

# involve

a journal of mathematics

## Editorial Board

Kenneth S. Berenhaut, *Managing Editor*

Colin Adams	Suzanne Lenhart
John V. Baxley	Chi-Kwong Li
Arthur T. Benjamin	Robert B. Lund
Martin Bohner	Gaven J. Martin
Nigel Boston	Mary Meyer
Amarjit S. Budhiraja	Emil Minchev
Pietro Cerone	Frank Morgan
Scott Chapman	Mohammad Sal Moslehian
Jem N. Corcoran	Zuhair Nashed
Toka Diagana	Ken Ono
Michael Dorff	Timothy E. O'Brien
Sever S. Dragomir	Joseph O'Rourke
Behrouz Emamizadeh	Yuval Peres
Joel Foisy	Y.-F. S. Pétermann
Errin W. Fulp	Robert J. Plemmons
Joseph Gallian	Carl B. Pomerance
Stephan R. Garcia	Bjorn Poonen
Anant Godbole	James Propp
Ron Gould	Józeph H. Przytycki
Andrew Granville	Richard Rebarber
Jerrold Griggs	Robert W. Robinson
Sat Gupta	Filip Saidak
Jim Haglund	James A. Sellers
Johnny Henderson	Andrew J. Serge
Jim Hoste	Ann Trenk
Natalia Hritonenko	Ravi Vakil
Glenn H. Hurlbert	Antonia Vecchio
Charles R. Johnson	Ram U. Verma
K. B. Kulasekera	John C. Wierman
Gerry Ladas	Michael E. Zieve
David Larson	



## EDITORS

### MANAGING EDITOR

Kenneth S. Berenhaut, Wake Forest University, USA, [berenhks@wfu.edu](mailto:berenhks@wfu.edu)

### BOARD OF EDITORS

Colin Adams	Williams College, USA <a href="mailto:colin.c.adams@williams.edu">colin.c.adams@williams.edu</a>	David Larson	Texas A&M University, USA <a href="mailto:larson@math.tamu.edu">larson@math.tamu.edu</a>
John V. Baxley	Wake Forest University, NC, USA <a href="mailto:baxley@wfu.edu">baxley@wfu.edu</a>	Suzanne Lenhart	University of Tennessee, USA <a href="mailto:lenhart@math.utk.edu">lenhart@math.utk.edu</a>
Arthur T. Benjamin	Harvey Mudd College, USA <a href="mailto:benjamin@hmc.edu">benjamin@hmc.edu</a>	Chi-Kwong Li	College of William and Mary, USA <a href="mailto:ckli@math.wm.edu">ckli@math.wm.edu</a>
Martin Bohner	Missouri U of Science and Technology, USA <a href="mailto:bohner@mst.edu">bohner@mst.edu</a>	Robert B. Lund	Clemson University, USA <a href="mailto:lund@clemson.edu">lund@clemson.edu</a>
Nigel Boston	University of Wisconsin, USA <a href="mailto:boston@math.wisc.edu">boston@math.wisc.edu</a>	Gaven J. Martin	Massey University, New Zealand <a href="mailto:g.j.martin@massey.ac.nz">g.j.martin@massey.ac.nz</a>
Amarjit S. Budhiraja	U of North Carolina, Chapel Hill, USA <a href="mailto:budhiraj@email.unc.edu">budhiraj@email.unc.edu</a>	Mary Meyer	Colorado State University, USA <a href="mailto:meyer@stat.colostate.edu">meyer@stat.colostate.edu</a>
Pietro Cerone	La Trobe University, Australia <a href="mailto:P.Cerone@latrobe.edu.au">P.Cerone@latrobe.edu.au</a>	Emil Minchev	Ruse, Bulgaria <a href="mailto:eminchev@hotmail.com">eminchev@hotmail.com</a>
Scott Chapman	Sam Houston State University, USA <a href="mailto:scott.chapman@shsu.edu">scott.chapman@shsu.edu</a>	Frank Morgan	Williams College, USA <a href="mailto:frank.morgan@williams.edu">frank.morgan@williams.edu</a>
Joshua N. Cooper	University of South Carolina, USA <a href="mailto:cooper@math.sc.edu">cooper@math.sc.edu</a>	Mohammad Sal Moslehian	Ferdowsi University of Mashhad, Iran <a href="mailto:moslehian@ferdowsi.um.ac.ir">moslehian@ferdowsi.um.ac.ir</a>
Jem N. Corcoran	University of Colorado, USA <a href="mailto:corcoran@colorado.edu">corcoran@colorado.edu</a>	Zuhair Nashed	University of Central Florida, USA <a href="mailto:znashed@mail.ucf.edu">znashed@mail.ucf.edu</a>
Toka Diagana	Howard University, USA <a href="mailto:tdiagana@howard.edu">tdiagana@howard.edu</a>	Ken Ono	Emory University, USA <a href="mailto:ono@mathcs.emory.edu">ono@mathcs.emory.edu</a>
Michael Dorff	Brigham Young University, USA <a href="mailto:mdorff@math.byu.edu">mdorff@math.byu.edu</a>	Timothy E. O'Brien	Loyola University Chicago, USA <a href="mailto:tbriell@luc.edu">tbriell@luc.edu</a>
Sever S. Dragomir	Victoria University, Australia <a href="mailto:sever@matilda.vu.edu.au">sever@matilda.vu.edu.au</a>	Joseph O'Rourke	Smith College, USA <a href="mailto:orourke@cs.smith.edu">orourke@cs.smith.edu</a>
Behrouz Emamizadeh	The Petroleum Institute, UAE <a href="mailto:bemamizadeh@pi.ac.ae">bemamizadeh@pi.ac.ae</a>	Yuval Peres	Microsoft Research, USA <a href="mailto:peres@microsoft.com">peres@microsoft.com</a>
Joel Foisy	SUNY Potsdam <a href="mailto:foisyjs@potsdam.edu">foisyjs@potsdam.edu</a>	Y.-F. S. Pétermann	Université de Genève, Switzerland <a href="mailto:petermann@math.unige.ch">petermann@math.unige.ch</a>
Errin W. Fulp	Wake Forest University, USA <a href="mailto:fulp@wfu.edu">fulp@wfu.edu</a>	Robert J. Plemmons	Wake Forest University, USA <a href="mailto:rplemmons@wfu.edu">rplemmons@wfu.edu</a>
Joseph Gallian	University of Minnesota Duluth, USA <a href="mailto:kgallian@d.umn.edu">kgallian@d.umn.edu</a>	Carl B. Pomerance	Dartmouth College, USA <a href="mailto:carl.pomerance@dartmouth.edu">carl.pomerance@dartmouth.edu</a>
Stephan R. Garcia	Pomona College, USA <a href="mailto:stephan.garcia@pomona.edu">stephan.garcia@pomona.edu</a>	Vadim Ponomarenko	San Diego State University, USA <a href="mailto:vadim@sciences.sdsu.edu">vadim@sciences.sdsu.edu</a>
Anant Godbole	East Tennessee State University, USA <a href="mailto:godbole@etsu.edu">godbole@etsu.edu</a>	Bjorn Poonen	UC Berkeley, USA <a href="mailto:poonen@math.berkeley.edu">poonen@math.berkeley.edu</a>
Ron Gould	Emory University, USA <a href="mailto:rg@mathcs.emory.edu">rg@mathcs.emory.edu</a>	James Propp	U Mass Lowell, USA <a href="mailto:jpropp@cs.uml.edu">jpropp@cs.uml.edu</a>
Andrew Granville	Université Montréal, Canada <a href="mailto:andrew.andrew@dms.umontreal.ca">andrew.andrew@dms.umontreal.ca</a>	József H. Przytycki	George Washington University, USA <a href="mailto:przytyck@gwu.edu">przytyck@gwu.edu</a>
Jerrold Griggs	University of South Carolina, USA <a href="mailto:griggs@math.sc.edu">griggs@math.sc.edu</a>	Richard Rebarber	University of Nebraska, USA <a href="mailto:rrebarbe@math.unl.edu">rrebarbe@math.unl.edu</a>
Sat Gupta	U of North Carolina, Greensboro, USA <a href="mailto:sgupta@uncg.edu">sgupta@uncg.edu</a>	Robert W. Robinson	University of Georgia, USA <a href="mailto:rwr@cs.uga.edu">rwr@cs.uga.edu</a>
Jim Haglund	University of Pennsylvania, USA <a href="mailto:jhaglund@math.upenn.edu">jhaglund@math.upenn.edu</a>	Filip Saidak	U of North Carolina, Greensboro, USA <a href="mailto:f_saidak@uncg.edu">f_saidak@uncg.edu</a>
Johnny Henderson	Baylor University, USA <a href="mailto:johnny_henderson@baylor.edu">johnny_henderson@baylor.edu</a>	James A. Sellers	Penn State University, USA <a href="mailto:sellersj@math.psu.edu">sellersj@math.psu.edu</a>
Jim Hoste	Pitzer College <a href="mailto:jhoste@pitzer.edu">jhoste@pitzer.edu</a>	Andrew J. Sterge	Honorary Editor <a href="mailto:andy@ajsterge.com">andy@ajsterge.com</a>
Natalia Hritonenko	Prairie View A&M University, USA <a href="mailto:nahritonenko@pvamu.edu">nahritonenko@pvamu.edu</a>	Ann Trenk	Wellesley College, USA <a href="mailto:atrenk@wellesley.edu">atrenk@wellesley.edu</a>
Glenn H. Hurlbert	Arizona State University, USA <a href="mailto:hurlbert@asu.edu">hurlbert@asu.edu</a>	Ravi Vakil	Stanford University, USA <a href="mailto:vakil@math.stanford.edu">vakill@math.stanford.edu</a>
Charles R. Johnson	College of William and Mary, USA <a href="mailto:crjohnso@math.wm.edu">crjohnso@math.wm.edu</a>	Antonia Vecchio	Consiglio Nazionale delle Ricerche, Italy <a href="mailto:antonia.vecchio@cnr.it">antonia.vecchio@cnr.it</a>
K. B. Kulasekera	Clemson University, USA <a href="mailto:kk@ces.clemson.edu">kk@ces.clemson.edu</a>	Ram U. Verma	University of Toledo, USA <a href="mailto:verma99@msn.com">verma99@msn.com</a>
Gerry Ladas	University of Rhode Island, USA <a href="mailto:gladas@math.uri.edu">gladas@math.uri.edu</a>	John C. Wierman	Johns Hopkins University, USA <a href="mailto:wierman@jhu.edu">wierman@jhu.edu</a>
		Michael E. Zieve	University of Michigan, USA <a href="mailto:zieve@umich.edu">zieve@umich.edu</a>

## PRODUCTION


Silvio Levy, Scientific Editor

See inside back cover or [msp.org/involve](http://msp.org/involve) for submission instructions. The subscription price for 2015 is US \$140/year for the electronic version, and \$190/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<sup>®</sup> from Mathematical Sciences Publishers.

PUBLISHED BY

 **mathematical sciences publishers**  
nonprofit scientific publishing

<http://msp.org/>

© 2015 Mathematical Sciences Publishers

# Colorability and determinants of $T(m, n, r, s)$ twisted torus knots for $n \equiv \pm 1 \pmod{m}$

Matt DeLong, Matthew Russell and Jonathan Schrock

(Communicated by Kenneth S. Berenhaut)

We develop theorems to compute the  $p$ -colorability of the families of  $T(m, n, r, s)$  twisted torus knots for  $n \equiv \pm 1 \pmod{m}$  by finding their determinants. Instead of the usual method of reducing crossing matrices to find the determinant, we describe a new method that is applicable for braid representations with full cycles and twists.

## 1. Introduction

In an undergraduate research project, Breiland, Oesper and Taalman [Breiland et al. 2009] used determinants to completely characterize the  $p$ -colorability of torus knots. Conceptually, twisted torus knots, a recent addition to the field first described by Dean [1996], are derived from torus knots. Thus, studying the determinants and  $p$ -colorability of twisted torus knots is a natural extension of [Breiland et al. 2009].

In our paper, we develop theorems for calculating the determinant of certain families of twisted torus knots  $T(m, n, r, s)$ , namely, when  $n \equiv \pm 1 \pmod{m}$ . Table 1 presents a summary of our results. The columns for  $m$ ,  $r$ , and  $s$  give the parity of those parameters (if the column for  $s$  is left blank, that means the parity of  $s$  has no effect on the formula for the determinant). The second column relates  $n$  to  $m$ , and the final column gives the determinant.

The organization of the paper is as follows. Section 2 provides background information and previously known results. Section 3 introduces a new method of finding the determinant of twisted torus knots and proves some preliminary results. In Section 4 we prove our main results. Finally, in Section 5, we conclude with suggestions for further research.

## 2. Background

**2A. Torus knots and twisted torus knots.** For  $m, n$  relatively prime, let  $T(m, n)$  represent the torus knot that circles the meridian of a torus  $m$  times and the longitude

---

*MSC2010:* primary 57M27; secondary 11C20, 05C15.

*Keywords:* knot theory, determinants, colorability, twisted torus knots.

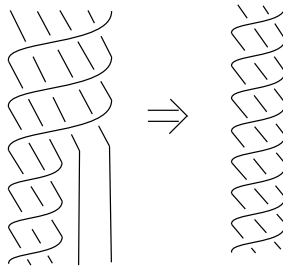
$m$	$n$	$r$	$s$	$\det(T(m, n, r, s))$
even	$mq \pm 1$	even		$ mq \pm 1 + rs \pm (m - r)qrs $
even	$mq \pm 1$	odd	odd	$ r \pm (mr - r^2 + 1)q $
even	$mq \pm 1$	odd	even	$ mq \pm 1 $
odd	$2mq \pm 1$	even		$ rs \pm 1 $
odd	$2mq \pm 1$	odd	odd	$r$
odd	$2mq \pm 1$	odd	even	$1$
odd	$(2q + 1)m \pm 1$	even		$ m \mp (m - r)rs $
odd	$(2q + 1)m \pm 1$	odd	odd	$ mr - r^2 + 1 $
odd	$(2q + 1)m \pm 1$	odd	even	$m$

**Table 1.** Summary of determinants of  $T(m, n, r, s)$  twisted torus knots with  $n \equiv \pm 1 \pmod{m}$ .

of a torus  $n$  times [Adams 2004].  $T(m, n)$  is the closure of the braid with  $m$  strands and  $n$  cycles, where we define a *cycle* on  $m$  strands as the passing of the right-most strand over the remaining  $m - 1$  strands.

A *twisted torus knot* can be constructed by beginning with the braid representation of a  $T(m, n)$  torus knot and then performing  $s$  full twists on  $r$  parallel strands [Champanerkar et al. 2004]. We denote a twisted torus knot by  $T(m, n, r, s)$ , where  $m$  is the total number of strands in the braid representation,  $n$  is the number of cycles on the  $m$  strands,  $r$  is the number of strands to be twisted, and  $s$  is the number of full twists on the  $r$  strands, as in Figure 1. Obviously,  $m$  and  $r$  must be positive and  $r \leq m$ . Both  $n$  and  $s$  can be positive or negative; hence there are four possibilities for the signs of the parameters. However, the determinant and  $p$ -colorability are the same for a knot and its mirror image, so we assume that  $n$  is positive throughout.

An important equivalence that we will use several times is described in the following theorem, which was shown by Dean [1996] for  $s = \pm 1$ . His arguments can be extended to any value for  $s$ .



**Figure 1.** The  $T(5, 4)$  torus knot changed into a  $T(5, 4, 3, 1)$  twisted torus knot.

**Theorem 2.1.** *The  $T(m, n, r, s)$  twisted torus knot is equivalent to the  $T(n, m, r, s)$  twisted torus knot.*

**2B. Colorability and determinants.** A knot is  $p$ -colorable if the strands in a projection of the knot can be labeled according to the following three conditions [Livingston 1993]. The first is that each strand must be labeled with an integer from 0 to  $p - 1$ . The second requires that at least two labels are distinct. The third requires that

$$x + y - 2z \equiv 0 \pmod{p} \tag{1}$$

at each crossing, where  $z$  is the label of the overstrand and  $x$  and  $y$  are the labels of the two understrands [loc. cit.]. Note that if a knot is colorable for some prime  $p$ , then it is colorable for any multiple of  $p$ .

A knot is  $p$ -colorable if and only if  $p$  divides the determinant of the knot. The *determinant* of a knot is the absolute value of the determinant of a minor crossing matrix constructed by removing a row and a column from the crossing matrix of a projection of the knot. A crossing matrix is a matrix representing the system of equations determined by requirement (1) at each crossing of a projection of the knot [loc. cit.].

The following result of Breiland et al. [2009] completely characterizes the colorability of torus knots. Recall that  $T(m, n)$  and  $T(n, m)$  are the same knot, so only two cases need to be considered.

**Theorem 2.2.** *Let  $T(m, n)$  be a torus knot and  $p$  a prime:*

- (i) *If  $m$  and  $n$  are both odd, then  $T(m, n)$  is not  $p$ -colorable.*
- (ii) *If  $m$  is odd and  $n$  is even, then  $T(m, n)$  is  $p$ -colorable if and only if  $p \mid m$ .*

Their proof was a direct consequence of the following lemma, which they proved by evaluating Alexander polynomials at  $t = -1$  [Livingston 1993].

**Lemma 2.3.** *For any torus knot  $T(m, n)$ ,*

- (i) *if  $m$  and  $n$  are odd, then  $\det(T(m, n)) = 1$ ;*
- (ii) *if  $m$  is odd and  $n$  is even, then  $\det(T(m, n)) = m$ .*

### 3. Methods

**3A. Computer experimentation.** We wrote a program in Matlab that input the four parameters of a twisted torus knot and output the determinant of a minor crossing matrix of the knot, which is equal to the determinant of the knot up to sign. Table 2 is a sample of the program’s output. The boldface lines identify the beginning of a new “family”, where we fix  $m$ ,  $n$ , and  $r$ , and let  $s$  vary.

$m$	$n$	$r$	$s$	$\det(C)$
<b>4</b>	<b>3</b>	<b>2</b>	<b>1</b>	<b>1</b>
4	3	2	2	-1
4	3	2	3	-3
4	3	2	4	-5
4	3	2	5	-7
<b>4</b>	<b>3</b>	<b>3</b>	<b>1</b>	<b>1</b>
4	3	3	2	3
4	3	3	3	1
4	3	3	4	3
4	3	3	5	1
<b>5</b>	<b>3</b>	<b>2</b>	<b>1</b>	<b>1</b>
5	3	2	2	3
5	3	2	3	5
5	3	2	4	7
5	3	2	5	9

$m$	$n$	$r$	$s$	$\det(C)$
<b>5</b>	<b>3</b>	<b>3</b>	<b>1</b>	<b>-3</b>
5	3	3	2	-1
5	3	3	3	-3
5	3	3	4	-1
5	3	3	5	-3
<b>5</b>	<b>3</b>	<b>4</b>	<b>1</b>	<b>-1</b>
5	3	4	2	-1
5	3	4	3	-1
5	3	4	4	-1
5	3	4	5	-1
<b>5</b>	<b>4</b>	<b>2</b>	<b>1</b>	<b>11</b>
5	4	2	2	17
5	4	2	3	23
5	4	2	4	29
5	4	2	5	35

**Table 2.** Experimental data on the determinants of twisted torus knot minor crossing matrices.

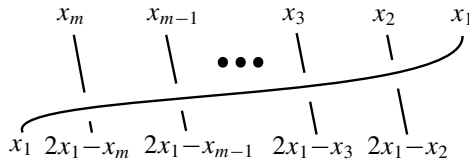
When  $r$  is even, the computed determinants of the  $T(m, n, r, s)$  twisted torus knots form an arithmetic progression in  $s$ . When  $r$  is odd, the computed determinants oscillate between two values as  $s$  varies. Two questions naturally arise: what determines the starting values and differences in the progressions and what determines the values in the oscillations? In trying to answer these questions, we were able to make conjectures for several families of twisted torus knots. The next two subsections develop the techniques that we used to prove our conjectures.

**3B. Definitions and notation.** We define a *coloring vector* as a vector

$$\mathbf{x} = (x_1, x_2, \dots, x_m)$$

that lists the colors of  $m$  strands of a twisted torus knot from right to left between two consecutive cycles (for example, see the top of Figure 2). We also define a *coloring matrix* as a matrix that operates on a coloring vector according to the coloring relation (1). A coloring matrix represents the changes that occur to the colors on the  $m$  strands after a specified number of cycles and/or twists.

We define  $\Gamma_m$  to be the coloring matrix that represents the change after one cycle of  $m$  strands. Therefore, for a twisted torus knot with  $m$  strands and  $n$  cycles, the coloring matrix that represents the changes through the torus part (the part above the twists) of the knot is  $\Gamma_m^n$ . The  $\Gamma_m$  matrix representing one cycle of an arbitrary



**Figure 2.** One cycle of an arbitrary knot.

knot is an  $m \times m$  matrix of the form

$$\Gamma_m = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & 0 \\ 2 & 0 & -1 & \cdots & 0 & 0 \\ 2 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 2 & 0 & 0 & \cdots & 0 & -1 \\ 1 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}, \tag{2}$$

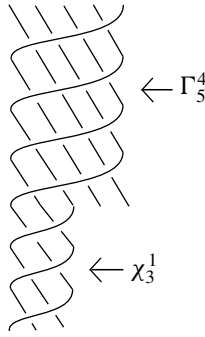
as can be seen from [Figure 2](#) (see also [\[Breiland et al. 2009\]](#)).

We define  $\chi_r$  as a coloring matrix that represents the change that occurs after one full twist of  $r$  strands in the lower part of a twisted torus knot projection. By definition,  $\chi_r = \Gamma_r^r$  since there will be  $r$  cycles on  $r$  strands in one full twist. Later in this section we will explore special properties of some powers of  $\chi_r$  matrices. Some of these properties have previously been stated by Przytycki [\[1998\]](#), using  $n$ -moves and half-twists.

Throughout, we will use  $\chi_r$  to symbolize the  $r \times r$  matrix that represents the changes occurring on only the  $r$  strands that are being twisted and also to symbolize the  $m \times m$  matrix that represents the changes on all  $m$  strands in the lower part of the diagram. In this case, the rightmost  $m - r$  strands are left unchanged, so this matrix will contain the original  $\chi_r$  matrix in the lower right, while also having 1s in the main diagonal from the upper left corner down to the start of the original  $\chi_r$  matrix. We hope that the distinction will be clear from the context.

If  $A_1, A_2, \dots, A_i$  are coloring matrices that represent all of the changes that occur to the coloring vectors, in order, from the top of a projection of a twisted torus knot to the bottom, then we can form an overall coloring matrix for the twisted torus knot  $A = A_i A_{i-1} \dots A_1$ . Then, if  $\mathbf{x}$  is the coloring vector at the top of the projection, the coloring vector  $\mathbf{x}'$  at the bottom of the projection can be found using  $A\mathbf{x} = \mathbf{x}' \pmod p$ . Thus, the twisted torus knot can be colored if and only if there exists a nonconstant vector  $\mathbf{x}$  such that  $A\mathbf{x} = \mathbf{x} \pmod p$ . In our calculations,  $A$  is generally equal to  $\chi_r^s \Gamma_m^n$  for the twisted torus knot  $T(m, n, r, s)$ . For an example, see [Figure 3](#).

**3C. Determinants.** The usual method of assessing  $p$ -colorability of a knot depends on the fact that the system of equations obtained from the coloring relation [\(1\)](#)



**Figure 3.** Coloring matrices for the  $T(5, 4, 3, 1)$  twisted torus knot.

at each crossing has a nontrivial solution mod  $p$  if and only if any minor of the crossing matrix of the knot has determinant divisible by  $p$  [Livingston 1993]. Here we describe a slightly different method for finding the determinant of a twisted torus knot that utilizes coloring matrices rather than crossing matrices. This method has the advantage of dealing with much smaller matrices, which have some very nice forms and useful properties.

Recall that a knot has a nontrivial  $p$ -coloring if and only if there is a nonconstant vector  $\mathbf{x}$  such that  $\mathbf{x} = A\mathbf{x} \pmod{p}$  for the coloring matrix  $A$ . So, we analyze the system of equations  $B\mathbf{x} = \mathbf{0} \pmod{p}$ , where  $B = A - I$ . Our treatment below of the matrix  $B$  mimics the usual treatment of a crossing matrix to find the determinant of a knot, as explained, for example, in [Livingston 1993].

First note that any constant vector  $\mathbf{x}$  satisfies  $A\mathbf{x} = \mathbf{x}$ , and so the system  $B\mathbf{x} = \mathbf{0}$  has nontrivial solutions. However, when considering colorability, we are only looking for nonconstant solutions. By linearity, any two solutions to  $B\mathbf{x} = \mathbf{0}$  can be added to yield another solution. Hence, if there were a nonconstant solution to  $B\mathbf{x} = \mathbf{0} \pmod{p}$ , then there must be one with  $x_i = 0$  for any choice of  $i$ .

Second, since the system  $B\mathbf{x} = \mathbf{0}$  has nontrivial solutions, the rows of  $B$  are linearly dependent. Moreover, as can be seen from the forms of the coloring matrices given in the sequel, and remembering that  $B = A - I$ , the matrix  $B$  has the property that multiplying every other row in the matrix by  $-1$  results in a matrix whose rows sum to the zero vector. This yields a dependence relation involving all the rows of  $B$ , and so any one of the equations represented by the matrix  $B$  is a result of the others.

Taking the two previous observations together, we note that in looking for nonconstant solutions, we can delete any row and any column from  $B$ , forming a minor that we denote as  $B'$ . Then, the knot has a nontrivial  $p$ -coloring if and only if  $p$  divides the determinant of  $B'$ . Moreover, since the matrix obtained from  $B$  by multiplying every other row by  $-1$  has the property that any row and any column sums to 0, the mod  $p$  rank is independent of which row and column are deleted [Livingston 1993].



This construction is the same as the “black-box approach” used by Kauffman and Lopes [2009] to find determinants of rational knots. There they argue that the absolute value of the determinant of what we are calling  $B'$  is equal to the classical determinant of the knot. We also note that the details of Oesper’s calculation [2005] of determinants of weaving knots show concretely, in a similar setting to ours, how the classical determinant is obtained from the determinant of a minor of what we are calling a coloring matrix.

**3D. Forms of matrices.** Recall that the coloring matrix  $\chi_k$  corresponds to a full twist on  $k$  strands. The form of  $\chi_k$  is

$$\begin{pmatrix} 1 & -2 & 2 & \cdots & 2 & -2 & 2 \\ 2 & -3 & 2 & \cdots & 2 & -2 & 2 \\ 2 & -2 & 1 & \cdots & 2 & -2 & 2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 2 & -2 & 2 & \cdots & 1 & -2 & 2 \\ 2 & -2 & 2 & \cdots & 2 & -3 & 2 \\ 2 & -2 & 2 & \cdots & 2 & -2 & 1 \end{pmatrix} \tag{3}$$

when  $k$  is odd, and

$$\begin{pmatrix} 3 & -2 & 2 & \cdots & 2 & -2 & 2 & -2 \\ 2 & -1 & 2 & \cdots & 2 & -2 & 2 & -2 \\ 2 & -2 & 3 & \cdots & 2 & -2 & 2 & -2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 2 & -2 & 2 & \cdots & 3 & -2 & 2 & -2 \\ 2 & -2 & 2 & \cdots & 2 & -1 & 2 & -2 \\ 2 & -2 & 2 & \cdots & 2 & -2 & 3 & -2 \\ 2 & -2 & 2 & \cdots & 2 & -2 & 2 & -1 \end{pmatrix} \tag{4}$$

when  $k$  is even, as can be shown by induction.

**3E. Properties of coloring matrices.** Let  $\chi_k$  be a coloring matrix, with  $k$  odd. Then,  $\chi_k$  has the form (3). Squaring this immediately yields the following lemma. Its corollary is similar to a result of Przytycki [1998].

**Lemma 3.1.** *For  $k$  odd, we have  $\chi_k^2 = I_k$ .*

**Corollary 3.2.** *An even twist of an odd number of strands applied to a  $p$ -colorable torus knot or twisted torus knot will result in a new knot that is also  $p$ -colorable.*

*Proof.* Since  $\chi_k^2 = I_k$  for  $k$  odd, it follows that any even twist of an odd number of strands will have the same colors at the top and bottom. □

By induction, one can see that the coloring matrix  $\chi_k^q$  for  $k$  even will have the form

$$\begin{pmatrix} 2q + 1 & -2q & 2q & \cdots & 2q & -2q \\ 2q & -2q + 1 & 2q & \cdots & 2q & -2q \\ 2q & -2q & 2q + 1 & \cdots & 2q & -2q \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 2q & -2q & 2q & \cdots & 2q + 1 & -2q \\ 2q & -2q & 2q & \cdots & 2q & -2q + 1 \end{pmatrix}. \tag{5}$$

Given this result, we can immediately prove another lemma. Again, a result similar to its corollary was also demonstrated by Przytycki [1998].

**Lemma 3.3.** *For  $k$  even, we have  $\chi_k^q \equiv I_k \pmod q$ .*

Obviously, we could have stated that for  $k$  even,  $\chi_k^q \equiv I_k \pmod{2q}$ . However, in this paper, we will only utilize the result as given in the lemma.

**Corollary 3.4.** *If the original torus knot was  $p$ -colorable, twisting an even number of strands  $s$  times, where  $p \mid s$ , will result in another  $p$ -colorable knot.*

*Proof.* We have  $\chi_k^s = \chi_k^{pj}$  for some  $j$ . Then,  $\chi_k^{pj} = I_k^j = I_k \pmod p$ . Therefore, when coloring mod  $p$ , the same colors will appear at the top and bottom of the twist.  $\square$

In our proofs, we will use a few special powers of the  $\Gamma_m$  matrices, which we now calculate. First, we find  $\Gamma_m^{mq+1}$  for  $m$  even. This is equal to  $\Gamma_m^{mq} \Gamma_m = \chi_m^q \Gamma_m$ . This is (5) times (2), which is

$$\begin{pmatrix} 2q + 2 & -2q - 1 & 2q & -2q & \cdots & -2q & 2q & -2q \\ 2q + 2 & -2q & 2q - 1 & -2q & \cdots & -2q & 2q & -2q \\ 2q + 2 & -2q & 2q & -2q - 1 & \cdots & -2q & 2q & -2q \\ 2q + 2 & -2q & 2q & -2q & \cdots & -2q & 2q & -2q \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 2q + 2 & -2q & 2q & -2q & \cdots & -2q & 2q - 1 & -2q \\ 2q + 2 & -2q & 2q & -2q & \cdots & -2q & 2q & -2q - 1 \\ 2q + 1 & -2q & 2q & -2q & \cdots & -2q & 2q & -2q \end{pmatrix}. \tag{6}$$

Here, we exhibit the form of  $\Gamma_m^{mq-1}$  for  $m$  even, which is

$$\begin{pmatrix} 2q & -2q & 2q & \cdots & 2q & -2q + 1 \\ 2q - 1 & -2q & 2q & \cdots & 2q & -2q + 2 \\ 2q & -2q - 1 & 2q & \cdots & 2q & -2q + 2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 2q & -2q & 2q & \cdots & 2q & -2q + 2 \\ 2q & -2q & 2q & \cdots & 2q - 1 & -2q + 2 \end{pmatrix}. \tag{7}$$

When we multiply (7) by (2), we obtain (5). Therefore, the matrix (7) has been shown to be  $\Gamma_m^{mq-1}$  since we have  $\Gamma_m^{mq-1} \Gamma_m = \Gamma_m^{mq} = \chi_m^q$  and  $\Gamma_m$  is invertible.

Finally, we calculate  $\Gamma_m^{2mq\pm 1}$  for  $m$  odd. Since  $\chi_m^{2q} = I_m$ ,

$$\Gamma_m^{2mq+1} = \Gamma_m^{2mq} \Gamma_m = I_m \Gamma_m = \Gamma_m. \tag{8}$$

Also,

$$\begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 1 \\ -1 & 0 & 0 & \cdots & 0 & 2 \\ 0 & -1 & 0 & \cdots & 0 & 2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 2 \\ 0 & 0 & 0 & \cdots & -1 & 2 \end{pmatrix} \tag{9}$$

times (2) is equal to  $I_m$ . Thus (9) is equal to  $\Gamma_m^{2mq-1}$  since  $\Gamma_m^{2mq-1} \Gamma_m = \Gamma_m^{2mq} = I_m$ .

### 4. Results

We now calculate the determinants of  $T(m, n, r, s)$ , for some families of the parameters. We find  $A = \chi_r^s \Gamma_m^n$  and then use the process from Section 3C to find the determinant of the knot by finding the determinant of a minor of  $A - I$ , which we do by row reduction. We use the second definition of  $\chi_r$  matrices given in Section 3B — that is, a  $\chi_r$  matrix is an  $m \times m$  matrix that contains  $m - r$  1s along the main diagonal and the rest of the nonzero entries in the lower right of the matrix. For  $r$  even, we have

$$\chi_r^s = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 2s + 1 & -2s & \cdots & 2s & -2s \\ 0 & 0 & \cdots & 0 & 2s & -2s + 1 & \cdots & 2s & -2s \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 2s & -2s & \cdots & 2s + 1 & -2s \\ 0 & 0 & \cdots & 0 & 2s & -2s & \cdots & 2s & -2s + 1 \end{pmatrix}. \tag{10}$$

Recall from Lemma 3.1 that  $\chi_r^2 = I_r$  for  $r$  odd. For  $r, s$  odd we have

$$\chi_r^s = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1 & -2 & \cdots & -2 & 2 \\ 0 & 0 & \cdots & 0 & 2 & -3 & \cdots & -2 & 2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 2 & -2 & \cdots & -3 & 2 \\ 0 & 0 & \cdots & 0 & 2 & -2 & \cdots & -2 & 1 \end{pmatrix}. \tag{11}$$

**4A.  $T(m, mq + 1, r, s)$  family with  $m$  even.** By [Theorem 2.1](#), the  $T(4, 5, 2, s)$  family of twisted torus knots is the same as the  $T(5, 4, 2, s)$  family of twisted torus knots. By [Table 2](#), we see that this family has determinants in an arithmetic progression with starting value 5 (the determinant of  $T(4, 5)$ ) and difference 6. This is a special case of the following theorem, which states that related families of twisted torus knots will have determinants in arithmetic progressions with starting values at the determinant of the (untwisted) torus knot and a difference that depends on  $m, n, r$ , and  $s$ .

**Theorem 4.1.** *A  $T(m, mq + 1, r, s)$  twisted torus knot, with  $m, r$  even and  $m > r$ , has determinant  $\Delta = |mq + 1 + rs + (m - r)qrs|$ .*

*Proof.* Multiply the  $\chi_r^s$  matrix [\(10\)](#) on the right by  $\Gamma_m^{mq+1}$  [\(6\)](#), yielding

$$\begin{pmatrix} 2q+2 & -2q-1 & \cdots & -2q & 2q & -2q & 2q & \cdots & 2q & -2q \\ 2q+2 & -2q & \cdots & -2q & 2q & -2q & 2q & \cdots & 2q & -2q \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 2q+2 & -2q & \cdots & -2q & 2q-1 & -2q & 2q & \cdots & 2q & -2q \\ 2q+2s+2 & -2q & \cdots & -2q & 2q & -2q-2s-1 & 2q+2s & \cdots & 2q+2s & -2q-2s \\ 2q+2s+2 & -2q & \cdots & -2q & 2q & -2q-2s & 2q+2s-1 & \cdots & 2q+2s & -2q-2s \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 2q+2s+2 & -2q & \cdots & -2q & 2q & -2q-2s & 2q+2s & \cdots & 2q+2s & -2q-2s-1 \\ 2q+2s+1 & -2q & \cdots & -2q & 2q & -2q-2s & 2q+2s & \cdots & 2q+2s & -2q-2s \end{pmatrix}.$$

Here,  $R_{m-r+1}$  is the first row with entries that contain an  $s$ . We subtract  $I_m$  and remove the first row and column:

$$\begin{pmatrix} -2q-1 & 2q-1 & \cdots & -2q & 2q & -2q & 2q & \cdots & 2q & -2q \\ -2q & 2q-1 & \cdots & -2q & 2q & -2q & 2q & \cdots & 2q & -2q \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -2q & 2q & \cdots & -2q-1 & 2q-1 & -2q & 2q & \cdots & 2q & -2q \\ -2q & 2q & \cdots & -2q & 2q-1 & -2q-2s-1 & 2q+2s & \cdots & 2q+2s & -2q-2s \\ -2q & 2q & \cdots & -2q & 2q & -2q-2s-1 & 2q+2s-1 & \cdots & 2q+2s & -2q-2s \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -2q & 2q & \cdots & -2q & 2q & -2q-2s & 2q+2s & \cdots & 2q+2s-1 & -2q-2s-1 \\ -2q & 2q & \cdots & -2q & 2q & -2q-2s & 2q+2s & \cdots & 2q+2s & -2q-2s-1 \end{pmatrix}.$$

To find the determinant of this matrix, we use elementary row operations to convert the matrix into an upper triangular matrix, whose determinant we can then easily compute by taking the product of the diagonal entries. Using the row operations

$R_1 \rightarrow R_1 - R_2, R_2 \rightarrow R_2 - R_3, \dots, R_{m-2} \rightarrow R_{m-2} - R_{m-1}$  yields the matrix

$$\begin{pmatrix} -1 & 0 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 0 & 1+2s & -2s & \cdots & 2s & -2s & 2s \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 0 & 1 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 0 \\ -2q & 2q & -2q & 2q & \cdots & -2q & 2q & -\alpha & \alpha & \cdots & -\alpha & \alpha & -\alpha - 1 \end{pmatrix},$$

where  $\alpha = 2q + 2s$ . (Note that the entries  $\pm 2s$  occur in row  $R_{m-r-1}$ .) We now reduce the last row using

$$R_{m-1} \rightarrow R_{m-1} + \sum_{i=1}^{(m-r)/2} 2iq(R_{2i} - R_{2i-1}),$$

$$R_{m-1} \rightarrow R_{m-1} + \sum_{i=1}^{(r-2)/2} ((m-r)(1+2is)q + 2i(q+s))(R_{m-r+2i} - R_{m-r+2i-1}).$$

This leaves us with

$$\begin{pmatrix} -1 & 0 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 0 & 1+2s & -2s & \cdots & 2s & -2s & 2s \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 0 & 1 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \Delta \end{pmatrix},$$

where

$$\Delta = -1 - 2q - 2s - q(2s)(m-r) - ((m-r)(1+(r-2)s)q + (r-2)(q+s)).$$

The determinant of this upper triangular matrix is  $\Delta$  since there are an even number of  $-1$ s along the diagonal. We can rewrite  $\Delta$  as  $-1 - mq - rs - (m - r)qrs$ . As we explained in [Section 3C](#), the determinant of the knot is the absolute value of the determinant of this matrix, so it follows that the determinant of the knot is equal to  $|1 + mq + rs + (m - r)qrs|$ .  $\square$

For these values of  $m$  and  $n$  but odd  $r$ , a different phenomenon results. For example, the  $T(5, 4, 3, s)$  family has determinants that oscillate between 5 (the determinant of  $T(5, 4)$ ) and 7. Next we show that this is representative of related families of twisted torus knots, which have determinants that oscillate between the determinant of the untwisted knot and another value that depends on  $m$ ,  $n$ , and  $r$ . We first prove the following lemma for  $s = 1$ .

**Lemma 4.2.** *A  $T(m, mq + 1, r, 1)$  twisted torus knot, with  $m$  even and  $r$  odd, has determinant  $\Delta = |r + (mr - r^2 + 1)q|$ .*

*Proof.* Multiply the  $\chi_r^s$  matrix by  $\Gamma_m^{mq+1}$ . This is (11) times (6), which equals

$$\begin{pmatrix} 2q+2 & -2q-1 & \cdots & 2q & -2q & 2q & -2q & \cdots & 2q & -2q \\ 2q+2 & -2q & \cdots & 2q & -2q & 2q & -2q & \cdots & 2q & -2q \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 2q+2 & -2q & \cdots & 2q & -2q-1 & 2q & -2q & \cdots & 2q & -2q \\ 2q & -2q & \cdots & 2q & -2q & 2q-1 & -2q+2 & \cdots & 2q-2 & -2q+2 \\ 2q & -2q & \cdots & 2q & -2q & 2q-2 & -2q+3 & \cdots & 2q-2 & -2q+2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 2q & -2q & \cdots & 2q & -2q & 2q-2 & -2q+2 & \cdots & 2q-1 & -2q+3 \\ 2q & -2q & \cdots & 2q & -2q & 2q-2 & -2q+2 & \cdots & 2q-2 & -2q+3 \\ 2q+1 & -2q & \cdots & 2q & -2q & 2q-2 & -2q+2 & \cdots & 2q-2 & -2q+2 \end{pmatrix}.$$

Note the change from row  $R_{m-r}$  to  $R_{m-r+1}$ . Subtract  $I_m$  and remove the first row and column:

$$\begin{pmatrix} -2q-1 & 2q-1 & \cdots & 2q & -2q & 2q & -2q & \cdots & 2q & -2q \\ -2q & 2q-1 & \cdots & 2q & -2q & 2q & -2q & \cdots & 2q & -2q \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -2q & 2q & \cdots & 2q-1 & -2q-1 & 2q & -2q & \cdots & 2q & -2q \\ -2q & 2q & \cdots & 2q & -2q-1 & 2q-1 & -2q+2 & \cdots & 2q-2 & -2q+2 \\ -2q & 2q & \cdots & 2q & -2q & 2q-3 & -2q+3 & \cdots & 2q-2 & -2q+2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -2q & 2q & \cdots & 2q & -2q & 2q-2 & -2q+2 & \cdots & 2q-1 & -2q+2 \\ -2q & 2q & \cdots & 2q & -2q & 2q-2 & -2q+2 & \cdots & 2q-3 & -2q+3 \\ -2q & 2q & \cdots & 2q & -2q & 2q-2 & -2q+2 & \cdots & 2q-2 & -2q+1 \end{pmatrix}.$$

Reducing with  $R_1 \rightarrow R_1 - R_2, R_2 \rightarrow R_2 - R_3, \dots, R_{m-2} \rightarrow R_{m-2} - R_{m-1}$  gives

$$\begin{pmatrix} -1 & 0 & 1 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -1 & 0 & 1 & -2 & \dots & -2 & 2 & -2 \\ 0 & 0 & 0 & 0 & \dots & 0 & -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & -1 & 2 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & -1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & -1 & 2 \\ -2q & 2q & -2q & 2q & \dots & 2q & -2q & 2q & -2 & -2q+2 & \dots & -2q+2 & 2q-2 & -2q+2 \end{pmatrix},$$

where the row containing the  $\pm 2s$  is  $R_{m-r-1}$ . We now reduce the last row using

$$\begin{aligned} R_{m-1} &\rightarrow R_{m-1} + \sum_{i=1}^{(m-r-1)/2} 2iq(R_{2i} - R_{2i-1}), \\ R_{m-1} &\rightarrow R_{m-1} + \sum_{i=1}^{(r-3)/2} (((2i+1)(m-r)+1)q+2i)R_{m-r+2i} \\ &\quad - \sum_{i=1}^{(r-1)/2} (((2i-1)(m-r)+1)q+2i)R_{m-r+2i-1}. \end{aligned}$$

We now have the upper triangular matrix

$$\begin{pmatrix} -1 & 0 & 1 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -1 & 0 & 1 & -2 & \dots & -2 & 2 & -2 \\ 0 & 0 & 0 & 0 & \dots & 0 & -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & -1 & 2 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & -1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & -1 & 2 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & \Delta \end{pmatrix},$$

where

$$\Delta = 1 - 2q - 2(m-r-1)q + ((r-2)(m-r)+1)q + r - 3 - 2(((r-2)(m-r)+1) + r - 1).$$

Since there are an even number of  $-1$ s on the diagonal, the determinant is  $\Delta$ , which simplifies to  $-r - (mr - r^2 + 1)q$ . The determinant of the knot is then  $|r + (mr - r^2 + 1)q|$ .  $\square$

This immediately leads into a theorem:

**Theorem 4.3.** *A  $T(m, mq + 1, r, s)$  twisted torus knot, with  $m$  even and  $r$  odd, has determinant  $\Delta = |r + (mr - r^2 + 1)q|$  if  $s$  is odd, and determinant  $\Delta = |mq + 1|$  if  $s$  is even.*

*Proof.* If  $s$  is odd,  $\chi_r^s$  will equal the one used in the proof of Lemma 4.2, so the determinant of  $T(m, mq + 1, r, s)$  would equal that of  $T(m, mq + 1, r, 1)$ . If  $s$  is even,  $\chi_r^s$  will be the identity, so the determinant of the knot would simply be the determinant of the  $T(m, mq + 1)$  torus knot, which is  $mq + 1$  by Lemma 2.3, since  $m$  is even and  $mq + 1$  is odd.  $\square$

**4B.  $T(m, mq - 1, r, s)$  family with  $m$  even.** We now proceed to investigate a similar family to the one just analyzed. In these proofs, instead of using some power of  $\Gamma_m$  that has a diagonal with  $-1$ s in it to the upper right of the main diagonal, as in (6), we utilize different powers of  $\Gamma_m$  that have the property that there is a diagonal with  $-1$ s in it to the lower left of the main diagonal, as in (7). By glancing at the values for the  $T(4, 3, 2, s)$  family in Table 2, we conjecture that we will have an arithmetic progression beginning at the determinant of the  $T(4, 3)$  torus knot. We now prove that this is the case.

**Theorem 4.4.** *A  $T(m, mq - 1, r, s)$  twisted torus knot, with  $m, r$  even, has determinant  $\Delta = |mq - 1 + rs - (m-r)qrs|$ .*

*Proof.* Multiply the  $\chi_r^s$  matrix by  $\Gamma_m^{mq-1}$ . This will be (10) times (7), which is

$$\begin{pmatrix} 2q & -2q & \cdots & 2q & -2q & 2q & -2q & \cdots & 2q & -2q+1 \\ 2q-1 & -2q & \cdots & 2q & -2q & 2q & -2q & \cdots & 2q & -2q+2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 2q & -2q & \cdots & 2q-1 & -2q & 2q & -2q & \cdots & 2q & -2q+2 \\ 2q & -2q & \cdots & 2q & -2q-2s-1 & 2q+2s & -2q-2s & \cdots & 2q+2s & -2q+2 \\ 2q & -2q & \cdots & 2q & -2q-2s & 2q+2s-1 & -2q-2s & \cdots & 2q+2s & -2q+2 \\ 2q & -2q & \cdots & 2q & -2q-2s & 2q+2s & -2q-2s-1 & \cdots & 2q+2s & -2q+2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 2q & -2q & \cdots & 2q & -2q-2s & 2q+2s & -2q-2s & \cdots & 2q+2s & -2q+2 \\ 2q & -2q & \cdots & 2q & -2q-2s & 2q+2s & -2q-2s & \cdots & 2q+2s-1 & -2q+2 \end{pmatrix}.$$



We subtract  $I_m$  from this. At this point, instead of deleting the first row and column as we have done previously, we choose to remove the last row and column:

$$\begin{pmatrix} 2q-1 & -2q & \cdots & 2q & -2q & 2q & \cdots & -2q & 2q \\ 2q-1 & -2q-1 & \cdots & 2q & -2q & 2q & \cdots & -2q & 2q \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 2q & -2q & \cdots & 2q-1 & -2q-1 & 2q & \cdots & -2q & 2q \\ 2q & -2q & \cdots & 2q & -2q-2s-1 & 2q+2s-1 & \cdots & -2q-2s & 2q+2s \\ 2q & -2q & \cdots & 2q & -2q-2s & 2q+2s-1 & \cdots & -2q-2s & 2q+2s \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 2q & -2q & \cdots & 2q & -2q-2s & 2q+2s & \cdots & -2q-2s-1 & 2q+2s \\ 2q & -2q & \cdots & 2q & -2q-2s & 2q+2s & \cdots & -2q-2s-1 & 2q+2s-1 \end{pmatrix}.$$

The first row with entries containing a term with an  $s$  is  $R_{m-r+1}$ . We now reduce using the row operations

$$\begin{aligned} R_2 &\rightarrow R_2 - R_3, & R_3 &\rightarrow R_3 - R_4, & \dots, & R_{m-2} &\rightarrow R_{m-2} - R_{m-1}, \\ R_{m-1} &\rightarrow R_{m-1} - R_1, & R_1 &\rightarrow R_1 + R_{m-1}. \end{aligned} \quad (12)$$

Additionally, we cyclically permute the rows by moving  $R_1$  to the bottom, while shifting all of the other rows up by one. This puts the diagonal of  $-1$ s on the main diagonal using an even number of switches. Thus, the determinant remains unchanged. The matrix becomes

$$\begin{pmatrix} -1 & 0 & 1 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 & 1 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 2s & -2s+1 & 2s & \cdots & -2s & 2s & -2s \\ 0 & 0 & 0 & \cdots & 0 & -1 & 0 & 1 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & -1 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & -1 & 0 & 1 \\ 1 & 0 & 0 & \cdots & 0 & -2s & 2s & -2s & \cdots & 2s & -2s-1 & 2s-1 \\ 2q & -2q & 2q & \cdots & 2q & -\alpha & \alpha & -\alpha & \cdots & \alpha & -\alpha-1 & \alpha-1 \end{pmatrix},$$

where  $R_{m-r-1}$  is the first row with entries  $\pm 2s$ . (As before,  $\alpha = 2q + 2s$ .) We now reduce  $R_{m-2}$  with

$$R_{m-2} \rightarrow R_{m-2} + \sum_{i=1}^{(m-2)/2} R_{2i-1}.$$

We then reduce  $R_{m-1}$  with

$$\begin{aligned}
 R_{m-1} &\rightarrow R_{m-1} + \sum_{i=1}^{(m-r-2)/2} 2iq(R_{2i-1} - R_{2i}) + (m-r)qR_{m-r-1}, \\
 R_{m-1} &\rightarrow R_{m-1} + \sum_{i=1}^{r/2} ((m-r)(2iqs - q) - (2i-2)q - 2is)R_{m-r-2+2i}, \\
 R_{m-1} &\rightarrow R_{m-1} - \sum_{i=1}^{(r-2)/2} ((m-r)(2iqs - q) - 2iq - 2is)R_{m-r-1+2i}.
 \end{aligned}$$

Now we have successfully reduced the matrix into an upper-triangular matrix

$$\begin{pmatrix}
 -1 & 0 & 1 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\
 0 & -1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\
 \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
 0 & 0 & 0 & \cdots & 0 & 1 & 0 & 1 & \cdots & 0 & 0 & 0 \\
 0 & 0 & 0 & \cdots & -1 & 2s & -2s + 1 & 2s & \cdots & -2s & 2s & -2s \\
 0 & 0 & 0 & \cdots & 0 & -1 & 0 & 1 & \cdots & 0 & 0 & 0 \\
 0 & 0 & 0 & \cdots & 0 & 0 & -1 & 0 & \cdots & 0 & 0 & 0 \\
 \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & 0 \\
 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & -1 & 0 & 1 \\
 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 0 \\
 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \Delta
 \end{pmatrix}.$$

with determinant

$$\Delta = 2q + 2s - 1 - 2s(m-r)q - ((m-r)((r-2)qs - q) - (r-2)q - (r-2)s).$$

As before, there are an even number of  $-1$ s on the diagonal, and the row operations did not affect the determinant. Simplifying  $\Delta$ , the determinant of the knot is  $|-1 + mq + rs - (m-r)qrs|$ . □

To investigate this family when  $r$  is odd, we begin with a lemma for the case  $s = 1$ .

**Lemma 4.5.** *A  $T(m, mq - 1, r, 1)$  twisted torus knot, with  $m$  even and  $r$  odd, has determinant  $\Delta = |r - (mr - r^2 + 1)q|$ .*

*Proof.* Multiply the  $\chi_r^s$  matrix by  $\Gamma_m^{mq-1}$ . This is (11) multiplied by (7), which gives

$$\begin{pmatrix} 2q & -2q & \cdots & -2q & 2q & -2q & 2q & \cdots & -2q & 2q & -2q+1 \\ 2q-1 & -2q & \cdots & -2q & 2q & -2q & 2q & \cdots & -2q & 2q & -2q+2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 2q & -2q & \cdots & -2q-1 & 2q & -2q & 2q & \cdots & -2q & 2q & -2q+2 \\ 2q & -2q & \cdots & -2q & 2q-1 & -2q+2 & 2q-2 & \cdots & -2q+2 & 2q-2 & -2q+2 \\ 2q & -2q & \cdots & -2q & 2q-2 & -2q+3 & 2q-2 & \cdots & -2q+2 & 2q-2 & -2q+2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 2q & -2q & \cdots & -2q & 2q-2 & -2q+2 & 2q-2 & \cdots & -2q+2 & 2q-2 & -2q+2 \\ 2q & -2q & \cdots & -2q & 2q-2 & -2q+2 & 2q-2 & \cdots & -2q+3 & 2q-2 & -2q+2 \\ 2q & -2q & \cdots & -2q & 2q-2 & -2q+2 & 2q-2 & \cdots & -2q+2 & 2q-1 & -2q+2 \end{pmatrix}.$$

As in the previous proof, we delete the last row and column after subtracting  $I_m$ :

$$\begin{pmatrix} 2q-1 & -2q & \cdots & -2q & 2q & -2q & 2q & \cdots & -2q & 2q \\ 2q-1 & -2q-1 & \cdots & -2q & 2q & -2q & 2q & \cdots & -2q & 2q \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 2q & -2q & \cdots & -2q-1 & 2q-1 & -2q & 2q & \cdots & -2q & 2q \\ 2q & -2q & \cdots & -2q & 2q-1 & -2q+1 & 2q-2 & \cdots & -2q+2 & 2q-2 \\ 2q & -2q & \cdots & -2q & 2q-2 & -2q+3 & 2q-3 & \cdots & -2q+2 & 2q-2 \\ 2q & -2q & \cdots & -2q & 2q-2 & -2q+2 & 2q-1 & \cdots & -2q+2 & 2q-2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 2q & -2q & \cdots & -2q & 2q-2 & -2q+2 & 2q-2 & \cdots & -2q+2 & 2q-2 \\ 2q & -2q & \cdots & -2q & 2q-2 & -2q+2 & 2q-2 & \cdots & -2q+1 & 2q-2 \\ 2q & -2q & \cdots & -2q & 2q-2 & -2q+2 & 2q-2 & \cdots & -2q+3 & 2q-3 \end{pmatrix}.$$

We apply the row operations given in (12). Also,  $R_1$  is moved to the bottom, and the other rows are shifted up one, giving

$$\begin{pmatrix} -1 & 0 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 0 & -1 & 2 & -2 & \cdots & -2 & 2 & -2 & 2 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & -2 & 1 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 1 & -2 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & -2 & 1 \\ -1 & 0 & 0 & 0 & \cdots & 0 & -2 & 2 & -2 & 2 & \cdots & 2 & -2 & 3 & -3 \\ 2q & -2q & 2q & -2q & \cdots & -2q & \beta & -\beta & \beta & -\beta & \cdots & -\beta & \beta & -\beta+1 & \beta-1 \end{pmatrix}.$$

Here,  $R_{m-r-1}$  contains the sequence of alternating  $\pm 2s$  and  $\beta = 2q - 2$ . The absolute value of the determinant is unchanged by these row operations. To reduce  $R_{m-2}$ , we use

$$R_{m-2} \rightarrow R_{m-2} + \sum_{i=1}^{(m-2)/2} R_{2i-1}.$$

In so doing, we find that adding  $R_{m-r-1}$  to it creates a lot of cancellation. For the last row, we use

$$R_{m-1} \rightarrow R_{m-1} + \sum_{i=1}^{(m-r-1)/2} 2qi(R_{2i-1} - R_{2i}),$$

$$R_{m-1} \rightarrow R_{m-1} - \sum_{i=1}^{(r-1)/2} (((2i-1)(m-r)+1)q-2i)R_{m-r-2+2i},$$

$$R_{m-1} \rightarrow R_{m-1} - \sum_{i=1}^{(r-1)/2} (((2i+1)(m-r)-1)q-2i)R_{m-r-1+2i}.$$

Our matrix has been transformed into

$$\begin{pmatrix} -1 & 0 & 1 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & \dots & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & \dots & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -1 & 0 & -1 & 2 & -2 & \dots & -2 & 2 & -2 & 2 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & -2 & 1 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & -2 & 1 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 & -2 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & \dots & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \Delta \end{pmatrix},$$

for

$$\Delta = 2q - 3 - 2(m-r-1)q - (((r-2)(m-r)+1)q - (r-1)) + 2((r(m-r)-1)q - (r-2)).$$

There are  $m-r-1$  entries of  $-1$  on the main diagonal. Since  $m-r-1$  is even, the determinant of this matrix is  $\Delta$ , which simplifies to  $-r + (mr - r^2 + 1)q$ . The determinant of the knot is then  $|r - (mr - r^2 + 1)q|$ . □

As in the proof of [Theorem 4.3](#), this lemma leads directly to a corresponding theorem.

**Theorem 4.6.** *A  $T(m, mq - 1, r, s)$  twisted torus knot, with  $m$  even and  $r$  odd, has determinant  $\Delta = |r - (mr - r^2 + 1)q|$  if  $s$  is odd, and determinant  $\Delta = |mq - 1|$  if  $s$  is even.*

**4C.  $T(m, 2mq + 1, r, s)$  family with  $m$  odd.** Now we begin our discussion of twisted torus knots when both  $m$  and  $n$  are odd. This represents a major change for two reasons. First, the  $T(m, n)$  torus knot that we begin with will no longer be  $p$ -colorable for any  $p$ ; by [Lemma 2.3](#), it will have a determinant of 1. Additionally, the powers of the  $\Gamma_m$  matrices that we use will no longer have  $qs$  in them. However, after examination of [Table 2](#), the trend of having either an oscillating pattern or an arithmetic progression appears to hold when  $m$  and  $n$  are both odd (the determinants of the  $T(5, 3, 4, s)$  family form an arithmetic progression with difference 0). Although the details are slightly different, the methods of this section closely follow those of [Section 4A](#). For space considerations, we suppress the matrices involved and only record the arithmetic details. We trust that the reader could supply the matrices if desired.

**Theorem 4.7.** *A  $T(m, 2mq + 1, r, s)$  twisted torus knot, with  $m$  odd,  $r$  even, and  $m > r$ , has determinant  $\Delta = |rs + 1|$ .*

*Proof.* Multiply the  $\chi_r^s$  matrix by  $\Gamma_m^{2mq+1}$ . By [\(8\)](#), this will be [\(10\)](#) times [\(2\)](#). As we did in [Section 4A](#), we will return to our method of subtracting  $I_m$  and removing the first row and column. We do not have to reduce any of the first  $m - r$  rows, as there are no entries to the left of the long diagonal in these rows. (The first row containing  $2s$  and  $-2s$  happens to be  $R_{m-r}$ .) Therefore, we use a different process of row operations, as we only will work with the last  $r$  rows, as follows:

$$\begin{aligned} R_{m-r+1} &\rightarrow R_{m-r+1} - R_{m-r+2}, \\ R_{m-r+2} &\rightarrow R_{m-r+2} - R_{m-r+3}, \dots R_{m-2} \rightarrow R_{m-2} - R_{m-1}. \end{aligned} \tag{13}$$

All that remains is to reduce  $R_{m-1}$ . Our procedure for doing this is

$$R_{m-1} \rightarrow R_{m-1} + \sum_{i=1}^{(r-2)/2} 2si(R_{m-r+2i} - R_{m-r+2i-1}).$$

This converts the matrix into an upper triangular matrix with an odd number of  $-1$ s along the diagonal and  $-\Delta = -2s - 1 - (r - 2)s$  as the only other diagonal entry. The determinant of this matrix is then  $\Delta = 1 + rs$ . The determinant of the knot is thus  $|1 + rs|$ . □

Similarly, we can prove that when  $r$  is odd the determinants will oscillate. However, they now oscillate between 1 and some other value, as the determinant of a  $T(m, 2mq + 1)$  torus knot is 1 by [Lemma 2.3](#), because both  $m$  and  $2mq + 1$  are odd.

**Lemma 4.8.** *A  $T(m, 2mq + 1, r, 1)$  twisted torus knot, with  $m, r$  odd, has determinant  $\Delta = r$ .*

*Proof.* Multiply the  $\chi_r^s$  matrix by  $\Gamma_m^{2mq+1}$ . By (8), we have (11) multiplied by (2). We subtract  $I_m$  and remove the first row and column. Again, we do not have to reduce the first  $m - r$  rows. (The first row with more than two entries is  $R_{m-r}$ .) We use the row operations given in (13) on the remaining rows.

The last row is the only one preventing an upper-triangular matrix. We remedy this with

$$R_{m-1} \rightarrow R_{m-1} - \sum_{i=1}^{(r-3)/2} 2i(R_{m-r+2i} + R_{m-r+2i-1}) - (r-1)R_{m-2}.$$

This leaves an upper triangular matrix with an odd number of  $-1$ s on the diagonal and  $-\Delta$  in the last diagonal entry, where  $-\Delta = 1 + (r-3) - 2(r-1)$ . The determinant of this upper triangular matrix is  $\Delta$ . Fortunately,  $\Delta$  simplifies to  $r$ . The determinant of the knot is then just  $r$ . (Note that  $r$  can never be negative, as it represents the number of strands.)  $\square$

Again this lemma leads to a full theorem.

**Theorem 4.9.** *A  $T(m, 2mq + 1, r, s)$  twisted torus knot, with  $m, r$  odd, has determinant  $\Delta = r$  if  $s$  is odd, and determinant  $\Delta = 1$  if  $s$  is even.*

**4D.  $T(m, 2mq - 1, r, s)$  family with  $m$  odd.** The final family that we will investigate with our procedure is the  $T(m, 2mq - 1, r, s)$  family. In many ways, these proofs correspond to those presented in [Section 4B](#), which deal with the  $T(m, mq - 1, r, s)$  family, just as the proofs from [Section 4C](#) correspond to those from [Section 4A](#). This is due to the fact that the diagonal with  $-1$ s is to the lower left of the main diagonal, instead of the upper right. As in the previous section we suppress the matrices to save space.

**Theorem 4.10.** *A  $T(m, 2mq - 1, r, s)$  twisted torus knot, with  $m$  odd,  $r$  even, and  $m > r$ , has determinant  $\Delta = |rs - 1|$ .*

*Proof.* Multiply the  $\chi_r^s$  matrix by  $\Gamma_m^{2mq-1}$ , which is (10) times (9). As in the proofs of [Theorem 4.4](#) and [Lemma 4.5](#), we opt to delete the last row and column after subtracting  $I_m$ . Here, the first row with entries  $\pm 2s$  is  $R_{m-r+1}$ . In this proof, we use a different method of turning this matrix into a triangular matrix. Instead of subtracting each row from the row above it and ending up with an upper triangular matrix, we choose to subtract each row from the row below it, eventually reaching

a lower triangular matrix. This avoids any need to cyclically permute the rows. Our row operations are

$$\begin{aligned} R_{m-1} &\rightarrow R_{m-1} - R_{m-2}, \\ R_{m-2} &\rightarrow R_{m-2} - R_{m-3}, \dots R_{m-r+2} \rightarrow R_{m-r+2} - R_{m-r+1}. \end{aligned} \tag{14}$$

Because of our different procedure, we must reduce  $R_{m-r+1}$  (not  $R_{m-1}$ ). We use

$$R_{m-r+1} \rightarrow R_{m-r+1} + \sum_{i=1}^{(r-2)/2} 2is(R_{m-2i+1} - R_{m-2i}).$$

This gives a lower triangular matrix with an odd number of  $-1$ s along the diagonal and  $-\Delta = 2s - 1 + (r - 2)s$  in row  $R_{m-r+1}$  as the only other entry on the diagonal. The determinant of this matrix is  $\Delta = -1 + rs$ , and so the determinant of the knot is  $|-1 + rs|$ . □

Our final proof of this type investigates a case where  $r$  is odd. Again, we are confirmed by [Table 2](#), in which one family satisfying the following conditions is  $T(5, 3, 3, s)$ .

**Lemma 4.11.** *A  $T(m, 2mq - 1, r, 1)$  twisted torus knot, with  $m, r$  odd, and  $m > r$ , has determinant  $\Delta = r$ .*

*Proof.* Multiply the  $\chi_r^s$  matrix by  $\Gamma_m^{2mq-1}$ . This will be (11) multiplied by (9). As in the proof of [Theorem 4.10](#), we subtract  $I_m$  and remove the last row and column. We again choose to subtract each row (beginning with  $R_{m-r+1}$ ) from the row below it, with the intention of finding a lower-triangular matrix. Our row operations are those given in (14).

All that remains is to reduce  $R_{m-r+1}$ , which we do with

$$R_{m-r+1} \rightarrow R_{m-r+1} - \sum_{i=1}^{(r-1)/2} 2i R_{m-2i+1} - \sum_{i=1}^{(r-3)/2} 2i R_{m-2i}.$$

This leaves a lower triangular matrix with an odd number of  $-1$ s along the diagonal, with the only other entry on the diagonal being  $-\Delta = 1 + (r - 3) - 2(r - 1)$  in  $R_{m-r+1}$ . The determinant of this matrix is  $\Delta = r$ . Thus, the determinant of the knot is  $r$  (which is always positive). □

Naturally, this lemma gives a similar theorem.

**Theorem 4.12.** *A  $T(m, 2mq - 1, r, s)$  twisted torus knot, with  $m, r$  odd, has determinant  $\Delta = r$  if  $s$  is odd, and determinant  $\Delta = 1$  if  $s$  is even.*

**4E.**  $T(m, (2q+1)m+1, r, s)$  and  $T(m, (2q+1)m-1, r, s)$  families with  $m$  odd.

In this section, we use our previous results to prove some important corollaries.

**Corollary 4.13.** *The determinant of a  $T(m, (2q+1)m+1, r, s)$  twisted torus knot is  $\Delta = |mr - r^2 + 1|$  for  $m, r, s$  odd, and  $\Delta = m$  for  $m, r$  odd and  $s$  even.*

*Proof.* First, consider the case of  $T(m, m+1, r, s)$ . Using [Theorem 2.1](#), we rewrite this knot as  $T(m+1, m, r, s)$ . By [Theorem 4.6](#), we see that its determinant is  $\Delta = |r - ((m+1)r - r^2 + 1)| = |mr - r^2 + 1|$  for  $s$  odd, and  $\Delta = m$  for  $s$  even. Therefore, these are the determinants for the  $T(m, m+1, r, s)$  knots. Since  $\chi_m^2 = I_m$  by [Lemma 3.1](#), adding  $2qm$  cycles doesn't change the determinant, so  $\det(T(m, (2q+1)m+1, r, s)) = \det(T(m+1, m, r, s))$  for any  $q$ .  $\square$

The following three corollaries similarly follow from [Theorems 4.4, 4.3, and 4.1](#).

**Corollary 4.14.** *The determinant of a  $T(m, (2q+1)m+1, r, s)$  twisted torus knot is  $\Delta = |m - (m-r)rs|$  for  $m$  odd and  $r$  even.*

**Corollary 4.15.** *The determinant of a  $T(m, (2q+1)m-1, r, s)$  twisted torus knot is  $\Delta = |mr - r^2 + 1|$  for  $m, r, s$  odd, and  $\Delta = m$  for  $m, r$  odd and  $s$  even.*

**Corollary 4.16.** *The determinant of a  $T(m, (2q+1)m-1, r, s)$  twisted torus knot is  $\Delta = |m + (m-r)rs|$  for  $m$  odd and  $r$  even.*

These four corollaries, together with the theorems presented in [Sections 4C and 4D](#), complete all cases when  $n \equiv \pm 1 \pmod{m}$  because if  $n \equiv \pm 1 \pmod{m}$ , then  $n \equiv \pm 1 \pmod{2m}$  or  $n \equiv \pm m + 1 \pmod{2m}$ . The theorems from [Sections 4C and 4D](#) took care of  $n \equiv \pm 1 \pmod{2m}$ , while the four corollaries here fully covered the cases  $n \equiv \pm m + 1 \pmod{2m}$ .

**4F. Counting  $p$ -colorings.** The  $p$ -nullity of a knot is the dimension of the mod  $p$  nullspace of a crossing matrix for the knot. A knot with  $p$ -nullity  $n$  has  $p^n - p$  different  $p$ -colorings because there are  $n$  strands that can be assigned any of  $p$  different colors, whereas the remaining strands are then determined (subtracting  $p$  discards the trivial ‘‘colorings’’) [[Brownell et al. 2006](#)]. Two colorings of a knot are *fundamentally different* if they are not simply permutations of each other. If two colorings are fundamentally different, then they belong to different  $p$ -coloring classes; otherwise, they are in the same  $p$ -coloring class. Breiland, et al. [[2009](#)] showed that if a torus knot is  $p$ -colorable, then it has only one nontrivial  $p$ -coloring class. Our methods show a similar result for the twisted torus knots that we analyzed.

**Theorem 4.17.** *If a twisted torus knot  $T(m, n, r, s)$ , with  $n \equiv \pm 1 \pmod{m}$ , is  $p$ -colorable, it has  $p^2 - p$  different  $p$ -colorings, and hence only one nontrivial  $p$ -coloring class.*

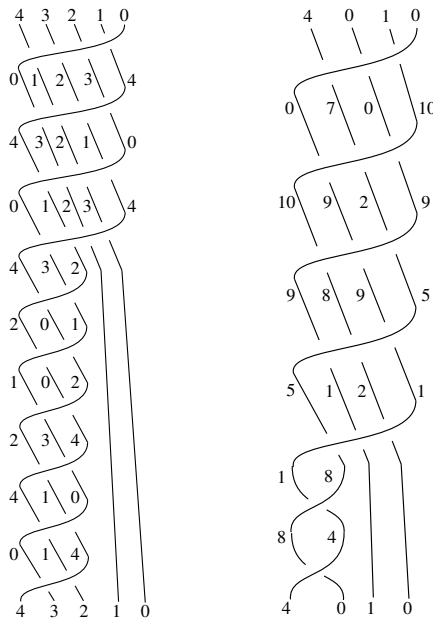


*Proof.* In each of our proofs,  $B'$  was converted into a triangular matrix by row reduction. Note that all of the row operations were valid mod  $p$  for any  $p$ , and so the mod  $p$  nullspace of the matrix was unchanged. After reduction, all but one of the entries on the main diagonal were equal to  $\pm 1$ . If the knot being analyzed was  $p$ -colorable — that is, if  $p \mid \Delta$  — then there was only one value on the diagonal of the reduced matrix that was divisible by  $p$ . Thus, in assigning the values of the labels to the top strands, there were two free variables: one for the deleted column, and one for the column containing  $\pm \Delta$ . This implies that the  $p$ -nullity of the knot was 2.  $\square$

### 5. Conclusion

While the theorems presented in this paper provide examples of determinants from each of the possible combinations of the parities of the parameters of twisted torus knots, they do not completely characterize the determinants of all twisted torus knots. A natural goal would be a complete characterization. It may be possible to generalize the methods presented in this paper to all twisted torus knots; however, the families investigated in this paper were chosen because their matrices allowed for straightforward row-reduction schemes.

Future research could also investigate the patterns in labelings of twisted torus knots, two examples of which are shown in Figure 4. Breiland et al. [2009] showed that all possible  $p$ -colorings of a torus knot were equivalent under permutation of



**Figure 4.** A 5-coloring of the  $T(5, 4, 3, 2)$  twisted torus knot and an 11-coloring of the  $T(4, 5, 2, 1)$  twisted torus knot.

the labels to a “main coloring,” which arose from labeling the uppermost strands of their projection with  $0, 1, \dots, p - 1$ , in that order. However, many  $p$ -colorable twisted torus knots cannot be colored in this fashion — for example, the  $T(4, 5, 2, 1)$  twisted torus knot, which has determinant 11 by [Theorem 4.1](#), cannot be 11-colored this way. Alternatively, the  $T(5, 4, 3, 2)$  twisted torus knot, which has determinant 5 by [Corollary 4.15](#), can be 5-colored using the main coloring. It would be interesting to determine which twisted torus knots can be  $p$ -colored using the main coloring.

### Acknowledgements

This research was funded by a Taylor University Step Grant. We are indebted to Thomas Mattman for suggesting this project. We are also grateful to Colin Adams, Thomas Mattman, Laura Taalman, Cornelia Van Cott, and the anonymous referee for providing helpful feedback.

### References

- [Adams 2004] C. C. Adams, *The knot book: An elementary introduction to the mathematical theory of knots*, American Mathematical Society, Providence, RI, 2004. [MR 2005b:57009](#) [Zbl 1065.57003](#)
- [Breiland et al. 2009] A.-L. Breiland, L. Oesper, and L. Taalman, “ $p$ -coloring classes of torus knots”, *Missouri J. Math. Sci.* **21**:2 (2009), 120–126. [MR 2010f:57011](#) [Zbl 1175.57011](#)
- [Brownell et al. 2006] K. Brownell, K. O’Neil, and L. Taalman, “Counting  $m$ -coloring classes of knots and links”, *Pi Mu Epsilon Journal* **12**:5 (2006), 265–278.
- [Champanerkar et al. 2004] A. Champanerkar, I. Kofman, and E. Patterson, “The next simplest hyperbolic knots”, *J. Knot Theory Ramifications* **13**:7 (2004), 965–987. [MR 2005k:57010](#) [Zbl 1064.57003](#)
- [Dean 1996] J. C. Dean, *Hyperbolic knots with small Seifert-fibered Dehn surgeries*, Ph.D. thesis, University of Texas at Austin, 1996.
- [Kauffman and Lopes 2009] L. Kauffman and P. Lopes, “Determinants of rational knots”, *Discrete Math. Theor. Comput. Sci.* **11**:2 (2009), 111–122. [MR 2010j:57007](#) [Zbl 1207.57020](#)
- [Livingston 1993] C. Livingston, *Knot theory*, Carus Mathematical Monographs **24**, Mathematical Association of America, Washington, DC, 1993. [MR 94m:57021](#) [Zbl 0887.57008](#)
- [Oesper 2005] L. Oesper,  *$p$ -colorings of weaving knots*, undergraduate thesis, Pomona College, 2005.
- [Przytycki 1998] J. H. Przytycki, “3-coloring and other elementary invariants of knots”, pp. 275–295 in *Knot theory* (Warsaw, 1995), Banach Center Publ. **42**, Polish Acad. Sci., Warsaw, 1998. [MR 1634462](#) [Zbl 0904.57002](#)

Received: 2009-11-10

Revised: 2013-06-27

Accepted: 2013-11-17

[mtdelong@taylor.edu](mailto:mtdelong@taylor.edu)

*Department of Mathematics, Taylor University,  
236 West Reade Avenue, Upland, IN 46989, United States*

[russell2@math.rutgers.edu](mailto:russell2@math.rutgers.edu)

*Department of Mathematics,  
Rutgers, The State University of New Jersey,  
110 Frelinghuysen Road, Piscataway, NJ 08854, United States*

[schrockj@ornl.gov](mailto:schrockj@ornl.gov)

*Oak Ridge National Laboratory, P.O. Box 2008 MS6164,  
Oak Ridge, TN 37831, United States*

# Parameter identification and sensitivity analysis to a thermal diffusivity inverse problem

Brian Leventhal, Xiaojing Fu, Kathleen Fowler and Owen Eslinger

(Communicated by Suzanne Lenhart)

The solution to inverse problems is an application shared by mathematicians, scientists, and engineers. For this work, a set of shallow soil temperatures measured at eight depths between 0 and 30 cm and sampled every five minutes over 24 hours is used to determine the diffusivity of the soil. Thermal diffusivity is a modeling parameter that impacts how heat flows through soil. In particular, it is not known in advance if the subsurface region is homogeneous or heterogeneous, which means the thermal diffusivity may or may not depend on depth. To this end, it is not clear which assumptions may apply to represent the physical system embedded within the parameter estimation problem. Analytic methods and a simulation based least-squares approach to approximate the diffusivity are compared to fit the temperature profiles to different heat flow models. The simulation is based on a spatially dependent, nonsteady-state discretization to a partial differential equation. To complete the work, a statistical sensitivity study using analysis of variance is used to understand how errors that arise in the modeling phase impact the final model. We show that for the analytic methods, errors in the initial fitting of the temperature data to sinusoidal boundary conditions can have a strong impact on the thermal diffusivity values. Our proposed framework shows that this soil sample is heterogeneous and that modeling depends significantly on data used as top and bottom boundary conditions. This work offers a protocol to determine the soil type and model sensitivities using analytic, numerical, and statistical approaches and is applicable to modifications of the classic heat equation found across disciplines.

## 1. Introduction

Inverse problems arise routinely across science and engineering disciplines. Using a mathematical approach to such parameter estimation problems avoids the tedious task of trial-and-error to match a mathematical model to experimental data. For this work, we consider a heat transport model in the shallow subsurface and use both

---

*MSC2010:* primary 35K05, 49N45, 62J10; secondary 35Q93.

*Keywords:* inverse problems, subsurface flow, sensitivity analysis.

analytic and numerical approaches to fit data. Part of the challenge is that the nature of the subsurface is not known in advance; thus it is not clear which model applies or whether assumptions made to apply analytic models are reasonable. In applying numerical approaches, assumptions on the types of boundary conditions can significantly impact the results. In the presence of such uncertainty and the possible addition of experimental error, the identified parameters may give suboptimal fits or provide values far from truth. This work offers a protocol to determine the soil type and model sensitivities using analytic, numerical, and statistical approaches by comparing common approaches to heat flow in the shallow subsurface and studying how choices made during the modeling phase can impact the results of the inverse problem.

The propagation of heat in the subsurface can be modeled by the second-order partial differential equation

$$\frac{\partial T}{\partial t} = K \frac{\partial^2 T}{\partial z^2}, \quad (1)$$

where  $T(z, t)$  is the time-dependent temperature distribution at depth  $z > 0$  for  $t > 0$ . Thermal diffusivity,  $K$  cm<sup>2</sup>/min, which describes how easily heat propagates through the medium, is proportionally related to thermal conductivity such that

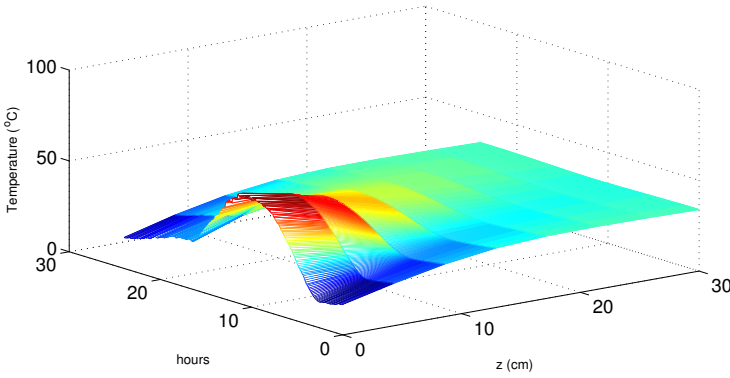
$$K = \frac{\hat{k}}{\rho c}, \quad (2)$$

where  $\hat{k}$  is the thermal conductivity,  $\rho$  is the density and  $c$  is the heat capacity. Although in (1),  $K$  is often assumed to be constant in practice, due to the complex nature of the subsurface,  $K$  is usually spatially dependent. We refer to these as homogeneous and heterogeneous soils respectively. Heat flow in the heterogeneous case would be described by

$$\frac{\partial T}{\partial t} = \frac{\partial}{\partial z} \left[ K \frac{\partial T}{\partial z} \right] = K \frac{\partial^2 T}{\partial z^2} + \frac{\partial K}{\partial z} \frac{\partial T}{\partial z}, \quad (3)$$

where now  $K = K(z)$ . Analytic solutions to various forms of these models exist [Carslaw and Jaeger 1986; Powers 2006; Narasimhan 2009] and have been studied for decades. Alternatively, given the spatially distributed thermal conductivity along with initial and boundary temperature, the temperature distribution over time can be approximated numerically.

The inverse of this problem is the focus of this study. Mathematical approaches can be used to help guide practitioners on the nature of the subsurface since it is not known in advance how much the soil type actually varies. Specifically, subsurface temperature data monitored at seven depths between 0 and 30 cm and logged over time were used to determine the thermal diffusivity of the test site. Figure 1 below



**Figure 1.** Temperature data.

shows temperature data as a function of time and depth. Analytic methods for determining  $K$  from temperature profiles have been proposed in the literature and implemented using data taken from the Loess Plateau in China [Gao et al. 2009]. Some of those methods are compared here but are based on the assumption that the soil is homogeneous. We compare these approaches to a simulation-based approach using a numerical approximation to the heterogeneous model in (3) and a minimization of the least-squares error between the model output and the temperature data. Since all of these methods include choices made during the modeling phase, we conduct a sensitivity study to understand how these choices impact the final model. The sensitivity study is based on a statistical analysis of variance.

We proceed by describing the methods used to determine the thermal diffusivity, both analytically and numerically, and then presenting those results in Sections 2 and 3. We follow with the sensitivity analysis in Section 4 and point the way towards future work in Section 5.

## 2. Analytic approaches

We consider four methods that approximate  $K$  values explicitly using temperature values at different depths. The methods are based on the homogeneous model in (1).

If we consider boundary conditions of the form

$$T(0, t) = T_a + A \sin(\omega t + \phi) \quad (4)$$

and

$$\lim_{z \rightarrow \infty} T(z, t) = T_a, \quad (5)$$

an analytic solution to (1) is given by

$$T(z, t) = T_a + A e^{-z/D} \sin\left(\omega t - \frac{z}{D} + \phi\right), \quad (6)$$

with  $D = \sqrt{2K/\omega}$ . Here, (4) states that the surface temperature varies as a sinusoidal function whose parameters include the time-average temperature  $T_a$  ( $^{\circ}\text{C}$ ), amplitude  $A$  ( $^{\circ}\text{C}$ ), radial frequency  $\omega$  ( $\text{rad s}^{-1}$ ) and phase constant  $\phi$  (rad). The bottom boundary condition (5) indicates that as depth increases sufficiently, the soil temperature is not affected by the surface temperature and thus maintains a constant value.

Four analytic methods were used to approximate the thermal conductivities at seven locations. The seven locations are between different depths, i.e., between 0 and 1 cm, between 1 and 5 cm, between 5 and 10 cm, and continuing until 30 cm deep. The methods described in [Gao et al. 2009; Horton et al. 1983] call for a homogeneous soil thermal conductivity profile. With thermal conductivity assumed homogeneous, the analytic methods call for only two depths to estimate the conductivity. To perform the analytic methods, the raw data temperatures need to be approximated by a sinusoidal curve of the form

$$T_1(z_1, t) = \bar{T}_1 + A_1 \sin(\omega t + \phi_1), \quad (7)$$

$$T_2(z_2, t) = \bar{T}_2 + A_2 \sin(\omega t + \phi_2), \quad (8)$$

where  $A_1, A_2$  are half of the difference between the daytime maximum and nighttime minimum amplitudes for the soil depths. Furthermore,  $\bar{T}_1, \bar{T}_2$  are the arithmetic averages of the daytime maximum soil temperature and the nighttime minimum soil temperature at depths  $z_1, z_2$ . The initial phases of the soil temperature,  $\phi_1$  and  $\phi_2$ , are obtained using a least-squares fit (as opposed to using a spline to fit the data) because numerical values for those parameters are needed in the analytic models to determine the conductivity. The resulting least-squares problem is nonlinear and a variety of optimization methods would apply. Since a genetic algorithm [Holland 1973] was being used in the project elsewhere, it was used here as well. Genetic algorithms require no gradient information for minimization and are thus attractive choices for an off-the-shelf optimization approach.

Since sinusoidal approximation is only needed at two depths for the analytic models, the two depths whose sinusoidal curves give the least error compared to the raw error are used to compute the thermal conductivity. With each producing a residual of  $10^{-1}$ , the data located one centimeter and five centimeters deep were used. Tables 1 and 2 show the results of the fit curve for each of the seven days of data. Table 3 shows the top boundary condition sinusoidal parameters as well.

The four methods considered for this experiment are the amplitude method, the phase method, the arctangent method and the logarithmic method. Essentially, if we assume that  $K$  is independent of depth (i.e., the media is homogeneous) and that the boundary temperature is sinusoidal, then the analytic solution of the one dimensional heat equation can be used to approximate  $K$ . The amplitude and phase methods are directly based on the analytic solution above. The arctangent and

day	$A_1$	$\omega$	$\phi_1$	$\bar{T}_1$
1	$1.45 \cdot 10^1$	$6.51 \cdot 10^{-3}$	3.13	$3.98 \cdot 10^1$
2	$1.54 \cdot 10^1$	$5.58 \cdot 10^{-3}$	5.58	$3.99 \cdot 10^1$
3	$1.45 \cdot 10^1$	$5.46 \cdot 10^{-3}$	5.46	$3.88 \cdot 10^1$
4	$1.38 \cdot 10^1$	$5.16 \cdot 10^{-3}$	5.16	$3.73 \cdot 10^1$
5	$1.58 \cdot 10^1$	$3.94 \cdot 10^{-3}$	3.94	$3.35 \cdot 10^1$
6	$1.64 \cdot 10^1$	$5.49 \cdot 10^{-3}$	5.49	$3.74 \cdot 10^1$
7	$1.73 \cdot 10^1$	$3.94 \cdot 10^{-3}$	2.44	$3.28 \cdot 10^1$

**Table 1.** Parameters obtained at a depth of 1 cm.

day	$A_2$	$\omega$	$\phi_2$	$\bar{T}_2$
1	$8.73 \cdot 10^1$	$5.75 \cdot 10^{-3}$	3.25	$3.75 \cdot 10^1$
2	$8.65 \cdot 10^1$	$5.69 \cdot 10^{-3}$	1.55	$3.80 \cdot 10^1$
3	$8.12 \cdot 10^1$	$5.11 \cdot 10^{-3}$	1.67	$3.69 \cdot 10^1$
4	$7.60 \cdot 10^1$	$5.02 \cdot 10^{-3}$	1.19	$3.62 \cdot 10^1$
5	$8.37 \cdot 10^1$	$5.60 \cdot 10^{-3}$	2.66	$3.66 \cdot 10^1$
6	$9.93 \cdot 10^1$	$3.87 \cdot 10^{-3}$	1.97	$3.31 \cdot 10^1$
7	$9.39 \cdot 10^1$	$4.51 \cdot 10^{-3}$	2.95	$3.47 \cdot 10^1$

**Table 2.** Parameters obtained at a depth of 5 cm.

day	amplitude	$\omega$	phase	$\bar{T}$
1	$1.74 \cdot 10^1$	$6.81 \cdot 10^{-3}$	3.00	$3.76 \cdot 10^1$
2	$2.14 \cdot 10^1$	$5.68 \cdot 10^{-3}$	2.33	$3.82 \cdot 10^1$
3	$1.96 \cdot 10^1$	$5.21 \cdot 10^{-3}$	2.18	$3.61 \cdot 10^1$
4	$1.84 \cdot 10^1$	$4.85 \cdot 10^{-3}$	2.81	$3.35 \cdot 10^1$
5	$2.05 \cdot 10^1$	$4.79 \cdot 10^{-3}$	2.40	$3.34 \cdot 10^1$
6	$2.28 \cdot 10^1$	$5.48 \cdot 10^{-3}$	2.59	$3.51 \cdot 10^1$
7	$2.26 \cdot 10^1$	$5.27 \cdot 10^{-3}$	2.92	$3.48 \cdot 10^1$

**Table 3.** Parameters obtained for the boundary condition.

logarithmic methods are based on the notion that a Fourier series can reduce errors introduced by the assumption that a single sinusoidal wave is sufficient to estimate the surface temperature. We state these approaches here and point the reader to [Gao et al. 2009; Horton et al. 1983] for more details.

**The amplitude method:**

$$K = \frac{\omega(z_1 - z_2)^2}{2 \ln(A_1/A_2)^2}. \quad (9)$$

**The phase method:**

$$K = \frac{\omega(z_1 - z_2)^2}{2(\phi_1 - \phi_2)^2}. \quad (10)$$

**The arctangent method:** This method is based on the notion that soil temperature can be described by a Fourier series,

$$T = \bar{T} + \sum_{i=1}^n (a_i \sin(i\omega t) + b_i \cos(i\omega t)).$$

With  $n = 2$ ,  $K$  can be estimated with

$$K = \left( \frac{\omega \Delta z^2}{2 \arctan \frac{(T_1 - T_3)(T'_2 - T'_4) - (T_2 - T_4)(T'_1 - T'_3)}{(T_1 - T_3)(T'_1 - T'_3) + (T_2 - T_4)(T'_2 - T'_4)}} \right)^2, \quad (11)$$

where temperatures  $T_j$  and  $T'_j$  are recorded at 6 hour time intervals and two different depths  $z_1, z_2$ .

**The logarithmic method:** Using the same assumptions as the arctangent method,  $K$  can be expressed as

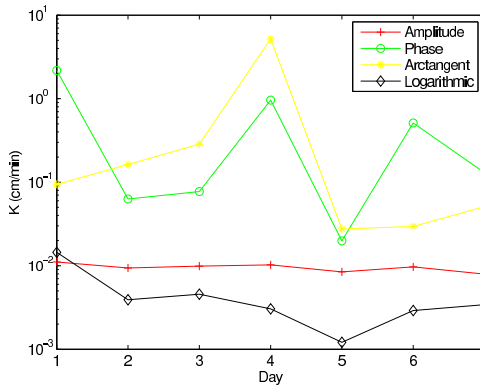
$$K = \left( \frac{0.012 \Delta z}{\ln \frac{(T_1 - T_3)^2 + (T_2 - T_4)^2}{(T'_1 - T'_3)^2 + (T'_2 - T'_4)^2}} \right)^2. \quad (12)$$

Table 4 shows the results for each method for the seven days studied. As can be observed, the amplitude method and logarithmic method estimate the thermal conductivity on the same order of magnitude over the seven days. The other two methods, phase and arctangent, estimate the thermal conductivities with significant variability. They do not hold the order of magnitude constant over the seven days, thus producing significantly different results from other methods.

day	amplitude	phase	arctangent	logarithm
1	$1.11 \cdot 10^{-2}$	2.18	$9.43 \cdot 10^{-2}$	$1.45 \cdot 10^{-2}$
2	$9.39 \cdot 10^{-3}$	$6.29 \cdot 10^{-2}$	$1.63 \cdot 10^{-1}$	$3.92 \cdot 10^{-3}$
3	$9.91 \cdot 10^{-3}$	$7.73 \cdot 10^{-2}$	$2.86 \cdot 10^{-1}$	$4.56 \cdot 10^{-3}$
4	$1.02 \cdot 10^{-3}$	$9.62 \cdot 10^{-1}$	5.19	$3.04 \cdot 10^{-3}$
5	$8.43 \cdot 10^{-3}$	$1.98 \cdot 10^{-2}$	$2.75 \cdot 10^{-2}$	$1.21 \cdot 10^{-3}$
6	$9.69 \cdot 10^{-3}$	$5.12 \cdot 10^{-1}$	$2.95 \cdot 10^{-2}$	$2.91 \cdot 10^{-3}$
7	$7.93 \cdot 10^{-3}$	$1.30 \cdot 10^{-1}$	$5.09 \cdot 10^{-2}$	$3.42 \cdot 10^{-3}$

**Table 4.** Estimated conductivities (cm/min).





**Figure 2.** Estimated conductivities.

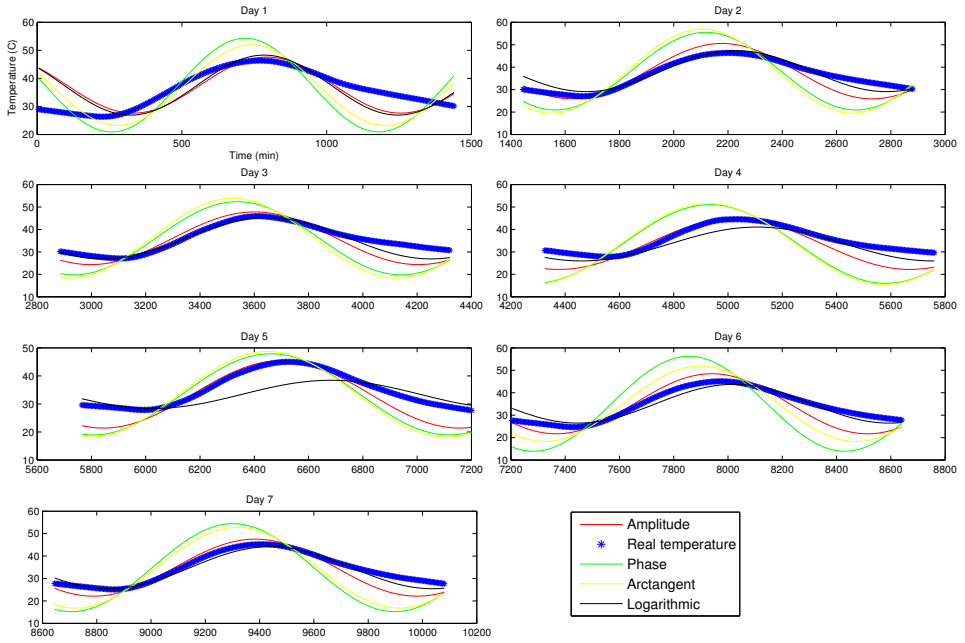
The thermal conductivity changes with days shown by Figure 2. As there are noteworthy differences between each method, the arctangent and phase methods seem to have the most significant change.

The thermal conductivities found can be used to determine temperature profiles for each method. As seen in Figure 3, not one method provides an accurate estimation of the data profile. Large errors in the original sinusoidal fitting or the inaccuracy of assuming thermal conductivity homogeneity could account for this difference in estimation. To this end, although attractive for their simplicity, the analytic methods do not provide an accurate approximation to the data.

The assumption that thermal conductivity is homogeneous throughout the soil may be inaccurate. Instead of assuming homogeneity from 0 to 30 cm, homogeneity can be assumed on small subintervals. This assumption is reasonable if the porous media is layered so that it is homogeneous in  $x - y$  directions and constant on layered intervals in the  $z$  direction. These results are shown in Table 5. By displaying a change in thermal conductivity with depth in Figure 5, results either confirm that homogeneity was an inaccurate assumption or that errors from the initial fitting are having an impact on the final values.

depth (cm)	amplitude	phase	arctangent	logarithmic
1–5	$8.50 \cdot 10^{-3}$	$7.20 \cdot 10^{-2}$	$2.52 \cdot 10^{-1}$	$2.90 \cdot 10^{-3}$
5–10	$1.70 \cdot 10^{-2}$	$3.70 \cdot 10^{-2}$	$2.40 \cdot 10^{-2}$	$1.40 \cdot 10^{-3}$
10–15	$7.70 \cdot 10^{-3}$	$3.17 \cdot 10^{-2}$	$8.73 \cdot 10^{-2}$	$3.20 \cdot 10^{-3}$
15–20	$3.58 \cdot 10^{-2}$	$8.60 \cdot 10^{-1}$	$1.64 \cdot 10^{-2}$	$5.79 \cdot 10^{-4}$
20–25	$6.19 \cdot 10^{-2}$	$6.03 \cdot 10^{-2}$	$8.09 \cdot 10^{-1}$	$3.88 \cdot 10^{-5}$
25–30	$2.05 \cdot 10^{-2}$	$2.88 \cdot 10^{-2}$	$4.90 \cdot 10^{-2}$	$4.26 \cdot 10^{-3}$

**Table 5.** Differences in conductivities at each depth on day 7.

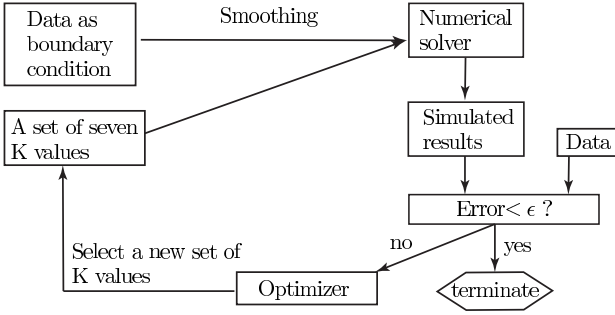


**Figure 3.** Temperature distributions.

We will show later that small variations in fitting the sinusoidal curve impact the analytic solutions greatly. Moreover, with regard to the inaccurate assumption of homogeneity, a new approach is taken below. We proceed by analyzing the simulation-based optimization approach to conduct the experiment with the assumption that thermal conductivities are heterogeneous.

### 3. Simulation-based approach

In the first approach, thermal conductivities are calculated using four analytic methods. However, the results indicate that the assumption of homogeneity may not be valid. To this end, an optimization framework where the least-squares error (LSE) between data and a simulated temperature profile facilitates the incorporation of spatially varying thermal conductivities. For the simulation, finite differences were used to discretize (3) in space with backward Euler in time. To validate the simulation tool, results were compared to a problem with a known solution using a forcing term  $f(z, t)$  on the right hand side and a known function  $K(z)$  to ensure accurate truncation error. To account for the fact that data would be used in the subsequent study for  $K$ , we use a spline to describe the variation of  $K$  in space and then differentiate it to obtain  $\partial K / \partial z$ .



**Figure 4.** Structure of the optimization scheme.

In this new context, the logged data  $T^{\text{obs}}$  is an  $N_t \times N_z$  matrix, where  $N_t$  is the number of time points and  $N_z$  is the number of spatial nodes. The least-squares problem is then

$$\min_{K \in \Omega} J(K) = \frac{\frac{1}{2} \sum_{i=1}^N (\hat{T}_i(K) - T_i^{\text{obs}})^2}{\frac{1}{2} \sum_{i=1}^N (T_i^{\text{obs}})^2}, \quad (13)$$

where  $\Omega$  represents reasonable bound constraints on  $K$ . The simulated temperature profile  $\hat{T}(K)$  is obtained by numerically solving the heat equation. Since the temperature at the surface has been recorded from the meteorology station, a Dirichlet boundary condition can be easily incorporated.

Because evaluation of the objective function in (13) requires output from a simulation tool, we use sampling methods for the optimization since gradient information is not available. To proceed, we use the same genetic algorithm that was used to fit the sinusoidal boundary condition from the above study. The optimization framework is displayed in Figure 4. At each iteration, the optimizer will pick a set of six  $K$  values based on the bounds  $\Omega$  and previous function evaluations. This vector of  $K$  values is then used as input for the numerical solver for the heat equation that outputs the temperature profile. The simulated profile is compared to actual data to obtain the error at the current iteration. The optimization terminates when the error becomes sufficiently small, resulting with the current set of  $K$  values as a potential optimal solution.

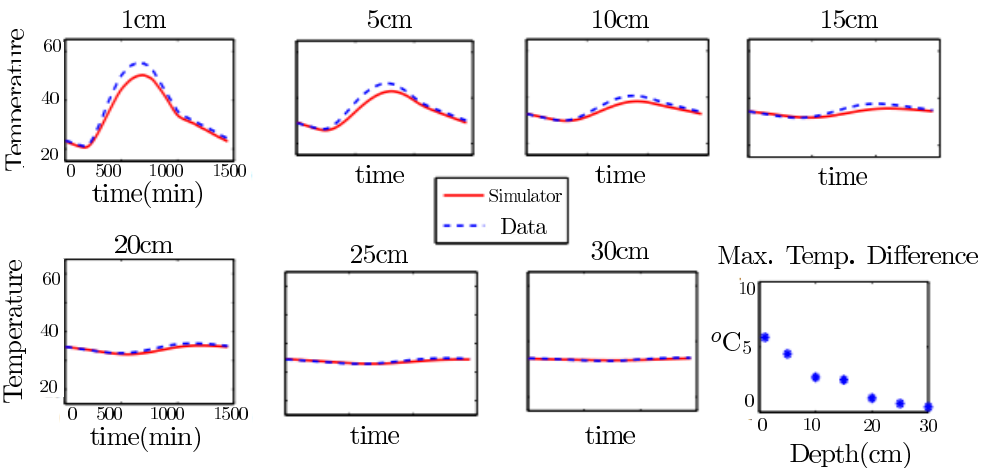
With the simulation tool in place, we fit the temperature profile in each layer at 24 hours by optimizing conductivities at 1, 5, 10, 15, 20, 25, and 30 cm. As a first attempt, we assumed that the conductivity varies linearly between these locations and was constant from 30 cm to the location of the bottom boundary condition and between the top of the domain and 1 cm. To this end, since the mean subsurface temperature is not known, we also include the temperature at the bottom depth as a decision variable. We used a depth of 70 cm to enforce the bottom boundary condition and used  $\Delta z = 0.1$  cm and  $\Delta t = 0.1$  minutes. The temperature data used

depth	$K$ (cm/min)
1 cm	$3.2809 \cdot 10^{-2}$
5 cm	$3.1440 \cdot 10^{-1}$
10 cm	$2.6486 \cdot 10^{-2}$
15 cm	$3.9509 \cdot 10^{-1}$
20 cm	$1.9201 \cdot 10^{-1}$
25 cm	$9.8476 \cdot 10^{-3}$
30 cm	$2.6704 \cdot 10^{-1}$

**Table 6.** Preliminary optimization results, for a temperature of  $35.3^{\circ}\text{C}$ ,  $\text{LSE} = 5.9633 \cdot 10^{-5}$  and  $E = 0.895^{\circ}\text{C}$ .

over space and time is given above in Figure 1. The temperatures range from about 13 to  $67^{\circ}\text{C}$  in the first 24 hours. Table 6 shows the optimal values obtained for each depth at the 24th hour. The last two rows show the least-squares error (LSE) and the maximum temperature difference (E) over each depth.

The results are promising and the temperature fit can be seen in Figure 5. The maximum error across all depths over time is only  $6.2^{\circ}\text{C}$ , which is a significantly lower than the corresponding first day results in Figure 3. These results confirm that the data likely corresponds to heterogeneous soil. However, in general, this is not known in advance. Thus, it is important to understand the strengths and weaknesses of all methods applied here. To this end, the sensitivity study presented in the next section quantifies how errors in these modeling components impact the overall quality of the inverse problem solution.



**Figure 5.** Comparison of simulated temperatures and data at each sensor location over time and maximum temperature difference.

## 4. Sensitivity analysis

Analysis of variance (ANOVA) is a way to determine whether model parameters have an effect on the model output by comparing the ratio of the variation between sample means to the variation within each sample. For this study, we consider how the parameters in both analytic and simulation-based approaches impact the estimation of  $K$  and the model fit. The starting point for the procedure is to sort each parameter into groups. Analysis is done by considering changes in a response as the group changes. Specifically, ANOVA is a hypothesis test with null hypothesis  $H_0 = \mu_1 = \mu_2 = \dots = \mu_k$ , where  $k$  is the number of experimental groups. Each  $\mu$  represents the mean of the single parameter, often called a factor, that is being found by the values in each experimental group. When rejecting the null hypothesis, the alternative hypothesis states that at least one mean is different from another; however, it does not specify which one. The experimental groups are different equally spaced intervals for a single variable. The ANOVA examines the source of variation by finding the sum of squares of deviation from the mean for each of these groups. Using a statistical F-test, the procedure is able to determine whether or not at least one mean is deviating from the others. The F-test will produce a  $p$ -value; if this value is below a significance of 0.05 then the null hypothesis is rejected. If the significance is above 0.05, the null hypothesis is failed to be rejected. For this work, we seek to understand the sensitivity of parameters for both the analytic and the numerical approaches to matching the temperature data.

**4.1. Sensitivity analysis of analytic methods.** Even if a soil sample is homogeneous, there could be errors within the initial sinusoidal fitting of the data due to experimental noise. A sensitivity study can be used to understand how errors in this fitting will impact the resulting temperature profile, in particular, if we consider a hypothetical problem with known model parameters. In other words we sampled variations of the parameters in (7) and (8) and determined how they impacted the ability to identify the conductivity. Specifically, we varied  $A_1$ ,  $A_2$ ,  $\bar{T}_1$ ,  $\bar{T}_2$ ,  $\phi_1$  and  $\phi_2$  and compared the calculated  $K$  to the known value. Using a Latin hypercube sampling (LHS) approach to assure a uniform distribution of selections with intervals surrounding the true values, we considered 1,600 values of each parameter. The bounds used for the LHS sampling are displayed in Table 7 as well as the true parameter value. Following the sampling, the parameters were grouped and an analysis of variance (ANOVA) was performed to show how errors in the initial least-squares fit impact the thermal conductivities from the four analytic methods.

We consider one response for each of the four analytic methods to determine  $K$ . These are found by taking the difference between the true conductivity and the conductivity found using the perturbed parameter values. Values for each of the independent parameters were grouped into eight subsets determined by equal sized

parameter	lower bound	upper bound
$A_1 = 5.5974^\circ\text{C}$	$3^\circ\text{C}$	$7^\circ\text{C}$
$A_2 = 2.2885^\circ\text{C}$	$1^\circ\text{C}$	$5^\circ\text{C}$
$\bar{T}_1 = 20^\circ\text{C}$	$18^\circ\text{C}$	$22^\circ\text{C}$
$\bar{T}_2 = 20^\circ\text{C}$	$18^\circ\text{C}$	$22^\circ\text{C}$
$\phi_1 = 0.776$	-1	1
$\phi_2 = -0.1880$	-1	1

**Table 7.** True parameter values and LHS bounds.

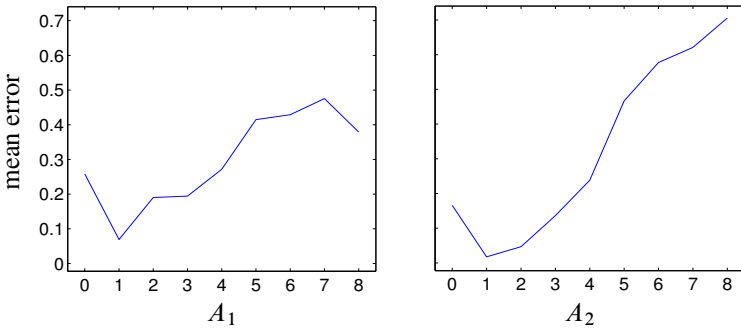
ranges within the lower and upper bound of the parameter. ANOVA compares the variance of the objective function within each group to that same variance between the groups. If this ratio is sufficiently small, then the objective function is sensitive to changes in that parameter. This test provides a  $p$ -value that establishes a confidence level for sensitivity.

ANOVA results are easily visualized through main effect plots, one developed for each parameter analyzed. Large changes in dependent variable values within each plot show the method is sensitive to changes of that independent parameter. In other words, a flat line means little sensitivity to variation of the parameter value. The vertical axis shows the mean value of the response for values of the parameter of that specific group. A  $p$ -value is found to numerically measure the sensitivity, with a  $p$ -value close to zero indicating that the parameter is sensitive. The main effects results of the analysis of variance for the amplitude, phase, arctangent and logarithmic methods are shown in Figures 6–9.

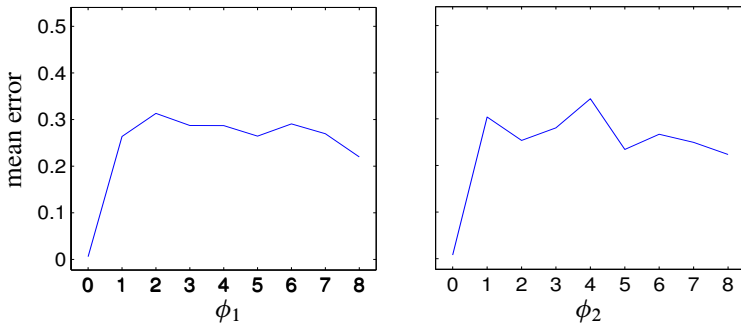
As seen, all methods are most sensitive to variations of the amplitude parameter. Thus, errors in estimating the amplitude result in large changes in thermal conductivity results from the four analytic methods. It appears that variations of the other parameters have an impact but are not nearly as significant as variations within the amplitude.

**4.2. Sensitivity analysis of a heterogeneous system.** The simulation-based approach uses an optimization algorithm to determine a temperature profile. This technique calls for variation with the bottom boundary condition and seven thermal conductivities. A similar study using ANOVA is conducted to understand the impact of each of these parameters on the model fit by considering the LSE as the output. As with the analytic results, a Latin hypercube sampling is used to sample all parameters. The bounds for the LHS were  $[20, 45]$  degrees Celsius for the bottom temperature and  $[10^{-4}, 10^{-1}]$  for each thermal conductivity.

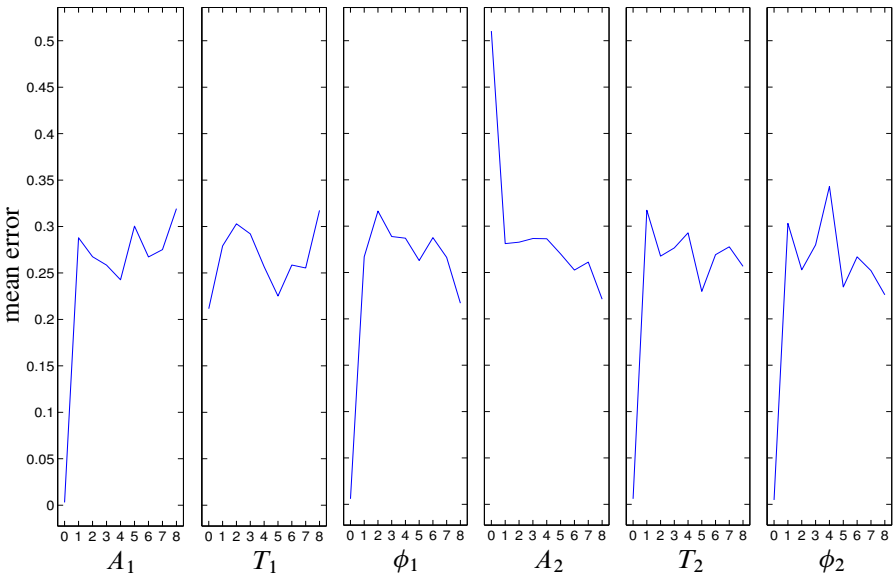
Parameters are considered to be sensitive if their corresponding  $p$ -value is less than 0.05;  $p$ -values are given in Table 8. ANOVA results are displayed visually



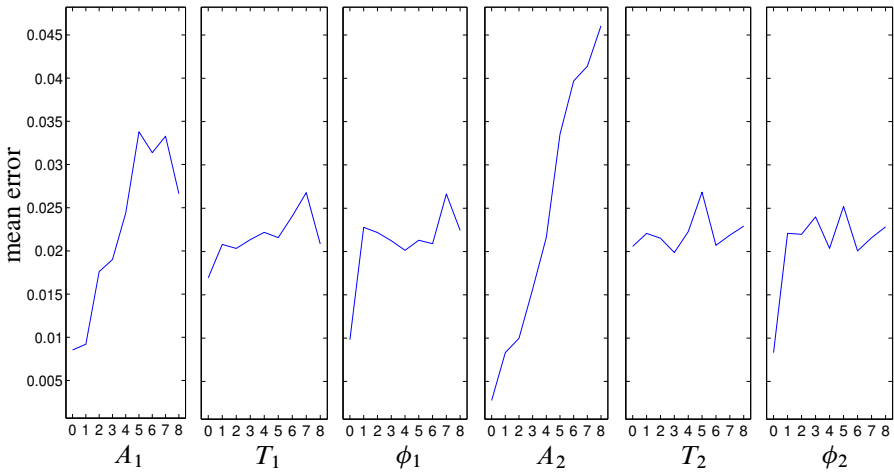
**Figure 6.** Amplitude method: both  $A_1, A_2$  result in  $p \approx 0$ .



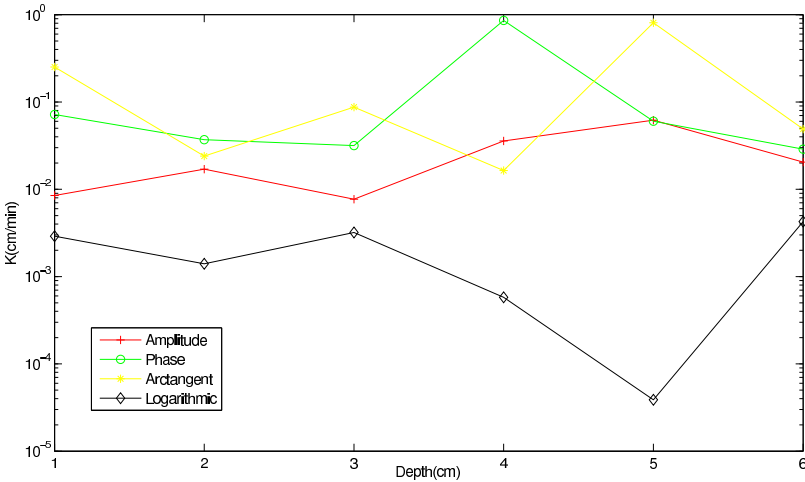
**Figure 7.** Phase method: both  $p$ -values  $> 0.05$ .



**Figure 8.** Arctangent method: all  $p$ -values  $> 0.05$ .



**Figure 9.** Logarithmic method: both  $A_1, A_2$  result in  $p \approx 0$ .



**Figure 10.** Day 7: conductivities at different depths.

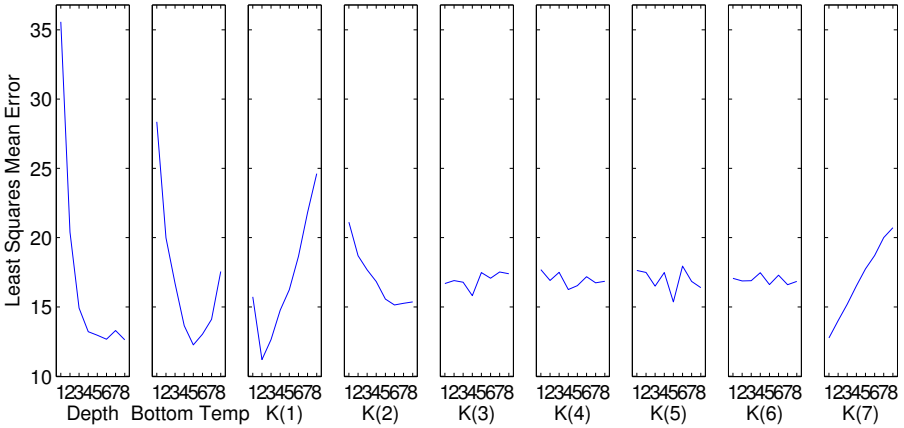
using main effects plots shown in [Figure 11](#). The horizontal axis shows the number of the group, where the intervals above were split into eight equal subintervals. The vertical axis is the average least-squares value corresponding to each group.

From these results, we can see that several of the conductivities are sensitive. Related work analyzes the impact of errors in the boundary and initial data on solution to the inverse problem [[Fu and Leventhal 2011](#)]. Here we find that our solutions are not sensitive to this boundary condition. Often in practice, ANOVA is done in advance to understand which model parameters should be included in the optimization and thereby reduce the size of the design space. In this context, the analysis can be used to weight those sensors more heavily in a subsequent optimization study.



parameter	<i>p</i> -value
depth	0
bot. temp (°C)	$6.22207 \cdot 10^{-1}$
$K_1$	$3.4870 \cdot 10^{-5}$
$K_5$	$1.4413 \cdot 10^{-1}$
$K_{10}$	$1.2058 \cdot 10^{-1}$
$K_{15}$	$2.6220 \cdot 10^{-2}$
$K_{20}$	$1.3701 \cdot 10^{-7}$
$K_{25}$	$5.2570 \cdot 10^{-1}$
$K_{30}$	0

**Table 8.** ANOVA results.



**Figure 11.** Main effects plots for simulation-based approach.

### 5. Conclusion

In this work, we have considered an inverse problem to determine the thermal conductivities for a heat transport model using temperature data in the shallow subsurface. Since it is not known in advance if the soil is homogeneous or heterogeneous, analytic and numerical approaches were used. Furthermore, sensitivity analyses can be paired with optimization and modeling problems to help understand how choices made during the solution procedure impact the quality of the results. These ideas provide a protocol for approaching these types of problems.

In this study, not one of the analytics methods for estimating thermal conductivity fit the temperature profile within the given degree of desired accuracy. Each parameter is significant in each method with amplitude being the most significant parameter. Thus small deviations in the amplitude cause large deviations within the

resulting thermal conductivity. The amplitude and logarithmic methods display the general trend of temperature values however still not within the desired error. The numerical approach gave satisfactory results and a significantly smaller error than the analytic methods, indicating that this data corresponds to heterogeneous soil.

## References

- [Carslaw and Jaeger 1986] H. S. Carslaw and J. C. Jaeger, *Conduction of heat in solids*, 2nd ed., Clarendon, Oxford, 1986. [Zbl 0584.73001](#)
- [Fu and Leventhal 2011] X. Fu and B. Leventhal, “Understanding the impact of boundary and initial condition errors on the solution to a thermal diffusivity inverse problem”, *SIAM Undergrad. Res. Online* **4** (2011), 156–174.
- [Gao et al. 2009] Z. Gao, L. Wang, and R. Horton, “A comparison of six algorithms to determine the soil thermal diffusivity at a site in the Loess Plateau of China”, *Hydrol. Earth Syst. Sci.* **6** (2009), 2247–2274.
- [Holland 1973] J. H. Holland, “Genetic algorithms and the optimal allocation of trials”, *SIAM J. Comput.* **2** (1973), 88–105. [MR 52 #12445a](#) [Zbl 0259.90031](#)
- [Horton et al. 1983] R. Horton, P. J. Wierenga, and D. R. Nielsen, “Evaluation of methods for determining the apparent thermal diffusivity of soil near the surface”, *Soil Sci. Soc. Am.* **47**:1 (1983), 25–32.
- [Narasimhan 2009] T. N. Narasimhan, “The dichotomous history of diffusion”, *Phys. Today* **62**:7 (2009), 48–53.
- [Powers 2006] D. L. Powers, *Boundary value problems and partial differential equations*, 5th ed., Elsevier, Amsterdam, 2006. Chapter 2. [MR 2008a:35001](#) [Zbl 1107.35001](#)

Received: 2011-04-28

Revised: 2013-11-11

Accepted: 2013-12-20

[leventbc@gmail.com](mailto:leventbc@gmail.com)

*Department of Psychology in Education,  
University of Pittsburgh, 5930 Wesley W. Posvar Hall,  
Pittsburgh, PA 15260, United States*

[rubyfu@mit.edu](mailto:rubyfu@mit.edu)

*Department of Civil and Environmental Engineering,  
Massachusetts Institute of Technology, 77 Massachusetts  
Avenue, Cambridge, MA 02139, United States*

[kfowler@clarkson.edu](mailto:kfowler@clarkson.edu)

*Department of Mathematics, Clarkson University,  
8 Clarkson Avenue, Potsdam, NY 13699, United States*

[owen.j.eslinger@usace.army.mil](mailto:owen.j.eslinger@usace.army.mil)

*US Army ERDC, 3909 Halls Ferry Road,  
Vicksburg, MS 39180, United States*

# A mathematical model for the emergence of HIV drug resistance during periodic bang-bang type antiretroviral treatment

Nicoleta Tarfulea and Paul Read

(Communicated by Suzanne Lenhart)

In treating HIV infection, strict adherence to drug therapy is crucial in maintaining a low viral load, but the high dosages required for this often have toxic side effects which make perfect adherence to antiretroviral therapy (ART) unsustainable. Moreover, even in the presence of drug therapy, ongoing viral replication can lead to the emergence of drug-resistant virus variants. We introduce a mathematical model that incorporates two viral strains, wild-type and drug-resistant, to theoretically and numerically investigate HIV pathogenesis during ART. A periodic model of bang-bang type is employed to estimate the drug efficacies. Furthermore, we numerically investigate the antiviral response and we characterize successful drugs or drug combination scenarios for both strains of the virus.

## 1. Introduction

Over the last few decades, the rapid spread of the human immunodeficiency virus (HIV) and the death toll of acquired immunodeficiency syndrome (AIDS) have motivated a great deal of scientific and medical research. Treatment of the HIV infection has traditionally consisted of antiretroviral therapy (ART), a regimen of pharmaceutical treatments that often produces unwanted physical side effects and can become costly over long periods of time. Moreover, strict adherence to drug therapy is crucial in maintaining a low viral load, but the high dosages required for this often have toxic side effects which make perfect adherence to ART unsustainable. This in turn leads to the development of resistant strains [Kepler and Perelson 1998; Kirschner and Webb 1997; Murray and Perelson 2005; Ribeiro et al. 1998]. Since its discovery in 1984, much research has been done and researchers have increased their understanding of the virus, and consequently drugs have been

---

*MSC2010:* primary 92D30; secondary 92B05, 34A34.

*Keywords:* HIV dynamics, time-varying antiretroviral treatment, drug resistance.

Read was supported by the US Department of Energy-Northwest Indiana Computational Grid Grant and a CURM mini-grant funded by the NSF grant DMS-0636648.

successful in the treatment but not the cure of the disease. In the last decade, it has become more and more evident that mathematical models are extremely useful in understanding of various biological processes. They create a powerful and inexpensive virtual laboratory where one can test and experiment different competing hypotheses.

When HIV enters the bloodstream, it primarily targets crucial components of the immune system [Fauci 1993], specifically, CD4+ T-cells or helper T-cells, whose function is to assist the response to bodily infections by releasing chemicals that signal other immune system cells, such as CD8+ (killer) T-cells, to kill infected cells or infectious particles [Bofill et al. 1992; Cohen and Boyle 2004; Fauci 1993; McMichael Winter 1996; Wilson et al. 2000; NHS 2008]. HIV is capable of infecting other immune cells, such as macrophages [Perelson and Nelson 1999], but the primary targets of infection are the CD4+ T-cells [Koup et al. 1994]. Hence, they play a central role in existing mathematical models [Adams et al. 2005; Burg et al. 2009; Huang 2008; Perelson et al. 1993; Perelson and Nelson 1999; Rapin et al. 2006; Rong et al. 2007a; 2007b; Tarfulea et al. 2011; Tarfulea 2011b; 2011a]. However, the most significant and threatening problem that HIV presents is its ability to continuously mutate in the body and form resistances to otherwise useful drugs [Shiri et al. 2005; Smith and Wahl 2005; Wahl and Nowak 2000].

Building upon the model introduced in [Tarfulea 2011b], we include two distinct viral strains (drug-sensitive and drug-resistant) and time-varying antiretroviral treatment of bang-bang type. This mathematical model is described by a system of six differential equations and is used to analyze the efficacy of different drug combinations in tandem with the evolution of the resistant strain in each case. We use the Floquet multipliers to investigate the stability properties of the infection-free steady state. We obtain the expected monotonicity property, namely if the treatment is periodic of bang-bang type and it can clear the infection, then the infection is cleared more rapidly if the treatment is more efficient or lasts longer. The multiple viral strains that this new model incorporates brought forth a much more useful understanding to the conditions faced by the antiretroviral drugs and the components of the infected immune system. Furthermore, we investigate the consequences of different scenarios of antiviral therapy, as well as the influence of different combinations of the major classes of drugs available for the treatment. We also study their impact on the evolution of the disease and determine a possible optimal treatment strategy that will lower the total viral load in the body. Thus, our model could be used to suggest which drugs or combination of drugs are optimal for a given patient, as well as to investigate the consequences of changing the treatment frequency or imperfect adherence. The effect of periodic treatment that includes pharmacokinetics on a multistrain model and the effect of STIs is an ongoing investigation.

variable	description
$T$	healthy T-cell concentration
$T_s$	drug-sensitive infected T-cell concentration
$T_r$	drug-resistant infected T-cell concentration
$V_s$	drug-sensitive virus concentration
$V_r$	drug-resistant virus concentration
$E$	concentration of CD8+ T-cells

**Table 1.** Variables used in the differential equation systems.

## 2. Formulation of the problem

**2.1. The mathematical model for the pretreatment case.** We now present the mathematical model for the dynamics of HIV before treatment (see [Tarfulea 2011b]). Building upon it, we will introduce in Section 2.2 the mathematical model with time-varying drug efficacies of bang-bang type.

A widely adopted mathematical model of HIV infection consists of a system of differential equations describing the evolution of the concentrations of healthy CD4+ T-cells, infected CD4+ T-cells, and free viruses in the body (see [Adams et al. 2005; Perelson et al. 1993; Perelson and Nelson 1999; Rapin et al. 2006; Rong et al. 2007a; 2007b; Stafford et al. 2000]).

The course of HIV infection varies widely across the infected population, and this is at least partially explained by individually specific immunological responses. The primary effector of the cell-mediated immune response is the CD8+ killer T-cells (CTLs). The CD8+ T-cell kills infected cells bearing a specific antigen. The activation of the killer T-cells is largely dependent upon the CD4+ helper T-cells, which direct the immune response. Thus, incorporation of cellular compartments representing both the helper and effector T-cells more completely represents the body's cellular immune system. In [Tarfulea et al. 2011], the authors consider a model for HIV dynamics which includes the CTLs' response.

To model the emergence of drug resistance and a possible treatment method, a new model is required which accounts for the presence of drug-sensitive and drug-resistant strains of the virus separately, rather than aggregating them. In this manner, one could determine whether a certain treatment regimen was producing an increase in the drug-resistant concentration of the virus over time, even if the population of the drug-sensitive HIV virus was declining. Treatments which cause the population of the drug-sensitive virus to decline, but allow the population of the drug-resistant virus to increase over time are postponing the inevitable, as they do not provide a long-term benefit to an individual infected with HIV. A model

incorporating two strains of HIV has been utilized in [Rong et al. 2007a] to model the effects of antiretroviral therapy (ART) on the appearance of drug-resistant strains of HIV. In [Tarfulea 2011b], the author considers the following model for HIV dynamics which includes the CTLs' response:

$$\begin{aligned}
 \frac{dT}{dt} &= \lambda_T - Td - k_s V_s T - k_r V_r T, \\
 \frac{dT_s}{dt} &= (1-u)k_s T V_s - \delta T_s - m_1 E T_s, \\
 \frac{dV_s}{dt} &= N_s \delta T_s - c V_s, \\
 \frac{dT_r}{dt} &= uk_s T V_s + k_r V_r T - \delta T_r - m_2 E T_r, \\
 \frac{dV_r}{dt} &= N_r \delta T_r - c V_r, \\
 \frac{dE}{dt} &= \lambda_E + c_E (T_s + T_r) - \delta_E E,
 \end{aligned} \tag{1}$$

together with initial data

$$T(0) = T_0, \quad T_s(0) = 0, \quad V_s(0) = V_0, \quad T_r(0) = 0, \quad V_r(0) = 0, \quad E(0) = E_0, \tag{2}$$

where  $T_0, V_0, E_0 > 0$ . The variables used in system (1) are described in Table 1 and the parameters used and their values are described in Table 2. Here  $u$  represents the rate at which drug-sensitive T-cells mutate to become drug-resistant, and it applies only when the two strains of virus differ by a single point mutation. HIV replicates at a very high rate in untreated patients. Thus, there is a realistic chance that drug-resistant variants exist even before the initiation of therapy [Ribeiro et al. 1998; Rong et al. 2007a]. Moreover, since the wild-type virus dominates the population before the initiation of therapy (see [Bonhoeffer et al. 2000; Nowak et al. 1997]), the mutation from drug-resistant to drug-sensitive is neglected. Also, it is assumed in this model that  $c$ , the clearance rate, and  $\delta$ , the infected T-cell death rate, are the same for both strains of virus.

System (1) has three possible positive steady states:

(1) The infection-free steady state:

$$S_0 := \left( T_0 = \frac{\lambda_T}{d}, T_{s0} = 0, V_{s0} = 0, T_{r0} = 0, V_{r0} = 0, E_0 = \frac{\lambda_E}{\delta_E} \right). \tag{3}$$

(2) The boundary steady state  $S_b$ , when only the drug-resistant strain is present:

$$S_b := (T_b, T_{sb}, V_{sb}, T_{rb}, V_{rb}, E_b), \tag{4}$$

where

$$T_{sb} = 0, \quad V_{sb} = 0,$$

$$T_{rb} = \frac{c}{N_r \delta} \frac{\lambda_T - d T_b}{k_r T_b}, \quad V_{rb} = \frac{\lambda_T - d T_b}{k_r T_b}, \quad E_b = \frac{\lambda_E}{\delta_E} + \frac{c_E}{\delta_E} \frac{\lambda_T - d T_b}{k_r T_b},$$

and  $T_b$  is the positive solution of the quadratic equation  $T^2 - A_b T - B_b = 0$ , where

$$A_b = \frac{c}{N_r \delta k_r} \left( \delta + m_2 \frac{\lambda_E}{\delta_E} - m_2 d \frac{c}{N_r \delta k_r} \frac{c_E}{\delta_E} \right) \quad \text{and} \quad B_b = m_2 \left( \frac{c}{N_r \delta k_r} \right)^2 \frac{c_E}{\delta_E} \lambda_T.$$

parameter	description	value	reference
$\lambda_T$	Recruitment rate of uninfected cells	$d \cdot T(0)$	1
$d$	Death rate of uninfected cells	$0.01 \text{ day}^{-1}$	1, 2
$k_s$	Infection rate of T-cells by the wild-type virus	$2.4 \cdot 10^{-5} \mu\text{l day}^{-1}$	1, 3, 4
$k_r$	Infection rate of T-cells by the drug-resistant virus	$2.4 \cdot 10^{-5} \mu\text{l day}^{-1}$	1, 3, 4
$\delta$	Death rate of infected cells	$0.3 \text{ day}^{-1}$	5
$m_1$	Immune-induced clearance rate for infected $T_s$ cells	$10^{-2} \mu\text{l day}^{-1}$	3
$m_2$	Immune-induced clearance rate for infected $T_r$ cells	$10^{-2} \mu\text{l day}^{-1}$	3
$N_s$	Virions produced per infected drug-sensitive cell	5000	1
$N_r$	Virions produced per infected drug-resistant cell	5000	1
$c$	Clearance rate of free virus	$23 \text{ day}^{-1}$	1
$\lambda_E$	Immune effector production (source) rate	$10^{-3} \mu\text{l day}^{-1}$	3
$c_E$	Stimulation of CTL proliferation	$0.3 \text{ day}^{-1}$	5
$\delta_E$	Death rate of immune effectors	$0.1 \text{ day}^{-1}$	3, 5
$u$	Mutation rate from sensitive strain to resistant strain	$3 \cdot 10^{-5}$	1

**Table 2.** Parameter definitions and values used in numerical simulations. Key for references: 1 = [Rong et al. 2007a]; 2 = [Mohri et al. 1998]; 3 = [Adams et al. 2005]; 4 = [Perelson et al. 1993]; 5 = [Bonhoeffer et al. 2000].

- (3) The interior steady state  $S_i$ , when both the wild-type and the resistant strains coexist:

$$S_i := (T_i, T_{si}, V_{si}, T_{ri}, V_{ri}, E_i), \quad (5)$$

where

$$T_i = \frac{\lambda_T c}{dc + \delta(k_s N_s T_{si} + k_r N_r T_{ri})}, \quad V_{si} = \frac{\delta N_s T_{si}}{c},$$

$$V_{ri} = \frac{\delta N_r T_{ri}}{c}, \quad E = \frac{\lambda_E + c_E(T_{si} + T_{ri})}{\delta_E},$$

and  $T_{si}$  and  $T_{ri}$  are the solutions of the system

$$\begin{cases} \frac{(1-u)k_s N_s \delta \lambda_T}{dc + \delta(k_s N_s T_s + k_r N_r T_r)} - \delta - \frac{m_1(\lambda_E + c_E(T_s + T_r))}{\delta_E} = 0, \\ \frac{\delta \lambda_T (uk_s N_s T_s + k_r N_r T_r)}{dc + \delta(k_s N_s T_s + k_r N_r T_r)} - \delta T_r - \frac{m_2(\lambda_E + c_E(T_s + T_r))T_r}{\delta_E} = 0. \end{cases} \quad (6)$$

In the special case that there is no mutation, i.e.,  $u = 0$ , the interior steady state  $S_i$  reduces to another boundary steady state  $S_w$ , when only the wild-type strain is present:

$$S_w := (T_w, T_{sw}, V_{sw}, T_{rw}, V_{rw}, E_w), \quad (7)$$

where

$$T_{rw} = 0, \quad V_{rw} = 0,$$

$$T_{sw} = \frac{c}{N_s \delta} \frac{\lambda_T - d T_w}{k_s T_w}, \quad V_{sw} = \frac{\lambda_T - d T_w}{k_s T_w}, \quad E_w = \frac{\lambda_E}{\delta_E} + \frac{c_E}{\delta_E} \frac{\lambda_T - d T_w}{k_s T_w},$$

and  $T_w$  is the positive solution of the quadratic equation  $T^2 - A_w b T - B_w = 0$ , where

$$A_w = \frac{c}{N_s \delta k_s} \left( \delta + m_1 \frac{\lambda_E}{\delta_E} - m_1 d \frac{c}{N_s \delta k_s} \frac{c_E}{\delta_E} \right) \quad \text{and} \quad B_w = m_1 \left( \frac{c}{N_s \delta k_s} \right)^2 \frac{c_E}{\delta_E} \lambda_T.$$

The other steady states  $S_0$  and  $S_b$  are the same.

Let

$$R_s := \frac{N_s \delta k_s \lambda_T}{cd(\delta + m_1 \frac{\lambda_E}{\delta_E})} \quad \text{and} \quad R_r := \frac{N_r \delta k_r \lambda_T}{cd(\delta + m_2 \frac{\lambda_E}{\delta_E})} \quad (8)$$

denote the basic reproductive ratios of the wild-type strain and the drug-resistant strain, respectively, and let  $\sigma = (k_s N_s)/(k_r N_r)$ . In [Tarfulea 2011b], it was shown that the infection-free steady state  $S_0$  is locally asymptotically stable if  $R_r < 1$  and  $R_s < 1/(1-u)$ , and it is unstable if  $R_r > 1$  or  $R_s > 1/(1-u)$ . In the case that  $u = 0$  in model (1) (i.e., there is no mutation), the infection-free steady state  $S_0$  is locally asymptotically stable if  $R_r < 1$  and  $R_s < 1$ , and it is unstable if  $R_r > 1$  or  $R_s > 1$ .



**2.2. Model with antiretroviral therapy.** There are two major classes of antiretroviral drugs which are utilized in HIV treatment: the reverse transcriptase inhibitors (RTI) and the protease inhibitors (PI). Combinations of these are used in a regimen known as highly active antiretroviral therapy (HAART) [Cohen and Boyle 2004; Cohen 2005a; 2005b; El-Sadr et al. 2006; Nowak et al. 1997; Sharomi and Gumel 2008] designed to limit the virus' ability to mutate and develop drug-resistant strains. Nucleoside reverse transcriptase inhibitors (NRTIs) and nonnucleoside reverse transcriptase inhibitors (NNRTIs) inhibit reverse transcription enzymes. Entry inhibitors prevent the virus from attaching to the surface of the lymphocytes. This class of drugs in our model would have an impact on reducing  $k_s$  and  $k_r$ , the infection rates for the wild-type and the drug-resistant viruses. Protease inhibitors inhibit the protein enzymes that cut viral proteins to the correct size. PIs go to work after the process of reverse transcription by inhibiting the activity of protease, an enzyme needed by the virus for the production of new virions in infected lymphocytes [Casiday and Frey 2001], and this would impact  $N_s$  and  $N_r$ , the number of virions produced per infected drug-sensitive and drug-resistant cell, respectively.

We study the antiretroviral drug therapy in this system by introducing drug-efficacy parameters, which are extensively used in numerous models, such as [Adams et al. 2005; Perelson and Nelson 1999; Rong et al. 2007a; 2007b]. We consider  $\varepsilon_{RT}^s$  and  $\varepsilon_{RT}^r$  to represent the efficacies of RTIs and  $\varepsilon_{PI}^s$  and  $\varepsilon_{PI}^r$  to be the efficacies of PIs, for drug-sensitive and drug-resistant strains. These drugs are incorporated into model (1) to obtain the following system (the initial condition used is the values for the infected steady state in the no-treatment case given by (5) and the parameter values used are from Table 2):

$$\begin{aligned}
\frac{dT}{dt} &= \lambda_T - Td - k_s(1 - \varepsilon_{RT}^s)V_sT - k_r(1 - \varepsilon_{RT}^r)V_rT, \\
\frac{dT_s}{dt} &= (1 - u)k_s(1 - \varepsilon_{RT}^s)TV_s - \delta T_s - m_1ET_s, \\
\frac{dV_s}{dt} &= N_s(1 - \varepsilon_{PI}^s)\delta T_s - cV_s, \\
\frac{dT_r}{dt} &= uk_s(1 - \varepsilon_{RT}^s)TV_s + k_r(1 - \varepsilon_{RT}^r)V_rT - \delta T_r - m_2ET_r, \\
\frac{dV_r}{dt} &= N_r(1 - \varepsilon_{PI}^r)\delta T_r - cV_r, \\
\frac{dE}{dt} &= \lambda_E + c_E(T_s + T_r) - \delta_E E.
\end{aligned} \tag{9}$$

The case of constant drug efficacies has been addressed in several models (see [Adams et al. 2005; Perelson and Nelson 1999; Rong et al. 2007a; 2007b; Tarfulea et al. 2011; Tarfulea 2011b]). In this case,  $\varepsilon_{RT}^s$ ,  $\varepsilon_{PI}^s$ ,  $\varepsilon_{RT}^r$ , and  $\varepsilon_{PI}^r$  lie in  $[0, 1]$ . In

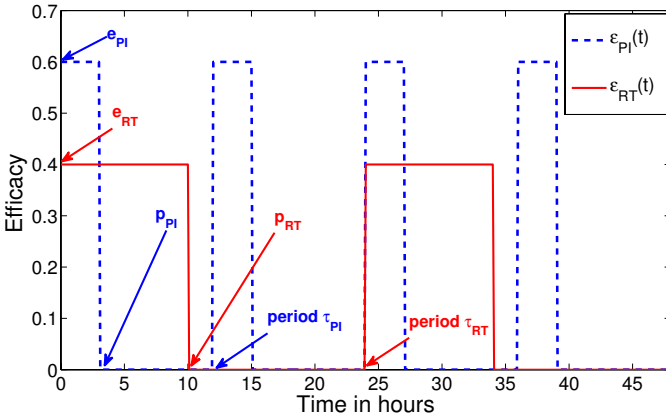
the case that all are zero, i.e., no treatment, we obtain system (1); if all are 1, then we obtain a complete cure of the disease since  $dV_s/dt < 0$  and  $dV_r/dt < 0$ . Moreover, we have that  $\varepsilon_{RT}^s > \varepsilon_{RT}^r$  and  $\varepsilon_{PI}^s > \varepsilon_{PI}^r$  since the wild-type virus is more susceptible to drugs. Therefore we can consider that  $\varepsilon_{RT}^r = \alpha \varepsilon_{RT}^s$  or that  $\varepsilon_{PI}^r = \alpha \varepsilon_{PI}^s$ , where  $0 < \alpha < 1$  and  $\alpha$  represents the HIV mutants' level of resistance; as  $\alpha$  decreases, there is more resistance to the used drug for the drug-resistant strains. However, in reality, the drug efficacies are not constant in time; thus the main purpose of this paper is to investigate the effect of including periodic antiretroviral therapy of bang-bang type.

### 3. Time-varying drug efficiency

In this section, we include time-varying drug efficacy functions to model various treatment regimens. Thereafter, we consider the model (9) where  $\varepsilon_{RT}^s(t)$ ,  $\varepsilon_{RT}^r(t)$ ,  $\varepsilon_{PI}^s(t)$ , and  $\varepsilon_{PI}^r(t)$  are functions of time with range the interval  $[0, 1]$  and they represent the time-varying drug efficacies of the RTIs and PIs for drug-sensitive and drug-resistant strains. When  $\varepsilon_{RT}^s(t)$ ,  $\varepsilon_{RT}^r(t)$  or  $\varepsilon_{PI}^s(t)$ ,  $\varepsilon_{PI}^r(t)$  are close to zero, the drug has almost no effect, while if they are near 1, the viral replication is almost completely inhibited. The shapes of these functions are determined by the pharmacokinetics that describe what happens to a drug after the moment of intake and before starting to be active at the infection site [De Leenheer 2009]. It is characterized by a fast rise to the peak value immediately after the drug intake, followed by a slower decay. Thus, we consider that each of the drug efficacies considered,  $\varepsilon_{RT}^s(t)$ ,  $\varepsilon_{RT}^r(t)$ ,  $\varepsilon_{PI}^s(t)$ , and  $\varepsilon_{PI}^r(t)$ , is periodic, that is  $\varepsilon_{RT}^s(t) = \varepsilon_{RT}^s(t + \tau_{RT}^s)$  and  $\varepsilon_{PI}^s(t) = \varepsilon_{PI}^s(t + \tau_{PI}^s)$  for all  $t$ , where  $\tau_{RT}^s, \tau_{PI}^s > 0$  are the principal periods for the RTIs and PIs for the sensitive strain. We have similar relations for the efficiency of the drug-resistant strain. For example, the period is 1 if medication is taken daily or 0.5 for a twice a day treatment schedule. Moreover, we assume the efficiency functions to be of the bang-bang type, i.e., at any time during treatment, the drug is either active or inactive. It is clear that is just an approximation of the real shape of  $\varepsilon(t)$  determined by the pharmacokinetics, but some key properties are to be revealed from this case. These functions are given by

$$\begin{aligned} \varepsilon_{RT}^s(t) &= \begin{cases} e_{RT}^s, & \text{for } t \in [0, p_{RT}^s], \\ 0, & \text{for } t \in (p_{RT}^s, \tau_{RT}^s), \end{cases} \\ \varepsilon_{PI}^s(t) &= \begin{cases} e_{PI}^s, & \text{for } t \in [0, p_{PI}^s], \\ 0, & \text{for } t \in (p_{PI}^s, \tau_{PI}^s), \end{cases} \end{aligned} \quad (10)$$

with a similar behavior for  $\varepsilon_{RT}^r$  and  $\varepsilon_{PI}^r$ . An example of such functions is illustrated in Figure 1. Here  $p_{RT}^s \in (0, \tau_{RT}^s)$  is the time duration when the RT drug is active with efficacy  $e_{RT}^s \in [0, 1]$ , and  $p_{PI}^s$  and  $e_{PI}^s$  are defined similarly. The drug is assumed to be totally inefficient during the remaining part of the corresponding



**Figure 1.** An example of periodic drug efficacies functions of the bang-bang type,  $\varepsilon_{RT}(t)$  (solid line) and  $\varepsilon_{PI}(t)$  (dotted line). Here RTI drug has the period  $\varepsilon_{RT} = 1$  (i.e., 24 h), is active for 10 h (i.e.,  $p_{RT} = 0.42$ ) with efficacy  $e_{RT} = 0.4$ ; PI drug has the period  $\varepsilon_{PI} = 0.5$  (i.e., 12 h), is active for 4 h (i.e.,  $p_{PI} = 0.17$ ) with efficacy  $e_{PI} = 0.6$ .

period. The same relations hold for drug-resistant drug efficacies. Furthermore, we have that  $\varepsilon_{RT}^s > \varepsilon_{RT}^r$  and  $\varepsilon_{PI}^s > \varepsilon_{PI}^r$  since the wild-type virus is more susceptible to drugs. Therefore, we can consider that  $\varepsilon_{RT}^r = \alpha_1 \varepsilon_{RT}^s$  or that  $\varepsilon_{PI}^r = \alpha_2 \varepsilon_{PI}^s$ , where  $0 < \alpha_1, \alpha_2 < 1$  and  $\alpha_1, \alpha_2$  represent the HIV mutants' level of resistance; as  $\alpha_1$  or  $\alpha_2$  decreases, there is more resistance to the used drug for the drug-resistant strains.

In order to compare our results with results from related models using constant efficacies, we define the average drug efficacy for each type of drug used, given by

$$\bar{\varepsilon}_{RT}^s := \frac{1}{\tau_{RT}^s} \int_0^{\tau_{RT}^s} \varepsilon_{RT}^s(t) dt \quad \text{and} \quad \bar{\varepsilon}_{PI}^s := \frac{1}{\tau_{PI}^s} \int_0^{\tau_{PI}^s} \varepsilon_{PI}^s(t) dt, \quad (11)$$

and thus,

$$\bar{\varepsilon}_{RT}^s = \frac{e_{RT}^s p_{RT}^s}{\tau_{RT}^s} \quad \text{and} \quad \bar{\varepsilon}_{PI}^s = \frac{e_{PI}^s p_{PI}^s}{\tau_{PI}^s},$$

for the sensitive strain, and

$$\bar{\varepsilon}_{RT}^r := \frac{1}{\tau_{RT}^r} \int_0^{\tau_{RT}^r} \varepsilon_{RT}^r(t) dt \quad \text{and} \quad \bar{\varepsilon}_{PI}^r := \frac{1}{\tau_{PI}^r} \int_0^{\tau_{PI}^r} \varepsilon_{PI}^r(t) dt, \quad (12)$$

and thus,

$$\bar{\varepsilon}_{RT}^r = \frac{e_{RT}^r p_{RT}^r}{\tau_{RT}^r} \quad \text{and} \quad \bar{\varepsilon}_{PI}^r = \frac{e_{PI}^r p_{PI}^r}{\tau_{PI}^r},$$

for the resistant strain. Moreover, we introduce the overall treatment effects

$$\varepsilon^s = 1 - (1 - \bar{\varepsilon}_{RT}^s)(1 - \bar{\varepsilon}_{PI}^s) \quad \text{and} \quad \varepsilon^r = 1 - (1 - \bar{\varepsilon}_{RT}^r)(1 - \bar{\varepsilon}_{PI}^r) \quad (13)$$

for the wild-type and mutant strains, respectively.

There are two parameters which can vary in the efficacies  $\varepsilon(t)$  (for both RTIs and PIs), namely the efficacy of the drug  $e$  and the time duration  $p$ . In the remaining part of this section, we investigate their effect on the Floquet multipliers of systems (15) and (16).

We begin by investigating the effect of only one drug in the system at a time. Let us assume first that the efficiencies  $\varepsilon_{RT}^s(t)$  and  $\varepsilon_{RT}^r(t)$  are periodic (as described above) and  $\varepsilon_{PI}^s(t) = 0$  and  $\varepsilon_{PI}^r(t) = 0$ , i.e., only RTIs are administered in the system. Notice that the infection-free steady state

$$S_0 = \left( T_0 = \frac{\lambda_T}{d}, T_{s0} = 0, V_{s0} = 0, T_{r0} = 0, V_{r0} = 0, E_0 = \frac{\lambda_E}{\delta} \right)$$

is still an equilibrium solution of the model (9), regardless the inclusion of the drug efficiency. Moreover, in our investigation we use only this steady state since its stability implies that the treatment can clear the infection. Thus, we linearize the system (9) about  $S_0$  and obtain the linear system

$$\frac{dx}{dt} = A(t)x, \quad (14)$$

where

$$A(t) = \begin{pmatrix} -d & 0 & -a_{RT}^s(t) & 0 & -a_{RT}^r(t) & 0 \\ 0 & -\delta - m_1 E_0 & (1-u)a_{RT}^s(t) & 0 & 0 & 0 \\ 0 & N_s \delta & -c & 0 & 0 & 0 \\ 0 & 0 & u a_{RT}^s(t) & -\delta - m_2 E_0 & a_{RT}^r(t) & 0 \\ 0 & 0 & 0 & N_r \delta & -c & 0 \\ 0 & c_E & 0 & c_E & 0 & -\delta_E \end{pmatrix},$$

with  $a_{RT}^s(t) = k_s(1 - \varepsilon_{RT}^s(t))T_0$  and  $a_{RT}^r(t) = k_r(1 - \varepsilon_{RT}^r(t))T_0$ . Here  $x$  is the six-dimensional vector function whose components are the perturbations corresponding to the main variables  $T$ ,  $T_s$ ,  $V_s$ ,  $T_r$ ,  $V_r$ , and  $E$ , respectively. The local stability properties of  $S_0$  for system (9) are determined by the Floquet multipliers of (14) (see [De Leenheer and Smith 2003]) which, given the block-triangular structure of  $A(t)$ , are  $e^{-d\tau}$ ,  $\lambda_2$ ,  $\lambda_3$ ,  $\lambda_4$ ,  $\lambda_5$ , and  $e^{-\delta_E \tau}$ , where  $\lambda_2$  and  $\lambda_3$  are the Floquet multipliers of the planar  $\tau$ -periodic system

$$\begin{pmatrix} \dot{x}_2 \\ \dot{x}_3 \end{pmatrix} = \begin{pmatrix} -\delta - m_1 E_0 & (1-u)k_s(1 - \varepsilon_{RT}^s(t))T_0 \\ N_s \delta & -c \end{pmatrix} \begin{pmatrix} x_2 \\ x_3 \end{pmatrix}, \quad (15)$$

and  $\lambda_4$  and  $\lambda_5$  are the Floquet multipliers of the planar  $\tau$ -periodic system

$$\begin{pmatrix} \dot{x}_4 \\ \dot{x}_5 \end{pmatrix} = \begin{pmatrix} -\delta - m_2 E_0 & (1-u)k_r(1 - \varepsilon_{RT}^r(t))T_0 \\ N_r \delta & -c \end{pmatrix} \begin{pmatrix} x_4 \\ x_5 \end{pmatrix}. \quad (16)$$

The infection-free steady state  $S_0$  is locally asymptotically stable for system (9) if the Floquet multipliers of system (14) are contained in the unit disk of the complex plane, which is satisfied if  $|\lambda_2|, |\lambda_3|, |\lambda_4|, |\lambda_5| < 1$ . Unfortunately it is well known that for general functions  $\varepsilon(t)$  this condition is difficult to verify. If we consider the drug efficacies  $\varepsilon_{RT}^s(t)$  and  $\varepsilon_{RT}^r(t)$  of the bang-bang form given by (10), we get that the Floquet multipliers  $\lambda_2$  and  $\lambda_3$  of system (15) are the eigenvalues of the matrix

$$\Phi(e_{RT}^s, p_{RT}^s) := \exp((\tau_{RT}^s - p_{RT}^s)B(0)) \exp(p_{RT}^s B(e_{RT}^s)), \quad (17)$$

where the matrix function  $B(\cdot)$  is defined by

$$B(e_{RT}^s) := \begin{pmatrix} -\delta - m_1 E_0 & (1-u)k_s(1 - e_{RT}^s)T_0 \\ N_s \delta & -c \end{pmatrix}, \quad (18)$$

for any value of  $e_{RT}^s$ . Using the approach in [De Leenheer and Smith 2003], we obtain that the Floquet multipliers are contained in the interior of the unit disk of the complex plane if and only if the spectral radius  $\rho(\Phi(e_{RT}^s, p_{RT}^s))$  of the matrix  $\Phi(e_{RT}^s, p_{RT}^s)$  is less than 1. Furthermore, by applying Proposition 2 in [De Leenheer and Smith 2003] to our system, we get the expected monotonicity properties: the spectral radius is decreasing in each of its arguments. That is, if the treatment is periodic of the bang-bang type and it can eradicate the virus, then the infection is cleared more rapidly when the treatment is more effective or it lasts longer. These effects are confirmed by the results obtained from the numerical investigations described in the second part of this section.

We obtain a similar result if we consider the effect of only PIs, in which case

$$B(e_{PI}^s) := \begin{pmatrix} -\delta - m_1 E_0 & (1-u)k_s T_0 \\ N_s(1 - e_{PI}^s)\delta & -c \end{pmatrix},$$

or if we consider a cocktail of drugs where both inhibitors are present, in which case

$$B(e_{RT}^s, e_{PI}^s) = \begin{pmatrix} -\delta - m_1 E_0 & (1-u)k_s(1 - e_{RT}^s)T_0 \\ N_s(1 - e_{PI}^s)\delta & -c \end{pmatrix}.$$

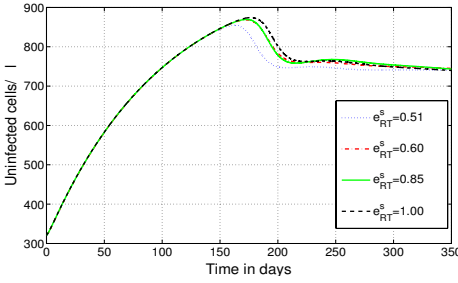
**3.1. Numerical results.** In this section, we analyze our results from the numerical investigations performed. We created MATLAB codes in order to solve the system numerically which allowed us to test and validate the mathematical mode and to explore various scenarios. We used `ode45` and `ode15s`, two MATLAB functions for the numerical solutions for our systems of differential equations (`ode45` is based on an explicit Runge–Kutta (4,5) formula, the Dormand–Prince pair, a one-step

solver that needs only the solution at the immediately preceding time point, whereas ode15s is a variable order solver based on the backward differentiation formulas, Gear's method, a multistep solver for stiff problems).

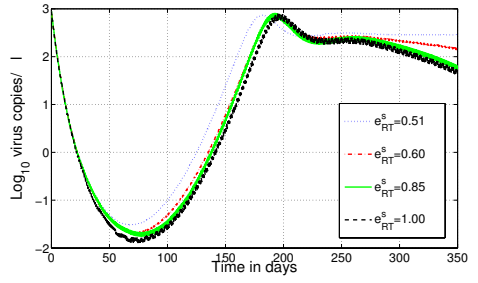
Our focus is placed on the following areas of interest: quantity of viral load and uninfected cell count for individual drug intake where average drug efficacies ( $\bar{\varepsilon}_{RT}^s$ ,  $\bar{\varepsilon}_{PI}^s$ ) are fixed and the time duration when the drug is active is varied, quantity of viral load and uninfected cell count for both classes of drugs taken in conjunction where drug efficacies are fixed and the time duration when the drug is active is varied, the effect on viral load and uninfected cell count for both drugs taken in conjunction where the ratio of their corresponding efficacies are varied over the same period, the effect on viral load and uninfected cell concentration while strictly varying the total efficacy of either drug, and the effect on viral load when the level of resistances ( $\alpha_1$ ,  $\alpha_2$ ) for the resistant-type viruses are varied.

We first consider a treatment scenario with only the reverse transcriptase inhibitor (RTI) drug where we fix the average efficacy,  $\bar{\varepsilon}_{RT}^s$  and vary the step-function parameters,  $e_{RT}^s$  and  $p_{RT}^s$ . Note that  $\varepsilon^s = 1 - (1 - \bar{\varepsilon}_{RT}^s)(1 - \bar{\varepsilon}_{PI}^s)$ , as defined by (13) (the same relation holds for  $\varepsilon^r$ ). We choose  $\varepsilon^s = 0.51$  and since we are only considering the RTI drug, we choose  $\bar{\varepsilon}_{PI}^s = 0.00$  and therefore  $\bar{\varepsilon}_{RT}^s = 0.51$ . We also note that in the periodic step-function, we have  $\bar{\varepsilon}_{RT}^s = (e_{RT}^s p_{RT}^s) / \tau_{RT}^s$ . We therefore pick the convenient ordered pair values for  $(e_{RT}^s, p_{RT}^s) \in \{(0.51, 1.00), (0.60, 0.85), (0.85, 0.60), (1.00, 0.51)\}$ . As intuition would lead us to expect, we see that the total viral load is lowest at the time when the drug is active is the largest (i.e., the case for which  $(e_{RT}^s, p_{RT}^s) = (0.51, 1.00)$ ). However, we also see the result in which the uninfected cell concentration has an inverse relationship to the viral load, due to the resistant strain virus. The wild-type viral load behaves similarly to the uninfected CD4+ T-cells. More specifically, the uninfected cell concentration peaks the highest and also converges to the highest steady state when the period over which the drug is released is the shortest (i.e., the case for which  $(e_{RT}^s, p_{RT}^s) = (1.00, 0.51)$ ) (see Figure 2). This is a result similar to the case when constant efficiencies  $\varepsilon_{RT}$ ,  $\varepsilon_{PI}(t)$  are used (see [Rong et al. 2007a; Tarfulea 2011b]). An analogous conclusion is obtained when investigating the effects on viral load and uninfected cell concentration when considering a treatment such that  $\bar{\varepsilon}_{RT}^s = 0.00$  and  $\bar{\varepsilon}_{PI}^s = 0.51$ , in other words, a treatment using only protease inhibitors (PIs) and varying the step-function parameters as done for RTIs. In all the above mentioned cases, we consider  $\alpha_1 = \alpha_2 = 0.2$ .

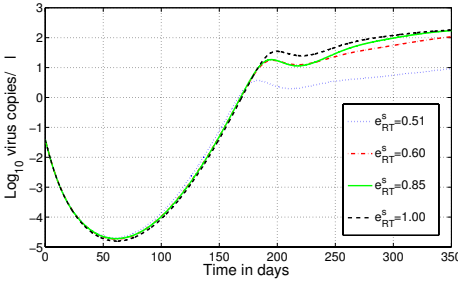
We now consider a treatment scenario in which RTIs and PIs are used in conjunction. Our first investigation begins with setting the efficacies of both drugs to be equal (i.e.,  $\varepsilon_{RT}^s = \varepsilon_{PI}^s$ ). Therefore, we again choose  $\varepsilon^s = 0.51$  (with  $\alpha_1 = \alpha_2 = 0.2$ ), and therefore it follows from  $\varepsilon^s = 1 - (1 - \bar{\varepsilon}_{RT}^s)(1 - \bar{\varepsilon}_{PI}^s)$  that  $\bar{\varepsilon}_{RT}^s = \bar{\varepsilon}_{PI}^s = 0.30$ . Thus, the equivalence of the ordered pairs  $(e_{RT}^s, p_{RT}^s)$  and  $(e_{PI}^s, p_{PI}^s)$  follows. We



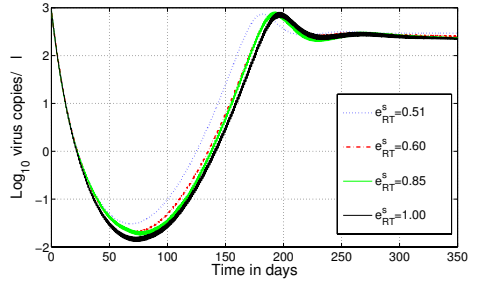
(a) Uninfected cells  $T(t)$ .



(b) Wild-type virus  $V_s(t)$ .



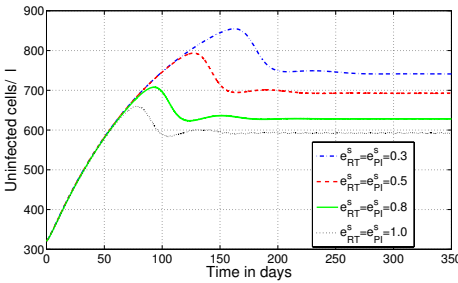
(c) Resistant virus  $V_r(t)$ .



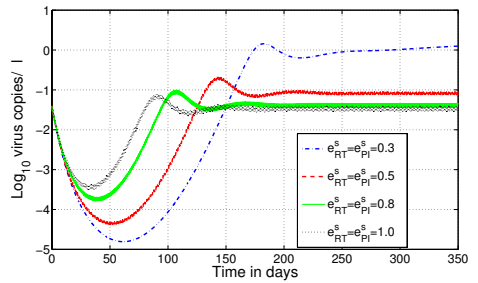
(d) Total virus.

**Figure 2.** Simulation over the first 350 days of infection with  $\bar{\epsilon}_{RT}^s = 0.51$  and  $\bar{\epsilon}_{PI}^s = 0.00$ ; thus  $\epsilon^s = 0.51$  (see text for details).

therefore choose the convenient ordered pair values for  $(e_{RT}^s, p_{RT}^s) = (e_{PI}^s, p_{PI}^s) \in \{(0.3, 1.0), (0.5, 0.6), (0.8, 0.375), (1.0, 0.3)\}$ . This simulation yields results which are opposed that of the results when the two drugs were used individually and are presented in [Figure 3](#) for the uninfected T-cell and resistant strain virus concentrations. The lowest viral peak with a convergence to the lowest steady state came from the highest drug efficacy and shortest time release period (i.e.,  $e_{RT}^s = e_{PI}^s = 1.0$  and  $p_{RT}^s = p_{PI}^s = 0.3$ ); that is, it is better if the drug is effective longer than if it

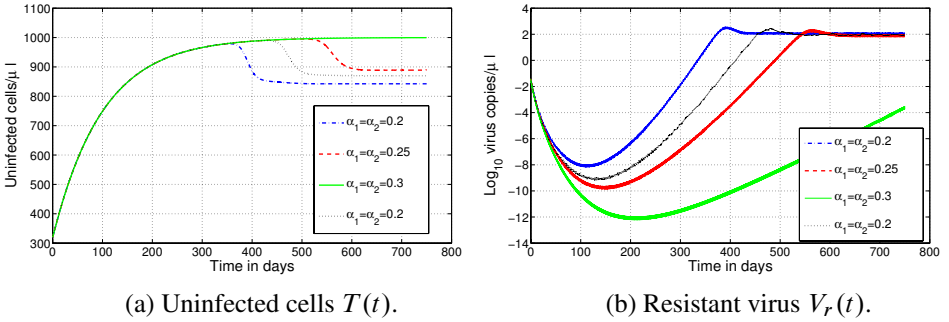


(a) Uninfected cells  $T(t)$ .



(b) Resistant virus  $V_r(t)$ .

**Figure 3.** Simulation over the first 350 days of drug treatment with  $\epsilon^s = 0.51$  and  $\bar{\epsilon}_{RT}^s = \bar{\epsilon}_{PI}^s = 0.3$ .



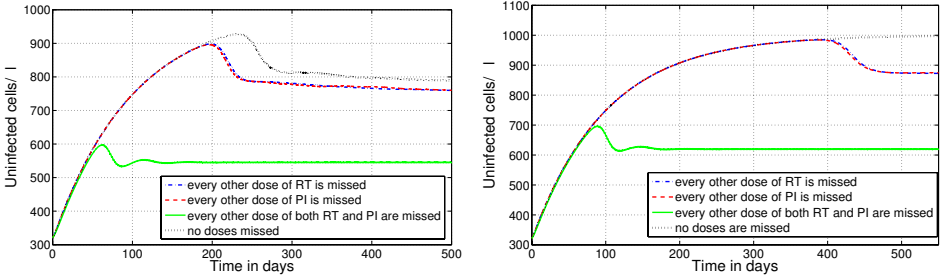
**Figure 4.** Simulation over the first 750 days of drug treatment, varying the HIV mutants' level of resistance. For all but black dotted line,  $\bar{\varepsilon}_{RT}^s = \bar{\varepsilon}_{PI}^s = 0.71$ , and for the black dotted line,  $\bar{\varepsilon}_{RT}^s = \bar{\varepsilon}_{PI}^s = 0.81$

has a higher peak. However, as in the case of the individual drug treatment cases, the uninfected cell concentration had inverse results to the resistant strain viral concentration and behaves similarly to the wild-type viral concentration.

Our next analysis considers the effects on the viral load and uninfected cell concentrations while varying the efficacies for both drugs. We continue to consider a fixed overall treatment effect where  $\varepsilon^s = 0.51$ . We then examine the average efficacy values  $(\bar{\varepsilon}_{RT}^s, \bar{\varepsilon}_{PI}^s) \in \{(0.51, 0.0), (0.41, 0.17), (0.3, 0.3), (0.17, 0.41)\}$ . Here again  $\alpha_1 = \alpha_2 = 0.2$ . We considered the results of  $(\bar{\varepsilon}_{RT}^s, \bar{\varepsilon}_{PI}^s) = (0.51, 0.0)$  in a previous section and used these values again for comparison. It is not surprising to see that this is, in fact, the least efficient scenario since the others involve a drug cocktail as opposed to this one-drug treatment. It is noted that the best result, having the lowest viral peak and convergent steady state with the highest uninfected cell concentration, comes from a drug cocktail in which the drug efficacy ratio (RTI:PI) is 1:4. Moreover, as we would expect, administering any cocktail of drugs with any chosen efficacies (without keeping a constant overall efficacy) gives better results than the individual classes of drugs alone.

Recall that we use resistance rates,  $\alpha_1$  and  $\alpha_2$ , such that  $\varepsilon_{RT}^r = \alpha_1 \varepsilon_{RT}^s$  and  $\varepsilon_{PI}^r = \alpha_2 \varepsilon_{PI}^s$ , with  $\alpha_1, \alpha_2 \in (0, 1)$ . We consider the effects on viral load for varying levels of resistance. We let  $\alpha_1, \alpha_2 \in \{0.25, 0.5, 0.75, 1.0\}$ . Note that when  $\alpha_1 = \alpha_2 = 1.0$ , the efficacy for the drugs against the mutant virus is equal to that of the drug-sensitive-type virus. We see the intuitive results that demonstrate that when  $\alpha_1, \alpha_2$  get closer to 1, the total viral load for the resistant-type, the mutant virus decreases. We next consider fixing one of the resistant rates (i.e., the resistant rate for one of the drugs) and vary the other. We observe the total viral load in the case where we fix  $\alpha_2 = 0.25$  and vary  $\alpha_1 \in \{0.25, 0.5, 0.75, 1.0\}$ . It is noted that, again, we see the lowest viral load is obtained when  $\alpha_1 = 1.0$ , and as  $\alpha_1$  becomes closer





(a)  $\bar{\epsilon}_{RT}^s = \bar{\epsilon}_{PI}^s = 0.51$  and  $\alpha_1 = \alpha_2 = 0.2$ . (b)  $\bar{\epsilon}_{RT}^s = \bar{\epsilon}_{PI}^s = 0.71$  and  $\alpha_1 = \alpha_2 = 0.3$ .

**Figure 5.** Uninfected T-cell concentration  $T(t)$  under suboptimal treatment.

to 1, the viral load of the resistant-type virus decreases. An analogous observation is made for fixing  $\alpha_1$  and varying  $\alpha_2$ . In Figure 4, we consider the efficacies for the two classes of drugs to be  $\bar{\epsilon}_{RT}^s = \bar{\epsilon}_{PI}^s = 0.71$ , which guarantees that the wild-type virus is suppressed. We let  $p_{RT}^s = p_{PI}^s = 0.71$  and we see that for values for  $\alpha_1$  and  $\alpha_2$  lower than 0.3, the drug-resistant strain persists. Moreover, if we increase the drug efficacies to 0.81, for  $\alpha_1 = \alpha_2 = 0.2$ , the drug-resistant strain still persists.

One of the critical obstacles to successful HIV drug therapy is the imperfect adherence to a prescribed drug regimen due to its complexity or severe side effects. Receiving treatment for HIV is expensive and people can be careless; therefore we want to look into the effects of missing doses. We investigate numerous efficacy combinations and RTI/PI individual and/or combined treatments. The results unanimously indicate that skipping a dose of either drug at any combination has certain undesirable effects which included a weaker drop in viral load and lower overall uninfected cell concentration. In Figure 5, we present the dynamics of uninfected T-cell concentration when every other dose of RTIs, PIs, or both are missed and compare with the dynamics of a regular treatment. In Figure 5(a) we consider  $\bar{\epsilon}_{RT}^s = \bar{\epsilon}_{PI}^s = 0.51$ ,  $p_{RT}^s = p_{PI}^s = 0.51$ , and  $\alpha_1 = \alpha_2 = 0.2$ , whereas in Figure 5(b) we consider  $\bar{\epsilon}_{RT}^s = \bar{\epsilon}_{PI}^s = 0.71$ ,  $p_{RT}^s = p_{PI}^s = 0.71$ , and  $\alpha_1 = \alpha_2 = 0.3$ . In the latter case, the viral load is eradicated under perfect adherence, but the uninfected T-cell concentration decreases and both strains of virus persist even when only one drug is missed.

### 4. Conclusions

We have developed and analyzed a mathematical model that accounts for multiple viral strains during the course of antiretroviral therapy with periodic antiretroviral therapy of bang-bang type. There were many different circumstances that we investigated thoroughly. The first area of interest was determining how the system behaves when only the presence of one antiretroviral class of drugs is used. This was done for each of the two classes of interest, namely protease inhibitors and

reverse transcriptase inhibitors. It was noted, based on the periodic step-function used for our analysis, that, upon taking only one of the two available drugs, when the efficacy of either drug was increased and the period over which the drug would be active, the total viral load decreased. There was an identical scenario for either drug taken alone.

Certainly, the optimal scenario for drug treatment is by means of a patient taking a cocktail of both classes of drugs. Therefore, it was of great importance to investigate the functionality of using both drugs of interest simultaneously. When the two drugs were taken in conjunction, they had an inverse effect on the infected body. In other words, when we increased initial efficacy of the drug cocktail and decreased the period, the total viral load decreased. For any of the scenarios investigated, however, the total uninfected cell count responded inversely to the response of the resistant strain viral load. This led us to the conclusion that the drug cocktail was not only the proper choice, but we also observed that it was most effective given at a 4:1 ratio (protease inhibitors: reverse transcriptase inhibitors). Furthermore, the examination of the effects of using different ratios of both drugs to further optimize the efficacy of the treatment was also of substantial interest. Scenarios for both varying efficacy and varying the level of resistance to the drug therapy by the drug-resistant-type virus were examined. As the level of drug efficacy increased, there were noticeable increases in the uninfected cell count as well as a stronger decrease in the total viral load. When the level of resistance was increased, we noted an increase in viral load as we expected. Although seemingly intuitive, we were also sure to investigate the functionality of the system when varying both drug efficacies, individually and in tandem, and the results of the evolution of drug resistance.

Given the staggering percentage of infected people that are either unable to obtain the appropriate drug therapies or simply cannot take all the recommended doses, we also numerically investigated the effect of imperfect adherence to the prescribed treatment regimen. That is, we investigated what would happen when someone is under a drug regimen and particular doses were skipped. The last area of results we obtained consisted of scenarios where the infected person missed a certain number of doses for either drug and for both drugs together. Skipping doses for either drug alone had nearly identical effects; there was significantly less of a drop in viral load and the uninfected cell count was much lower. The results of missing doses when the drug cocktail was being administered followed directly from the individual missed doses as well.

## Appendix

The following table contains all of the symbols used throughout the paper (in the order of appearance).

symbol	description
$T$	healthy T-cell concentration
$T_s$	drug-sensitive infected T-cell concentration
$T_r$	drug-resistant infected T-cell concentration
$V_s$	drug-sensitive virus concentration
$V_r$	drug-resistant virus concentration
$E$	concentration of CD8+ T-cells
$\lambda_T$	recruitment rate of uninfected cells
$d$	death rate of uninfected cells
$k_s$	infection rate of T-cells by the wild-type virus
$k_r$	infection rate of T-cells by the drug-resistant virus
$\delta$	death rate of infected cells
$m_1$	immune-induced clearance rate for infected $T_s$ cells
$m_2$	immune-induced clearance rate for infected $T_r$ cells
$N_s$	virions produced per infected drug-sensitive cell
$N_r$	virions produced per infected drug-resistant cell
$c$	clearance rate of free virus
$c_E$	stimulation of CTL proliferation
$\delta_E$	death rate of immune effectors
$u$	mutation rate from sensitive strain to resistant strain
$S_0$	vector $(T_0, T_{s0}, V_{s0}, V_{r0}, E_0)$ with the infection-free steady state
$S_b$	vector $(T_b, T_{sb}, V_{sb}, V_{rb}, E_b)$ with the boundary steady state
$S_i$	vector $(T_i, T_{si}, V_{si}, V_{ri}, E_i)$ with the interior steady state
$S_w$	vector $(T_w, T_{sw}, V_{sw}, V_{rw}, E_w)$ with the wild-type steady state
$R_s$	basic reproductive ratio of the wild-type strain
$R_r$	basic reproductive ratio of the drug-resistant strain
$\varepsilon_{RT}^s$	efficacy of RTIs for drug-sensitive strain
$\varepsilon_{RT}^r$	efficacy of RTIs for drug-resistant strain
$\varepsilon_{PI}^s$	efficacy of PIs for drug-sensitive strain
$\varepsilon_{PI}^r$	efficacy of PIs for drug-resistant strain
$\alpha$	HIV mutants' level of resistance
$\tau_{RT}^s$	principal period for the RT inhibitors for the sensitive strain
$\tau_{PI}^s$	principal period for the P inhibitors for the sensitive strain

symbol	description
$p_{RT}^s$	time duration when the RT drug for the sensitive strain is active
$p_{RT}^r$	time duration when the RT drug for the resistant strain is active
$p_{PI}^s$	time duration when the P drug for the sensitive strain is active
$p_{PI}^r$	time duration when the P drug for the resistant strain is active
$e_{RT}^s$	efficacy of RT drugs for the sensitive strain
$e_{RT}^r$	efficacy of RT drugs for the resistant strain
$e_{PI}^s$	efficacy of P drugs for the sensitive strain
$e_{PI}^r$	efficacy of P drugs for the resistant strain
$\alpha_1$	HIV mutants' level of resistance for the RT drug
$\alpha_2$	HIV mutants' level of resistance for the P drug
$\bar{e}_{RT}^s$	average efficacy of RT drugs for sensitive strain
$\bar{e}_{RT}^r$	average efficacy of RT drugs for resistant strain
$\bar{e}_{PI}^s$	average efficacy of P drugs for sensitive strain
$\bar{e}_{PI}^r$	average efficacy of P drugs for resistant strain
$\varepsilon^s$	overall treatment effect on the sensitive strain
$\varepsilon^r$	overall treatment effect on the resistant strain

## References

- [Adams et al. 2005] B. M. Adams, H. T. Banks, M. Davidian, H.-D. Kwon, H. T. Tran, S. N. Wynne, and E. S. Rosenberg, “HIV dynamics: modeling, data analysis, and optimal treatment protocols”, *J. Comput. Appl. Math.* **184**:1 (2005), 10–49. MR 2006h:92020 Zbl 1075.92030
- [Bofill et al. 1992] M. Bofill, G. Janossy, C. A. Lee, D. MacDonald-Burns, A. N. Phillips, C. Sabin, A. Timms, M. A. Johnson, and P. B. Kernoff, “Laboratory control values for CD4 and CD8 T lymphocytes: implications for HIV-1 diagnosis”, *Clin. Exp. Immunol.* **88**:2 (1992), 243–252.
- [Bonhoeffer et al. 2000] S. Bonhoeffer, M. Rembiszewski, G. M. Ortiz, and D. F. Nixon, “Risks and benefits of structured antiretroviral drug therapy interruptions in HIV-1 infection”, *AIDS* **14**:15 (2000), 2313–2322.
- [Burg et al. 2009] D. Burg, L. Rong, A. U. Neumann, and H. Dahari, “Mathematical modeling of viral kinetics under immune control during primary HIV-1 infection”, *J. Theor. Biol.* **259**:4 (2009), 751–759. MR 2973193
- [Casiday and Frey 2001] R. Casiday and R. Frey, “Drug strategies to target HIV: enzyme kinetics and enzyme inhibitors”, preprint, Washington University, St. Louis, MO, 2001, <http://www.chemistry.wustl.edu/~edudev/LabTutorials/HIV/DrugStrategies.html>.
- [Cohen 2005a] C. J. Cohen, “EuroSIDA study confirms ability of most HAART regimens to decrease rates of illness and death”, in *The 12th Conference on Retroviruses and Opportunistic Infections* (Boston, 2005), The Body Pro, New York, February 24, 2005.

- [Cohen 2005b] C. J. Cohen, “Ritonavir-boosted protease inhibitors, 1: Strategies for balancing efficacy with effects on lipids”, *AIDS Read.* **15**:9 (2005), 462–465, 470–471, 474, 477.
- [Cohen and Boyle 2004] C. J. Cohen and B. A. Boyle, “Antiretroviral therapy: the ‘when to start’ debates”, *Clin. Infect. Dis.* **39**:11 (2004), 1705–1708.
- [De Leenheer 2009] P. De Leenheer, “Within-host virus models with periodic antiviral therapy”, *Bull. Math. Biol.* **71**:1 (2009), 189–210. MR 2010c:92044 Zbl 1169.92024
- [De Leenheer and Smith 2003] P. De Leenheer and H. L. Smith, “Virus dynamics: a global analysis”, *SIAM J. Appl. Math.* **63**:4 (2003), 1313–1327. MR 2004b:34136 Zbl 1035.34045
- [El-Sadr et al. 2006] W. M. El-Sadr et al., “CD4+ count-guided interruption of antiretroviral treatment”, *N. Engl. J. Med.* **355**:22 (2006), 2283–2296.
- [Fauci 1993] A. S. Fauci, “Immunopathogenesis of HIV infection”, *J. Acquir. Immune Defic. Syndr.* **6**:6 (1993), 655–662.
- [Huang 2008] Y. Huang, “Long-term HIV dynamic models incorporating drug adherence and resistance to treatment for prediction of virological responses”, *Comput. Stat. Data Anal.* **52**:7 (2008), 3765–3778. MR 2427379 Zbl 05564736
- [Kepler and Perelson 1998] T. B. Kepler and A. S. Perelson, “Drug concentration heterogeneity facilitates the evolution of drug resistance”, *Proc. Natl. Acad. Sci. USA* **95**:20 (1998), 11514–11519. Zbl 0919.92023
- [Kirschner and Webb 1997] D. E. Kirschner and G. F. Webb, “Understanding drug resistance for monotherapy treatment of HIV infection”, *Bull. Math. Biol.* **59**:4 (1997), 763–786.
- [Koup et al. 1994] R. A. Koup, J. T. Safrit, Y. Cao, C. A. Andrews, G. McLeod, W. Borkowsky, C. Farthing, and D. D. Ho, “Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome”, *J. Virol.* **68**:7 (1994), 4650–4655.
- [McMichael Winter 1996] A. J. McMichael, “How HIV fools the immune system”, *MRC News* (Winter 1996).
- [Mohri et al. 1998] H. Mohri, S. Bonhoeffer, S. Monard, A. S. Perelson, and D. D. Ho, “Rapid turnover of T lymphocytes in SIV-infected rhesus macaques”, *Science* **279**:5354 (1998), 1223–1227.
- [Murray and Perelson 2005] J. M. Murray and A. S. Perelson, “Human immunodeficiency virus: quasi-species and drug resistance”, *Multiscale Model. Simul.* **3**:2 (2005), 300–311. MR 2122990 Zbl 1068.92027
- [NHS 2008] National Health Service, “HIV and AIDS”, 2008, <http://www.nhs.uk/Conditions/HIV>.
- [Nowak et al. 1997] M. A. Nowak, A. L. Lloyd, G. M. Vasquez, T. A. Wiltrot, L. M. Wahl, N. Bischofberger, J. Williams, A. Kinter, A. S. Fauci, V. M. Hirsch, and J. D. Lifson, “Viral dynamics of primary viremia and antiretroviral therapy in simian immunodeficiency virus infection”, *J. Virol.* **71**:10 (1997), 7518–7525.
- [Perelson and Nelson 1999] A. S. Perelson and P. W. Nelson, “Mathematical analysis of HIV-1 dynamics in vivo”, *SIAM Rev.* **41**:1 (1999), 3–44. MR 1669741 Zbl 1078.92502
- [Perelson et al. 1993] A. S. Perelson, D. E. Kirschner, and R. De Boer, “Dynamics of HIV infection of CD4+ T cells”, *Math. Biosci.* **114**:1 (1993), 81–125. Zbl 0796.92016
- [Rapin et al. 2006] N. Rapin, C. Kesmir, S. Frankild, M. Nielsen, C. Lundegaard, S. Brunak, and O. Lund, “Modelling the human immune system by combining bioinformatics and systems biology approaches”, *J. Biol. Phys.* **32**:3–4 (2006), 335–353.

- [Ribeiro et al. 1998] R. M. Ribeiro, S. Bonhoeffer, and M. A. Nowak, “The frequency of resistant mutant virus before antiviral therapy”, *AIDS* **12**:5 (1998), 461–465.
- [Rong et al. 2007a] L. Rong, Z. Feng, and A. S. Perelson, “Emergence of HIV-1 drug resistance during antiretroviral treatment”, *Bull. Math. Biol.* **69**:6 (2007), 2027–2060. MR 2008e:92022 Zbl 1298.92053
- [Rong et al. 2007b] L. Rong, M. A. Gilchrist, Z. Feng, and A. S. Perelson, “Modeling within-host HIV-1 dynamics and the evolution of drug resistance: trade-offs between viral enzyme function and drug susceptibility”, *J. Theor. Biol.* **247**:4 (2007), 804–818. MR 2009m:92133
- [Sharomi and Gumel 2008] O. Sharomi and A. B. Gumel, “Dynamical analysis of a multi-strain model of HIV in the presence of anti-retroviral drugs”, *J. Biol. Dyn.* **2**:3 (2008), 323–345. MR 2009g:92129 Zbl 1154.92033
- [Shiri et al. 2005] T. Shiri, W. Garira, and S. D. Musekwa, “A two-strain HIV-1 mathematical model to assess the effects of chemotherapy on disease parameters”, *Math. Biosci. Eng.* **2**:4 (2005), 811–832. MR 2170427 Zbl 1097.92031
- [Smith and Wahl 2005] R. J. Smith and L. M. Wahl, “Drug resistance in an immunological model of HIV-1 infection with impulsive drug effects”, *Bull. Math. Biol.* **67** (2005), 783–813. MR 2006k:34017
- [Stafford et al. 2000] M. A. Stafford, L. Corey, Y. Cao, E. S. Daar, D. D. Ho, and A. S. Perelson, “Modeling plasma virus concentration during primary HIV infection”, *J. Theor. Biol.* **203**:3 (2000), 285–301.
- [Tarfulea 2011a] N. E. Tarfulea, “A mathematical model for HIV treatment with time-varying antiretroviral therapy”, *Int. J. Comput. Math.* **88**:15 (2011), 3217–3235. MR 2834516 Zbl 1237.92034
- [Tarfulea 2011b] N. E. Tarfulea, “A mathematical model for the CTL effect on the drug resistance during antiretroviral treatment of HIV infection”, IMA Preprint Series #2377, Institute for Mathematics and its Applications, Minneapolis, MN, August 2011, <http://www.ima.umn.edu/preprints/pp2011/2377.pdf>.
- [Tarfulea et al. 2011] N. E. Tarfulea, A. Blink, E. Nelson, and D. Turpin, Jr., “A CTL-inclusive mathematical model for antiretroviral treatment of HIV infection”, *Int. J. Biomath.* **4**:1 (2011), 1–22. MR 2012c:92088
- [Wahl and Nowak 2000] L. M. Wahl and M. A. Nowak, “Adherence and drug resistance: predictions for therapy outcome”, *Proc. R. Soc. Lond. B* **267**:1445 (2000), 835–843.
- [Wilson et al. 2000] J. D. K. Wilson, G. S. Ogg, R. L. Allen, C. Davis, S. Shaunak, J. Downie, W. Dyer, C. Workman, J. S. Sullivan, A. J. McMichael, and S. L. Rowland-Jones, “Direct visualization of HIV-1-specific cytotoxic T lymphocytes during primary infection”, *AIDS* **14**:3 (2000), 225–233.

Received: 2011-07-06

Revised: 2013-08-08

Accepted: 2014-05-31

ntarfule@purduecal.edu

*Mathematics, Computer Science and Statistics,  
Purdue University Calumet, 2200 169th Street,  
Hammond, IN 46323, United States*

paulread820@gmail.com

*Purdue University Calumet, 2200 196th Street,  
Hammond, IN 46323, United States*

# An extension of Young's segregation game

Michael Borchert, Mark Burek, Rick Gillman and Spencer Roach

(Communicated by Kenneth S. Berenhaut)

In *Individual strategy and social structure* (2001), Young demonstrated that the stochastically stable configurations of his *segregation game* are precisely those that are segregated. This paper extends the work of Young to configurations involving three types of individuals. We show that the stochastically stable configurations in this more general setting are again precisely those that are segregated.

Schelling [1971] investigated self-organizing systems consisting of two groups of individuals, two of whom could trade locations at each discrete time interval to improve at least one's contentment level without diminishing the other's. He identified the equilibria of these systems under various conditions. Most of the time, these equilibria were more segregated in the sense that the individual members of each of the groups tended to gather in larger clusters rather than be uniformly mixed. Young [2001] used a Markov chain model to identify the stochastically stable equilibria of these self-organizing systems with two groups of individuals.

By an *equilibrium* we mean a state in which no pair of individuals exist who would prefer to trade positions. These equilibria are stable in sense that once one is reached, there will be no further change in the system.

However, if we allow for the possibility of error, that is, trades of pairs of individuals which do not benefit at least one of the two, without harming the other, it is possible to move from some equilibria to others. Those equilibria which remain stable in this more general context are called *stochastically stable* equilibria. They are precisely the segregated equilibria, those with all of the individuals of a group gathered into a single cluster.

After seeing this behavior modeled in a classroom activity, a student asked the faculty author of this paper whether the same phenomena happened if there were more than two types of individuals. Responding to that question, in [Burek et al. 2009] we showed that there are both segregated (all members of each group living next to each other in a single cluster) and non-segregated equilibria in such a model,

---

MSC2010: 60J10, 91A15, 91A22, 91C99.

Keywords: segregation game, Schelling, markov process.

consistent with the work of Schelling. In this paper, we will show that the segregated equilibria are the only stochastically stable equilibria, consistent with the work of Young.

A real world example of this type of self-organizing behavior was provided in 2004 when Bill Bishop received national attention when he made the following claim and coined the neologism the *big sort*: the phenomenon that Americans have been sorting themselves into increasingly homogeneous political communities according to city and even neighborhood. He published his argument in [Bishop 2009] using demographical data to justify his claims. Therefore, in recognition of Bishop’s work, we will refer to our three groups of people as Republicans, Democrats, and Libertarians.

### Terminology

Let R, D, and L represent a individual that is a Republican, Democrat, and Libertarian, respectively. A *configuration* is an linear arrangement of individuals members that contains at least four members from each party, with an explanation for this restriction being given later. In general, let  $r$ ,  $d$ , and  $l$  represent the total number of individuals in each of the Republican, Democrat, and Libertarian parties, respectively. We assume that our configurations are circular in the sense that the first and last individuals are assumed to be neighbors of each other; this allows us to not worry about end conditions. For instance, in the following configuration, the leftmost R is considered to be next to the rightmost L:

$RDLLLLLLLLLLLLLLRRRRRRDRDLDDLLLLRRRRRRRRRRL$ .

We consider the positions of the Republicans, Democrats, and Libertarians to be ordered in this configuration. Thus, the configuration above is distinct from the one obtained by shifting each individual nine positions to the right, displayed here:

$RRRRRRRRRRLDLLLLLLLLLLLLLLRRRRRRDRDLDDLLLL$ .

To avoid unnecessary repetition, we use exponential notation and define a cluster of  $Y^m$  to be a string of  $m$   $Y$ ’s in a row, where  $2 \leq m \leq q$  where  $q$  denotes the total number of members in  $Y$ ’s party. Thus, the first configuration displayed above can be somewhat more compactly conveyed as

$$RDL^{11}R^6DRDL D^3L^4R^{10}L.$$

While the positions are distinct, the individuals themselves are not distinguished beyond their party affiliation.

Given any configuration, we need to determine an individual’s contentment level. Measuring contentment was straightforward in [Young 2001] since Young only



considered two types of individuals: either you're next to at least one individual like yourself (and are content) or you are not (and are therefore not content). Introducing a third group adds a layer of complexity in the form of bias: which individuals (aside from those of your own party) do you prefer to be next to, which are you neutral towards, and which do you prefer not to be next to at all? In [Burek et al. 2009], we describe seven different scenarios with varying levels of bias. In this paper, our focus is on individuals who have no aversion towards individuals of either of the other two parties, but do have a preference for neighbors of their own party.

We can describe this low level of bias as follows, since we do not need to specify the utility functions for our purposes. Let  $X$ ,  $Y$ , and  $Z$  be arbitrary individuals, not necessarily of distinct parties. Given an individual  $Y$  in a configuration, consider the ordered triple consisting of  $Y$  and its immediate neighbors to the left and the right,  $X$  and  $Z$ , respectively.  $Y$  has the highest contentment if both  $X$  and  $Z$  are of the same party as  $Y$ .  $Y$  has a somewhat lower contentment level if exactly one of  $X$  and  $Z$  is of the same party as  $Y$ . Finally  $Y$  has the lowest contentment level if neither  $X$  nor  $Z$  is of the same party as  $Y$ . For example, in the configuration

$$RDL^{11}R^6DRDL D^3L^4R^{10}L,$$

the first  $D$  individual has the lowest contentment level and the second to last  $D$  has the highest contentment level. More than three levels of contentment would be possible were we to allow higher levels of bias, as described in [Burek et al. 2009].

Two individuals in a configuration are willing to trade positions if at least one of the individual's contentment level increases as a result of this trade, and the other individual's contentment level does not decrease as a result of the trade. We call this a *favorable trade*.

Notice that when two individuals trade positions, it moves us from the original configuration  $s$  to a new configuration  $s'$ . When we move forward to a new time period, a pair of individuals are randomly chosen from among those pairs for whom a favorable trade exists and these two individuals trade positions. Eventually, no favorable trades remain and the system reaches an *equilibrium configuration*. Some of these are *segregated equilibrium configurations*, in the form

$$R^r D^d L^l \quad \text{or} \quad R^r L^l D^d.$$

Segregated equilibrium configurations could start with Democrats or Libertarians as well. In particular, note that the configurations

$$L^l R^r D^d \quad \text{and} \quad D^d L^l R^r$$

can be obtained from the first segregated equilibrium configuration above by shifted positions to the right, but they can not be obtained from the second segregated equilibrium above. Thus there are two fundamental classes of segregated equilibria, those of the form  $RDL$  and those of the form  $RLD$ .

Other equilibria are *non-segregated*. In a non-segregated equilibrium, the members of at least one party are separated into two disjoint clusters, each of which contains at least two members. Some examples (with  $r = 6$ ,  $d = 10$ , and  $l = 8$ ) are

$$D^8 R^3 L^2 D^2 L^4 R^3 L^2, \quad R^3 D^3 L^8 D^5 R^5, \quad \text{and} \quad L^2 D^{10} L^6 R^6.$$

If there are only three members of a party, then they must be in a single cluster in every equilibria. Thus if there are only three members of each of the three parties, there are no non-segregated equilibria. Thus to ensure that we have non-segregated equilibria, we require that there be at least four individuals in each party.

We denote the set of all equilibrium configurations by  $E$ , the set of those equilibria that are segregated by  $E^S$  and those equilibria that are non-segregated by  $E^{NS}$ . Thus,  $E = E^S \cup E^{NS}$ .

In our discussion so far, we have only allowed favorable trades to occur. To investigate stochastically stable configurations, we need to allow the possibility of non-favorable trades to occur as well. We define three types of such trades. Let  $a$ ,  $b$ , and  $c$  denote positive real numbers such that  $0 < a < b < c$ . A *type a perturbation* occurs when two individuals trade with one individual's contentment level rising and the other's falling, or when two individuals trade with neither individual's contentment level changing. A *type b perturbation* occurs when two individuals trade positions such that one individual's contentment level decreases, but the other individual's contentment level remains constant. Finally, a *type c perturbation* occurs when two individuals trade positions such that both of the individuals' contentment levels go down.

### Markov chain model

Both the basic situation and the perturbed situation can be modeled as a Markov chain. In this section, we describe those models, identify their key properties, their relationship, and give the key theorem that we will use in our analysis. The reader interested in more detailed discussion of Markov chains should consult [Ghahramani 2005] or [Norris 1998] for an introduction to the subject, or [Ross 2000] for a more rigorous treatment.

We model the basic situation as a Markov chain,  $P$ , by letting the set of states,  $S$ , be the various configurations,  $s$ , of our neighborhoods. For each  $s$ , set  $p_{s,s'} = 0$  for any state  $s'$  such that there is no favorable trade which moves  $s$  to  $s'$ . For all other  $s'$ ,  $p_{s,s'} = k/n$ , where  $n$  is the number of favorable trades in  $s$  and  $k$  is the number

of favorable trades which move  $s$  to  $s'$ . As favorable trades occur, the system can be thought to randomly evolve over time, with at most one trade occurring during each time period.

If we allow the possibility of non-favorable trades as well, we can obtain a second Markov chain,  $P^\epsilon$ , in which a type  $a$  perturbation occurs with probability  $\epsilon^a$ , where  $1 > \epsilon > 0$ . A type  $b$  perturbation occurs with probability  $\epsilon^b$ , and a type  $c$  perturbation occurs with probability  $\epsilon^c$ . Favorable trades occur with equal probabilities which sum to  $1 - \sum pr(x)$ , where  $x$  ranges across all of the non-favorable trades. Thus  $P^\epsilon$  has the same state space as  $P$ , and

$$p_{s,s'} = \sum pr(y) + \sum pr(x),$$

where  $y$  ranges across all favorable trades moving  $s$  to  $s'$  and  $x$  ranges across all non-favorable trades doing the same.

We can say the following about  $P$  and  $P^\epsilon$ :

- (1) The absorbing states of  $P$  are precisely the equilibrium states.
- (2)  $P^\epsilon$  is irreducible.
- (3)  $P^\epsilon$  has a unique stationary distribution,  $\mu^\epsilon$ .
- (4)  $P^\epsilon$  satisfies  $\lim_{\epsilon \rightarrow 0} p_{s,s'}^\epsilon = p_{s,s'}$ , and there exists a unique  $r(s, s') > 0$  such that whenever  $p_{s,s'}^\epsilon > 0$  for some  $\epsilon > 0$ ,

$$0 < \lim_{\epsilon \rightarrow 0} \frac{p_{s,s'}^\epsilon}{\epsilon^{r(s,s')}} < \infty.$$

- (5)  $P^\epsilon$  is regular perturbed.

Briefly, these five items are justified as follows. In any non-equilibrium configuration, there are a finite number of favorable trades; as time advances and these trades happen, they are eventually depleted resulting in a configuration that is at equilibrium and is an absorbing state of  $P$ . Because all trades (favorable and non-favorable) have positive probability in  $P^\epsilon$ , there exists a positive probability that of moving from any configuration to any other configuration in the future. Hence  $P^\epsilon$  is irreducible. Further, since  $P^\epsilon$  has a finite state space, it has a unique stationary distribution. The first limit in item four follows from our definition of  $p_{s,s'}^\epsilon$ . The second limit follows from our assignments of probabilities to the various non-favorable trades. Finally, item five follows from items two and four.

In general,  $r(s, s')$  is called the resistance to moving from state  $s$  to state  $s'$ , and is the minimum, taken over all sequences of trades that begin in state  $s$  and end in state  $s'$ , of the sum of the resistances on the individual trades in the sequence. The values  $a$ ,  $b$ , and  $c$  are the resistance to the corresponding types of non-favorable trades. A favorable trade has resistance 0.

We now construct a graph theoretic model to compute the stochastically stable states of  $P^\epsilon$ . Recall that the only absorbing states of  $P$  are the equilibrium states in  $S$ . Denote these by  $E = \{z_1, z_2, z_3, \dots\}$ . Construct a weighted complete directed graph whose vertices are the elements of  $E$  and whose edges have weights equal to the resistances  $r(z_i, z_j)$ . A  $z$ -tree is a set of  $|E| - 1$  directed edges such that, from every vertex different from  $z \in E$ , there is a unique directed path in the tree to  $z$ . The resistance of a  $z$ -tree is the sum of the resistances on the edges that compose it. The *stochastic potential* of the state  $z$  is the minimum resistance over all  $z$ -trees.

Figure 1 illustrates one such tree. In this illustration,  $z$  is an *RLD* segregated equilibria, and each *RLD* and *RDL* vertex represents a one position shift from its parent vertex. The *ns* vertices represent generic non-segregated equilibria. The choice of edge weights,  $a$ ,  $b$ , and  $a + b$ , will be explained after Theorem 1.

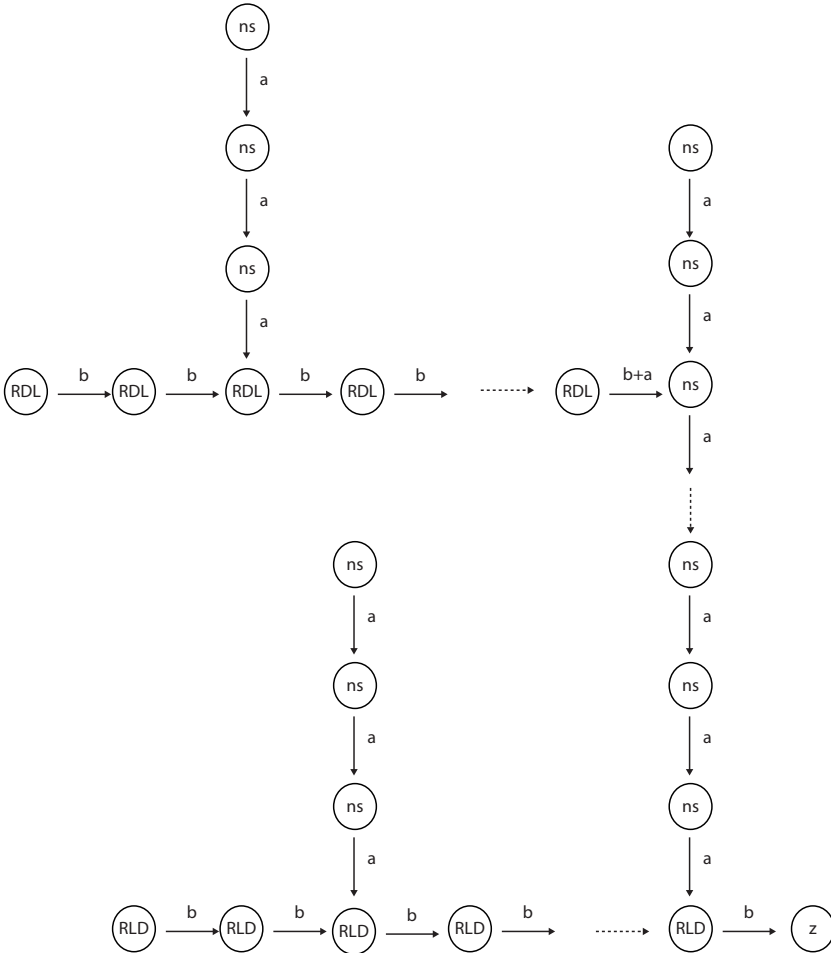


Figure 1. A  $z$ -tree for a *RLD* segregated equilibrium.

The stochastically stable states are those states that occur with positive probability in the long run while the probability of error,  $\epsilon$ , is small but non-vanishing. That is, the state  $s \in E$  is stochastically stable for the Markov chain  $P^\epsilon$  if

$$\lim_{\epsilon \rightarrow 0} \mu_s^\epsilon > 0,$$

where  $\mu^\epsilon$  is the unique stationary distribution of  $P^\epsilon$ . Young's theorem provides a method for determining these states, which is the goal of this paper.

**Young's theorem.** *Let  $P^\epsilon$  be a regular perturbed Markov chain and let  $\mu^\epsilon$  be the unique stationary distribution of  $P^\epsilon$  for each  $\epsilon > 0$ . Then the stochastically stable states are precisely those states that are absorbing states of  $P$  having minimum stochastic potential [Young 1993].*

### Main result

In this section, we construct  $z$ -trees for both segregated and non-segregated equilibria and demonstrate that the former have minimal stochastic potential. We begin by proving three lemmas which will develop our argument.

**Lemma 1.** *Given a non-segregated equilibrium, the resistance to moving to another equilibrium by making a trade which moves an individual from one cluster to another cluster of like individuals is  $a$ .*

*Proof.* Given a non-segregated equilibrium state, suppose that one party, say the  $R$ s, has at least two clusters. Then at least two of the  $R$  clusters have neighbor clusters of the same type, say  $L$ . Otherwise, there are exactly two  $R$  clusters, one with two  $D$  clusters as neighbors and the other with two  $L$  cluster neighbors. (The pattern is  $D - R - D - L - R - L$ .) In this case, we change our perspective to the two  $D$  clusters, which have a common  $R$  cluster as a neighbor.

There are three patterns possible for the two  $R$  clusters and their  $L$  cluster neighbors:

$$\begin{aligned} &L^{l_1-1} \mathbf{L}R^{r_1} \dots L^{l_2} \mathbf{R}R^{r_2-1}, \\ &L^{l_1-1} \mathbf{L}R^{r_1} \dots R^{r_2-1} \mathbf{R}L^{l_2}, \\ &R^{r_1-1} \mathbf{R}L^{l_1} \dots L^{l_2-1} \mathbf{L}R^{r_2}. \end{aligned}$$

In each case, trading the bold faced individuals shifts one individual from one cluster to another and results in a new equilibrium state. Each of these trades has resistance  $a$ . □

**Lemma 2.** *Given any segregated equilibrium, the minimum resistance to shifting to another segregated equilibrium is  $b$ .*

*Proof.* Consider the segregated equilibrium

$$R^{r-1} \mathbf{R} \mathbf{D}^d L^{l-1} \mathbf{L}.$$

We trade the boldfaced individuals, with resistance  $b$ , to get

$$R^{r-1} \mathbf{L} \mathbf{D}^d L^{l-1} \mathbf{R}.$$

However, this configuration is not an equilibrium. Therefore, we need to make a favorable trade, which has resistance 0, to return to equilibrium. The two individuals involved in this trade are indicated in bold:

$$R^{r-1} \mathbf{L} \mathbf{D}^{d-1} \mathbf{D} L^{l-1} \mathbf{R}.$$

Trading these two individuals results in the segregated configuration:

$$R^{r-1} \mathbf{D} \mathbf{D}^{d-1} \mathbf{L} L^{l-1} \mathbf{R}.$$

Note that the new configuration is the the original equilibrium configuration shifted one position to the left.

To obtain a smaller resistance, either one of the individuals trading had an increase in their contentment level while the second had a decrease, or neither of the individuals trading had any change in contentment level. However, when we begin with a segregated equilibrium, both cases imply that any individual who trades must trade with another individual of the same party, and that results in the same equilibrium after the trade as before. Thus it is not possible to shift from one segregated equilibrium to another with a resistance less than  $b$ .  $\square$

**Lemma 3.** *Given any equilibrium, the minimum resistance to creating a new cluster is  $b + a$ .*

*Proof.* Without loss of generality, consider an equilibrium containing the sequence

$$\dots L^{l_1-1} \mathbf{L} \mathbf{D}^{d_1-1} \mathbf{D} R^{r_1} \dots$$

In a trade between the bold  $L$  and the bold  $D$ ,  $L$ 's contentment level would drop, while  $D$ 's contentment level stays the same, resulting in a trade with resistance  $b$ . The resulting configuration,

$$\dots L^{l_1-1} \mathbf{D} \mathbf{D}^{d_1-1} \mathbf{L} R^{r_1} \dots$$

is not in equilibrium, so a trade between the second right-most  $L$  with the (new) right-most  $D$ , with resistance  $a$ , results in an equilibrium with an additional cluster of consisting of two  $L$ 's. This is the smallest resistance possible, since creating a new cluster requires isolating an individual and consequently lowering their contentment level, a type  $b$  perturbation. To return to an equilibrium state with this

new cluster, another trade must occur in order to have a second individual join the first. At best this is a type  $a$  perturbation.  $\square$

With these three lemmas in hand, we are ready to compute the minimum stochastic potential for the  $z$ -trees. In the proof of [Lemma 2](#), we assumed that the segregated configuration has the ordering  $RDL$  of clusters. The alternative ordering is  $RLD$ . Clearly the Lemma applies to this ordering as well. However, in both [Theorem 1](#) and [Theorem 2](#) below, we do need to treat the two orderings separately, so we let  $E_{RDL}^S$  and  $E_{RLD}^S$  denote the two sets of segregated equilibriums, respectively. We begin with  $z \in E^S$ .

**Theorem 1.** *For each  $z \in E^S$ , its stochastic potential is*

$$a \cdot |E^{NS}| + b \cdot (|E_{RDL}^S| - 1) + b \cdot (|E_{RLD}^S| - 1) + (b + a) = a \cdot |E^{NS}| + b \cdot (|E^S| - 1) + a.$$

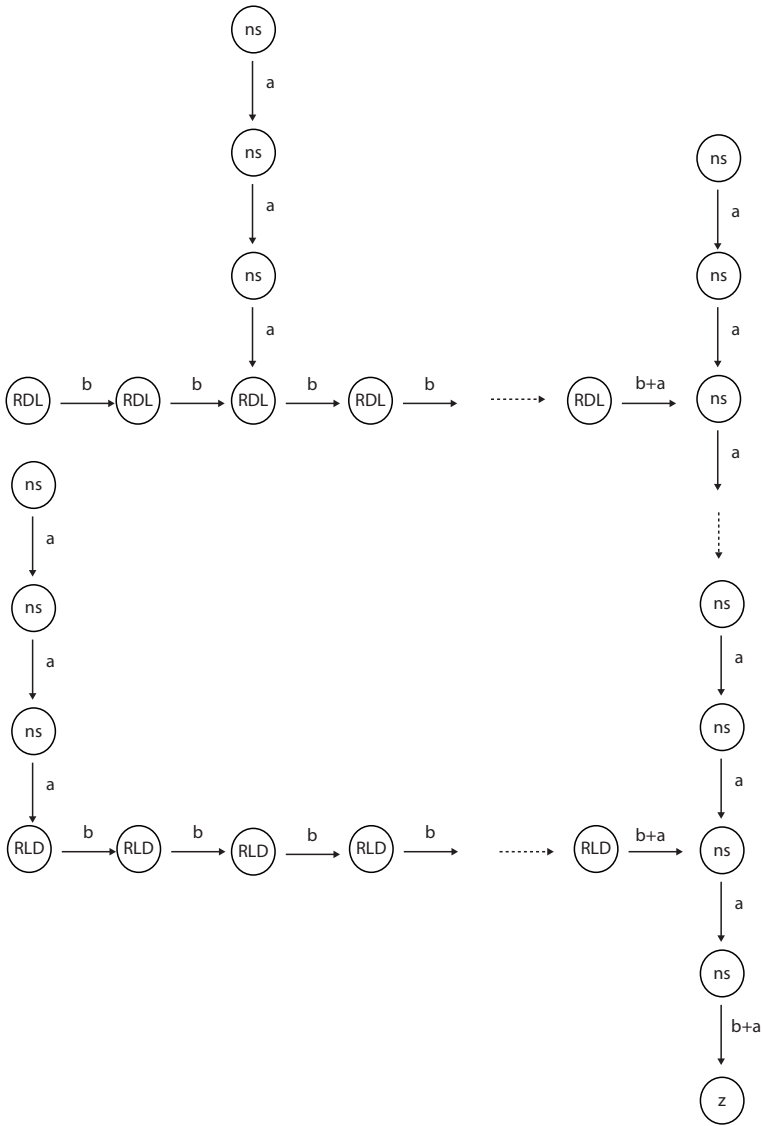
*Proof.* We will assume that  $z$  is an  $RLD$  type of segregated equilibrium. Each non-segregated equilibrium has an outbound edge to another equilibrium in which one of the clusters has one fewer individuals. By [Lemma 1](#), this edge has resistance (weight)  $a$ . All but two of the segregated equilibriums have an outbound edge to another segregated equilibrium, which rotates the positions of the individuals by one position. By [Lemma 2](#), each of these edges has resistance  $b$ . The first exception to the previous statement is the root equilibrium,  $z$ , which has no outbound edge associated with it. The second exception is the  $RDL$  equilibrium at which a new cluster is generated in order to begin the transition to an  $RLD$  equilibrium. By [Lemma 3](#), this particular equilibrium has an outbound edge that has resistance  $b + a$ . Summing the resistances on the various edges gives the result.  $\square$

[Figure 1](#) illustrates the proof for a typical  $z$ -tree, when  $z$  is a segregated equilibrium. The target  $RLD$  equilibrium is in the lower right corner, and the transitional  $RDL$  equilibrium has a resistance of  $b + a$ . In this illustration, each segregated equilibrium is rotated until it reaches  $z$ , or until the transitional configuration is reached. Each non-segregated state progressively moves to states with smaller and/or fewer clusters, eventually becoming segregated.

Next, we compute the minimum stochastic potential for an arbitrary  $z$ -tree where  $z$  is in  $E^{NS}$ . Notice that in [Theorem 1](#), we were able to calculate the minimum stochastic resistance precisely. In the following theorem, we are only able to determine a lower bound. This is because it may be required to create many new clusters, with the creation of each of these clusters increasing the sum given in the theorem. Fortunately, the result is sufficient for our purposes.

**Theorem 2.** *For each  $z \in E^{NS}$ , its stochastic potential is at least*

$$\begin{aligned} a \cdot (|E^{NS}| - 1) + b \cdot (|E_{RDL}^S| - 1) + (b + a) + b \cdot (|E_{RLD}^S| - 1) + (b + a) \\ = a \cdot |E^{NS}| + b \cdot (|E^S| - 1) + a + b. \end{aligned}$$



**Figure 2.** Minimal  $z$ -tree for a non-segregated equilibrium.

*Proof.* We will assume that  $z$  has only four clusters, the minimum possible in a non-segregated equilibrium. Each non-segregated equilibrium, other than  $z$ , has an outbound edge to another equilibrium in which one of the clusters has one fewer individuals. By Lemma 1, this edge has resistance  $a$ . All but two of the segregated equilibria has an outbound edge to another segregated equilibrium rotating the positions of the individuals by one position. By Lemma 2, each of these edges has resistance  $b$ . The two exceptions to the previous are the  $RLD$  equilibrium



and the *RDL* equilibrium at which new clusters are created; by [Lemma 3](#), these two equilibria have outbound edges with resistance  $b + a$ . Summing the resistance on the various edges gives the result.  $\square$

[Figure 2](#) illustrates the proof for a  $z$ -tree in which  $z$  is a non-segregated equilibrium. Again, the target non-segregated equilibrium is in the lower right corner, and the transitional *RLD* and *RDL* equilibria have resistance  $b + a$ .

Since the sum in [Theorem 1](#) is smaller than the sum in [Theorem 2](#), we are able to state our main result.

**Theorem 3.** *In segregation games with three types of individuals and the lowest level of bias, the stochastically stable equilibria are precisely those that are segregated.*

### Open questions

The model described in this paper assumes that no individuals have a bias against members of one of the other groups. In [\[Burek et al. 2009\]](#), we outline six other scenarios describing varying biases that are available among three groups. For example, would we get the same results in a scenario where Republicans and Democrats each prefer to live near Libertarians over each other, but Libertarians hold no such bias? What if Democrats prefer Republicans, Republicans prefer Libertarians, and Libertarians prefer Republicans? Demonstrating stochastic results similar to those presented in this paper would extend our model.

Furthermore, it would be interesting to extend the analysis in this paper to a 2-dimensional perspective. Doing so would allow for a more realistic geo-political interpretation of the results, such as that suggested by Bishop's work.

### References

- [Bishop 2009] B. Bishop, *The big sort: why the clustering of like-minded America is tearing us apart*, 1st ed., Houghton Mifflin Harcourt, Boston, MA, 2009.
- [Burek et al. 2009] M. Burek, M. McDonough, and S. Roach, "Isolation and contentment in segregation games with three types", *Ball State University Mathematics Exchange* **6**:1 (2009), 19–28.
- [Ghahramani 2005] S. Ghahramani, *Fundamentals of probability with stochastic processes*, 3rd ed., Pearson/Prentice Hall, Upper Saddle River, New Jersey, 2005.
- [Norris 1998] J. R. Norris, *Markov chains*, Reprint of 1997 ed., Cambridge Series in Statistical and Probabilistic Mathematics **2**, Cambridge University Press, Cambridge, 1998. [MR 99c:60144](#) [Zbl 0938.60058](#)
- [Ross 2000] S. M. Ross, *Introduction to probability models*, 7th ed., Harcourt/Academic Press, Burlington, MA, 2000. [MR 2001b:60002](#) [Zbl 0977.60001](#)
- [Schelling 1971] T. Schelling, "Dynamic models of segregation", *Journal of Mathematical Sociology* **1** (1971), 143–186.

[Young 1993] H. P. Young, “The evolution of conventions”, *Econometrica* **61**:1 (1993), 57–84.  
[MR 93j:92038](#) [Zbl 0773.90101](#)

[Young 2001] P. H. Young, *Individual strategy and social structure*, Princeton University Press, 2001.

Received: 2012-10-01    Revised: 2013-10-21    Accepted: 2014-02-25

[michael.borchert@valpo.edu](mailto:michael.borchert@valpo.edu)    *Valparaiso University, 1700 Chapel Drive,  
Valparaiso, IN 46383, United States*

[mark.burek@valpo.edu](mailto:mark.burek@valpo.edu)    *Valparaiso University, 1700 Chapel Drive,  
Valparaiso, IN 46383, United States*

[rick.gillman@valpo.edu](mailto:rick.gillman@valpo.edu)    *Valparaiso University, 1700 Chapel Drive,  
Valparaiso, IN 46383, United States*

[spencer.roach@valpo.edu](mailto:spencer.roach@valpo.edu)    *Valparaiso University, 1700 Chapel Drive,  
Valparaiso, IN 46383, United States*

# Embedding groups into distributive subsets of the monoid of binary operations

Gregory Mezera

(Communicated by Kenneth S. Berenhaut)

Let  $X$  be a set and  $\text{Bin}(X)$  the set of all binary operations on  $X$ . We say that  $S \subset \text{Bin}(X)$  is a distributive set of operations if all pairs of elements  $*_\alpha, *_\beta \in S$  are right distributive, that is,  $(a *_\alpha b) *_\beta c = (a *_\beta c) *_\alpha (b *_\beta c)$  (we allow  $*_\alpha = *_\beta$ ).

The question of which groups can be realized as distributive sets was asked by J. Przytycki. The initial guess that embedding into  $\text{Bin}(X)$  for some  $X$  holds for any  $G$  was complicated by an observation that if  $* \in S$  is idempotent ( $a * a = a$ ), then  $*$  commutes with every element of  $S$ . The first noncommutative subgroup of  $\text{Bin}(X)$  (the group  $S_3$ ) was found in October 2011 by Y. Berman.

Here we show that any group can be embedded in  $\text{Bin}(X)$  for  $X = G$  (as a set). We also discuss minimality of embeddings observing, in particular, that  $X$  with six elements is the smallest set such that  $\text{Bin}(X)$  contains a nonabelian subgroup.

1. Introduction	433
2. Regular distributive embedding	435
3. General conditions for a distributive embedding	435
4. Future directions; multiterm homology	436
Acknowledgements	437
References	437

## 1. Introduction

Let  $X$  be a set and  $\text{Bin}(X)$  the set of all distributive operations on  $X$ . We say that  $S \subset \text{Bin}(X)$  is a distributive set of operations if all pairs of elements  $*_\alpha, *_\beta \in S$  are right distributive, that is,  $(a *_\alpha b) *_\beta c = (a *_\beta c) *_\alpha (b *_\beta c)$  (we allow  $*_\alpha = *_\beta$ ). It was observed in [Przytycki 2011] (see also [Romanowska and Smith 1985]) that  $\text{Bin}(X)$  is a monoid with composition  $*_1 *_2$  given by  $a *_1 *_2 b = (a *_1 b) *_2 b$  and the identity  $*_0$  being the right trivial operation, that is,  $a *_0 b = a$  for any  $a, b \in X$ .

*MSC2010:* primary 55N35; secondary 18G60, 57M25.

*Keywords:* monoid of binary operations, distributive set, shelf, multishelf, distributive homology, embedding, group.

The submonoid of  $\text{Bin}(X)$  of all invertible elements in  $\text{Bin}(X)$  is a group denoted by  $\text{Bin}_{\text{inv}}(X)$ . If  $* \in \text{Bin}_{\text{inv}}(X)$  then  $*^{-1}$  is usually denoted by  $\bar{*}$ .

We say that a subset  $S \subset \text{Bin}(X)$  is a distributive set if all pairs of elements  $*_\alpha, *_\beta \in S$  are right distributive, that is,  $(a *_\alpha b) *_\beta c = (a *_\beta c) *_\alpha (b *_\beta c)$  (we allow  $*_\alpha = *_\beta$ ). Additionally,  $(X; S)$  is called a multishef<sup>1</sup>.

The following important basic lemma was proven in [Przytycki 2011]:

- Lemma 1.1.** (i) *If  $S$  is a distributive set and  $* \in S$  is invertible, then  $S \cup \{\bar{*}\}$  is also a distributive set.*
- (ii) *If  $S$  is a distributive set and  $M(S)$  is the monoid generated by  $S$ , then  $M(S)$  is a distributive monoid.*
- (iii) *If  $S$  is a distributive set of invertible operations and  $G(S)$  is the group generated by  $S$ , then  $G(S)$  is a distributive group.*

The question of which groups can be realized as distributive sets was asked by J. Przytycki. Soon after the definition of a distributive submonoid of  $\text{Bin}(X)$  was given in [Przytycki 2011], Michał Jabłonowski, a graduate student at Gdańsk University, noticed that any distributive monoid whose elements are idempotent operations is commutative.

**Proposition 1.2** [Przytycki 2011]. *Consider  $*_\alpha, *_\beta \in \text{Bin}(X)$  such that  $*_\beta$  is idempotent ( $a *_\beta a = a$ ) and distributive with respect to  $*_\alpha$ . Then  $*_\alpha$  and  $*_\beta$  commute. In particular:*

- (i) *If  $M$  is a distributive monoid and  $*_\beta \in M$  is an idempotent operation, then  $*_\beta$  is in the center of  $M$ .*
- (ii) *A distributive monoid whose elements are idempotent operations is commutative.*

*Proof.* We have  $(a *_\alpha b) *_\beta b \stackrel{\text{distrib}}{=} (a *_\beta b) *_\alpha (b *_\beta b) \stackrel{\text{idemp}}{=} (a *_\beta b) *_\alpha b. \quad \square$

A few months later, Agata Jastrzębska (also a graduate student at Gdańsk University) checked that any distributive group in  $\text{Bin}_{\text{inv}}(X)$  for  $|X| \leq 5$  is commutative.

The first noncommutative subgroup of  $\text{Bin}(X)$  (the group  $S_3$ ) was found in October 2011 by Yosef Berman. Soon after, Berman and Carl Hammarsten constructed an embedding of a general dihedral group  $D_{2\cdot n}$  in  $\text{Bin}(X)$  where  $X$  has  $2n$  elements. The embedding of Berman,  $\phi : D_{2\cdot 3} \rightarrow \text{Bin}(X)$ , is given as follows: if  $X = \{0, 1, 2, 3, 4, 5\}$  then the subgroup  $D_{2\cdot 3} \subset \text{Bin}(X)$  is generated by binary

---

<sup>1</sup>If  $(X; *)$  is a magma and  $*$  is a right self-distributive operation then  $(X; *)$  is called a shelf, the term coined by Alissa Crans [2004].

operations  $*_\tau$ , which generates reflection, and  $*_\sigma$ , which generates a 3-cycle;

$$*_\tau = \begin{pmatrix} 1 & 1 & 3 & 5 & 5 & 3 \\ 0 & 0 & 4 & 2 & 2 & 4 \\ 3 & 3 & 5 & 1 & 1 & 5 \\ 2 & 2 & 0 & 4 & 4 & 0 \\ 5 & 5 & 1 & 3 & 3 & 1 \\ 4 & 4 & 2 & 0 & 0 & 2 \end{pmatrix} \quad \text{and} \quad *_\sigma = \begin{pmatrix} 2 & 4 & 2 & 4 & 2 & 4 \\ 5 & 3 & 5 & 3 & 5 & 3 \\ 4 & 0 & 4 & 0 & 4 & 0 \\ 1 & 5 & 1 & 5 & 1 & 5 \\ 0 & 2 & 0 & 2 & 0 & 2 \\ 3 & 1 & 3 & 1 & 3 & 1 \end{pmatrix},$$

where  $i * j$  is placed in the  $i$ -th row and  $j$ -th column, and  $D_{2,3} = \{\tau, \sigma \mid \tau\sigma\tau = \sigma^{-1}\}$ .

### 2. Regular distributive embedding

We now show that any group  $G$  can be embedded in  $\text{Bin}(X)$  for some  $X$ .

**Theorem 2.1** (Regular embedding). *Every group  $G$  embeds in  $\text{Bin}(G)$ . This embedding (monomorphism),  $\phi^{\text{reg}} : G \rightarrow \text{Bin}(G)$ , sends  $g$  to  $*_g$ , where  $a *_g b = ab^{-1}gb$ .*

*Proof.* (i) We check that the set  $\{*_g\}_{g \in G}$  is a distributive set. We have

$$(a *_g b) *_g c = (ab^{-1}g_1b) *_g c = ab^{-1}g_1bc^{-1}g_2c,$$

and

$$(a *_g c) *_g (b *_g c) = (ac^{-1}g_2c) *_g (bc^{-1}g_2c) = ab^{-1}g_1bc^{-1}g_2c,$$

as needed.

(ii) Now we check that the map  $\phi^{\text{reg}}$  is a monomorphism. The image of the identity  $*_0$  is the identity in  $\text{Bin}(G)$ . Furthermore,  $a *_g b = ab^{-1}g_1g_2b$  and  $a *_g *_g b = (a *_g b) *_g b = ab^{-1}g_1bb^{-1}g_2b = ab^{-1}g_1g_2b$ , as needed. We have proven that  $\phi^{\text{reg}}$  is a homomorphism. To show that  $\phi^{\text{reg}}$  is a monomorphism, we substitute  $b = 1$  in the formula for  $a *_g b$  to get  $a *_g 1 = ag$ ; so different choices of  $g$  give different binary operations in  $\text{Bin}(G)$ . Notice that  $\phi^{\text{reg}}(g^{-1}) = \bar{*}_g$ .  $\square$

We call our embedding *regular*, analogous to the regular representation of a group. We do not claim that the regular embedding is minimal, so finding minimal distributive embeddings is a very interesting problem in itself.

### 3. General conditions for a distributive embedding

We now discuss a method that can be used to embed groups into subsets of  $\text{Bin}_{\text{inv}}(X)$  satisfying an arbitrary condition. We then use this method when the condition is right distributivity, which leads us to the regular distributive embedding of  $G$  in  $\text{Bin}(G)$  and should be a natural tool to look for minimal embeddings. For the group  $S_3$ , we know, by Jastrzebska’s calculations, that  $X$  consisting of six elements is the minimal set such that  $S_3$  embeds in  $\text{Bin}(X)$ .

We start from the following basic observation:

**Lemma 3.1.** *There is an isomorphism between  $\text{Bin}_{\text{inv}}(X)$  and  $S_{|X|}^{|X|}$ , where  $|X|$  is the cardinality of  $X$  and  $S_{|X|}$  is the group of permutations on set  $X$  ( i.e., bijections of the set  $X$ ). The isomorphism  $\alpha : \text{Bin}_{\text{inv}}(X) \rightarrow S_X^{|X|} = \prod_{y \in X} S_X^y$  is described as follows:  $\alpha(*) (y) : X \rightarrow X$  is the bijection where  $(\alpha(*) (y))(x) = x * y$ . In other words,  $\alpha(*) (y)$  is the bijection corresponding to the  $y$ -coordinate of  $S_X^{|X|}$ .*

Using the map  $\alpha$ , we can translate conditions on a set of binary operations in  $\text{Bin}(X)$  into a group-theoretic condition on (coordinates of) elements of  $S_X^{|X|}$ . With some work, we can use this to find an embedding of a group into  $\text{Bin}(X)$ . This is possible since the group axioms require that such an embedding must sit inside  $\text{Bin}_{\text{inv}}(X)$ . Let us consider distributive, invertible sets  $\mathcal{S}$  of binary operations in  $\text{Bin}_{\text{inv}}(X)$ . These are subsets  $\mathcal{S} \subseteq \text{Bin}_{\text{inv}}(X)$  that satisfy

$$(x *_i y) *_j z = (x *_j z) *_i (y *_j z) \quad \text{for all } *_i, *_j \in \mathcal{S} \text{ and } x, y, z \in X.$$

Let  $\sigma_{i,y} = p_y \alpha(*_i)$ , where  $p_y : S_X^{|X|} \rightarrow S_X$  is projection onto the  $y$ -th coordinate. Then translating the distributivity condition via  $\alpha$ ,

$$\sigma_{j,z}(x *_i y) = \sigma_{i,(y*_j z)}(x *_j z)$$

or

$$\sigma_{j,z}(\sigma_{i,y}(x)) = \sigma_{i,\sigma_{j,z}(y)}(\sigma_{j,z}(x)),$$

which leads to

$$\sigma_{i,\sigma_{j,z}(y)} = \sigma_{j,z} \sigma_{i,y} \sigma_{j,z}^{-1}.$$

Now the problem of embedding a group into  $\text{Bin}_{\text{inv}}(X)$  is reduced to finding subsets of  $S_{|X|}^{|X|}$  satisfying the condition above that are isomorphic to the group. We can then use tools of group theory (e.g., representation theory) to solve the problem. This process can be attempted for subsets of  $\text{Bin}_{\text{inv}}(X)$  satisfying any condition and leads to the embedding defined in the previous section for distributive subsets.

### 4. Future directions; multiterm homology

Przytycki [2011] defined multiterm homology for any distributive set. This provided motivation to have many examples of distributive sets. The regular embedding of a group (Theorem 2.1) provides an interesting family of distributive sets ripe for the study of their homology (compare with [Crans et al. 2014; Przytycki 2011; 2012; Przytycki and Putyra 2013; Przytycki and Sikora 2014]). As a nontrivial example, we propose computing  $n$ -term distributive homology related to the regular embedding of the cyclic group  $Z_n$ . Another problem related to Theorem 2.1 is determining which monoids are distributive submonoids of  $\text{Bin}(X)$ .

A key motivation is to use multiterm distributive homology in knot theory. This possibility arises from the relation of the third Reidemeister move with right distributivity (and eventually the Yang–Baxter operator) and the important work of Carter, Kamada, and Saito [2001] and other researchers on applications of quandle homology to knot theory.

### Acknowledgements

I was partially supported by the George Washington University Presidential Merit Fellowship.

I would like to thank Professor Józef Przytycki, Carl Hammarsten, and Krzysztof Putyra for helpful discussion, and Mieczysław Dąbkowski for his moral support.

### References

- [Carter et al. 2001] J. S. Carter, S. Kamada, and M. Saito, “Geometric interpretations of quandle homology”, *J. Knot Theory Ramifications* **10**:3 (2001), 345–386. MR 2002h:57009 Zbl 1002.57019
- [Crans 2004] A. S. Crans, *Lie 2-algebras*, Ph.D. thesis, University of California, Riverside, 2004, Available at <http://search.proquest.com/docview/305198326>. MR 2706291 Zbl 1057.17011
- [Crans et al. 2014] A. S. Crans, J. H. Przytycki, and K. K. Putyra, “Torsion in one-term distributive homology”, *Fund. Math.* **225** (2014), 75–94. MR 3205566 Zbl 06292117
- [Przytycki 2011] J. H. Przytycki, “Distributivity versus associativity in the homology theory of algebraic structures”, *Demonstratio Math.* **44**:4 (2011), 823–869. MR 2906433 Zbl 1286.55004
- [Przytycki 2012] J. H. Przytycki, *Teoria węzłów i związanych z nimi struktur dystrybutywnych*, University of Gdańsk Press, 2012.
- [Przytycki and Putyra 2013] J. H. Przytycki and K. K. Putyra, “Homology of distributive lattices”, *J. Homotopy Relat. Struct.* **8**:1 (2013), 35–65. MR 3031593 Zbl 1284.06016
- [Przytycki and Sikora 2014] J. H. Przytycki and A. S. Sikora, “Distributive products and their homology”, *Comm. Algebra* **42**:3 (2014), 1258–1269. MR 3169627 Zbl 06288603
- [Romanowska and Smith 1985] A. B. Romanowska and J. D. H. Smith, *Modal theory: an algebraic approach to order, geometry, and convexity*, Research and Exposition in Mathematics **9**, Heldermann Verlag, Berlin, 1985. MR 86k:08001 Zbl 0553.08001

Received: 2012-10-31

Revised: 2014-03-20

Accepted: 2014-04-27

[gregmezera@yahoo.com](mailto:gregmezera@yahoo.com)

*Department of Mathematics, George Washington University,  
Washington, DC 20052, United States*





# Persistence: a digit problem

Stephanie Perez and Robert Styer

(Communicated by Kenneth S. Berenhaut)

We examine the persistence of a number, defined as the number of iterations of the function which multiplies the digits of a number until one reaches a single digit number. We give numerical evidence supporting Sloane's 1973 conjecture that there exists a maximum persistence for every base. In particular, we give evidence that the maximum persistence in each base 2 through 12 is 1, 3, 3, 6, 5, 8, 6, 7, 11, 13, 7, respectively.

## 1. Introduction

Neil J. A. Sloane [1973] considered the function that multiplies the digits of a number and formally conjectured that the number of iterates needed to reach a fixed point is bounded. In particular, in base 10, he conjectured that one needs at most 11 iterates to reach a single digit. The problem did arise earlier; see [Gottlieb 1969, Problems 28–28; Beeler et al. 1972].

**Definition 1.** Let  $n = \sum_{j=0}^r d_j B^j$ , with  $0 \leq d_j < B$  for each  $d_j$ , be the base  $B$  expansion of  $n$ . We define the digital product function as  $f(n) = \prod_{j=0}^r d_j$ .

The persistence of a number  $n$  is defined as the minimum number  $k$  of iterates  $f^k(n) = d$  needed to reach a single digit  $d$ .

**Theorem 1.** *If  $n \geq B$ , then  $n > f(n)$ . If  $0 \leq n < B$ , then  $f(n) = n$  is a fixed point. Thus, every  $n$  has a finite persistence.*

*Proof.* Let  $n = \sum_{j=0}^r d_j B^j$ , with  $0 \leq d_j < B$  for each  $d_j$  and  $r > 0$ . Since  $r > 0$ ,

$$n \geq d_r B^r > d_r \prod_{j=0}^{r-1} d_j = f(n).$$

If  $n < B$ , then clearly  $f(n) = n$ . So, by induction on  $n$  one can show that every  $n$  has a finite persistence.  $\square$

For the remainder of this section, assume the base  $B$  equals 10.

MSC2010: 00A08, 97A20.

Keywords: persistence, digit problem, multiplicative persistence, iterated digit functions.

persistence	least $n$ with given persistence	$\ln \ln n$
2	25	1.1690
3	39	1.2984
4	77	1.4688
5	679	1.8750
6	6788	2.1774
7	68889	2.4106
8	2677889	2.6947
9	26888999	2.8395
10	3778888999	3.0934
11	277777788888899	3.5043

**Table 1.** Smallest number with a given persistence.

**Example.** Let  $n = 23487$ . Then

$$f(23487) = 2 \cdot 3 \cdot 4 \cdot 8 \cdot 7 = 1344,$$

$$f(1344) = 1 \cdot 3 \cdot 4 \cdot 4 = 48,$$

$$f(48) = 4 \cdot 8 = 32,$$

and finally,  $f(32) = 3 \cdot 2 = 6$ . In other words,  $f^4(23487) = 6$ , so 23487 has persistence 4.

One easily sees that  $n = 23114871$ ,  $n = 642227$  and  $n = 78432$  also have persistence 4 since each of these has  $f(n) = 1344$ . Thus, adding or removing the digit 1 does not change the persistence, nor does rearranging the digits or replacing digits that are products of smaller digits by these smaller digits.

In particular, since 288888899777777 has persistence 11, so do

$$1288888899777777, \quad 11288888899777777 \quad \text{and} \quad 111288888899777777,$$

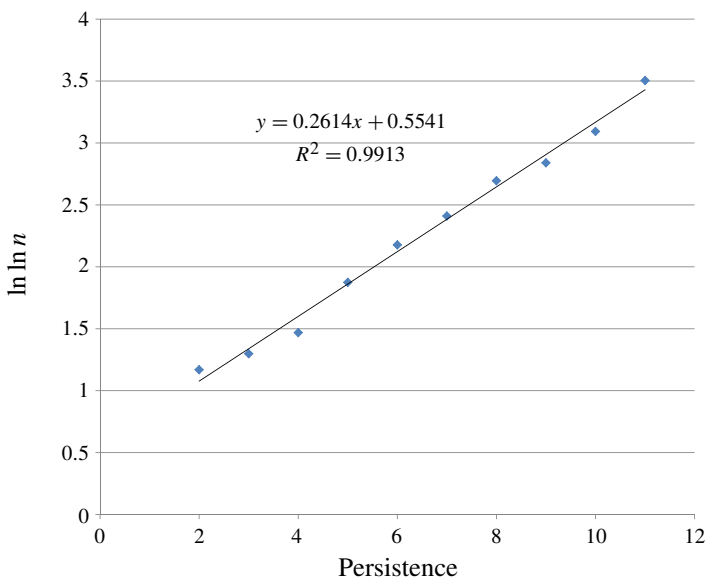
etc. Hence, there are an infinite number of integers with persistence 11.

We note some other immediate observations.

Let  $n = 543210$ . Then  $f(n) = 0$ , so it has persistence 1. More generally, any number with a 0 digit has persistence 1.

Let  $n = 54321$ . Then  $f(54321) = 120$ , so  $f^2(54321) = 0$ . More generally, in base 10, any number with a 5 digit, with an even digit, and with no 0 digit, has persistence 2.

Some preliminary calculations suggest that persistence depends on the size of the number. We list the smallest number with a given persistence (avoiding the contentious issue of defining the persistence of single digit numbers) in [Table 1](#).



**Figure 1.** The double logarithm of the smallest number with persistence  $p$  versus  $p$  seems linear.

Table 1 and Figure 1 might suggest that the persistence grows roughly as the double logarithm of the number; using a linear fit to the log-log of the data, one might expect to find a number of size about  $3 \cdot 10^{17}$  with persistence 12. Sloane [1973] showed, however, that no number less than  $10^{50}$  has persistence 12; this was extended by Carmody [2001] to  $10^{233}$ , and Diamond [2010] extended it to  $10^{333}$ , while we extend it to  $10^{1500}$ .

This paper has grown out of the senior research paper of the first author, intrigued by the mention of the problem in [Guy 2004, Problem F25].

## 2. Results

This section summarizes some results which give bounds for the persistence in various bases. We used Maple to calculate these results.

Since a large random number almost always has a 0 digit, we can prove the following theorem.

**Theorem 2.** *In any base  $B$ , the density of positive integers up to  $N$  with persistence greater than 1 approaches zero as  $N$  approaches infinity.*

*Proof.* Assume  $B > 2$ ; the next theorem deals with base  $B = 2$ .

Consider all numbers with  $k$  digits in base  $B$ , that is, all integers  $N$  with  $B^{k-1} \leq N < B^k$ . There are precisely  $(B-1)^k$  integers in this range without a 0 digit. Thus,

considering all integers in the range  $0 < N < B^k$ , there are

$$\sum_{j=1}^k (B-1)^j = \frac{(B-1)((B-1)^k - 1)}{B-2}$$

integers without a 0 digit. Thus, the density of integers with persistence greater than 1 up to  $B^k$  is

$$\frac{(B-1)((B-1)^k - 1)}{(B-2)B^k} = \frac{B-1}{B-2} \left( \left(1 - \frac{1}{B}\right)^k - \frac{1}{B^k} \right) < 2 \left(1 - \frac{1}{B}\right)^k.$$

As  $k$  approaches infinity, this last term goes to zero, proving the asymptotic density goes to zero.  $\square$

We now prove the well-known result that every number in base  $B = 2$  has persistence 1 (some authors define the persistence of a single digit to be 0, so we only consider numbers with two or more digits).

**Theorem 3.** *In base 2, each number  $n > 2$  has persistence 1.*

*Proof.* Either  $n$  has all digits equal to 1, in which case  $f(n) = 1$ , or  $n$  has at least one 0 digit, in which case  $f(n) = 0$ .  $\square$

Base 2 is the only base where we can prove Sloane's conjecture, but we can support his conjecture in other bases. In particular, Beeler and Gosper [1972, Item 57] showed that any number in base 3 with persistence greater than 3 must have more than 30739014 digits. We extend this to  $10^9$  digits.

**Theorem 4.** *In base 3, if  $n < 3^{10^9}$ , then  $n$  has persistence at most 3, and if  $n < 3^{10^9}$  has persistence 3, then  $f(n) = 2^3$  or  $2^{15}$ .*

*Proof.* As noted above, if  $n$  has a digit of 0, then it has persistence 1, and if  $n$  has a digit of 1, then the persistence is unchanged if we remove all 1 digits. Thus, we may assume  $n$  has every digit equal to 2, so  $f(n) = 2^k$  for some  $k$ . One can verify that the powers of 2 below 87 have persistence 1 except  $2^3$  and  $2^{15}$ , which have persistence 2. Beeler and Gosper showed that each power of 2 between  $2^{87}$  and  $2^{30739014}$  contains a 0 in its base 3 expansion, and hence has persistence 1. With today's faster computers, we easily extend this to all powers of 2 up to  $10^9$ .  $\square$

**Theorem 5.** *In base 4, if  $n < 4^{10^9}$ , then  $n$  has persistence at most 3. If  $n < 4^{10^9}$  has persistence 3, then  $f(n) = 2^a 3^b$ , where  $(a, b) = (0, 3), (1, 3), (1, 5), (0, 6), (0, 10),$  or  $(1, 11)$ .*

*Proof.* We have already noted that we need not consider any  $n$  with a digit of 0 or 1. Further, if  $n$  in base 4 has the digit 2 at least twice, then  $f(n)$  has low-order digit 0, so  $f(f(n)) = 0$ . Thus, we may assume  $n$  has at most one digit 2 and the rest of the digits are 3; in other words,  $f(n) = 2^a 3^b$  with  $a \in \{0, 1\}$ . We now calculate the

persistence of  $3^b$  and of  $2 \cdot 3^b$  for all  $b \leq 10^9$  and note that none have persistence greater than 1 except for the listed values. For  $b > 1000$ , we do not actually calculate the persistence; we merely verify that there is a 0 digit in the last 64 digits.  $\square$

**Theorem 6.** *In base 5, if  $n < 5^{10000}$ , then  $n$  has persistence at most 6. If  $n < 5^{10000}$  has persistence 6, then  $f(n) = 2^{40}3^2$ .*

*Proof.* As before, we need not consider any  $n$  with a digit of 0 or 1. If  $n$  has a digit of 4, we may replace it by two digits 2. Thus, we may assume  $n$  has all digits equal to 2 or 3, in other words,  $f(n) = 2^a3^b$  for  $a \geq 0$  and  $b \geq 0$ . We now calculate the persistence of  $2^a3^b$  for  $a$  and  $b$  with  $\lceil a/2 \rceil + b \leq 1000$ ; the factor of  $1/2$  arises because each digit 4 is replaced by two digits 2. For large  $a + b$ , we merely verify that there is a 0 digit in the last 64 digits. The calculations show that each such  $2^a3^b$  has persistence less than 5 except for  $2^{40}3^2$ , which has persistence 5; hence,  $n$  has persistence at most 6 for all  $n < 5^{10000}$ .  $\square$

**Theorem 7.** *In base 6, if  $n < 6^{10000}$ , then  $n$  has persistence at most 5. If  $n < 6^{10000}$  has persistence 5, then  $f(n) = 2^a5^b$ , where  $(a, b) = (7, 1), (1, 4), (0, 5), (7, 2), (4, 4), (9, 3), (7, 4), (0, 8),$  or  $(17, 2)$ .*

*Proof.* As before, we eliminate digits of 0 or 1, and replace digits of 4 by two digits 2. If  $n$  has a digit of 3 and an even digit, then  $f(f(n)) = 0$ , so we may assume  $n$  either has all digits equal to 2 or 5, or else  $n$  has all digits equal to 3 or 5. In other words,  $f(n) = 2^a5^b$  or  $3^a5^b$  for  $a \geq 0$  and  $b \geq 0$ . We now calculate the persistence of  $2^a5^b$  for  $a$  and  $b$  with  $\lceil a/2 \rceil + b \leq 10000$  (the factor of  $1/2$  covers the case where each digit 4 is replaced by two digits 2), and also calculate the persistence of  $3^a5^b$  where  $a + b \leq 10000$ . The calculations show that all such expressions have persistence less than 4 except for the listed values, which have persistence 4; hence,  $n$  has persistence at most 5 for all  $n < 6^{10000}$ .  $\square$

**Theorem 8.** *In base 7, if  $n < 7^{1000}$ , then  $n$  has persistence at most 8. If  $n < 7^{1000}$  has persistence 8, then  $f(n) = 2^a3^b5^c$ , where  $(a, b, c) = (9, 3, 12), (9, 17, 4), (11, 8, 10), (10, 20, 5), (10, 8, 16), (19, 25, 1), (1, 44, 0), (27, 0, 20), (39, 24, 1),$  or  $(11, 39, 3)$ .*

*Proof.* As before, we eliminate digits of 0 or 1, replace digits of 4 by two digits 2, and now also replace digits 6 by digits 2 and 3. So, we may assume  $n$  has all digits equal to 2, 3 or 5. In other words,  $f(n) = 2^a3^b5^c$  for  $a \geq 0, b \geq 0,$  and  $c \geq 0$ . We now calculate the persistence of  $2^a3^b5^c$ ; since we replaced digits of 4 by  $2 \cdot 2$  and digits of 6 by  $2 \cdot 3$ , we must consider  $a, b, c$  with

$$a + b + c - \min(a, b) - \left\lfloor \frac{a - \min(a, b)}{2} \right\rfloor \leq 1000.$$

We calculate the persistence of each such  $2^a 3^b 5^c$  to find that all such expressions have persistence less than 6 except for the listed values, which have persistence 6; hence,  $n$  has persistence at most 7 for all  $n < 7^{1000}$ .  $\square$

**Theorem 9.** *In base 8, if  $n < 8^{1000}$ , then  $n$  has persistence at most 6. If  $n < 8^{1000}$  has persistence 6, then  $f(n) = 3^3 5^4 7^2$ .*

*Proof.* As before, we eliminate digits of 0 or 1, replace digits of 4 by two digits 2, and now also replace digits 6 by digits 2 and 3. So, we may assume  $n$  has all digits equal to 2, 3, 5 or 7. If there are three or more digits 2, then  $f(f(n)) = 0$ . Therefore,

$$f(n) = 2^d 3^a 5^b 7^c \quad \text{for } a \geq 0, b \geq 0, c \geq 0, \text{ and } d \in \{0, 1, 2\}.$$

We consider  $a, b, c$  with  $a + b + c \leq 1000$  to guarantee we are considering up to 1000 digits. We calculate the persistence of each such  $2^d 3^a 5^b 7^c$  to find that all such expressions have persistence less than 5 except for  $3^3 5^4 7^2$ , which has persistence 5; hence,  $n$  has persistence at most 6 for all  $n < 8^{1000}$ .  $\square$

**Theorem 10.** *In base 9, if  $n < 9^{1000}$ , then  $n$  has persistence at most 7. If  $n < 9^{1000}$  has persistence 7, then  $f(n) = 2^a 5^b 7^c$ , where  $(a, b, c) = (1, 1, 5), (3, 3, 4), (24, 1, 1), (4, 6, 4), (11, 5, 3),$  or  $(16, 7, 1)$ .*

*Proof.* As before, we eliminate digits of 0 or 1, replace digits of 4 by two digits 2, replace digits 6 by digits 2 and 3, and now also replace 8 by three digits 2. So, we may assume  $n$  has all digits equal to 2, 3, 5 or 7. If there are two or more digits 3, then  $f(f(n)) = 0$ , so we may assume  $f(n) = 2^a 5^b 7^c$  or  $f(n) = 3 \cdot 2^a 5^b 7^c$  for  $a \geq 0, b \geq 0,$  and  $c \geq 0$ . We now calculate the persistence of  $3^d 2^a 5^b 7^c$  for  $d = 0$  or 1; in order to guarantee that we consider all numbers up to 1000 digits, we must consider  $a, b, c$  with  $\lceil a/3 \rceil + b + c \leq 1000$ . We calculate the persistence of each such  $3^d 2^a 5^b 7^c$  to find that all such expressions have persistence less than 6 except for the listed values (all having  $d = 0$ ), which have persistence 6; hence,  $n$  has persistence at most 7 for all  $n < 9^{1000}$ .  $\square$

We now deal with base 10. Diamond [2010] calculated the persistence of all numbers  $2^a 3^b 7^c$  and  $3^a 5^b 7^c$  with  $a \leq 1000, b \leq 1000$  and  $c \leq 1000$ . We verify his calculations and extend them to cover all numbers up to 1500 digits.

**Theorem 11.** *In base 10, if  $n < 10^{1500}$ , then  $n$  has persistence at most 11. If  $n < 10^{1500}$  has persistence 11, then  $f(n) = 2^4 3^{20} 7^5$  or  $2^{19} 3^4 7^6$ .*

*Proof.* As before, we eliminate digits of 0 or 1, replace digits of 4 by two digits 2, replace digits 6 by digits 2 and 3, replace the digit 8 by three digits 2, and now also replace 9 by two digits 3. In base 10, if we have both a digit 2 and a digit 5, then  $f(f(n)) = 0$ . So, we may assume  $f(n) = 2^a 3^b 7^c$  or  $f(n) = 3^a 5^b 7^c$  for  $a \geq 0, b \geq 0,$  and  $c \geq 0$ . To consider all  $n$  with less than 1500 digits, we only need to

consider  $f(n) = 2^a 3^b 7^c$  with  $\lfloor a/3 \rfloor + \lfloor b/2 \rfloor + c \leq 1500$ , as well as  $f(n) = 3^a 5^b 7^c$  with  $\lceil a/2 \rceil + b + c \leq 1500$ . We find that all such expressions have persistence at most 9, except for the listed exceptions which have persistence 10; hence,  $n$  has persistence at most 11 for all  $n < 10^{1500}$ .  $\square$

**Theorem 12.** *In base 11, if  $n < 11^{250}$ , then  $n$  has persistence at most 13. If  $n < 11^{250}$  has persistence 13, then  $f(n) = 2^{42} 3^{13} 5^{20} 7^{17}$ ,  $2^{91} 3^{37} 5^7 7^6$ , or  $2^{32} 3^3 5^{35} 7^{18}$ .*

*Proof.* As before, we eliminate digits of 0 or 1, replace digits of 4 by two digits 2, replace digits 6 by digits 2 and 3, replace the digit 8 by three digits 2, and now also replace 9 by two digits 3. We may assume  $f(n) = 2^a 3^b 5^c 7^d$  for  $a, b, c, d \geq 0$ . To consider all  $n$  with less than 250 digits, we only need to consider  $f(n) = 2^a 3^b 5^c 7^d$  with  $\lfloor a/3 \rfloor + \lfloor b/2 \rfloor + c + d \leq 250$ . We find that all such expressions have persistence at most 11, except for the listed exceptions which have persistence 12; hence,  $n$  has persistence at most 13 for all  $n < 11^{250}$ .  $\square$

**Theorem 13.** *In base 12, if  $n < 12^{250}$ , then  $n$  has persistence at most 7. If  $n < 12^{250}$  has persistence 7, then  $f(n) = 2^5 5^8 11^9$  or  $3^5 5^1 7^6$ .*

*Proof.* As before, we eliminate digits of 0 or 1, replace digits of 4 by two digits 2, replace digits 6 by digits 2 and 3, replace the digit 8 by three digits 2, and now also replace 9 by two digits 3. We may assume  $f(n) = 2^a 5^b 7^c 11^d$  or  $3^a 5^b 7^c 11^d$  or  $6 \cdot 3^a 5^b 7^c 11^d$  for  $a, b, c, d \geq 0$ . To consider all  $n$  with less than 250 digits, we only need to consider  $f(n) = 2^a 5^b 7^c 11^d$  with  $\lfloor a/3 \rfloor + b + c + d \leq 250$ , and for  $f(n) = 3^a 5^b 7^c 11^d$  or  $6 \cdot 3^a 5^b 7^c 11^d$ , we consider  $\lfloor a/2 \rfloor + b + c + d \leq 250$ . We find that all such expressions have persistence at most 5, except for the listed exceptions which have persistence 6; hence,  $n$  has persistence at most 7 for all  $n < 12^{250}$ .  $\square$

### 3. Conclusion

These calculations support Sloane's conjecture that the persistence is bounded for a given base. This makes sense since when a product of powers like  $2^a 3^b 7^c$  has many digits, one expects to find a 0 digit among them. For instance, in base 10, we saw that  $2^4 3^{20} 7^5 = 937638166841712$  has persistence 10, but

$$2^3 3^{20} 7^5 = 468819083420856, \quad 2^4 3^{19} 7^5 = 312546055613904,$$

$$2^4 3^{20} 7^4 = 133948309548816$$

all have a digit of 0. In general, almost all such powers will have a persistence of 1.

We used simple Maple programs, so the calculations for each theorem above took several hours to a few days to run on a laptop.

The first author tried to develop a method to work backwards, in order to answer questions such as which numbers iterate to the digit 1. We can devise many such interesting questions. Paul Erdős [Weisstein] asked what would happen if one

multiplies only the nonzero digits (i.e., ignore the zero digits). Presumably this Erdős multiplicative persistence is no longer bounded, and the question of which numbers iterate to the digit 1 becomes more interesting. See [Wagstaff 1981] for another fascinating variation. We hope this paper inspires others to pursue the many fascinating problems related to multiplicative persistence.

## References

- [Beeler et al. 1972] M. Beeler, R. Gosper, and R. Schroepfel, “**HAKMEM**”, MIT AI Memo 239, 1972, <http://www.inwap.com/pdp10/hbaker/hakmem/number.html#item56>.
- [Carmody 2001] P. Carmody, “**OEIS A003001, and a ‘zero-length message’**”, message on NMBR-THRY listserve, 23 July 2001, <http://goo.gl/55n3LP>.
- [Diamond 2010] M. R. Diamond, “**Multiplicative persistence base 10: some new null results**”, 2010, <http://www.markdiamond.com.au/download/joous-3-1-1.pdf>.
- [Gottlieb 1969] A. J. Gottlieb, “Bridge, group theory, and a jigsaw puzzle”, *Technology Rev.* **72** (1969), unpaginated.
- [Guy 2004] R. K. Guy, *Unsolved problems in number theory*, 3rd ed., Springer, New York, 2004. [MR 2005h:11003](#)
- [Sloane 1973] N. J. A. Sloane, “The persistence of a number”, *J. Recreational Math.* **6**:2 (1973), 97–98.
- [Wagstaff 1981] S. S. Wagstaff, Jr., “Iterating the product of shifted digits”, *Fibonacci Quart.* **19**:4 (1981), 340–347. [MR 83b:10012](#)
- [Weisstein] E. W. Weisstein, “**Multiplicative persistence**”, webpage, <http://mathworld.wolfram.com/MultiplicativePersistence.html>.

Received: 2013-05-19

Revised: 2013-09-09

Accepted: 2013-12-23

[sperez03@villanova.edu](mailto:sperez03@villanova.edu)

*Department of Mathematics and Statistics,  
Villanova University, 800 Lancaster Avenue,  
Villanova, PA 19085-1699, United States*

[robert.styer@villanova.edu](mailto:robert.styer@villanova.edu)

*Department of Mathematics and Statistics,  
Villanova University, 800 Lancaster Avenue,  
Villanova, PA 19085-1699, United States*



# A new partial ordering of knots

Arazelle Mendoza, Tara Sargent, John Travis Shrontz and Paul Drube

(Communicated by Józef H. Przytycki)

Our research concerns how knots behave under crossing changes. In particular, we investigate a partial ordering of alternating knots that results from performing crossing changes. A similar ordering was originally introduced by Kouki Taniyama in the paper “A partial order of knots”. We amend Taniyama’s partial ordering and present theorems about the structure of our ordering for more complicated knots. Our approach is largely graph theoretic, as we translate each knot diagram into one of two planar graphs by checkerboard coloring the plane. Of particular interest are the class of knots known as pretzel knots, as well as knots that have only one direct minor in the partial ordering.

## 1. Introduction

**Basic knot theory.** A *knot*  $K$  is a smooth embedding of a circle  $S^1$  in  $\mathbb{R}^3$ . Some of our results generalize to links. A *link*  $L$  is a smooth embedding of multiple disjoint copies of  $S^1$  in  $\mathbb{R}^3$ . Knot theorists generally do not want to work with 3-dimensional objects, which is why it is common to use knot diagrams. A *knot diagram*  $D$  of the knot  $K$  is a way of projecting  $K$  onto  $\mathbb{R}^2$ . This projection is one-to-one everywhere except a finite number of points called *crossings* where it is two-to-one. At every crossing there is an unbroken line for the overstrand and a broken line for the understrand. The overstrand corresponds to the arc that was initially closer to the viewer in  $\mathbb{R}^3$ .

A leading problem in knot theory is that one knot  $K$  may have many different diagrams that don’t look remotely similar. We then need methods to determine when two knot diagrams represent the same knot.

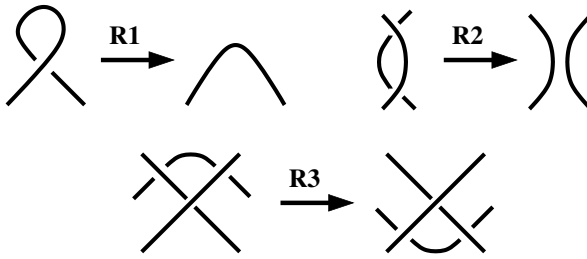
The required machinery to deal with this problem are the *Reidemeister moves*, which are a set of three moves that connect diagrams of the same knot. The Reidemeister moves are shown in [Figure 1](#). These moves are local, meaning the knot is unchanged outside of the exhibited region. The fundamental result about Reidemeister moves is the following:

---

*MSC2010:* primary 57M25; secondary 57M27.

*Keywords:* knots, links, crossing changes.

Research supported by NSF Grant DMS-0851721.



**Figure 1.** Reidemeister moves.

**Theorem 1.1** (Reidemeister). *Two diagrams  $D_1$  and  $D_2$  represent the same knot  $K$  if and only if they may be connected by a finite number of Reidemeister moves.*

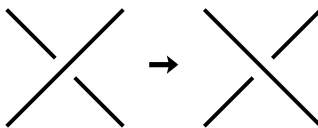
*Proof.* See [Reidemeister 1927] for a proof of this standard result.  $\square$

The *crossing number*  $c(K)$  of a knot  $K$  is the minimum number of crossings over all diagrams  $K$ . A *minimal knot diagram* is a diagram  $D$  where the number of crossings equals  $c(K)$ . The standard way to denote knots takes the form  $N_n$ , where  $N$  denotes the crossing number of the knot and the subscript  $n$  is a traditional ordering (which depends upon an invariant known as the determinant).

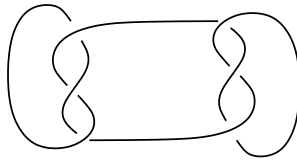
Our research concerns how knots behave under crossing changes. A *crossing change* is a local operation that flips the role of the overstrand and the understrand at a single crossing in a knot diagram. The most important thing to note here is that a crossing change may change the underlying knot. An example of a crossing change is shown in Figure 2. As with our images for the Reidemeister moves, it is assumed that the link is unchanged outside of the region shown.

We focus on the class of knots known as prime alternating knots since they have many nice properties that allow for stronger results. An *alternating knot* is a knot with an alternating diagram, which is a knot diagram that alternates between overstrands and understrands as one travels around the diagram in a fixed direction. A *prime knot* is a knot that cannot be drawn as a connect sum of two nontrivial knots (i.e., it doesn't look like two or more nontrivial knots that have been strung together). Figure 3 shows the granny knot, which is a connect sum of two trefoil knots  $3_1$ .

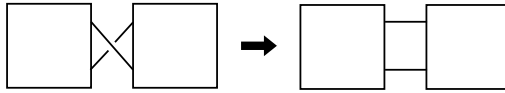
The following two theorems are important results that make prime alternating knots especially nice to work with.



**Figure 2.** Crossing change.



**Figure 3.** A nonprime knot.



**Figure 4.** A nugatory crossing and untwisting that crossing.

**Theorem 1.2** (Kauffman, Murasugi, and Thistlethwaite). *Let  $K$  be a prime alternating knot with diagram  $D$ . Then  $D$  is a minimal diagram for  $K$  if and only if  $D$  is a reduced alternating diagram.*

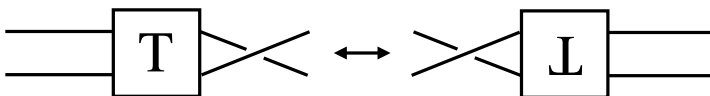
*Proof.* See [Adams 2004] for a proof of this foundational result. □

The term reduced above means that the diagram contains no nugatory crossings. A crossing in a diagram  $D$  is a *nugatory (removable) crossing* if removing a neighborhood of that crossing splits the knot diagram into two separate pieces. These are the crossings that can obviously be eliminated (via a 180-degree twist) to lower the crossing number of  $D$  without changing the underlying knot. See Figure 4.

**Theorem 1.3** (Tait’s flying conjecture, Menasco & Thistlethwaite). *Let  $D_1$  and  $D_2$  be two minimal diagrams of the same prime alternating knot  $K$  in  $S^2$ . Then  $D_1$  can be transformed into  $D_2$  via a series of flypes.*

*Proof.* See [Menasco and Thistlethwaite 1993] for the (surprisingly complex) proof of this result, which had eluded knot theorists for a century. □

An example of a *flype* is shown in Figure 5. This operation involves a 180-degree twist of the portion of the knot denoted by  $T$  (known as a tangle), effectively moving a single half-twist from one side of that tangle to the other side. A flype is usually a complex combination of Reidemeister moves, but just like the basic Reidemeister moves, it does not change the underlying knot.



**Figure 5.** Flype operation.

**A partial ordering of knots.** The starting point for our research was the partial ordering on knots defined by Kouki Taniyama [1989]. To distinguish Taniyama's ordering from our own, we will henceforth refer to this partial ordering as the T-order:

**Definition 1.4.** Let  $K_1$  and  $K_2$  be knots. The *T-order* defines  $K_1 \leq K_2$  if every diagram of  $K_2$  can be transformed into some diagram of  $K_1$  via some number of simultaneous crossing changes.

The number of simultaneous crossing changes required above depends upon the diagram of  $K_2$  chosen, and there may not be a systematic way to determine the required crossing changes in a given diagram of  $K_2$ .

We present a modified version of Taniyama's T-ordering that was also influenced by the distinct partial ordering of Ernst, Diao, and Stasiak [Diao et al. 2009]. We will call our ordering the V-order, in honor of Valparaiso University (the site of the REU where we conducted this research).

**Definition 1.5.** Let  $K_1$  and  $K_2$  be prime alternating knots. The *V-order* defines  $K_1$  to be a *V-minor* of  $K_2$  if there exists a minimal diagram of  $K_2$  that can be transformed into some diagram of  $K_1$  via simultaneous crossing changes. We then define  $(K_n, K_{n-1}, \dots, K_2, K_1)$  to be a *proper sequence* of knots if  $K_i$  is a V-minor of  $K_{i+1}$  for all  $i$ , and  $K_1 \leq K_2$  if there exists a proper sequence containing both  $K_1$  and  $K_2$ , where  $K_1$  appears to the right of  $K_2$ .

In this partial ordering (as was the case in Taniyama's original partial ordering), we do not differentiate between a knot, its reflection, and its reverse.

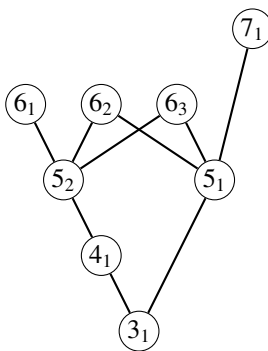
The reason that we present such a complicated definition involving proper sequences is to ensure that the resulting relation is transitive. One can quickly verify that the V-order defines a partial order of alternating knots, meaning that

- (1)  $K \leq K$  for all  $K$ ;
- (2) if  $K_1 \leq K_2$  and  $K_2 \leq K_3$ , then  $K_1 \leq K_3$ ;
- (3) if  $K_1 \leq K_2$  and  $K_2 \leq K_1$ , then  $K_1 = K_2$ .

It is the third condition in the partial ordering definition above that requires us to restrict our attention solely to prime alternating knots. There exist nonalternating knots such that  $K_1 \leq K_2$  and  $K_2 \leq K_1$ , yet  $K_1 \neq K_2$  (see Theorem 2.3 for more details).

We represent the V-order with a Hasse diagram, which is a graphical way to represent the relationships in the partial ordering. If two knots  $K_1$  and  $K_2$  are connected by a series of edges on the Hasse diagram, and if  $K_1$  lies below  $K_2$  on the edge, then  $K_1 \leq K_2$ . We manually verified that the V-order is identical to the T-order for the first eight nontrivial prime alternating knots (through  $7_1$ ), yielding the Hasse diagram in Figure 6.

Note that our ordering requires that we check only one minimal diagram of  $K_2$  to verify  $K_1 \leq K_2$ , while Taniyama's ordering requires that we check all diagrams



**Figure 6.** Partial ordering for the first eight prime knots.

of  $K_2$ . Also notice that if  $K_1 \leq K_2$  in the T-order, then  $K_1 \leq K_2$  in the V-order. The converse is not necessarily true a priori, although we conjecture that it is true for prime alternating knots (see [Conjecture 4.1](#)). The V-order relates to Ernst, Diao, and Stasiak’s work [[Diao et al. 2009](#)] in that their ordering allows for only one crossing change, while ours allows for multiple simultaneous crossing changes. This seemingly simple modification actually makes our ordering significantly more complicated, yet also helps our ordering maintain a closer relationship with Taniyama’s original ordering (which also allows for multiple simultaneous crossing changes).

In future sections, we will be especially interested in direct V-minors:

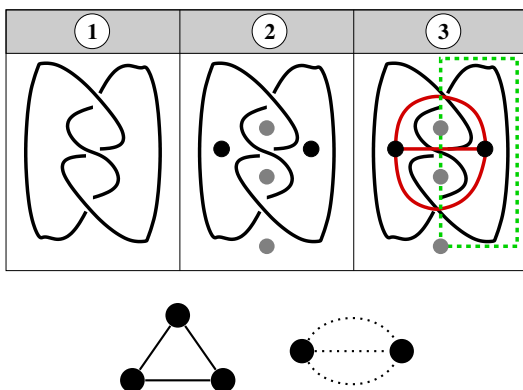
**Definition 1.6.**  $K_1$  is a *direct V-minor* of  $K_3$  if  $K_1 \leq K_3$  and there does not exist  $K_2$  ( $K_2 \neq K_1, K_3$ ) such that  $K_1 \leq K_2 \leq K_3$ .

**Definition 1.7.**  $K_1$  is a *remote V-minor* of  $K_3$  if  $K_1 \leq K_3$  and there exists  $K_2$  ( $K_2 \neq K_1, K_3$ ) such that  $K_1 \leq K_2 \leq K_3$ .

For example, as easily read from our Hasse diagram in [Figure 6](#), the knot  $3_1$  is a remote V-minor of  $7_1$  because  $3_1 \leq 5_1 \leq 7_1$ . However,  $3_1$  is a direct V-minor of  $5_1$  since there does not exist a distinct knot  $K$  such that  $3_1 \leq K \leq 5_1$ .

**Graph theoretical methods in knot theory.** By [Theorem 1.1](#), we know that any two diagrams of one knot may be connected via a series of Reidemeister moves, but it is tedious to constantly redraw the diagram every time we perform a Reidemeister move. To make calculations easier, we convert knot diagrams to a specific type of signed planar graph that contains all of the same information. The procedure for converting a knot diagram to a graph is as follows:

- (1) Checkboard color the regions of the plane in the complement of the knot diagram so that around each crossing there are two white regions and two gray regions. Then mark each crossing by dropping a line segment connecting the two regions that lie counterclockwise from the overstrand.



**Figure 7.** Checkerboard graphs of  $3_1$ .

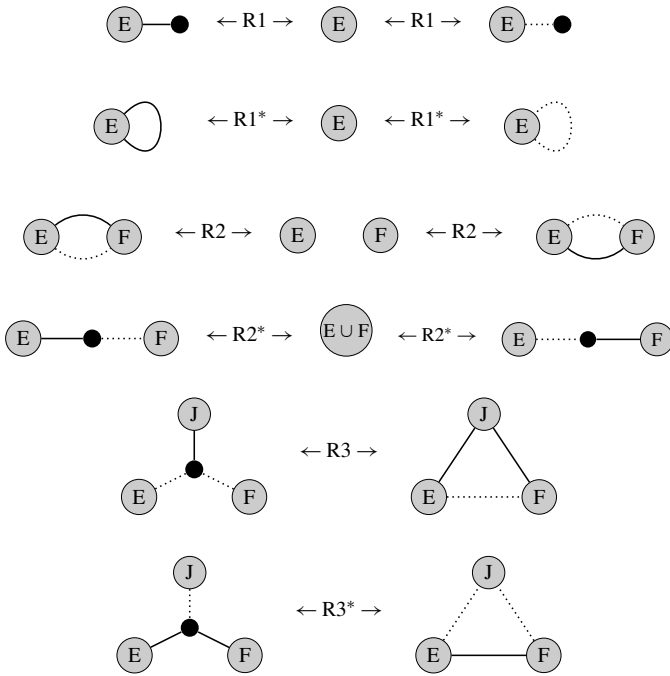
- (2) Pick one of the two colors. Place a vertex inside each region of this fixed color.
- (3) If two of the chosen regions share a crossing, add an edge between the corresponding vertices in the graph. This edge is solid if the marking associated with that crossing falls within the chosen regions and is dotted if the marking falls within the regions of the other color.

Since we had two choices in (2) above, we get two distinct graphs for any knot diagram. These graphs are always signed duals of one another. We illustrate this entire procedure for the trefoil knot  $3_1$  in [Figure 7](#), showing both of the resulting planar graphs in the second row.

Our next challenge is to determine how the Reidemeister moves for knot diagrams translate to checkerboard graphs, as we need a reliable way of determining when two graphs represent the same underlying knot. It is important to note that every Reidemeister move for knot diagrams actually corresponds to two graph Reidemeister moves that are duals of one another. We illustrate all of these graph Reidemeister moves in [Figure 8](#). In this figure, each diagram represents a local piece of the entire checkerboard graph.  $E$  and  $F$  represent nodes in the graph that may or may not have other edges entering them, while the small black vertices are adjacent only to the edges shown. In the second R2 move,  $E \cup F$  denotes that the central node is now adjacent to all edges that were formerly incident upon either the  $E$  node or the  $F$  node.

One more important thing to note is that both graph representations of an alternating diagram only have one type of edge (one of them has all solid edges, while the signed dual has all dotted edges). This makes alternating diagrams especially easy to identify from checkerboard graphs: you no longer have to trace along the entire diagram to see if the knot alternates between overstrands and understrands!

Since our research deals with how knots behave under crossing changes, we need to determine how a crossing change effects a knot diagram’s associated graph.

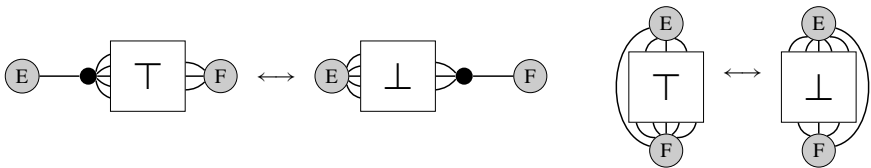


**Figure 8.** Reidemeister moves for graphs.

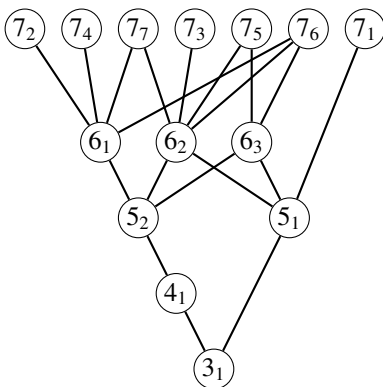
Crossing changes switch the roles of the overstrand and understrand at a single crossing. In either checkerboard graph for the diagram, this changes the marking on the associated edge and hence flips the type of edge that appears in the graph (dotted to solid, or solid to dotted).

Finally, we need to know how flypes effect our graphs. Figure 9 shows the graph representations of flypes. Just as with the Reidemeister moves, a flype has two different graph representations that are (signed) duals of one another.

Notice that, in the first flype equivalence, we are rearranging edges that separate the same two regions in our graph. In the second equivalence, we are rearranging edges that connect the same two vertices.



**Figure 9.** Graph equivalents of a flype.



**Figure 10.** The V-order for prime alternating knots through  $7_7$ .

### 2. Our partial ordering

Now we will investigate our V-order. Recall the definition:

**Definition 2.1.** Let  $K_1$  and  $K_2$  be prime alternating knots. The V-order defines  $K_1$  to be a V-minor of  $K_2$  if there exists a minimal diagram of  $K_2$  that can be transformed into some diagram of  $K_1$  via simultaneous crossing changes. We then define  $(K_n, K_{n-1}, \dots, K_2, K_1)$  to be a *proper sequence* of knots if  $K_i$  is a V-minor of  $K_{i+1}$  for all  $i$ , and  $K_1 \leq K_2$  if there exists a proper sequence containing both  $K_1$  and  $K_2$ , where  $K_1$  appears to the right of  $K_2$ .

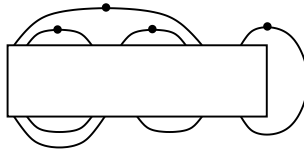
Our first goal was to directly expand the Hasse diagram of Section 1 up through 7-crossing prime alternating knots. If Conjecture 4.1 proves to be true, these results will translate into a direct extension of Taniyama’s original T-order.

In order to directly determine which knots were V-minors of a particular knot  $K$ , we exhaustively checked all possible ways to make simultaneous crossing changes on the graph for a fixed minimal diagram  $D$  of  $K$ . We checked all of the (combinatorially distinct) ways to make one crossing change at a time, and then two crossing changes at a time, etc., up to half of the crossing number of  $K$ . We did not need to change more than half of the crossings at a time because we do not distinguish between a knot and its reflection: if changing some set of crossings yields a diagram of  $K$ , then changing the complement of that set gives a diagram of the reflection of  $K$ .

Our updated Hasse diagram is shown in Figure 10. See the Appendix for the calculations that yielded this Hasse diagram.

**Invariants and the V-order.** The problem with the direct technique above is that there are an extremely large number of cases to check for each knot. In order to quickly eliminate many possible relationships in the V-order, we prove several





**Figure 11.** A knot diagram with  $br(D) = 4$ .

results about the ordering that involve knot invariants. A *knot invariant* is a function  $i : \kappa \rightarrow \alpha$  from the set of all knots  $\kappa$  to some algebraic structure  $\alpha$ . Distinct diagrams of the same knot must get sent to the same value by the invariant, so if an invariant gives different values for two diagrams, they cannot represent the same knot.

The knot invariants we work with are crossing number  $c(K)$ , bridge index  $br(K)$ , and braid index  $b(K)$ . It should be noted that some of our proofs in this section are similar to those presented in [Endo et al. 2010], where Endo, Itah, and Taniyama relate an entirely distinct partial ordering of links to common link invariants.

**Theorem 2.2.** *Let  $K_1, K_2$  be distinct knots with  $K_1 \leq K_2$ , then  $c(K_1) \leq c(K_2)$ .*

*Proof.* Let  $K_1$  and  $K_2$  be knots, where  $K_1 \leq K_2$  and  $c(K_2) = n$ . Then there exists a minimal diagram  $D_2$  of  $K_2$  that can be transformed into a diagram  $D_1$  of  $K_1$  via some number of simultaneous crossing changes. Now,  $D_1$  has  $n$  crossings, and thus the crossing number of  $K_1$  can be at most  $n$ . □

The following theorem, which was originally proven by Taniyama [1989], is more specific to our research since our V-order is restricted to alternating knots.

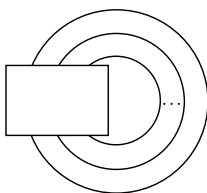
**Theorem 2.3.** *Let  $K_1, K_2$  be alternating knots with  $K_1 \leq K_2$ , then  $c(K_1) < c(K_2)$ .*

*Proof.* Let  $K_1$  and  $K_2$  be alternating knots, where  $K_1 \leq K_2$  and  $c(K_2) = n$ . Then there exists a minimal diagram  $D_2$  of  $K_2$  that can be transformed into a diagram  $D_1$  of  $K_1$  by simultaneously changing some but not all of the crossings in  $D_2$ . Now,  $D_1$  has  $n$  crossings, so by Theorem 1.2,  $D_1$  cannot be a minimal diagram of  $K_1$ . Thus  $c(K_1) < n$ . □

The second invariant we work with is the bridge number. The bridge number of a knot diagram  $D$  of  $K$  is the number of local maximums in  $D$  with respect to the  $y$ -coordinate in  $\mathbb{R}^2$  (the number of “top points” in the diagram). The *bridge index*  $br(K)$  of a knot  $K$  is the minimal bridge number over all diagrams of  $K$ . Note that, for every diagram  $D$  of  $K$ , there is one local minimum for every local maximum, so the bridge number could have been defined using local minimums.

An example of a knot diagram  $D$  with  $br(D) = 4$  is shown in Figure 11. Here the box represents some (possibly complex) part of the knot diagram that contains no local maxima or minima.

**Theorem 2.4.** *If  $K_1 \leq K_2$ , then  $br(K_1) \leq br(K_2)$ .*



**Figure 12.** A knot diagram with  $b(K) = 3$ .

*Proof.* Let  $K_1$  and  $K_2$  be knots, where  $K_1 \leq K_2$  and  $br(K_2) = n$ . Then there exists a minimal bridge diagram  $D_2$  of  $K_2$  that can be transformed into a diagram  $D_1$  of  $K_1$  via some number of simultaneous crossing changes. Since  $D_1$  has  $n$  local maxes, the bridge number of  $K_1$  can be at most  $n$ .  $\square$

The last invariant we work with is the braid index. The *braid index*  $b(K)$  is the minimal number of strands over all braid representations of a knot.

An example of a braid representation is shown in Figure 12. As with our figure for bridge number, the box represents some (possibly complex) part of the knot diagram that contains no local maxima or minima.

**Theorem 2.5.** *If  $K_1 \leq K_2$ , then  $b(K_1) \leq b(K_2)$ .*

*Proof.* Let  $K_1$  and  $K_2$  be knots where  $K_1 \leq K_2$  and  $b(K_2) = n$ . Then there exists a minimal braid diagram  $D_2$  of  $K_2$  (with  $n$  braid strands) that can be transformed into a diagram  $D_1$  of  $K_1$  via some number of simultaneous crossing changes. Since  $D_1$  has  $n$  braid strands, the braid index of  $K_1$  can be at most  $n$ .  $\square$

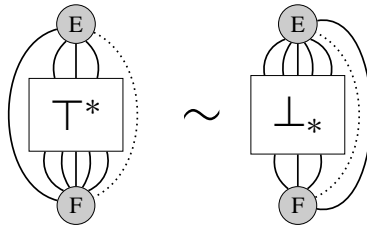
**Direct V-minors.** We now turn our attention to finding direct V-minors. Recall that  $K_1$  is a direct V-minor of  $K_3$  if  $K_1 \leq K_3$  and there does not exist a distinct  $K_2$  such that  $K_1 \leq K_2 \leq K_3$ . As we are restricting ourselves to prime alternating knots, we will search for direct minors by finding alternating knots  $K_1 \leq K_3$  such that  $c(K_1) = c(K_3) - 1$ . Theorem 2.3 ensures that all pairs of knots with this property yield a direct V-minor. Although this strategy won't find all direct V-minors, it will locate most of them (as you can tell from our expanded Hasse diagram, the vast majority of edges connect knots that differ by a crossing number of one).

Our primary tool in applying this strategy is the following theorem, which vastly limits the number of cases where  $c(K_1) = c(K_3) - 1$  is possible.

**Theorem 2.6.** *Let  $K_1$  and  $K_2$  be alternating knots with  $K_1 \leq K_2$ , and let  $G_2$  be any minimal graph of  $K_2$ .*

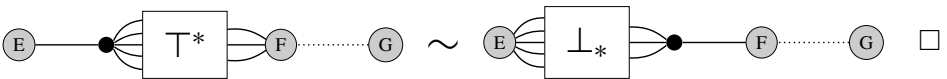
- (1) *In  $G_2$ , if we switch some but not all of the edges connecting two vertices, then  $c(K_1) \leq c(K_2) - 2$ .*
- (2) *In  $G_2$ , if we switch some but not all of the edges separating two regions, then  $c(K_1) \leq c(K_2) - 2$ .*

*Proof.* We are given that  $K_1$  and  $K_2$  are alternating knots with  $K_1 \leq K_2$ . Let  $G_2$  be an alternating graph of  $K_2$ . Switch some but not all of the edges connecting two vertices, so that those two vertices have at least one dotted edge and one solid edge between them. In general these edges need not be directly adjacent. If they are not directly adjacent, we can perform the flype below to make them adjacent:



After performing this flype we can always perform an R2 move, which will produce a graph with two edges less than the original  $G_2$ . Thus,  $K_1$  has at most  $c(K_2) - 2$  crossings and  $c(K_1) \leq c(K_2) - 2$ .

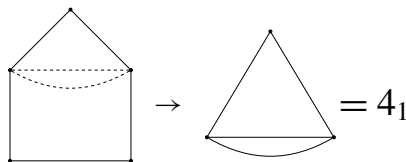
The proof for the case of switching some but not all of the edges separating two regions is similar to above. Now the relevant flype that yields an R2 move takes the form of the diagram below:



When searching for direct V-minors, we restrict our attention to the combinatorial cases that involve changing all crossings that connect a fixed pair of vertices or all crossings that separate a fixed pair of regions (or a multiple number of such cases). Using terminology from the literature, these cases correspond to changing all crossings in fixed number of twist boxes. These guidelines directly guided the calculations that we performed in the [Appendix](#).

It should be noted that the conditions from [Theorem 2.6](#), although necessary for obtaining a direct V-minor with  $c(K_1) = c(K_2) - 1$ , are not sufficient to guarantee that  $c(K_1) = c(K_2) - 1$ . Below is an example where we follow the conditions of [Theorem 2.6](#) but still end up with a knot such that  $c(K_1) \leq c(K_2) - 2$ .

**Example 2.7.** If we change both of the middle edges of the graph of  $7_5$ , we drop to the graph of  $4_1$ , which has  $c(4_1) = c(7_5) - 3$ .



### 3. Pretzel links and our partial ordering

**Basic properties.** A particularly simple class of links that behave nicely with respect to our partial ordering are pretzel links. A link is a *pretzel link* if it has a diagram that takes the form on the left side of Figure 13. Here the boxes represent twist boxes full of half-twists in either direction. Since it is sometimes difficult to tell whether a pretzel link is a one-component knot or a multiple-component link, all of our theorems in this subsection have been extended to alternating links.

If we take the gray regions from our checkerboard coloring on the left, we see that a pretzel link always has a graph of the form on the right side of Figure 13. Here the half-twists in the link diagram translate into parallel edges between adjacent vertices. We refer to graphs of this type as *polygonal graphs*. We denote the pretzel link of Figure 13 by  $P_v(x_1, x_2, x_3, \dots, x_v)$ , where  $v$  is the number of twist boxes in the link diagram (or the number of vertices in the associated polygonal graph) and  $x_i$  is an integer corresponding to the number of half-twists in each twist box (or the number of edges connecting the consecutive vertices  $v_i$  and  $v_{i+1}$ ). We define  $v_v$  to precede  $v_1$ . By convention,  $x_i$  will be negative if all of the edges in the given twist box are dotted, and positive if all of the edges are solid (if there are solid and dotted edges between two fixed vertices, we immediately eliminate them with an R2 move).

For example, in Figure 14 we have  $P_3(3, 3, 2) = 8_5$ . Notice that  $P_3(3, 3, 2) = P_3(3, 2, 3) = P(2, 3, 3)$ .

**Pretzel links and our partial order.** The reason that pretzel links are extremely nice in relation to our partial ordering is that many of them have only one or two direct V-minors (and almost all knots with only one or two direct V-minors appear to be pretzel knots; see Section 4). Here we present several theorems characterizing the role of several classes of pretzel links in our partial ordering.

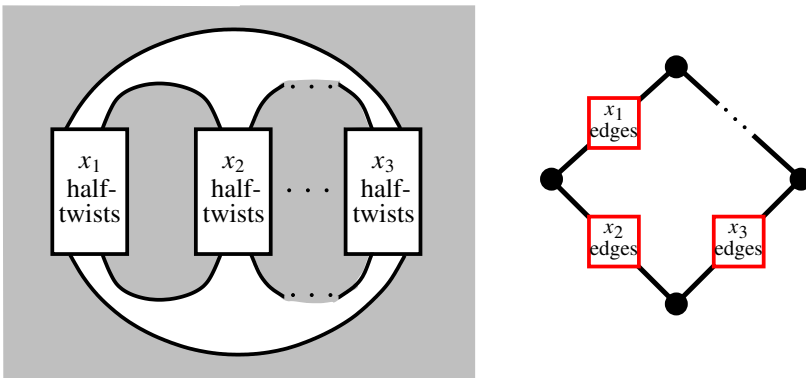
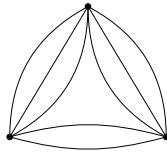


Figure 13. Pretzel knot diagram and its graph.



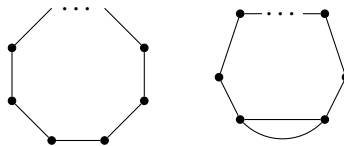
**Figure 14.**  $8_5 = P_3(3, 3, 2)$ .

The simplest class of pretzel links are  $(p, 2)$ -torus links. A  $(p, 2)$ -torus link is a link with only a single twist box, where  $p$  is the total number of half-twists in the twist box. They are so named because they fit upon the surface of a torus in  $\mathbb{R}^3$  and wrap around the torus  $p$  times in the meridian direction for every two times that they wrap around the torus in the longitudinal direction. If  $p$  is odd then the  $(p, 2)$ -torus link is a knot; if  $p$  is even then the  $(p, 2)$ -torus link is a two-component link. In terms of our pretzel link notation, the  $(p, 2)$ -torus link is  $P_p(1, 1, \dots, 1)$ . Figure 15 shows the general form for the checkerboard graph of a torus knot.

**Theorem 3.1.** *Every V-minor of the  $(p, 2)$ -torus link is a  $(q, 2)$ -torus link with  $q < p$ . Furthermore, the  $(p, 2)$ -torus link has a single direct V-minor in the  $(p - 2, 2)$ -torus link.*

*Proof.* Consider the graph  $P_p(1, \dots, 1)$  of the  $(p, 2)$ -torus link. If we change  $m < p$  crossings in the polygonal graph’s sole twist box, there will be a solid edge next to a dotted edge. This means that we can always perform an R2 move, removing edges in pairs until the edges are all solid or all dotted. Every time we perform an R2 move, we lose two edges. The resulting graph will always be of the form  $P_{p-2k}(1, \dots, 1)$ , where  $k$  is the minimum between the number of dotted edges and the number of solid edges that we start with.  $\square$

This theorem supports what we already found for the torus knots  $3_1, 5_1, 7_1$  in our Hasse diagram: the  $(p, 2)$ -torus knots line up in our Hasse diagram and have the smaller  $(p, 2)$ -torus knots below them in a line. Note that many non- $(p, 2)$ -torus knots may have a  $(p, 2)$ -torus knot as their V-minor: our theorem doesn’t work in the other direction.



**Figure 15.** Left:  $(p, 2)$ -torus knot checkerboard graph. Right: twist knot checkerboard graph.

Another basic class of pretzel links are twist links, which are always one-component knots. A *twist knot* is a pretzel link whose checkerboard graph is of the form shown in [Figure 15](#). Its two polygonal graphs are always of the form  $P_{c(K)-1}(2, 1, 1, \dots, 1)$  and  $P_3(c(K) - 2, 1, 1)$ . The smallest nontrivial twist knots are  $3_1 = P_3(1, 1, 1)$ ,  $4_1 = P_3(3, 1, 1)$ ,  $5_1 = P_3(4, 1, 1)$ , and  $6_1 = P_3(4, 1, 1)$ . Notice that  $3_1$  is both a twist knot and a  $(p, 2)$ -torus knot.

**Theorem 3.2.** *Every V-minor of the twist knot  $P_3(n, 1, 1)$  is a twist knot  $P_3(m, 1, 1)$  with  $m < n$ . Furthermore, the twist knot  $P_3(n, 1, 1)$  has a single direct V-minor in  $P_3(n - 1, 1, 1)$ .*

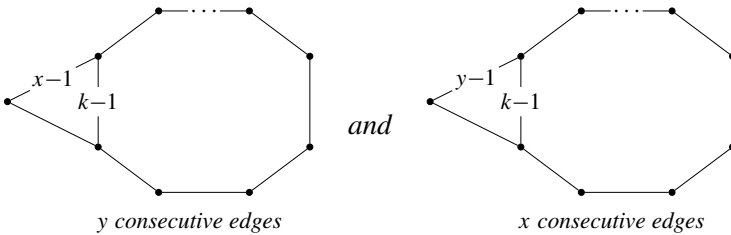
*Proof.* Changing  $m < n$  crossings in the big twist box always allows for R2 moves, similarly to [Theorem 3.1](#). The result is always a twist knot of the form  $P_3(n - 2k, 1, 1)$  for some integer  $k > 0$ . Changing one but not both of the remaining two crossings always results in the unknot (technically a twist knot), as an R2 move on the bottom allows us to completely untwist the knot. Changing both of the remaining crossings results in the direct V-minor  $P_3(n - 1, 1, 1)$ ; see the proof of [Theorem 3.3](#) for a more general demonstration of this fact. Changing both of the remaining two crossings and some number of crossings in the big twist box results in the same knot as changing the complement of these crossings, which falls into the same case as above. In every case, we are left with a twist knot.  $\square$

As with [Theorem 3.1](#), the implication of [Theorem 3.2](#) is easily seen in our Hasse diagram: the twist knots  $3_1, 4_1, 5_1$ , etc. line up along the left side of the diagram and only have other twist knots underneath them.

Theorems [3.1](#) and [3.2](#) are actually special cases of the theorem below, which gives a very broad class of pretzel links with only one or two direct V-minors:

**Theorem 3.3.** *Consider the pretzel link  $L = P_{k+2}(x, y, 1, 1, 1, \dots)$ , where  $k > 1$ .*

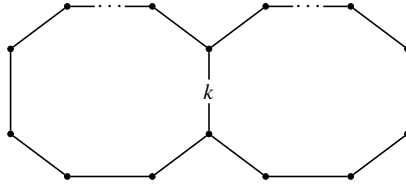
- (1) *If  $x, y \neq 1$ , then  $L$  has two direct V-minors, each of which has crossing number  $c(L) - 1$ . These two V-minors, which are equivalent if  $x = y$ , have (possibly nonpolygonal) graphs of the form*



*Here the  $x - 1, y - 1$ , and  $k - 1$  refer to that number of parallel strands.*

- (2) *If  $x = 1$ , then  $L$  has one direct V-minor of the form  $P_3(k, y - 1, 1)$ . Equivalently, if  $y = 1$ , then  $L$  has only one direct V-minor of the form  $P_3(k, x - 1, 1)$ .*

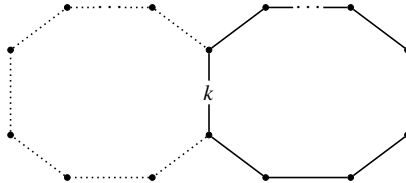
*Proof.* Given  $L$  as defined above, the dual graph of  $P_{k+2}(x, y, 1, 1, 1, \dots)$  is



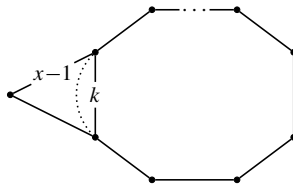
Here we have  $k$  parallel strands in the middle, a string of  $x$  consecutive strands of the left, and a string of  $y$  consecutive strands on the right. We choose to perform our possible crossing changes on this dual graph.

From [Theorem 2.6](#), we know that we can only achieve a direct V-minor  $L'$  with  $c(L') = c(L) - 1$  if we perform crossing changes on entire twist boxes. From the diagram above, we clearly have three twist boxes: one on the left, one on the right, and one with the  $k$  parallel strands down the middle. We then have three cases to check, corresponding to changing all of the crossings in each twist box (notice that, up to reflection, changing all crossings in two twist boxes yields the same knot as changing all of the crossings in the remaining twist box).

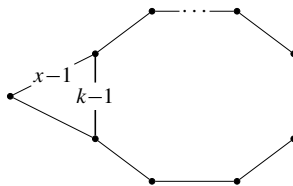
First we change all crossings on the left side, giving



After adding a free solid edge on the left side (corresponding to an R1 move), a series of R3 moves reduces the graph to

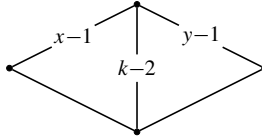


Notice that this graph has  $c(L) + 1$  edges. After performing an R2 move in the middle, we are left with the following graph with  $c(L) - 1$  edges, corresponding to the first direct V-minor from the theorem statement:



Changing all of the crossings on the right side of the original graph is equivalent to the above, and results in the second direct V-minor from the theorem statement.

Lastly, we consider changing all crossings in the middle twist box. This is equivalent (up to reflection) to changing all of the crossings on the left and on the right, which allows us to perform the procedure above two consecutive times to arrive at



This graph has  $c(L) - 2$  edges and is actually a direct V-minor of the two  $c(L) - 1$  crossing knots derived above. Hence it is a remote V-minor of our original link. Thus our link has only the two direct V-minors stated in the theorem.

Part (2) of the theorem is a special case of part (1). When  $x = 1$ , the string of consecutive edges in the right graph from the theorem statement is a single edge that adds to the twist box in the middle (which now has  $k$  parallel edges instead of  $k - 1$  parallel edges). The argument for  $y = 1$  is similar. □

### 4. Future work

Our work revealed several questions that we hope to address in future papers. The biggest open question that lay behind much of our research was what we referred to as the minimal conjecture.

**Conjecture 4.1** (The minimal conjecture). *Let  $K_2$  be a prime alternating knot (link) and let  $K_1$  be any knot (link). If there exists a minimal diagram of  $K_2$  that can be transformed into a diagram of  $K_1$  via some number of simultaneous crossing changes, then every diagram of  $K_2$  can be transformed into  $K_1$  via some number of simultaneous crossing changes.*

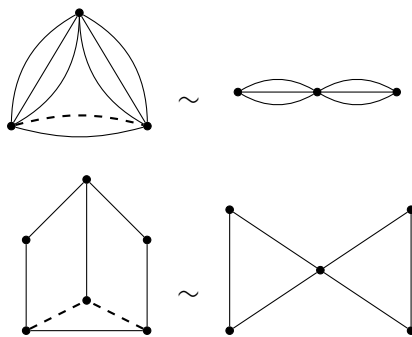
As noted earlier in the paper, if [Conjecture 4.1](#) is true, it implies that the V-order and T-order are equivalent for prime alternating knots. This means that our work would be a direct refinement of Taniyama’s original methods. Unfortunately, this conjecture seems to resist all direct methods of proof that we attempted.

In [Section 3](#), we produced many knots with only one direct V-minor. For knots with low crossing number, the only knots we found that had only one direct V-minor were pretzel knots. This begs the following conjecture.

**Conjecture 4.2.** *Pretzel knots are the only prime alternating knots with one direct V-minor.*

Below are a few additional general avenues of research that we may address in future research.





**Figure 16.** Top:  $8_5 \geq 3_1\#3_1$ . Bottom:  $8_{16} \geq 3_1\#3_1$ .

**Future Topic 4.3.** All  $(p, 2)$ -torus knots  $K$  lack direct V-minors  $K'$  with  $c(K') = c(K) - 1$ . Most other knots seem to have at least one V-minor with  $c(K') = c(K) - 1$ , but there are still examples of non- $(p, 2)$ -torus knots that fail in this regard. The knots  $8_5$  and  $8_{16}$  are non- $(p, 2)$ -torus knots  $K$  that have no direct V-minors  $K'$  with  $c(K') = c(K) - 1$ . Is there something special about these knots that we can generalize? Notice that these problematic eight-crossing knots are also the eight-crossing alternating knots with nonprime V-minors; see Figure 16.

Is it possible to expand our work to nonprime or nonalternating links? At the very least, is it possible to fully categorize which prime alternating knots have nonprime or nonalternating knots directly beneath them in our ordering?

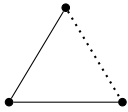

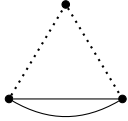
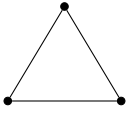
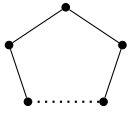
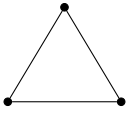
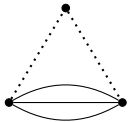
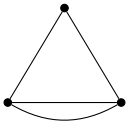
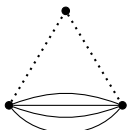
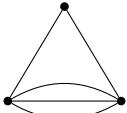
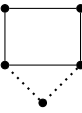
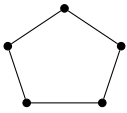
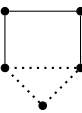
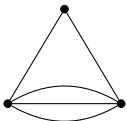
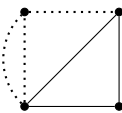
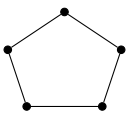
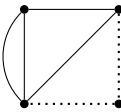
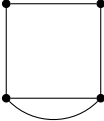
**Future Topic 4.4.** In relation to this final topic, we already have one result about the placement of nonalternating knots within the V-order:

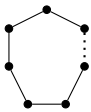
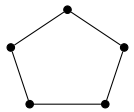
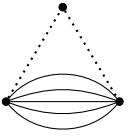
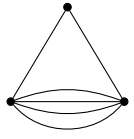
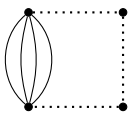
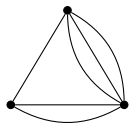
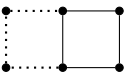
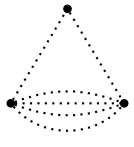
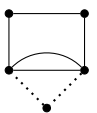
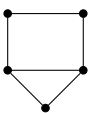
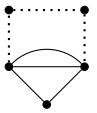
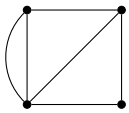
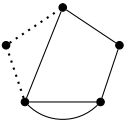
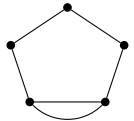
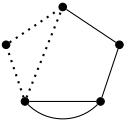
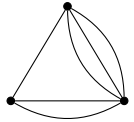
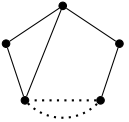
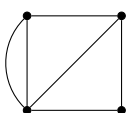
**Theorem 4.5.** *Let  $L_1$  be a nonalternating link with  $c(L_1) = n$ . Then there exists an alternating link  $L_2$ , where  $c(L_2) = n$ , such that  $L_1 \leq L_2$ .*

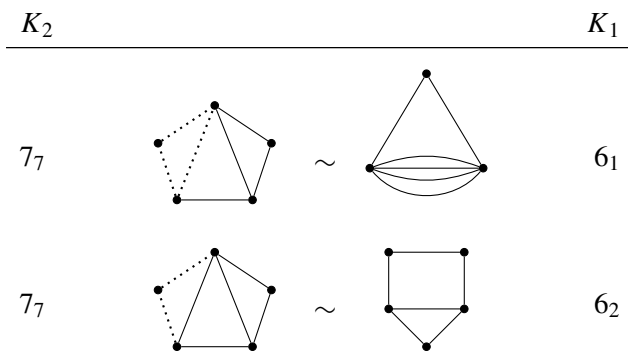
*Proof.* If  $L_1$  is a nonalternating link with  $c(L_1) = n$ , the minimal graph for  $L_1$  will have both dotted and solid edges with  $n$  edges total. If we change all the dotted edges to solid, we now have a graph of a link  $L_2$  with all solid edges. Since this projection is reduced alternating, Theorem 1.2 implies that this graph of  $L_2$  is minimal. So we have a minimal graph of  $L_2$  with crossing number  $n$ . We also can see that  $L_1 \leq L_2$  since we are able to transform a minimal diagram of  $L_2$  into  $L_1$  via crossing changes. □

### Appendix: Expansion of the Hasse diagram

Here we exhibit the calculations that yielded our expansion of the Hasse diagram in Section 2. For each edge in the diagram, which corresponds to  $K_1 \leq K_2$ , we show a minimal diagram of  $K_2$  with the crossing changes needed to produce the direct V-minor  $K_1$ .

$K_2$		$K_1$		
$3_1$		$\sim$		$0_1$
$4_1$		$\sim$		$3_1$
$5_1$		$\sim$		$3_1$
$5_2$		$\sim$		$4_1$
$6_1$		$\sim$		$5_2$
$6_2$		$\sim$		$5_1$
$6_2$		$\sim$		$5_2$
$6_3$		$\sim$		$5_1$
$6_3$		$\sim$		$5_2$

$K_2$		$K_1$		
$7_1$		$\sim$		$5_1$
$7_2$		$\sim$		$6_1$
$7_3$		$\sim$		$6_2$
$7_4$		$\sim$		$6_1$
$7_5$		$\sim$		$6_2$
$7_5$		$\sim$		$6_3$
$7_6$		$\sim$		$6_1$
$7_6$		$\sim$		$6_2$
$7_6$		$\sim$		$6_3$



### References

[Adams 2004] C. C. Adams, *The knot book: an elementary introduction to the mathematical theory of knots*, American Mathematical Society, Providence, RI, 2004. [MR 2005b:57009](#) [Zbl 1065.57003](#)

[Diao et al. 2009] Y. Diao, C. Ernst, and A. Stasiak, “A partial ordering of knots and links through diagrammatic unknotting”, *J. Knot Theory Ramifications* **18**:4 (2009), 505–522. [MR 2010i:57012](#) [Zbl 1200.57004](#)

[Endo et al. 2010] T. Endo, T. Itoh, and K. Taniyama, “A graph-theoretic approach to a partial order of knots and links”, *Topology Appl.* **157**:6 (2010), 1002–1010. [MR 2011c:57008](#) [Zbl 1196.57004](#)

[Menasco and Thistlethwaite 1993] W. Menasco and M. Thistlethwaite, “The classification of alternating links”, *Ann. of Math. (2)* **138**:1 (1993), 113–171. [MR 95g:57015](#) [Zbl 0809.57002](#)

[Reidemeister 1927] K. Reidemeister, “Knoten und Gruppen”, *Abh. Math. Sem. Univ. Hamburg* **5**:1 (1927), 7–23. [MR 3069461](#) [JFM 52.0578.04](#)

[Taniyama 1989] K. Taniyama, “A partial order of knots”, *Tokyo J. Math.* **12**:1 (1989), 205–229. [MR 90h:57008](#) [Zbl 0688.57006](#)

Received: 2013-06-21      Revised: 2014-03-30      Accepted: 2014-04-02

[arazelle.mendoza.09@cnu.edu](mailto:arazelle.mendoza.09@cnu.edu)      *Christopher Newport University, Newport News, VA 23606, United States*

[tara.sargent@clarke.edu](mailto:tara.sargent@clarke.edu)      *Clarke University, Dubuque, IA 52001, United States*

[jts0012@uah.edu](mailto:jts0012@uah.edu)      *University of Alabama in Huntsville, Huntsville, AL 35816, United States*

[paul.drube@valpo.edu](mailto:paul.drube@valpo.edu)      *Department of Mathematics and Computer Science, Valparaiso University, 1900 Chapel Drive, Valparaiso, IN 46383, United States*

# Two-parameter taxicab trigonometric functions

Kelly Delp and Michael Filipksi

(Communicated by Frank Morgan)

In this paper, we review some of the fundamental properties of the  $\ell^1$ , or taxicab, metric on  $\mathbb{R}^2$ . We define and give explicit formulas for two-parameter sine and cosine functions for this metric space. We also determine the maximum of these functions, which is greater than 1.

## 1. Introduction

The  $\ell^1$  metric on  $\mathbb{R}^2$ , the so-called taxicab metric, is often one of the first non-Euclidean metrics a mathematics student encounters. For any points  $p = (p_1, p_2)$  and  $q = (q_1, q_2)$  in  $\mathbb{R}^2$ , the metric is given by the formula

$$d_T(p, q) = |p_1 - q_1| + |p_2 - q_2|.$$

The  $\ell^1$  metric is just one metric in a class of metrics defined on  $\mathbb{R}^2$  known as *Minkowski metrics*; see [Álvarez Paiva and Thompson 2005] for an introduction to these metric spaces. Let  $\Omega$  be a closed, bounded convex set in  $\mathbb{R}^2$  which contains and is symmetric about the origin. The set  $\Omega$  defines a norm on  $\mathbb{R}^2$ , where  $\Omega$  is the unit disk. Given a norm  $\|\cdot\|$ , one can define a metric on  $\mathbb{R}^2$  by  $d(p, q) = \|p - q\|$ . Examples of Minkowski metrics include the  $\ell^p$  metrics, the  $\ell^\infty$  or max metric, and metrics with a unit disk that is a regular  $2n$ -gon.

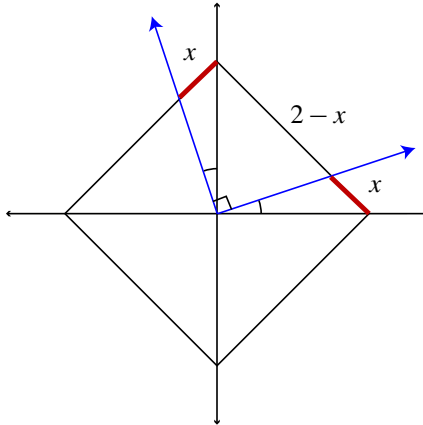
Length minimizing paths in the taxicab plane are not necessarily unique, so we use the vector space properties of  $\mathbb{R}^2$  and define *lines* to be the sets of points of the form  $L = \{t\mathbf{v} + \mathbf{b} \mid t \in \mathbb{R}\}$  for some fixed  $\mathbf{v}$  and  $\mathbf{b}$ . We can similarly define *line segments*, *triangles*, *rays*, and *angles* (pairs of rays sharing an initial point). We define the length of a line segment  $\overline{AB}$  to be the distance between the endpoints,  $d_T(A, B)$ .

Given a metric  $d$  on a set  $X$ , a circle  $C$  of radius  $r$  is the set of all points  $p \in X$  equidistant from a given point called the center. A circle in the taxicab metric is a square with diagonals parallel to the  $x$ - and  $y$ -axes. In Euclidean space there is an intrinsic notion of angle measure, radian measure, which is determined by the

---

MSC2010: 51F99, 52A21.

Keywords: taxicab trigonometry, Minkowski geometry.



**Figure 1.** Euclidean right angles have taxicab angle measure of 2.

length of a particular circle arc. We can similarly define an intrinsic angle measure in the taxicab plane, called *t-radians*.

**Definition 1.** Let  $C$  be a circle with radius  $r$  and center  $P$ . Given an angle with vertex  $P$ , let  $s$  be the length of the subtended arc. The *t-radian* measure,  $\theta$ , of a taxicab angle is given by

$$\theta = \frac{s}{r}.$$

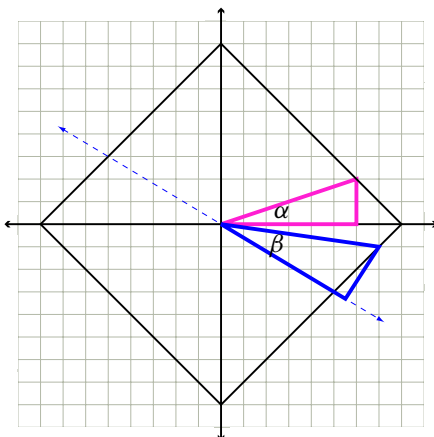
It is this notion of angle measure which was used in these previous works [Akça and Kaya 1997; Brisbin and Artola 1985; Thompson and Dray 2000] on taxicab trigonometry. Another well-studied angle measure in a Minkowski metric uses the *area* of the sector of the circle, rather than arc length, to define the angle measure. (Due to a theorem of Haar, any area measure  $\mu$  is proportional to Lebesgue measure; see [Álvarez Paiva and Thompson 2004] for a discussion of areas in normed spaces.) By Theorem 1 in [Düvelmeyer 2005], these two notions are equivalent (up to scale) because the taxicab circle is an example of an equiframed curve. See [Düvelmeyer 2005] for the definition of equiframed curve.

Note that an  $\ell^1$  circle has 8 *t-radians*, which means in this metric, 4 is the analogue of  $\pi$ . Some of the properties from Euclidean geometry have analogous statements which are true in the taxicab plane. We will use the following propositions.

**Proposition 2** [Thompson and Dray 2000, Theorem 4.2]. *The angle sum of a taxicab triangle is 4 t-radians.*

We define a taxicab right angle to be an angle with measure 2 *t-radians*, which, as in Euclidean geometry, is an angle which has measure equal to its supplement.

**Proposition 3** [Thompson and Dray 2000, Lemma 2.5]. *A Euclidean right angle has taxicab angle measure of 2 t-radians, and the converse is also true.*



**Figure 2.** An  $\ell^1$  circle with two right-angled triangles.

Figure 1 gives a sketch of a proof of Proposition 3.

Proposition 3 implies that the vectors  $\mathbf{x}$  and  $\mathbf{y}$  form a right angle in the taxicab plane if and only if they are orthogonal in the Euclidean sense. The study of different notions of orthogonality in Minkowski spaces is an active area of research. Two important orthogonality types in Minkowski spaces are Birkhoff orthogonality, ( $\mathbf{x} \perp \mathbf{y}$  if and only if  $\|\mathbf{x} - \alpha\mathbf{y}\| \geq \|\mathbf{x}\|$  for all  $\alpha$ ) and James (or isosceles) orthogonality ( $\mathbf{x} \perp \mathbf{y}$  if and only if  $\|\mathbf{x} + \mathbf{y}\| = \|\mathbf{x} - \mathbf{y}\|$ ). In the taxicab plane, Birkhoff orthogonality is not symmetric and James orthogonality is not invariant under scalar multiplication, which implies neither notion is equivalent to the definition of right angle that we use above; see the recent survey [Alonso et al. 2012] for an explanation of these facts and extensive discussion of orthogonality in normed linear spaces.

Not all angles in the taxicab geometry behave as nicely as right angles. In Figure 2, the Euclidean angles  $\alpha$  and  $\beta$  of the two triangles depicted are not equal, but the taxicab angle measure of both is  $\frac{1}{2}$ .

A taxicab right triangle is in *standard position* if the base of the triangle is parallel to the  $x$ -axis (see  $\alpha$ -triangle in Figure 2). For triangles in standard position, we can define the taxicab sine and cosine functions as we do in Euclidean geometry with the  $\cos \theta$  and  $\sin \theta$  equal to the  $x$ - and  $y$ -coordinates on the unit circle. Indeed, the piecewise linear formulas for these functions are given in [Thompson and Dray 2000; Akça and Kaya 1997] and with slightly different formulas in [Brisbin and Artola 1985]. However, if we define sine and cosine as ratios of sides of right triangles, considering only triangles in standard position will not give all possible values. To illustrate this, we refer again to Figure 2.

Both triangles are right triangles with hypotenuse (the side opposite the 2-radian angle) of length 1. Also, since  $\alpha$  and  $\beta$  both have angle measure  $\frac{1}{2}$ , the other

nonright angle is  $4 - 2 - \frac{1}{2} = \frac{3}{2}$ . In the  $\alpha$ -triangle, we compute the cosine of  $\alpha$  by taking the ratio of the lengths of the adjacent side and the hypotenuse, which is  $\frac{3}{4}$ . However, looking at the  $\beta$ -triangle, we see the vertex of the right angle falls outside of the unit circle, which implies that the length of the side adjacent to  $\beta$ , and therefore the cosine of  $\beta$ , is *greater* than 1.

A natural question arises: what is the maximum value of the cosine of an angle in the taxicab plane? In this paper, we define and give explicit formulas for two-parameter sine and cosine functions, describing the possible side ratios of right triangles in the taxicab plane. Using these formulas, we show the maximum value to be  $1/2 + 1/\sqrt{2}$ , which is greater than 1. Thus we obtain a quantitative measure of a difference between the Euclidean and taxicab plane.

We would like to thank the referee for pointing out many references on the geometry of Minkowski metric spaces, including [Thompson 1996]. In Chapter 8 of this text, Thompson defines two-parameter sine and cosine functions for general Minkowski spaces. For Thompson's function, the Minkowski cosine of two vectors is zero if and only if the vectors  $x_1$  and  $x_2$  are Birkhoff orthogonal. This property does not hold for our definition of cosine, so our functions are not a special case of those defined by Thompson, even up to scale. Using the sine function, Thompson defines an  $\alpha$  which measures how far the Minkowski space is from Euclidean space, leaving us with a question: is this  $\alpha$  related to the value we obtain for the maximum of our taxicab sine function?

## 2. A two-parameter sine and cosine function

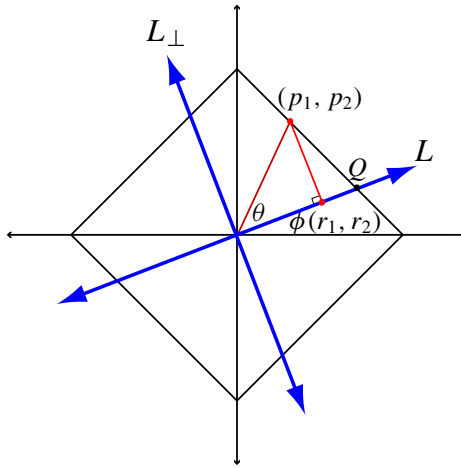
**Definition 4.** Given two metric spaces  $(X, d_1)$  and  $(Y, d_2)$ , a bijection  $f : X \rightarrow Y$  is an *isometry* if for any two points  $p, q \in X$ ,

$$d_1(p, q) = d_2(f(p), f(q)).$$

Given a metric space  $X$ , the set of all isometries  $\phi : X \rightarrow X$  forms a group, and the set of isometries that fix a point forms a subgroup of this group. An important subgroup is the set of isometries which fix the origin, which, by the Mazur–Ulam theorem (see [Thompson 1996, Chapter 3]), are linear. Using this fact and the fact that isometries map circles to circles with the same radius, one can see that the group of isometries that fix the origin of  $(\mathbb{R}^2, d_T)$  is the group of symmetries of a square, also called the dihedral group  $D_4$ . This includes the set of rotations (by  $0^\circ$ ,  $90^\circ$ ,  $180^\circ$ , and  $270^\circ$ ) and reflections across the  $x$ -axis,  $y$ -axis and the lines passing through the origin with slope  $\pm 1$ . The full group of isometries is the semidirect product  $D_4 \rtimes \mathbb{R}^2$ , which is proved in [Schattschneider 1984]. This group is generated by translations and isometries that fix the origin.

Two triangles  $T_1, T_2$  in the taxicab plane are *congruent* if there is a taxicab isometry  $\phi$  such that  $\phi(T_1) = T_2$ . Note that due to the rigidity of the isometry group,





**Figure 3.** Defining sine and cosine.

there is no taxicab isometry taking the  $\alpha$ -triangle in Figure 2 to the  $\beta$ -triangle, so there is no angle-side-angle theorem in taxicab geometry. We will define the taxicab sine and cosine functions to have two angle parameters; one parameter is the usual  $\theta$ -angle parameter measured from a fixed axis, and the other  $\phi$ -parameter will denote the “direction” of the triangle in the plane (see Figure 3).

Before giving the definition, we describe a notion of orthogonal projection in the taxicab plane. Let  $L$  be a line and  $P$  be a point. If  $P$  is on  $L$ , the *orthogonal projection of  $P$  onto  $L$*  is  $P$ . If  $P$  is not on  $L$ , the orthogonal projection is a unique point  $R$  on  $L$  for which the line segment  $\overline{OPR}$  makes a Euclidean right angle with  $L$ ; Proposition 3 implies that this point  $R$  is also the unique point on  $L$  which makes a taxicab right angle. The following definition, which is convenient for later proofs, may seem somewhat unnatural; we refer the reader to Propositions 6 and 7 which justify that this definition gives the desired “signed ratio” of side lengths.

**Definition 5.** Let  $L$  be the line through the origin  $O$  which makes reference angle  $\phi$  with the  $x$ -axis, where  $0 \leq \phi < 2$ , and let  $P = (p_1, p_2)$  be a point on the unit circle so that  $\overline{OP}$  makes angle  $\theta$  with  $L$ . Let  $R = (r_1, r_2)$  be the orthogonal projection of  $P$  onto  $L$ . We define the taxicab cosine and sine of angle  $\theta$  at reference angle  $\phi$  as

$$\text{tcos}_\phi \theta = r_1 + r_2, \quad \text{tsin}_\phi \theta = (r_1 - p_1) + (p_2 - r_2).$$

Given a right triangle  $T$  with hypotenuse of length 1, there is a taxicab isometry which maps  $T$  to a triangle of the form  $\triangle PRO$  given in Definition 5, so  $T$  is congruent to  $\triangle PRO$ .

Let  $L_\perp$  be the perpendicular to  $L$  which also passes through the origin. The lines  $L$  and  $L_\perp$  divide the plane into four quadrants, which we number I, II, III, IV in the usual way.

**Proposition 6.** *The value of  $t\cos_\phi \theta$  is positive for  $\theta$  in  $L$ - $L_\perp$  quadrants I and IV, and negative for  $\theta$  in quadrants II and III. Similarly  $t\sin_\phi \theta$  is positive for  $\theta$  in quadrants I and II, and negative for  $\theta$  in quadrants III and IV.*

*Proof.* Let  $P = (p_1, p_2)$  and  $R = (r_1, r_2)$  be as given in Definition 5. When  $\theta$  is in quadrants I and IV, as defined by  $L$  and  $L_\perp$ , the coordinate  $r_1$  is positive and  $r_2$  is nonnegative (when  $\phi = 0$ , the line  $L$  is the x-axis and  $r_2 = 0$ ). Therefore  $t\cos_\phi \theta$ , which is the sum of these coordinates, is positive. Similarly, when  $\theta$  is in quadrants II and III,  $r_1$  is negative and  $r_2$  is nonpositive; hence  $t\cos_\phi \theta$  is negative.

Recall that  $t\sin_\phi \theta = (r_1 - p_1) + (p_2 - r_2)$ . For a fixed  $\phi$ , the coordinates of  $P$  and  $R$  are continuous real-valued functions of  $\theta$ , and therefore the functions  $r_1 - p_1$  and  $p_2 - r_2$  are also continuous functions. When  $0 < \phi < 2$ , each of these functions is zero if and only if  $\theta = 4n$  for some integer  $n$ . This follows from the fact that the slope of  $L$  is positive, which implies that the line through  $P$  and  $R$  has negative slope; so  $p_1 = r_1$  or  $p_2 = r_2$  if and only if  $P = R$ . Therefore the sign of each of these functions,  $r_1 - p_1$  and  $p_2 - r_2$ , is constant for  $\theta$  in quadrants I and II. Picking a specific angle such as  $\theta = 2$  allows us to verify that both are positive, and therefore  $t\sin_\phi \theta$  is positive. Choosing an angle in the range  $4 < \theta < 8$  shows that both of these functions are negative, and therefore  $t\sin_\phi \theta$  is also negative when  $\theta$  is in quadrants III and IV.

When  $\phi = 0$ , we have that  $r_2 = 0$  and  $r_1 = p_1$ ; then  $t\sin_\phi \theta = p_2$ , and the result follows.  $\square$

**Proposition 7.** *In the right triangle made by  $P$ ,  $R$  and the origin  $O$ ,  $|t\cos_\phi \theta|$  gives the length of the side adjacent to  $\theta$ , and  $|t\sin_\phi \theta|$  gives the length of the opposite side.*

*Proof.* Fix an angle  $0 \leq \phi < 2$ . The length of the adjacent side is the distance from  $R$  to the origin, which is  $|r_1| + |r_2|$ . When  $\theta$  is in quadrants I and IV (defined by  $L$  and  $L_\perp$ ), both  $r_1$  and  $r_2$  are nonnegative, so

$$|r_1| + |r_2| = r_1 + r_2 = |t\cos_\phi \theta|.$$

When  $\theta$  lies in quadrants II and III, both  $r_1$  and  $r_2$  are nonpositive, so

$$|r_1| + |r_2| = -r_1 - r_2 = -(r_1 + r_2) = |t\cos_\phi \theta|.$$

The length of the side opposite of  $\theta$  in triangle  $OPR$  is given by the distance between  $P$  and  $R$ , which is  $|p_1 - r_1| + |p_2 - r_2|$ . Arguing as in Proposition 6, when  $\theta$  is in quadrants I and II, we have

$$|p_1 - r_1| + |p_2 - r_2| = (r_1 - p_1) + (p_2 - r_2) = |t\sin_\phi \theta|,$$

and when  $\theta$  is in quadrants III and IV,

$$\begin{aligned} |p_1 - r_1| + |p_2 - r_2| &= -(r_1 - p_1) - (p_2 - r_2) \\ &= -((r_1 - p_1) + (p_2 - r_2)) = |\operatorname{tsin}_\phi \theta|. \quad \square \end{aligned}$$

**Proposition 8.** *The following identities hold.*

$$\operatorname{tsin}_\phi(\theta - 4) = -\operatorname{tsin}_\phi \theta \quad \text{and} \quad \operatorname{tcos}_\phi(\theta - 4) = -\operatorname{tcos}_\phi \theta.$$

*Proof.* Let  $P$  and  $R$  be the points given in [Definition 5](#) corresponding to  $\theta$ , and  $P'$  and  $R'$  the points corresponding to  $\theta - 4$ . By [Proposition 3](#), taxicab angles of measure 2 are Euclidean right angles, which means  $P$  and  $P'$  are antipodal points on the unit circle and  $P' = -P$ . The map  $(x, y) \rightarrow (-x, -y)$  is an isometry of the taxicab plane which maps  $P$  to  $P'$ . Angles are defined by the metric, and therefore isometries preserve angle measure. It follows from the definition of  $R$  that  $R' = -R$ . Therefore,

$$\operatorname{tcos}_\phi(\theta - 4) = -r_1 - r_2 = -(r_1 + r_2) = -\operatorname{tcos}_\phi \theta$$

and

$$\begin{aligned} \operatorname{tsin}_\phi(\theta - 4) &= (-r_1 + p_1) + (-p_2 + r_2) \\ &= -[(r_1 - p_1) + (p_2 - r_2)] = -\operatorname{tsin}_\phi \theta. \quad \square \end{aligned}$$

### 3. Explicit formulas for sine and cosine functions

**Theorem 9.** *Let  $\phi$  be a taxicab reference angle such that  $0 \leq \phi < 2$  and let  $\theta$  be a taxicab angle measured relative to  $\phi$ . Let*

$$\alpha = \frac{1}{\phi^2 - 2\phi + 2},$$

which is well-defined for all  $\phi$  since  $\phi^2 - 2\phi + 2 > 0$ . The sine and cosine of  $\theta$  with reference angle  $\phi$  are given by

$$\operatorname{tsin}_\phi \theta = \begin{cases} \alpha \theta & \text{if } -\phi \leq \theta \leq 2 - \phi, \\ 1 + \alpha(\theta - 2)(\phi - 1) & \text{if } 2 - \phi \leq \theta \leq 4 - \phi, \\ \alpha(4 - \theta) & \text{if } 4 - \phi \leq \theta \leq 6 - \phi, \\ -1 + \alpha(6 - \theta)(\phi - 1) & \text{if } 6 - \phi \leq \theta \leq 8 - \phi, \end{cases}$$

and

$$\operatorname{tcos}_\phi \theta = \begin{cases} 1 + \alpha \theta(\phi - 1) & \text{if } -\phi \leq \theta \leq 2 - \phi, \\ \alpha(2 - \theta) & \text{if } 2 - \phi \leq \theta \leq 4 - \phi, \\ -1 + \alpha(4 - \theta)(\phi - 1) & \text{if } 4 - \phi \leq \theta \leq 6 - \phi, \\ \alpha(\theta - 6) & \text{if } 6 - \phi \leq \theta \leq 8 - \phi. \end{cases}$$

**Lemma 10.** *Let  $L$  be a line through the origin that makes angle  $\phi$  with the  $x$ -axis, where  $0 \leq \phi < 2$ . The point of intersection between  $L$  and the unit taxicab circle is*

$$Q = \left( \frac{2-\phi}{2}, \frac{\phi}{2} \right).$$

*Proof.* Let  $Q = (q_1, q_2)$ . Since  $Q$  lies on the unit circle and  $0 \leq \phi < 2$ , both coordinates are positive and

$$q_1 + q_2 = 1. \tag{1}$$

Since the radius of the unit circle is 1, the definition of angle implies that  $\phi$  is the distance between  $Q$  and  $(1, 0)$ . This distance is given by

$$|q_1 - 1| + |q_2 - 0| = 1 - q_1 + q_2 = \phi. \tag{2}$$

We solve the system of linear equations consisting of (1) and (2) for  $q_2$  by adding the two equations to get

$$q_2 = \frac{\phi}{2};$$

substituting  $q_2$  into (1) gives us  $q_1 = 1 - \phi/2$ , which is the desired result.  $\square$

**3.1. Proof of Theorem 9 for  $-\phi \leq \theta \leq 2 - \phi$ .** Let  $0 \leq \phi < 2$  and  $-\phi \leq \theta \leq 2 - \phi$ . We will determine the coordinates of  $P$  and  $R$ , given in Definition 5, as functions of  $\phi$  and  $\theta$ . Lemma 10 implies that the  $\phi$ -axis (line  $L$  in Figure 3) intersects the circle at

$$Q = \left( \frac{2-\phi}{2}, \frac{\phi}{2} \right).$$

Since the  $\phi$ -axis passes through the origin, we find that the equation is

$$L(x) = \frac{\phi}{2-\phi} x. \tag{3}$$

Next, we determine the coordinates of  $P$ , the point of intersection between the circle and the  $(\theta + \phi)$ -ray. Applying Lemma 10 again with angle  $\theta + \phi$  gives coordinates

$$P = \left( \frac{2-\phi-\theta}{2}, \frac{\phi+\theta}{2} \right).$$

Proposition 3 implies that Euclidean right angles are taxicab right angles. Therefore, to find the point  $R$  we determine the equation of the line perpendicular (in the usual Euclidean sense) to the  $\phi$ -axis,  $L_P$ , through point  $P$ . Since the  $\phi$ -axis has slope  $\phi/(2-\phi)$ ,  $L_P$  has slope  $(\phi-2)/\phi$ . Since we know the coordinates of

$P = (p_1, p_2)$  and the slope, we can determine the equation for  $L_P$ , which is

$$\begin{aligned} L_P(x) &= \left( \frac{\phi - 2}{\phi} \right) (x - p_1) + p_2 \\ &= \frac{(\phi - 2)x}{\phi} + \frac{(\phi - 2)(\theta + \phi - 2) + \phi(\theta + \phi)}{2\phi}. \end{aligned} \quad (4)$$

The point  $R$  is the intersection between the  $\phi$ -axis and  $L_P$ . Setting equations (3) and (4) equal to each other and solving for the  $x$ -coordinate of  $R$  yields

$$r_1 = \frac{2 - \phi}{2} + \frac{(2 - \phi)(\phi\theta - \theta)}{2(\phi^2 - 2\phi + 2)}.$$

Plugging  $r_1$  into  $L$  (or  $L_P$ ) gives the  $y$ -coordinate of  $R$ ,

$$r_2 = \frac{\phi}{2} + \frac{\phi^2\theta - \phi\theta}{2(\phi^2 - 2\phi + 2)}.$$

Thus, the coordinates of  $R$  are

$$R = \left( \frac{2 - \phi}{2} + \frac{(2 - \phi)(\phi\theta - \theta)}{2(\phi^2 - 2\phi + 2)}, \frac{\phi}{2} + \frac{\phi^2\theta - \phi\theta}{2(\phi^2 - 2\phi + 2)} \right).$$

The result now follows by using the coordinates of  $R$  and  $P$  to compute  $\text{tsin}_\phi \theta$  and  $\text{tcos}_\phi \theta$  by the formulas given in [Definition 5](#).  $\square$

**3.2. Proof for  $2 - \phi \leq \theta \leq 4 - \phi$ .** We again find the coordinates of  $P$  and  $R$  to compute  $\text{tsin}_\phi \theta$  and  $\text{tcos}_\phi \theta$ . When  $2 < \theta + \phi < 4$ , the point  $P$  is in the second quadrant (as defined by the  $x$ - and  $y$ -axes). Let  $\theta_1$  be the portion of  $\theta$  measured from the  $y$ -axis, so  $\theta_1 = \phi + \theta - 2$ .

Let  $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  be the map defined by  $(x, y) \mapsto (y, -x)$ . This map is an order 4 isometry of the  $\ell^1$  metric. Note that  $f(0, 1) = (1, 0)$  and  $f(P)$  is in the first quadrant. Since angle measure is defined by the metric, angle measure is preserved by isometries. We can therefore apply [Lemma 10](#) to  $f(P)$  to obtain the coordinates

$$f(P) = \left( \frac{2 - \theta_1}{2}, \frac{\theta_1}{2} \right).$$

To obtain the coordinates for  $P$  we apply the inverse map:

$$P = f^{-1} \left( \frac{2 - \theta_1}{2}, \frac{\theta_1}{2} \right) = \left( -\frac{\theta_1}{2}, \frac{2 - \theta_1}{2} \right) = \left( \frac{2 - \phi - \theta}{2}, \frac{4 - \phi - \theta}{2} \right).$$

To finish the proof for this interval, we use the same procedure as in the proof for the first interval; that is, we find the equation of the line perpendicular to the  $\phi$ -axis through  $P$  to determine the coordinates of the point  $R$ . The line through  $P$

perpendicular to  $L(x)$  is

$$\begin{aligned} L_P(x) &= \left( \frac{\phi - 2}{\phi} \right) (x - p_1) + p_2 \\ &= \frac{(\phi - 2)x}{\phi} + \frac{(\phi - 2)(\theta + \phi - 2) + \phi(4 - \theta - \phi)}{2\phi}. \end{aligned} \quad (5)$$

To find  $r_1$ , we set equations (3) and (5) equal to one another and solve for  $x$ , which gives

$$r_1 = \frac{(\phi - 2)(\theta - 2)}{2(\phi^2 - 2\phi + 2)}.$$

Plugging  $r_1$  into  $L(x)$  (Equation (3)) gives

$$r_2 = \frac{-\phi(\theta - 2)}{2(\phi^2 - 2\phi + 2)}.$$

The sine and cosine functions can now be computed from the formulas given in Definition 5.  $\square$

**3.3. Proof for  $4 - \phi \leq \theta \leq 8 - \phi$ .** We will use the symmetry of the functions to establish the formulas for the third and fourth intervals. Let  $\theta$  be in the given interval, and  $\theta^* = \theta - 4$ . Then  $-\phi \leq \theta^* \leq 4 - \phi$ . We have determined formulas for  $\text{tsin}_\phi(\theta^*)$  and  $\text{tcos}_\phi(\theta^*)$  in this interval, so applying Proposition 8 gives formulas for angle  $\theta$  in the remaining two intervals.  $\square$

It should be noted that our formulas are a generalization of those formulas in [Thompson and Dray 2000; Akça and Kaya 1997]; if  $\phi = 0$ , then  $\theta$  is in standard position and we obtain identical formulas.

## 4. Properties of the functions

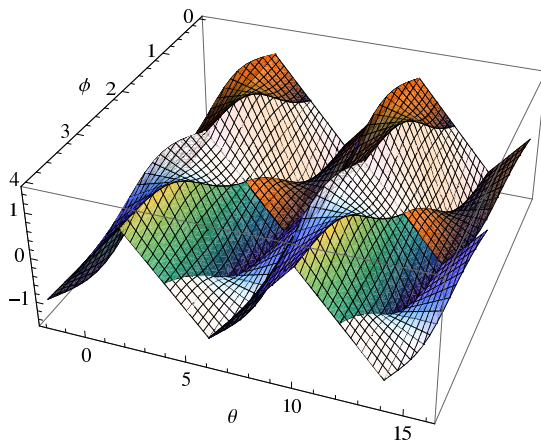
**4.1. Periodic extensions and graphs.** In Definition 5, the generalized sine and cosine functions were defined for all real numbers  $\theta$  and for values of  $\phi$  such that  $0 \leq \phi < 2$ . It is evident from the definition that the  $\theta$ -period of these functions is 8, so for any integer  $k$ ,

$$\text{tcos}_\phi(\theta + 8k) = \text{tcos}_\phi \theta \quad \text{and} \quad \text{tsin}_\phi(\theta + 8k) = \text{tsin}_\phi \theta.$$

There is a natural  $\phi$ -extension of these functions; since rotation by right angles gives isometries of the  $\ell^1$  metric, we extend the  $\phi$ -domain of the generalized sine and cosine functions to be  $\phi$ -periodic with period 2. Therefore, for any integer  $s$ ,

$$\text{tcos}_{\phi+2s} \theta = \text{tcos}_\phi \theta \quad \text{and} \quad \text{tsin}_{\phi+2s} \theta = \text{tsin}_\phi \theta.$$

It should be noted that the formulas for  $\text{tsin}_\phi \theta$  and  $\text{tcos}_\phi \theta$  given by  $P$  and  $R$  from Definition 5 are only valid for values of  $\phi$  in the first quadrant. Since Theorem 9



**Figure 4.** Graph of the generalized sine function.

gives explicit formulas for entire  $\phi$  and  $\theta$  periods, we may use this theorem and the two periodic properties stated above to give values for  $\text{tsin}_\phi \theta$  and  $\text{tcos}_\phi \theta$  for any  $(\phi, \theta) \in \mathbb{R} \times \mathbb{R}$ . Figure 4 contains a graph of  $\text{tsin}_\phi \theta$  for two periods of  $\phi$  and two periods of  $\theta$ .

Table 1 shows a family of cross-sections. Referring to the formulas in Theorem 9, we see that for a fixed  $\phi$  these functions are piecewise linear. We invite the interested reader to verify that these functions are constant when  $\theta = 2n$  for some integer  $n$ .

Recall that in the Euclidean metric,  $\sin(\theta + \pi/2) = \cos \theta$ . The cross-sections for the sine and cosine functions when  $\phi$  is fixed suggest a similar identity.

**Proposition 11.**  $\text{tsin}_\phi(\theta + 2) = \text{tcos}_\phi \theta$ .

*Proof.* While this identity follows from the symmetry of the space, Theorem 9 gives explicit formulas for  $\text{tsin}_\phi \theta$  and  $\text{tcos}_\phi \theta$ , so we need only check the formulas to verify this identity. Assume that  $0 \leq \phi < 2$  and  $-\phi \leq \theta \leq 2 - \phi$ , which implies  $2 - \phi \leq \theta + 2 \leq 4 - \phi$ . For angles in the interval  $[2 - \phi, 4 - \phi]$ ,

$$\text{tsin}_\phi \theta = 1 + \alpha(\theta - 2)(\phi - 1).$$

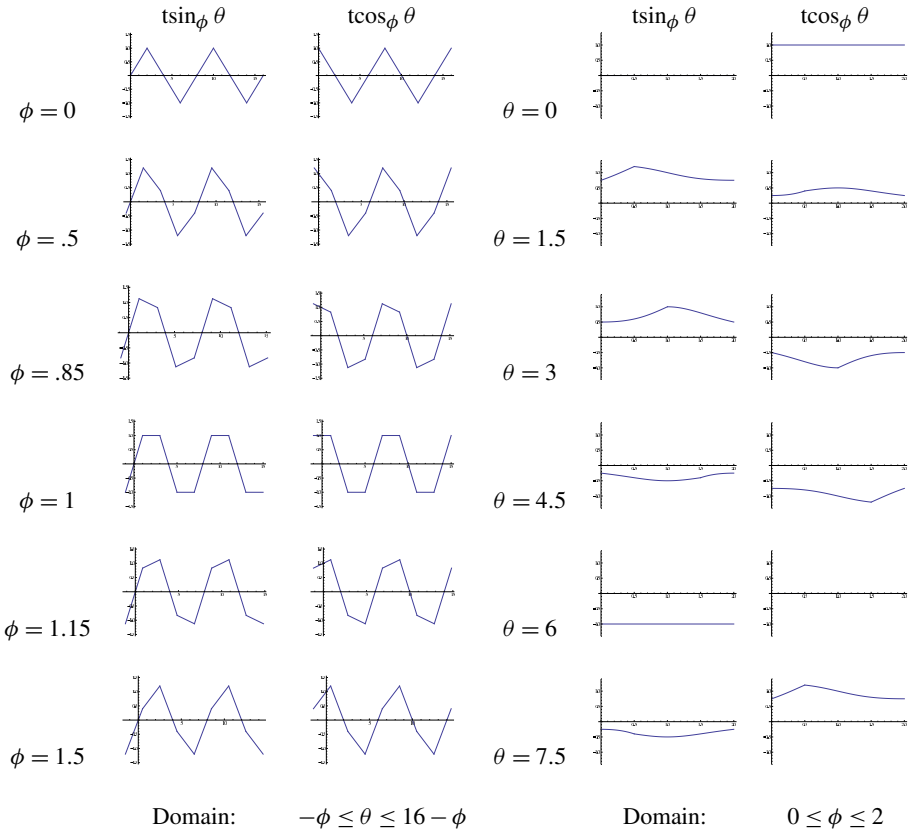
Therefore,

$$\text{tsin}_\phi(\theta + 2) = 1 + \alpha((\theta + 2) - 2)(\phi - 1) = 1 + \alpha \theta(\phi - 1),$$

which is equal to  $\text{tcos}_\phi \theta$  when  $-\phi \leq \theta \leq 2 - \phi$ . The other intervals can be verified similarly. □

**4.2. Maximum and minimum values.**

**Theorem 12.** *The maximum value of  $\text{tsin}_\phi \theta$  and  $\text{tcos}_\phi \theta$  is  $1/2 + 1/\sqrt{2}$ ; the minimum value is  $-(1/2 + 1/\sqrt{2})$ .*



**Table 1.** Cross sections.

*Proof.* By Proposition 11, the maximum of the sine function is equal to the maximum of the cosine function. Also, by Proposition 8, the minimum of the sine function is equal to the negative of the maximum. Therefore it is sufficient to verify the maximum of the sine function.

The sine function has a  $\theta$ -period of 8 and a  $\phi$ -period of 2. However, the maximum of the sine function must occur when sine is positive, and hence  $\theta$  must be in the interval  $[0, 4]$  by Proposition 6. It is therefore sufficient to find the maximum of  $\text{tsin}_\phi \theta$  on the region defined by  $0 \leq \phi \leq 2$  and  $0 \leq \theta \leq 4$ . We will use standard techniques from multivariable calculus to maximize this function.

As  $\text{tsin}_\phi \theta$  is piecewise defined, we will consider the intervals

$$[0, 2 - \phi], \quad [2 - \phi, 4 - \phi], \quad \text{and} \quad [4 - \phi, 4].$$

Recall that

$$\alpha = \frac{1}{\phi^2 - 2\phi + 2} = \frac{1}{(\phi - 1)^2 + 1},$$



which is positive for all  $\phi$ . When  $\theta$  is in the interval  $[0, 2 - \phi]$ , we have  $\text{tsin}_\phi \theta = \alpha\theta$ , and  $\theta$  in  $[4 - \phi, 4]$  implies  $\text{tsin}_\phi \theta = \alpha(4 - \theta)$ . The partial derivatives with respect to  $\theta$  of these functions are  $\alpha$  and  $-\alpha$ ; therefore,  $\text{tsin}_\phi \theta$  is increasing with respect to  $\theta$  on  $[0, 2 - \phi]$  and decreasing in  $\theta$  on  $[4 - \phi, 4]$ . This implies the absolute maximum of  $\text{tsin}_\phi \theta$  occurs when  $\theta$  is in the middle interval.

When  $2 - \phi \leq \theta \leq 4 - \phi$ ,

$$\text{tsin}_\phi \theta = 1 + \frac{(\theta - 2)(\phi - 1)}{\phi^2 - 2\phi + 2}.$$

The partial derivatives are

$$\begin{aligned} \frac{\partial}{\partial \phi} \left[ 1 + \frac{(\theta - 2)(\phi - 1)}{\phi^2 - 2\phi + 2} \right] &= \frac{(2\phi - \phi^2)(\theta - 2)}{(\phi^2 - 2\phi + 2)^2}, \\ \frac{\partial}{\partial \theta} \left[ 1 + \frac{(\theta - 2)(\phi - 1)}{\phi^2 - 2\phi + 2} \right] &= \frac{\phi - 1}{\phi^2 - 2\phi + 2}. \end{aligned}$$

These are both zero only when  $(\phi, \theta) = (1, 2)$ . In this case,  $\text{tsin}_1(2) = 1$ . We now check the boundary conditions.

When  $\phi = 0$ , we have  $2 \leq \theta \leq 4$  and  $\text{tsin}_\phi \theta = 2 + (-\theta/2)$ , which has a maximum of 1. Note that  $\text{tsin}_\phi \theta$  has the same maximum when  $\phi = 2$  because of the  $\phi$ -periodic property previously stated.

When  $\theta = 2 - \phi$ , we have

$$g(\phi) = \text{tsin}_\phi(2 - \phi) = 1 - \frac{(\phi - 1)\phi}{\phi^2 - 2\phi + 2}.$$

The derivative of this function is

$$g'(\phi) = \frac{\phi^2 - 4\phi + 2}{(\phi^2 - 2\phi + 2)^2}.$$

This function is zero when  $\phi = 2 \pm \sqrt{2}$ . Only one of these values,  $\phi = 2 - \sqrt{2}$ , is in the region under consideration. For this value of  $\phi$ , we have  $\theta = \sqrt{2}$  and we see the value of the sine function is

$$\text{tsin}_{2-\sqrt{2}} \sqrt{2} = 1/2 + 1/\sqrt{2}.$$

When  $\theta = 4 - \phi$ , we have

$$h(\phi) = \text{tsin}_\phi(4 - \phi) = 1 - \frac{(\phi - 2)(\phi - 1)}{\phi^2 - 2\phi + 2}.$$

The derivative of this function is

$$h'(\phi) = \frac{2 - \phi^2}{(\phi^2 - 2\phi + 2)^2}.$$

For values of  $\phi$  in the interval  $[0, 2]$ , this derivative is zero when  $\phi = \sqrt{2}$ . Then  $\theta = 4 - \sqrt{2}$ , and

$$\text{tsin}_{\sqrt{2}}(4 - \sqrt{2}) = \frac{1}{2} + \frac{1}{\sqrt{2}}.$$

We can therefore conclude for values in the region  $0 \leq \phi \leq 2$  and  $0 \leq \theta \leq 4$ , the function  $\text{tsin}_{\phi} \theta$  achieves its absolute maximum,  $1/2 + 1/\sqrt{2}$ , in two locations:  $(2 - \sqrt{2}, \sqrt{2})$  and  $(\sqrt{2}, 4 - \sqrt{2})$ .  $\square$

**Corollary 13.** *The hypotenuse of a right triangle in taxicab space is not always the longest side of the triangle.*

## References

- [Akça and Kaya 1997] Z. Akça and R. Kaya, “On the taxicab trigonometry”, *J. Inst. Math. Comput. Sci. Math. Ser.* **10**:3 (1997), 151–159. [MR 99c:51022](#) [Zbl 0926.51026](#)
- [Alonso et al. 2012] J. Alonso, H. Martini, and S. Wu, “On Birkhoff orthogonality and isosceles orthogonality in normed linear spaces”, *Aequationes Math.* **83**:1-2 (2012), 153–189. [MR 2012m:46001](#) [Zbl 1241.46006](#)
- [Álvarez Paiva and Thompson 2004] J. C. Álvarez Paiva and A. C. Thompson, “Volumes on normed and Finsler spaces”, pp. 1–48 in *A sampler of Riemann–Finsler geometry*, Math. Sci. Res. Inst. Publ. **50**, Cambridge Univ. Press, 2004. [MR 2006c:53079](#) [Zbl 1078.53072](#)
- [Álvarez Paiva and Thompson 2005] J. C. Álvarez Paiva and A. Thompson, “On the perimeter and area of the unit disc”, *Amer. Math. Monthly* **112**:2 (2005), 141–154. [MR 2005i:51012](#) [Zbl 1084.52007](#)
- [Brisbin and Artola 1985] R. Brisbin and P. Artola, “Taxicab trigonometry”, *Pi Mu Epsilon Journal* **8**:2 (1985), 89–95.
- [Düvelmeyer 2005] N. Düvelmeyer, “Angle measures and bisectors in Minkowski planes”, *Canad. Math. Bull.* **48**:4 (2005), 523–534. [MR 2006g:52008](#) [Zbl 1093.52002](#)
- [Schattschneider 1984] D. J. Schattschneider, “The taxicab group”, *Amer. Math. Monthly* **91**:7 (1984), 423–428. [MR 86b:51027](#) [Zbl 0564.51005](#)
- [Thompson 1996] A. C. Thompson, *Minkowski geometry*, Encyclopedia of Mathematics and its Applications **63**, Cambridge University Press, Cambridge, 1996. [MR 97f:52001](#) [Zbl 0868.52001](#)
- [Thompson and Dray 2000] K. Thompson and T. Dray, “Taxicab angles and trigonometry”, *Pi Mu Epsilon Journal* **11**:2 (2000), 87–96.

Received: 2013-06-24

Revised: 2013-11-06

Accepted: 2013-11-07

[kelly.delp@gmail.com](mailto:kelly.delp@gmail.com)

*Department of Mathematics, Ithaca College,  
201 Muller Center, Ithaca, NY 14850, United States*

[mgfilips@buffalo.edu](mailto:mgfilips@buffalo.edu)

*Department of Mathematics, University at Buffalo,  
244 Mathematics Building, Buffalo, NY 14260, United States*

# ${}_3F_2$ -hypergeometric functions and supersingular elliptic curves

Sarah Pitman

(Communicated by Ken Ono)

In recent work, Monks described the supersingular locus of families of elliptic curves in terms of  ${}_2F_1$ -hypergeometric functions. We lift his work to the level of  ${}_3F_2$ -hypergeometric functions by means of classical transformation laws and a theorem of Clausen.

## 1. Introduction and statement of results

Dating back to the works of Gauss, hypergeometric functions play an important role in mathematics. More recently, these complex functions and their analogs have been studied in terms of the complex periods of elliptic curves. The purpose of this paper is to further develop these sorts of connections. We begin by setting the notation and defining the hypergeometric functions which will be used throughout. If  $n$  is a nonnegative integer, we recall the Pochhammer symbol  $(\gamma)_n$ , defined by

$$(\gamma)_n := \begin{cases} 1 & \text{if } n = 0, \\ \gamma(\gamma + 1)(\gamma + 2) \cdots (\gamma + n - 1) & \text{if } n \geq 1. \end{cases}$$

The *classical hypergeometric function* in parameters  $\alpha_1, \dots, \alpha_h, \beta_1, \dots, \beta_j \in \mathbb{C}$  is defined by

$${}_hF_j^{\text{cl}} \left( \begin{matrix} \alpha_1 & \alpha_2 & \cdots & \alpha_h \\ \beta_1 & \cdots & \beta_j \end{matrix} \middle| x \right) := \sum_{n=0}^{\infty} \frac{(\alpha_1)_n (\alpha_2)_n (\alpha_3)_n \cdots (\alpha_h)_n}{(\beta_1)_n (\beta_2)_n \cdots (\beta_j)_n} \cdot \frac{x^n}{n!}.$$

We are interested in the hypergeometric functions

$${}_2F_1^{\text{cl}} \left( \begin{matrix} a & b \\ c \end{matrix} \middle| x \right) := \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \cdot \frac{x^n}{n!} \tag{1-1}$$

and

$${}_3F_2^{\text{cl}} \left( \begin{matrix} a & b & d \\ c & e \end{matrix} \middle| x \right) := \sum_{n=0}^{\infty} \frac{(a)_n (b)_n (d)_n}{(c)_n (e)_n} \cdot \frac{x^n}{n!}, \tag{1-2}$$

---

*MSC2010:* 11G20, 33C20.

*Keywords:* hypergeometric functions, supersingular, elliptic curves.

and their truncations modulo primes  $p$ . For any odd prime  $p$ , we define these truncations by

$${}_2F_1^{\text{tr}}\left(\begin{matrix} a & b \\ c \end{matrix} \middle| x\right)_p \equiv \sum_{n=0}^{(p-1)/2} \frac{(a)_n(b)_n}{(c)_n} \cdot \frac{x^n}{n!} \pmod{p} \tag{1-3}$$

and

$${}_3F_2^{\text{tr}}\left(\begin{matrix} a & b & d \\ c & e \end{matrix} \middle| x\right)_p \equiv \sum_{n=0}^{(p-1)/2} \frac{(a)_n(b)_n(d)_n}{(c)_n(e)_n} \cdot \frac{x^n}{n!} \pmod{p}. \tag{1-4}$$

Monks [2012] studied elliptic curves and their relation to  ${}_2F_1^{\text{tr}}$ -hypergeometric functions and proved that these polynomials give the supersingular loci of certain families of elliptic curves. Here we lift his work from  ${}_2F_1^{\text{tr}}$ - to  ${}_3F_2^{\text{tr}}$ -hypergeometric functions and establish a similar result for these hypergeometric functions with additional parameters.

**Remark.** We note that above,  $\text{tr}$  denotes the truncation of a hypergeometric series after  $x^{(p-1)/2}$ , but in [Monks 2012],  $\text{tr}$  implies truncation after  $x^{p-1}$ . We will see that the relevant polynomials agree when reduced modulo  $p$ .

Let  $p$  be an odd prime and let  $\mathbb{F}$  be a field of characteristic  $p$ . An elliptic curve  $E/\mathbb{F}$  is said to be *supersingular* if it has no  $p$ -torsion over  $\bar{\mathbb{F}}$ . In other words, there is no element of order  $p$  in the group  $E(\bar{\mathbb{F}})$ . This condition is dependent only on the  $j$ -invariant of  $E$ . There are only finitely many isomorphism classes of supersingular elliptic curves in  $\bar{\mathbb{F}}_p$ , which Kaneko and Zagier [1998] determined using the theory of modular forms.

Here we consider supersingular elliptic curves in certain families. A well-known subfamily of elliptic curves is the Legendre family, which is denoted by

$$E_{1/2}(\lambda) : y^2 = x(x - 1)(x - \lambda)$$

for  $\lambda \neq 0, 1$ . These curves can be studied by means of the *supersingular locus*

$$S_{p,1/2}(\lambda) := \prod_{\substack{\lambda_0 \in \bar{\mathbb{F}}_p \\ \text{supersingular } E_{1/2}(\lambda_0)}} (\lambda - \lambda_0).$$

These polynomials have coefficients in  $\mathbb{F}_p$ .

El-Guindy and Ono [2013] studied the family of elliptic curves defined by

$$E_{1/4}(\lambda) : y^2 = (x - 1)(x^2 + \lambda). \tag{1-5}$$

We also consider the following families of elliptic curves:

$$E_{1/3}(\lambda) : y^2 + \lambda y x + \lambda^2 y = x^3,$$

$$E_{1/12}(\lambda) : y^2 = 4x^3 - 27\lambda x - 27\lambda.$$

For  $i \in \{\frac{1}{4}, \frac{1}{3}, \frac{1}{12}\}$  and all primes  $p \geq 5$ , we let

$$S_{p,i}(\lambda) := \prod_{\substack{\lambda_0 \in \mathbb{F}_p \\ \text{supersingular } E_i(\lambda_0)}} (\lambda - \lambda_0). \tag{1-6}$$

Monks [2012] studied these families with respect to hypergeometric functions, and he showed that their supersingular loci are given by certain  ${}_2F_1$ -hypergeometric functions reduced modulo  $p$ . We extend these results of Monks, El-Guindy, and Ono to prove the following theorem. Assume the notation above.

**Theorem 1.1.** *The following are true:*

(1) *If  $p \geq 5$  is prime, then*

$$S_{p,1/4}(x)^2 \equiv (x + 1)^{(p-1)/2} \cdot {}_3F_2^{\text{tr}}\left(\begin{matrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 1 & 1 \end{matrix} \middle| \frac{x}{x+1}\right)_p \pmod{p}.$$

(2) *If  $p \geq 5$  is prime, then*

$$S_{p,1/3}(x)^2 \equiv x^{2 \cdot \lfloor p/3 \rfloor} \cdot {}_3F_2^{\text{tr}}\left(\begin{matrix} \frac{1}{3} & \frac{2}{3} & \frac{1}{2} \\ 1 & 1 \end{matrix} \middle| \frac{108x - 2916}{x^2}\right)_p \pmod{p}.$$

(3) *If  $p \geq 5$  is prime, then*

$$S_{p,1/12}(x)^2 \equiv (c_p^{-1})^2 \cdot x^{\lfloor p/6 \rfloor} \cdot {}_3F_2^{\text{tr}}\left(\begin{matrix} \frac{1}{6} & \frac{5}{6} & \frac{1}{2} \\ 1 & 1 \end{matrix} \middle| 1 - \frac{1}{x}\right)_p \pmod{p}.$$

Here

$$c_p = \left( \begin{matrix} 6 \lfloor \frac{p}{12} \rfloor + d_p \\ \lfloor \frac{p}{12} \rfloor \end{matrix} \right)$$

and  $d_p = 0, 2, 2, 4$  for  $p \equiv 1, 5, 7, 11 \pmod{12}$  respectively.

## 2. Nuts and bolts

**Statement of Clausen’s theorem and transformation laws.** Our main tools for establishing these congruences are a theorem of Clausen and two classical  ${}_2F_1^{\text{cl}}$  transformation laws. We make use of Clausen’s theorem [Bailey 1935] which gives the following equality of hypergeometric polynomials:

$${}_3F_2^{\text{cl}}\left(\begin{matrix} 2\alpha & 2\beta & \alpha + \beta \\ 2\alpha + 2\beta & \alpha + \beta + \frac{1}{2} \end{matrix} \middle| x\right) = {}_2F_1^{\text{cl}}\left(\begin{matrix} \alpha & \beta \\ \alpha + \beta + \frac{1}{2} \end{matrix} \middle| x\right)^2. \tag{2-1}$$

We also use two transformation laws in our proof so that we can apply (2-1) to the hypergeometric functions. The first, given in [Bailey 1935], states that

$${}_2F_1^{\text{cl}}\left(\begin{matrix} a & b \\ c \end{matrix} \middle| x\right) = (1-x)^{-a} \cdot {}_2F_1^{\text{cl}}\left(\begin{matrix} a & c-b \\ c \end{matrix} \middle| \frac{x}{x-1}\right). \tag{2-2}$$

The second, from Vidūnas [2009], gives that

$${}_2F_1^{\text{cl}}\left(a \quad \frac{b}{\frac{a+b+1}{2}} \mid x\right) = {}_2F_1^{\text{cl}}\left(\frac{a}{2} \quad \frac{\frac{b}{2}}{\frac{a+b+1}{2}} \mid 4x(1-x)\right). \tag{2-3}$$

**Elementary reduction modulo  $p$ .** By definition (1-4), we have that

$${}_3F_2^{\text{tr}}\left(\frac{1}{3} \quad \frac{2}{3} \quad \frac{1}{2} \mid \frac{108x-2916}{x^2}\right)_p \equiv \sum_{n=0}^{(p-1)/2} \frac{\left(\frac{1}{3}\right)_n \left(\frac{2}{3}\right)_n \left(\frac{1}{2}\right)_n \cdot (108x-2916)^n}{(n!)^3 \cdot x^{2n}} \pmod{p}.$$

For  $n > \lfloor p/3 \rfloor$ , any  $p$  will appear in the numerator of the expansion for  $\left(\frac{1}{3}\right)_n$ ,  $\left(\frac{2}{3}\right)_n$ , or  $\left(\frac{1}{2}\right)_n$ , so all of these terms will be congruent to 0 modulo  $p$  and will vanish. Thus we can simplify to

$${}_3F_2^{\text{tr}}\left(\frac{1}{3} \quad \frac{2}{3} \quad \frac{1}{2} \mid \frac{108x-2916}{x^2}\right)_p \equiv \sum_{n=0}^{\lfloor p/3 \rfloor} \frac{\left(\frac{1}{3}\right)_n \left(\frac{2}{3}\right)_n \left(\frac{1}{2}\right)_n \cdot (108x-2916)^n}{(n!)^3 \cdot x^{2n}} \pmod{p}. \tag{2-4}$$

Similarly by (1-4) we have that

$${}_3F_2^{\text{tr}}\left(\frac{1}{6} \quad \frac{5}{6} \quad \frac{1}{2} \mid 1 - \frac{1}{x}\right)_p \equiv \sum_{n=0}^{(p-1)/2} \frac{\left(\frac{1}{6}\right)_n \left(\frac{5}{6}\right)_n \left(\frac{1}{2}\right)_n \cdot \left(1 - \frac{1}{x}\right)^n}{(n!)^3} \pmod{p}.$$

For any  $n > \lfloor p/6 \rfloor$ ,  $p \equiv 1, 5 \pmod{6}$  will appear in the numerator of the expansion of  $\left(\frac{1}{6}\right)_n \left(\frac{5}{6}\right)_n \left(\frac{1}{2}\right)_n$  causing all of these sequential terms to be congruent to 0 modulo  $p$  and vanish, which gives

$${}_3F_2^{\text{tr}}\left(\frac{1}{6} \quad \frac{5}{6} \quad \frac{1}{2} \mid 1 - \frac{1}{x}\right)_p \equiv \sum_{n=0}^{\lfloor p/6 \rfloor} \frac{\left(\frac{1}{6}\right)_n \left(\frac{5}{6}\right)_n \left(\frac{1}{2}\right)_n \cdot \left(1 - \frac{1}{x}\right)^n}{(n!)^3} \pmod{p}. \tag{2-5}$$

**Work of Monks.** The proof of Theorem 1.1 relies on recent work of El-Guindy and Ono and Monks.

**Theorem 2.1** [Monks 2012, pp. 2–3]. *The following are true:*

(1) *If  $p \geq 5$  is prime,*

$$S_{p,1/4}(x) \equiv {}_2F_1^{\text{tr}}\left(\frac{1}{4} \quad \frac{3}{4} \mid -x\right)_p \pmod{p}. \tag{2-6}$$

(2) *If  $p \geq 5$  is prime,*

$$S_{p,1/3}(x) \equiv x^{\lfloor p/3 \rfloor} \cdot {}_2F_1^{\text{tr}}\left(\frac{1}{3} \quad \frac{2}{3} \mid \frac{27}{x}\right)_p \pmod{p}. \tag{2-7}$$

(3) For  $p \equiv 1, 5 \pmod{12}$  and prime,

$$S_{p,1/12}(x) \equiv c_p^{-1} \cdot x^{\lfloor p/12 \rfloor} \cdot {}_2F_1^{\text{tr}} \left( \begin{matrix} \frac{1}{12} & \frac{5}{12} \\ 1 \end{matrix} \middle| 1 - \frac{1}{x} \right)_p \pmod{p}. \quad (2-8)$$

(4) For  $p \equiv 7, 11 \pmod{12}$  and prime,

$$S_{p,1/12}(x) \equiv c_p^{-1} \cdot x^{\lfloor p/12 \rfloor} \cdot {}_2F_1^{\text{tr}} \left( \begin{matrix} \frac{7}{12} & \frac{11}{12} \\ 1 \end{matrix} \middle| 1 - \frac{1}{x} \right)_p \pmod{p}, \quad (2-9)$$

where

$$c_p = \binom{6 \lfloor \frac{p}{12} \rfloor + d_p}{\lfloor \frac{p}{12} \rfloor}$$

and  $d_p = 0, 2, 2, 4$  for  $p \equiv 1, 5, 7, 11 \pmod{12}$  respectively.

**Remark.** We note that (2-6) is a direct result of El-Guindy and Ono [2013] and is therefore not technically part of Monks’ theorem in [2012].

Squaring these supersingular loci in terms of the  ${}_2F_1^{\text{tr}}$ -hypergeometric functions, we obtain congruent  ${}_3F_2^{\text{tr}}$ -hypergeometric representations in Theorem 1.1.

### 3. Proof of Theorem 1.1

To prove Theorem 1.1, we show the first part using the results of El-Guindy and Ono. Then we calculate the equivalent statements for the remaining cases. We use classical  ${}_2F_1^{\text{cl}}$  transformation laws to obtain the necessary forms to use Clausen’s theorem, given in (2-1), and lift the  ${}_2F_1^{\text{tr}}$ -hypergeometric functions of Monks to equivalent  ${}_3F_2^{\text{tr}}$  representations. First we require the following descriptions of  ${}_2F_1^{\text{tr}}$ -hypergeometric functions:

**Lemma 3.1.** *The following are true:*

(1) If  $p \geq 5$  is an odd prime, then

$${}_2F_1^{\text{tr}} \left( \begin{matrix} \frac{1}{4} & \frac{3}{4} \\ 1 \end{matrix} \middle| -x \right)_p^2 \equiv (x+1)^{(p-1)/2} \cdot {}_3F_2^{\text{tr}} \left( \begin{matrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 1 & 1 \end{matrix} \middle| \frac{x}{x+1} \right)_p \pmod{p}.$$

(2) If  $p \geq 5$  is an odd prime, then

$${}_2F_1^{\text{tr}} \left( \begin{matrix} \frac{1}{3} & \frac{2}{3} \\ 1 \end{matrix} \middle| \frac{27}{x} \right)_p^2 \equiv {}_3F_2^{\text{tr}} \left( \begin{matrix} \frac{1}{3} & \frac{2}{3} & \frac{1}{2} \\ 1 & 1 \end{matrix} \middle| \frac{108x - 2916}{x^2} \right)_p \pmod{p}.$$

(3) For  $p \equiv 1, 5 \pmod{12}$ ,

$${}_2F_1^{\text{tr}} \left( \begin{matrix} \frac{1}{12} & \frac{5}{12} \\ 1 \end{matrix} \middle| 1 - \frac{1}{x} \right)_p^2 \equiv {}_3F_2^{\text{tr}} \left( \begin{matrix} \frac{1}{6} & \frac{5}{6} & \frac{1}{2} \\ 1 & 1 \end{matrix} \middle| 1 - \frac{1}{x} \right)_p \pmod{p}.$$

(4) For  $p \equiv 7, 11 \pmod{12}$ ,

$${}_2F_1^{\text{tr}}\left(\begin{matrix} \frac{7}{12} & \frac{11}{12} \\ 1 & 1 \end{matrix} \middle| 1 - \frac{1}{x}\right)_p \equiv x \cdot {}_3F_2^{\text{tr}}\left(\begin{matrix} \frac{1}{6} & \frac{5}{6} & \frac{1}{2} \\ 1 & 1 \end{matrix} \middle| 1 - \frac{1}{x}\right)_p \pmod{p}.$$

*Proof.* For brevity, we give the proof of (2). The remaining cases follow in a similar way. Applying the transformation law for  ${}_2F_1$ -hypergeometric functions given by (2-3) with  $a = \frac{1}{3}$ ,  $b = \frac{2}{3}$ , and  $x = 27/x$ , we see that

$${}_2F_1^{\text{cl}}\left(\begin{matrix} \frac{1}{3} & \frac{2}{3} \\ 1 \end{matrix} \middle| \frac{27}{x}\right) = {}_2F_1^{\text{cl}}\left(\begin{matrix} \frac{1}{6} & \frac{1}{3} \\ 1 \end{matrix} \middle| \frac{108x - 2916}{x^2}\right).$$

We then square both sides of this equation and apply Clausen’s theorem in (2-1) to the right-hand expression with  $\alpha = \frac{1}{6}$ ,  $\beta = \frac{1}{3}$ , and  $x = (108x - 2916)/x^2$  to obtain

$$\begin{aligned} {}_2F_1^{\text{cl}}\left(\begin{matrix} \frac{1}{3} & \frac{2}{3} \\ 1 \end{matrix} \middle| \frac{27}{x}\right)^2 &= {}_2F_1^{\text{cl}}\left(\begin{matrix} \frac{1}{6} & \frac{1}{3} \\ 1 \end{matrix} \middle| \frac{108x - 2916}{x^2}\right)^2 \\ &= {}_3F_2^{\text{cl}}\left(\begin{matrix} \frac{1}{3} & \frac{2}{3} & \frac{1}{2} \\ 1 & 1 \end{matrix} \middle| \frac{108x - 2916}{x^2}\right). \end{aligned} \tag{3-1}$$

By definition (1-1), when we expand the infinite hypergeometric series on the left-hand side of this equation, we obtain

$${}_2F_1^{\text{cl}}\left(\begin{matrix} \frac{1}{3} & \frac{2}{3} \\ 1 \end{matrix} \middle| \frac{27}{x}\right)^2 = \left(\sum_{N=0}^{\infty} \frac{\left(\frac{1}{3}\right)_N \left(\frac{2}{3}\right)_N}{(N!)^2} \cdot \left(\frac{27}{x}\right)^N\right)^2,$$

and when we expand the right hand side by definition (1-2) we get

$${}_3F_2^{\text{cl}}\left(\begin{matrix} \frac{1}{3} & \frac{2}{3} & \frac{1}{2} \\ 1 & 1 \end{matrix} \middle| \frac{108x - 2916}{x^2}\right) = \sum_{N=0}^{\infty} \frac{\left(\frac{1}{3}\right)_N \left(\frac{2}{3}\right)_N \left(\frac{1}{2}\right)_N}{(N!)^3} \cdot \left(\frac{108x - 2916}{x^2}\right)^N.$$

By (3-1), we have that these two infinite series expansions are equal and

$$\left(\sum_{N=0}^{\infty} \frac{\left(\frac{1}{3}\right)_N \left(\frac{2}{3}\right)_N}{(N!)^2} \cdot \left(\frac{27}{x}\right)^N\right)^2 = \sum_{N=0}^{\infty} \frac{\left(\frac{1}{3}\right)_N \left(\frac{2}{3}\right)_N \left(\frac{1}{2}\right)_N}{(N!)^3} \cdot \left(\frac{108x - 2916}{x^2}\right)^N. \tag{3-2}$$

This means that in both series expansions, the coefficients for  $x^{-N}$ , given by  $a(N)$  and  $b(N)$  respectively, are equal. More precisely, by squaring we have

$$a(N) = \sum_{n=0}^N \frac{\left(\frac{1}{3}\right)_n \left(\frac{2}{3}\right)_n}{(n!)^2} \cdot \frac{\left(\frac{1}{3}\right)_{N-n} \left(\frac{2}{3}\right)_{N-n}}{((N-n)!)^2} \cdot 27^N,$$



and by the binomial theorem,

$$b(N) = \sum_{n=\lceil N/2 \rceil}^N \frac{\left(\frac{1}{3}\right)_n \left(\frac{2}{3}\right)_n \left(\frac{1}{2}\right)_n}{(n!)^3} \cdot \binom{n}{2n-N} (108)^{2n-N} (-2916)^{N-n}.$$

We note that for  $b(N)$ , only  $n$  with  $\lceil N/2 \rceil \leq n \leq N$  will actually contribute to each coefficient value. When we truncate these series in (3-2) at  $N = p - 1$  (i.e., truncate at  $x^{1-p}$ ), all of the coefficients will still be equal. The truncation of the series can be explicitly expressed by

$$\begin{aligned} & \sum_{N=0}^{p-1} \sum_{n=0}^N \frac{\left(\frac{1}{3}\right)_n \left(\frac{2}{3}\right)_n}{(n!)^2} \cdot \frac{\left(\frac{1}{3}\right)_{N-n} \left(\frac{2}{3}\right)_{N-n}}{((N-n)!)^2} \cdot 27^N \cdot x^{-N} \\ &= \sum_{N=0}^{p-1} \sum_{n=\lceil N/2 \rceil}^N \frac{\left(\frac{1}{3}\right)_n \left(\frac{2}{3}\right)_n \left(\frac{1}{2}\right)_n}{(n!)^3} \cdot \binom{n}{2n-N} (108)^{2n-N} (-2916)^{N-n} \cdot x^{-N}. \end{aligned} \quad (3-3)$$

We observe that since  $N$ , and consequently  $n$ , will never exceed  $p - 1$ , all of these coefficients are  $p$ -integral since  $p$  does not appear in any of the denominators. Therefore we can take both sides of (3-3) modulo  $p$ . In fact, we know that a lot of terms will vanish modulo  $p$  because  $p$  will appear as a factor in the numerators of the coefficient expansions of these series given by  $a(N)$  and  $b(N)$ , making them congruent to 0. More specifically, this is the case for  $N$  with  $(p - 1)/2 < N \leq p - 1$  and  $n \geq (p - 1)/2$ . We can write these simplified congruences as

$$\begin{aligned} & \sum_{N=0}^{p-1} \sum_{n=0}^N \frac{\left(\frac{1}{3}\right)_n \left(\frac{2}{3}\right)_n}{(n!)^2} \cdot \frac{\left(\frac{1}{3}\right)_{N-n} \left(\frac{2}{3}\right)_{N-n}}{((N-n)!)^2} \cdot \left(\frac{27}{x}\right)^N \\ & \equiv \left( \sum_{N=0}^{(p-1)/2} \frac{\left(\frac{1}{3}\right)_N \left(\frac{2}{3}\right)_N}{(N!)^2} \cdot \left(\frac{27}{x}\right)^N \right)^2 \pmod{p} \end{aligned} \quad (3-4)$$

and

$$\begin{aligned} & \sum_{N=0}^{p-1} \sum_{n=\lceil N/2 \rceil}^N \frac{\left(\frac{1}{3}\right)_n \left(\frac{2}{3}\right)_n \left(\frac{1}{2}\right)_n}{(n!)^3} \cdot \binom{n}{2n-N} (108)^{2n-N} (-2916)^{N-n} \cdot x^{-N} \\ & \equiv \sum_{N=0}^{(p-1)/2} \frac{\left(\frac{1}{3}\right)_N \left(\frac{2}{3}\right)_N \left(\frac{1}{2}\right)_N}{(N!)^3} \cdot \left(\frac{108x - 2916}{x^2}\right)^N \pmod{p}. \end{aligned} \quad (3-5)$$

Finally, we see that the right-hand sides of (3-4) and (3-5) are congruent modulo  $p$  to the definitions of the truncated forms of the squares of the  ${}_2F_1$ - and

${}_3F_2$ -hypergeometric functions, respectively, given by:

$${}_2F_1^{\text{tr}}\left(\begin{matrix} \frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & 1 \end{matrix} \middle| \frac{27}{x} \right)_p \equiv \left( \sum_{N=0}^{(p-1)/2} \frac{\left(\frac{1}{3}\right)_N \left(\frac{2}{3}\right)_N}{(N!)^2} \cdot \left(\frac{27}{x}\right)^N \right)^2 \pmod{p}$$

and

$$\begin{aligned} {}_3F_2^{\text{tr}}\left(\begin{matrix} \frac{1}{3} & \frac{2}{3} & \frac{1}{2} \\ 1 & 1 & 1 \end{matrix} \middle| \frac{108x - 2916}{x^2} \right)_p \\ \equiv \sum_{N=0}^{(p-1)/2} \frac{\left(\frac{1}{3}\right)_N \left(\frac{2}{3}\right)_N \left(\frac{1}{2}\right)_N}{(N!)^3} \cdot \left(\frac{108x - 2916}{x^2}\right)^N \pmod{p}. \end{aligned}$$

It follows that

$${}_2F_1^{\text{tr}}\left(\begin{matrix} \frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & 1 \end{matrix} \middle| \frac{27}{x} \right)_p \equiv {}_3F_2^{\text{tr}}\left(\begin{matrix} \frac{1}{3} & \frac{2}{3} & \frac{1}{2} \\ 1 & 1 & 1 \end{matrix} \middle| \frac{108x - 2916}{x^2} \right)_p \pmod{p},$$

which completes the proof. □

**Proof of Theorem 1.1.** For the proof of (1), we begin with Lemma 3.1(1) which gives

$$(x + 1)^{(p-1)/2} \cdot {}_3F_2^{\text{tr}}\left(\begin{matrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 1 & 1 & 1 \end{matrix} \middle| \frac{x}{x+1} \right)_p \equiv {}_2F_1^{\text{tr}}\left(\begin{matrix} \frac{1}{4} & \frac{3}{4} \\ 1 & 1 \end{matrix} \middle| -x \right)_p^2 \pmod{p}.$$

Substituting the left-hand side of the above congruence into the square of (2-6), we obtain the congruence for the square of the supersingular locus  $S_{p,(1/4)}(x)^2$  for the family of elliptic curves given by  $E_{1/4}(\lambda)$ .

The remaining cases use the congruences of the supersingular loci given by Monks. We begin by squaring the  ${}_2F_1^{\text{tr}}$ -hypergeometric functions in (2-7)–(2-9). Squaring (2-7), we obtain

$$S_{p,1/3}(x)^2 \equiv x^{2 \cdot \lfloor p/3 \rfloor} \cdot {}_2F_1^{\text{tr}}\left(\begin{matrix} \frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & 1 \end{matrix} \middle| \frac{27}{x} \right)_p^2 \pmod{p}.$$

Then using the congruence in Lemma 3.1(2), we have

$$S_{p,1/3}(x)^2 \equiv x^{2 \cdot \lfloor p/3 \rfloor} \cdot {}_3F_2^{\text{tr}}\left(\begin{matrix} \frac{1}{3} & \frac{2}{3} & \frac{1}{2} \\ 1 & 1 & 1 \end{matrix} \middle| \frac{108x - 2916}{x^2} \right)_p \pmod{p},$$

completing the proof of (2).

In the third case, after squaring (2-8), we obtain

$$S_{p,1/12}(x)^2 \equiv (c_p^{-1})^2 \cdot x^{2 \cdot \lfloor p/12 \rfloor} \cdot {}_2F_1^{\text{tr}}\left(\begin{matrix} \frac{1}{12} & \frac{5}{12} \\ 1 & 1 \end{matrix} \middle| 1 - \frac{1}{x} \right)_p^2 \pmod{p}.$$

Then we use our congruence given in Lemma 3.1(3) and substitute the  ${}_3F_2$ -hypergeometric function to give

$$S_{p,1/12}(x)^2 \equiv (c_p^{-1})^2 \cdot x^{\lfloor p/6 \rfloor} \cdot {}_3F_2^{\text{tr}} \left( \begin{matrix} \frac{1}{6} & \frac{5}{6} & \frac{1}{2} \\ 1 & 1 \end{matrix} \middle| 1 - \frac{1}{x} \right)_p \pmod{p}.$$

We see in (3) and (4) of Lemma 3.1, for  $p \equiv 1, 5 \pmod{6}$ , the squared  ${}_2F_1^{\text{tr}}$ -hypergeometric functions are congruent apart from the  $x$  in (4). We combine these cases and alter the exponent of  $x$  to satisfy both, which then gives our result.

### 4. Examples

**Example.** Here we consider  $E_{1/12}(x)$  when  $p = 13$ . By Monks' theorem, we know that there is just one supersingular elliptic curve for  $E_{1/12}(x)$ . It turns out that  $E_{1/12}(3)$  is that supersingular elliptic curve. To see this, we note that  $E_{1/12}(3)$  over  $\mathbb{F}_{13}$  has 13 points including the point at infinity. By Monks, this implies that

$$S_{13,1/12}(x) \equiv (x - 3) \equiv (x + 10) \pmod{13}.$$

We square this to obtain

$$S_{13,1/12}(x)^2 \equiv (x + 10)^2 \equiv (x^2 + 20x + 100) \equiv x^2 + 7x + 9 \pmod{13}.$$

Using Theorem 1.1(3), we calculate

$$(c_{13}^{-1})^2 \cdot x^{\lfloor 13/6 \rfloor} \cdot {}_3F_2^{\text{tr}} \left( \begin{matrix} \frac{1}{6} & \frac{5}{6} & \frac{1}{2} \\ 1 & 1 \end{matrix} \middle| 1 - \frac{1}{x} \right)_{13} \pmod{13},$$

which gives  $(c_{13}^{-1})^2 \equiv \frac{1}{10} \pmod{13}$  and  $x^{\lfloor 13/6 \rfloor} = x^2$ . Substituting these values into our expression gives

$$\frac{1}{10} \cdot x^2 \cdot \left( 10 + \frac{5}{x} + \frac{12}{x^2} \right) \equiv x^2 + \frac{1}{2}x + \frac{6}{5} \equiv x^2 + 7x + 9 \pmod{13}.$$

This polynomial can be factored modulo 13 as

$$x^2 + 7x + 9 \equiv (x + 10)^2 \pmod{13},$$

which is what we found after directly squaring  $S_{13,1/12}(x)$ .

**Example.** We consider  $E_{1/12}(x)$  when  $p = 59$ . By Monks' theorem, we know that there are four supersingular elliptic curves for  $E_{1/12}(x)$ . Those supersingular elliptic curves are found to be  $E_{1/12}(32)$ ,  $E_{1/12}(35)$ ,  $E_{1/12}(24)$  and  $E_{1/12}(22)$ . To see this, we note that  $E_{1/12}(x)$  for  $x = 32, 35, 24$  and  $22$  over  $\mathbb{F}_{59}$  have 59 points

including the point at infinity. By Monks, this implies that

$$\begin{aligned} S_{59,1/12}(x) &\equiv (x - 32)(x - 35)(x - 24)(x - 22) \\ &\equiv (x + 27)(x + 24)(x + 35)(x + 37) \pmod{59}. \end{aligned}$$

After squaring this directly, we obtain

$$S_{59,1/12}(x)^2 \equiv (x + 27)^2(x + 24)^2(x + 35)^2(x + 37)^2 \pmod{59}. \quad (4-1)$$

Next using [Theorem 1.1\(3\)](#) we calculate

$$(c_{59}^{-1})^2 \cdot x^{\lfloor 59/6 \rfloor} \cdot {}_3F_2^{\text{tr}} \left( \begin{matrix} \frac{1}{6} & \frac{5}{6} & \frac{1}{2} \\ 1 & 1 \end{matrix} \middle| 1 - \frac{1}{x} \right)_{59} \pmod{59}.$$

For  $p = 59$ , we have  $(c_{59}^{-1})^2 = 15$  and  $x^{\lfloor 59/6 \rfloor} = x^9$ , so we obtain

$$\begin{aligned} 15 \cdot x^9 \cdot \left( \frac{4}{x} + \frac{40}{x^2} + \frac{3}{x^3} + \frac{16}{x^4} + \frac{38}{x^5} + \frac{56}{x^6} + \frac{16}{x^7} + \frac{28}{x^8} + \frac{36}{x^9} \right) \\ \equiv x^8 + 10x^7 + 45x^6 + 4x^5 + 39x^4 + 14x^3 + 4x^2 + 7x + 9 \pmod{59}. \end{aligned}$$

This polynomial of degree 8 can be factored as

$$(x + 27)^2(x + 24)^2(x + 35)^2(x + 37)^2 \pmod{59},$$

which is congruent modulo 59 to  $S_{59,1/12}(x)^2$  as given in (4-1).

## References

- [Bailey 1935] W. Bailey, *Generalized hypergeometric series*, Cambridge Univ. Press, 1935. Reprinted 1964, etc.
- [El-Guindy and Ono 2013] A. El-Guindy and K. Ono, “Hasse invariants for the Clausen elliptic curves”, *Ramanujan J.* **31**:1-2 (2013), 3–13. [MR 3048650](#)
- [Kaneko and Zagier 1998] M. Kaneko and D. Zagier, “Supersingular  $j$ -invariants, hypergeometric series, and Atkin’s orthogonal polynomials”, pp. 97–126 in *Computational perspectives on number theory* (Chicago, IL, 1995), edited by 99b:11064, AMS/IP Stud. Adv. Math. **7**, Amer. Math. Soc., Providence, RI, 1998. [MR 99b:11064](#)
- [Monks 2012] K. Monks, “On supersingular elliptic curves and hypergeometric functions”, *Involve* **5**:1 (2012), 99–113. [MR 2924318](#)
- [Vidūnas 2009] R. Vidūnas, “Algebraic transformations of Gauss hypergeometric functions”, *Funkcial. Ekvac.* **52**:2 (2009), 139–180. [MR 2010i:33012](#)

Received: 2013-07-17

Revised: 2013-09-02

Accepted: 2013-09-04

[spitman222@gmail.com](mailto:spitman222@gmail.com)

Emory University,  
Department of Mathematics and Computer Science,  
400 Dowman Drive, Atlanta, Georgia 30322, United States

# A contribution to the connections between Fibonacci numbers and matrix theory

Miriam Farber and Abraham Berman

(Communicated by Robert J. Plemmons)

We present a lovely connection between the Fibonacci numbers and the sums of inverses of  $(0, 1)$ -triangular matrices, namely, a number  $S$  is the sum of the entries of the inverse of an  $n \times n$   $(0, 1)$ -triangular matrix (for  $n \geq 3$ ) if and only if  $S$  is an integer between  $2 - F_{n-1}$  and  $2 + F_{n-1}$ . Corollaries include Fibonacci identities and a Fibonacci-type result on determinants of a special family of  $(1, 2)$ -matrices.

## 1. Introduction

One of the ways to motivate students' interest in linear algebra is to present interesting connections between matrices and the Fibonacci numbers

$$F_1 = F_2 = 1, \quad F_n = F_{n-1} + F_{n-2}, \quad n \geq 3.$$

For example, one can prove that  $F_n^2 - F_{n-1}F_{n+1} = (-1)^{n+1}$  by using induction and the fact that

$$\begin{aligned} \det \begin{pmatrix} F_n & F_{n-1} \\ F_{n+1} & F_n \end{pmatrix} &= \det \begin{pmatrix} F_n & F_{n-1} \\ F_{n+1} - F_n & F_n - F_{n-1} \end{pmatrix} \\ &= \det \begin{pmatrix} F_n & F_{n-1} \\ F_{n-1} & F_{n-2} \end{pmatrix} = -\det \begin{pmatrix} F_{n-1} & F_n \\ F_{n-2} & F_{n-1} \end{pmatrix}. \end{aligned}$$

Similarly, one can determine the exact value of the  $n$ -th Fibonacci number, by calculating the eigenvalues and the eigenvectors of  $\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$  and using the equation

$$\begin{pmatrix} F_n \\ F_{n-1} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} F_{n-1} \\ F_{n-2} \end{pmatrix} = \cdots = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}^{n-2} \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

*MSC2010:* 15A15, 11B39, 15A09, 15B99.

*Keywords:* Fibonacci numbers, Hessenberg matrix, sum of entries.

As another example of connections between Fibonacci numbers and matrix theory, consider lower triangular matrices of the form

$$\begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ -1 & 1 & 0 & \cdots & \cdots & 0 \\ -1 & -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & -1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -1 & -1 & 1 \end{pmatrix}.$$

The inverses of these matrices are of the form

$$\begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 1 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 2 & 1 & 1 & 0 & \cdots & \cdots & 0 \\ 3 & 2 & 1 & 1 & 0 & \ddots & \vdots \\ 5 & 3 & 2 & 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \cdots & 5 & 3 & 2 & 1 & 1 \end{pmatrix},$$

which, due to their remarkable structure, are known as Fibonacci matrices. Various properties of these matrices and their generalizations have been studied [Lee et al. 2002; Lee and Kim 2003; Wang and Wang 2008].

Fibonacci numbers are also widely used in algorithms in computer science [Atkins and Geist 1987; Knuth 1997], such as algorithms for finding extrema, merging files, searching in trees, etc. We provide here an example of their use in the searching of ordered arrays, described in [Atkins and Geist 1987]. Suppose that we have a sorted array with  $F_n - 1$  elements for some natural number  $n$  (we can always pad the array with dummy elements in order to achieve such number of elements); for example, let  $A = (0, 1, 2, 3, 5, 6, 9, 11, 15, 18, 20, 23)$  be an array with  $F_7 - 1 = 12$  elements. We would like to check whether 15 is in  $A$ . First compare 15 with the  $F_{7-1}$ -th entry. Since  $11 < 15$ , we can eliminate all the entries to the left of the  $F_{7-1}$ -th entry (including the  $F_{7-1}$ -th entry), and we are left with the array  $B = (15, 18, 20, 23)$  which contains  $F_5 - 1 = 4$  elements. We now compare the  $F_{5-1}$ -th entry in  $B$  with 15, and since  $20 > 15$ , we eliminate 20 and 23, and we are left with the array  $C = (15, 18)$  that has  $F_4 - 1$  entries. Finally, we compare the  $F_{4-1}$ -th entry of  $C$  to 15, and since  $18 > 15$ , we are left with 15 and have a match. The full algorithm is described in [Atkins and Geist 1987]. Another interesting connection between Fibonacci numbers and matrices is given in [Li 1993], where it is shown that the maximal determinant of an  $n \times n$   $(0, 1)$ -Hessenberg matrix is  $F_n$ .

Let  $S(X)$  denote the sum of the entries of a matrix  $X$ . Huang, Tam and Wu [Huang et al. 2013] show, among other results, that a number  $S$  is equal to  $S(A^{-1})$  for an adjacency matrix (a symmetric  $(0, 1)$ -matrix with trace zero)  $A$  if and only if  $S$  is rational. More generally, they ask what can be said about the sum of the entries of the inverse of a  $(0, 1)$ -matrix. We consider the class of triangular matrices and show that a number  $S$  is equal to  $S(A^{-1})$  for a triangular  $(0, 1)$ -matrix  $A$  if and only if  $S$  is an integer. This follows from our main result which shows that for  $n \geq 3$ , a number  $S$  is equal to  $S(A^{-1})$  for an  $n \times n$  triangular  $(0, 1)$ -matrix  $A$  if and only if

$$2 - F_{n-1} \leq S \leq 2 + F_{n-1}.$$

We use the following definitions and notation. Let  $e$  denote a vector of ones (so  $S(A) = e^T A e$ ) and  $A_n$  the set of  $n \times n$  invertible  $(0, 1)$ -upper triangular matrices. We will say that a matrix  $A \in A_n$ , where  $n \geq 3$ , is *maximizing* if  $S(A^{-1}) = 2 + F_{n-1}$  and *minimizing* if  $S(A^{-1}) = 2 - F_{n-1}$ , and refer to maximizing and minimizing matrices as *extremal matrices*. For a set of vectors  $V \subset R^n$ , a vector  $v \in V$  is *absolutely dominant* if for every  $u \in V$ ,  $|v_i| \geq |u_i|$ , where  $i = 1, 2, \dots, n$ .

We will use the following well-known properties of Fibonacci numbers (see, for example, [Vorobiev 2002]):

- Lemma 1.1.** (i)  $1 + \sum_{k=1}^n F_k = F_{n+2}$ ;  
 (ii)  $1 + \sum_{k=1}^n F_{2k} = F_{2n+1}$ ;  
 (iii)  $\sum_{k=1}^n F_{2k-1} = F_{2n}$ .

The main result of the paper is proved in Section 2. In Section 3, we describe a construction of extremal matrices with a beautiful Fibonacci pattern in their inverses, and use it to obtain several Fibonacci identities. We conclude with a Fibonacci-type result on determinants of  $(1, 2)$ -matrices in spirit of the result in [Li 1993].

## 2. The main result

**Theorem 2.1.** *Let  $n \geq 3$ . Then  $S = S(A^{-1})$  for some  $A \in A_n$  if and only if  $S$  is an integer between  $2 - F_{n-1}$  and  $2 + F_{n-1}$ ; that is,  $2 - F_{n-1} \leq S \leq 2 + F_{n-1}$ .*

*Proof.* Obviously,  $S(A^{-1})$  must be an integer since  $A^{-1} = \text{adj}(A)/\det(A)$  and  $\det(A) = 1$ . The main part of the proof consists of showing

- (a)  $\max_{A \in A_n} S(A^{-1}) = 2 + F_{n-1}$ ,
- (b)  $\min_{A \in A_n} S(A^{-1}) = 2 - F_{n-1}$ , and
- (c) for every integer  $S$  between  $2 - F_{n-1}$  and  $2 + F_{n-1}$ , there exists  $A \in A_n$  such that  $S(A^{-1}) = S$ .

To show (a) and (b) we prove the following lemma.

**Lemma 2.2.** *Let  $V = \{e^T A^{-1} \mid A \in A_n\}$ . For the purposes of this lemma only, we will let  $F_0 = -1$  (note that this is not a Fibonacci number). Then  $v = (v_i)$ , where  $v_i = (-1)^i F_{i-1}$ , is an absolutely dominant vector of  $V$ .*

*Proof.* For  $n = 1$ , we have  $V = \{(1)\}$ ; for  $n = 2$ , we have  $V = \{(1 \ 1), (1 \ 0)\}$ ; and for  $n = 3$ , we have  $V = \{(1 \ 1 \ 1), (1 \ 0 \ 1), (1 \ 1 \ 0), (1 \ 0 \ 0), (1 \ 1 \ -1)\}$ . Therefore the statement holds for  $n = 1, 2, 3$ . To prove the lemma for  $n \geq 4$ , we will use induction. Suppose the lemma is true for  $k < n$ .

We will now show that the vector  $v$ , defined in the lemma, is an absolutely dominant vector of the set  $V = \{e^T A^{-1} \mid A \in A_n\}$ . Let  $A \in A_n$ . Then  $A$  is of the form

$$\begin{pmatrix} C & \alpha & \beta \\ 0 & 1 & x \\ 0 & 0 & 1 \end{pmatrix},$$

where  $C \in A_{n-2}$ ,  $\alpha, \beta \in \{0, 1\}^{n-2}$ , and  $x \in \{0, 1\}$ . Therefore,

$$A^{-1} = \begin{pmatrix} C^{-1} & -C^{-1}(\alpha \ \beta) \begin{pmatrix} 1 & -x \\ 0 & 1 \end{pmatrix} \\ 0 & \begin{pmatrix} 1 & -x \\ 0 & 1 \end{pmatrix} \end{pmatrix} = \begin{pmatrix} C^{-1} & -C^{-1}(\alpha \ \beta - x\alpha) \\ 0 & \begin{pmatrix} 1 & -x \\ 0 & 1 \end{pmatrix} \end{pmatrix}.$$

We will use the following notation:

$$\begin{aligned} e^T C^{-1} &= (c_1 \ c_2 \ \dots \ c_{n-2}), \quad \alpha = (\alpha_1 \ \alpha_2 \ \dots \ \alpha_{n-2})^T, \\ \beta &= (\beta_1 \ \beta_2 \ \dots \ \beta_{n-2})^T. \end{aligned}$$

So

$$e^T A^{-1} = \left( c_1 \ c_2 \ \dots \ c_{n-2} \ 1 - \sum_{i=1}^{n-2} \alpha_i c_i \ 1 - x - \sum_{i=1}^{n-2} c_i(\beta_i - x\alpha_i) \right).$$

Consider the  $n$ -th entry of  $e^T A^{-1}$ . Since  $c_1 = 1$ ,  $n \geq 4$ , and  $-1 \leq \beta_i - x\alpha_i \leq 1$  for all  $1 \leq i \leq n-2$ , it is easy to see that

$$-\sum_{i=1}^{n-2} |c_i| \leq 1 - x - \sum_{i=1}^{n-2} c_i(\beta_i - x\alpha_i) \leq \sum_{i=1}^{n-2} |c_i|$$

for all possible  $x, \alpha_i, \beta_i \in \{0, 1\}$ , where  $1 \leq i \leq n-2$ . Since  $\beta_i - \alpha_i \in \{-1, 0, 1\}$ , it is possible to achieve equality in each inequality by taking

$$x = 1 \quad \text{and} \quad \text{sign}(\beta_i - \alpha_i) = \text{sign}(c_i), \quad 1 \leq i \leq n-2 \tag{1}$$

in the first, and

$$x = 1 \quad \text{and} \quad \text{sign}(\beta_i - \alpha_i) = -\text{sign}(c_i), \quad 1 \leq i \leq n-2. \tag{2}$$



in the second. Now, since  $|\sum_{i=1}^{n-2} |c_i|| = |\sum_{i=1}^{n-2} c_i|$ , we get that if  $A \in A_n$  is a matrix for which  $e^T A^{-1}$  is an absolutely dominant vector, its  $n$ -th entry must be equal to either

$$-\sum_{i=1}^{n-2} |c_i| \tag{3}$$

or

$$\sum_{i=1}^{n-2} |c_i|. \tag{4}$$

Note that the maximal value of (3) is obtained by taking  $C$  such that  $e^T C^{-1}$  is an absolutely dominant vector of the set  $V = \{e^T A^{-1} \mid A \in A_{n-2}\}$  (and all the absolutely dominant vectors will give the same value). The same is true of the minimal value of (4). By the inductive hypothesis and using Lemma 1.1, the maximal value of (4) is

$$\sum_{i=1}^{n-2} |c_i| = 1 + \sum_{i=1}^{n-3} F_i = F_{n-1}$$

(and this value may be achieved by choosing an appropriate  $C$ ). Similarly, the minimal value of (3) is  $-F_{n-1}$ . Let us now consider the  $(n-1)$ -th entry of  $e^T A^{-1}$ . By the inductive hypothesis, its absolute value is bounded from above by  $F_{n-2}$ . By taking  $C \in A_{n-2}$  such that  $e^T C^{-1}$  is an absolutely dominant vector, choosing  $\alpha, \beta$  such that either (1) or (2) is satisfied and using Lemma 1.1 and the inductive hypothesis, we get that the  $(n-1)$ -th entry of  $e^T A^{-1}$  is equal to either

$$1 - \sum_{i=1}^{n-2} \alpha_i c_i = 1 - \sum_{k=1}^{\lfloor \frac{n-3}{2} \rfloor} c_{2k+1} = 1 + \sum_{k=1}^{\lfloor \frac{n-3}{2} \rfloor} F_{2k} = F_{2\lfloor \frac{n-3}{2} \rfloor + 1}, \tag{5}$$

or

$$1 - \sum_{i=1}^{n-2} \alpha_i c_i = 1 - c_1 - \sum_{k=1}^{\lfloor \frac{n-2}{2} \rfloor} c_{2k} = - \sum_{k=1}^{\lfloor \frac{n-2}{2} \rfloor} F_{2k-1} = -F_{2\lfloor \frac{n-2}{2} \rfloor}. \tag{6}$$

Note that if  $n$  is odd then expression (5) is equal to  $F_{n-2}$ , and if  $n$  is even then expression (6) is equal to  $-F_{n-2}$ . In sum, using the inductive hypothesis, we showed that the largest possible absolute value of the  $n$ -th entry of  $e^T A^{-1}$  (such that  $A \in A_n$ ) is  $F_{n-1}$ . In this case, we showed that it is possible to choose  $\alpha$  such that the absolute value of the  $(n-1)$ -th entry of  $e^T A^{-1}$  is  $F_{n-2}$ , the largest possible absolute value due to the inductive hypothesis. Therefore, we showed that the vector  $v$ , defined in the lemma, is an absolutely dominant vector for  $V = \{e^T A^{-1} \mid A \in A_n\}$ .  $\square$

We are now ready to prove (a) and (b). We represent  $A \in A_n$  in the same form as in Lemma 2.2, and

$$\begin{aligned} e^T A^{-1} e &= e^T \begin{pmatrix} C^{-1} & -C^{-1}(\alpha \ \beta - x\alpha) \\ 0 & \begin{pmatrix} 1 & -x \\ 0 & 1 \end{pmatrix} \end{pmatrix} e \\ &= 2 - x + e^T C^{-1} e - e^T C^{-1}(\beta + (1 - x)\alpha) \\ &= 2 - x + e^T C^{-1}(e - \alpha - \beta + x\alpha). \end{aligned}$$

Let  $u = e - \alpha - \beta + x\alpha$ . Note that if  $x = 1$  then  $u \in \{0, 1\}^{n-2}$ , and if  $x = 0$  then  $u \in \{-1, 0, 1\}^{n-2}$ . In addition, note that

$$\begin{aligned} \max\{2 - x + e^T C^{-1} u \mid x = 0, \alpha, \beta \in \{0, 1\}^{n-2}\} \\ \geq \max\{2 - x + e^T C^{-1} u \mid x = 1, \alpha, \beta \in \{0, 1\}^{n-2}\}. \end{aligned} \tag{7}$$

Now, since  $C \in A_{n-2}$ , the first entry of  $e^T C^{-1}$  is 1. If  $x = 0$ , then in order to minimize the value of  $e^T C^{-1} u$ , we have to take the first entries of  $\alpha$  and  $\beta$  to be 1. On the other hand, if  $x = 1$ , then in order to minimize the value of  $e^T C^{-1} u$ , we have to take the first entries of  $\beta$  to be 1. The difference between these two cases is 1, and therefore

$$\begin{aligned} \min\{2 - x + e^T C^{-1} u \mid x = 0, \alpha, \beta \in \{0, 1\}^{n-2}\} \\ \leq \min\{2 - x + e^T C^{-1} u \mid x = 1, \alpha, \beta \in \{0, 1\}^{n-2}\}. \end{aligned} \tag{8}$$

Since we are only interested in the minimal and the maximal values of  $e^T A^{-1} e$ , we may assume, by (7) and (8), that  $x = 0$ . Therefore,  $e^T A^{-1} e = 2 + e^T C^{-1}(e - \alpha - \beta)$ . Using the notation of Lemma 2.2, we get

$$\min\{2 + e^T C^{-1}(e - \alpha - \beta) \mid \alpha, \beta \in \{0, 1\}^{n-2}\} = 2 - \sum_{i=1}^{n-2} |c_i| \tag{9}$$

and

$$\max\{2 + e^T C^{-1}(e - \alpha - \beta) \mid \alpha, \beta \in \{0, 1\}^{n-2}\} = 2 + \sum_{i=1}^{n-2} |c_i|. \tag{10}$$

Therefore, the minimal and the maximal values of  $e^T A^{-1} e$  are achieved by taking  $C$  such that  $e^T C^{-1}$  is an absolutely dominant vector of  $\{e^T A^{-1} \mid A \in A_{n-2}\}$ . Hence,

by Lemmas 2.2 and 1.1,

$$\begin{aligned} \max_{A \in A_n} S(A^{-1}) &= \max \{2 + e^T C^{-1}(e - \alpha - \beta) \mid \alpha, \beta \in \{0, 1\}^{n-2}, C \in A_{n-2}\} \\ &= 3 + \sum_{i=1}^{n-3} F_i = 2 + F_{n-1}, \end{aligned}$$

and similarly,

$$\min_{A \in A_n} S(A^{-1}) = 1 - \sum_{i=1}^{n-3} F_i = 2 - F_{n-1}.$$

It is well known that every natural number is the sum of distinct Fibonacci numbers. For the proof of (c), we need a slightly stronger observation.

**Lemma 2.3.** *Let  $M$  be a natural number, and let  $n$  be an integer for which  $F_{n-1} \leq M < F_n$ . Then  $M$  can be represented as a sum of distinct Fibonacci elements from the set  $\{F_1, F_2, \dots, F_{n-2}\}$ .*

*Proof.* For  $M = 1$ , the statement is true. Proceeding by induction, assume that it is true for all integers less than  $M$ . Let  $n$  be an integer for which  $F_{n-1} \leq M < F_n$ . Since  $M < F_n$ , we get that  $M < F_{n-2} + F_{n-1}$ , and hence  $M - F_{n-2} < F_{n-1}$ . Therefore, there exists  $k$  with  $n - 1 \geq k > 0$  such that  $F_{k-1} \leq M - F_{n-2} < F_k$ , and hence by the inductive hypothesis,  $M - F_{n-2}$  can be represented as a sum of distinct Fibonacci elements from the set  $\{F_1, F_2, \dots, F_{k-2}\}$ . Since  $n - 1 \geq k$ , we have  $n - 3 \geq k - 2$ , and so  $M$  can be represented as a sum of distinct Fibonacci elements from the set  $\{F_1, F_2, \dots, F_{n-2}\}$ .  $\square$

We conclude the proof of Theorem 2.1 by proving (c). Let  $S = 2 + T$ , where  $-F_{n-1} \leq T \leq F_{n-1}$ . The cases  $T = F_{n-1}$  and  $T = -F_{n-1}$  were proved in (a) and (b). For  $T = 0$ , let  $A$  be a triangular Toeplitz matrix with first row  $(1 \ 0 \ 1 \ 0 \ 0 \ \dots \ 0)$ . Then  $S(A^{-1}) = 2$ . Similarly, it is easy to prove the claim for any  $S$  between 1 and  $n$ . For the other integers in  $[2 - F_{n-1}, 2 + F_{n-1}]$  (and also for  $1, 2, \dots, n$ ), let us consider the expression in (10). It is easy to see that in fact by choosing appropriate  $\alpha$  and  $\beta$  (and  $C$  such that  $e^T C^{-1}$  is an absolutely dominant vector),  $e^T C^{-1}(e - \alpha - \beta)$  can achieve any value of the form

$$\alpha_1 + \sum_{i=2}^{n-2} \alpha_i F_{i-1},$$

where  $\alpha_i \in \{0, 1\}$  for all  $1 \leq i \leq n - 2$ . Note that by Lemma 2.3, there exists appropriate set  $\{\alpha_i\}_{i=1}^{n-2}$  such that

$$T = \sum_{i=2}^{n-2} \alpha_i F_{i-1} \quad (\text{we may choose } \alpha_1 = 0).$$

Hence, for this choice of  $C$ ,  $\alpha$  and  $\beta$ , we get  $A$  such that  $S = T + 2 = e^T A^{-1} e$ . We obtain a similar result for the case  $S = 2 - T$ , where  $0 \leq T \leq F_{n-1}$ , by looking at expression (9), and this completes the proof.  $\square$

As an analogy to the result on rational numbers of [Huang et al. 2013] mentioned in the introduction, we now have the following corollary.

**Corollary 2.4.** *A number  $S$  is equal to  $S(A^{-1})$  for a  $(0, 1)$ -triangular matrix  $A$  if and only if  $S$  is an integer.*

Define  $G_n$  to be the set of  $n \times n$  matrices of the form  $I + B$ , where  $B$  is an  $n \times n$  upper triangular nilpotent matrix with entries from the interval  $[0, 1]$ . Then, using the fact that for an invertible matrix  $A$ ,  $A^{-1} = \text{adj}(A)/\det(A)$ , and that for  $A \in G_n$ ,  $\det(A) = 1$ , we have  $A^{-1} = \text{adj}(A)$  for  $A \in G_n$ . Thus, since  $S(A^{-1})$  is linear in each one of the entries in such a matrix  $A$ , we conclude the following:

**Corollary 2.5.**  $\max_{A \in G_n} S(A^{-1}) = 2 + F_{n-1}$  and  $\min_{A \in G_n} S(A^{-1}) = 2 - F_{n-1}$ .

**Remark 2.6.** For a general  $n \times n$  invertible  $(0, 1)$ -matrix  $A$  (which is not necessarily triangular), the question regarding the minimal or the maximal value that  $S(A^{-1})$  may obtain is still open. For  $n = 3, 4, 5, 6$ , the extremal values are exactly the same as in the triangular case. However, for  $n = 7$ , there exist  $n \times n$  invertible  $(0, 1)$ -matrices  $M$  and  $N$  (which are presented below) such that  $S(M^{-1}) = -7$  and  $S(N^{-1}) = 11$ , whereas in the triangular case, the minimal and the maximal values are  $-6$  and  $10$ , respectively.

$$M = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad N = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

For larger values of  $n$ , the difference between the general and the triangular case gets bigger.

### 3. Extremal matrices

Recall that an invertible triangular  $n \times n$   $(0, 1)$ -matrix  $A$  is extremal if

$$e^T A^{-1} e = 2 \pm F_{n-1}.$$

The matrices  $I_3$  and  $I_4$  are maximizing matrices. The matrices

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

are minimizing matrices.

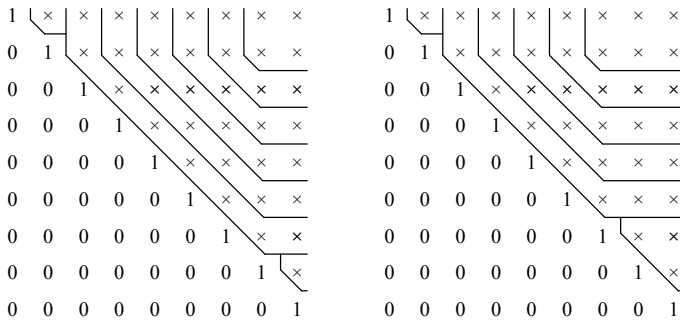
Following the proof of [Theorem 2.1](#), we can construct extremal matrices for  $n \geq 5$  that have a beautiful Fibonacci pattern in their inverses. For  $l = 2, 3$ , partition the off-diagonal entries of an upper triangular  $n \times n$  matrix into  $n - l$  sets,  $S_0, S_1, \dots, S_{n-l-1}$ . The set  $S_{n-l-1}$  consists of the entries in the first two rows of the last  $l$  columns. For  $i = 1, 2, \dots, n - l - 2$ , the set  $S_i$  consists of the entries immediately to the left or immediately below the entries in  $S_{i+1}$ , and  $S_0$  consists of all the remaining entries which are above the main diagonal (two if  $l = 2$  and four if  $l = 3$ ). For example, in the case that  $n = 9$ , [Figure 1](#) (left) presents the partition in the case  $l = 2$ , and [Figure 1](#) (right) presents the partition in the case  $l = 3$ .

Let  $A$  be an invertible  $(0, 1)$ -upper triangular matrix, where the entries in  $S_i$  are taken modulo 2. It follows from the proof of [Theorem 2.1](#) that  $A^{-1}$  is an  $n \times n$  upper triangular matrix where the diagonal entries are 1, the entries in  $S_0$  are 0, and the entries in  $S_i$  for  $i \geq 1$  are  $(-1)^i F_i$ . For example, when  $n = 9, l = 2$ ,

$$A = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad A^{-1} = \begin{pmatrix} 1 & 0 & -1 & 1 & -2 & 3 & -5 & 8 & 8 \\ 0 & 1 & -1 & 1 & -2 & 3 & -5 & 8 & 8 \\ 0 & 0 & 1 & -1 & 1 & -2 & 3 & -5 & -5 \\ 0 & 0 & 0 & 1 & -1 & 1 & -2 & 3 & 3 \\ 0 & 0 & 0 & 0 & 1 & -1 & 1 & -2 & -2 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix};$$

and when  $n = 9, l = 3$ ,

$$A = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad A^{-1} = \begin{pmatrix} 1 & 0 & -1 & 1 & -2 & 3 & -5 & -5 & -5 \\ 0 & 1 & -1 & 1 & -2 & 3 & -5 & -5 & -5 \\ 0 & 0 & 1 & -1 & 1 & -2 & 3 & 3 & 3 \\ 0 & 0 & 0 & 1 & -1 & 1 & -2 & -2 & -2 \\ 0 & 0 & 0 & 0 & 1 & -1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$



**Figure 1.** The partition in the cases  $l = 2$  (left) and  $l = 3$  (right).

In general, if  $n + l$  is even,  $e^T A^{-1} e = 2 - F_{n-1}$ , and hence  $A$  is a minimizing extremal matrix (this also includes the case  $n = 4$ ). If  $n + l$  is odd,  $e^T A^{-1} e = 2 + F_{n-1}$ , and hence  $A$  is a maximizing extremal matrix. Using these equalities, we obtain the following Fibonacci identities:

**Corollary 3.1.**  $\sum_{i=1}^{n-4} (n-i)(-1)^i F_i + 4(-1)^{n-3} F_{n-3} = (-1)^{n-1} F_{n-1} - (n-2)$ .

**Corollary 3.2.**  $\sum_{i=1}^{n-5} (n-i)(-1)^i F_i + 6(-1)^{n-4} F_{n-4} = (-1)^n F_{n-1} - (n-2)$ .

### 4. Determinants of (1, 2)-matrices

In [Huang et al. 2013], the following remark, which follows from Cramer’s rule and the multilinearity of the determinant, was presented:

**Remark 4.1.** For any nonsingular matrix  $A$ ,

$$S(A^{-1}) = \frac{\det(A + J) - \det(A)}{\det(A)},$$

where  $J$  is the matrix whose entries are all 1.

Recall that it was proved in [Li 1993] that the maximal determinant of an  $n \times n$  Hessenberg (0,1)-matrix is  $F_n$ . Using our main result and Remark 4.1, we obtain another family of matrices whose determinants are strongly related to the Fibonacci sequence.

Let  $W_n$  be the family of  $n \times n$  matrices such that for any  $A \in W_n$ ,

$$A_{ij} = \begin{cases} 1 & \text{if } j > i, \\ 2 & \text{if } j = i, \\ 1 \text{ or } 2 & \text{if } j < i. \end{cases}$$

From Remark 4.1 and Theorem 2.1, we obtain the following corollary:

**Corollary 4.2.** Let  $n \geq 3$ . Then  $S = \det(A)$  for some  $A \in W_n$  if and only if  $S$  is an integer that satisfies  $3 - F_{n-1} \leq S \leq 3 + F_{n-1}$ .

## References

- [Atkins and Geist 1987] J. Atkins and R. Geist, “Fibonacci numbers and computer algorithms”, *College Math. J.* **18** (1987), 328–336.
- [Huang et al. 2013] L.-H. Huang, B.-S. Tam, and S.-H. Wu, “Graphs whose adjacency matrices have rank equal to the number of distinct nonzero rows”, *Linear Algebra Appl.* **438**:10 (2013), 4008–4040. [MR 3034513](#)
- [Knuth 1997] D. E. Knuth, *The art of computer programming, I: Fundamental algorithms*, 3rd ed., Addison-Wesley, Reading, MA, 1997. [MR 3077152](#)
- [Lee and Kim 2003] G.-Y. Lee and J.-S. Kim, “The linear algebra of the  $k$ -Fibonacci matrix”, *Linear Algebra Appl.* **373** (2003), 75–87. [MR 2004j:15028](#)
- [Lee et al. 2002] G.-Y. Lee, J.-S. Kim, and S.-G. Lee, “Factorizations and eigenvalues of Fibonacci and symmetric Fibonacci matrices”, *Fibonacci Quart.* **40**:3 (2002), 203–211. [MR 2003k:11024](#)
- [Li 1993] C. Li, “The maximum determinant of an  $n \times n$  lower Hessenberg  $(0, 1)$  matrix”, *Linear Algebra Appl.* **183** (1993), 147–153. [MR 94b:15006](#)
- [Vorobiev 2002] N. N. Vorobiev, *Fibonacci numbers*, Birkhäuser, Basel, 2002. [MR 2003m:11024](#)
- [Wang and Wang 2008] W. Wang and T. Wang, “Identities via Bell matrix and Fibonacci matrix”, *Discrete Appl. Math.* **156**:14 (2008), 2793–2803. [MR 2009g:05018](#)

Received: 2013-08-19

Revised: 2013-10-28

Accepted: 2013-11-05

[miriamf@technion.technion.ac.il](mailto:miriamf@technion.technion.ac.il) *Department of Mathematics, Technion–Israel Institute of Technology, Technion City, Haifa 3200003, Israel*

[berman@technion.technion.ac.il](mailto:berman@technion.technion.ac.il) *Department of Mathematics, Technion–Israel Institute of Technology, Technion City, Haifa 3200003, Israel*





# Stick numbers in the simple hexagonal lattice

Ryan Bailey, Hans Chaumont, Melanie Dennis, Jennifer McLoud-Mann,  
Elise McMahan, Sara Melvin and Geoffrey Schuette

(Communicated by Colin Adams)

In the simple hexagonal lattice, bridge number is used to establish a lower bound on stick numbers of knots. This result aids in giving a new proof that the minimal stick number is 11. In addition, the authors establish upper bounds for the stick number of a composite knot. Constructions for  $(p, p+1)$ -torus knots and some 3-bridge knots are given requiring one more stick than the lower bound guarantees.

## 1. Introduction

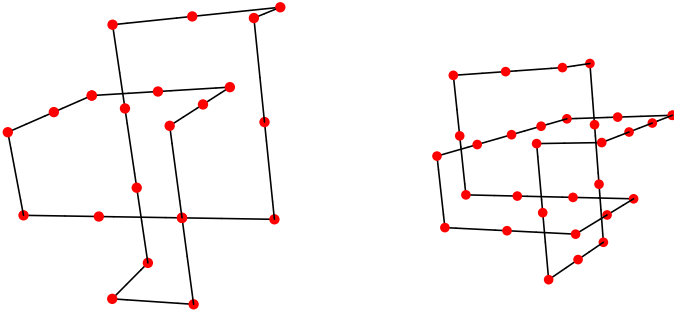
Most results concerning lattice knots have focused on knots in the simple cubic lattice,  $sc$  or  $\mathbb{Z}^3$ . Various lower and upper bounds for stick number in the cubic lattice have been given in [Adams et al. 2012; Janse van Rensburg and Promislow 1999; Hong et al. 2013]. Minimal stick numbers for the  $3_1$  and  $4_1$  knots are 12 and 14 [Huh and Oh 2005]; see Figure 1. The stick number for a  $(p, p+1)$ -torus knot is  $6p$  for  $p \geq 2$  [Adams et al. 2012]. Work has also been done for the minimum stick number of the composition of two knots [Adams et al. 1997; 2012]. Relatively little is known about analogous results in the simple hexagonal lattice. Mann, McLoud-Mann and Milan [Mann et al. 2012] show that the minimum number of sticks to create a nontrivial knot is 11.

In this paper, we will answer some questions regarding the simple hexagonal lattice. In Section 3, we establish a lower bound on the stick number in terms on the bridge number. In Section 4, we give the idea of a new proof of the result in [Mann et al. 2012]. In Section 5, we give an upper bound for the stick number of a composite knot. In Section 6, we catalog results about the stick number of  $(p, p+1)$ -torus knots, some 3-bridge knots, and particular composite knots.

---

*MSC2010:* 57M50.

*Keywords:* lattice knots, stick number, composition, bridge number.



**Figure 1.** Minimal stick  $3_1$  (left) and  $4_1$  (right) knots in the simple cubic lattice.

### 2. Some preliminaries

We will adopt notation for the simple hexagonal lattice from [Mann et al. 2012], which we include here for completeness. The simple hexagonal lattice is defined to be the set of all integral combinations of vectors

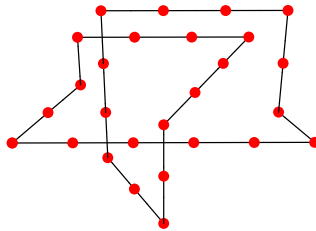
$$x = \langle 1, 0, 0 \rangle, \quad y = \langle \frac{1}{2}, \frac{\sqrt{3}}{2}, 0 \rangle, \quad w = \langle 0, 0, 1 \rangle;$$

that is,

$$\text{sh} = \{a\langle 1, 0, 0 \rangle + b\langle \frac{1}{2}, \frac{\sqrt{3}}{2}, 0 \rangle + c\langle 0, 0, 1 \rangle \mid a, b, c \in \mathbb{Z}\}.$$

Further, let  $X = -x$ ,  $Y = -y$ ,  $W = -w$ ,  $z = \langle -\frac{1}{2}, \frac{\sqrt{3}}{2}, 0 \rangle$ , and  $Z = -z$  so that we can describe a polygon by a string of vectors. In Figure 2, the polygon may be written as  $x^5 z w^2 X^3 W^3 Z^2 w^2 y^3 X^3 W Y^2$ .

A maximal segment in a polygon  $\mathcal{P}$  which is parallel to  $x = \langle 1, 0, 0 \rangle$  will be called an  $x$ -stick. Similarly, define  $y$ -,  $z$ -, and  $w$ -sticks to be maximal segments in  $\mathcal{P}$  which are parallel to  $\langle \frac{1}{2}, \frac{\sqrt{3}}{2}, 0 \rangle$ ,  $\langle -\frac{1}{2}, \frac{\sqrt{3}}{2}, 0 \rangle$ , and  $\langle 0, 0, 1 \rangle$ , respectively. A closed nonintersecting polygon formed from  $x$ -,  $y$ -,  $z$ -, and  $w$ -sticks is called an sh lattice knot. The number of  $x$ -,  $y$ -,  $z$ -, and  $w$ -sticks in a polygon  $\mathcal{P}$  will be denoted  $|\mathcal{P}|_x$ ,  $|\mathcal{P}|_y$ ,  $|\mathcal{P}|_z$ , and  $|\mathcal{P}|_w$ , respectively, and the total number of sticks used will be  $|\mathcal{P}|$ .



**Figure 2.** A trefoil knot in the simple hexagonal lattice.

The stick number of a knot type  $K$  in the lattice, denoted  $s[K]$ , is the minimum number of sticks required to form a polygon of type  $K$ . In Figure 2,  $|\mathcal{P}|_x = 3$ ,  $|\mathcal{P}|_y = 2$ ,  $|\mathcal{P}|_z = 2$ ,  $|\mathcal{P}|_w = 4$ , and  $|\mathcal{P}| = 11$ . Further, observe that  $s[3_1] \leq 11$ .

### 3. Lower bound for stick numbers

Janse van Rensburg and Promislow [1999] established the lower bound for the stick number of a knot in the simple cubic lattice with three directions  $x = \langle 1, 0, 0 \rangle$ ,  $y = \langle 0, 1, 0 \rangle$ , and  $z = \langle 0, 0, 1 \rangle$ ; it was  $6b[K]$  where  $b[K]$  is the bridge number of the knot  $K$  (the minimum number of local maxima of any projection of a knot onto any single vector). The proof guaranteed  $2b[K]$  sticks in each of the three directions. Indeed, maximums in the up-down direction, or  $z$ -direction, will occur in  $xy$ -planes and each maximum will have two  $z$ -sticks at the ends of the arc containing the maximum in the  $xy$ -plane. We give a similar result here for the simple hexagonal lattice.

**Theorem 1** (lower bound for stick numbers). *For any knot  $K$  in the simple hexagonal lattice,  $s[K] \geq 5b[K]$ .*

*Proof.* A maximum in the  $w$ -direction, occurring in an  $xy$ -plane, will have two  $w$ -sticks at the ends of the arc containing the maximum in the  $xy$ -plane. Note that using a  $z$ -stick at the end of the arc would keep you in the same  $xy$ -plane. Since there are at least  $b[K]$  maxima, we have  $|\mathcal{P}|_w \geq 2b[K]$ .

A maxima occurring in an  $xw$ -plane will have two sticks at the ends of the arc containing the maximum in the  $xw$ -plane — these sticks can be  $y$ - or  $z$ -sticks. Since there are at least  $b[K]$  maxima, we have  $|\mathcal{P}|_y + |\mathcal{P}|_z \geq 2b[K]$ . One also considers maxima occurring in  $yw$ -planes and  $zw$ -planes to get two more inequalities summarized below:

$$|\mathcal{P}|_w \geq 2b[K], \tag{1}$$

$$|\mathcal{P}|_y + |\mathcal{P}|_z \geq 2b[K], \tag{2}$$

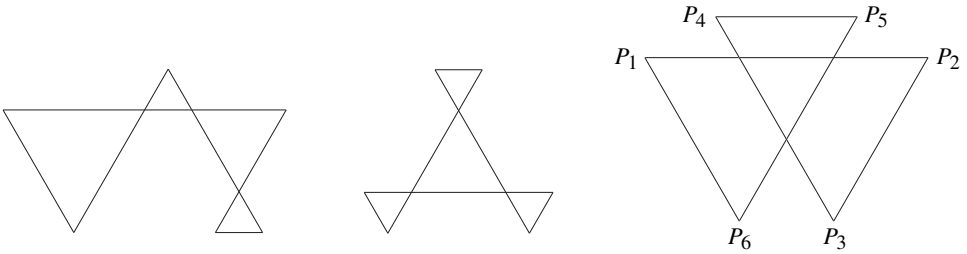
$$|\mathcal{P}|_x + |\mathcal{P}|_z \geq 2b[K], \tag{3}$$

$$|\mathcal{P}|_x + |\mathcal{P}|_y \geq 2b[K]. \tag{4}$$

Summing inequalities (2)–(4) and dividing by 2 yields  $|\mathcal{P}|_x + |\mathcal{P}|_y + |\mathcal{P}|_z \geq 3b[K]$ . Then adding inequality (1) gives  $|\mathcal{P}| = |\mathcal{P}|_x + |\mathcal{P}|_y + |\mathcal{P}|_z + |\mathcal{P}|_w \geq 5b[K]$ .  $\square$

At this point, we can say that the stick number of any nontrivial knot in the simple hexagonal lattice is at least 10. However, in [Mann et al. 2012], it was shown to be 11. In the next section we show that any polygon constructed with ten sticks in the simple hexagonal lattice is the trivial polygon. Before we proceed, we point out what must happen if  $|\mathcal{P}| = 5b[K]$ .

**Corollary 2.** *If  $|\mathcal{P}| = 5b[K]$ , then  $|\mathcal{P}|_x = |\mathcal{P}|_y = |\mathcal{P}|_z = \frac{1}{2}|\mathcal{P}|_w = b[K]$ .*



**Figure 3.** Three crossing projections of ten stick sh knots.

*Proof.* Suppose  $|\mathcal{P}|_x \neq b[K]$ ,  $|\mathcal{P}|_y \neq b[K]$ ,  $|\mathcal{P}|_z \neq b[K]$ , or  $|\mathcal{P}|_w \neq 2b[K]$ . If  $|\mathcal{P}|_w > 2b[K]$  is combined with  $|\mathcal{P}|_x + |\mathcal{P}|_y + |\mathcal{P}|_z \geq 3b[K]$ , the argument above yields  $|\mathcal{P}| > 5b[K]$ . For the remainder of the argument we may assume  $|\mathcal{P}|_w = 2b[K]$ .

If  $|\mathcal{P}|_x < b[K]$ , then  $|\mathcal{P}|_x = b[K] - n$  for some  $n > 0$ . Inequalities (3) and (4) imply that  $|\mathcal{P}|_y \geq b[K] + n$  and  $|\mathcal{P}|_z \geq b[K] + n$ . Thus  $|\mathcal{P}| \geq 5b[K] + n > 5b[K]$ . Following a similar argument, if  $|\mathcal{P}|_y < b[K]$  or  $|\mathcal{P}|_z < b[K]$ , then  $|\mathcal{P}| > 5b[K]$ . Hence for the remainder of the argument we may assume  $|\mathcal{P}|_x \geq b[K]$ ,  $|\mathcal{P}|_y \geq b[K]$  and  $|\mathcal{P}|_z \geq b[K]$ . Observe that since one of these inequalities is strict from our original assumption, it must happen that  $|\mathcal{P}| > 5b[K]$ . □

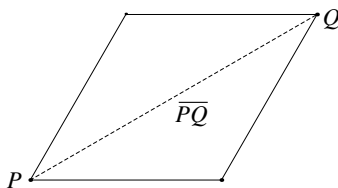
### 4. Stick number of the lattice

As mentioned in the previous section, the stick number of any nontrivial knot in the simple hexagonal lattice is at least 10. The work in this section will show that a simple hexagonal knot constructed with ten sticks (necessarily using two  $x$ -sticks, two  $y$ -sticks, two  $z$ -sticks, and four  $w$ -sticks from Corollary 2) is the trivial knot. This, along with the eleven-stick trefoil in Figure 2, will establish the following result.

**Theorem 3** (minimum stick number in the simple hexagonal lattice). *In the simple hexagonal lattice, the stick number of any nontrivial knot is at least 11.*

Given a ten-stick knot  $K$  using two  $x$ -sticks, two  $y$ -sticks, two  $z$ -sticks, and four  $w$ -sticks, consider the projection of  $K$  onto the  $xy$ -plane. If the projection contains two line segments laying on top of one another or multiple crossings at one point, then do a slight perturbation of the knot before projecting. If the projection contains less than three crossings, then the knot is trivial. There are only a few possibilities for projections containing three crossings; see Figure 3 for representative projections.

The first two projections are the trivial knot. For the last projection, it must have alternating crossings to be a nontrivial knot. However, it cannot have alternating crossings in the hexagonal lattice. Indeed, label the endpoints of the projection  $P_1, P_2, P_3, P_4, P_5,$  and  $P_6$  as in Figure 3. Without loss of generality, suppose that  $P_1 P_2$  on level  $i$  crosses over  $P_3 P_4$  on level  $j$ ; that is,  $i > j$ . Alternating crossings gives



**Figure 4.** Connecting sh lattice points  $P$  and  $Q$  with two sticks.

that  $P_3P_4$  on level  $j$  crosses over  $P_5P_6$  on level  $k$  and  $P_5P_6$  crosses over  $P_1P_2$ . This gives  $i > j > k > i$ .

### 5. Upper bound for stick composition

In order to compose sh knots we must identify places on the knots to compose them; these will be called *configurations*. To achieve the highest reduction of sticks and edges in the composition of sh lattice knots, we will compose knots with configurations in planes parallel to the  $xy$ -plane. In particular, we will compose with configurations in the top  $xy$ -plane or the bottom  $xy$ -plane of a knot.

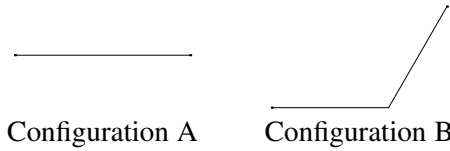
Suppose  $K$  is a minimal stick conformation in the sh lattice — that is, it can't be constructed with fewer sticks. If  $K$  contains more than one connected component in the top  $xy$ -plane, then the vertical sticks for one connected component can be lengthened in order to push that connected component to a higher  $xy$ -plane without increasing the number of sticks used to create  $K$ . Thus one may assume that the top  $xy$ -plane (and similarly the bottom  $xy$ -plane) contains only one connected component. The two endpoints  $P$  and  $Q$  of the connected component can either be connected via one stick or two sticks since there are no other components to avoid when creating a path. To see this, consider the angles between the vector  $\overrightarrow{PQ}$  and the vectors  $\pm\langle 1, 0, 0 \rangle$ ,  $\pm\langle \frac{1}{2}, \frac{\sqrt{3}}{2}, 0 \rangle$ ,  $\pm\langle -\frac{1}{2}, \frac{\sqrt{3}}{2}, 0 \rangle$ . If one of the angles is zero, then  $P$  and  $Q$  are connected with one stick. If not, then we construct a parallelogram with  $P$  and  $Q$  on opposite corners using the two vectors which yield the smallest two angles from above. Note that  $\overrightarrow{PQ}$  forms the major axis of the parallelogram. In this situation  $P$  and  $Q$  can be connected via two sticks. An example is given in Figure 4.

Thus after possibly rotating the knot around the  $z$ -axis, we have two possible configurations occurring in the top or bottom  $xy$ -plane as shown in Figure 5.

**Theorem 4.** *Given knots  $K$  and  $L$  in the simple hexagonal lattice,*

$$s[K\#L] \leq s[K] + s[L] - 3.$$

*Proof.* Let  $K$  and  $L$  be two knots in minimal stick conformations in the simple hexagonal lattice. We will compose  $K$  along a configuration in the bottom  $xy$ -plane and  $L$  along a configuration in the top  $xy$ -plane. Finally, when expressing  $K$  and  $L$

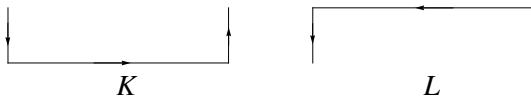


**Figure 5.** Configurations in sh.

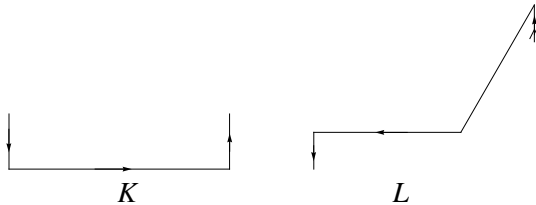
as strings we will choose convenient starting places and orientations to allow for easier composition.

**Case 1.** Suppose  $K$  and  $L$  both have type A configurations. Then the bottom and top configurations of  $K$  and  $L$ , respectively, can be viewed as in Figure 6. Let  $K = sx^n$  and  $L = X^m t$ , where the strings  $s$  and  $t$  represent what remains of  $K$  and  $L$  after the type A configurations are removed. Note that  $s$  will begin with a  $w$  and end with a  $W$ , whereas  $t$  will begin with a  $W$  and end with a  $w$ . Assuming that  $n \neq m$ , we scale  $K$  by  $m$  and scale  $L$  by  $n$ . We have  $K = \tilde{s}x^{nm}$  and  $L = X^{nm}\tilde{t}$ , where  $\tilde{s}$  represents  $s$  scaled by  $m$  and  $\tilde{t}$  represents  $t$  scaled by  $n$ . (In the case that  $n = m$ ,  $\tilde{s} = s$  and  $\tilde{t} = t$ .) We may now compose  $K$  and  $L$ , and write  $K\#L = \tilde{s}\tilde{t}$ . At first glance it may seem that we have removed only two sticks (from the  $x$ s and  $X$ s). However, we have removed two more sticks. The end of  $\tilde{s}$  and the beginning of  $\tilde{t}$  have combined into one stick instead of two. Similarly the end of  $\tilde{t}$  and beginning of  $\tilde{s}$  have combined into one stick. Thus we have a reduction of four sticks for this case. That is,  $s[K\#L] \leq s[K] + s[L] - 4$ .

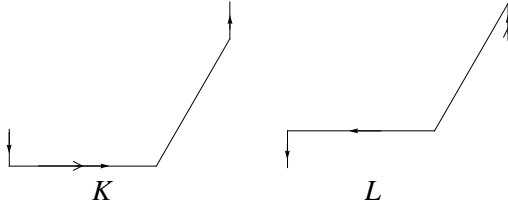
**Case 2.** Suppose  $K$  has a type A configuration and  $L$  has a type B configuration. Then the bottom and top configurations of  $K$  and  $L$ , respectively, can be viewed as in Figure 7. Let  $K = sx^n$  and  $L = X^m t Y^p$ , where strings  $s$  and  $t$  represent what



**Figure 6.**  $K$  and  $L$  with type A configurations: bottom and top, respectively.



**Figure 7.**  $K$  with type A configuration and  $L$  with type B configuration: bottom and top, respectively.



**Figure 8.**  $K$  and  $L$  with type  $B$  configurations: bottom and top, respectively.

remains of  $K$  and  $L$  after the type A and B configurations are removed. Note that  $s$  will begin with a  $w$  and end with a  $W$ , whereas  $t$  will begin with a  $W$  and end with a  $w$ . Assuming that  $n \neq m$ , we scale  $K$  by  $m$  and scale  $L$  by  $n$ . We have  $K = \tilde{s}x^{nm}$  and  $L = X^{nm}\tilde{t}Y^{np}$ , where  $\tilde{s}$  represents  $s$  scaled by  $m$  and  $\tilde{t}$  represents  $t$  scaled by  $n$ . (In the case that  $n = m$ ,  $\tilde{s} = s$  and  $\tilde{t} = t$ .) We may now compose  $K$  and  $L$ , and write  $K\#L = \tilde{s}\tilde{t}Y^{np}$ . Thus we have a reduction of three sticks for this case—the first for the  $x$ s, the second for the  $X$ s and the third for putting end of  $\tilde{s}$  together with beginning of  $\tilde{t}$ . Therefore  $s[K\#L] \leq s[K] + s[L] - 3$ .

**Case 3.** Suppose  $K$  and  $L$  both have type B configurations. Then the bottom and top configurations of  $K$  and  $L$ , respectively, can be viewed as in Figure 8. Let  $K = y^m s x^n$  and  $L = X^p t Y^q$ , where the strings  $s$  and  $t$  represent what remains of  $K$  and  $L$  after the type B configurations are removed. Note that  $s$  will begin with a  $w$  and end with a  $W$ , whereas  $t$  will begin with a  $W$  and end with a  $w$ . Assuming that  $n \neq p$ , we scale  $K$  by  $p$  and scale  $L$  by  $n$  to obtain  $K = y^{mp} \tilde{s} x^{np}$  and  $L = X^{np} \tilde{t} Y^{nq}$ , with  $\tilde{s}$  being  $s$  scaled by  $p$ , and  $\tilde{t}$  being  $t$  scaled by  $n$ . We may now compose  $K$  and  $L$ , and write

$$K\#L = \begin{cases} y^{mp-nq} \tilde{s}\tilde{t} & \text{if } mp > nq, \\ \tilde{s}\tilde{t}Y^{nq-mp} & \text{if } mp < nq, \\ \tilde{s}\tilde{t} & \text{if } mp = nq. \end{cases}$$

Thus we have a reduction of at least three sticks for  $mp \neq nq$  and a reduction of at least six sticks for  $mp = nq$ . In other words,

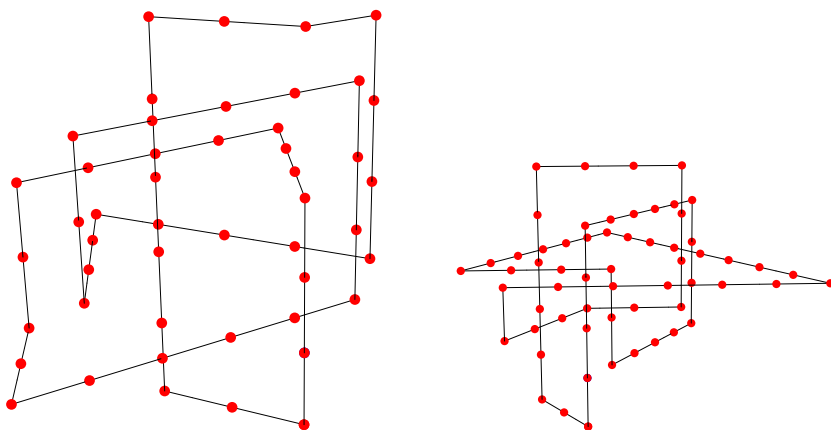
$$s[K\#L] \leq \begin{cases} s[K] + s[L] - 3 & \text{if } mp \neq nq, \\ s[K] + s[L] - 6 & \text{if } mp = nq. \end{cases}$$

Thus we have a minimum reduction of three sticks over all cases. Hence,

$$s[K\#L] \leq s[K] + s[L] - 3. \quad \square$$

### 6. Knot constructions

Adams, Chu, Crawford, Jensen, Siegel and Zhang [Adams et al. 2012] use constructions combined with the lower bound on stick number to establish that the stick number of the 3-bridge knots  $8_{20}$ ,  $8_{21}$ , and  $9_{46}$  are 18 in the simple cubic



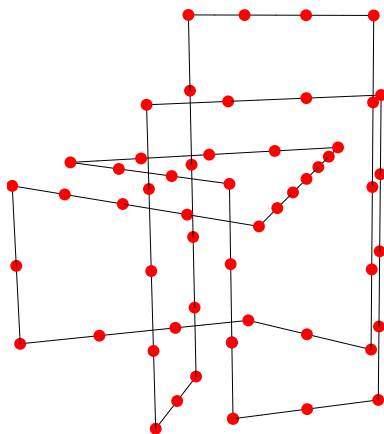
**Figure 9.** 16-stick hexagonal  $8_{20}$  knot (left) and  $8_{21}$  knot (right).

lattice. In a similar manner, one considers these knots in the simple hexagonal lattice. Figures 9 and 10 show these knots built with 16 sticks. Inspection of these knot constructions does not yield any obvious one stick reductions. Using the constructions and Theorem 1, one gets the following theorem.

**Theorem 5.** *In the simple hexagonal lattice, knots  $8_{20}$ ,  $8_{21}$ , and  $9_{46}$  have stick number either 15 or 16.*

Another use of knot construction combined with using the lower bound for stick number can be seen with  $(p, p+1)$ -torus knots.

**Theorem 6** (stick number for  $(p, p+1)$ -torus knots). *For a  $(p, p+1)$ -torus knot  $K$ ,  $5p \leq s[K] \leq 5p + 1$ .*



**Figure 10.** 16-stick hexagonal  $9_{46}$  knot.



*Proof.* Consider a  $(p, p+1)$ -torus knot  $K$  which can be constructed in the simple hexagonal lattice in the following way:

$$Y w^p X^{3+p(p-1)/2} y^p W x^{3+\alpha} \prod_{i=0}^{p-2} (Y^{3-i+\alpha} w^{2i+2} z^{2-i+\alpha} W^{2i+3} x^{3-i+\alpha}),$$

where  $\alpha = (p-2)(p-1)/2$  and an exponent on a letter refers to the edge length of the stick. Notice there are  $5p+1$  sticks used in this construction. In [Schubert 1954], it is shown that  $b[K] = p$ . Using Theorem 1, we have  $s[K] \geq 5p$ . Therefore,  $s[K] = 5p$  or  $s[K] = 5p+1$ .  $\square$

**Corollary 7.** For a  $(p, p+1)$ -torus knot  $K$ ,  $10p-5 \leq s[K\#K] \leq 10p-4$ .

*Proof.* Using two configurations of type B, one sees from Theorem 4 that

$$s[K\#K] \leq 2(5p+1) - 6 = 10p-4.$$

On the other hand, [Schubert 1954] says

$$b[K\#K] = 2b[K] - 1 = 2p - 1,$$

and Theorem 1 yields

$$s[K\#K] \geq 5b[K] \geq 10p - 5. \quad \square$$

## 7. Further work

With all the constructions in the previous section where it is not obvious how to reduce the stick number, it leads one to conjecture that the stick number of a knot is one more than five times its bridge number. It would be nice to prove this improved lower bound or find an example to demonstrate why the standing lower bound is sharp.

**Conjecture.** For any knot  $K$  in the simple hexagonal lattice,  $s[K] \geq 5b[K] + 1$ .

One could try to extend the results to other lattices such as the face-centered cubic lattice and the body-centered cubic lattice. Preliminary investigations of lower bounds for minimal stick number are not great; following similar inequality arguments for these two lattices yields lower bounds of 7 and 8 respectively for 2-bridge knots but has been conjectured to be 9 and 12 via knot constructions [Mann et al. 2012]. A cursory inspection of upper bounds for stick numbers of composite knots suggests that one cannot do better than being subadditive. That is, the stick number of a composite knot is less than or equal to the sum of the stick numbers.

## Acknowledgements

We would like to thank the reviewer for very helpful comments. We would also like to thank the NSF for its support; all authors were supported by DMS NSF grant 1062740 during the summers of 2011 and 2013.

## References

- [Adams et al. 1997] C. C. Adams, B. M. Brennan, D. L. Greilsheimer, and A. K. Woo, “Stick numbers and composition of knots and links”, *J. Knot Theory Ramifications* **6**:2 (1997), 149–161. [MR 98h:57010](#) [Zbl 0884.57005](#)
- [Adams et al. 2012] C. Adams, M. Chu, T. Crawford, S. Jensen, K. Siegel, and L. Zhang, “Stick index of knots and links in the cubic lattice”, *J. Knot Theory Ramifications* **21**:5 (2012), 1250041. [MR 2902272](#) [Zbl 1239.57008](#)
- [Hong et al. 2013] K. Hong, S. No, and S. Oh, “Upper bound on lattice stick number of knots”, *Math. Proc. Cambridge Philos. Soc.* **155**:1 (2013), 173–179. [MR 3065265](#) [Zbl 1270.57022](#)
- [Huh and Oh 2005] Y. Huh and S. Oh, “Lattice stick numbers of small knots”, *J. Knot Theory Ramifications* **14**:7 (2005), 859–867. [MR 2006g:57011](#) [Zbl 1085.57005](#)
- [Mann et al. 2012] C. E. Mann, J. C. McCloud-Mann, and D. P. Milan, “The stick number for the simple hexagonal lattice”, *J. Knot Theory Ramifications* **21**:14 (2012), 1250120. [MR 3021758](#) [Zbl 1270.57029](#)
- [Janse van Rensburg and Promislow 1999] E. J. Janse van Rensburg and S. D. Promislow, “The curvature of lattice knots”, *J. Knot Theory Ramifications* **8**:4 (1999), 463–490. [MR 2000i:57009](#) [Zbl 0940.57013](#)
- [Schubert 1954] H. Schubert, “Über eine numerische Knoteninvariante”, *Math. Z.* **61** (1954), 245–288. [MR 17,292a](#) [Zbl 0058.17403](#)

Received: 2013-10-21      Revised: 2014-05-21      Accepted: 2014-05-23

[rlb3624@utexas.edu](mailto:rlb3624@utexas.edu)

*The University of Texas at Austin,  
Department of Mathematics, 1 University Station C1200,  
Austin, TX 78712, United States*

[chaumont@math.wisc.edu](mailto:chaumont@math.wisc.edu)

*Department of Mathematics,  
University of Wisconsin–Madison, 480 Lincoln Drive,  
Madison, WI 53706, United States*

[melanie.n.dennis.gr@dartmouth.edu](mailto:melanie.n.dennis.gr@dartmouth.edu)

*Department of Mathematics, Dartmouth College,  
27 North Main Street, Hanover, NH 03755, United States*

[jmcloud@uw.edu](mailto:jmcloud@uw.edu)

*Division of Engineering and Mathematics,  
University of Washington Bothell, Box 358538,  
18115 Campus Way NE, Bothell, WA 98011, United States*

[elisemc93@gmail.com](mailto:elisemc93@gmail.com)

*Manteca, CA 95337, United States*

[smelvin@uttyler.edu](mailto:smelvin@uttyler.edu)

*Department of Mathematics, The University of Texas at Tyler,  
3900 University Boulevard, Tyler, TX 75799, United States*

[geoffrey.schuetter@mavs.uta.edu](mailto:geoffrey.schuetter@mavs.uta.edu)

*Department of Mathematics, The University of Texas  
at Arlington, 411 South Nedderman Drive, 478 Pickard Hall,  
Arlington, TX 76019, United States*

# On the number of pairwise touching simplices

Bas Lemmens and Christopher Parsons

(Communicated by Kenneth S. Berenhaut)

In this note, it is shown that the maximum number of pairwise touching translates of an  $n$ -simplex is at least  $n + 3$  for  $n = 7$ , and for all  $n \geq 5$  such that  $n \equiv 1 \pmod{4}$ . The current best known lower bound for general  $n$  is  $n + 2$ . For  $n = 2^k - 1$  and  $k \geq 2$ , we will also present an alternative construction to give  $n + 2$  touching simplices using Hadamard matrices.

## 1. Introduction

A classic problem in discrete geometry is to determine for a given convex body  $K$  in  $\mathbb{R}^n$  the maximum number of pairwise touching translates of  $K$ . This number is called the *touching number of  $K$*  and is denoted by  $t(K)$ . It is well known that for any convex body  $K$  in  $\mathbb{R}^n$ ,

$$t(K) \leq 2^n,$$

and equality holds if, and only if,  $K$  is a parallelotope; see [Danzer and Grünbaum 1962; Petty 1971; Soltan 1975]. On the other hand, it is unknown if for each convex body  $K$  in  $\mathbb{R}^n$  the inequality  $t(K) \geq n + 1$  holds when  $n \geq 4$ ; see [Bezdek 2010, Section 2.3].

This paper concerns the touching number of  $n$ -dimensional simplices,  $\Delta_n$ . This number was studied by Koolen, Laurent and Schrijver [Koolen et al. 2000]. They showed, among other things, that  $t(\Delta_n) \geq n + 2$  for all  $n \geq 3$  and  $t(\Delta_3) = 5$ , see Figure 1. Lemmens [2007] gave examples that showed that  $t(\Delta_4) \geq 7$  and  $t(\Delta_5) \geq 9$ .

The main goal of this short note is to present a construction that gives the following small improvement of the lower bound for  $t(\Delta_n)$ .

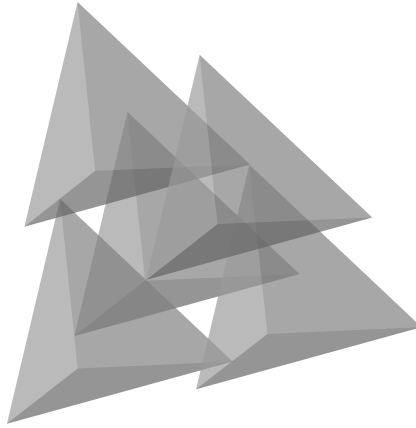
**Theorem 1.1.** *For  $n = 7$  and  $n \equiv 1 \pmod{4}$ , with  $n \geq 5$ , we have that*

$$t(\Delta_n) \geq n + 3.$$

*MSC2010:* primary 52C17; secondary 05B40, 46B20.

*Keywords:* touching number, simplices, equilateral sets,  $\ell_1$ -norm.

Parsons is grateful for the support from School of Mathematics, Statistics and Actuarial Science at the University of Kent and from the HE STEM project “Communicating Mathematical Sciences”.



**Figure 1.** Five pairwise touching tetrahedra.

The problem of determining  $t(\Delta_n)$  is known to be equivalent to finding the maximum size of  $\ell_1$ -norm equilateral sets in a hyperplane [Koolen et al. 2000; Lemmens 2007]. We will discuss the equivalence between these two problems in the next section.

## 2. Equilateral sets

A convex body  $K$  in  $\mathbb{R}^n$  which is centrally symmetric, i.e.,  $x \in K$  if and only if  $-x \in K$ , is the unit ball of a norm  $\|\cdot\|_K$  on  $\mathbb{R}^n$ . Indeed, for  $x \in \mathbb{R}^n$ , we can define the norm by

$$\|x\|_K = \inf\{\lambda > 0 : x \in \lambda K\}.$$

A set  $S$  in a normed space  $(\mathbb{R}^n, \|\cdot\|)$  is called an *equilateral set* if there exists a constant  $\delta > 0$  such that

$$\|s - t\| = \delta \quad \text{for all } s \neq t \text{ in } S.$$

The maximum size of an equilateral set in  $(\mathbb{R}^n, \|\cdot\|)$  is the *equilateral dimension* of  $(\mathbb{R}^n, \|\cdot\|)$  and is denoted by  $e(\mathbb{R}^n, \|\cdot\|)$ . Note that the constant  $\delta > 0$  does not play a role, as we can always scale the equilateral set. Clearly, if  $K$  is a centrally symmetric body in  $\mathbb{R}^n$ , then  $S = \{s_1, \dots, s_p\}$  is an equilateral set in  $(\mathbb{R}^n, \|\cdot\|_K)$  with pairwise distance 2 if, and only if, the set of unit balls with centers  $s_1, \dots, s_p$  is a configuration of  $p$  pairwise touching translates of  $K$ .

The equilateral dimension has been studied for many normed spaces; see, for example, [Alon and Pudlák 2003; Swanepoel 2004a; Swanepoel 2004b]. Particular attention has been given to so-called  $\ell_p$ -norms which are defined as follows. For  $1 \leq p < \infty$ , the  $\ell_p$ -norm on  $\mathbb{R}^n$  is given by  $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$ . For the  $\ell_1$ -norm,

$n = 5$	$n = 6$	$n = 8$
(4, 0, 1, 1, 2)	(4, 0, 1, 1, 1, 1)	(0, 4, 2, 2, 0, 4, 2, 2)
(0, 4, 1, 1, 2)	(0, 4, 1, 1, 1, 1)	(4, 0, 2, 2, 4, 0, 2, 2)
(1, 1, 4, 0, 2)	(1, 1, 4, 0, 1, 1)	(2, 2, 0, 4, 2, 2, 0, 4)
(1, 1, 0, 4, 2)	(1, 1, 0, 4, 1, 1)	(2, 2, 4, 0, 2, 2, 4, 0)
(2, 2, 0, 0, 4)	(1, 1, 1, 1, 4, 0)	(8, 2, 1, 1, 0, 2, 1, 1)
(0, 0, 2, 2, 4)	(1, 1, 1, 1, 0, 4)	(4, 4, 4, 4, 0, 0, 0, 0)
(2, 2, 2, 2, 0)	(2, 2, 2, 2, 0, 0)	(4, 4, 0, 0, 4, 4, 0, 0)
	(2, 2, 0, 0, 2, 2)	(4, 4, 0, 0, 0, 0, 4, 4)
	(0, 0, 2, 2, 2, 2)	(4, 0, 4, 0, 0, 4, 0, 4)
		(4, 0, 0, 4, 0, 4, 4, 0)

**Table 1.** Equilateral sets.

it has been conjectured by Kusner that  $e(\mathbb{R}^n, \|\cdot\|_1) = 2n$  [Guy 1983], but at present this has only been confirmed for  $1 \leq n \leq 4$ ; see [Bandelt et al. 1998; Koolen et al. 2000]. Obviously,  $2n$  is a lower bound for  $e(\mathbb{R}^n, \|\cdot\|_1)$ , as the set of standard basis vectors and their opposites form an equilateral set. The best known upper bound is  $Cn \log n$ , where  $C > 0$  is a constant, which was obtained using probabilistic methods by Alon and Pudlák [2003].

Finding the touching number for the  $n$ -dimensional simplex turns out to be equivalent to determining the maximum size of an  $\ell_1$ -norm equilateral set contained in a hyperplane. More precisely, if  $h(n)$  is the maximum size of an  $\ell_1$ -norm equilateral set in  $H_\alpha = \{x \in \mathbb{R}^n : \sum_i x_i = \alpha\}$  for some  $\alpha \in \mathbb{R}$ , then

$$t(\Delta_n) = h(n + 1) \quad \text{for all } n \geq 1; \tag{2-1}$$

see [Koolen et al. 2000; Lemmens 2007]. For example, the  $\ell_1$ -norm equilateral set

$$S = \{(2, 0, 1, 1), (0, 2, 1, 1), (1, 1, 2, 0), (1, 1, 0, 2), (2, 2, 0, 0)\} \tag{2-2}$$

in the hyperplane  $H_4 \subseteq \mathbb{R}^4$  corresponds to the configuration of five pairwise touching translates of a tetrahedron depicted in Figure 1. The examples of equilateral sets in Table 1 were found with the aid of a computer. In particular, we see that  $t(\Delta_7) \geq 10$ , which settles the  $n = 7$  case in Theorem 1.1.

It is interesting to note that in these examples all the nonzero coordinates are powers of 2. We have looked into those type of examples in more detail, which let to the construction in Proposition 4.1. At present, however, we have no clear understanding of why these coordinate values generate large examples.

Before we prove Theorem 1.1, we mention that the inequalities

$$h(n) \leq e(\mathbb{R}^n, \|\cdot\|_1) \leq h(2n - 1)$$

are known to hold for all  $n \geq 1$  [Koolen et al. 2000; Lemmens 2007]. Thus,  $e(\mathbb{R}^n, \|\cdot\|_1)$  grows linearly in  $n$  if, and only if,  $h(n)$  does.

### 3. Proof of Theorem 1.1

For each  $n \equiv 2 \pmod{4}$  with  $n \geq 6$ , we shall construct an  $\ell_1$ -norm equilateral set in  $H_\alpha = \{x \in \mathbb{R}^n : \sum_i x_i = \alpha\}$  of size  $n + 2$ , where  $\alpha = (n - 2)^2/2$ . The result then follows from Equation (2-1). So let  $n \equiv 2 \pmod{4}$  with  $n \geq 6$ . Define

$$\begin{aligned} v^1 &= (b, 0, a, a, \dots, a, a), \\ v^2 &= (0, b, a, a, \dots, a, a), \\ v^3 &= (a, a, b, 0, \dots, a, a), \\ v^4 &= (a, a, 0, b, \dots, a, a), \\ &\vdots \\ v^{n-1} &= (a, a, a, a, \dots, b, 0), \\ v^n &= (a, a, a, a, \dots, 0, b), \end{aligned}$$

in  $\mathbb{R}^n$ , where  $a = (n - 4)/2$  and  $b = n - 2$ . Furthermore let

$$v^{n+1} = (\overbrace{y, y, \dots, y}^k, \overbrace{z, z, \dots, z}^{n-k}) \quad \text{and} \quad v^{n+2} = (\overbrace{z, z, \dots, z}^{n-k}, \overbrace{y, y, \dots, y}^k)$$

in  $\mathbb{R}^n$ . We now show that if we take

$$k = \frac{n-2}{2}, \quad y = \frac{n-6}{2}, \quad \text{and} \quad z = \frac{n-2}{2},$$

then  $V = \{v^1, \dots, v^{n+2}\}$  is an  $\ell_1$ -norm equilateral set in  $H_\alpha$ , where  $\alpha = (n - 2)^2/2$  and the distance is  $2(n - 2)$ .

To verify this we note first that  $b \geq z \geq a \geq y \geq 0$ . For  $i = 1, \dots, n$ , the coefficient sum of  $v^i$  is given by

$$b + (n - 2)a = n - 2 + \frac{(n-2)(n-4)}{2} = \frac{(n-2)^2}{2}.$$

Similarly the coefficient sum for the vectors  $v^{n+1}$  and  $v^{n+2}$  is equal to

$$(n - k)z + ky = \frac{(n+2)(n-2)}{4} + \frac{(n-2)(n-6)}{4} = \frac{(n-2)^2}{2}.$$

Let  $1 \leq i \neq j \leq n$ . For  $i = 2k - 1$  and  $j = 2k$ , the distance between  $v^i$  and  $v^j$  is given by

$$\|v^i - v^j\|_1 = |b - 0| + |0 - b| = 2(n - 2),$$

and for all other  $i \neq j$ ,

$$\|v^i - v^j\|_1 = |b - a| + |0 - a| + |a - b| + |a - 0| = 2(b - a) + 2a = 2(n - 2).$$

Also

$$\|v^{n+1} - v^{n+2}\|_1 = k|z - y| + k|y - z| = (n - 2)\left(\frac{n-2}{2} - \frac{n-6}{2}\right) = 2(n - 2).$$

Finally the distance between any of the first  $n$  vectors and the last two is calculated as in either the case of  $v^1$  and  $v^{n+1}$ ,

$$\begin{aligned} \|v^1 - v^{n+1}\|_1 &= |b - y| + |0 - y| + (k - 2)|a - y| + (n - k)|a - z| \\ &= n - 2 + \frac{n-6}{2} + \frac{n+2}{2} \\ &= 2(n - 2), \end{aligned}$$

or, as in the case of  $v^1$  and  $v^{n+2}$ ,

$$\begin{aligned} \|v^1 - v^{n+2}\|_1 &= |b - z| + |0 - z| + (n - k - 2)|a - z| + k|a - y| \\ &= n - 2 + \frac{n-2}{2} + \frac{n-2}{2} \\ &= 2(n - 2). \end{aligned}$$

Thus,  $V$  is an  $\ell_1$ -norm equilateral set in  $H_\alpha$  of size  $n + 2$ . [Table 2](#) shows examples in dimensions  $n = 6, 10$  and  $14$ .

#### 4. Hadamard matrices

In this section, we will give an alternative construction that shows that  $t(\Delta_n) \geq n + 2$  for all  $n = 2^k - 1$  with  $k \geq 2$  using  $\ell_1$ -norm equilateral sets and Hadamard matrices. Recall that an  $n \times n$  matrix  $H = [h_{ij}]$  with entries  $h_{ij} \in \{-1, 1\}$  for all  $i$  and  $j$  is called a *Hadamard matrix* if  $HH^T = nI$ . There exists a simple well-known construction of Hadamard matrices of size  $2^k$ . Define  $H_1 = [1]$  and

$$H_{2^{k+1}} = \begin{bmatrix} H_{2^k} & H_{2^k} \\ H_{2^k} & -H_{2^k} \end{bmatrix}$$

for all  $k \geq 1$ . So,

$$H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad H_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}, \quad \dots$$

$n = 6$	$n = 10$	$n = 14$
(4, 0, 1, 1, 1, 1)	(8, 0, 3, 3, 3, 3, 3, 3, 3, 3)	(12, 0, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5)
(0, 4, 1, 1, 1, 1)	(0, 8, 3, 3, 3, 3, 3, 3, 3, 3)	(0, 12, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5)
(1, 1, 4, 0, 1, 1)	(3, 3, 8, 0, 3, 3, 3, 3, 3, 3)	(5, 5, 12, 0, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5)
(1, 1, 0, 4, 1, 1)	(3, 3, 0, 8, 3, 3, 3, 3, 3, 3)	(5, 5, 0, 12, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5)
(1, 1, 1, 1, 4, 0)	(3, 3, 3, 3, 8, 0, 3, 3, 3, 3)	(5, 5, 5, 5, 12, 0, 5, 5, 5, 5, 5, 5, 5, 5)
(1, 1, 1, 1, 0, 4)	(3, 3, 3, 3, 0, 8, 3, 3, 3, 3)	(5, 5, 5, 5, 0, 12, 5, 5, 5, 5, 5, 5, 5, 5)
(2, 2, 2, 2, 0, 0)	(3, 3, 3, 3, 3, 3, 8, 0, 3, 3)	(5, 5, 5, 5, 5, 5, 12, 0, 5, 5, 5, 5, 5, 5)
(0, 0, 2, 2, 2, 2)	(3, 3, 3, 3, 3, 3, 0, 8, 3, 3)	(5, 5, 5, 5, 5, 5, 0, 12, 5, 5, 5, 5, 5, 5)
	(3, 3, 3, 3, 3, 3, 3, 3, 8, 0)	(5, 5, 5, 5, 5, 5, 5, 5, 12, 0, 5, 5, 5, 5)
	(3, 3, 3, 3, 3, 3, 3, 3, 0, 8)	(5, 5, 5, 5, 5, 5, 5, 5, 0, 12, 5, 5, 5, 5)
	(4, 4, 4, 4, 4, 4, 2, 2, 2, 2)	(5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 12, 0, 5, 5)
	(2, 2, 2, 2, 4, 4, 4, 4, 4, 4)	(5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 0, 12, 5, 5)
		(5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 12, 0)
		(5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 0, 12)
		(6, 6, 6, 6, 6, 6, 6, 6, 6, 4, 4, 4, 4, 4, 4)
		(4, 4, 4, 4, 4, 4, 4, 6, 6, 6, 6, 6, 6, 6, 6)

**Table 2.** Equilateral sets of size  $n + 2$ .

Now suppose  $k \geq 2$ . Let  $v^1, \dots, v^{2^k} \in \mathbb{R}^{2^k}$  denote the rows of the Hadamard matrix  $H_{2^k}$ , and define the set

$$V_k = \{v^3\} \cup \{v^i : i = 5, \dots, 2^k\}.$$

Furthermore, let  $W_k = \{w^1, w^2, w^3, w^4\} \in \mathbb{R}^{2^k}$  be given by

$$\begin{aligned} w^1 &= (1, -1, 0, 0, 1, -1, 0, 0, \dots, 1, -1, 0, 0), \\ w^2 &= (-1, 1, 0, 0, -1, 1, 0, 0, \dots, -1, 1, 0, 0), \\ w^3 &= (0, 0, 1, -1, 0, 0, 1, -1, \dots, 0, 0, 1, -1), \\ w^4 &= (0, 0, -1, 1, 0, 0, -1, 1, \dots, 0, 0, -1, 1). \end{aligned}$$

**Proposition 4.1.** *For each  $k \geq 2$ , the set  $V_k \cup W_k$  is an  $\ell_1$ -norm equilateral set of size  $2^k + 1$  in  $H_0 = \{x \in \mathbb{R}^{2^k} : \sum_i x_i = 0\}$ .*

*Proof.* Let  $k \geq 2$ . It is easy to show that each  $u \in V_k \cup W_k$  lies in  $H_0$ . Also note that any two distinct points  $v^i$  and  $v^j$  in  $V_k$  satisfy

$$\|v^i - v^j\|_1 = 2^k,$$

as the rows in  $H_{2^k}$  differ in exactly  $2^{k-1}$  places. The reader can check that

$$\|w^i - w^j\|_1 = 2^k \quad \text{for all } 1 \leq i \neq j \leq 4.$$



So, it remains to show that

$$\|v^i - w^j\|_1 = 2^k \quad \text{for all } v^i \in V_k \text{ and } w^j \in W_k. \quad (4-1)$$

We use induction on  $k$ . Note that if  $k = 2$ , we have that

$$V_2 \cup W_2 = \{(1, 1, -1, -1), (1, -1, 0, 0), (-1, 1, 0, 0), (0, 0, 1, -1), (0, 0, -1, 1)\},$$

which is an  $\ell_1$ -norm equilateral set with distance 4. Now suppose that (4-1) holds for  $k$ . Denote the points in  $V_{k+1}$  by  $\bar{v}^i$  and the points in  $W_{k+1}$  by  $\bar{w}^j$ . Note that for  $j = 1, \dots, 4$ , we have  $\bar{w}^j = (w^j, w^j)$ , where  $w^j \in W_k$ . Also observe that for  $i = 3, 5, \dots, 2^k$ , we have  $\bar{v}^i = (v^i, v^i)$ , and for  $i = 2^k + 1, \dots, 2^{k+1}$ , we have  $\bar{v}^i = (v^{i-2^k}, -v^{i-2^k})$ , where  $v^i \in V_k$ .

So, for  $i = 3, 5, \dots, 2^k$  and  $j = 1, \dots, 4$ , we have that

$$\|\bar{v}^i - \bar{w}^j\|_1 = \sum_{l=1}^{2^{k+1}} |\bar{v}_l^i - \bar{w}_l^j| = 2 \sum_{l=1}^{2^k} |v_l^i - w_l^j| = 2 \cdot 2^k = 2^{k+1}$$

by the induction hypothesis. Also for  $i = 2^k + 1, \dots, 2^{k+1}$  and  $j = 1, \dots, 4$ , we have that

$$\|\bar{v}^i - \bar{w}^j\|_1 = \sum_{l=1}^{2^k} (|v_l^{i-2^k} - w_l^j| + |v_l^{i-2^k} + w_l^j|) = \sum_{l=1}^{2^k} (1 - w_l^j + 1 + w_l^j) = 2^{k+1},$$

as  $v_l^i \in \{-1, 1\}$  and  $-1 \leq w_l^j \leq 1$  for all  $l$ .  $\square$

The reader should note that the equilateral set  $V_k \cup W_k$  can be seen as a generalization of the equilateral set  $S$  in (2-2), as  $V_2 \cup W_2 = S - (1, 1, 1, 1)$ . Furthermore, the example in Table 1 with  $n = 8$  is also of this type, if one ignores the point  $(8, 2, 1, 1, 0, 2, 1, 1)$ .

## References

- [Alon and Pudlák 2003] N. Alon and P. Pudlák, “Equilateral sets in  $\ell_p^n$ ”, *Geom. Funct. Anal.* **13**:3 (2003), 467–482. [MR 2004h:46011](#) [Zbl 1034.46015](#)
- [Bandelt et al. 1998] H.-J. Bandelt, V. Chepoi, and M. Laurent, “Embedding into rectilinear spaces”, *Discrete Comput. Geom.* **19**:4 (1998), 595–604. [MR 99d:51017](#) [Zbl 0973.51012](#)
- [Bezdek 2010] K. Bezdek, *Classical topics in discrete geometry*, Springer, New York, 2010. [MR 2011j:52014](#) [Zbl 1207.52001](#)
- [Danzer and Grünbaum 1962] L. Danzer and B. Grünbaum, “Über zwei Probleme bezüglich konvexer Körper von P. Erdős und von V. L. Klee”, *Math. Z.* **79** (1962), 95–99. [MR 25 #1488](#) [Zbl 0188.27602](#)
- [Guy 1983] R. K. Guy, “An olla-podrida of open problems, often oddly posed”, *Amer. Math. Monthly* **90**:3 (1983), 196–200. [MR 1540158](#)
- [Koolen et al. 2000] J. Koolen, M. Laurent, and A. Schrijver, “Equilateral dimension of the rectilinear space”, *Des. Codes Cryptogr.* **21**:1-3 (2000), 149–164. [MR 2001j:52013](#) [Zbl 0970.51016](#)

- [Lemmens 2007] B. Lemmens, “Variations of a combinatorial problem on finite sets”, *Elem. Math.* **62**:2 (2007), 59–67. [MR 2008e:05017](#) [Zbl 1135.05008](#)
- [Petty 1971] C. M. Petty, “Equilateral sets in Minkowski spaces”, *Proc. Amer. Math. Soc.* **29** (1971), 369–374. [MR 43 #1051](#) [Zbl 0214.20801](#)
- [Soltan 1975] P. S. Soltan, “Analogues of regular simplexes in normed spaces”, *Dokl. Akad. Nauk SSSR* **222**:6 (1975), 1303–1305. In Russian; translated in *Soviet Math. Dokl.* **16**:3 (1975), 787–789. [MR 52 #4127](#) [Zbl 0338.46025](#)
- [Swanepoel 2004a] K. J. Swanepoel, “Equilateral sets in finite-dimensional normed spaces”, pp. 195–237 in *Seminar of mathematical analysis*, edited by D. Girela Álvarez et al., Colección Abierta **71**, Univ. Sevilla Secr. Publ., Seville, 2004. [MR 2005j:46009](#) [Zbl 1071.52008](#)
- [Swanepoel 2004b] K. J. Swanepoel, “A problem of Kusner on equilateral sets”, *Arch. Math. (Basel)* **83**:2 (2004), 164–170. [MR 2005i:52024](#) [Zbl 1062.52017](#)

Received: 2013-12-17

Accepted: 2014-02-23

[b.lemmens@kent.ac.uk](mailto:b.lemmens@kent.ac.uk)

*School of Mathematics, Statistics & Actuarial Science,  
University of Kent, Cornwallis Building, Canterbury,  
CT2 7NF, United Kingdom*

[cmp37@kent.ac.uk](mailto:cmp37@kent.ac.uk)

*School of Mathematics, Statistics & Actuarial Science,  
University of Kent, Cornwallis Building, Canterbury,  
CT2 7NF, United Kingdom*

# The zipper foldings of the diamond

Erin W. Chambers, Di Fang, Kyle A. Sykes,  
Cynthia M. Traub and Philip Trettenero

(Communicated by Kenneth S. Berenhaut)

In this paper, we classify and compute the convex foldings of a particular rhombus that are obtained via a zipper folding along the boundary of the shape. In the process, we explore computational aspects of this problem; in particular, we outline several useful techniques for computing both the edge set of the final polyhedron and its three-dimensional coordinates. We partition the set of possible zipper starting points into subintervals representing equivalence classes induced by these edge sets. In addition, we explore nonconvex foldings of this shape which are obtained by using a zipper starting point outside of the interval corresponding to a set of edges where the polygon folds to a convex polyhedron; surprisingly, this results in multiple families of nonconvex and easily computable polyhedra.

## 1. Introduction

A folding of a polygon is a gluing together of the points on the perimeter to form a polyhedron. A theorem of Alexandrov [1958] shows that as long as the sum of the angles at every glued point is no more than  $2\pi$ , every folding of a convex polygon leads to unique convex polyhedron (in which a doubly covered polygon is considered a flat polyhedron). If a folding meets the requirements for Alexandrov's theorem then we are given the existence of a convex polyhedron corresponding to the folding. A more recent constructive proof by Bobenko and Izmistiev [2008] allows for the explicit construction of a polyhedron by solving a certain differential equation. An implementation of the constructive algorithm has been coded by Stefan Sechelmann<sup>1</sup>. Given any input triangulation of the polygon with gluing instructions, the implementation will output the final polyhedron. However, this particular implementation does not return the corresponding triangulation on the polygon. The algorithm runs in pseudopolynomial time since the algorithm must take the initial triangulation and flip it to a geodesic triangulation, which is not, in general, a polynomial time operation [Kane et al. 2009].

*MSC2010:* 68U05.

*Keywords:* computational geometry, folding algorithms, combinatorial geometry.

<sup>1</sup>[www3.math.tu-berlin.de/geometrie/ps/software.shtml#AlexandrovPolyhedron](http://www3.math.tu-berlin.de/geometrie/ps/software.shtml#AlexandrovPolyhedron)

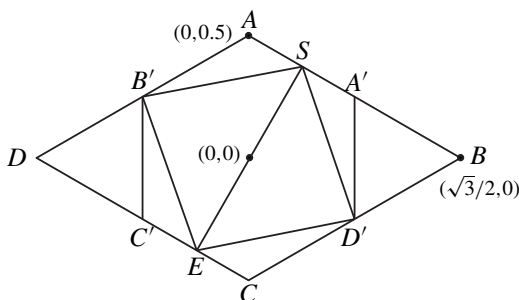
We seek a more combinatorial approach to computing this information. Given a set of gluing rules corresponding to a zipper folding, we outline an approach for computing the crease patterns (i.e., characterize and predict the combinatorial structure of edges and faces) as well as the exact location of the vertices.

**Related work.** Work has been done towards determining all the combinatorially different convex polyhedra obtained via foldings, primarily for regular convex polygons as well as a few other shapes such as the Latin cross [Lubiw and O'Rourke 1996; Alexander et al. 2003; Akiyama and Nakamura 2003; 2004; 2005]. In each work, the authors must determine the set of line segments in the polygon which become edges in the final polyhedron; we refer to these edges as the *crease pattern* for the shape. Note that the crease pattern may not contain all boundary edges of the original polygon; see the left picture in Figure 6 for an example of when the polyhedral edges cross the boundary of the original polygon.

In [Alexander et al. 2003] (and later in [Demaine and O'Rourke 2007]), all (combinatorially distinct) convex polyhedra that are foldable from a square are determined using a combinatorial structure called gluing trees. Crease patterns and reconstructions of the folded polyhedra are also given, making the study of foldings of the square complete. In [Akiyama and Nakamura 2003; 2004; 2005], the focus is on determining all foldings of regular  $n$ -gons, without focusing on reconstructing the actual polyhedron. There is also related work which examines when the Platonic solids can be unzipped to a polygonal net and reziped into a doubly covered flat polygon [O'Rourke 2010]; another paper considers finding different tetrahedra which unzip to a common polygonal net [O'Rourke 2011]. A complete analysis of polyhedra that are zipper foldable from the  $1 \times 2$  rectangle is given in [Schwent 2013], utilizing the techniques outlined here.

As previously mentioned, our primary goal here is to seek a simpler combinatorial approach to verify correct crease patterns in a restricted type of folding. To that end, we consider a restricted class of foldings using the perimeter-halving method, where the perimeter of the polygon is identified starting from a specified point gluing together points equidistant from the starting point (as measured along the perimeter), which zips up the boundary of the polygon into a polyhedron. We will use the term *zipper foldings*, which was first introduced in Demaine et al. [2010]; a related special case is the class of pita polyhedra which arise from zipper folding regular polygons [Demaine and O'Rourke 2007]. As far as the actual resulting polyhedra, surprisingly little is known even about this simple class of foldings aside from the previously mentioned papers which (like ours) consider a particular shape and examine it in detail [Akiyama and Nakamura 2003; 2004; 2005; Alexander et al. 2003].

**Predicting creases.** As was observed by Alexandrov and noted in [Alexander et al. 2003], there are a finite number of possible crease patterns. However, in our



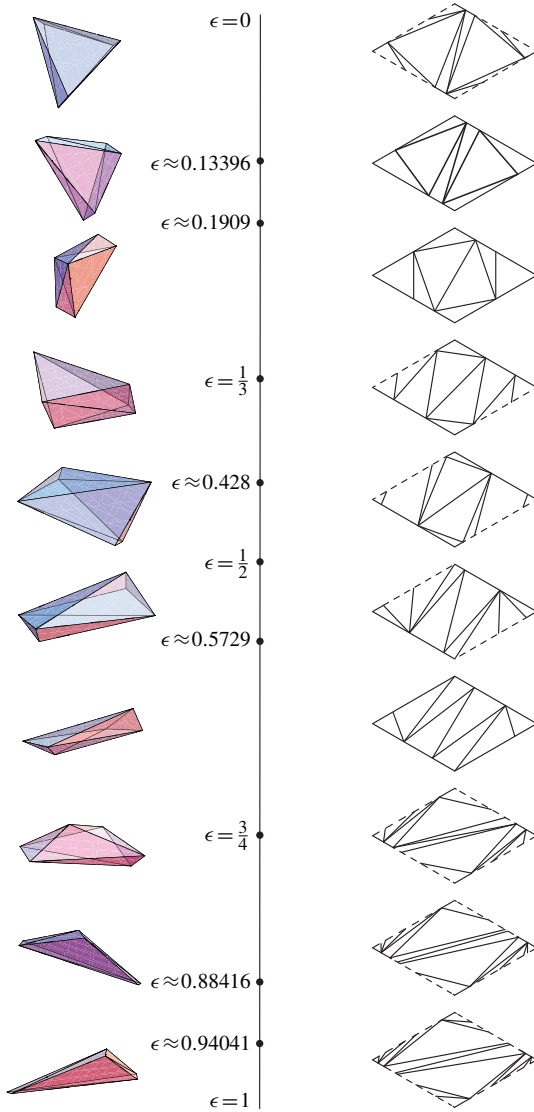
**Figure 1.** The crease pattern at  $\epsilon = \frac{1}{4}$ , which has a 3-regular adjacency graph.

experience, verifying or discounting a crease pattern is surprisingly difficult in more complex polygons since checking a crease pattern either involves seeing if a paper model will fold (highly prone to error) or attempting to compute the folding in a program such as Mathematica (which can lead to numerical issues). As a result, most prior combinatorial work on computing zipper foldings was done using ad hoc methods.

We describe now points on the polygon which will be of interest as we construct the corresponding polyhedron via zipper folding. Our initial polygon is the equilateral rhombus (or diamond) centered on the origin with unit edge length and interior angles  $60^\circ$  and  $120^\circ$ . Label the vertices  $A, B, C, D$ ; see Figure 1 for an example of the labeling. Let the starting point of our zipping  $S$  be a point on edge  $AB$  located at  $(\sqrt{3}\epsilon/2, (1-\epsilon)/2)$ , and refer to the location of  $S$  by this  $\epsilon$ . Note that  $0 \leq \epsilon \leq 1$ , and the location of  $S$  is distance  $\epsilon$  from point  $A$  along edge  $AB$ . The point  $E$  on edge  $CD$  is the reflection of  $S$  through the origin, which is where the zipper ends. For any such folding, we will use  $A', B', C', D'$  to denote the points on the boundary of the polygon which glue to  $A, B, C, D$ , respectively.

As previously mentioned, for polygonal foldings in general, it is known that if the requirements for Alexandrov's theorem are satisfied then there exists a valid folding. Here, we present several computational reconstruction techniques which may be of interest in this area. We also develop several methods to prove that a particular crease pattern is valid as the starting point moves along a continuous interval on the boundary; previous papers seem to have relied on numerical approximation to verify validity, which will reach its limit as the polygon becomes more complex. We also classify all the zipper foldings resulting from the diamond outlined above.

**Theorem 1.** *There are 21 combinatorially distinct convex polyhedra resulting from zipper foldings of a diamond. There are 7 polyhedra which have nontriangular faces and 4 flat polyhedra, all of which occur at isolated points where the crease pattern changes.*



**Figure 2.** All the crease patterns for the zipper foldings of the diamond as the start point  $S$  varies by distance  $\epsilon$  from point  $A$  along edge  $AB$ . Images are taken from sample values between each transition point, marked with solid dots. Dashed lines indicate that a crease extends over an edge. The polyhedra shown between transition points correspond to the respective crease patterns. The symmetry of our input shape allows us to study all zipper foldings by varying the location of  $S$  along one edge; moving  $S$  to another edge will give a combinatorially equivalent folding.

The polyhedra shown in [Figure 2](#) represent the 10 polyhedra with triangular faces, and the solid dots represent the 11 isolated polyhedra noted above. The polyhedra with triangular faces form octahedra. Together, these represent all the zipper foldings of the diamond.

## 2. Computing the foldings

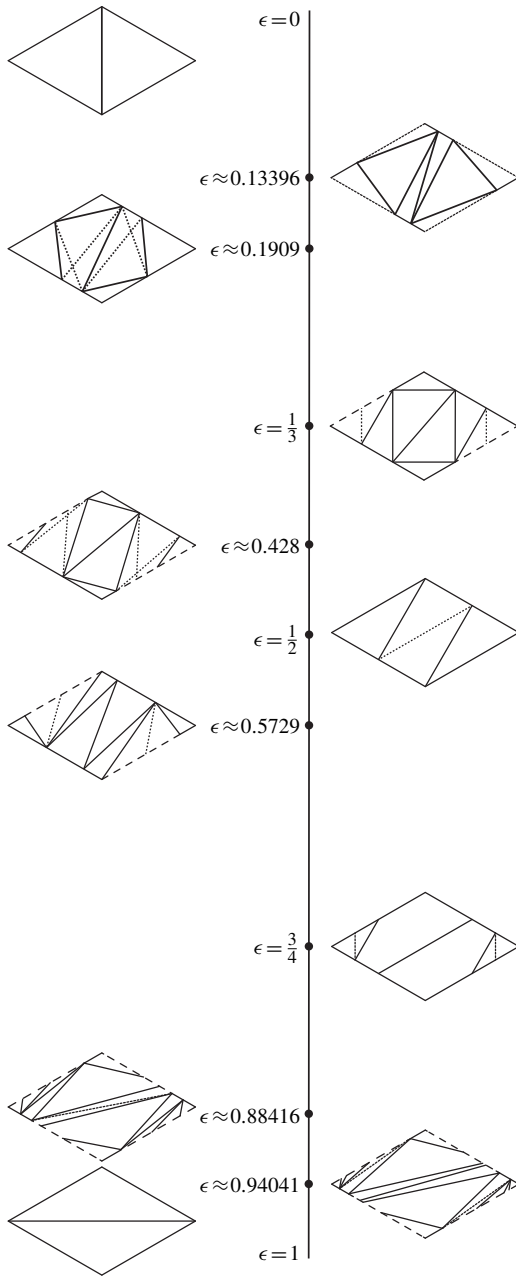
In our foldings, all polyhedra will have at most 6 vertices, resulting from gluing each of  $A$ ,  $B$ ,  $C$ , and  $D$  to some other point on the perimeter, as well as the vertices  $S$  and  $E$ . We are often interested in the actual adjacencies in the final folded polyhedron; this network of edges forms an adjacency graph, often called the graph of the polyhedron, on the (at most) 6 vertices. We refer to this adjacency graph as the *crease pattern*.

Our techniques for computing these foldings break down into several relevant categories. The first (and simplest) are the flat foldings when the entire polygon folds into a doubly covered polygon. For example, when  $\epsilon = 0$ , the vertices  $B$  and  $D$  zip together and the result is a flat doubly covered regular triangle; flat foldings also occur when  $\epsilon = 0.5, 0.75$ , and  $1$ .

The remaining cases in our computation are handled based on whether the graph of the polyhedron is 4-regular or not; if not, in our shape, as well as in the  $1 \times 2$  rectangle studied in [[Schwent 2013](#)], the graph will always consist of vertices of degrees 3, 4, and 5. When degree-3 vertices exist, as discussed in [Section 2.1](#), computing the crease pattern is much simpler since it is not difficult to verify that the underlying structure of the polyhedron can be decomposed into several tetrahedra. The more complex 4-regular case requires additional techniques to calculate exactly; we detail these techniques in [Section 2.2](#). In addition, further complexity arises when the boundary edges of the initial polygon do not become edges in the final polyhedron; see, for example, the crease patterns in [Figure 2](#) which are nearest to  $\epsilon = 1$ . These patterns, which occur much more often in this shape than previous related work, required an extra set of tools to calculate correct crease patterns and 3-D realizations. In [Section 2.3](#), we examine these tools which require zipper folding a related *nonconvex* polygon to yield the same polyhedron.

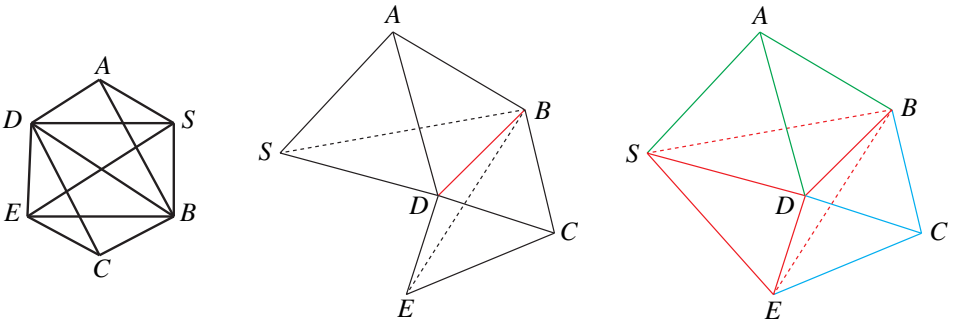
In [Figure 3](#), we show the creases with marked points for the places where the crease pattern undergoes a combinatorial change, which we call a *transition*. Note that (as in previous work) at most of these transitions, two triangles become coplanar to form a quadrilateral (indicated as a dotted line) and then the opposite quadrilateral diagonal appears in the polyhedron. All other transitions occur when the polyhedron folds to a flat doubly covered polygon.

**2.1. Degree-3 vertices in the pattern.** Crease patterns with at least one degree-3 vertex are substantially easier to realize in  $\mathbb{R}^3$ , computationally speaking. In this



**Figure 3.** All the crease patterns at each transition point (which are marked by black dots). Dashed lines indicate that a crease extends over an edge. Dotted lines indicate interior diagonals of a quadrilateral face that are realized as polyhedral edges on either side of the transition point.





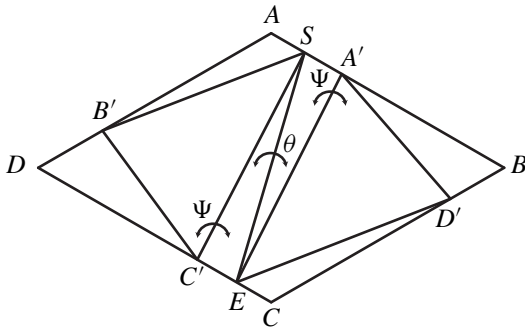
**Figure 4.** Left: The graph of the polyhedron for  $\epsilon = \frac{1}{4}$ . Middle: The two outer tetrahedra of the polyhedron joined along the common edge  $\overline{BD}$ . Right: The final polyhedron decomposed into three tetrahedra.

shape, this results from the fact that when we have such a graph with our setup, we can decompose the final polyhedron into three tetrahedra (two outer tetrahedra and the inner tetrahedron). In Figure 4, we show the adjacency graph of the polyhedron generated when  $\epsilon = \frac{1}{4}$ , the reconstruction of the two outer tetrahedra, and the final polyhedron decomposed into three tetrahedra, where the inner tetrahedron is composed of two triangles from the outer tetrahedra which meet on an edge, plus a single additional edge.

For values of  $\epsilon$  in intervals with a degree-3 vertex in the crease pattern, we wrote code to find exact coordinates for the three-dimensional polyhedron that results. Reconstructing a tetrahedron using adjacencies and edge lengths is not difficult to do, so the general approach we used was to reconstruct the inside tetrahedron shown in Figure 4, and then reconstruct the outer tetrahedra. We next illustrate this process via an example.

Consider the crease pattern at  $\epsilon = \frac{1}{4}$ . This crease pattern contains the edges  $SE$  and  $BD$ . In its initial configuration, we note that points  $B$  and  $D$  are both adjacent to  $SE$  as well as to each other. We can leave edge  $SE$  fixed in the  $z=0$  plane. Rotate points  $B'$  and  $D'$  by  $\theta$  about edge  $SE$  into the positive  $z$ -direction. We solve for the value of  $\theta$  which positions points  $B'$  and  $D'$  at the correct final distance  $|B'D'| = 2\epsilon$  from each other; this establishes a central tetrahedron within our final polygon. Vertex  $A$  is adjacent to  $B'$ ,  $D'$ ,  $S$ , and hence can be located by solving a system of three distance equations. Similarly,  $C$  is adjacent to  $B'$ ,  $D'$ ,  $E$ . The resulting figure is convex and has edge lengths that match those from the polygonal net.

To extend from a specific value of  $\epsilon$  to the entire interval containing  $\epsilon$ , we note that the ability to construct the central tetrahedron  $BDSE$  for  $S$  corresponding to a specific  $0.1909 < \epsilon < \frac{1}{3}$  can be verified via an intermediate value theorem argument. Simply measure the distance between  $B$  and  $D$  when the dihedral angle at  $SE$  is 0

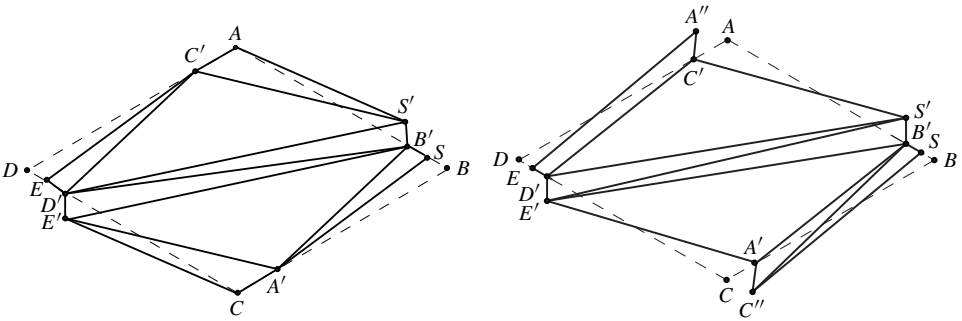


**Figure 5.** A crease pattern whose adjacency graph is 4-regular to illustrate the reconstruction process.

and again when it is  $\pi$ . If the desired length of  $BD$  is between these two values, then folding over  $SE$  by some angle  $0 \leq \theta \leq \pi$  will attain the correct length for  $BD$ . Then, checking the angle criterion given in Lemma 2 (see the Appendix) confirms that locations for points  $A$  and  $C$  can be found that realize all desired distances. It remains only to verify that the resulting polyhedron is convex. Since for  $\epsilon$ -values in this interval, the orthogonal projection of  $A$  onto the plane containing triangle  $BDS$  is interior to triangle  $BDS$ , the final polyhedron will be convex.

**2.2. 4-regular graph of the polyhedron.** In folding patterns where all vertices are degree-4, realization of the vertices in  $\mathbb{R}^3$  is not as simple as the degree-3 case. In [Demaine and O’Rourke 2007], the authors describe a method for constructing an octahedron by splitting it into two smaller hexahedra which share an edge that is an internal diagonal of the octahedron. They vary the length of this edge until the dihedral angles of the faces incident to the edge match. We utilize a different method that also reduces a partial polyhedron to one parameter of change. We illustrate this for  $\epsilon$  in the interval  $0.13396 < \epsilon < 0.1909$ ; the crease pattern for this range is shown in Figure 5.

We consider the following flex over edge  $SE$ . Fold triangles  $SEA'$  and  $SEC'$  upward from the  $z = 0$  plane, each by angle  $\theta$ , leaving edge  $SE$  fixed in the plane. Each choice of  $\theta$  results in a fixed measure of  $\angle A'SC'$  and  $\angle A'EC'$ . The two remaining triangular faces  $SB'C'$  and  $ASB'$  which are incident to  $S$  (or respectively  $E$ ) are uniquely configurable into a shell comprised of faces of the final convex polyhedron. That is, of the two locations in  $\mathbb{R}^3$  for point  $B_\theta$  that give correct distances for segments  $AB$ ,  $BC$ ,  $BS$ , only one is extendable into a convex polyhedron. We similarly find a location for  $D_\theta$ . (The subscripts here serve as a reminder that the locations of  $B_\theta$  and  $D_\theta$  depend on the initial flex by angle  $\theta$ .) Note that this convex shell contains six of the eight faces of the final polyhedron; the two missing faces must share a common edge. We then vary  $\theta$  to realize the correct length for this



**Figure 6.** The crease patterns for  $\epsilon = \frac{7}{8}$  (left) and the range just below 1 (right) rearranged to a nonconvex polygon.

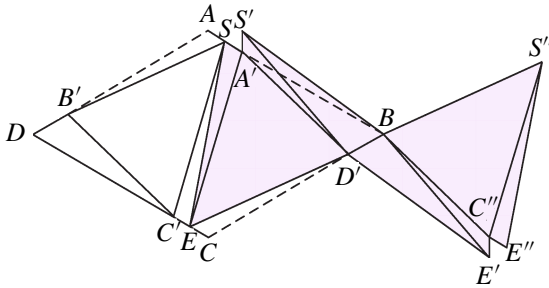
missing common edge; if no such  $\theta$  exists, then we can reject this crease pattern. Moreover, we can also reject the pattern if the final folding results in a nonconvex polyhedron; else, it must realize a convex folding of the initial crease pattern.

**2.3. Creases over the boundary of the polygon.** While the particular approach varies slightly, this process from the previous section can be repeated for any 4-regular graph of the polyhedron. However, some complications arise when the crease pattern is more complex. For example, consider the crease pattern when the source of the zipper  $S$  is near  $B$ . In this case, many of the edges in the final polyhedron actually cross an edge of the initial polygon since not all of the polygon's edges are edges of the final polyhedron. (This also occurs at several other positions; see Figure 2.) Computationally speaking, these patterns are more difficult because a single crease is split into different segments inside the polygon. In order to compute these foldings, we altered the original polygon to be *nonconvex* and verified the crease pattern in this related polygon.

One example occurs when the zipper point reaches near point  $B$  in our shape for the crease patterns above  $\epsilon = 0.75$ ; see Figure 6 for the pattern at  $\epsilon = \frac{7}{8}$ . Here, the creases cross over the edges  $AB'$  (and by symmetry also  $BA'$ ) as well as  $C'D$  (and  $DC'$ ). Using the original gluing information, we reconstructed an equivalent nonconvex polygon which folded to an identical polyhedron and allowed for easier computation, given the symmetry and reduction in the number of creases.

This set of crossings becomes even more drastic as the zipper point nears  $B$ , which is a vertex of high curvature. The crossing edges do not change, but the rearranged figure becomes more complex due to extra crossings, and the nonconvex polygon in turn becomes more complex. See Figure 6 for the final rearranged figure just below  $\epsilon = 1$ .

A very different example occurs for the crease pattern at  $\epsilon = \frac{1}{10}$ ; here, instead of keeping the shape close to the original diamond, we more drastically rearrange



**Figure 7.** A crease pattern for a convex polyhedron when  $\epsilon = \frac{1}{10}$ . The initial net is diamond  $ABCD$ . The shaded faces are shown rearranged to match the algorithmic approach to reconstruction.

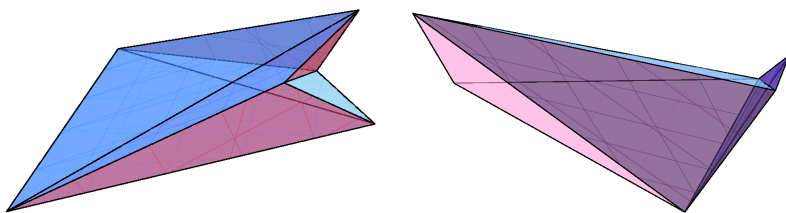
to take advantage of symmetry when computing the folding. This crease pattern contains the edges  $SE$  and  $BD$ . However, in its initial configuration, we note that points  $A'$  and  $C'$  are both adjacent to  $SE$ , but  $A$  and  $C$  are not adjacent to each other in the final polyhedron. We use the gluing instructions to rearrange the triangular faces so that they are as in [Figure 7](#). Now vertices  $S$  and  $E$  are both incident to edge  $BD$ , so we can fold this polygon symmetrically, leaving edge  $BD$  fixed in the  $z = 0$  plane and proceed exactly as outlined in the  $\epsilon = \frac{1}{4}$  case described in [Section 2.1](#).

For nets where the vertices incident to the crease through  $(0, 0)$  are adjacent, this rearranging of the net is not always needed. We do take advantage of this rearrangement technique whenever a three-dimensional edge of our final polyhedron intersects a two-dimensional boundary edge of our initial polygon. Since the gluing instructions are preserved, this is merely a bookkeeping tool that allows for easier computations.

### 3. Nonconvex polyhedra

One interesting result of our investigation of this pattern is a natural classification of some types of *nonconvex* foldings, which to the best of our knowledge have not been a focus of investigation in related work on zipper foldings. It is known, of course, that convex shapes will fold to nonconvex polyhedra, and work has been done on counting the number of foldings of a shape; see, for example, [\[Demaine et al. 2000\]](#). In addition, recent work has focused on unfolding a polyhedron to a convex shape and then refolding it to a different (convex) polyhedron [\[Demaine et al. 2012\]](#); in contrast, our results consider zipper folding a convex planar shape to one or more nonconvex polyhedra, which seems to be an interesting variant of refold rigidity.

The main point of interest is how easy these nonconvex foldings are to find computationally speaking. These foldings result from pushing a particular crease pattern past the point where two faces become coplanar and a flip in the crease



**Figure 8.** Left: A nonconvex folding when the zipper source is at  $\epsilon = .36$ . Right: A nonconvex folding when  $\epsilon = \frac{1}{6}$ ; here, the folding results in a flat flap, indicating that this crease pattern will not fold to a polyhedron at all when pushed lower than  $\frac{1}{6}$ .

pattern occurs. In our experiments, the primary method to establish the validity of a crease pattern is by finding a solution in  $\mathbb{R}^3$  to a system of quadratic equations defining pairwise distances, then checking the convexity of the resulting polyhedron. These nonconvex foldings appeared when the code for computing a solution ran successfully but failed the convexity check.

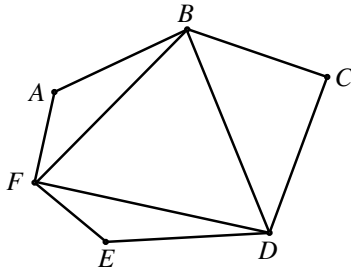
For an example of these, consider [Figure 8](#). In the example shown on the left, we consider when the zipper point is at  $\epsilon = 0.36$ . However, instead of using the correct crease pattern shown in [Figure 2](#), we are instead using the crease pattern for the interval below  $\epsilon = \frac{1}{3}$ . Similarly, on the right side of [Figure 8](#), we have a nonconvex folding when  $\epsilon = \frac{1}{6}$ , but the crease pattern used is the one that is valid for the interval above  $\epsilon \approx 0.1909$ . In this second picture, we have actually pushed the nonconvex folding as far as it will extend, since using this crease pattern for any lower value of  $\epsilon$  will result in an invalid folding (where the polygon self-intersects).

These calculations lead us to conjecture that any valid crease pattern over an interval will fold to a nonconvex polyhedron for some value of  $\epsilon$  close to the interval of convexity. This conjecture is certainly true in our shape (except for near flat foldings), and it seems likely to hold for other shapes since convexity does not impact the existence of a solution.

#### 4. Future directions

We have focused here on zipper foldings of this particular shape. In much of the previous work in this area, all the foldings of a convex shape have been determined using techniques such as gluing trees. Using those techniques to calculate all the convex foldings of this diamond remains an area to address.

Another interesting question is to determine the relationship between the zipper foldings we discuss here with the zipper foldings of the square, as they appear similar [[Alexander et al. 2003](#); [Demaine and O'Rourke 2007](#)]. A link between the diamond and square zipper foldings might give a list of constraints for when two



**Figure 9.** Corresponding figure for [Lemma 2](#).

similar figures have similar foldings. Related lines of questioning could be asked about rhombi in general. For instance, if the edge lengths are similarly defined, are the values of  $\epsilon$  where transitions occur similar?

The nonconvex foldings described in [Section 3](#) are also perhaps worth further investigation in other shapes. It would also be interesting to examine when these nonconvex foldings cease to be valid, and to try to discover how many valid (nonconvex) crease patterns might be present at a particular zipper point.

### Appendix: Realizing tetrahedra

In our discussion of calculating the folding where there is a degree-3 vertex in the graph of the polyhedron, we need a characterization of when a set of vertices and edges can be realized in  $\mathbb{R}^3$  as a tetrahedron. We then use this to help us discover the entire range along which the tetrahedron is present in the final folding. We summarize this tool in the following lemma:

**Lemma 2.** *A net of four triangles as shown in [Figure 9](#) will fold to a tetrahedron if*

- (1) *lengths of corresponding sides are equal ( $|AF| = |EF|$ ,  $|AB| = |BC|$ , and  $|CD| = |DE|$ );*
- (2) *at each vertex, the angle of the base is less than the sum of the other two incident face angles.*

*Proof.* It is clear that the first condition is necessary since, within the tetrahedron, points  $A$ ,  $C$ ,  $E$  will all be identified and thus corresponding edge lengths must be the same. To verify the second condition, we show that, without loss of generality, all points incident to  $F$  can be realized in three dimensions. Assume  $\angle BFD \leq \angle AFB + \angle DFE$ . If this condition were not met, then no position of point  $A$  rotated over segment  $BF$  will coincide with any position of point  $E$  rotated over  $DF$ . When the angle criterion is satisfied, let  $X$  be a point in  $\mathbb{R}^3$  where points  $A$  and  $E$  coincide after rotations over  $BF$  and  $DF$  respectively. We know by condition (1) that  $|XB| = |AB| = |CB|$  and  $|XD| = |ED| = |CD|$ , so triangle  $\triangle BCD$  will fold into position over edge  $BD$  with  $C$  identified with point  $X$  to complete the tetrahedron.  $\square$

## Acknowledgements

This research was supported in part by the National Science Foundation under Grant No. CCF 1054779, as well as an REU supplemental to that grant.

## References

- [Akiyama and Nakamura 2003] J. Akiyama and G. Nakamura, “Foldings of regular polygons to convex polyhedra, II: Regular pentagons”, *J. Indones. Math. Soc.* **9**:2 (2003), 89–99. [MR 2005a:52013](#) [Zbl 1102.52300](#)
- [Akiyama and Nakamura 2004] J. Akiyama and G. Nakamura, “Foldings of regular polygons to convex polyhedra, III: Regular hexagons and regular  $n$ -gons,  $n \geq 7s$ ”, *Thai J. Math.* **2**:1 (2004), 1–15. [Zbl 1066.52014](#)
- [Akiyama and Nakamura 2005] J. Akiyama and G. Nakamura, “Foldings of regular polygons to convex polyhedra, I: Equilateral triangles”, pp. 34–43 in *Combinatorial geometry and graph theory* (Bandung, 2003), edited by J. Akiyama et al., Lecture Notes in Comput. Sci. **3330**, Springer, Berlin, 2005. [MR 2006e:52004](#) [Zbl 1117.52004](#)
- [Alexander et al. 2003] R. Alexander, H. Dyson, and J. O’Rourke, “The foldings of a square to convex polyhedra”, pp. 38–50 in *Discrete and computational geometry* (Tokyo, 2002), edited by J. Akiyama and M. Kano, Lecture Notes in Comput. Sci. **2866**, Springer, Berlin, 2003. [MR 2005g:52047](#) [Zbl 1179.52027](#)
- [Alexandrov 1958] A. D. Alexandrov, *Konvexe polyeder*, Akademie, Berlin, 1958. [MR 19,1192c](#) [Zbl 0079.16303](#)
- [Bobenko and Izvestiev 2008] A. I. Bobenko and I. Izvestiev, “Alexandrov’s theorem, weighted Delaunay triangulations, and mixed volumes”, *Ann. Inst. Fourier (Grenoble)* **58**:2 (2008), 447–505. [MR 2009j:52016](#) [Zbl 1154.52005](#)
- [Demaine and O’Rourke 2007] E. D. Demaine and J. O’Rourke, *Geometric folding algorithms: Linkages, origami, polyhedra*, Cambridge University Press, 2007. [MR 2008g:52001](#) [Zbl 1135.52009](#)
- [Demaine et al. 2000] E. D. Demaine, M. L. Demaine, A. Lubiw, and J. O’Rourke, “Examples, counterexamples, and enumeration results for foldings and unfoldings between polygons and polytopes”, Technical Report 069, Smith College, 2000. [arXiv cs/0007019](#)
- [Demaine et al. 2010] E. D. Demaine, M. L. Demaine, A. Lubiw, A. Shallit, and J. L. Shallit, “Zipper unfoldings of polyhedral complexes”, pp. 219–222 in *Proceedings of the 22nd Canadian Conference on Computational Geometry* (Winnipeg MB, 2010), 2010.
- [Demaine et al. 2012] E. D. Demaine, M. L. Demaine, J. ichi Itoh, A. Lubiw, C. Nara, and J. O’Rourke, “Refold rigidity of convex polyhedra”, preprint, 2012, available at [http://madalgo.au.dk/fileadmin/madalgo/OA\\_PDF\\_s/J119.pdf](http://madalgo.au.dk/fileadmin/madalgo/OA_PDF_s/J119.pdf). In Abstracts from the 28th European Workshop on Computational Geometry.
- [Kane et al. 2009] D. Kane, G. N. Price, and E. D. Demaine, “A pseudopolynomial algorithm for Alexandrov’s theorem”, pp. 435–446 in *Algorithms and data structures* (Banff, Alberta, 2009), edited by F. Dehne et al., Lecture Notes in Comput. Sci. **5664**, Springer, Berlin, 2009. [MR 2550627](#) [Zbl 1253.65028](#)
- [Lubiw and O’Rourke 1996] A. Lubiw and J. O’Rourke, “When can a polygon fold to a polytope?”, Technical Report 048, Smith College, 1996, available at <http://cs.smith.edu/~orourke/Papers/folding.ps.Z>. Presented at AMS Conference, 1996.
- [O’Rourke 2010] J. O’Rourke, “Flat zipper-unfolding pairs for platonic solids”, preprint, 2010. [arXiv 1010.2450](#)

[O'Rourke 2011] J. O'Rourke, "Common edge-unzippings for tetrahedra", preprint, 2011. [arXiv 1105.5401](#)

[Schwent 2013] K. Schwent, *Perimeter-halving the one-by-two rectangle*, Master's thesis, Southern Illinois University Edwardsville, 2013.

Received: 2014-01-31    Revised: 2014-03-20    Accepted: 2014-07-01

[echambe5@slu.edu](mailto:echambe5@slu.edu)    *Department of Mathematics and Computer Science,  
Saint Louis University, St. Louis, MO 63103, United States*

[dfangphone@gmail.com](mailto:dfangphone@gmail.com)    *Department of Mathematics and Computer Science,  
Saint Louis University, St. Louis, MO 63103, United States*

[ksykes2@slu.edu](mailto:ksykes2@slu.edu)    *Department of Mathematics and Computer Science,  
Saint Louis University, St. Louis, MO 63103, United States*

[cytraub@siue.edu](mailto:cytraub@siue.edu)    *Department of Mathematics and Statistics, Southern Illinois  
University Edwardsville, Edwardsville, IL 62026, United States*

[tretten2@illinois.edu](mailto:tretten2@illinois.edu)    *University of Illinois, Urbana, IL 61801, United States*



# On distance labelings of amalgamations and injective labelings of general graphs

Nathaniel Karst, Jessica Oehrlein, Denise Sakai Troxell and Junjie Zhu

(Communicated by Jerrold Griggs)

An  $L(2, 1)$ -labeling of a graph  $G$  is a function assigning a nonnegative integer to each vertex such that adjacent vertices are labeled with integers differing by at least 2 and vertices at distance two are labeled with integers differing by at least 1. The minimum span across all  $L(2, 1)$ -labelings of  $G$  is denoted  $\lambda(G)$ . An  $L'(2, 1)$ -labeling of  $G$  and the number  $\lambda'(G)$  are defined analogously, with the additional restriction that the labelings must be injective. We determine  $\lambda(H)$  when  $H$  is a join-page amalgamation of graphs, which is defined as follows: given  $p \geq 2$ ,  $H$  is obtained from the pairwise disjoint union of graphs  $H_0, H_1, \dots, H_p$  by adding all the edges between a vertex in  $H_0$  and a vertex in  $H_i$  for  $i = 1, 2, \dots, p$ . Motivated by these join-page amalgamations and the partial relationships between  $\lambda(G)$  and  $\lambda'(G)$  for general graphs  $G$  provided by Chang and Kuo, we go on to show that  $\lambda'(G) = \max\{n_G - 1, \lambda(G)\}$ , where  $n_G$  is the number of vertices in  $G$ .

## 1. Introduction

In a well-studied model of the classic channel assignment problem introduced in [Hale 1980], each vertex of a graph  $G$  represents a transmitter in a communications network, and edges connect vertices corresponding to transmitters operating in close proximity which must receive sufficiently different frequencies to avoid interference. In a simplified instance of the problem, a frequency assignment is represented by an  $L(2, 1)$ -labeling of  $G$ , which is a function  $f$  from the vertex set to the nonnegative integers such that  $|f(x) - f(y)| \geq 2$  if vertices  $x$  and  $y$  are adjacent and  $|f(x) - f(y)| \geq 1$  if  $x$  and  $y$  are at distance two.  $L(2, 1)$ -labelings and their variations have been studied extensively since their introduction in [Griggs and Yeh 1992] (see the surveys [Calamoneri 2011; Griggs and Král 2009; Yeh 2006]) and continue to generate a rich literature to this date (see a sample of the

---

MSC2010: primary 68R10, 94C15; secondary 05C15, 05C78.

Keywords:  $L(2, 1)$ -labeling, distance two labeling, injective  $L(2, 1)$ -labeling, amalgamation of graphs, channel assignment problem.

most recent works in [Calamoneri 2013; Franks 2015; Karst et al. 2015; Li and Zhou 2013; Lin and Dai 2015; Lu and Zhou 2013; Shao and Solis-Oba 2013]).

An  $L(2, 1)$ -labeling of a graph  $G$  that uses labels in the set  $\{0, 1, \dots, k\}$  will be called a  $k$ - $L(2, 1)$ -labeling. The minimum  $k$  so that  $G$  has a  $k$ - $L(2, 1)$ -labeling is called the  $\lambda$ -number of  $G$ , denoted by  $\lambda(G)$ . Griggs and Yeh [1992] conjectured that  $\lambda(G) \leq \Delta^2(G)$ , where  $\Delta(G)$  denotes the maximum degree of  $G$ . This conjecture holds for  $\Delta(G) \geq 10^{69}$  [Havet et al. 2012], but it remains open even when  $\Delta(G) = 3$ . The best general upper bound yet established is  $\lambda(G) \leq \Delta^2(G) + \Delta(G) - 2$  [Gonçalves 2008]. Recently, it has been proven that this conjecture also holds for small enough graphs, namely, graphs with at most  $(\lfloor \Delta(G)/2 \rfloor + 1)(\Delta^2(G) - \Delta(G) + 1) - 1$  vertices [Franks 2015]. As the general problem of determining  $\lambda(G)$  is NP-hard [Georges et al. 1994], a significant body of literature has focused on finding bounds or exact  $\lambda$ -numbers for particular classes of graphs. In particular, [Adams et al. 2013] focused on the amalgamations of graphs.

**Definition 1.1.** Let  $H_1, H_2, \dots, H_p$  be  $p \geq 2$  graphs each containing a fixed induced subgraph isomorphic to a graph  $H_0$ . The *amalgamation* of  $H_1, H_2, \dots, H_p$  along  $H_0$  is the simple graph  $H = \text{Amalg}(H_0; H_1, H_2, \dots, H_p)$  obtained by identifying  $H_1, H_2, \dots, H_p$  at the vertices in the fixed subgraphs isomorphic to  $H_0$  in each  $H_1, H_2, \dots, H_p$  respectively.  $H_0$  is referred to as the *spine* and  $H_k$  as the  $k$ -th *page* of the amalgamation for  $k = 1, 2, \dots, p$ . (We refer the reader to [Adams et al. 2013] for some concrete examples.)

In [Adams et al. 2013], upper bounds for the  $\lambda$ -number of the amalgamation of graphs along a given graph were established by determining the exact  $\lambda$ -number of amalgamations of complete graphs along a complete graph. They also provided the exact  $\lambda$ -numbers of amalgamations of rectangular grids along a path, or more specifically, of the Cartesian products of a path and a star with spokes of arbitrary lengths. This focus on the Cartesian products motivated us to investigate amalgamations of the join of graphs.

**Definition 1.2.** Let  $G_1$  and  $G_2$  be two disjoint graphs. The *union*  $G_1 \cup G_2$  is the graph with vertex (resp., edge) set equal to the union of the vertex (resp., edge) sets of  $G_1$  and  $G_2$ . The *join*  $G_1 + G_2$  is obtained from  $G_1 \cup G_2$  by adding an edge between each vertex in  $G_1$  and each vertex in  $G_2$ .

**Definition 1.3.** Let  $G_0, G_1$ , and  $G_2$  be pairwise disjoint graphs. The graph  $G = \text{Amalg}(G_0; G_0 + G_1, G_0 + G_2)$  is called a *join-page amalgamation* of  $G_1, G_2$  along  $G_0$ . Note that  $G$  is isomorphic to  $G_0 + (G_1 \cup G_2)$ .

Definitions 1.2 and 1.3 can be extended for more than two graphs  $G_1, G_2$ . The  $\lambda$ -numbers of the union and join of graphs are well known as stated in the next two results.

**Result 1.4** [Chang and Kuo 1996, Lemma 3.1]. *For any two graphs  $G$  and  $H$ ,  $\lambda(G \cup H) = \max\{\lambda(G), \lambda(H)\}$ .*

**Result 1.5** [Georges et al. 1994, Corollary 4.6]. *For any two graphs  $G$  and  $H$  with  $n_G$  and  $n_H$  vertices respectively,*

$$\lambda(G + H) = \max\{n_G - 1, \lambda(G)\} + \max\{n_H - 1, \lambda(H)\} + 2.$$

In Section 2, we provide the exact  $\lambda$ -number for all join-page amalgamations. Motivated by a connection between this  $\lambda$ -number and the minimum span over injective  $L(2, 1)$ -labelings, Section 3 revisits these labelings for general graphs which were first introduced in [Chang and Kuo 1996]. More specifically, we establish a new exact relationship between the  $\lambda$ -number of a graph and the minimum span over all injective  $L(2, 1)$ -labelings of this graph.

## 2. The $\lambda$ -number of join-page amalgamations

**Theorem 2.1.** *Let  $G = \text{Amalg}(G_0; G_0 + G_1, G_0 + G_2, \dots, G_0 + G_p)$  be a join-page amalgamation, where  $G_i$  is a graph with  $n_i \geq 1$  vertices for  $i = 0, 1, \dots, p \geq 2$  so that  $n_1 \geq n_j$  for  $j = 2, 3, \dots, p$ , and let  $n = n_1 + n_2 + \dots + n_p$ . Then,*

$$\lambda(G) = \max\{n_0 - 1, \lambda(G_0)\} + \max\{n - 1, \lambda(G_1)\} + 2.$$

*Proof.* Since  $G$  is isomorphic to  $G_0 + (G_1 \cup G_2 \cup \dots \cup G_p)$ , using Results 1.4 and 1.5,

$$\begin{aligned} \lambda(G) &= \lambda(G_0 + (G_1 \cup G_2 \cup \dots \cup G_p)) \\ &= \max\{n_0 - 1, \lambda(G_0)\} + \max\{n - 1, \lambda(G_1 \cup G_2 \cup \dots \cup G_p)\} + 2 \\ &= \max\{n_0 - 1, \lambda(G_0)\} + \max\{n - 1, \lambda(G_1), \lambda(G_2), \dots, \lambda(G_p)\} + 2. \end{aligned}$$

For  $i = 2, 3, \dots, p$ , we have  $\lambda(G_i) \leq \lambda(K_{n_i}) = 2n_i - 2 \leq n_1 + n_i - 2 < n - 1$ , where  $K_{n_i}$  denotes the complete graph with  $n_i$  vertices, and therefore

$$\max\{n - 1, \lambda(G_1), \lambda(G_2), \dots, \lambda(G_p)\} = \max\{n - 1, \lambda(G_1)\},$$

and the desired result follows.  $\square$

It is worth noting that Theorem 2.1 implies that  $\lambda(G)$  depends on the number of vertices in  $G_2, G_3, \dots, G_p$  but not on their particular  $\lambda$ -numbers.

The following corollary is equivalent to Theorem 2.3 in [Adams et al. 2013] but with an alternative and more compact proof.

**Corollary 2.2.** *Let  $G = \text{Amalg}(K_0; K_0 + K_1, K_0 + K_2, \dots, K_0 + K_p)$  be a join-page amalgamation, where  $K_i$  is the complete graph with  $n_i \geq 1$  vertices for  $i = 0, 1, \dots, p \geq 2$  so that  $n_1 \geq n_j$  for  $j = 2, 3, \dots, p$ , and let  $n = n_1 + n_2 + \dots + n_p$ . Then  $\lambda(G) = 2n_0 + \max\{n - 1, 2n_1 - 2\}$ .*

*Proof.* By [Theorem 2.1](#),

$$\begin{aligned} \lambda(G) &= \max\{n_0 - 1, \lambda(K_0)\} + \max\{n - 1, \lambda(K_1)\} + 2 \\ &= \max\{n_0 - 1, 2n_0 - 2\} + \max\{n - 1, 2n_1 - 2\} + 2 \\ &= 2n_0 - 2 + \max\{n - 1, 2n_1 - 2\} + 2 \\ &= 2n_0 + \max\{n - 1, 2n_1 - 2\}. \end{aligned} \quad \square$$

### 3. A connection between join-page amalgamation and injective $L(2, 1)$ -labelings

When examining the  $L(2, 1)$ -labelings of a join-page amalgamation of the form  $G = \text{Amalg}(G_0; G_0 + G_1, G_0 + G_2, \dots, G_0 + G_p)$ , as described in [Theorem 2.1](#) in [Section 2](#), we noticed that we could extend an injective  $L(2, 1)$ -labeling of  $G_0$  of minimum span over all its injective labelings to a  $\lambda(G)$ - $L(2, 1)$ -labeling of the entire  $G$ . We suspected that this was not a coincidence, which led us to revisit the following variation of  $L(2, 1)$ -labelings introduced in [\[Chang and Kuo 1996\]](#).

**Definition 3.1.** An  $L'(2, 1)$ -labeling of a graph  $G$  is an injective  $L(2, 1)$ -labeling of  $G$ . The definitions of  $k$ - $L'(2, 1)$ -labeling,  $\lambda'$ -number and  $\lambda'(G)$  are analogous to those of  $k$ - $L(2, 1)$ -labeling,  $\lambda$ -number, and  $\lambda(G)$  when restricted to injective labelings.

The following basic properties were previously known.

**Result 3.2** [\[Chang and Kuo 1996, Lemmas 2.1, 2.2, 2.3\]](#). *For any graph  $G$  with  $n_G$  vertices,*

- (i)  $\lambda'(H) \leq \lambda'(G)$  for any subgraph  $H$  of  $G$ ;
- (ii)  $\lambda(G) \leq \lambda'(G)$  with equality if  $G$  has diameter at most two; and
- (iii)  $c(G) = \lambda'(G^c) - n_G + 2$ , where  $c(G)$  is the path covering number of  $G$ , i.e., the smallest number of vertex-disjoint paths needed to cover all the vertices of the graph  $G$ , and  $G^c$  is the complement of  $G$ .

In [Theorem 3.4](#), we will strengthen [Result 3.2\(ii\)](#) by providing a surprisingly simple exact relationship between  $\lambda(G)$  and  $\lambda'(G)$  for any graph  $G$ . We will be using the following auxiliary result in the proof of [Theorem 3.4](#).

**Result 3.3** [\[Georges et al. 1994, Theorem 1.1\]](#). *For any graph  $G$  on  $n_G$  vertices,*

- (i)  $\lambda(G) \leq n_G - 1$  if and only if  $c(G^c) = 1$ ; and
- (ii)  $\lambda(G) = n_G + c(G^c) - 2$  if and only if  $c(G^c) \geq 2$ .

**Theorem 3.4.** *For any graph  $G$  with  $n_G$  vertices,*

$$\lambda'(G) = \max\{n_G - 1, \lambda(G)\}.$$

*Proof.* Suppose  $\lambda(G) \leq n_G - 1$ . By [Result 3.3\(i\)](#),  $c(G^c) = 1$ , and [Result 3.2\(iii\)](#) implies  $1 = c(G^c) = \lambda'(G) - n_G + 2$ . Therefore,

$$\lambda'(G) = n_G - 1 = \max\{n_G - 1, \lambda(G)\}.$$

Assume, on the other hand, that  $\lambda(G) > n_G - 1$ . Item [\(i\)](#) in [Result 3.3](#) implies  $c(G^c) \geq 2$ , and item [\(ii\)](#) implies  $\lambda(G) = n_G + c(G^c) - 2$ , or equivalently,  $c(G^c) = \lambda(G) - n_G + 2$ . Finally, [Result 3.2\(iii\)](#) implies

$$\begin{aligned} \lambda'(G) &= c(G^c) + n_G - 2 \\ &= (\lambda(G) - n_G + 2) + n_G - 2 = \lambda(G) = \max\{n_G - 1, \lambda(G)\}. \quad \square \end{aligned}$$

In view of [Theorem 3.4](#), the general problem of determining the  $\lambda'$ -number of graphs is as complex as determining their  $\lambda$ -numbers, which, as mentioned previously, is known to be an NP-hard problem. Furthermore, the exact  $\lambda'$ -numbers of families of graphs, such as the ones derived in [[Chang and Kuo 1996](#)] using more involved techniques (e.g., paths, cycles, union and join of two graphs), can be readily obtained using [Theorem 3.4](#) and the vast list of known exact  $\lambda$ -numbers in the  $L(2, 1)$ -labeling literature.

If  $G = \text{Amalg}(G_0; G_0 + G_1, G_0 + G_2, \dots, G_0 + G_p)$  and we apply [Theorem 3.4](#) to  $G_0$  in [Theorem 2.1](#), we obtain a relationship between  $\lambda(G)$  and  $\lambda'(G_0)$ , confirming the connection between injective  $L(2, 1)$ -labelings of  $G_0$  and  $L(2, 1)$ -labelings of  $G$  we mentioned in the first paragraph of this section. The following corollary provides this relationship.

**Corollary 3.5.** *Let  $G = \text{Amalg}(G_0; G_0 + G_1, G_0 + G_2, \dots, G_0 + G_p)$  be a join-page amalgamation, where  $G_i$  is a graph with  $n_i$  vertices for  $i = 0, 1, \dots, p \geq 2$  so that  $n_1 \geq n_j$  for  $j = 2, 3, \dots, p$ , and let  $n = n_1 + n_2 + \dots + n_p$ . Then  $\lambda(G) = \lambda'(G_0) + \max\{n - 1, \lambda(G_1)\} + 2$ .*

## Acknowledgements

The authors would like to thank Sarah Spence Adams for handling administrative requirements regarding student research credits. Denise Sakai Troxell would like to thank Babson College for its support through the Babson Research Scholar award.

## References

- [Adams et al. 2013] S. S. Adams, N. Howell, N. Karst, D. S. Troxell, and J. Zhu, “On the  $L(2, 1)$ -labelings of amalgamations of graphs”, *Discrete Appl. Math.* **161**:7-8 (2013), 881–888. [MR 3030574](#) [Zbl 1263.05086](#)
- [Calamoneri 2011] T. Calamoneri, “The  $L(h, k)$ -labelling problem: An updated survey and annotated bibliography”, *Comput. J.* **54**:8 (2011), 1344–1371.

- [Calamoneri 2013] T. Calamoneri, “Optimal  $L(\delta_1, \delta_2, 1)$ -labeling of eight-regular grids”, *Inform. Process. Lett.* **113**:10-11 (2013), 361–364. MR 3037462 Zbl 06329871
- [Chang and Kuo 1996] G. J. Chang and D. Kuo, “The  $L(2, 1)$ -labeling problem on graphs”, *SIAM J. Discrete Math.* **9**:2 (1996), 309–316. MR 97b:05132 Zbl 0860.05064
- [Franks 2015] C. Franks, “The delta square conjecture holds for graphs of small order”, *Involve: J. Math.* **9**:2 (2015), to be supplied by the publisher.
- [Georges et al. 1994] J. P. Georges, D. W. Mauro, and M. A. Whittlesey, “Relating path coverings to vertex labellings with a condition at distance two”, *Discrete Math.* **135**:1-3 (1994), 103–111. MR 96b:05150 Zbl 0811.05058
- [Gonçalves 2008] D. Gonçalves, “On the  $L(p, 1)$ -labelling of graphs”, *Discrete Math.* **308**:8 (2008), 1405–1414. MR 2008k:05185 Zbl 1135.05065
- [Griggs and Král 2009] J. R. Griggs and D. Král, “Graph labellings with variable weights, a survey”, *Discrete Appl. Math.* **157**:12 (2009), 2646–2658. MR 2010m:05275 Zbl 1211.05145
- [Griggs and Yeh 1992] J. R. Griggs and R. K. Yeh, “Labelling graphs with a condition at distance 2”, *SIAM J. Discrete Math.* **5**:4 (1992), 586–595. MR 93h:05141 Zbl 0767.05080
- [Hale 1980] W. K. Hale, “Frequency assignment: Theory and applications”, *Proc. IEEE* **68**:12 (1980), 1497–1514.
- [Havet et al. 2012] F. Havet, B. Reed, and J.-S. Sereni, “Griggs and Yeh’s conjecture and  $L(p, 1)$ -labelings”, *SIAM J. Discrete Math.* **26**:1 (2012), 145–168. MR 2902638 Zbl 1245.05110
- [Karst et al. 2015] N. Karst, J. Oehrlein, D. S. Troxell, and J. Zhu, “ $L(d, 1)$ -labelings of the edge-path-replacement by factorization of graphs”, *J. Comb. Opt.* **30**:1 (2015), 34–41. MR 3352872
- [Li and Zhou 2013] X. Li and S. Zhou, “Labeling outerplanar graphs with maximum degree three”, *Discrete Appl. Math.* **161**:1-2 (2013), 200–211. MR 2973362 Zbl 06109944
- [Lin and Dai 2015] W. Lin and B. Dai, “On  $(s, t)$ -relaxed  $L(2, 1)$ -labelings of the triangular lattice”, *J. Comb. Optim.* **29**:3 (2015), 655–669. MR 3316710 Zbl 06435135
- [Lu and Zhou 2013] C. Lu and Q. Zhou, “Path covering number and  $L(2, 1)$ -labeling number of graphs”, *Discrete Appl. Math.* **161**:13-14 (2013), 2062–2074. MR 3057011 Zbl 1286.05150
- [Shao and Solis-Oba 2013] Z. Shao and R. Solis-Oba, “ $L(2, 1)$ -labelings on the modular product of two graphs”, *Theoret. Comput. Sci.* **487** (2013), 74–81. MR 3049272 Zbl 1283.05246
- [Yeh 2006] R. K. Yeh, “A survey on labeling graphs with a condition at distance two”, *Discrete Math.* **306**:12 (2006), 1217–1231. MR 2007g:05167 Zbl 1094.05047

Received: 2014-02-03

Revised: 2014-05-24

Accepted: 2014-05-31

nkarst@babson.edu

*Mathematics and Sciences Division, Babson College, Babson Park, MA 02457, United States*

jessica.oehrlein@students.olin.edu

*Franklin W. Olin College of Engineering, Olin Way, Needham, MA 02492, United States*

troxell@babson.edu

*Mathematics and Sciences Division, Babson College, Babson Park, MA 02457, United States*

jjzhu@stanford.edu

*Department of Electrical Engineering, Stanford University, Stanford, CA 94305, United States*

## Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the Submission page at the [Involve website](#).

**Originality.** Submission of a manuscript acknowledges that the manuscript is original and 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  $\text{\LaTeX}$  but submissions in other varieties of  $\text{\TeX}$ , 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  $\text{\BibTeX}$  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](mailto: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

2015

vol. 8

no. 3

Colorability and determinants of $T(m, n, r, s)$ twisted torus knots for $n \equiv \pm 1 \pmod{m}$	361
MATT DELONG, MATTHEW RUSSELL AND JONATHAN SCHROCK	
Parameter identification and sensitivity analysis to a thermal diffusivity inverse problem	385
BRIAN LEVENTHAL, XIAOJING FU, KATHLEEN FOWLER AND OWEN ESLINGER	
A mathematical model for the emergence of HIV drug resistance during periodic bang-bang type antiretroviral treatment	401
NICOLETA TARFULEA AND PAUL READ	
An extension of Young's segregation game	421
MICHAEL BORCHERT, MARK BUREK, RICK GILLMAN AND SPENCER ROACH	
Embedding groups into distributive subsets of the monoid of binary operations	433
GREGORY MEZERA	
Persistence: a digit problem	439
STEPHANIE PEREZ AND ROBERT STYER	
A new partial ordering of knots	447
ARAZELLE MENDOZA, TARA SARGENT, JOHN TRAVIS SHRONTZ AND PAUL DRUBE	
Two-parameter taxicab trigonometric functions	467
KELLY DELP AND MICHAEL FILIPSKI	
${}_3F_2$ -hypergeometric functions and supersingular elliptic curves	481
SARAH PITMAN	
A contribution to the connections between Fibonacci numbers and matrix theory	491
MIRIAM FARBER AND ABRAHAM BERMAN	
Stick numbers in the simple hexagonal lattice	503
RYAN BAILEY, HANS CHAUMONT, MELANIE DENNIS, JENNIFER MCLLOUD-MANN, ELISE MCMAHON, SARA MELVIN AND GEOFFREY SCHUETTE	
On the number of pairwise touching simplices	513
BAS LEMMENS AND CHRISTOPHER PARSONS	
The zipper foldings of the diamond	521
ERIN W. CHAMBERS, DI FANG, KYLE A. SYKES, CYNTHIA M. TRAUB AND PHILIP TRETTENERO	
On distance labelings of amalgamations and injective labelings of general graphs	535
NATHANIEL KARST, JESSICA OEHRLEIN, DENISE SAKAI TROXELL AND JUNJIE ZHU	