding the weights of the incident edges of each vertex) were distinct. Such a weighted graph was called *irregular*. This concept could be looked at in another manner, however. In particular, let $\mathbb{N}$ denote the set of positive integers and let E_v denote the set of edges of G incident with a vertex v. An edge coloring $c : E(G) \to \mathbb{N}$, where adjacent edges may be colored the same, is said to be *vertex-distinguishing* if the coloring $c' : V(G) \to \mathbb{N}$ induced by c and defined by $c'(v) = \sum_{e \in E_v} c(e)$ has the property that $c'(x) \neq c'(y)$ for every two distinct vertices x and y of G. The main emphasis of this research dealt with minimizing the largest color assigned to the edges of the graph to produce an irregular graph.

MSC2010: primary 05C15; secondary 05C05.

Keywords: rooted trees, closed modular colorings, closed modular k-coloring, closed modular chromatic number.

Vertex-distinguishing colorings have received increased attention during the past 25 years (see [Escuadro et al. 2007]).

Rosa [1967] introduced a vertex labeling that induces an *edge-distinguishing* labeling defined by subtracting labels. In particular, for a graph G of size m, a vertex labeling (an injective function) $f : V(G) \to \{0, 1, \dots, m\}$ was called a *β-valuation* if the induced edge labeling $f' : E(G) \to \{1, 2, \dots, m\}$ defined by $f'(uv) = |f(u) - f(v)|$ was bijective. Golomb [1972] called a β-valuation a *graceful labeling* and a graph possessing a graceful labeling a *graceful graph*. It is this terminology that became standard. Much research has been done on graceful graphs. A popular conjecture in graph theory, due to Anton Kotzig and Gerhard Ringel, is the following.

The Graceful Tree Conjecture. *Every nontrivial tree is graceful.*

Jothi [1991] introduced a concept that, in a certain sense, reverses the roles of vertices and edges in graceful labelings (see also [Gallian 1998]). For a connected graph G of order $n \geq 3$, let $f : E(G) \to \mathbb{Z}_n$ be an edge labeling of G that induces a bijective function $f' : V(G) \to \mathbb{Z}_n$ defined by $f'(v) = \sum_{e \in E_v} f(e)$ for each vertex v of G. Such a labeling f is called a *modular edge-graceful labeling*, while a graph possessing such a labeling is called *modular edge-graceful* (see [Jones et al. 2013]). Verifying a conjecture by Gnana Jothi on trees, Jones et al. [2012] showed not only that every tree of order $n \geq 3$ is modular edge-graceful if and only if $n \not\equiv 2 \pmod 4$ but a connected graph of order $n \geq 3$ is modular edge-graceful if and only if $n \not\equiv 2 \pmod 4$.

Many of these weighting or labeling concepts were later interpreted as coloring concepts with the resulting vertex-distinguishing labeling becoming a vertex-distinguishing coloring. A *neighbor-distinguishing coloring* is a coloring in which every pair of adjacent vertices are colored differently. Such a coloring is more commonly called a *proper coloring*. The minimum number of colors in a proper vertex coloring of a graph G is its chromatic number $\chi(G)$.

In 2004 a neighbor-distinguishing edge coloring $c : E(G) \to \{1, 2, \dots, k\}$ of a graph G was introduced (see [Chartrand and Zhang 2009, p. 385]) in which an induced vertex coloring $s : V(G) \to \mathbb{N}$ is defined by $s(v) = \sum_{e \in E_v} c(e)$ for each vertex v of G. The minimum k for which such a neighbor-distinguishing coloring exists is called the *sum distinguishing index*, denoted by $\mathrm{sd}(G)$ of G. This is therefore the proper coloring analogue of the irregular weighting mentioned earlier. Karoński et al. [2004] showed that, if $\chi(G) \leq 3$, then $\mathrm{sd}(G) \leq 3$. Addario-Berry et al. [2005] showed that, for every connected graph G of order at least 3, $\mathrm{sd}(G) \leq 4$. In fact, Karoński et al. [2004] made the following conjecture, which has acquired a name used by many.

The 1-2-3 Conjecture. *If G is a connected graph of order* 3 *or more*, *then* $\mathrm{sd}(G) \leq 3$.

A number of neighbor-distinguishing vertex colorings different from standard proper colorings have been introduced in the literature (see [Chartrand and Zhang 2009, pp. 379–385], for example). Chartrand et al. [2010] introduced a neighbor-distinguishing vertex coloring of a graph based on sums of colors. For a nontrivial connected graph G, let $c : V(G) \to \mathbb{N}$ be a vertex coloring of G where adjacent vertices may be colored the same. If k colors are used by c, then c is a *k-coloring* of G. The *color sum* $\sigma(v)$ of a vertex v is defined by $\sigma(v) = \sum_{u \in N(v)} c(u)$ where $N(v)$ denotes the neighborhood of v (the set of vertices adjacent to v). If $\sigma(u) \neq \sigma(v)$ for every two adjacent vertices u and v of G, then c is neighbor-distinguishing and is called a *sigma coloring* of G. The minimum number of colors required in a sigma coloring of a graph G is called the *sigma chromatic number* of G and is denoted by $\sigma(G)$. Chartrand et al. [2010] showed that, for each pair a, b of positive integers with $a \leq b$, there is a connected graph G with $\sigma(G) = a$ and $\chi(G) = b$.

Chartrand et al. [2012] introduced another neighbor-distinguishing vertex coloring that is closely related to colorings discussed above. For a nontrivial connected graph G, let $c : V(G) \to \mathbb{Z}_k$ $(k \geq 2)$ be a vertex coloring where adjacent vertices may be assigned the same color. The coloring c induces another vertex coloring $c' : V(G) \to \mathbb{Z}_k$, where

$$c'(v) = \sum_{u \in N[v]} c(u), \tag{1}$$

where $N[v] = N(v) \cup \{v\}$ is the closed neighborhood of v and the sum in (1) is performed in $\mathbb{Z}_k$. A coloring c of G is called a *closed modular k-coloring* if for every pair x, y of adjacent vertices in G either $c'(x) \neq c'(y)$ or $N[x] = N[y]$, in the latter case of which we must have $c'(x) = c'(y)$. Closed modular colorings of graphs were introduced in [Chartrand et al. 2012] and inspired by a domination problem. The minimum k for which G has a closed modular k-coloring is called the *closed modular chromatic number* of G and is denoted by $\overline{\text{mc}}(G)$. Chartrand et al. [2012] observed that the nontrivial complete graphs are the only nontrivial connected graphs G for which $\overline{\text{mc}}(G) = 1$. Two vertices u and v in a connected graph G are *twins* if u and v have the same neighbors in $V(G) - \{u, v\}$. If u and v are adjacent, they are referred to as *true twins*, while if u and v are nonadjacent, they are *false twins*. If u and v are adjacent vertices of a graph G such that $N[u] = N[v]$ (that is, u and v are true twins), then $c'(u) = c'(v)$ for every vertex coloring c of G. The following result appeared in [Chartrand et al. 2012].

Proposition 1.1. *If G is a nontrivial connected graph, then $\overline{\text{mc}}(G)$ exists. Furthermore, if G contains no true twins, then $\overline{\text{mc}}(G) \geq \chi(G)$.*

To illustrate these concepts, consider the bipartite graph G of Figure 1. Since $\chi(G) = 2$ and G has no true twins, it follows that $\overline{\text{mc}}(G) \geq 2$ by Proposition 1.1.

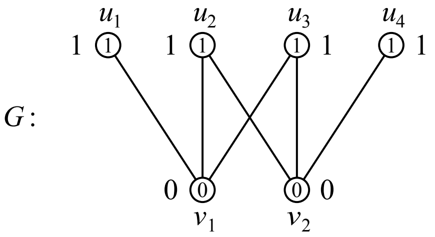


Figure 1. A graph G with $\chi(G) = 2$ and $\overline{\text{mc}}(G) = 3$.

In fact $\overline{\text{mc}}(G) = 3$. Figure 1 shows a closed modular 3-coloring of G (where the color of a vertex is placed within the vertex) together with the color $c'(v)$ for each vertex v of G (where the color $c'(v)$ of a vertex is placed next to the vertex).

For an edge uv of a graph G, the graph G/uv obtained from G by *contracting the edge* uv has the vertex set $V(G)$ in which u and v are identified. If we denote the vertex $u = v$ in G/uv by w, then $V(G/uv) = (V(G) \cup \{w\}) - \{u, v\}$ and the edge set of G/uv is

$$\begin{aligned} E(G/uv) = \{xy : xy \in E(G),\ x, y \in V(G) - \{u, v\}\} \\ \cup \{wx : ux \in E(G) \text{ or } vx \in E(G), x \in V(G) - \{u, v\}\}. \end{aligned}$$

The graph G/uv is referred to as an *elementary contraction* of G. For a nontrivial connected graph G, define the *true twins closure* $\text{TC}(G)$ of G as the graph obtained from G by a sequence of elementary contractions of pairs of true twins in G until no such pair remains. In particular, if G contains no true twins, then $\text{TC}(G) = G$. Thus $\text{TC}(G)$ is a minor of G. Chartrand et al. [2012] showed that $\overline{\text{mc}}(G) = \overline{\text{mc}}(\text{TC}(G))$ for every nontrivial connected graph G. Therefore, it suffices to consider nontrivial connected graphs containing no true twins.

Closed modular chromatic numbers were determined for several classes of regular graphs in [Chartrand et al. 2012]. In particular, it was shown that, for each integer $k \geq 2$, if G is a regular complete k-partite graph such that each of its partite sets has at least $2k + 1$ vertices, then $\overline{\text{mc}}(G) \leq 2\chi(G) - 1$ and this bound is sharp.

In [Phinezy and Zhang 2013], we investigated the closed modular chromatic number for trees and determined it for several classes of trees. For each tree T in these classes, either $\overline{\text{mc}}(T) = 2$ or $\overline{\text{mc}}(T) = 3$. Indeed, this is conjectured to be true in great generality:

Conjecture 1.2 [Phinezy and Zhang 2013]. *If T is a tree of order* 3 *or more, then* $\overline{\text{mc}}(T) \leq 3$.

In the paper cited, we showed that Conjecture 1.2 is true if 3 is replaced by 4. In this work, we investigate the closed modular chromatic numbers of rooted trees and confirm Conjecture 1.2 for several classes of rooted trees, including well-studied complete r-ary trees. We refer to [Chartrand et al. 2011] for graph theory notation and terminology not described in this paper. All trees under consideration in this work are rooted trees of order at least 3.

2. Rooted trees

Let T be a rooted tree of order at least 3 having the root v. For each integer i with $0 \le i \le e(v)$, where $e(v)$ is the distance between v and a vertex farthest from v, let

$$V_i = \{x \in V(T) : d(v, x) = i\}.$$

If $x \in V_i$ where $0 \le i \le e(v)$, then x is *at level i*. If $x \in V_i$ $(0 \le i \le e(v) - 1)$ is adjacent to $y \in V_{i+1}$, then x is the *parent* of y and y is a *child* of x. A vertex z is a *descendant* of x (and x is an *ancestor* of z) if the $x - z$ path in T lies below x. In this section, we show that, if T is a rooted tree of order at least 3 such that the numbers of children of all vertices of T have the same parity and no vertex of T has exactly one child, then either $\overline{\text{mc}}(T) = 2$ or $\overline{\text{mc}}(T) = 3$. In order to do this, we first present a result on a special class of trees, which was established in [Phinezy and Zhang 2013]. A *caterpillar* is a tree of order 3 or more, the removal of whose end-vertices produces a path called the *spine* of the caterpillar. Thus every path and star (of order at least 3) and every double star (a tree of diameter 3) is a caterpillar.

Theorem 2.1. *If T is a caterpillar of order at least* 3, *then* $\overline{\text{mc}}(T) \le 3$.

Theorem 2.2. *Let T be a rooted tree of order at least* 3.

(a) *If each vertex of T has an even number of children*, *then* $\overline{\text{mc}}(T) = 2$.

(b) *If each vertex of T has either no child or an odd number of children and no vertex has exactly one child*, *then* $\overline{\text{mc}}(T) \le 3$.

Proof. Suppose that v is the root of T. For each integer i with $0 \le i \le e(v)$, let

$$V_i = \{x \in V(T) : d(v, x) = i\}.$$

To verify (a), define the coloring $c : V(G) \to \mathbb{Z}_2$ by

$$c(x) = \begin{cases} 1 & \text{if } x \in V_i \text{ where } i \equiv 0, 1 \pmod 4, \\ 0 & \text{if } x \in V_i \text{ where } i \equiv 2, 3 \pmod 4. \end{cases} \tag{2}$$

Then $c'(x) = 1$ if $x \in V_i$ and i is even and $c'(x) = 0$ if $x \in V_i$ and i is odd. Thus c is a closed modular 2-coloring and so $\overline{\text{mc}}(T) = 2$ if each vertex of T has an even number of children.

To verify (b), we proceed by strong induction. If T is a star, then $\overline{\text{mc}}(T) \le 3$ by Theorem 2.1. Assume for an integer $n \ge 4$ that, if each vertex of a tree of order at most n has either no child or an odd number of children and no vertex has exactly one child, then the closed modular chromatic number of the tree is at most 3. Let T be a tree of order $n+1$ such that each vertex of T has either no child or an odd number of children and no vertex has exactly one child. We may assume that T is not a star. Let x be a peripheral vertex of T; then x is an end-vertex of T. Suppose that x is a child of the vertex y in T. Since each vertex of T has either no child or an odd number of children and no vertex of T has exactly one child, it follows that y has an odd number $r \geq 3$ of children; say $x = x_1, x_2, \ldots, x_r$ are children of y. Then each child of y is an end-vertex of T. Let $X = \{x = x_1, x_2, \ldots, x_r\}$. Consider $T^* = T - X$ which is a tree of order less than $n+1$ such that each vertex of T^* has either no child or an odd number of children and no vertex of T^* has exactly one child. By the induction hypothesis, T^* has a closed modular 3-coloring $c : V(T^*) \to \mathbb{Z}_3$. Next, we show that T has a closed modular 3-coloring $c_T : V(T) \to \mathbb{Z}_3$ such that $c_T(u) = c(u)$ and $c'_T(u) = c'(u)$ for each $u \in V(T^*)$. Since r is odd, $r \equiv 1, 3, 5 \pmod 6$. We consider these three cases.

Case 1: $r \equiv 1 \pmod 6$. In this case, $r \geq 7$. We define c_T on X such that $c'_T(y) = c'(y)$. If $c(y) \neq c'(y)$, then c_T assigns the color 0 to x_i for $1 \leq i \leq r$. Hence $c'_T(x_i) = c(y) \neq c'(y)$ for $1 \leq i \leq r$. If $c(y) = c'(y)$, then c_T assigns the color 2 to x_1 and x_2 and the color 1 to x_i for $3 \leq i \leq r$. Hence $c'_T(x_i) = c'(y) + 2 \neq c'(y)$ for $i = 1, 2$ and $c'_T(x_i) = c'(y) + 1 \neq c'(y)$ for $3 \leq i \leq r$.

Case 2: $r \equiv 3 \pmod 6$. We define c_T on X such that $c'_T(y) = c'(y)$. If $c(y) \neq c'(y)$, then c_T assigns the color 0 to x_i for $1 \leq i \leq r$. Hence $c'_T(x_i) = c(y) \neq c'(y)$ for $1 \leq i \leq r$. If $c(y) = c'(y)$, then c_T assigns the color 1 to x_i for $1 \leq i \leq r$. Hence $c'_T(x_i) = c'(y) + 1 \neq c'(y)$ for $1 \leq i \leq r$.

Case 3: $r \equiv 5 \pmod 6$. We define c_T on X such that $c'_T(y) = c'(y)$. If $c(y) \neq c'(y)$, then c_T assigns the color 0 to x_i for $1 \leq i \leq r$. Hence $c'_T(x_i) = c(y) \neq c'(y)$ for $1 \leq i \leq r$. If $c(y) = c'(y)$, then c_T assigns the color 2 to x_1 and assigns the color 1 to x_i for $2 \leq i \leq r$. Hence $c'_T(x_1) = c'(y) + 2$ and $c'_T(x_i) = c'(y) + 1$ for $2 \leq i \leq r$.

In each case, c_T is a closed modular 3-coloring of T and so $\overline{\mathrm{mc}}(T) \leq 3$. □

Theorem 2.2 provides the closed modular chromatic numbers for a well-known class of rooted trees. A rooted tree T is a *complete r-ary tree* for some integer $r \geq 2$ if every vertex of T has either r children or no child. The following is a consequence of Theorem 2.2.

Corollary 2.3. *For an integer $r \geq 2$, let T be a complete r-ary tree.*

(a) *If r is even, then* $\overline{\mathrm{mc}}(T) = 2$.

(b) *If r is odd, then* $\overline{\mathrm{mc}}(T) \leq 3$.

In the view of Theorem 2.2, it would be useful to introduce an additional terminology. A rooted tree T of order at least 3 is *even* if every vertex of T has an even number of children, while T is *odd* if every vertex of T has an odd number of children. It then follows by Theorem 2.2 that $\overline{\mathrm{mc}}(T) = 2$ if T is an even rooted tree and $\overline{\mathrm{mc}}(T) \le 3$ if T is an odd rooted tree and no vertex of T has exactly one child.

3. Odd rooted trees

In this section we investigate the closed modular colorings of odd rooted trees of order at least 3. We will see that, if the locations of leaves of an odd rooted tree T are given, then in some cases it is possible to determine the exact value of $\overline{\mathrm{mc}}(T)$. For each integer $p \in \{0, 1, 2, 3, 4, 5\}$, an odd rooted tree T of order at least 3 having root v is said to be of *type* p if $d(v, u) \equiv p \pmod 6$ for every leaf u in T. We now determine all odd rooted trees of type p were $0 \le p \le 5$ that have closed modular chromatic number 2.

Theorem 3.1. *For each integer $p \in \{0, 1, 2, 3, 4, 5\}$, let T be an odd rooted tree of order at least 3 that is of type p. Then $\overline{\mathrm{mc}}(T) = 2$ if and only if $p \ne 1$.*

Proof. Suppose that v is the root of T. For each integer i with $0 \le i \le e(v)$, let $V_i = \{x \in V(T) : d(v, x) = i\}$. First, suppose that $0 \le p \le 5$ and $p \ne 1$. We show $\overline{\mathrm{mc}}(T) = 2$. Since $\chi(T) = 2$ for every nontrivial tree T, it suffices to construct a closed modular 2-coloring $c : V(T) \to \mathbb{Z}_2$ of T. We consider three cases, according to the values of p.

Case 1: $p = 0$. In this case, a coloring $c : V(T) \to \mathbb{Z}_2$ is defined by

$$c(x) = \begin{cases} 0 & \text{if } x \in V_i \text{ and } i \equiv 0, 1, 5 \pmod 6, \\ 1 & \text{if } x \in V_i \text{ and } i \equiv 2, 3, 4 \pmod 6. \end{cases}$$

Then the induced coloring $c' : V(T) \to \mathbb{Z}_2$ is defined as

$$c'(x) = \begin{cases} 0 & \text{if } x \in V_i \text{ and } i \text{ is even}, \\ 1 & \text{if } x \in V_i \text{ and } i \text{ is odd}. \end{cases} \tag{3}$$

Case 2: $p \equiv 2, 3, 4 \pmod 6$. In this case, a coloring $c : V(T) \to \mathbb{Z}_2$ is defined by

$$c(x) = \begin{cases} 1 & \text{if } x \in V_i \text{ and } i \equiv 0, 1, 2 \pmod 6, \\ 0 & \text{if } x \in V_i \text{ and } i \equiv 3, 4, 5 \pmod 6. \end{cases}$$

Then the induced coloring $c' : V(T) \to \mathbb{Z}_2$ is defined as in (3).

Case 3: $p \equiv 5 \pmod 6$. In this case, a coloring $c : V(T) \to \mathbb{Z}_2$ is defined by

$$c(x) = \begin{cases} 0 & \text{if } x \in V_i \text{ and } i \equiv 0, 4, 5 \pmod 6, \\ 1 & \text{if } x \in V_i \text{ and } i \equiv 1, 2, 3 \pmod 6. \end{cases}$$

Then the induced coloring $c' : V(T) \to \mathbb{Z}_2$ is defined as

$$c'(x) = \begin{cases} 1 & \text{if } x \in V_i \text{ and } i \text{ is even}, \\ 0 & \text{if } x \in V_i \text{ and } i \text{ is odd}. \end{cases}$$

Thus c is a closed modular 2-coloring of T and so $\overline{\mathrm{mc}}(T) = 2$.

For the converse, suppose that T is an odd rooted tree of order at least 3 that is of type 1. Thus, if u is a leaf of T, then $u \in V_k$ for some integer k, where then $1 \leq k \leq e(v)$ and $k \equiv 1 \pmod 6$. We show that $\overline{\mathrm{mc}}(T) \neq 2$. Assume, to the contrary, that there is a closed modular 2-coloring $c : V(T) \to \mathbb{Z}_2$ of T. Then $c'(v) = 0$ or $c'(v) = 1$. We consider these two cases.

Case 1: $c'(v) = 0$. Thus $c'(x) = 0$ if $x \in V_i$ and i is even and $c'(x) = 1$ if $x \in V_i$ and i is odd. Since $c(v) \in \{0, 1\}$, there are two subcases.

Subcase 1.1: $c(v) = 0$. Since $c'(v) = 0$ and $c(v) = 0$, there is $v_1 \in V_1$ such that $c(v_1) = 0$. Since $c'(v_1) = 1$ and $c(v) = c(v_1) = 0$, there is $v_2 \in V_2$ such that $c(v_2) = 1$. Since $c'(v_2) = 0$, $c(v_1) = 0$ and $c(v_2) = 1$, there is $v_3 \in V_3$ such that $c(v_3) = 1$. Observe for each $i \geq 3$ that $c(v_i)$ is uniquely determined by $c'(v_{i-1})$, $c(v_{i-2})$ and $c(v_{i-1})$. Repeating this procedure, we obtain a path $P_k = (v_1, v_2, \ldots, v_k)$ in T such that (1) v_k is a leaf of T, $d(v, v_i) = i$ for $1 \leq i \leq k$ and $k \equiv 1 \pmod 6$ and (2) the color sequence $s_c = (c(v_1), c(v_2), \ldots, c(v_k))$ of the coloring c on the path P_k is

$$s_c = (0, \underline{1, 1, 1, 0, 0, 0}, \underline{1, 1, 1, 0, 0, 0}, \ldots, \underline{1, 1, 1, 0, 0, 0}).$$

Hence $(c(v_{k-2}), c(v_{k-1}), c(v_k)) = (0, 0, 0)$. However, then $c'(v_{k-1}) = c'(v_k) = 0$, which is a contradiction.

Subcase 1.2: $c(v) = 1$. By the same argument as in Subcase 1.1, we conclude that there must be a path $P_k = (v_1, v_2, \ldots, v_k)$ in T such that (1) v_k is a leaf of T, $d(v, v_i) = i$ for $1 \leq i \leq k$ and $k \equiv 1 \pmod 6$ and (2) the color sequence $s_c = (c(v_1), c(v_2), \ldots, c(v_k))$ of the coloring c on the path P_k is

$$s_c = (1, \underline{1, 0, 0, 0, 1, 1}, \underline{1, 0, 0, 0, 1, 1}, \ldots, \underline{1, 0, 0, 0, 1, 1}).$$

Hence $(c(v_{k-2}), c(v_{k-1}), c(v_k)) = (0, 1, 1)$. However, then $c'(v_{k-1}) = c'(v_k) = 0$, which is a contradiction.

Case 2: $c'(v) = 1$. Thus $c'(x) = 1$ if $x \in V_i$ and i is even and $c'(x) = 0$ if $x \in V_i$ and i is odd. Since $c(v) \in \{0, 1\}$, there are two subcases.

Subcase 2.1: $c(v) = 0$. Since $c'(v) = 1$ and $c(v) = 0$, there is $v_1 \in V_1$ such that $c(v_1) = 1$. Since $c'(v_1) = 0$, $c(v) = 0$ and $c(v_1) = 1$, there is $v_2 \in V_2$ such that $c(v_2) = 1$. Since $c'(v_2) = 1$, $c(v_1) = c(v_2) = 1$, there is $v_3 \in V_3$ such that $c(v_3) = 1$. Observe for each $i \geq 3$ that $c(v_i)$ is uniquely determined by $c'(v_{i-1})$, $c(v_{i-2})$ and $c(v_{i-1})$. Repeating this procedure, we obtain a path $P_k = (v_1, v_2, \ldots, v_k)$ in T such that (1) v_k is a leaf of T, $d(v, v_i) = i$ for $1 \leq i \leq k$ and $k \equiv 1 \pmod 6$ and (2) the color sequence $s_c = (c(v_1), c(v_2), \ldots, c(v_k))$ of the coloring c on the path P_k is

$$s_c = (1, \underline{1, 1, 0, 0, 0, 1}, \underline{1, 1, 0, 0, 0, 1}, \ldots, \underline{1, 1, 0, 0, 0, 1}).$$

Hence $(c(v_{k-2}), c(v_{k-1}), c(v_k)) = (0, 0, 1)$. However, then $c'(v_{k-1}) = c'(v_k) = 1$, which is a contradiction.

Subcase 2.2: $c(v) = 1$. By the same argument as in Subcase 2.1, we conclude that there must be a path $P_k = (v_1, v_2, \dots, v_k)$ in T such that (1) v_k is a leaf of T, $d(v, v_i) = i$ for $1 \le i \le k$ and $k \equiv 1 \pmod 6$ and (2) the color sequence $s_c = (c(v_1), c(v_2), \dots, c(v_k))$ of the coloring c on the path P_k is

$$s_c = (0, \underline{1, 0, 1, 0, 1, 0}, \underline{1, 0, 1, 0, 1, 0}, \dots, \underline{1, 0, 1, 0, 1, 0}).$$

Hence $(c(v_{k-2}), c(v_{k-1}), c(v_k)) = (0, 1, 0)$. However, then $c'(v_{k-1}) = c'(v_k) = 1$, which is a contradiction. □

By Theorem 3.1, if T is an odd rooted tree of order at least 3 that is of type 1, then $\overline{\text{mc}}(T) \ge 3$. On the other hand, every odd rooted tree of order at least 3 we have encountered that is of type 1 has closed modular chromatic number 3. Furthermore, the following is a consequence of Theorems 2.2 and 3.1.

Corollary 3.2. *If T is an odd rooted tree of order at least 3 that is of type 1 such that no vertex has exactly one child, then $\overline{\text{mc}}(T) = 3$.*

By Theorem 3.1, if p is an integer with $0 \le p \le 5$ and $p \ne 1$, then every odd rooted tree of order at least 3 that is of type p has closed modular chromatic number 2. This gives rise to the question:

> If $S \subseteq \{0, 1, 2, 3, 4, 5\}$ where $|S| \ge 2$ and $1 \notin S$ and T is an odd rooted tree of order at least 3 having root v such that, for every leaf u in T, $d(v, u) \equiv p \pmod 6$ for some $p \in S$, then is it necessary that $\overline{\text{mc}}(T) = 2$?

The answer to this question is no, as we show next. First, it will be convenient to introduce an additional definition. For a nonempty subset $S \subseteq \{0, 2, 3, 4, 5\}$, an odd rooted tree T having root v is said to be of *type* S if, for every leaf u in T, $d(v, u) \equiv p \pmod 6$ for some $p \in S$ and, for each $p \in S$, there is at least one leaf u in T such that $d(v, u) \equiv p \pmod 6$. In particular, if $S = \{p\}$ where $p \in \{0, 2, 3, 4, 5\}$, then T is of type p. We first consider odd rooted trees of type S, where $S = \{2, 5\}$ or $S = \{0, 3\}$. In the next two results, we show that if $S = \{2, 5\}$ or $S = \{0, 3\}$, then it is possible for an odd rooted tree T of type S to have $\overline{\text{mc}}(T) = 3$.

Theorem 3.3. *For $S = \{2, 5\}$, there are odd rooted trees of type S such that $\overline{\text{mc}}(T) = 3$.*

Proof. Consider the tree T in Figure 2, each of whose leaves are at level 2 or at level 5. We show $\overline{\text{mc}}(T) = 3$. For each integer i with $0 \le i \le 5$, let $V_i = \{x \in V(T) : d(v, x) = i\}$. Thus, if x is a leaf of T, then $x \in V_2$ or $x \in V_5$. By Corollary 2.3, $\overline{\text{mc}}(T) \le 3$. It remains to show that $\overline{\text{mc}}(T) \ne 2$. Assume, to the contrary, that there is a closed modular 2-coloring $c : V(T) \to \mathbb{Z}_2$ of T. Thus $c(v) = 0$ or $c(v) = 1$. We consider these two cases.

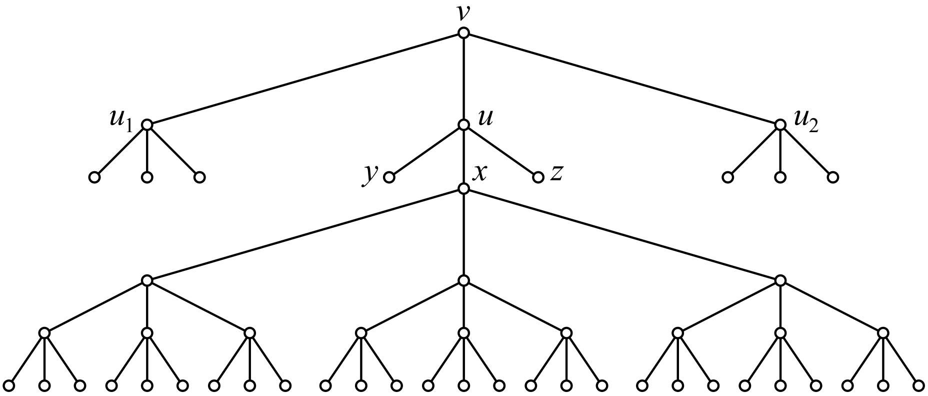


Figure 2. A tree T with $\overline{\text{mc}}(T) = 3$.

Case 1: $c(v) = 0$. Since either $c'(v) = 0$ or $c'(v) = 1$, there are two subcases.

Subcase 1.1: $c'(v) = 0$. Thus $c'(w) = 0$ if $w \in V_i$ and i is even and $c'(w) = 1$ if $w \in V_i$ and i is odd. Furthermore, $c(u_1) = 0$ or $c(u_1) = 1$. First, assume that $c(u_1) = 0$. Since $c'(u_1) = 1$ and $c(v) = 0$, there is a child w of u_1 such that $c(w) = 1$. However, then $c'(w) = c'(u_1) = 1$, a contradiction. Next, assume that $c(u_1) = 1$. Since $c'(u_1) = 1$ and $c(v) = 0$, there is a child w of u_1 such that $c(w) = 0$. However, then $c'(w) = c'(u_1) = 1$, a contradiction.

Subcase 1.2: $c'(v) = 1$. Thus $c'(w) = 0$ if $w \in V_i$ and i is odd and $c'(w) = 1$ if $w \in V_i$ and i is even. Furthermore, $c(u_1) = 1$ or $c(u_1) = 0$. First, assume that $c(u_1) = 1$. Since $c'(u_1) = 0$ and $c(v) = 0$, there is a child w of u_1 such that $c(w) = 1$. However, then $c'(w) = c'(u_1) = 0$, a contradiction. Next, assume that $c(u_1) = 0$. Since $c'(u_1) = 0$ and $c(v) = c(u_1) = 0$, there is a child w of u_1 such that $c(w) = 0$. However, then $c'(w) = c'(u_1) = 0$, a contradiction.

Case 2: $c(v) = 1$. Since either $c'(v) = 0$ or $c'(v) = 1$, there are two subcases.

Subcase 2.1: $c'(v) = 0$. Then $c(u) = 1$ or $c(u) = 0$. First, assume that $c(u) = 1$. Since $c'(u) = 1$ and $c(v) = 1$, there is a child w of u such that $c(w) = 1$. We claim that $c(y) \neq 0$ and $c(z) \neq 0$, for otherwise, say $c(y) = 0$. Then $c'(u) = c'(y) = 1$, a contradiction. Thus $c(y) = c(z) = 1$, as claimed, which implies that $c(x) = 1$. Since $c'(x) = 0$, there is a child w of x such that $c(w) = 0$. Since $c'(w) = 1$, there is a child w_1 of w such that $c(w_1) = 0$. Since $c'(w_1) = 0$, there is a child w_2 of w_1 such that $c(w_2) = 0$. However, then $c'(w_1) = c'(w_2) = 0$, a contradiction. Next, assume that $c(u) = 0$. We saw that $c(y) \neq 1$ and $c(z) \neq 1$ and so $c(y) = c(z) = 0$. Since $c'(u) = 1$, it follows that $c(x) = 0$. Since $c'(x) = 1$, there is a child w of x such that $c(w) = 0$. Since $c'(w) = 1$, there is a child w_1 of w such that $c(w_1) = 1$.

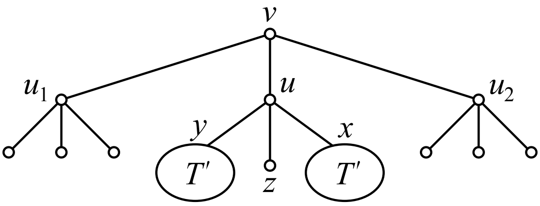


Figure 3. A tree T^* of type $S = \{2, 5\}$ with $\overline{\mathrm{mc}}(T^*) = 2$.

Since $c'(w_1) = 0$, there is a child w_2 of w_1 such that $c(w_2) = 1$. However, then $c'(w_1) = c'(w_2) = 0$, a contradiction.

Subcase 2.2: $c'(v) = 1$. Then $c(u) = 1$ or $c(u) = 0$. We consider these two possibilities.

Subcase 2.2.1: $c(u) = 1$. Now either $c(x) = 0$ or $c(x) = 1$. First assume that $c(x) = 0$. Since $c'(x) = 1$ and $c(u) = 1$, there is a child w of x such that $c(w) = 0$. Since $c'(w) = 0$ and $c(x) = 0$, there is a child w_1 of w such that $c(w_1) = 0$. Since $c'(w_1) = 1$ and $c(w_1) = 0$, there is a child w_2 of w_1 such that $c(w_2) = 1$. However, then $c'(w_1) = c'(w_2) = 1$, a contradiction. Next, assume that $c(x) = 1$. Since $c(u) = 1$ and $c'(u) = 0$, one of y and z must be colored 1, say $c(y) = 1$. However, then $c'(y) = c'(u) = 0$, a contradiction.

Subcase 2.2.2: $c(u) = 0$. Now either $c(x) = 0$ or $c(x) = 1$. First assume that $c(x) = 0$. Since $c'(u) = 0$, exactly one of y and z is colored 1, say $c(y) = 1$ and $c(z) = 0$. However, then $c'(z) = c'(u) = 0$, a contradiction. Next, assume that $c(x) = 1$. Since $c'(x) = c(x) = 1$, there is a child w of x such that $c(w) = 0$. Since $c'(w) = 0$, $c(x) = 1$ and $c(w) = 0$, there is a child w_1 of w such that $c(w_1) = 1$. Since $c'(w_1) = 1$, there is a child w_2 of w_1 such that $c(w_2) = 0$. However, then $c'(w_1) = c'(w_2) = 1$, a contradiction. □

By Theorem 3.3, despite the fact that every odd rooted tree of type 2 or type 5 has closed modular chromatic number 2, there are odd rooted trees T of type $S = \{2, 5\}$ for which $\overline{\mathrm{mc}}(T) = 3$. On the other hand, there are odd rooted trees of type $S = \{2, 5\}$ having closed modular chromatic number 2. For example, we start with the tree T in Figure 2. Let T' be the subtree of T whose vertex set consists of x and all descendants of x. Then the tree T^* is constructed from T of Figure 2 by replacing y with a copy of T' (see Figure 3). The coloring $c : V(T) \to \mathbb{Z}_2$ defined by assigning the color 0 to each vertex in $\{u, u_1, u_2, x, y\}$ and assigning the color 1 to the remaining vertices of T^* is a closed modular 2-coloring. Therefore, $\overline{\mathrm{mc}}(T^*) = 2$.

Theorem 3.4. *For $S = \{0, 3\}$, there are odd trees T of type S such that* $\overline{\mathrm{mc}}(T) = 3$.

Proof. Consider the tree T of Figure 4, each of whose leaves are at level 3 or at level 6. We show that $\overline{\mathrm{mc}}(T) = 3$. For each integer i with $0 \leq i \leq 6$, let

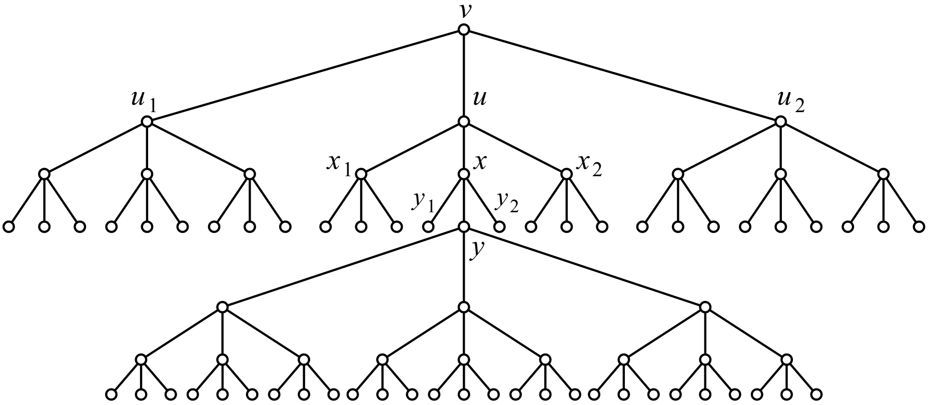


Figure 4. A tree T with $\overline{\mathrm{mc}}(T) = 3$.

$V_i = \{x \in V(T) : d(v, x) = i\}$. If x is a leaf of T, then $x \in V_3$ or $x \in V_6$. By Corollary 2.3, $\overline{\mathrm{mc}}(T) \le 3$. Thus it remains to show that $\overline{\mathrm{mc}}(T) \ne 2$. Assume, to the contrary, that there is a closed modular 2-coloring $c : V(T) \to \mathbb{Z}_2$ of T. Thus $c(v) = 0$ or $c(v) = 1$. We consider these two cases.

Case 1: $c(v) = 0$. Since either $c'(v) = 0$ or $c'(v) = 1$, there are two subcases.

Subcase 1.1: $c'(v) = 0$. In this case, there is a child of v that is colored 0. First, assume that $c(u_1) = 0$. Since $c'(u_1) = 1$, there is a child w of u_1 such that $c(w) = 1$. Since $c'(w) = 0$ and $c(w) = 1$, there is a child w_1 of w such that $c(w_1) = 1$. However, then $c'(w_1) = c'(w) = 0$, a contradiction. Thus $c(u_1) = 1$ and, similarly, $c(u_2) = 1$. This implies that $c(u)$ must be 0. Note that $c(x_1) = 1$ or $c(x_1) = 0$. If $c(x_1) = 1$, then there is a child w of x_1 such that $c(w) = 1$. However, then $c'(x_1) = c'(w) = 0$, a contradiction. If $c(x_1) = 0$, then there is a child w of x_1 such that $c(w) = 0$. However, then $c'(x_1) = c'(w) = 0$, a contradiction.

Subcase 1.2: $c'(v) = 1$. Since $c(v) = 0$, either exactly one or exactly three children of v must be colored 1. First, suppose that $c(u_1) = 0$. Then there is a child w of u_1 such that $c(w) = 0$. Since $c'(w) = 1$, there is a child w_1 of w such that $c(w_1) = 1$. However, then $c'(w_1) = c'(w) = 1$, a contradiction. Thus $c(u_1) = 1$ and, similarly, $c(u_2) = 1$. This implies that $c(u)$ must be 1. Note that $c(x) = 1$ or $c(x) = 0$. We consider these two subcases.

Subcase 1.2.1: $c(x) = 1$. In this case, either exactly one or exactly three children of x must be colored 1. Since $c'(x) = 1$, it follows $c(y_1) = 1$ (for otherwise, $c'(y_1) = 1$). Similarly $c(y_2) = 1$. Thus $c(y)$ must be 1. Since $c'(y) = 0$, there is a child w of y such that $c(w) = 0$. Since $c'(w) = 1$, there is a child w_1 of w such that $c(w_1) = 0$. This in turn implies that there is a child w_2 of w_1 such that $c(w_2) = 0$. However, then $c'(w_1) = c'(w_2) = 0$, a contradiction.

Subcase 1.2.2: $c(x) = 0$. Since $c'(x) = 1$, it follows that $c(y_1) = c(y_2) = 0$, which implies that $c(y) = 0$. Since $c'(y) = c(y) = 0$, there is a child w of y such that $c(w) = 0$. Since $c'(w) = 1$, there is a child w_1 of w such that $c(w_1) = 1$. This in turn implies that there is a child w_2 of w_1 such that $c(w_2) = 1$. However, then $c'(w_1) = c'(w_2) = 0$, a contradiction.

Case 2: $c(v) = 1$. Since either $c'(v) = 0$ or $c'(v) = 1$, there are two subcases.

Subcase 2.1: $c'(v) = 0$. Note that $c(u_1) = 0$ or $c(u_1) = 1$. First, assume that $c(u_1) = 0$. Since $c'(u_1) = 1$, there is a child w of u_1 such that $c(w) = 0$. Since $c'(w) = 0$, there is a child w_1 of w such that $c(w_1) = 0$. However, then $c'(w_1) = c'(w) = 0$, a contradiction. Thus $c(u_1) = 1$ and, similarly, $c(u_2) = 1$. This implies that $c(u)$ must be 1. Note that $c(x) = 0$ or $c(x) = 1$. There are two subcases.

Subcase 2.1.1: $c(x) = 0$. If $c(y_1) = 0$, then $c'(y) = c'(x) = 0$, a contradiction. Thus $c(y_1) = 1$ and, similarly, $c(y_2) = 1$. This implies that $c(y)$ must be 1. Since $c'(y) = 1$, there is a child w of y such that $c(w) = 0$. Since $c'(w) = 0$ and $c(y) = 1$, there is a child w_1 of w such that $c(w_1) = 1$. This in turn implies that there is a child w_2 of w_1 such that $c(w_2) = 0$. However, then $c'(w_1) = c'(w_2) = 1$, a contradiction.

Subcase 2.1.2: $c(x) = 1$. If $c(y_1) = 1$, then $c'(y) = c'(x) = 0$, a contradiction. Thus $c(y_1) = 0$ and, similarly, $c(y_2) = 0$. This implies that $c(y)$ must be 0. Since $c'(y) = 1$ and $c(x) = 1$, there is a child w of y such that $c(w) = 0$. Since $c'(w) = 0$, there is a child w_1 of w such that $c(w_1) = 0$. This in turn implies that there is a child w_2 of w_1 such that $c(w_2) = 0$. However, then $c'(w_1) = c'(w_2) = 1$, a contradiction.

Subcase 2.2: $c'(v) = 1$. Note that $c(u_1) = 0$ or $c(u_1) = 1$. First, assume that $c(u_1) = 0$. Since $c'(u_1) = c(u_1) = 0$ and $c(v) = 1$, there is a child w of u_1 such that $c(w) = 1$. Since $c'(w) = 1$, there is a child w_1 of w such that $c(w_1) = 0$. However, then $c'(w_1) = c'(w) = 1$, a contradiction. Thus $c(u_1) = 1$ and, similarly, $c(u_2) = 1$. This implies that $c(u)$ must be 0. Note that $c(x_1) = 0$ or $c(x_1) = 1$. Furthermore, $c'(x_1) = 1$. If $c(x_1) = 0$, then there is a child w of x_1 such that $c(w) = 1$. However, then $c'(x_1) = c'(w) = 1$, a contradiction. If $c(x_1) = 1$, then there is a child w of x_1 such that $c(w) = 0$. However, then $c'(x_1) = c'(w) = 1$, a contradiction. □

As with the case when $S = \{2, 5\}$, there are odd rooted trees of type $S = \{0, 3\}$ having closed modular chromatic number 2. For example, we start with the tree T in Figure 4. Let T' be the subtree of T whose vertex set consists of y and all descendants of y. Then the tree T^* is constructed from T of Figure 4 by replacing y_1 with a copy of T' as we did in the case when $S = \{2, 5\}$ (see Figure 3). The coloring $c : V(T) \to \mathbb{Z}_2$ defined by assigning the color 0 to each vertex in $\{v, y, y_1\} \cup V_5$ and assigning the color 1 to the remaining vertices of T^* is a closed modular 2-coloring. Therefore, $\overline{\text{mc}}(T^*) = 2$.

Next, we show that, if S is a nonempty subset of $\{0, 2, 3, 4, 5\}$ such that S contains at most one of 2 and 5 and at most one of 0 and 3, then every odd rooted tree of type S has closed modular chromatic number 2.

Theorem 3.5. *Let S be a nonempty subset of* $\{0, 2, 3, 4, 5\}$ *such that S contains at most one of* 2 *and* 5 *and at most one of* 0 *and* 3. *If T is an odd rooted tree of order at least* 3 *that is of type S, then* $\overline{\mathrm{mc}}(T) = 2$.

Proof. By Theorem 3.1, we may assume that $|S| \geq 2$. Since $|S \cap \{2, 5\}| \leq 1$ and $|S \cap \{0, 3\}| \leq 1$, it follows that $|S| \leq 3$. Thus we consider two cases, according to whether $|S| = 3$ or $|S| = 2$.

Case 1: $|S| = 3$. Then S is one of the sets $\{0, 2, 4\}$, $\{0, 4, 5\}$, $\{2, 3, 4\}$, $\{3, 4, 5\}$. Since $\chi(T) = 2$ for every nontrivial tree T, it suffices to show that there is a closed modular 2-coloring by Proposition 1.1. For each integer i with $0 \leq i \leq e(v)$, let

$$V_i = \{x \in V(T) : d(v, x) = i\}.$$

First, suppose that $S = \{0, 2, 4\}$. Define a coloring $c : V(T) \to \mathbb{Z}_2$ by

$$c(x) = \begin{cases} 0 & \text{if } x \in V_i \text{ and } i \text{ is odd,} \\ 1 & \text{if } x \in V_i \text{ and } i \text{ is even.} \end{cases}$$

Then $c'(x) = c(x)$ for each $x \in V(T)$. Next, suppose that S is one of $\{0, 4, 5\}$ and $\{2, 3, 4\}$. If $S = \{0, 4, 5\}$, then define a coloring $c : V(T) \to \mathbb{Z}_2$ by

$$c(x) = \begin{cases} 0 & \text{if } x \in V_i \text{ where } i \equiv 0, 1, 5 \pmod 6, \\ 1 & \text{if } x \in V_i \text{ where } i \equiv 2, 3, 4 \pmod 6. \end{cases}$$

If $S = \{2, 3, 4\}$, then define a coloring $c : V(T) \to \mathbb{Z}_2$ by

$$c(x) = \begin{cases} 0 & \text{if } x \in V_i \text{ where } i \equiv 3, 4, 5 \pmod 6, \\ 1 & \text{if } x \in V_i \text{ where } i \equiv 0, 1, 2 \pmod 6. \end{cases}$$

In either case, $c'(x) = 0$ if $x \in V_i$ and i is even and $c'(x) = 1$ if $x \in V_i$ and i is odd. Finally, suppose that $S = \{3, 4, 5\}$. Define a coloring $c : V(T) \to \mathbb{Z}_2$ by

$$c(x) = \begin{cases} 0 & \text{if } x \in V_i \text{ where } i \equiv 0, 4, 5 \pmod 6, \\ 1 & \text{if } x \in V_i \text{ where } i \equiv 1, 2, 3 \pmod 6. \end{cases}$$

Then $c'(x) = 0$ if $x \in V_i$ and i is odd and $c'(x) = 1$ if $x \in V_i$ and i is even. In each case, c is a closed modular 2-coloring of T and so $\overline{\mathrm{mc}}(T) = 2$.

Case 2: $|S| = 2$. Then S is a 2-element subset of one of the sets $\{0, 2, 4\}$, $\{0, 4, 5\}$, $\{2, 3, 4\}$, $\{3, 4, 5\}$ in Case 1. Observe that the closed modular 2-colorings described in Case 1 will provide closed modular 2-colorings for this case. For example, if S is a 2-element subset of $S' = \{0, 2, 4\}$, then a closed modular 2-coloring of a tree of type S' described in Case 1 provides a closed modular 2-coloring of T. Therefore, $\overline{\mathrm{mc}}(T) = 2$ in this case as well. □

References

[Addario-Berry et al. 2005] L. Addario-Berry, R. E. L. Aldred, K. Dalal, and B. A. Reed, "Vertex colouring edge partitions", *J. Combin. Theory Ser. B* **94**:2 (2005), 237–244. MR 2006e:05057 Zbl 1074.05031

[Chartrand and Zhang 2009] G. Chartrand and P. Zhang, *Chromatic graph theory*, Chapman & Hall, Boca Raton, FL, 2009. MR 2009k:05003 Zbl 1169.05001

[Chartrand et al. 1988] G. Chartrand, M. S. Jacobson, J. Lehel, O. R. Oellermann, S. Ruiz, and F. Saba, "Irregular networks", pp. 197–210 in *250th Anniversary Conference on Graph Theory* (Fort Wayne, IN, 1986), edited by K. S. Bagga et al., Congr. Numer. **64**, 1988. MR 90b:05073 Zbl 0671.05060

[Chartrand et al. 2010] G. Chartrand, F. Okamoto, and P. Zhang, "The sigma chromatic number of a graph", *Graphs Combin.* **26**:6 (2010), 755–773. MR 2011h:05087 Zbl 1207.05049

[Chartrand et al. 2011] G. Chartrand, L. Lesniak, and P. Zhang, *Graphs & digraphs*, 5th ed., Chapman & Hall, Boca Raton, FL, 2011. MR 2012c:05001 Zbl 1211.05001

[Chartrand et al. 2012] G. Chartrand, B. Phinezy, and P. Zhang, "On closed modular colorings of regular graphs", *Bull. Inst. Combin Appl.* **66** (2012), 7–32.

[Escuadro et al. 2007] H. Escuadro, F. Okamoto, and P. Zhang, "Vertex-distinguishing colorings of graphs: a survey of recent developments", *AKCE Int. J. Graphs Comb.* **4**:3 (2007), 277–299. MR 2384885 Zbl 1143.05305

[Gallian 1998] J. A. Gallian, "A dynamic survey of graph labeling", *Electron. J. Combin.* **5** (1998), Dynamic Survey 6, 43 pp. MR 99m:05141 Zbl 0953.05067

[Golomb 1972] S. W. Golomb, "How to number a graph", pp. 23–37 in *Graph theory and computing*, edited by R. C. Read, Academic, New York, 1972. MR 49 #4863 Zbl 0293.05150

[Jones et al. 2012] R. Jones, K. Kolasinski, and P. Zhang, "A proof of the modular edge-graceful trees conjecture", *J. Combin. Math. Combin. Comput.* **80** (2012), 445–455. MR 2918802 Zbl 1247.05207

[Jones et al. 2013] R. Jones, K. Kolasinski, F. Okamoto, and P. Zhang, "On modular edge-graceful graphs", *Graphs and Combin.* **29**:4 (2013), 901–912.

[Jothi 1991] R. B. G. Jothi, *Topics in graph theory*, Ph.D. thesis, Madurai Kamaraj University, 1991.

[Karoński et al. 2004] M. Karoński, T. Łuczak, and A. Thomason, "Edge weights and vertex colours", *J. Combin. Theory Ser. B* **91**:1 (2004), 151–157. MR 2004k:05087 Zbl 1042.05045

[Phinezy and Zhang 2013] B. Phinezy and P. Zhang, "On closed modular colorings of trees", *Discuss. Math. Graph Theory* **33**:2 (2013), 411–428.

[Rosa 1967] A. Rosa, "On certain valuations of the vertices of a graph", pp. 349–355 in *Theory of graphs* (Rome, 1966), edited by P. Rosenstiehl, Gordon and Breach, New York, 1967. MR 36 #6319 Zbl 0193.53204

Received: 2012-02-01 Accepted: 2012-08-09

bryan.a.phinezy@wmich.edu *Department of Mathematics, Western Michigan University,*
Kalamazoo, MI 49008, United States

Ping.zhang@wmich.edu *Department of Mathematics, Western Michigan University,*
1903 W. Michigan Avenue, Kalamazoo, MI 49008,
United States

INVOLVE 6:1 (2013)
dx.doi.org/10.2140/involve.2013.6.99

Iterations of quadratic polynomials over finite fields

William Worden

(Communicated by Michael Zieve)

Given a map $f : \mathbb{Z} \to \mathbb{Z}$ and an initial argument α, we can iterate the map to get a finite forward orbit modulo a prime p. In particular, for a quadratic map $f(z) = z^2 + c$, where c is constant, work by Pollard suggests that the forward orbit should have length on the order of $\sqrt{p}$. We give a heuristic argument that suggests that the statistical properties of this orbit might be very similar to the birthday problem random variable X_n, for an $n = p$ day year, and offer considerable experimental evidence that the limiting distribution of the orbit lengths, divided by $\sqrt{p}$, for $p \leq x$ as $x \to \infty$, converges to the limiting distribution of $X_n/\sqrt{n}$, as $n \to \infty$.

1. Introduction

Let $f \in \mathbb{Z}[z]$ be a polynomial and let $\alpha \in \mathbb{Z}$. We define the orbit of α under f to be

$$\mathbb{O}_f(\alpha) = \{f^n(\alpha) : n = 0, 1, 2, 3, \dots\},$$

and for each prime p we define the orbit modulo p of α under f to be

$$\mathbb{O}_f^p(\alpha) = \{f^n(\alpha) \bmod p : n = 0, 1, 2, 3, \dots\},$$

where f^n is the n-th iterate of f:

$$f^n = \underbrace{f \circ f \circ \cdots \circ f}_{n},$$

and $f^0(\alpha) = \alpha$. For a fixed f and α and a given prime p, let m_p be the size of $\mathbb{O}_f^p(\alpha)$.

If f is a random map, i.e., a map chosen from the uniformly distributed set consisting of all maps from $\mathbb{F}_p$ into $\mathbb{F}_p$ (see [Harris 1960]), then the values of $f^n(\alpha)$

MSC2010: primary 37P05; secondary 11B37.

Keywords: arithmetic dynamics, birthday problem, forward orbit modulo p, random maps.

This research was conducted while the author was an undergraduate at The City College of New York, CUNY. It was funded by a Rich Summer Internship grant from the Dr. Barnett and Jean-Hollander Rich Scholarship Fund. We are very grateful for the opportunity that this grant allowed, and would like to thank the selection committee and those who have made contributions to the fund.

are uniformly distributed for all n, and all α, and so the probability that $f^0(\alpha)$, $f^1(\alpha)$, $f^2(\alpha), \dots, f^k(\alpha)$ are all different is

$$1 \cdot \frac{p-1}{p} \cdot \frac{p-2}{p} \cdot \dots \cdot \frac{p-k}{p} = \frac{(p-1)!}{p^k(p-k-1)!},$$

since, once α is fixed, there are $p-1$ choices for $f^1(\alpha)$, $p-2$ choices for $f^2(\alpha)$, and so on. Therefore, in this case the probability that (at least) two of $f^0(\alpha)$, $f^1(\alpha)$, $f^2(\alpha), \dots, f^k(\alpha)$ are equal is

$$q_k^{(p)} = 1 - \frac{(p-1)!}{p^k(p-k-1)!}.$$

By an analogous argument, $q_k^{(p)}$ is also the probability that, among k people, two people have the same birthday, where p is the number of days in a year. Framing this a little differently, we let the random variable X_n be the number of times that we must sample (uniformly, with replacement) from the set $\{1, 2, 3, \dots, n\}$ to get a repetition. Since it is known that the expected value of this variable is on the order of $\sqrt{n}$, we look instead at the variable $X_n/\sqrt{n}$.

In light of the above heuristic, we might expect that, for a fixed polynomial f and initial value α, $m_p/\sqrt{p}$ will, on average, "behave" similarly to $X_n/\sqrt{n}$. In particular, we might guess that the limiting distribution of $m_p/\sqrt{p}$, for $p \le x$, $x \to \infty$, will be similar to the limiting distribution of $X_n/\sqrt{n}$, as $n \to \infty$. We note that the above heuristic is not new; similar arguments have been given by Pollard [1975], Bach [1991], and Brent [1980] to name a few, leading to conjectures that m_p is on average approximately equal to $\sqrt{(\pi/2)\,p}$.

We also consider a related question. For a fixed $f \in \mathbb{Z}[z]$, $\alpha \in \mathbb{Z}$, let

$$\mathcal{Q}_{f,\alpha}(x) = \{p \le x : f^n(\alpha) \equiv 0 \pmod{p} \text{ for some } n = 0, 1, 2, \dots\}.$$

That is, $\mathcal{Q}_{f,\alpha}(x)$ is the set of primes p less than or equal to x such that 0 appears in the orbit modulo p of α under f. In particular, we are interested in the size of $\mathcal{Q}_{f,\alpha}(x)$. Since, for a given prime p, the proportion of elements mod p in the orbit of α under f is m_p/p, we hypothesize that $|\mathcal{Q}_{f,\alpha}(x)|$ will grow at a rate proportional to m_p/p. Therefore, if we are correct that m_p will grow at a rate proportional to $\sqrt{p}$, we might expect that

$$|\mathcal{Q}_{f,\alpha}(x)| = \sum_{p \le x} \frac{m_p}{p} \approx c \cdot \frac{\sqrt{x}}{\log x}$$

for some constant $c \in \mathbb{R}$. The approximation above is discussed further in Section 3, where we derive the appropriate constant c.

In the following we take an experimental approach to studying properties of the set $m_p/\sqrt{p}$. For selected maps f and initial values α, we compute the orbits modulo p for all $p \le 2^{25}$. In particular, given these orbits we can find the moments of $m_p/\sqrt{p}$, and the length of $\mathcal{Q}_{f,\alpha}(x)$. As we will demonstrate in the sections to follow, our results give strong support to the above heuristic, and lead us to make the following conjectures:

Conjecture 1. *Let $f(z) = z^2 + c$ and $\alpha \in \mathbb{Z}$ be such that*

(1) $c \in \mathbb{Z} \setminus \{0, -2\}$,

(2) $\alpha \ne \pm\frac{1}{2}(1 \pm \sqrt{1-4c})$, $\quad \alpha \ne \pm\frac{1}{2}(1 \pm \sqrt{-3-4c})$, $\quad \alpha \ne 0, \pm 1$ *when* $c = -1$,

and let the orbit length m_p be as defined above. Then, as $x \to \infty$, the distribution of $m_p/\sqrt{p}$ converges, independent of f and α, to a continuous distribution $F(t) = 1 - e^{-t^2/2}$, $t \ge 0$. In particular, the r-th moments of $m_p/\sqrt{p}$ are given by $\mu_r = r(r-2)(r-4)\cdots 2$ for r even, and $\mu_r = r(r-2)(r-4)\cdots 1 \cdot \sqrt{\frac{\pi}{2}}$ for r odd.

The motivation for the result conjectured above is elaborated upon in Section 2, and the need to include conditions (1) and (2) for both conjectures is explained in Section 4.

Conjecture 2. *Let $f(z) = z^2 + c$ and $\alpha \in \mathbb{Z}$ be such that conditions* (1) *and* (2) *of Conjecture 1 hold, and $\alpha^2 \ne -c$. Define*

$$\mathcal{Q}_{f,\alpha}(x) = \{p \le x : f^n(\alpha) \equiv 0 \pmod{p} \text{ for some } n \ge 0\}.$$

Then

$$\lim_{x\to\infty} |\mathcal{Q}_{f,\alpha}(x)| \frac{\log x}{\sqrt{x}} = \sqrt{2\pi}.$$

2. Length of the orbit modulo p and the birthday problem

Let E_k be the k-th number drawn uniformly from the set $\{1, 2, 3, \ldots, n\}$, with replacement, and let X_n be as defined in Section 1. Then for $k \le n$ we have

$$\begin{aligned} P(X_n > k) &= P(E_1, \ldots, E_k \text{ all take different values}) \\ &= \prod_{j=2}^{k} \bigl(1 - P(E_j = E_i \text{ for some } i < j)\bigr) \\ &= \prod_{j=2}^{k} \left(1 - \frac{j-1}{n}\right) = \exp \sum_{j=1}^{k-1} \log(1 - j/n). \end{aligned}$$

So as $n \to \infty$, we have the following for $0 \le t \le \sqrt{n}$:

$$\begin{aligned} \lim_{n\to\infty} P(X_n/\sqrt{n} > t) = \lim_{n\to\infty} P(X_n > t\sqrt{n}) &= \lim_{n\to\infty} \exp \sum_{j=1}^{\lfloor t\sqrt{n} \rfloor} \log(1 - j/n) \\ &= \lim_{n\to\infty} \exp\left(-\sum_{j=1}^{\lfloor t\sqrt{n} \rfloor} \sum_{k=1}^{\infty} \frac{(j/n)^k}{k}\right), \end{aligned}$$

where we have used the power series representation for $\log(1-j/n)$ in the third line. Switching the order of summation, and pulling the first term of the sum over k out of the exponential, we have

$$\begin{aligned}
&\lim_{n\to\infty} P(X_n/\sqrt{n} > t)\\
&\qquad = \lim_{n\to\infty} \exp\Biggl(-\sum_{j=1}^{\lfloor t\sqrt{n}\rfloor} j/n\Biggr)\cdot \lim_{n\to\infty} \exp\Biggl(-\sum_{k=2}^{\infty}\sum_{j=1}^{\lfloor t\sqrt{n}\rfloor} \frac{(j/n)^k}{k}\Biggr)\\
&\qquad \approx \lim_{n\to\infty} \exp\biggl(-\frac{t\sqrt{n}(t\sqrt{n}+1)}{2n}\biggr)\cdot \lim_{n\to\infty} \exp\biggl(-\sum_{k=1}^{\infty} O\Bigl(\frac{t^{k+2}}{kn^{k/2}}\Bigr)\biggr)\\
&\qquad \approx e^{-t^2/2}\cdot \exp\sum_{k=1}^{\infty}\lim_{n\to\infty} O\Bigl(\frac{t^{k+2}}{kn^{k/2}}\Bigr) = e^{-t^2/2},
\end{aligned}$$

where the second line follows because, in general, $\sum_{j=1}^{m} j^k$ is a polynomial in m of degree $k+1$, and the third line, where we have brought the limit inside the sum, follows from the monotone convergence theorem. Therefore

$$\lim_{n\to\infty} P(X_n/\sqrt{n} \le t) = 1 - e^{-t^2/2},$$

so we see that the distribution of $X_n/\sqrt{n}$ converges to a distribution function $F(t) = 1-e^{-t^2/2}$, which has an associated density function $f(t) = F'(t) = te^{-t^2/2}$. To support our conjecture in Section 1 — that $F(t)$ is the limiting distribution of $m_p/\sqrt{p}$, as $x\to\infty$ — we compare the moments of $m_p/\sqrt{p}$, which we compute in Section 5 for large x, to the limiting moments of $X_n/\sqrt{n}$, as $n\to\infty$. With the limiting density function $f(t)$ of $X_n/\sqrt{n}$ in hand we can derive a general expression for the r-th moment:

$$\begin{aligned}
\mu_r &= \int_0^\infty t^r f(t)\,\mathrm{d}t = \int_0^\infty t^{r+1}e^{-t^2/2}\,\mathrm{d}t\\
&= -t^r e^{-t^2/2}\big|_0^\infty + r\int_0^\infty t^{r-1}e^{-t^2/2}\,\mathrm{d}t = r\int_0^\infty t^{r-1}e^{-t^2/2}\,\mathrm{d}t,
\end{aligned}$$

where r applications of l'Hôpital's rule give us 0 for the $-t^r e^{-t^2/2}$ term. We continue applying integration by parts as above until we get

$$\begin{aligned}
\mu_r &= r(r-2)(r-4)\cdots 2\cdot\int_0^\infty te^{-t^2/2}\,\mathrm{d}t \quad \text{if } r \text{ is even},\\
\mu_r &= r(r-2)(r-4)\cdots 1\cdot\int_0^\infty e^{-t^2/2}\,\mathrm{d}t \quad \text{if } r \text{ is odd}.
\end{aligned}$$

The first integral above evaluates to $-e^{-t^2/2}\big|_0^\infty = 1$, and the second integral we evaluate as follows:

$$
\begin{aligned}
& & I &= \int_0^\infty e^{-t^2/2}\,\mathrm{d}t \\
&\Longrightarrow & (2I)^2 &= \left(\int_{-\infty}^\infty e^{-t^2/2}\,\mathrm{d}t\right)^2 \\
& & &= \int_{-\infty}^\infty e^{-x^2/2}\,\mathrm{d}x \cdot \int_{-\infty}^\infty e^{-y^2/2}\,\mathrm{d}y \\
& & &= \int_{-\infty}^\infty \int_{-\infty}^\infty e^{-(x^2+y^2)/2}\,\mathrm{d}x\,\mathrm{d}y \\
& & &= \int_{r=0}^\infty \int_{\theta=0}^{2\pi} re^{-r^2/2}\,\mathrm{d}r\,\mathrm{d}\theta = 2\pi \\
&\Longrightarrow & I &= \sqrt{\frac{\pi}{2}}.
\end{aligned}
$$

Therefore the r-th moments of the limiting distribution of $X_n/\sqrt{n}$, as $n \to \infty$, are given by

$$
\begin{aligned}
\mu_r &= r(r-2)(r-4)\cdots 2 && \text{if } r \text{ is even,} \\
\mu_r &= r(r-2)(r-4)\cdots 1\cdot\sqrt{\pi/2} && \text{if } r \text{ is odd.}
\end{aligned}
$$

For the first four moments this gives us $\mu_1 = \sqrt{\pi/2}$, $\mu_2 = 2$, $\mu_3 = 3\sqrt{\pi/2}$, $\mu_4 = 8$. Therefore, to support our claim in Conjecture 1 we must provide evidence that the moments of $m_p/\sqrt{p}$ are converging, as $x \to \infty$, to the moments μ_r above. In our computations we use the following expression for the r-th moments of $m_p/\sqrt{p}$:

$$M_r = \frac{1}{|\{p \le x\}|}\sum_{p\le x}\left(\frac{m_p}{\sqrt{p}}\right)^r.$$

3. Iterates of f congruent to zero modulo p

In this section we consider the quantity $|\mathcal{Q}_{f,\alpha}(x)|(\log x)/\sqrt{x}$, as defined in Section 1. Assuming that the probability that $0 \in \mathcal{O}_f^p(\alpha)$ is m_p/p, and that M_1 will converge to $\sqrt{\pi/2}$, we define

$$G(x) = \frac{\log x}{\sqrt{x}}\sum_{p\le x}\frac{\sqrt{\pi/2}}{\sqrt{p}},$$

and make a guess that

$$\lim_{x\to\infty}|\mathcal{Q}_{f,\alpha}(x)|\frac{\log x}{\sqrt{x}} = \lim_{x\to\infty} G(x). \tag{1}$$

If we let $\pi(x) = \sum_{k\le x} a(k)$, where $a(k) = 1$ if k is prime and 0 otherwise, and define $f(x) = 1/\sqrt{x}$, then Stieltjes integration by parts gives

$$\sum_{p \le x} \frac{1}{\sqrt{p}} = \frac{\pi(x)}{\sqrt{x}} - \frac{1}{\sqrt{2}} + \frac{1}{2} \int_2^x \frac{\pi(t)}{t^{3/2}} \, \mathrm{d}t,$$

which implies

$$\lim_{x\to\infty} \frac{\log x}{\sqrt{x}} \sum_{p \le x} \frac{1}{\sqrt{p}} = 1 + \lim_{x\to\infty} \frac{\log x}{2\sqrt{x}} \int_2^x \frac{\pi(t)}{t^{3/2}} \, \mathrm{d}t. \tag{2}$$

Now, for $x \ge 55$, we can bound $\pi(x)$ by the inequalities

$$\frac{x}{\log x + 2} < \pi(x) < \frac{x}{\log x - 4};$$

see [Rosser 1941]. Hence, if we shift the lower limit of integration in (2) to 55, changing the value of the integral only by an additive constant which will vanish in the limit, we can write

$$\begin{aligned} \lim_{x\to\infty} \frac{\log x}{2\sqrt{x}} \int_{55}^x \frac{1}{\sqrt{t}(\log t + 2)} \, \mathrm{d}t &\le \lim_{x\to\infty} \frac{\log x}{2\sqrt{x}} \int_{55}^x \frac{\pi(t)}{t^{3/2}} \, \mathrm{d}t \\ &\le \lim_{x\to\infty} \frac{\log x}{2\sqrt{x}} \int_{55}^x \frac{1}{\sqrt{t}(\log t - 4)} \, \mathrm{d}t. \end{aligned} \tag{3}$$

Consider the limit on the left. The integral diverges, since the integrand exceeds $1/t$ everywhere — indeed, $\sqrt{t}(\log t + 2) < t$ for $t \ge 55$. Hence the limit has the form ∞/∞, where the denominator comes from expressing the quotient before the integral as the inverse of $2\sqrt{x}/\log x$. It follows that the limit on the left equals

$$\lim_{x\to\infty} \frac{1}{\sqrt{x}(\log x + 2)} \cdot \frac{\sqrt{x}\log^2 x}{\log x - 2} = \lim_{x\to\infty} \frac{\log^2 x}{\log^2 x - 4} = 1.$$

An analogous reasoning shows that the rightmost limit in (3) is equal to

$$\lim_{x\to\infty} \frac{1}{\sqrt{x}(\log x - 4)} \cdot \frac{\sqrt{x}\log^2 x}{\log x - 2} = \lim_{x\to\infty} \frac{\log^2 x}{\log^2 x - 6\log x + 8} = 1.$$

Therefore so is the limit in the middle. In other words, $\lim\limits_{x\to\infty} G(x) = 2\sqrt{\pi/2} = \sqrt{2\pi}$, and our guess (1) becomes

$$\lim_{x\to\infty} |\mathcal{Q}_{f,\alpha}(x)| \frac{\log x}{\sqrt{x}} = \sqrt{2\pi}.$$

As we test our hypothesis, it should be kept in mind that $\lim_{x\to\infty} G(x)$ converges very slowly. Since the values of x for which $|\mathcal{Q}_{f,\alpha}(x)|(\log x)/\sqrt{x}$ can actually be computed (in a reasonable amount of time) are relatively small, the largest being 2^{27}, we compare our computations to $G(x)$, rather than the limit $\sqrt{2\pi}$.

4. Some special cases

In this paper we consider polynomials of the form $f(z) = z^2 + c$, with $z, c \in \mathbb{Z}$, and initial argument values $\alpha \in \mathbb{Z}$. But for certain f, α pairs we find that we end up with a finite (over $\mathbb{Z}$) orbit, a condition which is clearly incompatible with our hypotheses outlined in Sections 2 and 3, since m_p will have a fixed bound for all primes p. In this section we classify these exceptional pairs f, α.

Proposition 1. *Let* $\mathbb{O}_f(\alpha) = \{f^n(\alpha) : n = 0, 1, 2, 3, \dots\}$ *be the orbit of* α *under* f, *where* $f(z) = z^2 + c$, $c \in \mathbb{Z}$, *and* $\alpha \in \mathbb{Z}$. *Then* $\mathbb{O}_f(\alpha)$ *is finite if and only if one of the following hold*:

$$\begin{aligned} &\text{(i)} && \alpha = \pm\tfrac{1}{2}(1 \pm \sqrt{1-4c}),\\ &\text{(ii)} && \alpha = \pm\tfrac{1}{2}(1 \pm \sqrt{-3-4c}),\\ &\text{(iii)} && \alpha \in \{0, 1, -1\} \quad \textit{and} \quad c \in \{0, -1, -2\}. \end{aligned}$$

Proof. First we prove the converse, which is easier. Assumption (i) gives us the solutions to $\alpha^2 \pm \alpha + c = 0$, and this equation implies that $\alpha^2 + c = \pm\alpha$, which implies that the orbit is finite. Assumption (ii) gives the solutions to $\alpha^2 \pm \alpha + c + 1 = 0$, and this equation implies that $\alpha^2 + c = \pm\alpha - 1$. With one more iteration we get

$$(\alpha^2 + c)^2 + c = (\pm\alpha - 1)^2 + c = \alpha^2 \mp 2\alpha + 1 - \alpha^2 \pm \alpha - 1 = \mp 2\alpha \pm \alpha = \pm\alpha,$$

which again implies that the orbit is finite. As for (iii), testing all possible α, c combinations will quickly convince the reader that the orbits are finite in all cases.

Now suppose that $\mathbb{O}_f(\alpha)$ is finite. First we make some simplifications. Since the orbits of α and $-\alpha$ will be identical except for the sign of the first element $f^0 = \alpha$, we may consider only nonnegative values of α. Also, since it is obvious that $c \in \{0, -1\}$ will have infinite orbit for $\alpha \geq 2$, and that $c \geq 1$ will have infinite orbit for all α, we consider only $c \leq -2$. We claim that $\mathbb{O}_f(\alpha)$ finite implies $\sqrt{-c} - 1 < \alpha < \sqrt{-c} + 1$. If this were not true, then we would have either $\alpha = \lceil\sqrt{-c}\rceil + b$ or $\alpha = \lfloor\sqrt{-c}\rfloor - b$ for some $b \in \mathbb{N}$, giving us

$$\alpha = \lceil\sqrt{-c}\rceil + b \implies \alpha^2 + c = (\lceil\sqrt{-c}\rceil)^2 + 2b\lceil\sqrt{-c}\rceil + b^2 + c > \lceil\sqrt{-c}\rceil + b,$$
$$\alpha = \lfloor\sqrt{-c}\rfloor - b \implies \alpha^2 + c = (\lfloor\sqrt{-c}\rfloor)^2 - 2b\lfloor\sqrt{-c}\rfloor + b^2 + c < -2b\lfloor\sqrt{-c}\rfloor + c.$$

The first of these immediately implies that the iterates of f are unbounded since they are strictly increasing. In the second case iterating once more gives us

$$(\alpha^2 + c)^2 + c > 4b^2\lfloor\sqrt{-c}\rfloor^2 - 4bc\lfloor\sqrt{-c}\rfloor + c^2 > \lfloor\sqrt{-c}\rfloor + b,$$

where the inequality reverses since $\alpha^2 + c < -2b\lfloor\sqrt{-c}\rfloor + c < 0$, and the second inequality follows since $c \leq -2$. Again we can conclude that the iterates of f are unbounded, and so we have shown that $\mathbb{O}_f(\alpha)$ finite implies $\sqrt{-c} - 1 < \alpha < \sqrt{-c} + 1$.

For any c, there are at most two integers that satisfy the preceding inequality, $\lfloor\sqrt{-c}\rfloor$ and $\lceil\sqrt{-c}\rceil$, so any member of $\mathbb{O}_f(\alpha)$ must be one of $\pm\lfloor\sqrt{-c}\rfloor$, $\pm\lceil\sqrt{-c}\rceil$, since otherwise the iterates of f will be unbounded. Since we know $\alpha \in \mathbb{O}_f(\alpha)$, the condition above implies that $\mathbb{O}_f(\alpha) \subset \{\alpha, -\alpha, \alpha - 1, -\alpha - 1\}$ or $\mathbb{O}_f(\alpha) \subset \{\alpha, -\alpha, \alpha + 1, -\alpha + 1\}$. However, we can rule out the latter case since

$$\begin{aligned} & \alpha^2 + c = \pm\alpha + 1 \\ \Longrightarrow\quad & (\alpha^2 + c)^2 + c = \pm 3\alpha + 2 \\ \Longrightarrow\quad & ((\alpha^2 + c)^2 + c)^2 + c = 7\alpha^2 \pm 13\alpha + 5 > 2\alpha + 5 > \pm\alpha + 1 > \pm\alpha, \end{aligned}$$

where the first inequality follows since in this case $c \le -2 \Longrightarrow \alpha \ge 2$. Therefore the iterates are unbounded in this case, and we are left with the following:

$$\begin{aligned} & \alpha^2 + c = \pm\alpha && \text{or}\quad \alpha^2 + c = \pm\alpha - 1 \\ \Longrightarrow\quad & \alpha^2 \pm \alpha + c = 0 && \text{or}\quad \alpha^2 \pm \alpha + c + 1 = 0 \\ \Longrightarrow\quad & \alpha = \pm\tfrac{1}{2}(1 \pm \sqrt{1 - 4c}) && \text{or}\quad \alpha = \pm\tfrac{1}{2}(1 \pm \sqrt{-3 - 4c}). \end{aligned}$$ □

This proposition is the basis for the second condition necessary for Conjectures 1 and 2; we now turn to the first condition, that $c \notin \{0, -2\}$. These two cases behave strikingly differently from the others studied, because they come from homomorphisms of the multiplicative group. The map $z^2 - 2$ is a Chebyshev polynomial, and so is connected to z^2 via a homomorphism from $\mathbb{C}^*$ to $\mathbb{C}^*/\{z \sim z^{-1}\}$ [Silverman 2007, pp. 29–30]. On a finite field $\mathbb{F}_p$, this means that the behavior of $z^2 - 2$ will be very similar to that of z^2. For $c = 0$ it is clear that $|\mathcal{Q}_{f,\alpha}(x)|$ will not grow as expected, since we'll have $p \in \mathcal{Q}_{f,\alpha}(x)$ if and only if p divides α. On the other hand, the length m_p of the orbit modulo p will grow much faster than we expect. Vasiga and Shallit [2004] studied these two cases in some depth, showing that, for a given prime p, if $(p-1)/2$ is prime and 2 is a primitive root modulo $(p-1)/2$, then $\sum_{0\le\alpha<p} m_p$ is at least on the order of p^2. Heuristics by Hardy and Littlewood [1923], along with Artin's conjecture, suggest that the number of primes less than x that satisfy this property is on the order of $x/(\log x)^2$, and thus the density of these primes is on the order of $1/\log x$. If we sum p^2, for $p \le x$, and multiply by $1/\log x$, we get something on the order of $x^3/(\log x)^2$, and dividing this by the sum $\sum_{p\le x}\sum_{0\le\alpha<p} 1 \sim x^2/\log x$ gives us an average orbit length on the order of $x/\log x$. Note that this estimate only takes into account primes with the aforementioned property, and assumes that all other primes have orbit length 0, so we should expect this to be a low estimate. Indeed, the limited experimentation we did on this question suggests that the average orbit length is closer to $x/(\log x)^{3/4}$.

Finally, Conjecture 2 requires an additional condition, that $\alpha^2 \ne -c$. If we disregard this condition we will have cases where $0 \in \mathcal{Q}_{f,\alpha}(x)$ for all p, which clearly conflicts with our claim. To see that the $f^0 = \alpha$ is the only iterate whose square can be equal to $-c$, suppose that the contrary is true, i.e., that we have $(f^l)^2 = -c$ for some $l \in \mathbb{Z}$; then, letting $f^k = f^{l-1}$, we have

$$\begin{aligned} \bigl((f^k)^2 + c\bigr)^2 + c &= 0, \\ (f^k)^4 + 2c(f^k)^2 + c^2 + c &= 0, \\ c^2 + (2(f^k)^2 + 1)c + (f^k)^4 &= 0. \end{aligned}$$

Therefore the quadratic formula gives us

$$c = \frac{-2(f^k)^2 - 1 \pm \sqrt{(2f^k)^2 + 1}}{2},$$

which is not an integer unless $f^k = 0$, in which case $(f^l)^2 = c^2 = -c \implies c = -1$. It is easy to see that this implies $\alpha \in \{0, 1\}$, and this case has already been excluded by Proposition 1(iii).

5. Results

First we consider the first four moments of $m_p/\sqrt{p}$, as discussed in Section 2, for $f(z) = z^2 + c$, where $c = \pm 1, +2, \pm 3$, and initial arguments $\alpha = 1, 2, \dots, 9$. Of these we can exclude $\alpha = 1, 2$ when $f(z) = z^2 - 3$, and $\alpha = 1$ when $f(z) = z^2 - 1$, because these (f, α) combinations have finite orbits, as discussed above. For the other 42 combinations, we find that our experimental results support our hypotheses very well. For the first moment we expected the limit to be $\sqrt{\pi/2} = 1.25331413\dots$, and for all (f, α) tested, M_1 was between 1.25138 and 1.25351 for $x = 2^{25}$, with an average value of 1.25279. Table 1 gives these figures along with the standard

	mean	stand dev	min	max
M_1	1.252795789	0.000518158	1.251387582	1.253505370
$\lvert\sqrt{\pi/2} - M_1\rvert$	0.000544827	0.000490241	0.000000052	0.001926555
M_2	1.998325027	0.001690776	1.993860194	2.000539507
$\lvert 2 - M_2\rvert$	0.001810403	0.001544894	0.000034079	0.006139806
M_3	3.755044605	0.004998323	3.742341997	3.762285912
$\lvert 3\sqrt{\pi/2} - M_3\rvert$	0.005419269	0.004427558	0.000121838	0.017600415
M_4	7.985456401	0.014915109	7.948531018	8.008817811
$\lvert 8 - M_4\rvert$	0.016278430	0.012999594	0.000149649	0.051468982

Table 1. Moments of $m_p/\sqrt{p}$ for $x = 2^{25}$ and distance from predicted limit. For comparison, $\sqrt{\pi/2} \sim 1.25331413731550$.

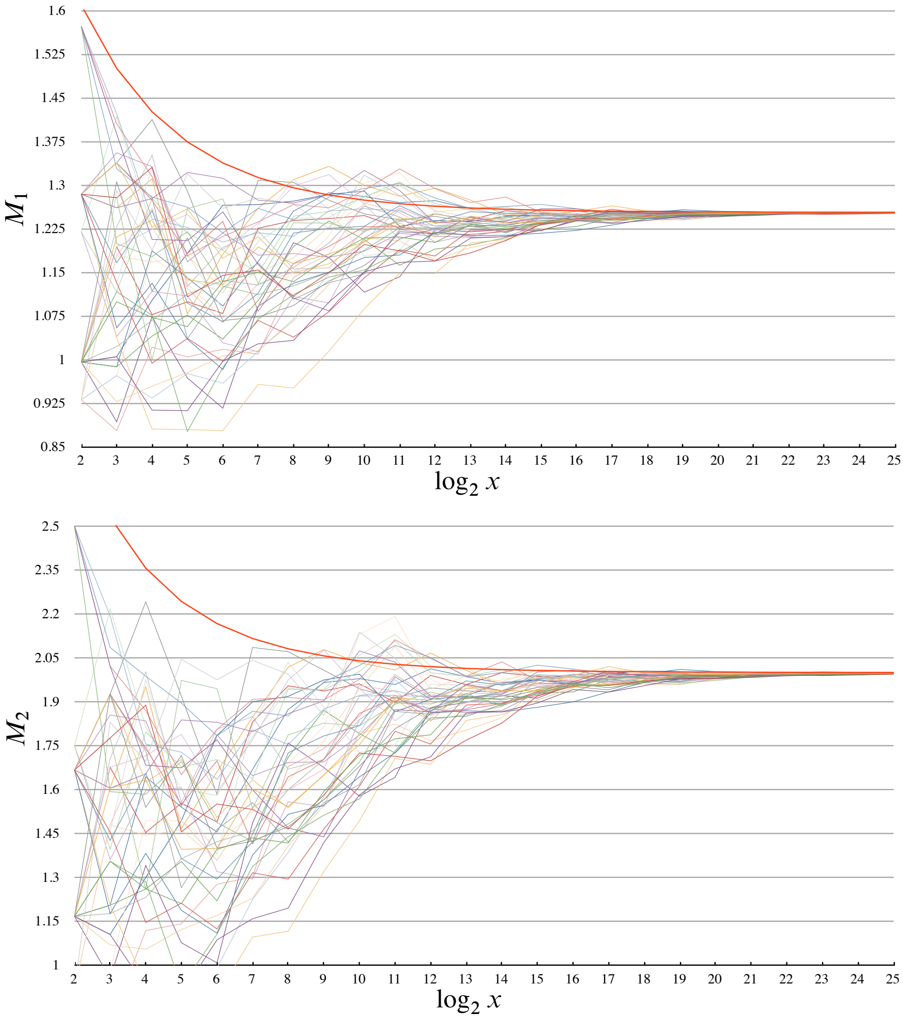


Figure 1. The first and second moments, M_1 and M_2, of $X/\sqrt{n}$ (thicker red lines) and $m_p/\sqrt{p}$ (thin lines) for all (f, α) tested.

deviation of the set of results for each moment. It also shows the mean, standard deviation, minimum, and maximum of the set

$$\{|\sqrt{\pi/2} - M_1| : x = 2^{25}, \text{for } (f, \alpha) \text{ tested}\},$$

and similarly for the second, third and fourth moments. Our complete results are depicted graphically in Figures 1 and 2, for the first, second, third, and fourth moments. In each of these graphs the heavier red curve is the respective moment of $X_n/\sqrt{n}$, for $n = x$. Notice that the y-axes of these graphs are not scaled equally with respect to each other (they are stretched by a factor of two for each subsequent

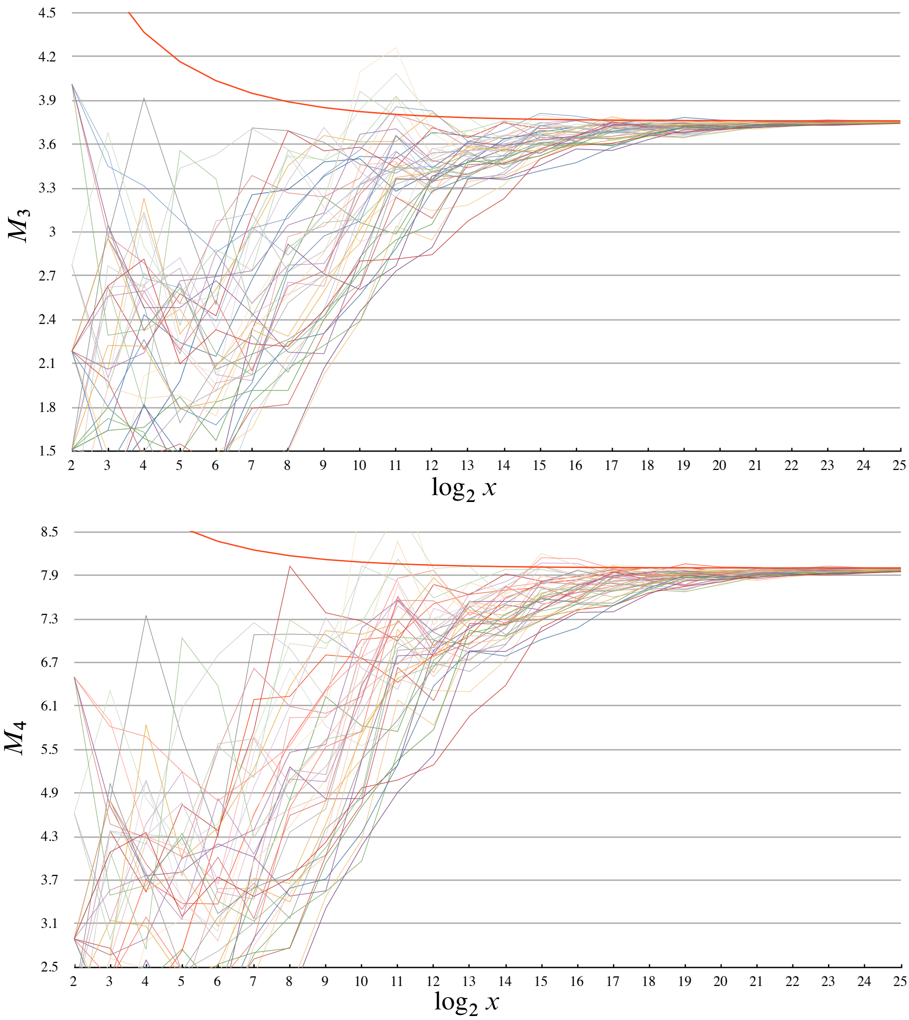


Figure 2. The third and fourth moments, M_3 and M_4, of $X/\sqrt{n}$ (thicker red lines) and $m_p/\sqrt{p}$ (thin lines) for all (f, α) tested.

moment graph), so if we're interested in comparing how quickly two of the moments converge, Table 1 will be more helpful.

The apparent common limit of the moments of $m_p/\sqrt{p}$ and $X/\sqrt{n}$ suggests that the limiting distributions of $m_p/\sqrt{p}$, as $x \to \infty$, and the random variable $X_n/\sqrt{n}$, as $n \to \infty$, are the same. For the variable $X/\sqrt{n}$ we showed in Section 2 that, as $n \to \infty$, the distribution $P(X_n/\sqrt{n} < t)$ converges to the function $F(t) = 1 - e^{-t^2/2}$. As the histogram in Figure 3 shows, the density function $f(t) = F'(t)$ approximates quite well the distribution of $m_p/\sqrt{p}$ for $x = 10^8$, $f(z) = z^2 + 1$, $\alpha = 3$. These results give considerable support to our first conjecture, stated in Section 1.

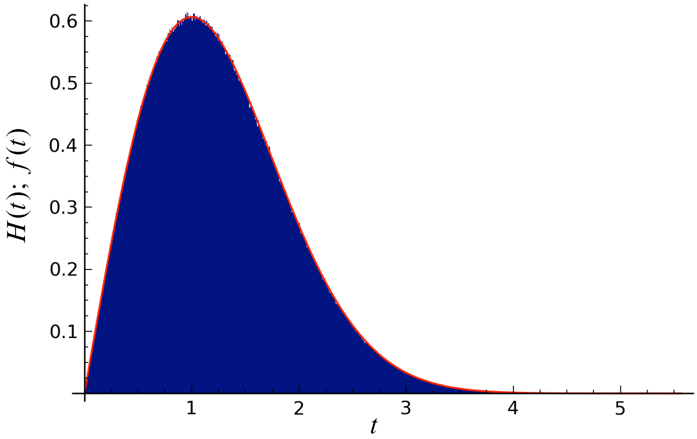


Figure 3. Histogram, $H(t)$, of the distribution of $m_p/\sqrt{p}$ (blue) for $x = 10^8$, $f(z) = z^2 + 1$, $\alpha = 3$, superimposed on the graph of $f(t) = te^{-t^2/2}$ (red). Here

$$H(t : wk \le t < w(k+1)) = \frac{|\{p \le 10^8 : wk \le m_p/\sqrt{p} < w(k+1)\}|}{w \cdot |\{p \le 10^8\}|},$$

for $k \in \mathbb{N}$. Each bar of the histogram has width $w \approx 5.6/800$.

To test the hypothesis discussed in Section 3, we compute $|\mathcal{Q}_{f,\alpha}(x)|(\log x)/\sqrt{x}$ for (f, α) as described above and $x \in \{2, 2^2, \ldots, 2^{27}\}$. Table 2 shows that, although our results are still fairly widely dispersed at $x = 2^{27}$, the average of the results for this x value is very close to $G(x)$, and the standard deviation is decreasing in general as x increases, as is the error of the mean from $G(x)$. As we mentioned earlier, $\lim_{x\to\infty} G(x)$ converges very slowly, and, as the table shows, even for x as large as 2^{27} we still have $|G(x) - \sqrt{2\pi}| \sim 0.36$, so we are not too surprised to see such a wide range in our results for this x value. That is, intuitively, it seems we should not expect our results to be very tightly grouped until we are close to the limiting value, $\sqrt{2\pi}$. Figure 4 gives a graphical representation of all (f, α) tested, for x from 4 to 2^{27}. On this graph the red and blue lines are $G(x)$ and the mean from Table 2, respectively. From this data, it seems reasonable to suppose that $|\mathcal{Q}_{f,\alpha}(x)|(\log x)/\sqrt{x}$ will eventually converge to $\sqrt{2\pi}$, independent of f, α, and so we make our second conjecture as stated in Section 1.

Joseph Silverman [2008] carried out computations that lead to a conjecture (in a more general setting) that under certain restrictions the set $\{p : m_p \le p^{1/2-\epsilon}\}$ will have density 0 for $\epsilon > 0$. This conjecture agrees with our own results, and in fact, if Conjecture 1 were proven, a less general version of Silverman's conjecture would readily follow. Computations of a similar nature to ours were also carried out in [Benedetto et al. 2013], with results that are compatible with our own.

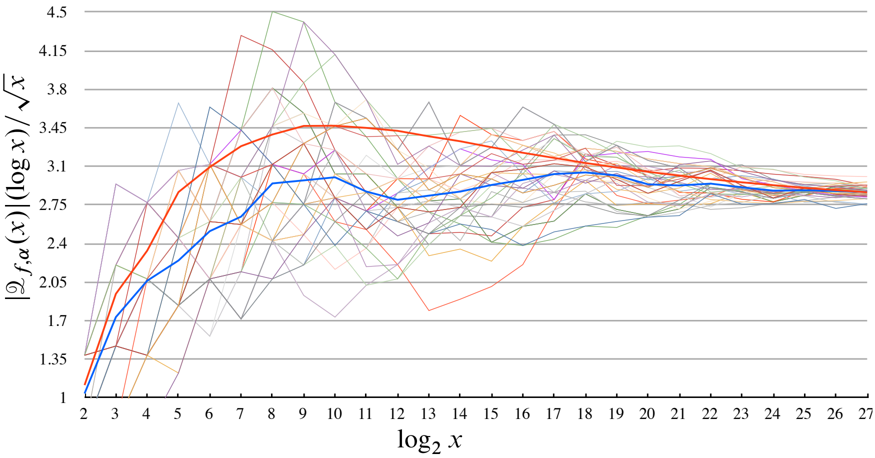


Figure 4. Graphs of $\mathcal{Q}_{f,\alpha}(x)(\log x)/\sqrt{x}$ for all 42 (f,α) combinations tested (thinner lines), the mean of these graphs (thick blue line), and our guess $G(x)$ (thick red line).

Acknowledgements

We would like to thank our advisor Gautam Chinta, whose guidance and assistance throughout the research process were instrumental to the project's success, and whose invaluable feedback during the drafting of this paper improved its quality considerably. In addition, we thank Prof. Hutz for calling our attention to the paper [Benedetto et al. 2013], where the authors carry out similar computations and obtain results compatible with our computations, and we thank Prof. Silverman for directing us to the publication in which his article [Silverman 2008] appeared.

References

[Bach 1991] E. Bach, "Toward a theory of Pollard's rho method", *Inform. and Comput.* **90**:2 (1991), 139–155. MR 92a:11151 Zbl 0716.11065

[Benedetto et al. 2013] R. L. Benedetto, D. Ghioca, B. Hutz, P. Kurlberg, T. Scanlon, and T. J. Tucker, "Periods of rational maps modulo primes", *Math. Ann.* **355**:2 (2013), 637–660. MR 3010142 Zbl 06133902

[Brent 1980] R. P. Brent, "An improved Monte Carlo factorization algorithm", *BIT* **20**:2 (1980), 176–184. MR 82a:10007 Zbl 0439.65001

[Hardy and Littlewood 1923] G. H. Hardy and J. E. Littlewood, "Some problems of "Partitio numerorum", III: On the expression of a number as a sum of primes", *Acta Math.* **44** (1923), 1–70. MR 1555183 Zbl 48.0143.04

[Harris 1960] B. Harris, "Probability distributions related to random mappings", *Ann. Math. Statist.* **31** (1960), 1045–1062. MR 22 #9993 Zbl 0158.34905

[Pollard 1975] J. M. Pollard, "A Monte Carlo method for factorization", *Nordisk Tidskr. Informationsbehandling* (*BIT*) **15**:3 (1975), 331–334. MR 52 #13611 Zbl 0312.10006

		$\lvert\mathcal{Q}_{f,\alpha}(x)\rvert(\log x)/\sqrt{x}$ for all (f,α) tested				
x	$G(x)^a$	mean	stand dev	min	max	$\lvert G(x)-\text{mean}\rvert$
2^{10}	3.46925	3.00157	0.51687	1.73287	4.11556	0.46767
2^{11}	3.45003	2.87221	0.46933	2.02178	3.70660	0.57781
2^{12}	3.42304	2.79734	0.34646	2.07944	3.37909	0.62570
2^{13}	3.37313	2.83502	0.37519	1.79203	3.68363	0.53811
2^{14}	3.32854	2.87187	0.30915	1.89532	3.56321	0.45667
2^{15}	3.27415	2.93202	0.35071	2.01030	3.44622	0.34213
2^{16}	3.22425	2.97785	0.33397	2.20941	3.63902	0.24640
2^{17}	3.17737	3.03080	0.28861	2.44107	3.48260	0.14657
2^{18}	3.13140	3.04548	0.18849	2.55869	3.38722	0.08593
2^{19}	3.09045	3.01797	0.21004	2.54637	3.32847	0.07248
2^{20}	3.05137	2.93775	0.17602	2.63992	3.27620	0.11362
2^{21}	3.01654	2.92737	0.15786	2.65359	3.28683	0.08917
2^{22}	2.98475	2.94397	0.12988	2.71031	3.21664	0.04077
2^{23}	2.95589	2.91037	0.09402	2.72467	3.11548	0.04552
2^{24}	2.92986	2.88060	0.08428	2.76176	3.07449	0.04925
2^{25}	2.90646	2.88289	0.05445	2.77612	3.02741	0.02357
2^{26}	2.88509	2.87751	0.05869	2.71911	3.01390	0.00759
2^{27}	2.86578	2.86821	0.05418	2.74621	3.00790	0.00243

Table 2. A comparison of our experimental results to our guess, $G(x)=\frac{\log x}{\sqrt{x}}\sum\limits_{p\le x}\frac{\sqrt{\pi/2}}{\sqrt{p}}$.

[Rosser 1941] B. Rosser, "Explicit bounds for some functions of prime numbers", *Amer. J. Math.* **63** (1941), 211–232. MR 2,150e Zbl 0024.25004

[Silverman 2007] J. H. Silverman, *The arithmetic of dynamical systems*, Graduate Texts in Mathematics **241**, Springer, New York, 2007. MR 2008c:11002 Zbl 1130.37001

[Silverman 2008] J. H. Silverman, "Variation of periods modulo p in arithmetic dynamics", *New York J. Math.* **14** (2008), 601–616. MR 2009h:11100 Zbl 1153.11028

[Vasiga and Shallit 2004] T. Vasiga and J. Shallit, "On the iteration of certain quadratic maps over GF(p)", *Discrete Math.* **277**:1-3 (2004), 219–240. MR 2004k:05104 Zbl 1045.11086

Received: 2012-03-02 Revised: 2012-04-25 Accepted: 2012-05-10

william.worden@temple.edu
Temple University, Wachman Hall Rm. 517,
1805 N. Broad St., Philadelphia, PA 19122, United States

mathematical sciences publishers msp

INVOLVE 6:1 (2013)
dx.doi.org/10.2140/involve.2013.6.113

Positive solutions to singular third-order boundary value problems on purely discrete time scales

Courtney DeHoet, Curtis Kunkel and Ashley Martin

(Communicated by Johnny Henderson)

We study singular discrete third-order boundary value problems with mixed boundary conditions of the form

$$-u^{\Delta\Delta\Delta}(t_{i-2}) + f\big(t_i, u(t_i), u^{\Delta}(t_{i-1}), u^{\Delta\Delta}(t_{i-2})\big) = 0,$$
$$u^{\Delta\Delta}(t_0) = u^{\Delta}(t_{n+1}) = u(t_{n+2}) = 0,$$

over a finite discrete interval $\{t_0, t_1, \ldots, t_n, t_{n+1}, t_{n+2}\}$. We prove the existence of a positive solution by means of the lower and upper solutions method and the Brouwer fixed point theorem in conjunction with perturbation methods to approximate regular problems.

1. Preliminaries

This paper is something of an extension of [Rachůnková and Rachůnek 2006] and [Kunkel 2006; 2008]. Rachůnková and Rachůnek studied a second-order singular boundary value problem for the discrete p-Laplacian, $\phi_p(x) = |x|^{p-2}x$, $p > 1$. In particular, they dealt with the discrete boundary value problem

$$\Delta\big(\phi_p(\Delta u(t-1))\big) + f(t, u(t), \Delta u(t-1)) = 0, \quad t \in [1, T+1],$$
$$\Delta u(0) = u(T+2) = 0,$$

in which $f(t, x_1, x_2)$ was singular in x_1. In [Kunkel 2006] this was extended to the third-order case, but only for $p = 2$; that is, boundary value problem treated was

$$-\Delta\Delta\Delta u(t-2) + f(t, u(t), \Delta u(t-1), \Delta\Delta u(t-2)) = 0, \quad t \in [2, T+1],$$
$$\Delta\Delta u(0) = \Delta u(T+2) = u(T+3) = 0.$$

In [Kunkel 2008], by contrast, the extension was to a second-order singular discrete

MSC2010: 39A10, 34B16, 34B18.

Keywords: singular discrete boundary value problem, mixed conditions, lower and upper solutions, Brouwer fixed point theorem, approximate regular problems.

boundary value problem with nonuniform step size:

$$u^{\Delta\Delta}(t_{i-1}) + f(t_i, u(t_i), u^{\Delta}(t_{i-1})) = 0, \quad t_i \in [2, T+1],$$
$$u^{\Delta}(t_0) = u(t_{n+1}) = 0.$$

The analysis in the present paper relies heavily on a lower and upper solutions method in conjunction with an application of the Brouwer fixed point theorem [Zeidler 1986]. We consider only the singular third-order boundary value problem, while letting our function range over a discrete interval with nonuniform step size. We will provide definitions of appropriate lower and upper solutions. The lower and upper solutions will be applied to nonsingular perturbations of our nonlinear problem, ultimately giving rise to our boundary value problem by passing to the limit.

Various forms of the lower and upper solutions method have been used extensively in establishing solutions of boundary value problems for finite difference equations. Examples include [Henderson and Kunkel 2006; Kunkel 2006; Rachůnková and Rachůnek 2006]; we mention especially [Jiang et al. 2005], which deals with singular discrete boundary value problems using the method. Other outstanding works where lower and upper solution methods have been employed to obtain solutions of boundary value problems for finite difference equations include [Agarwal et al. 1999; 2003; 2004; 2005; Agarwal and Wong 1997; Cabada 2011; Henderson and Thompson 2002; Kelley and Peterson 2001; O'Regan and El-Gebeily 2008; Pao 1985; Peterson et al. 2004; Zhang et al. 2002].

Singular discrete boundary value problems also have received a good deal of attention. As representative works, we suggest [Agarwal et al. 1999; 2005; 2008; Agarwal and Wong 1997; Akın-Bohner et al. 2003; Atici et al. 2003; Jódar 1987; Jódar et al. 1992; Naidu and Kailasa Rao 1982; Peterson et al. 2004; Rachůnková and Rachůnek 2009; Yuan et al. 2008; Zheng et al. 2011; Zhang et al. 2002].

We now state the definitions that are used in the remainder of the paper.

Definition 1.1. For $0 \le i \le n+2$, let $t_i \in \mathbb{R}$, where $t_0 < t_1 < \cdots < t_{n+1} < t_{n+2}$. Define the discrete intervals

$$\mathbb{T} := [t_0, t_{n+2}] = \{t_0, t_1, \dots, t_{n+1}, t_{n+2}\},$$
$$\mathbb{T}^{\circ} := [t_2, t_{n+1}] = \{t_2, t_3, \dots, t_n, t_{n+1}\}.$$

Definition 1.2. For the function $u : \mathbb{T} \to \mathbb{R}$, define the delta derivative [Bohner and Peterson 2001], u^{Δ}, by

$$u^{\Delta}(t_i) := \frac{u(t_{i+1}) - u(t_i)}{t_{i+1} - t_i}, \quad t_i \in \mathbb{T}^{\circ} \cup \{t_0, t_{n+1}\}.$$

We make note that $u^{\Delta\Delta}(t_i) = (u^{\Delta})^{\Delta}(t_i)$.

Consider the third-order nonlinear discrete dynamic

$$u^{\Delta\Delta\Delta}(t_{i-2})+f(t_i,u(t_i),u^{\Delta}(t_{i-1}),u^{\Delta\Delta}(t_{i-2}))=0,\quad t_i\in\mathbb{T}^{\circ},\tag{1}$$

with mixed boundary conditions

$$u^{\Delta\Delta}(t_0)=u^{\Delta}(t_{n+1})=u(t_{n+2})=0.\tag{2}$$

Our goal is to prove the existence of a positive solution of problem (1), (2).

Definition 1.3. By a solution of problem (1), (2), we mean a function $u:\mathbb{T}^{\circ}\to\mathbb{R}$ such that u satisfies the discrete dynamic (1) on $\mathbb{T}^{\circ}$ and the boundary conditions (2). If $u(t)>0$ for $t\in\mathbb{T}^{\circ}$, we say u is a positive solution of the problem (1), (2).

Definition 1.4. Let $\mathfrak{D}\subseteq\mathbb{R}^3$. We say that f is continuous on $\mathbb{T}\times\mathfrak{D}$ if $f(t_i,x,y,z)$ is defined on $t_i\in\mathbb{T}^{\circ}$ and $(x,y,z)\in\mathfrak{D}$, and if $f(t_i,x,y,z)$ is continuous on $\mathfrak{D}$ for each $t_i\in\mathbb{T}^{\circ}$.

We make the following assumptions throughout:

(A) $\mathfrak{D}=(0,\infty)\times\mathbb{R}^2$.

(B) f is continuous on $\mathbb{T}^{\circ}\times\mathfrak{D}$.

(C) $f(t_i,x,y,z)$ has a singularity at $x=0$; i.e., $\limsup_{x\to 0^+}|f(t_i,x,y,z)|=\infty$ for each $t_i\in\mathbb{T}^{\circ}$ and for some $(y,z)\in\mathbb{R}^2$.

2. Lower and upper solutions method for regular problems

Let us first consider the regular difference equation

$$u^{\Delta\Delta\Delta}(t_{i-2})+h(t_i,u(t_i),u^{\Delta}(t_{i-1}),u^{\Delta\Delta}(t_{i-2}))=0,\quad t_i\in\mathbb{T}^{\circ},\tag{3}$$

where h is continuous on $\mathbb{T}^{\circ}\times\mathbb{R}^3$, along with the boundary conditions (2). We establish a lower and upper solutions method for the regular problem (3), (2).

Definition 2.1. We call $\alpha:\mathbb{T}\to\mathbb{R}$ a lower solution of (3), (2) if

$$\alpha^{\Delta\Delta\Delta}(t_{i-2})+h(t_i,\alpha(t_i),\alpha^{\Delta}(t_{i-1}),\alpha^{\Delta\Delta\Delta}(t_{i-2}))\geq 0,\quad t_i\in\mathbb{T}^{\circ}\tag{4}$$

and α satisfies boundary conditions

$$\begin{aligned}\alpha^{\Delta\Delta}(t_0)&\leq 0,\\ \alpha^{\Delta}(t_{n+1})&\geq 0,\\ \alpha(t_{n+2})&\leq 0.\end{aligned}\tag{5}$$

Definition 2.2. We call $\beta : \mathbb{T} \to \mathbb{R}$ an upper solution of (3), (2) if

$$\beta^{\Delta\Delta\Delta}(t_{i-2}) + h(t_i, \beta(t_i), \beta^{\Delta}(t_{i-1}), \beta^{\Delta\Delta}(t_{i-2})) \le 0, \quad t_i \in \mathbb{T}^{\circ} \tag{6}$$

and β satisfies boundary conditions

$$\begin{aligned} \beta^{\Delta\Delta}(t_0) &\ge 0, \\ \beta^{\Delta}(t_{n+1}) &\le 0, \\ \beta(t_{n+2}) &\ge 0. \end{aligned} \tag{7}$$

Theorem 2.1 (lower and upper solutions method). *Let α and β be lower and upper solutions of* (3), (2), *respectively, with $\alpha \le \beta$ on $\mathbb{T}^{\circ}$. Let $h(t_i, x, y, z)$ be continuous on $\mathbb{T}^{\circ} \times \mathbb{R}^3$ and nonincreasing in its z variable. Then* (3), (2) *has a solution u satisfying*

$$\alpha(t) \le u(t) \le \beta(t), \quad t \in \mathbb{T}.$$

Proof. We proceed through a sequence of steps involving modifications of the function h.

Step 1. For $t_i \in \mathbb{T}^{\circ}$, $(x, y, z) \in \mathbb{R}^3$, define

$$\tilde{h}\left(t_i, x, y, \frac{y-z}{t_{i-1}-t_{i-2}}\right)$$
$$= \begin{cases} h\left(t_i, \beta(t_i), \beta^{\Delta}(t_{i-1}), \dfrac{\beta^{\Delta}(t_{i-1})-\sigma(t_{i-1}, z)}{t_{i-1}-t_{i-2}}\right) + \dfrac{\beta^{\Delta}(t_{i-1})-y}{\beta^{\Delta}(t_{i-1})-y+1}, & y < \beta^{\Delta}(t_{i-1}), \\ h\left(t_i, x, y, \dfrac{y-\sigma(t_{i-2}, z)}{t_{i-1}-t_{i-2}}\right), & \beta^{\Delta}(t_{i-1}) \le y \le \alpha^{\Delta}(t_{i-1}), \\ h\left(t_i, \alpha(t_i), \alpha^{\Delta}(t_{i-1}), \dfrac{\alpha^{\Delta}(t_{i-1})-\sigma(t_{i-1}, z)}{t_{i-1}-t_{i-2}}\right) + \dfrac{y-\alpha^{\Delta}(t_{i-1})}{y-\alpha^{\Delta}(t_{i-1})+1}, & y > \alpha^{\Delta}(t_{i-1}), \end{cases}$$

where

$$\sigma(t_{i-2}, z) = \begin{cases} \alpha^{\Delta}(t_{i-2}), & z > \alpha^{\Delta}(t_{i-2}), \\ z, & \beta^{\Delta}(t_{i-2}) \le z \le \alpha^{\Delta}(t_{i-2}), \\ \beta^{\Delta}(t_{i-2}), & z < \beta^{\Delta}(t_{i-2}). \end{cases}$$

By its construction, $\tilde{h}$ is continuous on $\mathbb{T}^{\circ} \times \mathbb{R}^3$ and there exists $M > 0$ so that

$$|\tilde{h}(t_i, x, y, z)| \le M, \quad t_i \in \mathbb{T}^{\circ}, \ (x, y, z) \in \mathbb{R}^3.$$

We now study the auxiliary equation

$$u^{\Delta\Delta\Delta}(t_{i-2}) + \tilde{h}(t_i, u(t_i), u^{\Delta}(t_{i-1}), u^{\Delta\Delta}(t_{i-2})) = 0, \quad t_i \in \mathbb{T}^{\circ}, \tag{8}$$

with boundary conditions (2). Our immediate goal is to prove the existence of a solution of (8), (2).

Step 2. The Brouwer fixed point theorem states that, for

$$K = \{(x_1), \dots, (x_n) : c_i \le x_i \le d_i, i = 1, \dots, n\},$$

if $T : K \to K$ is continuous, then T has a fixed point in K. To this end, define

$$E = \{u : \mathbb{T} \to \mathbb{R} : u^{\Delta\Delta}(t_0) = u^{\Delta}(t_{n+1}) = u(t_{n+2}) = 0\}$$

and also define

$$\|u\| = \max\{|u(t_i)| : t_i \in \mathbb{T}\}.$$

This makes E into a Banach space. We define an operator $\mathcal{T} : E \to E$ by

$$(\mathcal{T}u)(t_m) = -\sum_{k=m}^{n+1}(t_{k+1}-t_k)\sum_{j=k}^{n}(t_{j+1}-t_j)\sum_{i=1}^{j}(t_i-t_{i-1})\tilde{h}\big(t_{i+1}, u(t_{i+1}), u^{\Delta}(t_i), u^{\Delta\Delta}(t_{i-1})\big). \quad (9)$$

$\mathcal{T}$ is a continuous operator.

From the bounds placed on $\tilde{h}$ in Step 1 and from (9), if $r > (t_{n+1} - t_0)^3 M$, then $\mathcal{T}(\overline{B(r)}) \subset \overline{B(r)}$, where $B(r) = \{u \in E : \|u\| < r\}$. Therefore, by the Brouwer fixed point theorem [Zeidler 1986], there exists $u \in \overline{B(r)}$ such that $u = \mathcal{T}u$.

Step 3. We now show that u is a fixed point of $\mathcal{T}$ if and only if u is a solution of (8), (2).

First assume $u = \mathcal{T}u$. Then $u \in E$ and thus satisfies (2).

Further,

$$\begin{aligned}
&u^{\Delta}(t_{m-2}) \\
&= \frac{u(t_{m-1})-u(t_{m-2})}{t_{m-1}-t_{m-2}} \\
&= -\frac{\sum\limits_{k=m-1}^{n+1}(t_{k+1}-t_k)\sum\limits_{j=k}^{n}(t_{j+1}-t_j)\sum\limits_{i=1}^{j}(t_i-t_{i-1})\tilde{h}\big(t_{i+1}, u(t_{i+1}), u^{\Delta}(t_i), u^{\Delta\Delta}(t_{i-1})\big)}{t_m-t_{m-1}} \\
&\quad + \frac{\sum\limits_{k=m-2}^{n+1}(t_{k+1}-t_k)\sum\limits_{j=k}^{n}(t_{j+1}-t_j)\sum\limits_{i=1}^{i}(t_i-t_{i-1})\tilde{h}\big(t_{i+1}, u(t_{i+1}), u^{\Delta}(t_i), u^{\Delta\Delta}(t_{i-1})\big)}{t_{m-1}-t_{m-2}} \\
&= \frac{(t_{m-1}-t_{m-2})\sum\limits_{j=m-2}^{n}(t_{j+1}-t_j)\sum\limits_{i=1}^{j}(t_i-t_{i-1})\tilde{h}\big(t_{i+1}, u(t_{i+1}), u^{\Delta}(t_i), u^{\Delta\Delta}(t_{i-1})\big)}{t_{m-1}-t_{m-2}} \\
&= \sum_{j=m-2}^{n}(t_{j+1}-t_j)\sum_{i=1}^{j}(t_i-t_{i-1})\tilde{h}\big(t_{i+1}, u(t_{i+1}), u^{\Delta}(t_i), u^{\Delta\Delta}(t_{i-1})\big).
\end{aligned}$$

We also have

$$\begin{aligned}
u^{\Delta\Delta}(t_{m-2}) &= \frac{u^{\Delta}(t_{m-1}) - u^{\Delta}(t_{m-2})}{t_{m-1}-t_{m-2}} \\
&= \frac{\sum_{j=m-1}^{n}(t_{j+1}-t_j)\sum_{i=1}^{j}(t_i-t_{i-1})\tilde{h}\big(t_{i+1}, u(t_{i+1}), u^{\Delta}(t_i), u^{\Delta\Delta}(t_{i-1})\big)}{t_{m-1}-t_{m-2}} \\
&\quad - \frac{\sum_{j=m-2}^{n}(t_{j+1}-t_j)\sum_{i=1}^{j}(t_i-t_{i-1})\tilde{h}\big(t_{i+1}, u(t_{i+1}), u^{\Delta}(t_i), u^{\Delta\Delta}(t_{i-1})\big)}{t_{m-1}-t_{m-2}} \\
&= -\frac{(t_{m-1}-t_{m-2})\sum_{i=1}^{m-2}(t_i-t_{i-1})\tilde{h}\big(t_{i+1}, u(t_{i+1}), u^{\Delta}(t_i), u^{\Delta\Delta}(t_{i-1})\big)}{t_{m-1}-t_{m-2}} \\
&= -\sum_{i=1}^{m-2}(t_i-t_{i-1})\tilde{h}\big(t_{i+1}, u(t_{i+1}), u^{\Delta}(t_i), u^{\Delta\Delta}(t_{i-1})\big)
\end{aligned}$$

and

$$\begin{aligned}
u^{\Delta\Delta\Delta}(t_{m-2}) &= \frac{u^{\Delta\Delta}(t_{m-1}) - u^{\Delta\Delta}(t_{m-2})}{t_{m-1}-t_{m-2}} \\
&= \frac{-\sum_{i=1}^{m-1}(t_i-t_{i-1})\tilde{h}\big(t_{i+1}, u(t_{i+1}), u^{\Delta}(t_i), u^{\Delta\Delta}(t_{i-1})\big)}{t_{m-1}-t_{m-2}} \\
&\qquad + \frac{\sum_{i=1}^{i-1}(t_i-t_{i-1})\tilde{h}\big(t_{i+1}, u(t_{i+1}), u^{\Delta}(t_i), u^{\Delta\Delta}(t_{i-1})\big)}{t_{m-1}-t_{m-2}} \\
&= \frac{-(t_{m-1}-t_{m-2})\tilde{h}\big(t_m, u(t_m), u^{\Delta}(t_{m-1}), u^{\Delta\Delta}(t_{m-2})\big)}{t_{m-1}-t_{m-2}} \\
&= -\tilde{h}\big(t_m, u(t_m), u^{\Delta}(t_{m-1}), u^{\Delta\Delta}(t_{m-2})\big).
\end{aligned}$$

This implies that $u^{\Delta\Delta\Delta}(t_{m-2})+\tilde{h}(t_m, u(t_m), u^{\Delta}(t_{m-1}), u^{\Delta\Delta}(t_{m-2}))=0$ and, thus, $u(t)$ solves problem (8), (2).

On the other hand, let $u(t)$ solve (8), (2).

Then, for $i = 1, 2, \ldots, n$,

$$u^{\Delta\Delta}(t_i) - u^{\Delta\Delta}(t_{i-1}) = (t_i - t_{i-1})u^{\Delta\Delta\Delta}(t_{i-1}),$$

which means, for each $i = 1, 2, \ldots, n$,

$$\begin{aligned}
u^{\Delta\Delta}(t_i) - u^{\Delta\Delta}(t_i - 1) &= (t_i - t_{i-1})u^{\Delta\Delta\Delta}(t_{i-1}) \\
&= -(t_i - t_{i-1})\tilde{h}\big(t_{i+1}, u(t_{i+1}), u^{\Delta}(t_i), u^{\Delta}(t_{i-1})\big).
\end{aligned}$$

By $u^{\Delta\Delta}(t_0) = 0$ and summing the above equations from $i = 1$ to $i = j$, where $j = 1, 2, \dots, n$, we have

$$u^{\Delta\Delta}(t_j) = -\sum_{i=1}^{j}(t_i - t_{i-1})\tilde{h}\big(t_{i+1}, u(t_{i+1}), u^{\Delta}(t_i), u^{\Delta\Delta}(t_{i-1})\big). \tag{10}$$

Also, for $j = 0, 1, \dots, n$,

$$u^{\Delta}(t_{j+1}) - u^{\Delta}(t_j) = (t_{j+1} - t_j)u^{\Delta\Delta}(t_j).$$

Taking the sum of the above equations from $j = k$ to $j = n$, where $k = 0, 1, \dots, n$, and by $u^{\Delta}(t_{n+1}) = 0$ and (10), we have

$$u^{\Delta}(t_k) = \sum_{j=k}^{n}(t_{j+1} - t_j)\sum_{i=1}^{j}(t_i - t_{i-1})\tilde{h}\big(t_{i+1}, u(t_{i+1}), u^{\Delta}(t_i), u^{\Delta\Delta}(t_{i-1})\big). \tag{11}$$

Similarly, for $k = 0, 1, \dots, n+1$,

$$u(t_{k+1}) - u(t_k) = (t_{k+1} - t_k)u^{\Delta}(t_k).$$

Add the above equations from $k = m$ to $k = n+1$, where $m = 0, 1, \dots, n+2$, and by (11) and $u(t_{n+2}) = 0$, we have

$$-\sum_{k=m}^{n+1}(t_{k+1} - (t_k)\sum_{j=k}^{n}(t_{j+1} - t_j)\sum_{i=1}^{j}(t_i - t_{i-1})\tilde{h}\big(t_{i+1}, u(t_{i+1}), u^{\Delta}(t_i), u^{\Delta\Delta}(t_{i-1})\big).$$

Thus, $u = Tu$ and the claim holds.

Step 4. We now show that solutions $u(t)$ of (8), (2) satisfy

$$\alpha(t) \le u(t) \le \beta(t), \quad t \in \mathbb{T}.$$

Consider the case of obtaining $u(t) \le \beta(t)$. Let $v^{\Delta}(t) = \beta^{\Delta}(t) - u^{\Delta}(t)$. For the sake of establishing a contradiction, assume that

$$\max\{v^{\Delta}(t) : t \in \mathbb{T}\} := v^{\Delta}(l) > 0.$$

From the boundary conditions (2) and (7), we see that $l \equiv l_i \in \mathbb{T}^{\circ}$. Thus, $v^{\Delta}(l_{i+1}) \le v^{\Delta}(l_i)$ and $v^{\Delta}(l_{i-1}) \le v^{\Delta}(l_i)$. Therefore, $v^{\Delta\Delta}(l_i) \le 0$ and $v^{\Delta\Delta}(l_{i-1}) \ge 0$. This in turn implies that $v^{\Delta\Delta\Delta}(l_{i-1}) \le 0$. Consequently,

$$u^{\Delta\Delta\Delta}(l_{i-1}) \ge \beta^{\Delta\Delta\Delta}(l_{i-1}). \tag{12}$$

On the other hand, since h is nonincreasing in its fourth variable, we have from (3) that

$$\begin{aligned}
&\beta^{\Delta\Delta\Delta}(l_{i-1}) - u^{\Delta\Delta\Delta}(l_{i-1})\\
&\quad = \tilde{h}\big(l_{i+1}, u(l_{i+1}), u^{\Delta}(l), u^{\Delta\Delta}(l_{i-1})\big) + \beta^{\Delta\Delta\Delta}(l_{i-1})\\
&\quad = h(l_{i+1}, \beta(l_{i+1}), \beta^{\Delta}(l_i), \frac{\beta^{\Delta}(l_i) - \sigma(l_{i-1}), u(l_{i-1})}{l_i - l_{i-1}}) + \frac{v^{\Delta}(l)}{v^{\Delta}(l)+1} + \beta^{\Delta\Delta\Delta}(l_{i-1})\\
&\quad \geq h(l_{i+1}, \beta(l_{i+1}), \beta^{\Delta}(l), \beta^{\Delta\Delta}(l_{i-1})) + \frac{v^{\Delta}(l)}{v^{\Delta}(l)+1} + \beta^{\Delta\Delta\Delta}(l_{i-1})\\
&\quad \geq -\beta^{\Delta\Delta\Delta}(l_{i-1}) + \frac{v^{\Delta}(l)}{v^{\Delta}(l)+1} + \beta^{\Delta\Delta\Delta}(l_{i-1})\\
&\quad = \frac{v^{\Delta}(l)}{v^{\Delta}(l)+1} > 0.
\end{aligned}$$

Hence, $u^{\Delta\Delta\Delta}(l_{i-1}) < \beta^{\Delta\Delta\Delta}(l_{i-1})$, but this contradicts (12). Therefore, $v^{\Delta}(l) \leq 0$. This implies that $u^{\Delta}(l) \geq \beta^{\Delta}(l)$, and hence

$$\sum_{l=t}^{t_{n+2}} (t_i - t_{i-1})\beta^{\Delta}(l) \leq \sum_{l=t}^{t_{n+2}} (t_i - t_{i-1})u^{\Delta}(l).$$

This, in turn, yields

$$\begin{aligned}
&\beta(t_{n+2}) - \beta(t) \leq u(t_{n+2}) - u(t), && u(t) \leq \beta(t) - \beta(t_{n+2}),\\
&\beta(t_{n+2}) - \beta(t) \leq -u(t), && u(t) \leq \beta(t).
\end{aligned}$$

A similar argument shows that $\alpha(t) \leq u(t)$, $t \in \mathbb{T}$.

Thus, the conclusion of the theorem holds and our proof is complete. □

3. Existence result

In this section, we make use of Theorem 2.1 to obtain positive solutions of the singular problem (1), (2). In particular, in applying Theorem 2.1, we deal with a sequence of regular perturbations of (1), (2). Ultimately, we obtain a desired solution of (1), (2) by passing to the limit on a sequence of solutions for the perturbations.

Theorem 3.1. *Assume conditions* (A), (B), *and* (C) *hold, along with the following*:

(D) *there exists* $c \in (0, \infty)$ *so that* $f(t_i, c, 0, 0) \leq 0$ *for all* $t \in \mathbb{T}^{\circ}$;

(E) $f(t_i, x, y, z)$ *is nonincreasing in its* z *variable for* $t_i \in \mathbb{T}^{\circ}$ *and* $x \in (0, c]$;

(F) $\lim_{x\to 0^+} f(t_i, x, y, z) = \infty$ *for* $t_i \in \mathbb{T}^{\circ}$, $y \in \left(-\frac{c}{r}, \frac{c}{r}\right)$, *where r is sufficiently large.*

Then (1), (2) *has a solution u satisfying*

$$0 < u(t) \leq c, \quad t_i \in \mathbb{T}^\circ.$$

Proof. Again, for the proof, we proceed through a sequence of steps.

Step 1. For $l \in \mathbb{N}$, $t_i \in \mathbb{T}^\circ$, $(x, y, z) \in \mathbb{R}^3$, define

$$f_l(t_i, x, y, z) = \begin{cases} f(t_i, |x|, y, z), & |x| \geq \frac{1}{l}, \\ f(t_i, \frac{1}{l}, y, z), & |x| < \frac{1}{l}. \end{cases}$$

Then f_l is continuous on $\mathbb{T}^\circ \times \mathbb{R}^3$ and nonincreasing for $t_i \in \mathbb{T}^\circ$, $x \in [-c, c]$.

Assumption (F) implies that there exists l_0 such that, for all $l \geq l_0$,

$$f_l(t_i, c, 0, 0) = f(t_i, c, 0, 0) > 0, \quad t_i \in \mathbb{T}^\circ.$$

Consider, for each $l \geq l_0$,

$$u^{\Delta\Delta\Delta}(t_{i-2}) + f_l(t_i, u(t_i), u^{\Delta}(t_{i-1}), u^{\Delta\Delta}(t_{i-2})) = 0, \quad t_i \in \mathbb{T}^\circ. \tag{13}$$

Define $\alpha(t) = 0$ and $\beta(t) = c$. Then α and β are lower and upper solutions for (13), (2) and $\alpha(t) \leq \beta(t)$ on $\mathbb{T}^\circ$. Thus, by Theorem 2.1, there exists u_l a solution of (13), (2) satisfying $0 \leq u_l(t) \leq c$, $t_i \in \mathbb{T}$, $l \geq l_0$. Consequently,

$$|u_l^{\Delta}(t_i)| \leq \frac{c}{(t_i - t_{i-1})}, \quad t_i \in \mathbb{T}^\circ. \tag{14}$$

Step 2. Let $l \in \mathbb{N}$, $l \geq l_0$. Since $u_l(t)$ solves (13), we get, from work similar to that exhibited in Theorem 2.1,

$$u_l^{\Delta}(t_m) = \sum_{j=1}^{n}(t_{j+1} - t_j)\sum_{i=1}^{j}(t_i - t_{i-1}) f_l\big(t_i, u_l(t_i), u_l^{\Delta}(t_{i-1}), u_l^{\Delta\Delta}(t_{i-2})\big) \tag{15}$$

for $t_m \in \mathbb{T}^\circ$. By assumption (F), there exists $\varepsilon_1 \in (0, 1/l_0)$ such that, if $l \geq 1/\varepsilon_1$,

$$f_l(t_2, x, y, z) > \frac{c}{t_2 - t_1}, \quad x \in (0, \varepsilon_1],\ y \in (-c, c). \tag{16}$$

For the sake of establishing a contradiction, assume that $u_l(t_1) < \varepsilon_1$ for $l \geq 1/\varepsilon_1$. Then, by (15) and (16),

$$\begin{aligned}
u_l^{\Delta}(t_1) &= -\sum_{j=1}^{n}(t_{j+1}-t_j)\sum_{i=1}^{j}(t_i-t_{i-1})f_l\big(t_i, u_k(t_i), u_k^{\Delta}(t_{i-1}), u_l^{\Delta\Delta}(t_{i-2})\big)\\
&\geq f_l\big(t_i, u_l(t_i), u_l^{\Delta}(t_{i-1}), u_l^{\Delta\Delta}(t_{i-2})\big)\\
&\qquad\qquad + \sum_{j=2}^{n}(t_{j+1}-t_j)\sum_{i=1}^{j}(t_i-t_{i-1})f_l\big(t_i, u_l(t_i), u_l^{\Delta}(t_{i-1}), u_l^{\Delta\Delta}(t_{i-2})\big)\\
&\geq f_l\big(t_i, u_l(t_i), u_l^{\Delta}(t_{i-1}), u_l^{\Delta\Delta}(t_{i-2})\big)\\
&\geq \frac{c}{t_2-t_1} = -\frac{c}{r}.
\end{aligned}$$

But this contradicts (14). Hence $u_l(t_1) \geq \varepsilon_1$ for all $l \geq 1/\varepsilon_1$.

Define $a_2 = \max\{|f_l(t_2, x, y, z)| : x \in [\varepsilon_1, c],\ y \in (-c, c)\}$. By assumption (F), there exists $\varepsilon_2 \in (0, \varepsilon_1]$ such that, if $l \geq 1/\varepsilon_2$ and $u_l < \varepsilon_2$, then

$$f_l(t_3, x, y, z) > \frac{c}{t_3-t_2} - T(a_2), \quad x \in (0, \varepsilon_2],\ y \in (-c, c). \tag{17}$$

For the sake of establishing a contradiction, assume that, for $l \geq 1/\varepsilon_2$, we have $u_l(t_2) < \varepsilon_2$. Then, by (15) and (17), we have

$$\begin{aligned}
u_l^{\Delta}(t_2) &= \sum_{j=1}^{n}(t_{j+1}-t_j)\sum_{i=1}^{j}(t_i-t_{i-1})f_l\big(t_i, u_l(t_i), u_l^{\Delta}(t_{i-1}), u_l^{\Delta\Delta}(t_{i-2})\big)\\
&= \sum_{j=2}^{n}(t_{j+1}-t_j)\sum_{i=2}^{j}(t_i-t_{i-1})f_l\big(t_i, u_l(t_i), u_l^{\Delta}(t_{i-1})u_l^{\Delta\Delta}(t_{i-2})\big)\\
&\qquad + Tf_l\big(t_2, u_l(t_2), u_l^{\Delta}(t_1), u_l^{\Delta\Delta}(t_0)\big)\\
&= \sum_{j=3}^{n}(t_{j+1}-t_j)\sum_{i=2}^{j}(t_i-t_{i-1})f_l\big(t_i, u_k(t_i), u_l^{\Delta}(t_{i-1}), u_l^{\Delta\Delta}(t_{i-2})\big)\\
&\qquad + f_l\big(t_3, u_l(t_3), u_l^{\Delta}(t_2), u_l^{\Delta\Delta}(t_1)\big) + Tf_l\big(t_2, u_l(t_2), u_l^{\Delta}(t_1), u_l^{\Delta\Delta}(t_0)\big)\\
&\geq \sum_{j=3}^{n}(t_{j+1}-t_j)\sum_{i=2}^{j}(t_i-t_{i-1})f_l\big(t_i, u_k(t_i), u_l^{\Delta}(t_{i-1}), u_l^{\Delta\Delta}(t_{i-2})\big)\\
&\qquad + f_k\big(t_2, u_k(t_2), u_k^{\Delta}(t_1)\big)f_l\big(t_3, u_l(t_3), u_l^{\Delta}(t_2), u_l^{\Delta\Delta}(t_1)\big) + Ta_2\\
&> \sum_{j=3}^{n}(t_{j+1}-t_j)\sum_{i=2}^{j}(t_i-t_{i-1})f_l\big(t_i, u_k(t_i), u_l^{\Delta}(t_{i-1}), u_l^{\Delta\Delta}(t_{i-2})\big) + \frac{c}{t_3-t_2}\\
&> \frac{c}{t_3-t_2}.
\end{aligned}$$

But this contradicts (14). Hence $u_l(t_2) \geq \varepsilon_2$ for all $l \geq 1/\varepsilon_2$.

Continuing similarly for $t = 3, 4, \dots, nT$, we get $0 < \varepsilon_T < \cdots < \varepsilon_2 < \varepsilon_1$ such that $u_l(t_i) \geq \varepsilon_T$ for $t_i \in T$.

For $2 \leq i \leq n-1$, set

$$m_i = \max\bigl\{|f_l(t_i, x, y, z)| : x \in [\varepsilon_i, c],\, y \in (-c, c)\bigr\}.$$

By assumption (F), there exists $\varepsilon_n \in (0, \varepsilon_{n-1}]$ such that, if $l \geq 1/\varepsilon_n$ and $u_l(t_n) < \varepsilon_n$, then

$$f_l(t_n, x, y, z) > \frac{c}{t_n - t_{n-1}} - \sum_{i=2}^{n-1} m_i. \tag{18}$$

For the sake of establishing a contradiction, assume that, for $l \geq 1/\varepsilon_n$, we have $u_l(t_n) < \varepsilon_n$. Then, by (15) and (18), we have

$$\begin{aligned}
u_l^{\Delta}(t_n) &= \sum_{j=n+1}^{n+1} (t_{j+1} - t_j) \sum_{i=2}^{j} (t_i - t_{i-1}) f_l\bigl(t_i, u(t_i), u^{\Delta}(t_i), u^{\Delta\Delta}(t_{i-2})\bigr) \\
&= (t_{n+2} - t_{n+1}) \sum_{i=2}^{n+1} (t_i - t_{i-1}) f_l\bigl(t_i, u(t_i), u^{\Delta}(t_{i-1}), u^{\Delta\Delta}(t_{i-2})\bigr) \\
&= (t_{n+2} - t_{n+1}) \sum_{i=2}^{n} (t_i - t_{i-1}) f_l\bigl(t_i, u(t_i), u^{\Delta}(t_{i-1}), u^{\Delta\Delta}(t_{i-2})\bigr) \\
&\quad + f_l\bigl(t_{n+1}, u(t_{n+1}), u^{\Delta}(t_n), u^{\Delta\Delta}(t_{n-1})\bigr) \\
&> \sum_{i=2}^{n-1} (m_i) + \frac{c}{t_n - t_{n-1}} - \sum_{i=2}^{n-1} (m_i) \\
&= \frac{c}{t_n - t_{n-1}}.
\end{aligned}$$

But this contradicts (14). Hence $u_l(t_n) \geq \varepsilon_n$ for all $l \geq 1/\varepsilon_n$. Therefore, by letting $\varepsilon = \varepsilon_n$, we get

$$0 < \varepsilon \leq u_l(t_i) \leq c, \quad t \in \mathbb{T}^{\circ},\ l \geq \frac{1}{\varepsilon}. \tag{19}$$

Since $u_l(t_i)$ satisfies (19) and (2), we can choose a subsequence $\{u_{l_k}(t)\} \subset \{u_l(t_i)\}$ such that $\lim_{k\to\infty} u_{l_k}(t) = u(t_i)$, $t \in \mathbb{T}^{\circ}$, $u(t_i) \in E$, where E is as defined in Step 2 of Theorem 2.1. Moreover, (15) yields, for each sufficiently large k,

$$u_{l_k}^{\Delta}(t_i) = \sum_{j=t_i+1}^{n} (t_{j+1} - t_j) \sum_{i=2}^{j} (t_i - t_{i-1}) f\bigl(t_i, u_{l_k}(t_i), u_{l_k}^{\Delta}(t_{i-1}), u_{l_k}^{\Delta\Delta}(t_{i-2})\bigr),$$

and so, letting $l \to \infty$ and from the continuity of f, we get

$$u^{\Delta}(t_i) = \sum_{t_i+1}^{n}(t_{j+1} - t_j) \sum_{i=2}^{j}(t_i - t_{i-1}) f\big(t_i, u(t_i), u^{\Delta}(t_{i-1}), u^{\Delta\Delta}(t_{i-2})\big).$$

Consequently,

$$u^{\Delta\Delta}(t_{i-1}) = \sum_{i=2}^{j}(t_i - t_{i-1}) f\big(t_i, u(t_i), u^{\Delta}(t_{i-1}), u^{\Delta\Delta}(t_{i-2})\big).$$

Thus,

$$u^{\Delta\Delta\Delta}(t_{i-2}) = -f\big(t_i, u(t_i), u^{\Delta}(t_{i-1}), u^{\Delta\Delta}(t_{i-2})\big).$$

Therefore, u solves (1), and, by (19), our theorem holds. □

References

[Agarwal and Wong 1997] R. P. Agarwal and P. J. Y. Wong, *Advanced topics in difference equations*, Mathematics and its Applications **404**, Kluwer Academic Publishers Group, Dordrecht, 1997. MR 98i:39001 Zbl 0878.39001

[Agarwal et al. 1999] R. P. Agarwal, D. O'Regan, and P. J. Y. Wong, *Positive solutions of differential, difference and integral equations*, Dordrecht: Kluwer Academic Publishers, 1999. MR 2000a:34046 Zbl 1157.34301

[Agarwal et al. 2003] R. P. Agarwal, A. Cabada, and V. Otero-Espinar, "Existence and uniqueness results for n-th order nonlinear difference equations in presence of lower and upper solutions", *Arch. Inequal. Appl.* **1**:3-4 (2003), 421–431. MR 2004i:39011 Zbl 1049.39001

[Agarwal et al. 2004] R. P. Agarwal, A. Cabada, V. Otero-Espinar, and S. Dontha, "Existence and uniqueness of solutions for anti-periodic difference equations", *Arch. Inequal. Appl.* **2**:4 (2004), 397–411. MR 2005h:39005 Zbl 1087.39001

[Agarwal et al. 2005] R. P. Agarwal, D. O'Regan, and P. J. Y. Wong, "Existence of constant-sign solutions to a system of difference equations: The semipositone and singular case", *J. Difference Equ. Appl.* **11**:2 (2005), 151–171. MR 2005i:65217 Zbl 1066.39001

[Agarwal et al. 2008] R. P. Agarwal, D. O'Regan, and S. Stanêk, "An existence principle for nonlocal difference boundary value problems with φ-Laplacian and its application to singular problems", *Adv. Difference Equ.* **2008** (2008), 14 p. MR 2009h:39004 Zbl 1146.39026

[Akın-Bohner et al. 2003] E. Akın-Bohner, F. M. Atıcı, and B. Kaymakçalan, "Lower and upper solutions of boundary value problems", pp. 165–188 in *Advances in dynamic equations on time scales*, edited by M. Bohner and A. C. Peterson, Birkhäuser, Boston, MA, 2003. MR 1962548

[Atici et al. 2003] F. M. Atici, A. Cabada, and V. Otero-Espinar, "Criteria for existence and nonexistence of positive solutions to a discrete periodic boundary value problem", *J. Difference Equ. Appl.* **9**:9 (2003), 765–775. MR 2004f:39010 Zbl 1056.39016

[Bohner and Peterson 2001] M. Bohner and A. Peterson, *Dynamic equations on time scales: An introduction with applications*, Birkhäuser, Boston, MA, 2001. MR 2002c:34002 Zbl 0978.39001

[Cabada 2011] A. Cabada, "An overview of the lower and upper solutions method with nonlinear boundary value conditions", *Bound. Value Probl.* (2011), Art. ID 893753, 18. MR 2719294 Zbl 1230.34001

[Henderson and Kunkel 2006] J. Henderson and C. J. Kunkel, "Singular discrete higher order boundary value problems", *Int. J. Difference Equ.* **1**:1 (2006), 119–133. MR 2008b:39014 Zbl 1128.39011

[Henderson and Thompson 2002] J. Henderson and H. B. Thompson, "Existence of multiple solutions for second-order discrete boundary value problems", *Comput. Math. Appl.* **43**:10-11 (2002), 1239–1248. MR 2003f:39004 Zbl 1005.39014

[Jiang et al. 2005] D. Q. Jiang, D. O'Regan, and R. P. Agarwal, "A generalized upper and lower solution method for singular discrete boundary value problems for the one-dimensional p-Laplacian", *J. Appl. Anal.* **11**:1 (2005), 35–47. MR 2006c:39005 Zbl 1086.39022

[Jódar 1987] L. Jódar, "Singular bilateral boundary value problems for discrete generalized Lyapunov matrix equations", *Stochastica* **11**:1 (1987), 45–52. MR 89m:15010 Zbl 0659.15010

[Jódar et al. 1992] L. Jódar, E. Navarro, and J. L. Morera, "A closed-form solution of singular regular higher-order difference initial and boundary value problems", *Appl. Math. Comput.* **48**:2-3 (1992), 153–166. MR 93a:39011 Zbl 0768.39002

[Kelley and Peterson 2001] W. G. Kelley and A. C. Peterson, *Difference equations: An introduction with applications*, 2nd ed., Harcourt/Academic Press, San Diego, CA, 2001. MR 2001i:39001 Zbl 0970.39001

[Kunkel 2006] C. J. Kunkel, "Singular discrete third order boundary value problems", *Comm. Appl. Nonlinear Anal.* **13**:3 (2006), 27–38. MR 2007b:39004 Zbl 1109.39008

[Kunkel 2008] C. J. Kunkel, "Singular second order boundary value problems on purely discrete time scales", *J. Difference Equ. Appl.* **14**:4 (2008), 411–420. MR 2009f:39002 Zbl 1138.39019

[Naidu and Kailasa Rao 1982] D. S. Naidu and A. Kailasa Rao, "Singular perturbation methods for a class of initial- and boundary-value problems in discrete systems", *Internat. J. Control* **36**:1 (1982), 77–94. MR 84e:39003 Zbl 0484.93051

[O'Regan and El-Gebeily 2008] D. O'Regan and M. El-Gebeily, "Existence, upper and lower solutions and quasilinearization for singular differential equations", *IMA J. Appl. Math.* **73**:2 (2008), 323–344. MR 2009d:34035 Zbl 1202.34053

[Pao 1985] C. V. Pao, "Monotone iterative methods for finite difference system of reaction-diffusion equations", *Numer. Math.* **46**:4 (1985), 571–586. MR 86h:65156 Zbl 0589.65072

[Peterson et al. 2004] A. C. Peterson, Y. N. Raffoul, and C. C. Tisdell, "Three point boundary value problems on time scales", *J. Difference Equ. Appl.* **10**:9 (2004), 843–849. MR 2005g:34036 Zbl 1078.39016

[Rachůnková and Rachůnek 2006] I. Rachůnková and L. Rachůnek, "Singular discrete second order BVPs with p-Laplacian", *J. Difference Equ. Appl.* **12**:8 (2006), 811–819. MR 2007c:39027 Zbl 1106.39021

[Rachůnková and Rachůnek 2009] I. Rachůnková and L. Rachůnek, "Singular discrete and continuous mixed boundary value problems", *Math. Comput. Modelling* **49**:3-4 (2009), 413–422. MR 2009k:34039 Zbl 1173.34010

[Yuan et al. 2008] C. Yuan, D. Jiang, and Y. Zhang, "Existence and uniqueness of solutions for singular higher order continuous and discrete boundary value problems", *Bound. Value Probl.* (2008), Art. ID 123823, 11. MR 2008m:34048 Zbl 1154.34315

[Zeidler 1986] E. Zeidler, *Nonlinear functional analysis and its applications, I: Fixed-point theorems*, Springer, New York, 1986. MR 87f:47083 Zbl 0583.47050

[Zhang et al. 2002] B. Zhang, L. Kong, Y. Sun, and X. Deng, "Existence of positive solutions for BVPs of fourth-order difference equations", *Appl. Math. Comput.* **131**:2-3 (2002), 583–591. MR 2004c:39034 Zbl 1025.39006

[Zheng et al. 2011] B. Zheng, H. Xiao, and H. Shi, "Existence of positive, negative, and sign-changing solutions to discrete boundary value problems", *Boundary Value Problems* **2011** (2011), Art. ID 172818, 19. MR 2011m:39005 Zbl 1216.39011

Received: 2012-06-13 Revised: 2012-07-23 Accepted: 2012-07-23

coundeho@ut.utm.edu *Department of Mathematics and Statistics, University of Tennessee at Martin, Martin, TN 38238, United States*

ckunkel@utm.edu *Department of Mathematics and Statistics, University of Tennessee at Martin, Martin, TN 38238, United States*

ashnpoor@ut.utm.edu *Department of Mathematics and Statistics, University of Tennessee at Martin, Martin, TN 38238, United States*

mathematical sciences publishers msp

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

2013 vol. 6 no. 1

1944-4176(2013)6:1;1-A