

# 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	József H. Przytycki
Andrew Granville	Richard Rebarber
Jerrold Griggs	Robert W. Robinson
Sat Gupta	Filip Saidak
Jim Haglund	James A. Sellers
Johnny Henderson	Andrew J. Sterge
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	



# involve

msp.org/involve

## EDITORS

### MANAGING EDITOR

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

### BOARD OF EDITORS

Colin Adams	Williams College, USA colin.c.adams@williams.edu	David Larson	Texas A&M University, USA larson@math.tamu.edu
John V. Baxley	Wake Forest University, NC, USA baxley@wfu.edu	Suzanne Lenhart	University of Tennessee, USA lenhart@math.utk.edu
Arthur T. Benjamin	Harvey Mudd College, USA benjamin@hmc.edu	Chi-Kwong Li	College of William and Mary, USA ckli@math.wm.edu
Martin Bohner	Missouri U of Science and Technology, USA bohner@mst.edu	Robert B. Lund	Clemson University, USA lund@clemson.edu
Nigel Boston	University of Wisconsin, USA boston@math.wisc.edu	Gaven J. Martin	Massey University, New Zealand g.j.martin@massey.ac.nz
Amarjit S. Budhiraja	U of North Carolina, Chapel Hill, USA budhiraj@email.unc.edu	Mary Meyer	Colorado State University, USA meyer@stat.colostate.edu
Pietro Cerone	La Trobe University, Australia P.Cerone@latrobe.edu.au	Emil Minchev	Ruse, Bulgaria eminchev@hotmail.com
Scott Chapman	Sam Houston State University, USA scott.chapman@shsu.edu	Frank Morgan	Williams College, USA frank.morgan@williams.edu
Joshua N. Cooper	University of South Carolina, USA cooper@math.sc.edu	Mohammad Sal Moselehian	Ferdowsi University of Mashhad, Iran moslehian@ferdowsi.um.ac.ir
Jem N. Corcoran	University of Colorado, USA corcoran@colorado.edu	Zuhair Nashed	University of Central Florida, USA znashed@mail.ucf.edu
Toka Diagana	Howard University, USA tdiagana@howard.edu	Ken Ono	Emory University, USA ono@mathcs.emory.edu
Michael Dorff	Brigham Young University, USA mdorff@math.byu.edu	Timothy E. O'Brien	Loyola University Chicago, USA tobrie1@luc.edu
Sever S. Dragomir	Victoria University, Australia sever@matilda.vu.edu.au	Joseph O'Rourke	Smith College, USA orourke@cs.smith.edu
Behrouz Emamizadeh	The Petroleum Institute, UAE bemamizadeh@pi.ac.ae	Yuval Peres	Microsoft Research, USA peres@microsoft.com
Joel Foisy	SUNY Potsdam foisyjs@potsgdam.edu	Y.-F. S. Pétermann	Université de Genève, Switzerland petermann@math.unige.ch
Errin W. Fulp	Wake Forest University, USA fulp@wfu.edu	Robert J. Plemmons	Wake Forest University, USA rplemmons@wfu.edu
Joseph Gallian	University of Minnesota Duluth, USA jgallian@d.umn.edu	Carl B. Pomerance	Dartmouth College, USA carl.pomerance@dartmouth.edu
Stephan R. Garcia	Pomona College, USA stephan.garcia@pomona.edu	Vadim Ponomarenko	San Diego State University, USA vadim@sciences.sdsu.edu
Anant Godbole	East Tennessee State University, USA godbole@etsu.edu	Bjorn Poonen	UC Berkeley, USA poonen@math.berkeley.edu
Ron Gould	Emory University, USA rg@mathcs.emory.edu	James Propp	U Mass Lowell, USA jpropp@cs.uml.edu
Andrew Granville	Université Montréal, Canada andrew@dms.umontreal.ca	József H. Przytycki	George Washington University, USA przytyck@gwu.edu
Jerrold Griggs	University of South Carolina, USA griggs@math.sc.edu	Richard Rebarber	University of Nebraska, USA rrebarbe@math.unl.edu
Sat Gupta	U of North Carolina, Greensboro, USA sngupta@uncg.edu	Robert W. Robinson	University of Georgia, USA rwr@cs.uga.edu
Jim Haglund	University of Pennsylvania, USA jhaglund@math.upenn.edu	Filip Saidak	U of North Carolina, Greensboro, USA f_saidak@uncg.edu
Johnny Henderson	Baylor University, USA johnny_henderson@baylor.edu	James A. Sellers	Penn State University, USA sellersj@math.psu.edu
Jim Hoste	Pitzer College jhoste@pitzer.edu	Andrew J. Sterge	Honorary Editor andy@ajsterge.com
Natalia Hritonenko	Prairie View A&M University, USA nahritonenko@pvamu.edu	Ann Trenk	Wellesley College, USA atrenk@wellesley.edu
Glenn H. Hurlbert	Arizona State University, USA hurlbert@asu.edu	Ravi Vakil	Stanford University, USA vakil@math.stanford.edu
Charles R. Johnson	College of William and Mary, USA crjohnso@math.wm.edu	Antonia Vecchio	Consiglio Nazionale delle Ricerche, Italy antonia.vecchio@cnrit
K. B. Kulasekera	Clemson University, USA kk@ces.clemson.edu	Ram U. Verma	University of Toledo, USA verma99@msn.com
Gerry Ladas	University of Rhode Island, USA gladas@math.uri.edu	John C. Wierman	Johns Hopkins University, USA wierman@jhu.edu
		Michael E. Zieve	University of Michigan, USA zieve@umich.edu

## PRODUCTION


Silvio Levy, Scientific Editor

See inside back cover or [msp.org/involve](http://msp.org/involve) for submission instructions. The subscription price for 2014 is US \$120/year for the electronic version, and \$165/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/>

© 2014 Mathematical Sciences Publishers

# Whitehead graphs and separability in rank two

Matt Clay, John Conant and Nivetha Ramasubramanian

(Communicated by Gaven Martin)

By applying an algorithm of Stallings regarding separability of elements in a free group, we give an alternative approach to that of Osborne and Zieschang in describing all primitive elements in the free group of rank 2. As a result, we give a proof of a classical result of Nielsen, used by Osborne and Zieschang in their work, that the only automorphisms of  $F_2$  that act trivially on the abelianization are those defined by conjugation. Finally, we compute the probability that a Whitehead graph in rank 2 contains a cut vertex. We show that this probability is approximately  $1/l^2$ , where  $l$  is the number of edges in the graph.

## 1. Introduction

The free group of rank  $n$ ,  $F_n$ , is the set of reduced words in a fixed alphabet  $\{x_1, x_1^{-1}, \dots, x_n, x_n^{-1}\}$  with group operation concatenation followed by free reduction. A word is *reduced* if it does not contain any of the two letter subwords  $x_i x_i^{-1}$ ,  $x_i^{-1} x_i$  for  $i = 1, \dots, n$ . *Free reduction* is the process of repeatedly removing such two-letter subwords. When the rank is small, we usually denote  $x_1 = a$ ,  $x_2 = b$ , et cetera. Free groups form an important class of groups due to their connections with low-dimensional topology and geometry and also as every group is the quotient of two free groups (though possibly of infinite rank).

A subset of  $F_n$  with  $n$  elements that generates  $F_n$  is called a *basis*. In other words, given a basis  $\{a_1, \dots, a_n\} \subset F_n$ , we can uniquely express any element  $g \in F_n$  as a (reduced) word in the alphabet  $\{a_1, a_1^{-1}, \dots, a_n, a_n^{-1}\}$ . We call such an expression the word representing  $g$  in the given basis.

Of particular interest are the elements that are part of some basis. Such elements are called *primitive*. Whitehead [1936] described an algorithm to determine whether or not a word in a given basis represents a primitive element.

---

*MSC2010:* primary 20E05; secondary 20F65.

*Keywords:* free groups, primitive elements.

This work is partially supported by NSF grant DMS-1006898 and by the Richard J. Cook and Teresa M. Lahti Student-Faculty Research Endowment.

Osborne and Zieschang [1981] gave a complete construction of primitive elements in rank 2. First they define a collection of primitive elements, indexed by an ordered pair of relatively prime integers. The relatively prime pair is the abelianization of the given element. Next, they quote a result of Nielsen [1917] (see also [Lyndon and Schupp 2001]) that up to conjugacy, primitive elements in  $F_2$  are uniquely determined by their abelianization and that their abelianization is a relatively prime pair of integers. Thus, the list of primitive elements described by Osborne and Zieschang contains exactly one representative from each conjugacy class of a primitive element.

There is an alternative viewpoint due to Cohen, Metzler and Zimmermann [Cohen et al. 1981]. Their idea is to use Whitehead's algorithm to give a narrow condition that the exponents of primitive elements in  $F_2$  need to satisfy. They do not give a complete characterization in the sense that there exist elements in  $F_2$  that are not primitive but that satisfy their condition.

Several other results about the form of primitive elements in rank 2 are known. See for instance [Kassel and Reutenauer 2007; Piggott 2006].

One purpose of this article is to show that Whitehead graphs can be used to recover Osborne and Zieschang's construction and in turn give an alternative proof of the above-quoted result of Nielsen used by Osborne and Zieschang. In fact, we consider a slightly more general notion than primitivity, called *separable* (definitions appear in Section 2). Stallings [1999] proved a version of Whitehead's algorithm for determining when a given word in a basis represents a separable element. We review this algorithm in Section 3 and include proofs of two propositions in [Stallings 1999] that are left as exercises for the reader. In Section 4, we show how to use this algorithm to determine all the primitive elements in rank 2.

The other purpose is to explore the nongenericity of the separable property for an element of  $F_2$ . Borovik, Myasnikov and Shpilrain [Borovik et al. 2002] prove that the likelihood that a word in  $F_n$  of length  $k$  is separable decays to 0 exponentially in  $k$ . Actually, their proof as stated is about primitive elements, but an examination of their proof shows that it applies to separable elements as well. We consider a property of Whitehead graphs that is shared by all separable elements and indeed is the backbone of Stallings' algorithm. This property is the existence of a cut vertex. We show in Section 5 that the likelihood that a Whitehead graph of an element in  $F_2$  with  $l$  edges has a cut vertex decays to 0 as  $1/l^2$ .

## 2. Preliminaries

### *Separability.*

**Definition 2.1.** An element  $g \in F_n$  is *separable* if there is a basis  $\{a_1, a_2, \dots, a_n\}$  for  $F_n$  such that the word representing  $g$  in this basis omits one of the  $a_i$ .

In [Stallings 1999], the notion of separability is defined for sets of elements in  $F_n$ . Our work in Section 4 can easily be adapted to this more general setting.

It is clear that the notion of separability is a conjugacy invariant. We recall that conjugacy classes of  $F_n$  can be identified with *reduced cyclic words*. These are reduced words considered as written on a circle and therefore there is no start or end to the word.

**Example 2.2.** Consider  $F_2$  with basis  $\{a, b\}$ . Clearly, the words  $a, b, a^2, a^{-1}$  and  $b^{-1}$  are separable. It is not obvious to recognize, but these words are separable:  $ab, ba$  and  $b^{-1}a$ . Indeed, using Whitehead automorphisms (Example 2.5) one can see that  $\{ab, b\}, \{ba, b\}$  and  $\{b^{-1}a, b\}$  are all bases for  $F_2$ . With respect to these respective bases, the elements are clearly separable.

To show that an element is not separable, we must show that no basis as in Definition 2.1 exists. As there are infinitely many bases for  $F_n$ , we must have an effective algorithm that can tell us when to stop looking for such a basis. This is what Stallings' algorithm (Section 3) does for us. Using this, we will show that  $ab^{-3}ab^{-1}$  and  $aba^{-1}b^{-1}$  are not separable. See Example 3.3.

**Remark 2.3.** In rank 2, there is a connection between separable elements and primitive elements. An element  $g \in F_n$  is *primitive* if there exists a basis  $\{a_1, a_2, \dots, a_n\}$  such that the word representing  $g$  in this basis is one of the  $a_i$  or its inverse. In rank 2, an element is separable if and only if it is a nontrivial power of a primitive element.

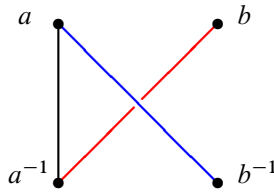
**Whitehead automorphisms.** Like for vector spaces in linear algebra, changing from one basis of  $F_n$  to another involves applying an automorphism of  $F_n$ . The *Whitehead automorphisms* are analogous to elementary matrices in linear algebra in the sense that every automorphism of  $F_n$  can be expressed as a product of Whitehead automorphisms [Whitehead 1936].

Given a basis  $\mathcal{A} = \{a_1, \dots, a_n\}$ , by  $\overline{\mathcal{A}}$  we denote the set  $\{a_1^{-1}, \dots, a_n^{-1}\}$ .

**Definition 2.4.** Let  $\mathcal{A}$  be a basis for  $F_n$  and decompose  $\mathcal{A} \cup \overline{\mathcal{A}} = Y \cup Z$  such that there is a  $v \in Y$  with  $v^{-1} \in Z$ . The *Whitehead automorphism*  $\phi = \phi_{(Y,Z,v)}$  is defined on  $x \in \mathcal{A} \cup \overline{\mathcal{A}}$ :

- (i) If  $x, x^{-1} \in Y$ , then  $\phi(x) = x$ .
- (ii) If  $x, x^{-1} \in Z$ , then  $\phi(x) = vxv^{-1}$ .
- (iii) If  $x = v$  or  $x = v^{-1}$ , then  $\phi(x) = x$ .
- (iv) If  $x \in Y$  and  $x^{-1} \in Z$ , then  $\phi(x) = vx$ .
- (v) If  $x^{-1} \in Y$  and  $x \in Z$ , then  $\phi(x) = xv^{-1}$ .

The map  $\phi$  is extended as a homomorphism to the rest of  $F_n$ .



**Figure 1.** The Whitehead graph for  $aba \in F_2$ .

**Example 2.5.** Consider the Whitehead automorphism  $\phi_{(Y,Z,v)}$  defined using the basis  $\{a, b\}$  of  $F_2$  where  $Y = \{a^{-1}, b^{-1}\}$ ,  $Z = \{a, b\}$  and  $v = b^{-1}$ . This automorphism sends the basis  $\{a, b\}$  to  $\{ab, b\}$ .

**Remark 2.6.** Let  $\{a_1, \dots, a_n\}$  be a basis for  $F_n$ . Suppose  $\phi$  is an automorphism of  $F_n$  and  $g \in F_n$  is such that the word representing  $\phi(g)$  omits one of the  $a_i$ , i.e.,  $\phi(g)$  is separable. Then by considering the basis  $\{\phi^{-1}(a_1), \dots, \phi^{-1}(a_n)\}$  we can witness that  $g$  is separable as well. In other words, if we can find some automorphism that removes all the occurrences of one of the basis elements from  $g$ , then  $g$  is separable. See Example 3.2.

**Whitehead graphs.** The key tool for detecting separability is the *Whitehead graph*.

**Definition 2.7.** Let  $\mathcal{A}$  be a basis for the free group  $F_n$ . Given an element  $g \in F_n$  whose conjugacy class is represented by the cyclic word  $w$  in the basis  $\mathcal{A}$ , we define the *Whitehead graph* of  $g$ , denoted  $\text{Wh}_{\mathcal{A}}(g)$ , by

(vertices)  $\mathcal{A} \cup \overline{\mathcal{A}}$ ,

(edges) between  $u, v \in \mathcal{A} \cup \overline{\mathcal{A}}$  for each instance of  $uv^{-1}$  as a subword of  $w$ .

**Example 2.8.** Consider the word  $aba \in F_2$ . The vertices for  $\text{Wh}_{\{a,b\}}(aba)$  are denoted  $a, a^{-1}, b, b^{-1}$ . The edges are determined as follows:

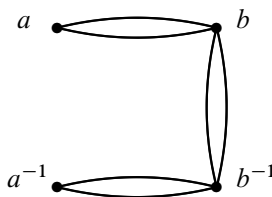
- First edge:** the subword  $ab$  gives an edge from  $a$  to  $b^{-1}$ .
- Second edge:** the subword  $ba$  gives an edge from  $b$  to  $a^{-1}$ .
- Third edge:** the subword  $aa$  gives an edge from  $a$  to  $a^{-1}$ .

This Whitehead graph is shown in Figure 1.

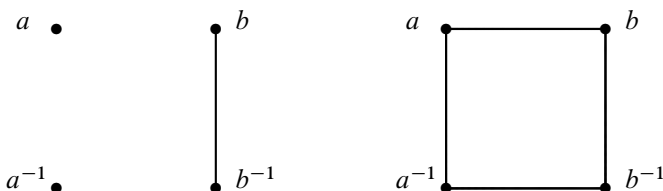
**Remark 2.9.** An important property of Whitehead graphs to note is that the valence of a vertex  $v$  is the same as the valence of the vertex  $v^{-1}$ . This observation plays a key role in Sections 4 and 5.

**Example 2.10.** The Whitehead graph  $\text{Wh}_{\{a,b\}}(ab^{-3}ab^{-1})$  is shown in Figure 2.

The following definitions, applied to Whitehead graphs, will be used in Section 3 to determine whether a word is separable.



**Figure 2.** The Whitehead graph for  $ab^{-3}ab^{-1} \in F_2$ .



**Figure 3.** Left: the Whitehead graph of  $b$  is disconnected. Right: the Whitehead graph of  $ab^{-1}a^{-1}b$  is connected and does not have a cut vertex.

**Definition 2.11.** A graph is connected if there is an edge path from any vertex to any other vertex in the graph.

The *trivial graph* is the graph with a single vertex and no edges.

**Definition 2.12.** A cut vertex  $v$  of a graph  $\Gamma$  is a vertex such that the graph decomposes into two nontrivial graphs  $\Gamma_1$  and  $\Gamma_2$  which intersect only at  $v$ . In other words, any edge path from a vertex of  $\Gamma_1$  to a vertex in  $\Gamma_2$  must go through  $v$ .

We remark that a disconnected Whitehead graph always has a cut vertex.

Figures 1 and 2 show Whitehead graphs that are connected and have a cut vertex. Figure 3 shows examples of Whitehead graphs that are respectively disconnected and connected without a cut vertex.

**Remark 2.13.** In terms of the Whitehead graph, an element  $g \in F_n$  is separable if there is a basis  $\mathcal{A}$  such that  $\text{Wh}_{\mathcal{A}}(g)$  has an isolated vertex. The isolated vertex exactly corresponds to the omitted basis element.

### 3. Stallings' algorithm

There is an algorithm due to Stallings [1999] that determines whether or not a word is separable. A flowchart for the algorithm is depicted in Figure 5. We will describe the algorithm in more detail, work out a couple of examples and provide proofs to a couple of the steps that are omitted in [Stallings 1999].

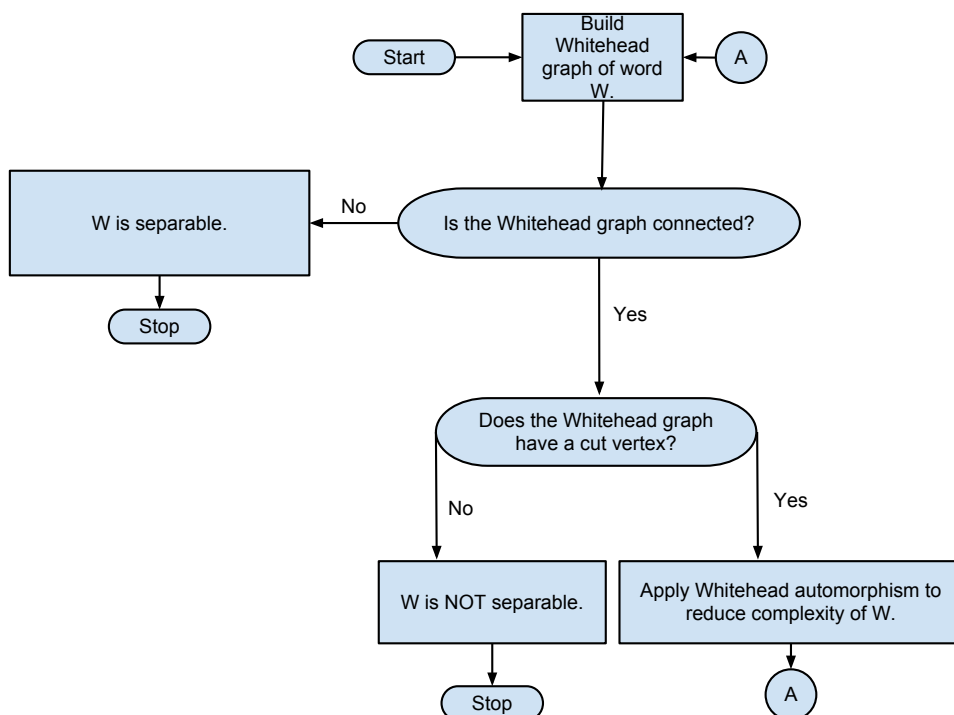
The important theorem needed to use the algorithm is the following.

**Theorem 3.1** [Stallings 1999, Theorem 2.4]. *If  $g \in F_n$  is separable, then the Whitehead graph of  $g$  in any basis contains a cut vertex.*

Using the contrapositive of this theorem, we can see that  $ab^{-1}a^{-1}b$  is not a separable element of  $F_2$ , as its Whitehead graph in Figure 3, right, does not have a cut vertex. In general, an element that is not separable may have a Whitehead graph with respect to some basis that does have a cut vertex. To determine that the element is not separable, we need to find a basis in which its Whitehead graph does not have cut a vertex.

**Stallings' algorithm.** To determine whether a reduced cyclic word  $w$  in some basis  $\mathcal{A}$  is separable or not, we start by constructing the Whitehead graph of  $w$  and determine if the graph is connected. If the graph is not connected, then Proposition 3.5 shows that after possibly applying a single Whitehead automorphism, the new Whitehead graph has an isolated vertex and hence  $w$  is separable (Remark 2.13).

If the graph is connected, then we determine if the graph has a cut vertex. If not, then by Theorem 3.1,  $w$  is not separable. If it does have a cut vertex, then by Proposition 3.6 there is a Whitehead automorphism  $\phi$  such that the complexity of



**Figure 5.** Flowchart for Stallings' algorithm.



$\phi(w)$  (that is, the length of the cyclic word representing it) is strictly less than the complexity of  $w$ . We now repeat the algorithm using the word  $w'$ .

Now in order for the algorithm to work, we need to know that it will terminate. That is precisely what Proposition 3.6 assures us. Since the complexity will be reduced, we know that eventually either the Whitehead graph will either be disconnected, or it will be connected without a cut vertex.

We now present an example of both a separable word and nonseparable word.

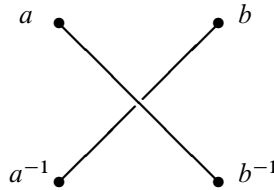
**Example 3.2.** The Whitehead graph of  $aba$  is shown in Figure 1. This graph has a cut vertex at  $a^{-1}$ . (The vertex  $a$  is also a cut vertex.) According to Proposition 3.6, we should apply the Whitehead automorphism with  $Y = \{a^{-1}, b\}$ ,  $Z = \{a, b^{-1}\}$ ,  $v = a$  to reduce the complexity. The automorphism is given by

$$a \mapsto a, \quad b \mapsto a^{-1}b. \tag{1}$$

Applying the automorphism to  $aba$ , we get

$$aba \mapsto a(a^{-1}b)a = ba.$$

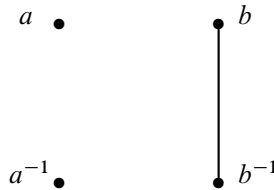
The graph of this new word  $ba$  is



This graph is disconnected, and thus by Proposition 3.5 we know that  $ba$  and hence  $aba$  is separable. We can apply the Whitehead automorphism using  $Y = \{a, b^{-1}\}$ ,  $Z = \{a^{-1}, b\}$ ,  $v = a$  to see this explicitly. This is the automorphism:

$$a \mapsto a, \quad b \mapsto ba^{-1}. \tag{2}$$

Applying this automorphism, we have  $ba \mapsto (ba^{-1})a = b$ . The Whitehead graph of  $b$  looks like



So  $aba$  is separable, as there is an isolated vertex in this graph. By working backwards, applying the inverse automorphism of (2) and then the inverse automorphism

of (1) to  $\{a, b\}$ , we can find a basis in which  $aba$  omits an element. The inverse to (2) is

$$a \mapsto a, \quad b \mapsto ba. \tag{3}$$

Applying this automorphism followed by the inverse to (1), given by

$$a \mapsto a, \quad b \mapsto ab, \tag{4}$$

we get  $a \mapsto a \mapsto a$  and  $b \mapsto ba \mapsto aba$ . It is clear, in terms of the basis  $\{a, aba\}$ , that  $aba$  is separable.

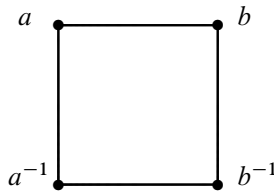
**Example 3.3.** Applying the algorithm to  $ab^{-3}ab^{-1}$ , we can show that this word is not separable. The Whitehead graph for this word is shown in Figure 2. Both  $b$  and  $b^{-1}$  are cut vertices; we choose  $b^{-1}$  to define our Whitehead automorphism. According to Proposition 3.6, we use the automorphism defined by the data  $Y = \{a^{-1}, b^{-1}\}$ ,  $Z = \{a, b\}$ ,  $v = b^{-1}$ . This is the automorphism:

$$a \mapsto ab, \quad b \mapsto b. \tag{5}$$

Applying this automorphism, we get

$$ab^{-3}ab^{-1} \mapsto (ab)b^{-3}(ab)b^{-1} = ab^{-2}a.$$

The Whitehead graph of  $ab^{-2}a$  is this:



This graph does not have a cut vertex, so  $ab^{-3}ab^{-1}$  is not separable.

Stallings provides examples to convince the reader of the validity of the steps:

- (i) disconnected  $\implies$  separable [Stallings 1999, Proposition 2.2];
- (ii) cut vertex  $\implies$  reduce complexity [Stallings 1999, Proposition 2.3].

However, he does not provide proofs. We will give proofs of these steps here. First, we prove a lemma that makes the arguments easier. The lemma shows that when the Whitehead graph has cut vertex  $v$ , subwords without  $v^{\pm 1}$  behave like single elements.

**Lemma 3.4.** *Suppose  $\mathcal{A}$  is a basis for  $F_n$ , let  $Y, Z$  be subsets of  $\mathcal{A} \cup \overline{\mathcal{A}}$  and let  $v \in \mathcal{A} \cup \overline{\mathcal{A}}$  define a Whitehead automorphism  $\phi = \phi_{(Y,Z,v)}$ . Suppose  $w = w_1w_2 \cdots w_k$  is a word over the basis  $\mathcal{A}$  such that  $w_i \neq v^{\pm 1}$  for all  $i = 1, \dots, k$ . Further suppose that either  $w_i, w_{i+1}^{-1} \in Y$  or  $w_i, w_{i+1}^{-1} \in Z$  for each  $i = 1, \dots, k-1$ .*

- (i) If  $w_1^{-1}, w_k \in Y$ , then  $\phi(w) = w$ .
- (ii) If  $w_1^{-1}, w_k \in Z$ , then  $\phi(w) = vwv^{-1}$ .
- (iii) If  $w_k \in Y$  and  $w_1^{-1} \in Z$ , then  $\phi(w) = vw$ .
- (iv) If  $w_1^{-1} \in Y, w_k \in Z$ , then  $\phi(w) = wv^{-1}$ .

*Proof.* We will prove this by induction on  $k$ . If  $k = 1$ , this is just the definition of the Whitehead automorphism  $\phi_{(Y,Z,v)}$  applied to  $w = w_1$ .

Now given  $w = w_1 \cdots w_{k-1} w_k$ , we have  $\phi(w_1 \cdots w_{k-1}) = v^{\epsilon_1} w_1 \cdots w_{k-1} v^{-\epsilon_2}$  by induction, where  $\epsilon_1, \epsilon_2$  are either 0 or 1 depending if  $w_1^{-1}$  and  $w_{k-1}$  are in  $Y$  or  $Z$ , respectively. Since  $w_k^{-1}$  is in  $Z$  if and only if  $w_{k-1}$  is in  $Z$ , we have  $\phi(w_k) = v^{\epsilon_2} w_k v^{-\epsilon_3}$  for some  $\epsilon_3$  equal to either 0 or 1 depending if  $w_k$  is in  $Y$  or  $Z$ . Hence

$$\begin{aligned} \phi(w) &= \phi(w_1 \cdots w_{k-1})\phi(w_k) \\ &= v^{\epsilon_1} w_1 \cdots w_{k-1} v^{-\epsilon_2} \cdot v^{\epsilon_2} w_k v^{-\epsilon_3} \\ &= v^{\epsilon_1} w v^{-\epsilon_3}. \end{aligned}$$

This proves the lemma. □

**Proposition 3.5** [Stallings 1999, Proposition 2.2]. *Suppose  $\mathcal{A}$  is a basis for  $F_n$  and  $w$  is a word in this basis such that the Whitehead graph  $\text{Wh}_{\mathcal{A}}(w)$  does not have an isolated vertex and is not connected. Then  $w$  is separable. Specifically, separate the vertices of  $\text{Wh}_{\mathcal{A}}(w)$  into two subsets  $Y$  and  $Z$  such that there is no edge from a vertex in  $Y$  to a vertex in  $Z$ . Then there is a vertex  $v \in Y$  such that  $v^{-1} \in Z$  and the Whitehead graph of  $\phi_{(Y,Z,v)}(w)$  has  $v$  an isolated vertex.*

*Proof.* If for all  $v \in \mathcal{A}$  there is an edge between  $v$  and  $v^{-1}$  in  $\text{Wh}_{\mathcal{A}}(w)$ , then we claim that the graph is connected. Indeed, let  $\Gamma$  be the graph obtained by collapsing all the edges between  $v$  and  $v^{-1}$  for each  $v \in \mathcal{A}$ , and denote the image vertices by the element of the basis. Then  $\Gamma$  has the same number of connected components as  $\text{Wh}_{\mathcal{A}}(w)$ . But now reading off the elements of the basis  $\mathcal{A}$  in the order in which they appear in  $w$  traces out a path in  $\Gamma$ . As there are no isolated vertices, every element in the basis appears along the path. Thus  $\Gamma$  and hence  $\text{Wh}_{\mathcal{A}}(w)$  is connected.

Hence, we have some vertex  $v$  as in the statement. By conjugating  $w$ , we can write  $w = w_1 v^{\epsilon_1} w_2 v^{\epsilon_2} \cdots w_k v^{\epsilon_k}$ , where  $\epsilon_i \in \{-1, 1\}$  and  $v$  and  $v^{-1}$  do not appear in any of the  $w_i$ 's. Indeed, as there is no edge between  $v$  and  $v^{-1}$ ,  $v$  can only appear in  $w$  to the power 1 or  $-1$ . Notice, the  $w_i$ 's satisfy the hypotheses of Lemma 3.4 using  $\phi = \phi_{(Y,Z,v)}$ .

Let  $X$  represent either  $Y$  or  $Z$ . We will write  $w_i \in X$  to mean that when writing  $w_i = u_1 u_2 \cdots u_k$  as a word in the basis  $\mathcal{A}$ , we have  $u_k \in X$ . Similarly,  $w_i^{-1} \in X$  means that  $u_1^{-1} \in X$ . By Lemma 3.4, this is sufficient to specify the image of  $w_i$  under  $\phi$ .

Suppose  $i = 1, \dots, k-1$ . If  $\epsilon_i = 1$ , then  $w_i \in Z$  and  $w_{i+1}^{-1} \in Y$ , hence

$$\phi(w_i v w_{i+1}) = (v^{\kappa_1} w_i v^{-1}) v (w_{i+1} v^{-\kappa_2}) = v^{\kappa_1} w_i w_{i+1} v^{-\kappa_2},$$

where  $\kappa_1, \kappa_2 \in \{0, 1\}$ . Likewise, if  $\epsilon_i = -1$ , then  $w_i \in Y$  and  $w_{i+1}^{-1} \in Z$ , hence

$$\phi(w_i v^{-1} w_{i+1}) = (v^{\kappa_1} w_i) v^{-1} (v w_{i+1} v^{-\kappa_2}) = v^{\kappa_1} w_i w_{i+1} v^{-\kappa_2},$$

where again  $\kappa_1, \kappa_2 \in \{0, 1\}$ .

These equations hold true for  $i = k$  interpreting  $w_{k+1}$  as  $w_1$ . Therefore, the cyclic word representing  $\phi(w)$  is  $w_1 \cdots w_k$ .  $\square$

**Proposition 3.6** [Stallings 1999, Proposition 2.3]. *Suppose  $\mathcal{A}$  is a basis for  $F_n$  and  $w$  is a word in this basis such that the Whitehead graph  $\text{Wh}_{\mathcal{A}}(w)$  is connected and that  $v$  is a cut vertex decomposing  $\text{Wh}_{\mathcal{A}}(w)$  into two nontrivial subgraphs  $\Gamma_1$  and  $\Gamma_2$ , which only intersect at  $v$ . Suppose that  $\Gamma_2$  contains the vertex  $v^{-1}$ . Let  $Y$  be the set of vertices of  $\Gamma_1$ , and  $Z$  the set of vertices of  $\Gamma_2$  with the vertex  $v$  removed. Then the complexity of  $\phi_{(Y,Z,v)}(w)$  is strictly less than the complexity of  $w$ .*

*Proof.* We can conjugate  $w$  to have form  $w = w_1 v^{n_1} \cdots w_k v^{n_k}$ , where  $n_i \neq 0$  for all  $i$  and  $v^{\pm 1}$  does not appear in any of the  $w_i$ 's. As in Proposition 3.5, the  $w_i$ 's satisfy the hypotheses of Lemma 3.4 using  $\phi = \phi_{(Y,Z,v)}$ . We continue to use the convention  $w_i \in Y$ , et cetera, from the proof of Proposition 3.5.

Suppose  $i = 1, \dots, k-1$ . If  $n_i > 0$ , then  $w_i \in Z$ . If  $w_{i+1}^{-1} \in Y$ , then

$$\phi(w_i v^{n_i} w_{i+1}) = (v^{\kappa_1} w_i v^{-1}) v^{n_i} (w_{i+1} v^{-\kappa_2}) = v^{\kappa_1} w_i v^{n_i-1} w_{i+1} v^{-\kappa_2},$$

where  $\kappa_1, \kappa_2 \in \{0, 1\}$ . Otherwise,  $w_{i+1}^{-1} \in Z$  and then

$$\phi(w_i v^{n_i} w_{i+1}) = (v^{\kappa_1} w_i v^{-1}) v^{n_i} (v w_{i+1} v^{-\kappa_2}) = v^{\kappa_1} w_i v^{n_i} w_{i+1} v^{-\kappa_2},$$

where  $\kappa_1, \kappa_2 \in \{0, 1\}$ .

Likewise, if  $n_i < 0$ , then  $w_{i+1}^{-1} \in Z$ . If  $w_i \in Y$ , then

$$\phi(w_i v^{n_i} w_{i+1}) = (v^{\kappa_1} w_i) v^{n_i} (v w_{i+1} v^{-\kappa_2}) = v^{\kappa_1} w_i v^{n_i+1} w_{i+1} v^{-\kappa_2},$$

where  $\kappa_1, \kappa_2 \in \{0, 1\}$ . Otherwise,  $w_i \in Z$  and then

$$\phi(w_i v^{n_i} w_{i+1}) = (v^{\kappa_1} w_i v^{-1}) v^{n_i} (v w_{i+1} v^{-\kappa_2}) = v^{\kappa_1} w_i v^{n_i} w_{i+1} v^{-\kappa_2},$$

where again  $\kappa_1, \kappa_2 \in \{0, 1\}$ .

Like in Proposition 3.5, for  $i = k$  these equations hold interpreting  $w_{k+1} = w_1$ . Thus, we see that the length of the cyclic word representing  $\phi(w)$  is reduced every time either  $w_i \in Y$  or  $w_{i+1}^{-1} \in Y$ . This is the number of edges adjacent to  $v$  that are in  $\Gamma_1$ .  $\square$

Using Stallings' algorithm, we can compute the length of the shortest word in any basis that is not separable.

**Theorem 3.7.** *Let  $g \in F_n$  be an element that is not separable. Then with respect to any basis of  $F_n$ , the length of the word representing  $g$  is at least  $2n$ . Furthermore, there is a word of length  $2n$  that represents an element that is not separable.*

*Proof.* Let  $w$  be a word in some basis of  $F_n$  with length at most  $2n - 2$ . Let  $\Gamma$  be the Whitehead graph of  $w$ . Then  $\Gamma$  will have  $2n$  vertices. Before we add any edges to  $\Gamma$ , we can count each vertex as a connected component. So the initial number of connected components is  $2n$ , and as long as the number of components is greater than 1, we know that  $\Gamma$  is disconnected. Each edge added to  $\Gamma$  will be adjacent with two vertices which are either previously connected or disconnected. If the former occurs, then the number of components does not change. If the latter occurs, then the number of components is reduced by 1. Since  $w$  has at most  $2n - 2$  edges, the fewest number of components of  $\Gamma$  is  $2n - (2n - 2) = 2$ . So we know that the Whitehead graph is disconnected for all words of length at most  $2n - 2$ , and hence by Proposition 3.5, every word of length at most  $2n - 2$  represents a separable element.

Now suppose the length of  $w$  is  $2n - 1$ . After adding  $2n - 2$  edges, the Whitehead graph  $\Gamma$  will be disconnected. Then when we add the last edge,  $\Gamma$  will either become connected or remain disconnected. If  $\Gamma$  becomes connected, we know that at least one of the vertices adjacent to the last edge added will be a cut vertex. Then by Proposition 3.6 we can reduce the complexity of  $w$ . Since all shorter words will have a disconnected Whitehead graph by the above paragraph, we know that  $w$  represents a separable element.

This proves the first statement of the theorem. Now we will construct a word of length  $2n$  that represents an element that is not separable.

Fix a basis  $\mathcal{A} = \{a_1, \dots, a_n\}$  and define a word  $w$  in this basis by

$$w = a_1^{-1} a_2 \dots a_n^{(-1)^n} a_n^{(-1)^n} \dots a_2 a_1^{-1}.$$

We claim that the Whitehead graph is a circuit that contains every vertex. Let  $1 \leq i < n$ . Then  $w$  will contain either  $a_i a_{i+1}^{-1}$  and  $a_{i+1}^{-1} a_i$  or  $a_i^{-1} a_{i+1}$  and  $a_{i+1} a_i^{-1}$  depending on if  $i$  is even or odd. In both cases the Whitehead graph will have edges between  $a_i$  and  $a_{i+1}$  and between  $a_i^{-1}$  and  $a_{i+1}^{-1}$ . Then since  $a_1^{-1}$  is on either side of  $w$  we will have an edge from  $a_1$  to  $a_1^{-1}$ . Additionally, the  $a_n^{\pm 2}$  in the center will add an edge from  $a_n$  to  $a_n^{-1}$ . This creates a circular graph which is connected without any cut vertices. So by Theorem 3.1,  $w$  represents an element that is not separable. □

In contrast with the fact that the likelihood of an element being separable decays to 0 as the word length increases [Borovik et al. 2002], the likelihood that a word

of length  $2n$  in  $F_n$  is not separable decays to 0 as  $n \rightarrow \infty$ . Let  $\Sigma(l, n)$  denote the words of  $F_n$  of length  $l$  and  $N(l, n)$  the subset that represent elements that are not separable.

**Theorem 3.8.** 
$$\lim_{n \rightarrow \infty} \frac{\#|N(2n, n)|}{\#|\Sigma(2n, n)|} = 0.$$

*Proof.* As we saw in the proof of Theorem 3.7, if a word  $w$  of length  $2n$  is not separable, then its Whitehead graph is a circuit that contains every vertex. Hence for each element  $a_i$  of the basis, two elements (possibly the same) from  $\{a_i, a_i^{-1}\}$  appear in  $w$ . This gives  $2^{2n}$  choices. Multiplying this by the number of ways to order the  $2n$  elements, we see that

$$\#|N(2n, n)| \leq 2^{2n} (2n)!.$$

It is well known that the number of words of length  $l$  in rank  $n$  is

$$\#|\Sigma(l, n)| = 2n(2n - 1)^{l-1}.$$

Therefore

$$\frac{\#|N(2n, n)|}{\#|\Sigma(2n, n)|} \leq \frac{2^{2n} (2n)!}{2n(2n - 1)^{2n-1}} \leq \frac{2^{2n} (2n)!}{(2n - 1)^{2n}}.$$

We will prove the theorem by showing this last ratio converges to 0.

Let us consider the series

$$\sum \frac{2^{2n} (2n)!}{(2n - 1)^{2n}}.$$

We now show that this series converges. By applying the ratio test, we get

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{\frac{2^{2n+2} (2n+2)!}{(2n+1)^{2n+2}}}{\frac{2^{2n} (2n)!}{(2n-1)^{2n}}} &= \lim_{n \rightarrow \infty} \frac{2^{2n+2} (2n+2)! (2n-1)^{2n}}{(2n+1)^{2n+2} 2^{2n} (2n)!} \\ &= \lim_{n \rightarrow \infty} \frac{2^2 (2n+2)(2n+1) (2n-1)^{2n}}{(2n+1)(2n+1) (2n+1)^{2n}} \\ &= 4 \lim_{n \rightarrow \infty} \frac{(2n-1)^{2n}}{(2n+1)^{2n}}. \end{aligned}$$

Upon substitution of  $x = 2n$ , this becomes

$$4 \lim_{x \rightarrow \infty} \frac{(x-1)^x}{(x+1)^x} = 4 \lim_{x \rightarrow \infty} \exp\left(\ln\left(\frac{x-1}{x+1}\right)^x\right) = 4e^{\lim_{x \rightarrow \infty} x(\ln(x-1) - \ln(x+1))}.$$

Now we apply l'Hospital's rule to the exponent:

$$\lim_{x \rightarrow \infty} x(\ln(x-1) - \ln(x+1)) = \lim_{x \rightarrow \infty} \frac{\frac{1}{x-1} - \frac{1}{x+1}}{\frac{-1}{x^2}} = \lim_{x \rightarrow \infty} \frac{-2x^2}{x^2 - 1} = -2.$$

Hence the limit of the ratio of successive terms is  $4e^{-2} < 1$ . So by the ratio test, the series  $\sum 2^{2n}(2n)!/(2n-1)^{2n}$  converges.  $\square$

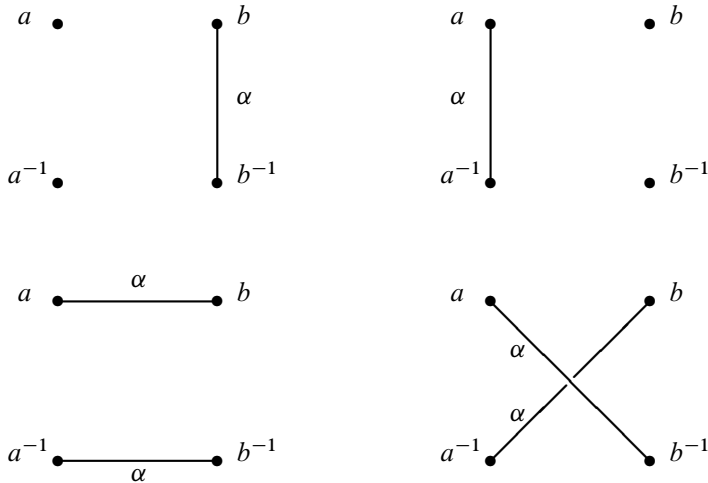
### 4. Separability in $F_2$

By Theorem 3.1, if an element is separable, then with respect to any basis its Whitehead graph has a cut vertex. In rank 2, this means that the Whitehead graph has one of the eight forms depicted in Figures 6 and 7. The labels  $\alpha, \beta$  represent the multiplicity of an edge. Notice that we used that in a Whitehead graph the vertices  $v$  and  $v^{-1}$  have the same valence. This rules out, for instance, the  $\sqcup$ -shaped graph with edges only between  $a$  and  $a^{-1}$ ,  $a^{-1}$  and  $b^{-1}$ , and  $b$  and  $b^{-1}$ . The labels on the graphs also reflect this observation.

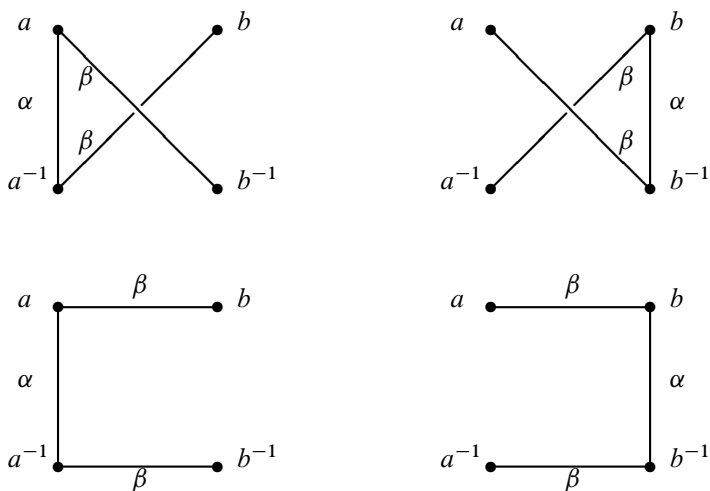
We make the following simple observations. These observations appear in [Cohen et al. 1981] as well.

**Lemma 4.1.** *Suppose  $g \in F_2$  is separable. Let  $w$  be the cyclic word representing the conjugacy class of  $g$ .*

- (i) *If  $a^k$  appears as a subword of  $w$ , where  $|k| > 1$ , then for every nontrivial subword of the form  $b^m$ , we have  $m = \pm 1$ . Similarly, if  $b^m$  appears as a subword of  $w$ , where  $|m| > 1$ , then for every nontrivial subword of the form  $a^k$ , we have  $k = \pm 1$ .*



**Figure 6.** Disconnected Whitehead graphs in rank 2.



**Figure 7.** Connected Whitehead graphs with a cut vertex in rank 2.

- (ii) If  $a^{k_1}$  and  $a^{k_2}$  are nontrivial subwords of  $w$ , then  $k_1 k_2 > 0$ . Similarly, if  $b^{m_1}$  and  $b^{m_2}$  are nontrivial subwords of  $w$ , then  $m_1 m_2 > 0$ .

*Proof.* Item (i) is clear, as in all the Whitehead graphs in Figures 6 and 7 there never appear edges both between  $a$  and  $a^{-1}$  and between  $b$  and  $b^{-1}$ . Thus either  $a$  or  $b$  can appear to a power other than  $\pm 1$ , but not both.

Item (ii) is also clear if the Whitehead graph for  $g$  is as in Figure 6, since in this case  $w$  is either  $a^{\pm\alpha}$ ,  $b^{\pm\alpha}$ ,  $(ab^{-1})^{\pm\alpha}$  or  $(ab)^{\pm\alpha}$ .

Suppose the Whitehead graph for  $g$  is the one depicted in the top left corner of Figure 7. Suppose both  $b$  and  $b^{-1}$  appeared as subwords of  $w$ . Then we have a subword of the form  $ba^k b^{-1}$ , where  $k \neq 0$ . The shape of the Whitehead graph applied to the initial  $ba^k$  forces  $k > 0$ , whereas applied to the latter  $a^k b^{-1}$  forces  $k < 0$ . This is a contradiction. A similar argument works if there is a subword of the form  $ab^{\pm 1} a^{-1}$ .

The other three Whitehead graphs are dealt with similarly by permuting  $a \leftrightarrow b$  and/or  $a \leftrightarrow a^{-1}$ . □

Let  $S^{+,+}(l, \alpha, \beta)$  be the set of cyclic words of length  $l$  that are separable, where any power of  $a$  or  $b$  that appears is positive and where  $\alpha$  and  $\beta$  are the amount of  $a$ 's and  $b$ 's, respectively. We allow for the possibility that  $\alpha$  or  $\beta$  is negative, in which case  $S^{+,+}(l, \alpha, \beta) = \emptyset$ . Notice that  $l = \alpha + \beta$ .

Likewise define  $S^{-,+}(l, \alpha, \beta)$  as the set of cyclic words of length  $l$  that are separable and only use  $a^{-1}$  and  $b$ . Define  $S^{+,-}(l, \alpha, \beta)$  and  $S^{-,-}(l, \alpha, \beta)$  in a similar fashion. By Lemma 4.1, we have that every cyclic word that is separable is contained in one of these four sets. By  $S$  we denote one of  $S^{+,+}$ ,  $S^{-,+}$ ,  $S^{+,-}$  or  $S^{-,-}$ .



Our goal is to show that there is exactly one element in  $S(l, \alpha, \beta)$  (Theorem 4.3). We will use an inductive argument based on the following proposition.

**Proposition 4.2.** *Suppose  $\alpha, \beta \geq 0$ . Then*

$$\#|S(l, \alpha, \beta)| = \max\{\#|S(l - \alpha, \alpha, \beta - \alpha)|, \#|S(l - \beta, \alpha - \beta, \beta)|\}.$$

*Proof.* To simplify the argument, assume  $S = S^{+,+}$ . The other three cases are similar. Since  $\#|S(l, \alpha, \beta)| = \#|S(l, \beta, \alpha)|$ , without loss of generality we can assume  $\alpha \geq \beta$ .

If  $\beta = 0$ , then  $S(l, \alpha, \beta) = S(l - \beta, \alpha - \beta, \beta)$  and  $S(l - \alpha, \alpha, \beta - \alpha) = \emptyset$  and so the proposition holds. Notice that  $S(l, l, 0) = \{a^l\}$ .

Now we assume that  $\alpha = \beta > 0$ . The Whitehead graph of any  $x \in S(l, \alpha, \beta)$  is the bottom right graph of Figure 6 (recall we are assuming that  $S = S^{+,+}$ ). Thus we must have that  $x$  can be represented by the cyclic word  $(ab)^\alpha$ , and hence  $\#|S(l, \alpha, \beta)| = 1$ . As

$$S(l - \alpha, \alpha, \beta - \alpha) = S(\alpha, \alpha, 0) \quad \text{and} \quad S(l - \beta, \alpha - \beta, \beta) = S(\beta, 0, \beta),$$

the proposition holds.

We are left with the case that  $\alpha > \beta > 0$ . Therefore, by Lemma 4.1, each  $x \in S(l, \alpha, \beta)$  is represented by a cyclic word of the form

$$a^{\alpha_1} b a^{\alpha_2} b \dots a^{\alpha_\beta} b,$$

where  $\alpha_1 + \alpha_2 + \dots + \alpha_\beta = \alpha$  and each  $\alpha_i > 0$ . We apply Proposition 3.6 in this case using  $Y = \{a^{-1}, b\}$ ,  $Z = \{a, b^{-1}\}$  and  $v = a^{-1}$ . This gives the Whitehead automorphism  $\phi$  of  $F_2$  defined by  $\phi(a) = a$  and  $\phi(b) = a^{-1}b$ . When we apply  $\phi$  to a word we will reduce its length and number of  $a$ 's by  $\beta$ . So for each  $x \in S(l, \alpha, \beta)$ , we have  $\phi(x) \in S(l - \beta, \alpha - \beta, \beta)$ . Therefore

$$\#|S(l, \alpha, \beta)| \leq \#|S(l - \beta, \alpha - \beta, \beta)|.$$

To see the opposite inequality, we consider the automorphism  $\phi^{-1}$ . This is the map  $\phi^{-1}(a) = a$  and  $\phi^{-1}(b) = ab$ . Then applying  $\phi^{-1}$  to an element

$$x \in S(l - \beta, \alpha - \beta, \beta)$$

will increase the number of  $a$ 's and the length of  $x$  by  $\beta$  (recall we are assume that  $S = S^{+,+}$ ). So for each  $x \in S(l - \beta, \alpha - \beta, \beta)$ , we have  $\phi^{-1}(x) \in S(l, \alpha, \beta)$ . Thus

$$\#|S(l - \beta, \alpha - \beta, \beta)| \leq \#|S(l, \alpha, \beta)|,$$

and therefore

$$\#|S(l - \beta, \alpha - \beta, \beta)| = \#|S(l, \alpha, \beta)|.$$

Notice that  $\#|S(l - \alpha, \alpha, \beta - \alpha)| = 0$  as  $\beta - \alpha < 0$ . Thus

$$\#|S(l, \alpha, \beta)| = \max\{\#|S(l - \alpha, \alpha, \beta - \alpha)|, \#|S(l - \beta, \alpha - \beta, \beta)|\}. \quad \square$$

**Theorem 4.3.** *Suppose  $\alpha, \beta \geq 0$ . Then*

$$\#|S(l, \alpha, \beta)| = 1.$$

*Proof.* As in Proposition 4.2, we assume that  $S = S^{+,+}$ .

Recall from the proof of Proposition 4.2 that

$$S(\alpha, \alpha, 0) = \{a^\alpha\} \quad \text{and} \quad S(\alpha, 0, \alpha) = \{b^\alpha\},$$

for all  $\alpha > 0$ . Hence, the Theorem holds for these special cases.

If  $\alpha \geq \beta > 0$ , then by Proposition 4.2,

$$\#|S(l, \alpha, \beta)| = \#|S(l - \beta, \alpha - \beta, \beta)|.$$

Likewise, if  $\beta \geq \alpha > 0$ , then by Proposition 4.2,

$$\#|S(l, \alpha, \beta)| = \#|S(l - \alpha, \alpha, \beta - \alpha)|.$$

Applying these repeatedly and using the Euclidean algorithm, we see

$$\#|S(l, \alpha, \beta)| = \#|S(d, d, 0)| = 1,$$

where  $d = \gcd(\alpha, \beta)$ . □

Theorem 4.3 allows us to give an alternative proof to a classical result of Nielsen [1917]. First, we offer a corollary from which we will deduce Nielsen’s result. Let  $A: F_2 \rightarrow \mathbb{Z}^2$  denote the abelianization map. Given a word  $w$  in the basis  $\{a, b\}$ , this is the map

$$A(w) = \begin{bmatrix} \exp_a(w) \\ \exp_b(w) \end{bmatrix},$$

where  $\exp_a(w)$  is the exponent sum of  $a$  in  $w$ , i.e., the number of  $a$ ’s that appear minus the number of  $a^{-1}$ ’s. The function  $\exp_b(w)$  is defined similarly.

**Corollary 4.4.** *Let  $g, h \in F_2$  be separable. Then  $A(g) = A(h)$  if and only if  $g$  and  $h$  are conjugate. Moreover, every nonzero element in  $\mathbb{Z}^2$  is the image of some separable element and a separable element  $g \in F_2$  is primitive if and only if the greatest common divisor of the components of  $A(g)$  is 1.*

*Proof.* If  $g$  and  $h$  are separable, then the cyclic words representing their conjugacy classes belong to  $S_1(l_1, \alpha_1, \beta_1)$  and  $S_2(l_2, \alpha_2, \beta_2)$ , respectively, where  $S_1$  and  $S_2$  denote one of  $S^{+,+}$ ,  $S^{-,+}$ ,  $S^{+,-}$  or  $S^{-,-}$ . As the abelianization of an element in  $S^{\pm,\pm}(l, \alpha, \beta)$  is  $\begin{bmatrix} \pm\alpha \\ \pm\beta \end{bmatrix}$ , if  $A(g) = A(h)$ , then  $S_1(l_1, \alpha_1, \beta_1) = S_2(l_2, \alpha_2, \beta_2)$ . By Theorem 4.3, this implies that  $g$  and  $h$  are conjugate.

The second part of the corollary can be seen by running the Euclidean algorithm that arises in Theorem 4.3 in reverse. We will explicitly show this in Theorem 4.6. □

As the subgroup of commutators  $[F_2, F_2]$  is characteristic, an automorphism of  $F_2$  defines an automorphism of  $\mathbb{Z}^2$ . This defines a homomorphism

$$\rho: \text{Aut}(F_2) \rightarrow \text{Aut}(\mathbb{Z}^2) = \text{GL}(2, \mathbb{Z}).$$

This homomorphism satisfies  $A \circ \phi = \rho(\phi) \circ A$ . In terms of matrices, this map is defined by

$$\rho(\phi) = \begin{bmatrix} \exp_a(\phi(a)) & \exp_a(\phi(b)) \\ \exp_b(\phi(a)) & \exp_b(\phi(b)) \end{bmatrix}.$$

**Corollary 4.5** [Nielsen 1917]. *Let  $\phi \in \text{Aut}(F_2)$ . If  $\rho(\phi) = \text{Id}$ , then there is a  $g \in F_2$  such that  $\phi(x) = gxg^{-1}$ .*

*Proof.* If  $\rho(\phi) = \text{Id}$ , then as  $\phi(a)$  is primitive and  $A\phi(a) = \rho(\phi)A(a) = A(a)$ , we have that  $\phi(a)$  is conjugate to  $a$  by Corollary 4.4. Say  $\phi(a) = g_1 a g_1^{-1}$ . Likewise, we have that  $\phi(b) = g_2 b g_2^{-1}$ . Define  $\psi \in \text{Aut}(F_2)$  by  $\psi(x) = g_1^{-1} x g_1$ . Thus  $\psi\phi(a) = a$  and  $\psi\phi(b) = g_3 b g_3^{-1}$ , where  $g_3 = g_1^{-1} g_2$ . As  $\psi\phi$  is an automorphism of  $F_2$ , the set  $\{a, g_3 b g_3^{-1}\}$  is a basis for  $F_2$ ; in particular, this set generates  $F_2$ . Using a method such as Stallings' foldings [Stallings 1983], it is clear that this is only possible if  $g_3 = a^k$  for some  $k$ . Thus  $\phi(x) = g_1 a^k x a^{-k} g_1^{-1}$ . □

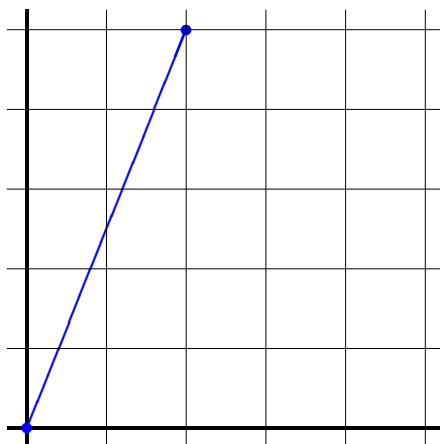
We will now give an explicit description of the cyclic word in  $S(l, \alpha, \beta)$  when  $\text{gcd}(\alpha, \beta) = 1$ . When the  $\text{gcd}(\alpha, \beta) = d \neq 1$ , the cyclic word is obtained by taking the  $d$ -th power of the cyclic word in  $S(l/d, \alpha/d, \beta/d)$ . Our description matches that of Osborne and Zieschang [1981].

For simplicity, we assume  $S = S^{+,+}$ . Let  $\begin{bmatrix} \alpha \\ \beta \end{bmatrix} \in \mathbb{Z}^2$  be such that  $\alpha, \beta \geq 1$  and  $\text{gcd}(\alpha, \beta) = 1$ . Let  $L_{\alpha,\beta}$  denote the line segment in  $\mathbb{R}^2$  from  $(0, 0)$  to  $(\alpha, \beta)$ . Define  $v_{\alpha,\beta}$  as the word in  $\{a, b\}$  where an  $a$  appears for each integer vertical line  $L_{\alpha,\beta}$  crosses and a  $b$  appears for each integer horizontal line  $L_{\alpha,\beta}$  crosses. The letters appear in the order of the lines  $L_{\alpha,\beta}$  crosses. As  $\text{gcd}(\alpha, \beta) = 1$ , the interior of  $L_{\alpha,\beta}$  does not simultaneously cross both an integer horizontal line and a integer vertical line. See Figure 8.

Now define  $w_{\alpha,\beta} = a v_{\alpha,\beta} b$ . Also define  $w_{1,0} = a$  and  $w_{0,1} = b$ . In the case that  $\alpha$  or  $\beta$  are negative, the words  $v_{\alpha,\beta}$  and  $w_{\alpha,\beta}$  are defined analogously.

**Theorem 4.6.** *Suppose  $\alpha, \beta \geq 0$  and that  $\text{gcd}(\alpha, \beta) = 1$ . The unique cyclic word in  $S^{+,+}(l, \alpha, \beta)$  is determined by  $w_{\alpha,\beta}$ .*

*Proof.* For simplicity we denote  $S = S^{+,+}$ . If  $\alpha = 0$  or  $\beta = 0$  then the theorem is clear. Likewise if  $\alpha = \beta = 1$ . In this case,  $v_{1,1}$  is the empty word and therefore



**Figure 8.** The line segment  $L_{2,5}$  and the word  $v_{2,5} = bbabb$ .

$w_{1,1} = ab$ . The cyclic word determined by  $w_{1,1}$  is the unique separable word in  $S(2, 1, 1)$ .

Assume that  $\alpha > \beta > 0$ . We show that  $w_{\alpha-\beta,\beta} = \phi(w_{\alpha,\beta})$ , where  $\phi$  is the Whitehead automorphism from the proof of Proposition 4.2, namely  $\phi(a) = a$  and  $\phi(b) = a^{-1}b$ . Since  $\alpha > \beta$ , each of the  $b$ 's in  $w_{\alpha,\beta}$  is isolated, as crossing two adjacent horizontal lines without crossing a vertical line implies the slope of  $L_{\alpha,\beta}$  is greater than 1, i.e.,  $\beta/\alpha > 1$ .

Thus it is clear that both  $w_{\alpha-\beta,\beta}$  and  $\phi(w_{\alpha,\beta})$  contain the same number of  $a$ 's and  $b$ 's, namely  $\alpha - \beta$  and  $\beta$ , respectively. The difference between  $w_{\alpha,\beta}$  and  $\phi(w_{\alpha,\beta})$  is one fewer  $a$  between adjacent  $b$ 's.

Notice that for  $i = 0, \dots, \beta - 1$ , the number of  $a$ 's between the  $i$ -th and  $(i + 1)$ -st  $b$  of  $v_{\alpha,\beta}$  is  $\langle (i + 1)\alpha/\beta \rangle - \langle i\alpha/\beta \rangle$ , where  $\langle x \rangle$  is the largest integer strictly less<sup>1</sup> than  $x$ . The 0-th  $b$  is interpreted as the beginning of  $v_{\alpha,\beta}$  and the  $\beta$ -th  $b$  is interpreted as the end of  $v_{\alpha,\beta}$ . Indeed,  $x = \langle i\alpha/\beta \rangle$  is the vertical line crossed by  $L_{\alpha,\beta}$  immediately preceding crossing the horizontal line  $y = i$ . Hence, we observe that the number of  $a$ 's between the  $i$ -th and  $(i + 1)$ -st  $b$  of  $v_{\alpha-\beta,\beta}$  is

$$\begin{aligned} \left\langle \frac{(i+1)(\alpha-\beta)}{\beta} \right\rangle - \left\langle \frac{i(\alpha-\beta)}{\beta} \right\rangle &= \left( \left\langle \frac{(i+1)\alpha}{\beta} \right\rangle - (i+1) \right) - \left( \left\langle \frac{i\alpha}{\beta} \right\rangle - i \right) \\ &= \left\langle \frac{(i+1)\alpha}{\beta} \right\rangle - \left\langle \frac{i\alpha}{\beta} \right\rangle - 1. \end{aligned}$$

This shows that  $\phi(w_{\alpha,\beta}) = w_{\alpha-\beta,\beta}$ .

If  $\beta > \alpha > 0$ , we have  $w_{\alpha,\beta-\alpha} = \psi(w_{\alpha,\beta})$  as above, where  $\psi(a) = ab^{-1}$  and  $\psi(b) = b$ .

<sup>1</sup>We use this variant of the floor function to avoid having to subtract 1 in the case  $i = \beta - 1$ .

By induction, this shows that  $w_{\alpha,\beta}$  is separable. By construction, the length of  $w_{\alpha,\beta}$  is  $l = \alpha + \beta$  and this word contains  $\alpha$   $a$ 's and  $\beta$   $b$ 's. Hence, the cyclic word determined by  $w_{\alpha,\beta}$  is the unique word in  $S(l, \alpha, \beta)$ .  $\square$

We end this section by showing that the above analysis allows for an exact count of the number of separable cyclic words of a given length. Let  $S^{+,+}(l)$  be the set of all positive conjugacy classes of length  $l$  that are separable. Then  $S^{+,+}(l)$  is the disjoint union

$$S^{+,+}(l) = S^{+,+}(l, 0, l) \cup S^{+,+}(l, 1, l-1) \cup \dots \cup S^{+,+}(l, l-1, 1) \cup S^{+,+}(l, l, 0).$$

So

$$\#|S^{+,+}(l)| = \sum_{\alpha=0}^l \#|S^{+,+}(l, \alpha, l-\alpha)| = l + 1.$$

Likewise, we can define  $S^{-,+}(l)$ ,  $S^{+,-}(l)$  and  $S^{-,-}(l)$ . The cardinality of each of these sets is also  $l + 1$ . Notice that  $S^{+,+}(l) \cap S^{+,-}(l) = \{a^l\}$ . There are three similar equations regarding the other intersections.

**Theorem 4.7.** *The number of cyclic words of length  $l$  in  $F_2$  that are separable is  $4l$ .*

### 5. Whitehead graphs in $F_2$

In this final section we will explore to what extent the decay in the likelihood of an element being separable is a property of Whitehead graphs in rank 2.

Let  $\text{WhG}(l)$  denote the set of Whitehead graphs in ranks 2 with  $l$  edges. Let  $\text{Dis}(l)$  denote the subset that are disconnected and let  $\text{Cut}(l)$  denote the subset that are connected with a cut vertex.

By counting the number for each  $l$  we arrive at:

**Theorem 5.1.**  $\#|\text{Dis}(l)| + \#|\text{Cut}(l)| = 2l.$

*Proof.* First, separate the equation into two parts:  $\#|\text{Cut}(l)|$  and  $\#|\text{Dis}(l)|$ ; we compute each separately.

To compute  $\#|\text{Dis}(l)|$ , we refer to Figure 6. When  $l$  is even, each of the 4 forms can appear ( $\alpha = l$  in the top two and  $\alpha = l/2$  in the bottom two), and when  $l$  is odd, only the top two forms appear ( $\alpha = l$ ). Hence

$$\#|\text{Dis}(l)| = \begin{cases} 4 & \text{if } l \text{ is even,} \\ 2 & \text{if } l \text{ is odd.} \end{cases} \tag{6}$$

To compute  $\#|\text{Cut}(l)|$ , we again consider two cases depending on if  $l$  is even or odd. Referring to Figure 7, we must have  $l = \alpha + 2\beta$ .

When  $l$  is odd, as  $l = \alpha + 2\beta$ ,  $l$  is odd too. The least odd number that  $\alpha$  can be is 1, in this case  $\beta = (l - 1)/2$ . Therefore, the range of  $\beta$  when  $l$  is odd is

$$1 \leq \beta \leq \frac{l-1}{2}.$$

Each value of  $\beta$  results in four distinct graphs in  $\text{Cut}(l)$ .

When  $l$  is even, we have the same equation as above,  $l = \alpha + 2\beta$ , but the least even number that  $\alpha$  can be is 2, in this case  $\beta = (l - 2)/2$ . So the range of  $\beta$  when  $l$  is even is

$$1 \leq \beta \leq \frac{l-2}{2}.$$

Again, each value of  $\beta$  corresponds to four distinct graphs in  $\text{Cut}(l)$ . Combining these calculations, we have

$$\#\text{Cut}(l) = \begin{cases} 2l - 4 & \text{if } l \text{ is even,} \\ 2l - 2 & \text{if } l \text{ is odd.} \end{cases} \tag{7}$$

Combining (6) and (7) we get  $\#\text{Dis}(l) + \#\text{Cut}(l) = 2l$ . □

**Remark 5.2.** Comparing Theorems 4.7 and 5.1, we see that for each Whitehead graph in  $\text{Dis}(l) \cup \text{Cut}(l)$  there are exactly two separable conjugacy classes associated to that graph. These two conjugacy classes are related by inversion.

Next we count the total number of Whitehead graphs in rank 2 by taking combinations of the graphs in Figure 6. Again, we are using the observation that in a Whitehead graph the valence of the vertex  $v$  is the same as the valence of the vertex  $v^{-1}$ .

**Theorem 5.3.**  $\#\text{WhG}(l) = \begin{cases} \frac{1}{24}(l^3 + 9l^2 + 26l + 24) & \text{if } l \text{ is even,} \\ \frac{1}{24}(l^3 + 9l^2 + 23l + 15) & \text{if } l \text{ is odd.} \end{cases}$

*Proof.* We begin by constructing a generating function  $f(x)$  for  $\#\text{WhG}(l)$  [Brualdi 2010, Section 7.4]. A Whitehead graph with  $l$  edges is formed by combining graphs in Figure 6 with  $\alpha = 1$ . Each graph from the top row contributes one edge and each graph from the bottom row contributes two edges. Hence

$$f(x) = (1 + x + x^2 + \dots)(1 + x + x^2 + \dots)(1 + x^2 + x^4 + \dots)(1 + x^2 + x^4 + \dots).$$

Then  $\#\text{WhG}(l)$  is the coefficient of  $x^l$  in  $f(x)$ . In order to compute this coefficient, we will compute the Taylor series for  $f$  centered at 0. To compute  $f^{(l)}$ , we rewrite  $f$  and take the partial fraction decomposition:

$$\begin{aligned} f(x) &= \frac{1}{(1-x)^2(1-x^2)^2} \\ &= \frac{1}{8} \left( \frac{1}{1+x} + \frac{1}{1-x} \right) + \frac{1}{16} \left( \frac{1}{(1+x)^2} + \frac{3}{(1-x)^2} \right) + \frac{1}{4} \left( \frac{1}{(1-x)^3} + \frac{1}{(1-x)^4} \right). \end{aligned}$$

The  $l$ -th derivative of  $f$  at 0 is

$$f^{(l)}(0) = \frac{1}{8}l!((-1)^l + 1) + \frac{1}{16}(l+1)!((-1)^l + 3) + \frac{1}{4}\left(\frac{(l+2)!}{2} + \frac{(l+3)!}{6}\right).$$

After dividing by  $l!$ , the equation simplifies to

$$\frac{f^{(l)}(0)}{l!} = \frac{1}{8}((-1)^l + 1) + \frac{1}{16}(l+1)((-1)^l + 3) + \frac{1}{4}\left(\frac{(l+1)(l+2)}{2} + \frac{(l+1)(l+2)(l+3)}{6}\right).$$

We will have two cases, looking at the equation above, for  $(-1)^{\text{even}} = 1$  and  $(-1)^{\text{odd}} = -1$ . Thus

$$\frac{f^{(l)}(0)}{l!} = \begin{cases} \frac{1}{24}(l^3 + 9l^2 + 26l + 24) & \text{if } l \text{ is even,} \\ \frac{1}{24}(l^3 + 9l^2 + 23l + 15) & \text{if } l \text{ is odd.} \end{cases}$$

As  $\#\text{WhG}(l) = f^{(l)}(0)/l!$ , the proof is complete.  $\square$

Notice that although the likelihood of a cyclically reduced word being separable decays to 0 exponentially in the length of the word [Borovik et al. 2002], the likelihood of a Whitehead graph containing a cut vertex approaches 0 like  $1/l^2$ , where  $l$  is the number of edges of the graph.

### Acknowledgement

This research was conducted as a Student/Faculty Summer Research project at Allegheny College. The authors thank Allegheny College for its support.

### References

- [Borovik et al. 2002] A. V. Borovik, A. G. Myasnikov, and V. Shpilrain, “Measuring sets in infinite groups”, pp. 21–42 in *Computational and statistical group theory* (Las Vegas, NV/Hoboken, NJ, 2001), edited by R. Gilman et al., Contemp. Math. **298**, Amer. Math. Soc., Providence, RI, 2002. MR 2003m:20024 Zbl 1022.20010
- [Brualdi 2010] R. A. Brualdi, *Introductory combinatorics*, 5th ed., Pearson Prentice Hall, Upper Saddle River, NJ, 2010. MR 2012a:05001 Zbl 0915.05001
- [Cohen et al. 1981] M. Cohen, W. Metzler, and A. Zimmermann, “What does a basis of  $F(a, b)$  look like?”, *Math. Ann.* **257**:4 (1981), 435–445. MR 82m:20028 Zbl 0458.20028
- [Kassel and Reutenauer 2007] C. Kassel and C. Reutenauer, “Sturmian morphisms, the braid group  $B_4$ , Christoffel words and bases of  $F_2$ ”, *Ann. Mat. Pura Appl.* (4) **186**:2 (2007), 317–339. MR 2007j:20029 Zbl 1150.05042
- [Lyndon and Schupp 2001] R. C. Lyndon and P. E. Schupp, *Combinatorial group theory*, Springer, Berlin, 2001. MR 2001i:20064 Zbl 0997.20037
- [Nielsen 1917] J. Nielsen, “Die Isomorphismen der allgemeinen, unendlichen Gruppe mit zwei Erzeugenden”, *Math. Ann.* **78**:1 (1917), 385–397. MR 1511907

- [Osborne and Zieschang 1981] R. P. Osborne and H. Zieschang, “Primitives in the free group on two generators”, *Invent. Math.* **63**:1 (1981), 17–24. MR 82i:20042 Zbl 0438.20017
- [Piggott 2006] A. Piggott, “Palindromic primitives and palindromic bases in the free group of rank two”, *J. Algebra* **304**:1 (2006), 359–366. MR 2007g:20026 Zbl 1111.20028
- [Stallings 1983] J. R. Stallings, “Topology of finite graphs”, *Invent. Math.* **71**:3 (1983), 551–565. MR 85m:05037a Zbl 0521.20013
- [Stallings 1999] J. R. Stallings, “Whitehead graphs on handlebodies”, pp. 317–330 in *Geometric group theory down under: proceedings of a special year in geometric group theory* (Canberra, 1996), edited by J. Cossey et al., de Gruyter, Berlin, 1999. MR 2000e:20002 Zbl 1127.57300
- [Whitehead 1936] J. H. C. Whitehead, “On equivalent sets of elements in a free group”, *Ann. of Math.* (2) **37**:4 (1936), 782–800. MR 1503309 Zbl 0015.24804

Received: 2012-03-14      Accepted: 2012-12-27

mattclay@uark.edu	<i>Department of Mathematical Sciences, University of Arkansas, SCEN 301, Fayetteville, AR 72701, United States</i>
jconan01@rams.shepherd.edu	<i>Department of Education, Shepherd University, Shepherdstown, WV 25443, United States</i>
nivetharamas@gmail.com	<i>Clinton Township, MI 48036, United States</i>



# Perimeter-minimizing pentagonal tilings

Ping Ngai Chung, Miguel A. Fernandez,  
Niralee Shah, Luis Sordo Vieira and Elena Wikner

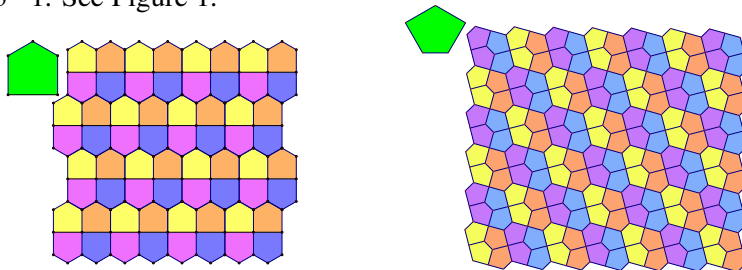
(Communicated by Michael Dorff)

We provide examples of perimeter-minimizing tilings of the plane by convex pentagons and examples of perimeter-minimizing tilings of certain small flat tori.

## 1. Introduction

**Cairo-prismatic tilings.** Thomas C. Hales [2001] proved the *honeycomb conjecture*, which says that regular hexagons provide a least-perimeter unit-area way to tile the plane. Squares and equilateral triangles, though less efficient than hexagons, provide a least-perimeter unit-area tiling by quadrilaterals and triangles.

It is interesting to ask about a least-perimeter unit-area *pentagonal* tiling, because regular pentagons do not tile the plane. Chung et al. [2012, Theorem 3.5] proved that among all convex unit-area pentagonal tilings of the plane and of appropriate flat tori, there are two that minimize perimeter: the *Cairo* and *prismatic* pentagons, defined as the unit-area pentagons having only  $90^\circ$  and  $120^\circ$  angles and circumscribed to a circle. The prismatic pentagon has adjacent right angles, and its sides (starting from the vertex along the axis of symmetry) are in the ratio  $1 : \frac{1}{2}(\sqrt{3}+1) : \sqrt{3}$ , while the Cairo pentagon has nonadjacent right angles, and its sides are in the ratio  $1 : 1 : \sqrt{3}-1$ . See Figure 1.



**Figure 1.** In green, the prismatic pentagon (left) and the Cairo pentagon (right). Each is a minimum-perimeter pentagonal tiler.

*MSC2010:* primary 52C20; secondary 52C05.

*Keywords:* tilings, pentagon, isoperimetric.

Building on this, we show in Propositions 2.3 and 2.4 below that each of these two polygons admits a unique monohedral edge-to-edge tiling. We also discuss mixed Cairo-prismatic tilings. Such tilings have been known at least since Marjorie Rice discovered a nonperiodic,  $D_6$ -symmetry example, first published as Figure 15 in [Schattschneider 1981] and shown also in [Chung et al. 2012], together with a number of other examples we have found (see Figures 11–23 below for a sampling).

**Restrictions on nonconvex tilings.** Can one beat the Cairo and prismatic pentagons by allowing nonconvex pentagonal shapes? In [Chung et al. 2012] (remark after Theorem 3.5) we conjectured that the answer is negative. Here we make some progress toward a proof: Proposition 2.10 below states that any tiling by unit-area nonconvex pentagons must have perimeter greater than a Cairo or prismatic tiling. Proposition 2.11 states that if a mixture of unit-area convex and nonconvex pentagons is perimeter-minimizing, then the ratio of the numbers of convex to nonconvex pentagons must be greater than 2.6. This provides a tool to investigate the problem further in Section 3.

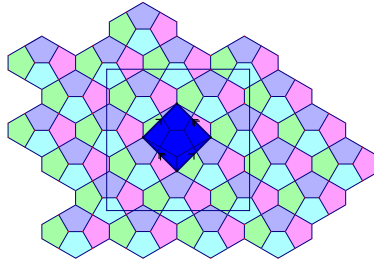
**Minimal tilings on flat tori.** Section 3 considers minimal tilings of small flat tori, which correspond to doubly periodic planar tilings. Proposition 3.3 states that the unique perimeter-minimizing edge-to-edge pentagonal tiling of a certain flat torus of area 2 is by prismatic pentagons, as shown in Figure 26. Similarly, Proposition 3.9 states that the unique perimeter-minimizing pentagonal tiling of the square torus of area 4 is by Cairo pentagons, as shown in Figure 2. Both these results allow for mixtures of nonconvex and convex pentagons.

The proofs depend on two main lemmas: Lemma 3.5 places a lower bound on the perimeter of a unit-area convex pentagon with a given small angle  $\alpha$ , and Lemma 3.8 places a lower bound on the perimeter of a nonconvex pentagon given two edges and the included angle.

We follow Proposition 3.9 by considering minimal tilings on other small flat tori and flat Klein bottles. Conjecture 3.4 proposes the minimal pentagonal tiling of the square torus of area 2. Conjecture 3.12 proposes minimal polygonal tilings for the square tori of areas 2, 3 and 4. Proposition 3.13 provides a lower bound on the perimeters of tilings of flat Klein bottles, and provides perimeter-minimizing tilings of many flat Klein bottles, as in Figure 35, right.

## 2. Cairo-prismatic tilings

**Definition 2.1.** A *tiling* is a decomposition of a surface into a union of simply connected disjoint open sets and their boundaries. The closure of each open set is called a *tile*. This paper focuses on tilings by unit-area pentagons. The Cairo and prismatic pentagons were defined on page 453; both have perimeter  $\sqrt{2}(1 + \sqrt{3}) \approx 3.86$ . We define the *perimeter ratio* of a tiling of the plane to be the limit superior as



**Figure 2.** Cairo tiling on square torus of area 4.

$r$  approaches infinity of the perimeter of the tiling inside a disc of radius  $r$  centered at the origin divided by  $\pi r^2$ . A *monohedral tiling* is a tiling by a single prototile.

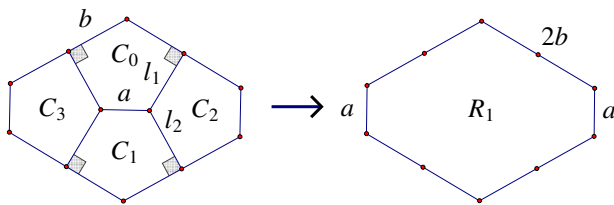
**Proposition 2.2** [Chung et al. 2012, Theorem 3.5]. *Perimeter-minimizing tilings of the plane by unit-area convex polygons with at most five sides are given by Cairo and prismatic tiles, as in Figure 1.*

**Remark.** Chung et al. remark that every doubly periodic perimeter-minimizing tiling by convex pentagons consists of Cairo and prismatic tiles. Of course, if allowed to break symmetry, one can alter a compact region arbitrarily without changing the limiting perimeter ratio.

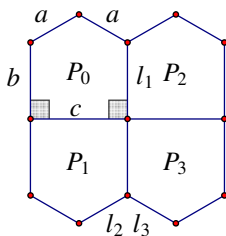
Next we state and prove Propositions 2.3 and 2.4.

**Proposition 2.3.** *The Cairo pentagonal tiling shown in Figure 1, right, is the unique tiling by Cairo pentagons.*

*Proof.* Consider a single Cairo prototile  $C_0$  with one side of length  $a$  and four sides of length  $b$ . The side of length  $a$  determines the orientation of the adjacent Cairo tile  $C_1$ , as shown in Figure 3. Furthermore, in a Cairo tile there is only one  $120^\circ$  angle with adjacent sides both of length  $b$ , which determines the orientations of tiles  $C_2$  and  $C_3$ . Thus, as shown in Figure 3, each Cairo tile must be in a hexagon with opposite edges of length  $a$  and the other four edges of equal length  $2b$ . We call this unit  $R_1$ . Using tiles  $C_2$  and  $C_3$ ,  $R_1$  determines vertical hexagons to its left and right. Similarly, these vertical hexagons each determine two horizontal hexagons above and below themselves. Therefore,  $R_1$  determines horizontal hexagons to its upper and lower left and right. Similarly, each of those horizontal hexagons determine four adjacent horizontal hexagons (among which are the other two hexagonal neighbors of  $R_1$ ). Continuing in this manner, a complete tiling by these horizontal hexagons is determined, and therefore the unique tiling, up to isometry, by Cairo pentagons is the tiling shown in Figure 1, right.  $\square$



**Figure 3.** A Cairo tiling must be a tiling by hexagons with opposite edges of length  $a$  and the other four edges of equal length  $2b$  ( $R_1$ ).



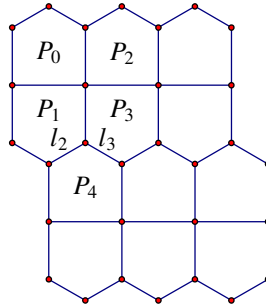
**Figure 4.** A prismatic pentagonal tiling must be a tiling by the hexagon with opposite edges of length  $2b$  and the other four edges of equal length  $a$ .

Notice that if we do not require edge-to-edge, then the prismatic tiling is not unique, as shown in Figure 10. On the other hand, if edge-to-edge is required, then the edge between the right angles has the unique length and determines the tiling:

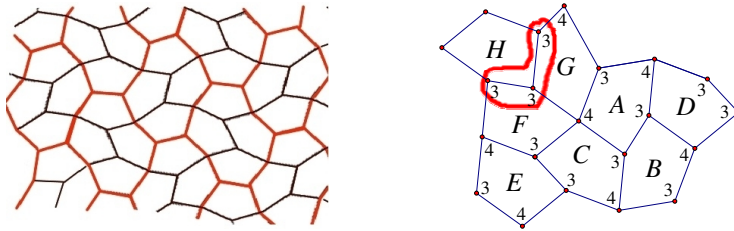
**Proposition 2.4.** *The prismatic tiling as shown in Figure 1, left, is the unique edge-to-edge tiling by prismatic pentagons.*

*Proof.* Consider a single prismatic prototile  $P_0$ . The unique edge of length  $c$  determines the orientation of the adjacent prismatic tile  $P_1$ . Thus, a prismatic pentagonal tiling must be a tiling by the hexagon with opposite edges of length  $2b$  and the other four edges of equal length  $a$ , as shown in Figure 4. Furthermore, the length  $b$  of edge  $l_1$  and the adjacent  $90^\circ$  angle determine the orientation of the adjacent prismatic tile  $P_2$ , which in turn determines the orientation of  $P_3$ . Continuing in this way, we construct a row of hexagons, each consisting of two prismatic pentagons, as shown in Figure 4. Note that the edges of length  $l_2$  and  $l_3$  determine the orientation of  $P_4$ . By a similar argument to that for  $P_0$ , the edge of length  $c$  of the tile  $P_4$  determines the orientation of the adjacent prismatic tile, establishing another row of two-prismatic hexagons, as shown in Figure 5. Continuing in this manner, we find that the unique edge-to-edge tiling by prismatic pentagons is, up to isometry, the prismatic tiling defined in Figure 1, right.  $\square$

For a monohedral tiling, denote the ordered degrees  $v_1, v_2, \dots, v_n$  of the vertices of a tile by  $[v_1, v_2, \dots, v_n]$ . Given the characterization of perimeter-minimizing



**Figure 5.** The edges of length  $l_2$  and  $l_3$  determine the orientation of  $P_4$ , which determines another row of two-prismatic hexagons.



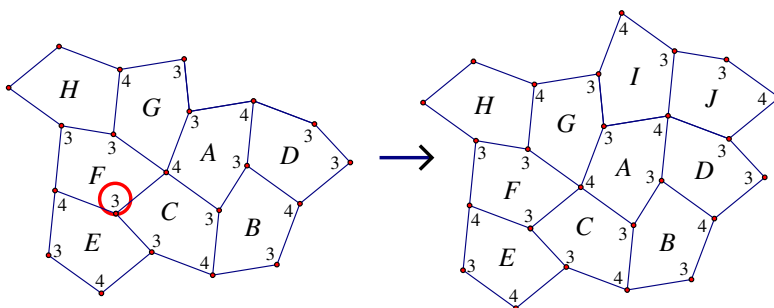
**Figure 6.** Left: the unique monohedral tiling with vertex degrees  $[3, 3, 4, 3, 4]$  (picture is taken from an earlier version of [Li et al. 2011]). Right: illustration of a contradiction in the proof of Proposition 2.5.

tilings by convex pentagons by Chung et al. [2012, Theorem 3.5], it is natural to ask whether the Cairo and prismatic tilings are the only monohedral tilings with the vertex degrees  $[3, 3, 4, 3, 4]$  and  $[3, 3, 3, 4, 4]$ , respectively.

Proposition 2.5 shows that any monohedral tiling with vertex degrees  $[3, 3, 4, 3, 4]$  is combinatorially equivalent to the tiling in Figure 6, left. Indeed, there are only two such tilings up to linear equivalence. Proposition 2.6 shows uncountably many distinct edge-to-edge tilings with vertex degrees  $[3, 3, 3, 4, 4]$  that are not equivalent under a linear map.

**Proposition 2.5.** *The tiling in Figure 6, left, is the unique tiling, up to combinatorial equivalence, by congruent tiles with vertex degrees  $[3, 3, 4, 3, 4]$ .*

*Proof.* Consider a single prototile  $A$  with vertex degrees  $[3, 3, 4, 3, 4]$  as shown in Figure 7. Then the edge connecting the vertices of degree 3 determines the degrees of the other three vertices of the adjacent tile  $B$ . Furthermore, each vertex of degree 3 shared by tiles  $A$  and  $B$ , as well as the two vertices of degree 4 adjacent to them, determine the degrees of the remaining two vertices of tiles  $C$  and  $D$ . Now, given the edge of  $C$  connecting the vertices of degree 3, the remaining three vertices

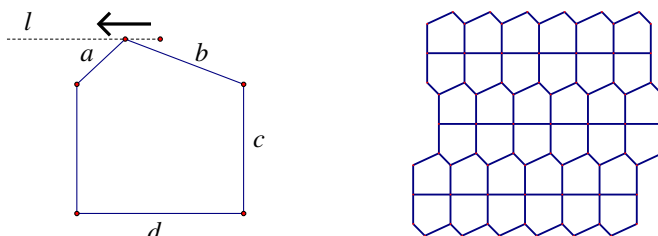


**Figure 7.** Unique monohedral tiling with vertex degrees  $[3, 3, 4, 3, 4]$  up to combinatorial equivalence.

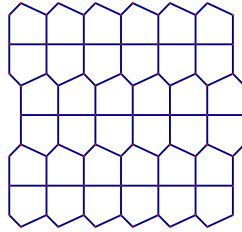
of the tile  $E$  are determined. The vertex of degree 3 circled in red and the adjacent vertices of degree 4 determine the degrees of the remaining two vertices of  $F$ . Now consider the tile  $G$ , which has three vertices determined by the adjacent tiles  $A$  and  $F$ . Suppose the vertex adjacent to the vertex of degree 3 shared with  $F$  was degree 3. Then the tile  $H$  adjacent to  $F$  and  $G$  would have three adjacent vertices of degree 3, a contradiction (Figure 6, right). Thus, this vertex in  $G$  must be of degree 4 and the remaining vertex is of degree 3. By a parallel argument, the vertex degrees of tile  $I$  are determined. Thus, the degrees of the vertices of all the tiles adjacent to  $A$  are determined. Continuing in this way, we can construct the tiling show in Figure 6, left. It follows that this is the unique tiling up to combinatorial equivalence by congruent tiles with vertex degrees  $[3, 3, 4, 3, 4]$ .  $\square$

**Proposition 2.6.** *Given vertex degrees  $[3, 3, 3, 4, 4]$ , there are uncountably many monohedral edge-to-edge tilings that are not equivalent under a linear mapping.*

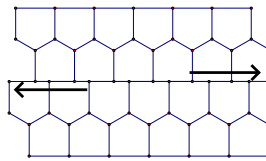
*Proof.* Take the prismatic tile and deform the line segments  $a$  and  $b$  of the pentagon as shown in Figure 8, left. Notice that each point on the line  $l$  has a different corresponding pentagon. There exists a monohedral tiling for each one of these prototiles with vertex degrees  $[3, 3, 3, 4, 4]$  (Figure 8, right). Therefore, there are



**Figure 8.** Left: different tiles with vertex degrees  $[3, 3, 3, 4, 4]$ . Right: a monohedral tiling with vertex degrees  $[3, 3, 3, 4, 4]$ .



**Figure 9.** Another construction of a nonequivalent monohedral tiling with vertex degrees  $[3, 3, 3, 4, 4]$ .



**Figure 10.** Prismatic tilings that are not edge-to-edge.

uncountably many monohedral, edge-to-edge tilings that are not equivalent under a linear mapping.  $\square$

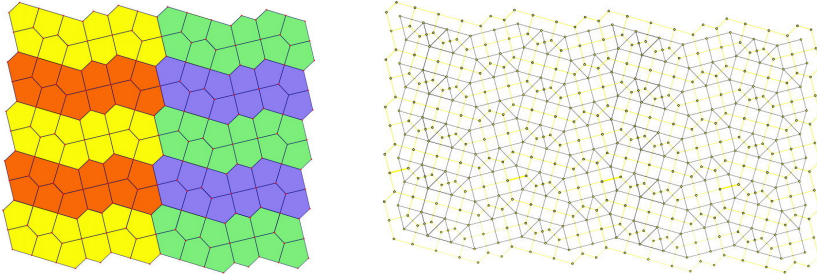
**Remark.** Another construction of a nonequivalent monohedral tiling with vertex degrees  $[3, 3, 3, 4, 4]$  is as shown in Figure 9. Similarly, for the prismatic tiling, one can translate a row of prismatic tiles sideways as in Figure 10. There may be many more.

To further the results of [Chung et al. 2012] we found many examples of tilings by mixtures of Cairo and prismatic pentagons. These tilings now appear in [Chung et al. 2012]. Meanwhile we discovered a Cairo-prismatic tiling by Marjorie Rice [ $\geq 2014$ ; Schattschneider 1981, Figure 15] who constructed this tiling even before Chung et al. proved that the Cairo and prismatic tilings are perimeter-minimizing for convex pentagons. We classify these tilings by their wallpaper groups in Figures 11–17. Daniel Huson [ $\geq 2014$ ] has found tilings by many other pairs of prototiles.

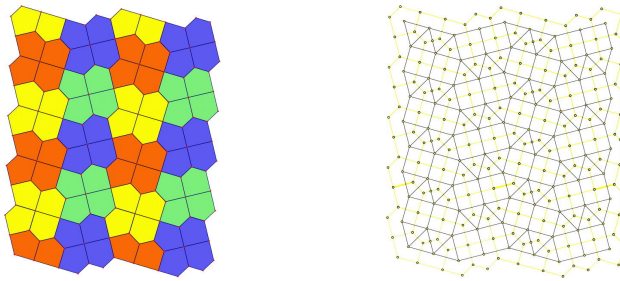
**Definition 2.7** [Schattschneider 1978]. *Wallpaper groups* are groups of isometries which leave a tiling invariant under linear combinations of two linearly independent translation vectors. Note that any such tiling is doubly periodic and therefore tiles a flat torus.

For a chart of the 17 wallpaper groups and their respective symmetries, refer to [Schattschneider 1978] or [Wikipedia Commons 2011].

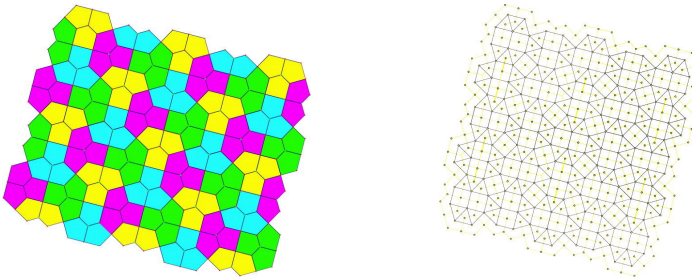
**Example 2.8.** Tilings of flat tori are sometimes categorized by their symmetries. Figures 11–17 are examples of Cairo and prismatic tilings for the listed wallpaper group.



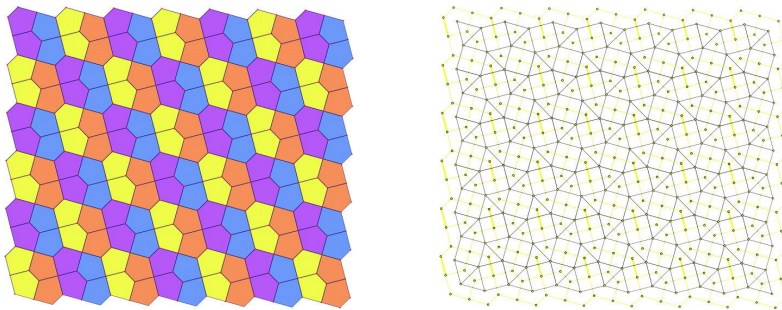
**Figure 11.** Tiling (left) and dual tiling (right) with  $p1$ .



**Figure 12.** Tiling (left) and dual tiling (right) with  $p2$ .

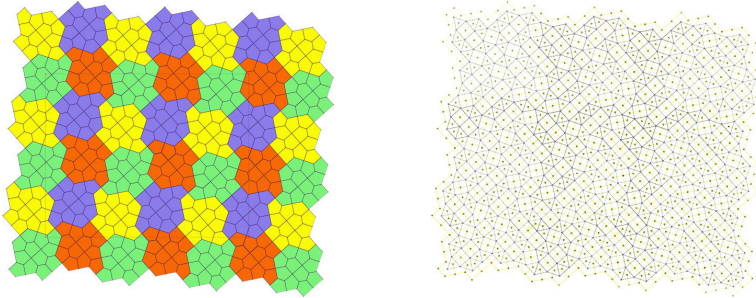


**Figure 13.** Tiling (left) and dual tiling (right) with  $p4g$ .

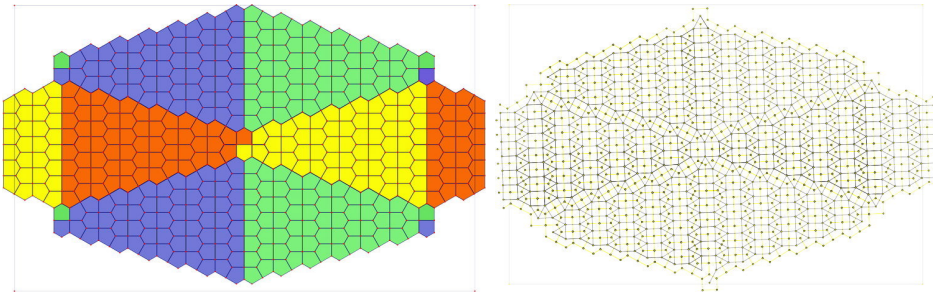


**Figure 14.** Cairo tiling with  $p4g$  (left) and its dual (right).

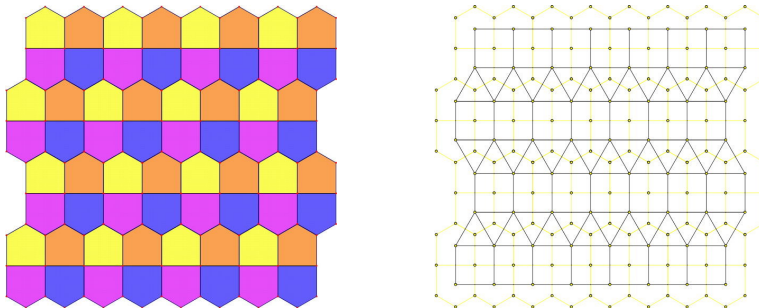




**Figure 15.** Spaceship tiling with  $p4g$  (left) and its dual (right).



**Figure 16.** Christmas tree tiling with  $cmm$  (left) and its dual (right).

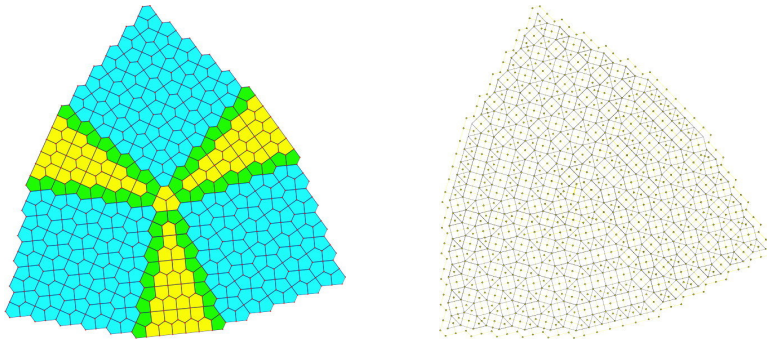


**Figure 17.** Prismatic tiling with  $cmm$  (left) and its dual (right).

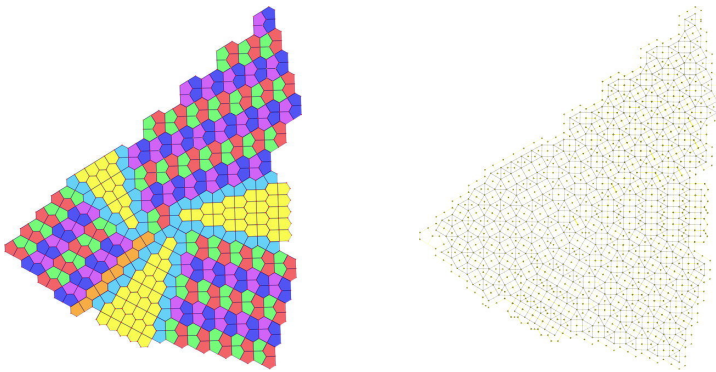
**Remark.** We think Figure 12, left, gives the unique planar tiling with fundamental region consisting of only two Cairo tiles and two prismatic tiles, but we didn't need this fact.

**Example 2.9.** Figures 18–23 are examples of Cairo-prismatic tilings which are not doubly periodic and therefore do not belong to a wallpaper group.

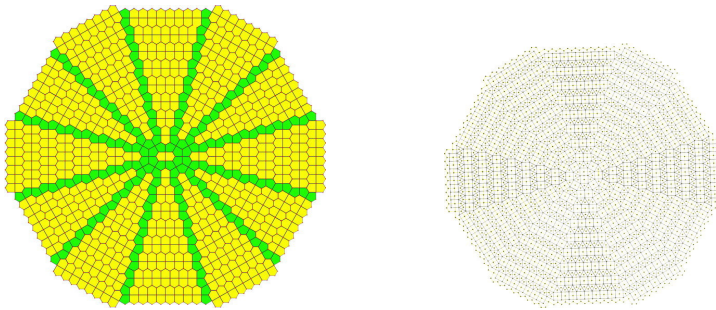
While it is still unknown whether the Cairo and prismatic tilings are perimeter-minimizing on the plane when we allow for mixtures of convex and nonconvex



**Figure 18.** Windmill tiling (left) and its dual (right).



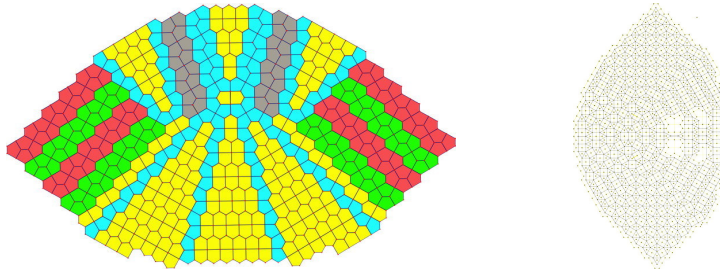
**Figure 19.** Chaos tiling (left) and its dual (right).



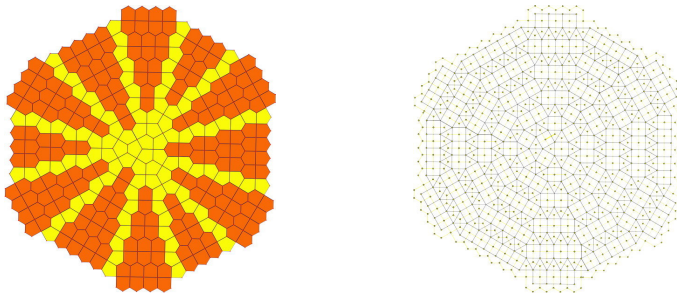
**Figure 20.** Plaza tiling (left) and its dual (right).

pentagons, we place bounds on the ratio of convex to nonconvex pentagons in order to rule out mixtures on certain flat tori.

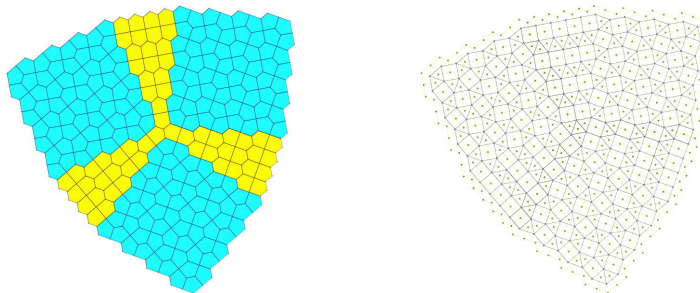
**Proposition 2.10.** *A tiling by unit-area nonconvex pentagons has more perimeter per tile than a Cairo or prismatic pentagon.*



**Figure 21.** Bunny tiling (left) and its dual (right).



**Figure 22.** Tiling by Marjorie Rice from [Schattschneider 1981, Figure 15] (left) and its dual (right).



**Figure 23.** Waterwheel tiling (left) and its dual (right).

*Proof.* Indeed, we may show that any unit-area nonconvex pentagon has perimeter greater than 4. Any nonconvex pentagon has more perimeter than its convex hull, which is a quadrilateral or triangle and hence has at least the perimeter of a square or equilateral triangle.  $\square$

**Proposition 2.11.** *In a perimeter-minimizing tiling by unit-area pentagons, the ratio of convex to nonconvex pentagons must be greater than 2.6.*

*Proof.* The perimeters of a regular pentagon, Cairo and prismatic pentagons, and the unit square are  $P_0 = 2\sqrt{5}\sqrt[4]{5} - 2\sqrt{5} > 3.81$ ,  $P_1 = 2\sqrt{2 + \sqrt{3}} > 3.86$ , and 4.

Since all nonconvex pentagons must have perimeter greater than or equal to that of the unit square, we consider the limit case in which the perimeter of the nonconvex pentagons is  $P_2 = 4$ . The convex pentagons have perimeter greater than or equal to that of the regular pentagon,  $P_0$ . Note that

$$\frac{P_2 - P_1}{P_1 - P_0} > 2.6,$$

and thus the ratio of regular pentagons to squares must be greater than 2.6. It follows that the ratio of convex to nonconvex pentagons must be greater than 2.6.  $\square$

**Lemma 2.12.** *For a flat torus of area 2, a tiling by nonconvex and convex unit-area pentagons has perimeter greater than 3.9.*

*Proof.* The perimeter of any nonconvex unit-area pentagon must be greater than or equal to the perimeter of the unit square, and the perimeter of any convex unit-area pentagon must be greater than or equal to that of the regular pentagon,  $2\sqrt{5}\sqrt[4]{5} - 2\sqrt{5} > 3.8$ . Thus, the total perimeter of a tiling by both nonconvex and convex pentagons on the appropriate torus of area 2 must be greater than

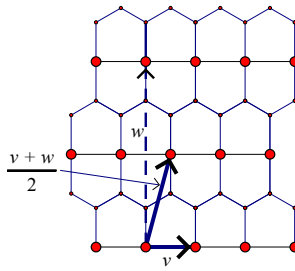
$$\frac{3.8 + 4}{2} = 3.9. \quad \square$$

### 3. Minimal tilings on flat tori

This section identifies unique optimal tilings for some small flat tori. Our main result, Proposition 3.9, states that the unique perimeter-minimizing unit-area pentagonal tiling of the square torus of area 4 is by Cairo pentagons as shown in Figure 2. Similarly, Conjecture 3.4 states that the minimal pentagonal tiling of the square torus of area 2 is by squares as in Figure 27. Proposition 3.3 states that the unique perimeter-minimizing unit-area pentagonal tiling of a certain flat torus of area 2 is by prismatic pentagons (Figure 26).

Similarly, we investigate minimal polygonal tilings of other small flat tori and Klein bottles (Figure 35). Wedd [2009] proposes that a regular hexagonal torus of area  $A \in \mathbb{N}$  can be tiled by regular hexagons if and only if  $A = x^2 + xy + y^2$  where  $x, y \in \mathbb{N}$ ,  $A \neq 0$  (Proposition 3.10). By [Hales 2001, Theorem 3], it would follow that these are unique perimeter-minimizing tilings (Proposition 3.11). Many flat tori cannot be tiled by regular hexagons. We investigate the square tori of areas 2, 3, 4 and conjecture that the minimal tilings are as shown in Figures 33 and 34.

In these proofs, we do not assume that the tiles are convex. Hales [2001, Theorem 1] proved that one cannot improve on the regular hexagonal tiling by mixing in nonconvex tiles. Similarly, Chung et al. [2012, Section 1] conjectured that one cannot improve on Cairo-prismatic tilings by mixing in nonconvex pentagons. We prove such results for some small flat tori.



**Figure 24.** Lattice of prismatic tiling.

Proposition 3.1 characterizes flat tori that can be tiled by prismatic pentagons:

**Proposition 3.1.** *A flat torus can be tiled by prismatic pentagons if and only if its fundamental polygon is determined by integer linear combinations of*

$$\langle \sqrt{6} - \sqrt{2}, 0 \rangle \quad \text{and} \quad \langle (\sqrt{6} - \sqrt{2})/2, (\sqrt{6} + \sqrt{2})/2 \rangle.$$

*Proof.* The prismatic tiling is the unique way to tile the plane with prismatic pentagons [Chung et al. 2012, Proposition 2.2]. The lattice of the tiling is generated by the two vectors  $\langle \sqrt{6} - \sqrt{2}, 0 \rangle$  and  $\langle (\sqrt{6} - \sqrt{2})/2, (\sqrt{6} + \sqrt{2})/2 \rangle$ , as shown in Figure 24. Therefore flat tori that can be tiled only by prismatic pentagons must have fundamental region determined by integer linear combinations of  $\langle \sqrt{6} - \sqrt{2}, 0 \rangle$  and  $\langle (\sqrt{6} - \sqrt{2})/2, (\sqrt{6} + \sqrt{2})/2 \rangle$ .

On the other hand, given a flat torus with fundamental polygon determined by integer linear combinations of  $\langle (\sqrt{6} - \sqrt{2}), 0 \rangle$  and  $\langle (\sqrt{6} - \sqrt{2})/2, (\sqrt{6} + \sqrt{2})/2 \rangle$ , one can cut it into many congruent parallelograms determined by vectors  $\langle (\sqrt{6} - \sqrt{2}), 0 \rangle$  and  $\langle (\sqrt{6} - \sqrt{2})/2, (\sqrt{6} + \sqrt{2})/2 \rangle$ , as shown in Figure 25. Each small parallelogram can be tiled by two prismatic pentagons. Thus the whole fundamental polygon can be tiled by prismatic pentagons as well.  $\square$

**Corollary 3.2.** *Prismatic pentagons do not tile a square torus.*

*Proof.* Suppose there exists a square torus that can be tiled by prismatic pentagons.

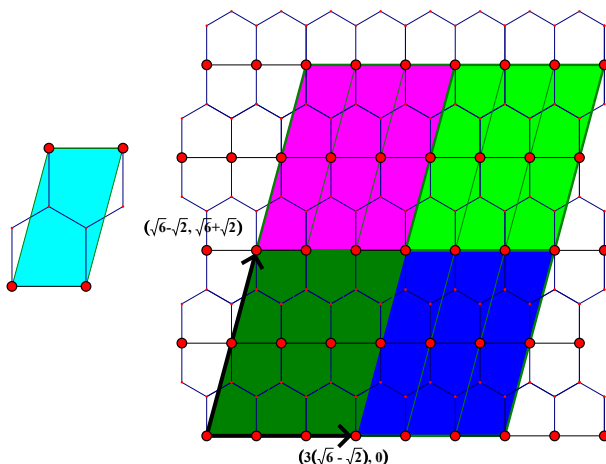
Let  $\mathbf{v} = \langle \sqrt{6} - \sqrt{2}, 0 \rangle$  and  $\mathbf{w} = \langle 0, \sqrt{6} + \sqrt{2} \rangle$  (Figure 24). By Proposition 3.1, a torus tiled by prismatic pentagons has fundamental polygon determined by integer linear combinations of  $\mathbf{v}$  and  $(\mathbf{v} + \mathbf{w})/2$ . Let  $a\mathbf{v} + b\mathbf{w}$  and  $c\mathbf{v} + d\mathbf{w}$  be the two linearly independent vectors determining the fundamental polygon, where  $a$  and  $b$  are either both integers or both half-integers, and similarly for  $c$  and  $d$ .

Therefore

$$(a\mathbf{v} + b\mathbf{w}) \cdot (c\mathbf{v} + d\mathbf{w}) = 0 \quad \text{and} \quad |a\mathbf{v} + b\mathbf{w}| = |c\mathbf{v} + d\mathbf{w}|$$

which implies

$$ac|\mathbf{v}|^2 + bd|\mathbf{w}|^2 = 0$$



**Figure 25.** The unique perimeter-minimizing pentagonal tiling of a certain flat torus (Proposition 3.1).

since  $\mathbf{v}$  and  $\mathbf{w}$  are orthogonal to each other. Since  $|\mathbf{v}|^2 = 8 - 4\sqrt{3}$  is not a rational multiple of  $|\mathbf{w}|^2 = 8 + 4\sqrt{3}$ ,  $ac = bd = 0$ . Since  $a\mathbf{v} + b\mathbf{w}$  and  $c\mathbf{v} + d\mathbf{w}$  are nonzero, either  $a = d = 0$  or  $b = c = 0$ . Without loss of generality assume that  $b = c = 0$ .

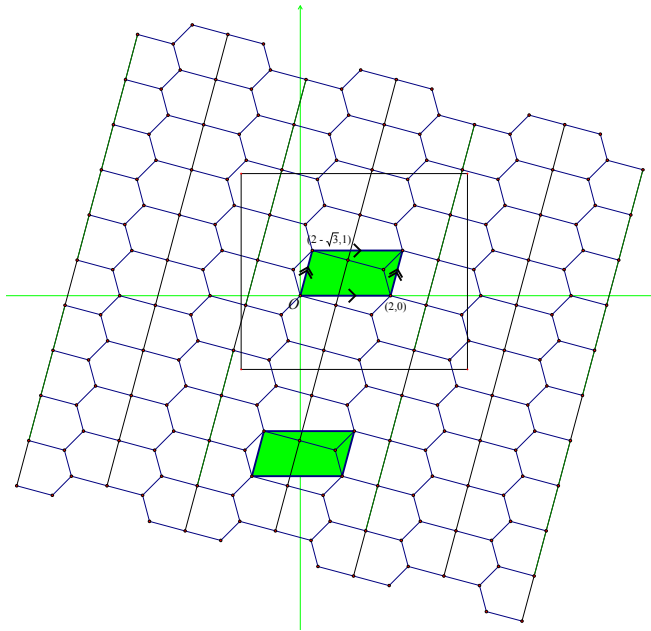
As a result,  $(\sqrt{6} - \sqrt{2})a = |a\mathbf{v}| = |d\mathbf{w}| = (\sqrt{6} + \sqrt{2})d$ . Since  $(\sqrt{6} - \sqrt{2})$  is not a rational multiple of  $(\sqrt{6} + \sqrt{2})$ , at least one of  $a$  and  $d$  is irrational, which contradicts their definitions. Therefore no square torus can be tiled by prismatic pentagons.  $\square$

Proposition 3.3 provides an example of a special hexagonal torus utilizing only prismatic pentagons.

**Proposition 3.3.** *For a flat torus of area 2 defined by the two vectors  $\langle 2 - \sqrt{3}, 1 \rangle$  and  $\langle 2, 0 \rangle$ , the prismatic tiling is the unique perimeter-minimizing edge-to-edge pentagonal tiling as shown in Figure 26.*

*Proof.* Note that it is possible to tile this torus with only prismatic tiles (Figure 26), and this tiling has total perimeter  $2\sqrt{2} + \sqrt{3} < 3.87$ . Since a tiling by nonconvex and convex unit-area pentagons has perimeter greater than 3.9 on a flat torus of area 2 by Lemma 2.12, a perimeter-minimizing tiling is by two convex pentagonal tiles. Since this tiling must be doubly periodic, by [Chung et al. 2012, remark after Theorem 3.5] perimeter-minimizing tilings by convex pentagons are uniquely given by Cairo and prismatic tiles.

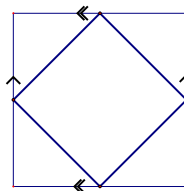
If there is at least one prismatic tile in the tiling, note that the base of this tile, which has length  $\sqrt{6} - \sqrt{2}$ , is unique among all the edges of the Cairo and prismatic pentagons. Since the tiling is edge-to-edge, the prismatic pentagon is consecutive to another prismatic pentagon rotated  $180^\circ$ . Thus these two pentagons tile the torus of area 2.



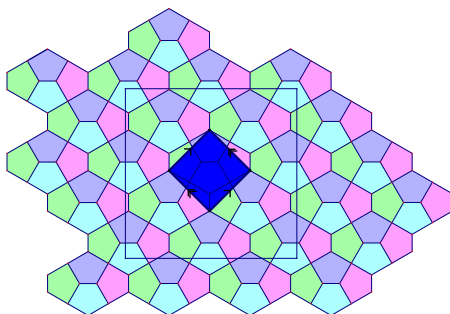
**Figure 26.** Prismatic Tiling on a flat torus of area 2 defined by the two vectors  $\langle 2 - \sqrt{3}, 1 \rangle$  and  $\langle 2, 0 \rangle$ .

If there is no prismatic tile, then the tiling consists of two Cairo pentagons. Since the short edge of the Cairo pentagon is unique, each Cairo pentagon is connected to another Cairo pentagon rotated  $180^\circ$ . Each of the two  $120^\circ$  angles between these two Cairo pentagons with two long adjacent edges can only fit in a third Cairo pentagon rotated  $90^\circ$ . Thus we have at least three pentagons that are not translational images of each other. Therefore they cannot all tile a flat torus of area 2.  $\square$

**Conjecture 3.4.** For a square torus of area 2, the unit-area square tiling is the unique perimeter-minimizing pentagonal tiling (Figure 27).



**Figure 27.** Conjectured perimeter-minimizing pentagonal tiling for the square torus of area 2.



**Figure 28.** Cairo tiling on square torus of area 4.

We now prove that for the square torus of area 4, the perimeter-minimizing tiling is given by Cairo tiles as in Figure 28. In the process we prove some bounds on the perimeters of certain classes of pentagons.

**Lemma 3.5.** *A unit-area convex pentagon with one of the angles  $\alpha \in (0, \pi)$  has perimeter greater than or equal to*

$$P(\alpha) = 2 \sqrt{\tan \frac{\pi - \alpha}{2} + 4 \tan \frac{\pi + \alpha}{8}}.$$

*Proof.* By [Chung et al. 2012, Proposition 3.1], the uniquely perimeter-minimizing pentagon with angles  $a_i, i = 1, 2, \dots, 5$ , has perimeter

$$2 \sqrt{\sum_{i=1}^5 \cot \frac{a_i}{2}}.$$

Since the function  $\cot$  is strictly convex up to  $\pi/2$ , fixing an angle  $a_1 = \alpha$  and taking the other angles to be equal will give the minimal perimeter. Thus the minimum perimeter is

$$2 \sqrt{\cot \frac{\alpha}{2} + 4 \cot \frac{3\pi - \alpha}{8}} = 2 \sqrt{\tan \frac{\pi - \alpha}{2} + 4 \tan \frac{\pi + \alpha}{8}}. \quad \square$$

**Lemma 3.6.** *A unit-area pentagon with one of the edges  $s$  has perimeter greater than or equal to*

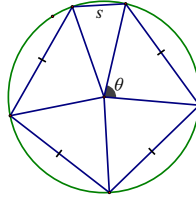
$$\frac{2\sqrt{2}(4 \sin(\theta/2) + \sin 2\theta)}{\sqrt{4 \sin \theta - \sin 4\theta}},$$

where  $\theta \in (0, \pi/2)$  is the only root of the equation

$$\frac{2\sqrt{2} \sin 2\theta}{\sqrt{4 \sin \theta - \sin 4\theta}} = s.$$

In fact,  $s$  is a strictly decreasing function of  $\theta$ .





**Figure 29.** Perimeter-minimizing unit-area pentagon given an edge length.

*Proof.* First of all it is well known that, given the edge lengths  $s_i, i = 1, 2, \dots, 5$ , of a pentagon, the one inscribed in a circle has the maximum area. The area is

$$\frac{1}{2}r^2 \sum_{i=1}^5 \sin \theta_i,$$

where  $\theta_i$  is the angle at center corresponding to the edge  $s_i$ .

Since  $\sin$  is strictly concave down in the range  $[0, \pi]$ , the more nearly equal the angles, the larger the area given a fixed perimeter. Therefore, fixing one edge, the unit-area pentagon inscribed in a circle with the four other edges of equal length has the minimum perimeter.

Let  $r$  be the radius of the circumcircle and  $\theta \in (0, \pi/2)$  be the angle at center which corresponds to one of the 4 edges of same length (Figure 29).

Then the perimeter is

$$P = 2r \left( 4 \sin \frac{\theta}{2} + \sin \frac{2\pi - 4\theta}{2} \right) = 2r \left( 4 \sin \frac{\theta}{2} + \sin 2\theta \right),$$

and the area is

$$A = \frac{1}{2}r^2 (4 \sin \theta + \sin(2\pi - 4\theta)) = \frac{1}{2}r^2 (4 \sin \theta - \sin 4\theta) = 1,$$

since the pentagon has unit area. After substitution of  $r$ , we get

$$P = \frac{2\sqrt{2}(4 \sin(\theta/2) + \sin 2\theta)}{\sqrt{4 \sin \theta - \sin 4\theta}}.$$

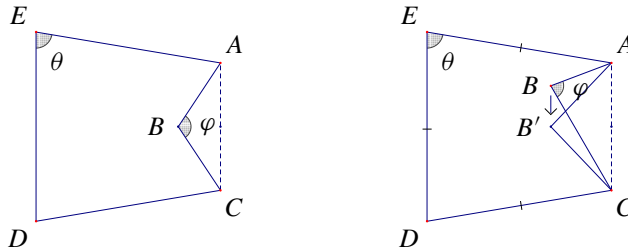
On the other hand,

$$s = 2r \sin \frac{2\pi - 4\theta}{2} = 2r \sin 2\theta = \frac{2\sqrt{2} \sin 2\theta}{\sqrt{4 \sin \theta - \sin 4\theta}}.$$

Hence

$$\frac{ds}{d\theta} = - \frac{16\sqrt{2}(7 \cos(\theta/2) + 3 \cos(3\theta/2)) \sin^3(\theta/2)}{(4 \sin \theta - \sin 4\theta)^{3/2}} < 0,$$

for  $0 < \theta < \pi/2$ . Therefore  $s$  strictly decreases as  $\theta$  increases in the range  $(0, \pi/2)$ ; thus there is only one value of  $\theta$  for a given  $s$ . □



**Figure 30.** Left: perimeter-minimizing pentagon in Lemma 3.7. Right: translation of  $B$  to  $B'$  to reduce perimeter without affecting the area.

**Lemma 3.7.** *A nonconvex unit-area pentagon  $ABCDE$  (Figure 30, left) satisfying*

$$CD = DE = EA, \quad AB = BC, \quad \text{and} \quad DE \parallel AC$$

*has perimeter*

$$P(\theta, \varphi) = \sqrt{\frac{(3 + (1 - 2 \cos \theta) / \sin(\varphi/2))^2}{(1 - \cos \theta) \sin \theta - (0.5 - \cos \theta)^2 \cot(\varphi/2)'}}$$

where  $\varphi = \angle ABC < \pi$  and  $\theta = \angle AED = \angle CDE$ ,  $\pi/3 < \theta < \pi$ . In particular, for a fixed value of  $\theta$ ,  $P$  is a decreasing function of  $\varphi$ .

*Proof.* The formula for  $P(\theta, \varphi)$  can be determined by direct calculation. Notice that, if we fix all the vertices except  $B$ , and increase the value of  $\varphi$  up to  $\pi$ , we will get less perimeter but more area. Thus we can scale the pentagon to get unit area with less perimeter. Therefore the perimeter  $P$  is a decreasing function of  $\varphi$ .  $\square$

**Lemma 3.8.** *Given positive real constants  $p, q, \varphi_0$  with  $p < q$  and  $\pi/2 < \varphi_0 \leq \pi$ , a unit-area nonconvex pentagon with an angle  $\phi = 2\pi - \varphi$  and two edges  $a, b$  adjacent to this angle that satisfies*

$$p \leq a \leq b \leq q, \quad \frac{\pi}{2} \leq \varphi \leq \varphi_0,$$

*has perimeter not less than*

$$\inf_{\pi/3 < \theta < \pi} P(\theta, \varphi'(p, q, \varphi_0)),$$

where

$$\varphi'(a, b, \varphi) = 2 \tan^{-1} \left( \frac{a^2 + b^2 - 2ab \cos \varphi}{2ab \sin \varphi} \right).$$

*Proof.* For a nonconvex pentagon  $ABCDE$  with  $\angle ABC = \varphi \leq \pi$ ,  $|\vec{AB}| = a$ ,  $|\vec{BC}| = b$ , consider a line  $\ell$  parallel to  $\vec{AC}$  and a point  $B'$  on  $\ell$  that satisfies

$|\overrightarrow{AB'}| = |\overrightarrow{B'C}|$ .  $AB'CDE$  is also unit-area, but with less perimeter than  $ABCDE$  unless  $B = B'$  (Figure 30, right).

Let  $\angle AB'C = \varphi' \leq \pi$ ,  $c = |\overrightarrow{AC}|$  and  $h$  be the distance from  $B$  to  $AC$ . Then

$$c^2 = a^2 + b^2 - 2ab \cos \varphi, \quad ch = ab \sin \varphi, \quad c = 2h \tan \frac{\varphi'}{2}.$$

After simplification,

$$\tan \frac{\varphi'}{2} = \frac{a^2 + b^2 - 2ab \cos \varphi}{2ab \sin \varphi}.$$

Hence,

$$\begin{aligned} \frac{\partial}{\partial a} \tan \frac{\varphi'}{2} &= \frac{1}{2b \sin \varphi} \left( 1 - \frac{b^2}{a^2} \right) \leq 0, \\ \frac{\partial}{\partial b} \tan \frac{\varphi'}{2} &= \frac{1}{2a \sin \varphi} \left( 1 - \frac{a^2}{b^2} \right) \geq 0, \\ \frac{\partial}{\partial \varphi} \tan \frac{\varphi'}{2} &= \csc^2 \varphi \left( 1 - \frac{a^2 + b^2}{2ab} \cos \varphi \right) > 0, \end{aligned}$$

since  $a \leq b$  and  $\varphi \geq \pi/2$ . Therefore the maximum of  $\tan(\varphi'/2)$  is attained at the point where  $(a, b, \varphi) = (p, q, \varphi_0)$ . Since  $\tan$  is an increasing function in the range  $(0, \pi/2)$ , the maximum of  $\varphi'$  is attained at the same point; i.e.,

$$\varphi'(a, b, \varphi) \leq \varphi'(p, q, \varphi_0),$$

for all  $p \leq a \leq b \leq q$  and  $\pi/2 \leq \varphi \leq \varphi_0$ .

By Lemma 3.7, the perimeter of  $AB'CDE$  is a decreasing function of  $\angle AB'C$ ; therefore

$$\begin{aligned} \text{perim}(ABCDE) &\geq \text{perim}(AB'CDE) \\ &= P(\angle AED, \varphi'(a, b, \varphi)) \\ &\geq P(\angle AED, \varphi'(p, q, \varphi_0)) \\ &\geq \inf_{\pi/3 < \theta < \pi} P(\theta, \varphi'(p, q, \varphi_0)). \quad \square \end{aligned}$$

**Proposition 3.9.** *For a square torus of area 4, the Cairo tiling is the unique perimeter-minimizing edge-to-edge pentagonal tiling.*

*Proof.* Recall that the lower bounds on the perimeters of a convex and nonconvex unit-area pentagon are

$$P_{\text{convex}} \geq 2\sqrt{5} \sqrt[4]{5 - 2\sqrt{5}} > 3.81193, \quad P_{\text{nonconvex}} > 4,$$

since the perimeter-minimizing convex pentagon is a regular pentagon and the perimeter-minimizing nonconvex pentagon has more perimeter than a square. The

perimeter of each Cairo pentagon is  $2\sqrt{2 + \sqrt{3}} < 3.86371$ ; thus the total perimeter of the Cairo tiling on a square torus of area 4 is

$$P_C = \frac{1}{2} \times 4(2\sqrt{2 + \sqrt{3}}) < 7.72742.$$

Therefore the perimeter-minimizing tiling has total perimeter less than 7.72742.

If there are at least two nonconvex pentagons in a pentagonal tiling, the total perimeter will be greater than  $(3.81193 \times 2 + 4 \times 2)/2 = 7.81193 > 7.72742$ ; thus it is not perimeter-minimizing. Therefore a perimeter-minimizing tiling has at most one nonconvex pentagonal tile.

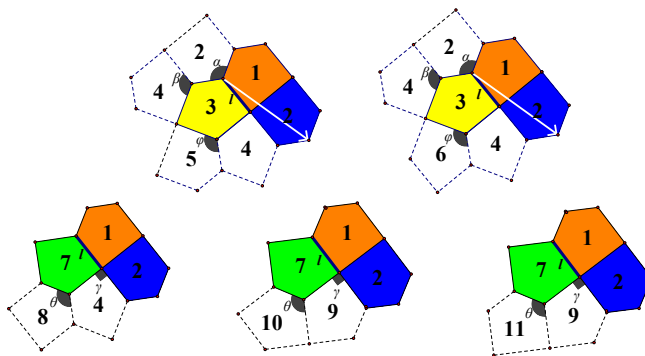
**Case 1: There is no nonconvex pentagonal tile.** Note that it is possible to tile this torus with only Cairo tiles (Figure 28). Since this tiling must be doubly periodic, by [Chung et al. 2012, remark after Theorem 3.5] perimeter-minimizing tilings by convex pentagons are uniquely given by Cairo and prismatic tiles.

*Case 1.1: The tiling consists of only Cairo pentagons.* This corresponds to a doubly periodic planar tiling by Cairo pentagons. By [Chung et al. 2012, Proposition 2.2] the Cairo tiling is the unique pentagonal planar tiling by Cairo tiles. Thus Figure 28 shows the unique perimeter-minimizing tiling in this case.

*Case 1.2: The tiling consists of only prismatic pentagons.* By Corollary 3.2, a square torus cannot be tiled by prismatic pentagons.

*Case 1.3: The tiling consists of both Cairo and prismatic pentagons.* Note that the length of the base of the prismatic pentagon is unique among all the edges of the Cairo and prismatic pentagons. Since the tiling is edge-to-edge, the prismatic pentagon is consecutive to another prismatic pentagon rotated 180°. The rest of the proof is presented with reference to Figure 31.

The pentagons are labeled with the same number if and only if they are trans-



**Figure 31.** Proof of Case 1.3: A square torus of area 4 cannot be tiled by a mixture of Cairo and prismatic pentagons.

lational images. Label the two prismatic pentagons as 1 and 2. Consider the edge  $l$  of pentagon 1. There are only two ways to put a Cairo pentagon next to edge  $l$ , namely pentagon 3 or 7, so that the degrees of the two ends of  $l$  match up.

If pentagon 3 is used, note that there is only one way to put a Cairo or prismatic pentagon at angle  $\alpha$ , which is to put another prismatic pentagon 2. The only way to put a Cairo or prismatic pentagon at angle  $\beta$  is to put a Cairo pentagon 4.

Now we have four unit-area tiles that are not translational images of each other; therefore they are all the tiles of the square torus of area 4. As a result, the vector (white arrow) that brings one of the pentagons labeled 2 to the other is indeed a translation vector of the tiling. Thus pentagon 4 is also brought right next to the other pentagon 2. At angle  $\varphi$ , there are two ways to put a Cairo or prismatic pentagon, namely pentagon 5 or 6. On the other hand, neither of them is a translational image of pentagon 1, 2, 3 or 4; thus this cannot be a tiling of a square torus of area 4. If pentagon 7 is used, consider angle  $\gamma$ . There are two ways to put a Cairo or prismatic pentagon, namely pentagon 4 or 9. If pentagon 4 is used, there is only one way to put a Cairo or prismatic pentagon at angle  $\theta$ : pentagon 8. In this case we have five tiles, no two of which are translational images of each other; thus they cannot tile the square torus of area 4. Thus pentagon 9 has to be put at angle  $\gamma$ . There are two ways to put a Cairo or prismatic pentagon at angle  $\theta$ , namely pentagon 10 or 11. However, neither of them is a translational image of any of the other four pentagons. Thus, again, this cannot tile the square torus of area 4.

Therefore we can conclude that we cannot tile the square torus of area 4 with both Cairo and prismatic pentagons.

**Case 2: There is one nonconvex pentagonal tile.** Since the area of the torus is 4, there are three convex pentagonal tiles.

If there is a convex tile with perimeter greater than or equal to 3.831, the total perimeter will be greater than  $(3.81193 \times 2 + 3.831 + 4)/2 = 7.72743 > 7.72742$ , thus not perimeter-minimizing. Therefore the perimeter of each convex tile satisfies

$$3.81193 < P_{\text{convex}} < 3.831.$$

If the nonconvex tile has perimeter greater than or equal to 4.0191, the total perimeter will be greater than  $(3.81193 \times 3 + 4.0191)/2 = 7.727445 > 7.72742$ , thus not perimeter-minimizing. Therefore the perimeter of the nonconvex tile satisfies

$$4 < P_{\text{nonconvex}} < 4.0191.$$

Consider the interior angles of the convex tiles. Note that for  $\alpha \geq 1.5792$  or  $\alpha \leq 2.2341$ , a convex pentagon with an angle  $\alpha$  has perimeter greater than or equal to  $P(\alpha) > 3.831$  by Lemma 3.5. Thus the convex tiles in a perimeter-minimizing tiling have interior angles within the range  $(1.5792, 2.2341)$ .

Now consider the edge lengths of the convex tiles. By Lemma 3.6 and the fact that  $3.81193 < P < 3.831$ ,

$$\frac{2\sqrt{2}(4 \sin(\theta/2) + \sin 2\theta)}{\sqrt{4 \sin \theta - \sin 4\theta}} < 3.831;$$

thus

$$1.1565 < \theta < 1.3564.$$

Since  $s$  is a decreasing function of  $\theta$ , we can get a bound on the edge length  $s$  of a convex tile:

$$0.5444 < s < 0.9659.$$

Consider the nonconvex pentagonal tile  $ABCDE$  with a reflex  $\angle ABC$ . Since the interior angles of the convex tiles are greater than  $1.5792 > \pi/2$ , if the point  $B$  has degree at least 3, the total angle at that point will be greater than  $\pi + (\pi/2) \times 2 = 2\pi$ , a contradiction. Therefore the nonconvex angle has degree 2. In this case  $\angle ABC$  is an interior angle of a convex tile; thus

$$1.5792 < \angle ABC < 2.2341 \quad \text{and} \quad 0.5444 < AB, BC < 0.9659.$$

By Lemma 3.8, the perimeter of  $ABCDE$  is not less than

$$\inf_{\pi/3 < \theta < \pi} P(\theta, \varphi'(0.5444, 0.9659, 2.2341)),$$

where

$$\varphi'(a, b, \varphi) = 2 \tan^{-1} \frac{a^2 + b^2 - 2ab \cos \varphi}{2ab \sin \varphi}$$

and

$$P(\theta, \varphi) = \sqrt{\frac{4(1.5 + (0.5 - \cos \theta) / \sin(\varphi/2))^2}{(1 - \cos \theta) \sin \theta - (0.5 - \cos \theta)^2 \cot(\varphi/2)}}.$$

After direct computation, the minimum is greater than  $4.078 > 4.0191$ . Therefore, in this case, the tiling will not be perimeter-minimizing.  $\square$

Some regular hexagonal tori can be tiled by regular hexagons. Wedd [2009] states without proof that for a regular hexagonal torus of area  $A$ , where  $A > 0$ , a tiling by regular hexagons exists if and only if  $A = x^2 + xy + y^2$  for some nonnegative integers  $x, y$ . We give a complete proof of Wedd's statement. By [Hales 2001, Theorem 3], this would be the unique perimeter-minimizing tiling of such a torus.

**Proposition 3.10** [Wedd 2009]. *A regular hexagonal torus with area  $A$ , where  $A \neq 0$ , can be tiled by regular hexagons with unit area if and only if  $A = x^2 + xy + y^2$ , where  $x, y$  are nonnegative integers.*

*Proof.* First we shall prove that a regular hexagonal torus with  $A = x^2 + xy + y^2$  can be tiled by regular hexagons. Our approach is to start with a unit-area hexagonal tiling of the plane, then find a regular hexagon with area  $A$  so that the tiling also tiles the torus formed by identifying opposite edges of this hexagon.

On a planar tiling by unit-area regular hexagons, construct the complex plane as follows: pick the center of one of the hexagons as the origin, and one of the vertices of the same hexagon as the point  $p$ , where  $p \in \mathbb{R}$  ( $p > 0$ ) is the distance between the vertex and the center. Let  $\zeta = e^{\pi i/3}$  be a primitive sixth root of unity. Consider the triangular lattice  $L$  formed by the vertices and centers of all the hexagons. It is generated by two points  $p$  and  $q$ , where  $q = p\zeta$ . Let  $x, y$  be nonnegative integers such that  $A = x^2 + xy + y^2$ . Note that

$$|xp + yq| = |x + y\zeta|p = p\sqrt{x^2 + y^2 + xy} = p\sqrt{A}.$$

Therefore the area of the regular hexagon with vertices

$$xp + yq, (xp + yq)\zeta, (xp + yq)\zeta^2, (xp + yq)\zeta^3, (xp + yq)\zeta^4, (xp + yq)\zeta^5$$

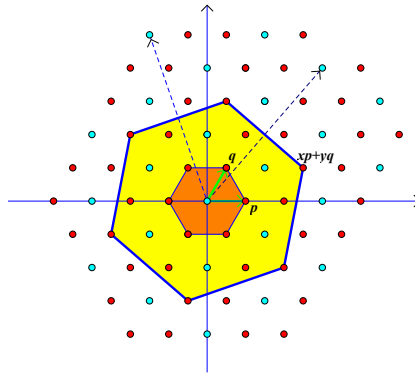
is  $A$  times the area of the hexagon with vertices  $p, p\zeta, p\zeta^2, p\zeta^3, p\zeta^4, p\zeta^5$ , which is precisely the unit-area hexagon centered at the origin. It is left to verify that the torus  $T$  formed by identifying opposite edges of the big hexagon is tiled by the unit-area hexagons.

Note that a tiling of a flat torus corresponds to a doubly periodic tiling of the plane, with translational vectors given by the vectors that define the fundamental parallelogram. In our case, the fundamental parallelogram of the torus  $T$  is spanned by  $(xp + yq) + (xp + yq)\zeta$  and  $(xp + yq)\zeta + (xp + yq)\zeta^2$ , as one may verify. These two vectors send a lattice point in  $L$  to another lattice point in  $L$ . Furthermore, in the lattice  $L$ , a point  $ap + bq$ , with  $a, b \in \mathbb{Z}$ , is the center of a hexagonal tile if and only if  $3 \mid a - b$ . Since

$$\begin{aligned} (xp + yq) + (xp + yq)\zeta &= (x - y)p + (2y + x)q, \\ (xp + yq)\zeta + (xp + yq)\zeta^2 &= (-x - 2y)p + (2x + y)q, \end{aligned}$$

and 3 divides both  $(2y + x) - (x - y)$  and  $(2x + y) - (-x - 2y)$ , both vectors send centers to centers and vertices to vertices. As a result, these two vectors send  $L$  to itself. Therefore we can conclude that the unit-area hexagonal tiling is a tiling of the torus  $T$ . This is the desired tiling of the regular hexagonal torus with area  $A = x^2 + xy + y^2$ .

The converse is true since the vertices of the big hexagon all lie on the lattice  $L$ . Scale the lattice  $L$  so that  $p = 1$ ; then the distance between any two lattice points is  $x^2 + y^2 + 2xy \cos(4\pi/3) = x^2 + xy + y^2$  for some nonnegative integers  $x$  and  $y$ .



**Figure 32.** Illustration of the proof of Proposition 3.10 in the case when  $A = 7$ .

Thus the area of the original hexagon will be  $x^2 + xy + y^2$  times the area of each hexagonal tile, which is 1 in our case, as desired.  $\square$

**Proposition 3.11.** *A regular hexagonal tiling is the unique perimeter-minimizing unit-area tiling of any regular hexagonal torus of area  $A = x^2 + xy + y^2$ , where  $x, y$  are nonnegative integers.*

*Proof.* This is a direct result of [Hales 2001, Theorem 3] and Proposition 3.10.  $\square$

**Remark.** There exists no flat torus whose fundamental polygon is a pentagon with interior angles less than  $\pi$ .

To investigate polygonal tilings of quadrilateral tori we also present conjectures on the best way to tile a square torus. While the Cairo tiling is the best pentagonal tiling of the square torus of area 4, it may not be the perimeter-minimizing tiling of the square torus in general.

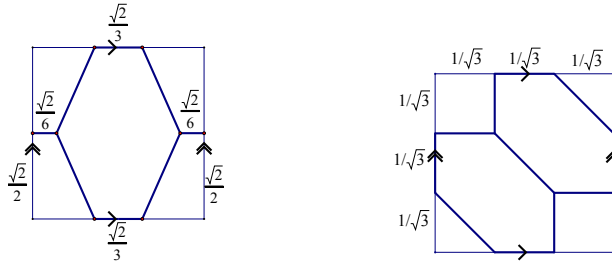
**Conjecture 3.12.** For the square tori of areas 2, 3, and 4, perimeter-minimizing unit-area tilings are given by the hexagonal tilings with dimensions as shown in Figures 33 and 34.

We conclude with some results on Klein bottles.

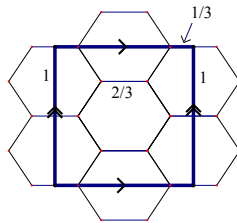
**Proposition 3.13.** *For a flat Klein bottle of integer area  $A$ , any unit-area tiling has perimeter greater than or equal to  $A/2$  times the perimeter of a unit-area regular hexagon. Equality is attained if and only if each tile is a regular hexagon.*

*Proof.* Since there exists a flat torus that double covers each flat Klein bottle, as shown in Figure 35, left, each tiling of a Klein bottle corresponds to a tiling of a certain flat torus. If there existed a tiling of a Klein bottle that had less perimeter than  $A/2$  times the perimeter of a unit-area regular hexagon, then there would exist a tiling of a flat torus that has a smaller perimeter ratio than the regular hexagonal

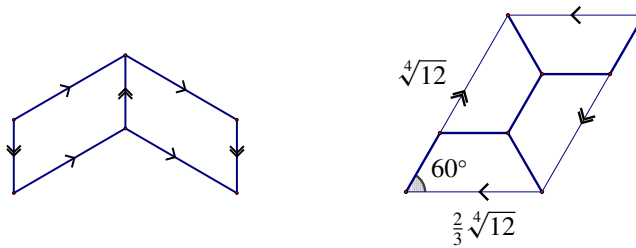




**Figure 33.** Proposed perimeter-minimizing tilings for the square torus of area 2 (left) and the square torus of area 3 (right).



**Figure 34.** Proposed perimeter-minimizing tiling for the square torus of area 4.



**Figure 35.** Left: every Klein bottle is double-covered by some flat torus. Right perimeter-minimizing tiling of a certain Klein bottle of area 2.

tiling. This would contradict the honeycomb conjecture [Hales 2001, Theorem 3]. □

**Example 3.14.** For the Klein bottle of area 2 of Figure 35, right, the unique perimeter-minimizing tiling is by regular hexagons.

### Acknowledgements

We thank our advisor Frank Morgan for his patience and invaluable input. We also thank the 2010 SMALL Geometry Group, Zane Martin of the 2012 SMALL

Geometry Group, Doris Schattschneider, Marjorie Senechal, Daniel Huson, and Chaim Goodman-Strauss for their help.

For funding, we thank the National Science Foundation for grants to Professor Morgan and to the Williams College SMALL Research Experience for Undergraduates; Williams College for additional funding; and the Mathematical Association of America for supporting our trip to talk at the 2011 MathFest.

## References

- [Chung et al. 2012] P. N. Chung, M. A. Fernandez, Y. Li, M. Mara, F. Morgan, I. R. Plata, N. Shah, L. S. Vieira, and E. Wikner, “Isoperimetric pentagonal tilings”, *Notices Amer. Math. Soc.* **59**:5 (2012), 632–640. MR 2954290 Zbl 06092121
- [Hales 2001] T. C. Hales, “The honeycomb conjecture”, *Discrete Comput. Geom.* **25**:1 (2001), 1–22. MR 2002a:52020 Zbl 1007.52008
- [Huson and Westphal  $\geq$  2014] D. Huson and K. Westphal, “2DTiler”, interactive tiling program, Eberhard Karls Universität Tübingen, <http://goo.gl/pl9Wq>.
- [Li et al. 2011] Y. Li, M. Mara, I. R. Plata, and E. Wikner, “Optimal planar tilings with vertex penalties”, preprint, 2011.
- [Rice  $\geq$  2014] M. Rice, “Intriguing tessellations”, web article, <http://goo.gl/uKavk>.
- [Schattschneider 1978] D. Schattschneider, “The plane symmetry groups: their recognition and notation”, *Amer. Math. Monthly* **85**:6 (1978), 439–450. MR 57 #17476 Zbl 0381.20036
- [Schattschneider 1981] D. Schattschneider, “In praise of amateurs”, pp. 140–166 in *The Mathematical Gardner*, edited by D. Klarner, Prindle, Weber & Schmidt, Boston, 1981.
- [Wedd 2009] N. S. Wedd, “Regular maps in the torus, with hexagonal faces”, website, 2009, <http://www.weddslist.com/groups/genus/1/hex.php>.
- [Wikipedia Commons 2011] Wikipedia, “Wallpaper group”, website, 2011, [http://en.wikipedia.org/wiki/Wallpaper\\_group](http://en.wikipedia.org/wiki/Wallpaper_group).

Received: 2012-03-29 Accepted: 2012-05-26

briancpn@mit.edu	<i>Department of Mathematics, Massachusetts Institute of Technology, 305 Memorial Drive, Cambridge, MA 02139, United States</i>
maf2831@truman.edu	<i>Mathematics and Computer Science Department, Truman State University, 100 E. Normal Avenue, Kirksville, MO 63501, United States</i>
niraleekshah@gmail.com	<i>Department of Mathematics and Statistics, Bronfman Science Center, Williams College, Williamstown, MA 01267, United States</i>
dw8603@wayne.edu	<i>Department of Mathematics, Wayne State University, 656 W. Kirby, Detroit, MI 48202, United States</i>
elena.wikner@gmail.com	<i>Department of Mathematics and Statistics, Bronfman Science Center, Williams College, Williamstown, MA 01267, United States</i>

# Discrete time optimal control applied to pest control problems

Wandi Ding, Raymond Hendon, Brandon Cathey,  
Evan Lancaster and Robert Germick

(Communicated by Suzanne Lenhart)

We apply discrete time optimal control theory to the mathematical modeling of pest control. Two scenarios: biological control and the combination of pesticide and biological control are considered. The goal is maximizing the “valuable” population, minimizing the pest population and the cost to apply the control strategies. Using the extension of Pontryagin’s maximum principle to discrete system, the adjoint systems and the characterization of the optimal pest controls are derived. Numerical simulations of various cases are provided to show the effectiveness of our methods.

## 1. Introduction

Pesticides and biological control are two popular ways of pest control. One of the conventional applications of control uses pesticides. The detrimental effects to local ecologies of overuse of pesticides has been widely documented, therefore, the conservation, introduction, and restocking of a pest’s natural enemies has become increasingly popular. Biological control is the use of living organisms to suppress the population of a specific pest organism, making it less abundant or less damaging than it would otherwise be [Eilenberg et al. 2001]. It is an environmentally sound and effective means of reducing or mitigating pests and pest effects through the use of natural enemies, and biological control has successfully contributed to the protection of the flora and fauna of many natural ecosystems [Driesche et al. 2010; Driesche 1994].

This study will focus on developing and analyzing two mathematical models for pest control using biocontrol and the combination of the pesticide and the biocontrol, while finding the optimal pest control strategies.

But biological control is both powerful and risky. Biological control agents may negatively affect native species directly or indirectly. Historically biological

---

*MSC2010:* 39A10, 49K99, 92D25.

*Keywords:* optimal control, biological pest control, discrete model, pesticide.

control introductions were not regulated the way they are today, and some horrible mistakes were made in the name of biological control (e.g., cane toads in Australia). Hawkins and Cornell [1999] gathered together recent theoretical developments and provide a guide to the critical issues that need to be considered in applying theory to biological control, they pointed out by developing theories based on fundamental population principles and the biological characteristics of the pest and agent, we can gain a much better understanding of when and how to use biological control.

A lot of studies done in this field have focused on the continuous predator-prey models, which are based on the assumption that population changes are always occurring. While this may be true for humans (births and deaths are fairly well distributed over time), many species have well-defined cycles of reproductions (births and deaths generally occur over a season or period of a few weeks or months). This fact causes us to focus on a discrete model over a continuous one for these biological systems.

The efficacy with which one is able to reduce a pest population is always subject to the amount of resources available to control that population. Due to cost and environmental consideration, it may be more appropriate to release a smaller amount of the predator population into the ecosystem, and then add to that amount incrementally for a given time frame to reduce the pest population more gradually. The costs involved will be substantially less in this case because the predator population will grow on its own, thus reducing the need to introduce more predators artificially, and it will be beneficial to the natural ecosystems.

Optimal control theory for discrete systems is well developed [Clark 1990; Sethi and Thompson 2006], but there are very few applications in pest control problems. Tang and Cheke [2008] studied integrated pest control problems using both continuous and discrete host-parasitoid models. Jang and Yu [2012] proposed a simple discrete time host-parasitoid model and derived an optimal control model using a chemical as a control for the hosts. They conclude that applying a chemical to eliminate the hosts directly may be a more effective control strategy than using the parasitoids to indirectly suppress the hosts. Whittle et al. [2007] use a discrete-time optimal control model to provide management for an invasive species consisting of a large main focus and several smaller outlier populations. Dabbs [2010] presents discrete time pest control models using three different growth functions: logistic, Beverton–Holt and Ricker spawner-recruit functions and compares the optimal control strategies respectively. Berryman [Hawkins and Cornell 1999] provides a review of the historic development of the ecological theory that relates to biological control, focusing on discrete time models that best describe systems in which the insects reproduce seasonally. He presents the control theory and the theory of predator-prey dynamics which are the key elements of the theory of biological control.

The paper is organized as follows: in Section 2 we present the optimal biological control problem, derive the adjoint equations and the characterization of the control, and give numerical results. In Section 3, we formulate the optimal dual control problems, derive the necessary conditions of optimal control and give some numerical results.

## 2. Optimal control using biological control

**2.1. The biological control problem.** Biological control of pests has been practiced in greenhouse as well as in field crops. For example *E. formosa* and *P. persimilis* have been used as biological control agents to reduce parasites over different crops such as tomatoes and cucumbers [van Lenteren and Woets 1988]. In our model, the valuable population, pest population and the predator (biological control agent) population are represented by

$$x = (x_0, x_1, \dots, x_T), \quad y = (y_0, y_1, \dots, y_T), \quad z = (z_0, z_1, \dots, z_T),$$

respectively, where the subscripts represent the time steps. The control satisfies

$$U_1 = \{u = (u_0, u_1, \dots, u_{T-1}) \in \mathbb{R}^T \mid 0 \leq u_k \leq M, k = 0, 1, \dots, T - 1\},$$

with  $M$  the maximum control effort.

The model is, for  $k = 0, 1, \dots, T - 1$  and given  $x_0, y_0, z_0$ ,

$$\begin{aligned} x_{k+1} &= x_k + r x_k(1 - x_k) - c_1 x_k y_k, \\ y_{k+1} &= d y_k + c_2 x_k y_k - c_3 y_k z_k, \\ z_{k+1} &= z_k - m z_k + c_4 y_k z_k + u_k z_k, \end{aligned} \tag{2-1}$$

where  $r$  and  $d$  are the intrinsic growth rates for the valuable population and pest population respectively,  $m$  is the death rate of the predator (biological control agent), the constants  $c_i, i = 1, \dots, 4$  are the interaction coefficients between the species. We apply the control  $u_k$  to increase the growth rate of the predator at each time step, for example, we can import the natural enemies of the pest or supplement the existing predators.

The goal is to maximize

$$\sum_{k=0}^{T-1} B_1 x_k - B_2 y_k - B_3 z_k - \frac{1}{2} A u_k^2 \tag{2-2}$$

over  $u \in U_1$ , with  $A > 0, B_i > 0, i = 1, 2, 3$  constants; that is, we want to maximize the valuable population while minimizing the pest population and the cost of applying the biological control, we also minimize the predator population

for environmental consideration over the entire time period. We choose a quadratic cost for simplicity and other forms could be treated.

We will use the extension of Pontryagin's maximum principle (PMP) [Lenhart and Workman 2007; Pontryagin et al. 1962; Sethi and Thompson 2006] for the optimal control of discrete system. The technique involves the use of adjoint functions, which append the discrete system (2-1) to the maximization of the objective functional (2-2). PMP gives the optimality system of difference equations consisting of the state and adjoint difference equations coupled with the control characterization. Note that the adjoint equations have final time boundary conditions while the state equations have initial conditions. The key idea is that the adjoint method provides us with the gradient of the cost function needed for the maximization procedure. We note that an optimal control exists due to the finite dimensional structure of this system.

Applying the extension of Pontryagin's maximum principle for discrete systems [Lenhart and Workman 2007; Pontryagin et al. 1962; Sethi and Thompson 2006], we form the Hamiltonian:

$$H_k = B_1 x_k - B_2 y_k - B_3 z_k - \frac{1}{2} A u_k^2 + \lambda_{1,k+1} (x_k + r x_k (1 - x_k) - c_1 x_k y_k) + \lambda_{2,k+1} (d y_k + c_2 x_k y_k - c_3 y_k z_k) + \lambda_{3,k+1} (z_k - m z_k + c_4 y_k z_k + u_k z_k), \quad (2-3)$$

which is used to derive the necessary conditions in the next theorem.

**Theorem 2.1.** *Given an optimal control  $u^* \in U_1$  and the corresponding states  $x^*, y^*, z^*$  from (2-1), there exist adjoint functions  $\lambda_i, i = 1, 2, 3$  satisfying:*

$$\begin{aligned} \lambda_{1,k} &= B_1 + \lambda_{1,k+1} (1 + r - 2r x_k^* - c_1 y_k^*) + \lambda_{2,k+1} c_2 y_k^*, \\ \lambda_{2,k} &= -B_2 - \lambda_{1,k+1} c_1 x_k^* + \lambda_{2,k+1} (d + c_2 x_k^* - c_3 z_k^*) + \lambda_{3,k+1} c_4 z_k^*, \\ \lambda_{3,k} &= -B_3 - \lambda_{2,k+1} c_3 y_k^* + \lambda_{3,k+1} (1 - m + c_4 y_k^* + u_k^*), \\ \lambda_{1,T} &= \lambda_{2,T} = \lambda_{3,T} = 0. \end{aligned} \quad (2-4)$$

Furthermore, the characterization of  $u_k^*$  is

$$u_k^* = \min\{\max\{\lambda_{3,k+1} z_k^* / A, 0\}, M\}. \quad (2-5)$$

*Proof.* Using the extension of Pontryagin's maximum principle for discrete systems [Lenhart and Workman 2007; Pontryagin et al. 1962; Sethi and Thompson 2006], we have

$$\begin{aligned} \lambda_{1,k} &= \frac{\partial H_k}{\partial x_k} = B_1 + \lambda_{1,k+1} (1 + r - 2r x_k^* - c_1 y_k^*) + \lambda_{2,k+1} c_2 y_k^*, \\ \lambda_{2,k} &= \frac{\partial H_k}{\partial y_k} = -B_2 - \lambda_{1,k+1} c_1 x_k^* + \lambda_{2,k+1} (d + c_2 x_k^* - c_3 z_k^*) + \lambda_{3,k+1} c_4 z_k^*, \\ \lambda_{3,k} &= \frac{\partial H_k}{\partial z_k} = -B_3 - \lambda_{2,k+1} c_3 y_k^* + \lambda_{3,k+1} (1 - m + c_4 y_k^* + u_k^*). \end{aligned} \quad (2-6)$$

In addition, the transversality conditions are

$$\lambda_{1,T} = \lambda_{2,T} = \lambda_{3,T} = 0.$$

Using

$$\frac{\partial H_k}{\partial u_k} = -Au_k + \lambda_{3,k+1}z_k,$$

and  $\partial H_k/\partial u_k = 0$  at  $u^*$  on the interior of the control set, we have the control characterization

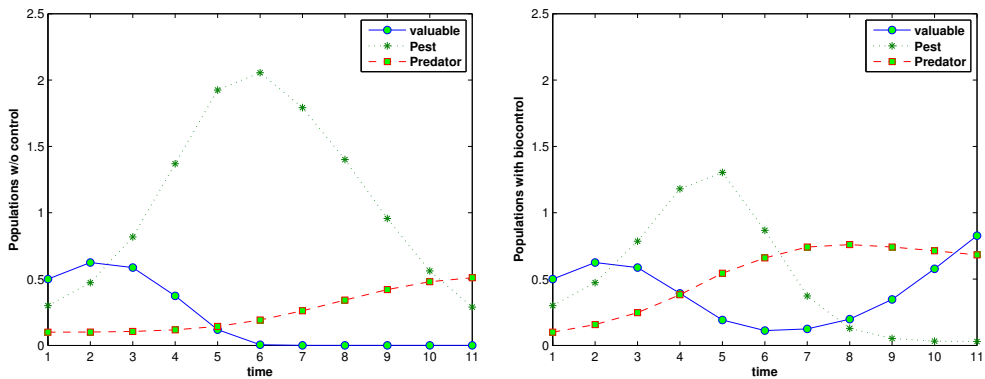
$$u_k^* = \min\{\max\{\lambda_{3,k+1}z_k^*/A, 0\}, M\}. \tag{2-7}$$

This concludes the proof. □

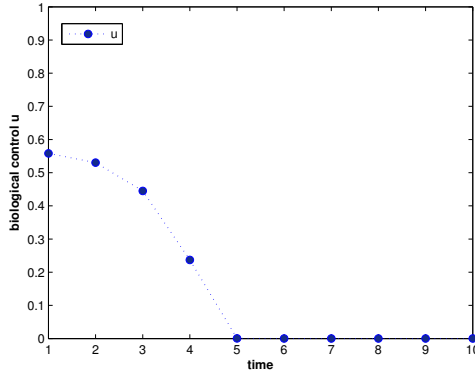
The optimality system consists of the state equations (2-1) with initial conditions and adjoint equations (2-4) with the final time conditions and with the characterization of the optimal control (2-5).

**2.2. Numerical results.** To solve the optimal biological control problem numerically, due to the boundary conditions being at the initial time for the states and at the final time for adjoints, an iterative method is used to solve this optimality system. Given initial guesses for the control and the state equations, the state system (2-1) is solved forward in time, and the adjoint system (2-4) is solved backward in time. The control is updated using the characterization (2-7) with the newly found state and adjoint values, and the iteration repeats until convergence occurs. See [Lenhart and Workman 2007] for details of this method.

Figure 1 (left) gives the valuable, pest and predator populations without the application of the control. Without the control, the pest population increases, killing off the valuable population, for  $r = d = 1.1$ ,  $m = 0.05$ ,  $c_1 = 1$ ,  $c_2 = c_3 = 1.2$ ,



**Figure 1.** Valuable, pest, predator populations without biological control (left) and with biological control (right;  $A = 5$ ).



**Figure 2.** Optimal biological control,  $A = 5$ .

$c_4 = 0.2$ . In contrast, with the biological control, the growth of the pest population decreases, allowing the valuable population to grow quickly; see Figure 1 (right). Figure 2 gives the optimal biocontrol result for  $A = 5$ ,  $M = 1$ ,  $B_i = 1$ ,  $i = 1, 2, 3$ . We see the optimal biological control effort is gradually decreasing and we don't apply it after time step 5.

### 3. Optimal control using dual control

**3.1. The dual control problem.** Now we attempt to control the pest population using both biological control and pesticide at the same time.

The controls satisfy

$$U_2 = \{(u_{i,0}, u_{i,1}, \dots, u_{i,T-1}) \in \mathbb{R}^T \mid 0 \leq u_{i,k} \leq M, i = 1, 2, k = 0, 1, \dots, T - 1\},$$

with  $M$  the maximum control effort.

The model is, for  $k = 0, 1, \dots, T - 1$  and given  $x_0, y_0, z_0$ ,

$$\begin{aligned} x_{k+1} &= x_k + rx_k(1 - x_k) - c_1x_ky_k, \\ y_{k+1} &= dy_k + c_2x_ky_k - c_3y_kz_k - u_{1,k}y_k, \\ z_{k+1} &= z_k - mz_k + c_4y_kz_k + u_{2,k}z_k, \end{aligned} \tag{3-1}$$

where the control  $u_{1,k}$  is the pesticide and we also apply the control  $u_{2,k}$  to increase the growth rate of the predator (biological control agent) at each time step.

The goal is to maximize

$$\sum_{k=0}^{T-1} B_1x_k - B_2y_k - B_3z_k - \frac{1}{2}A_1u_{1,k}^2 - \frac{1}{2}A_2u_{2,k}^2, \tag{3-2}$$

with  $B_i > 0$ ,  $i = 1, 2, 3$ ,  $A_j > 0$ ,  $i = 1, 2$ ; that is, we want to maximize the valuable population while minimizing the pest population and the cost of applying



the pesticide and biological control, we also minimize the predator population for environmental purpose over the entire time period.

Applying the extension of Pontryagin’s maximum principle for discrete systems [Lenhart and Workman 2007; Pontryagin et al. 1962; Sethi and Thompson 2006], we form the Hamiltonian:

$$\begin{aligned}
 H_k = & B_1x_k - B_2y_k - B_3z_k - \frac{1}{2}A_1u_{1,k}^2 - \frac{1}{2}A_2u_{2,k}^2 \\
 & + \lambda_{1,k+1}(x_k + rx_k(1 - x_k) - c_1x_ky_k) \\
 & + \lambda_{2,k+1}(dy_k + c_2x_ky_k - c_3y_kz_k - u_{1,k}y_k) \\
 & + \lambda_{3,k+1}(z_k - mz_k + c_4y_kz_k + u_{2,k}z_k), \quad (3-3)
 \end{aligned}$$

which is used to derive the necessary conditions in the next theorem.

**Theorem 3.1.** *Given optimal controls  $u_i^* \in U_2, i = 1, 2$  and the corresponding states  $x^*, y^*, z^*$  from (3-1), there exist adjoint functions  $\lambda_i, i = 1, 2, 3$  satisfying*

$$\begin{aligned}
 \lambda_{1,k} = & B_1 + \lambda_{1,k+1}(1 + r - 2rx_k^* - c_1y_k^*) + \lambda_{2,k+1}c_2y_k^*, \\
 \lambda_{2,k} = & -B_2 - \lambda_{1,k+1}c_1x_k^* + \lambda_{2,k+1}(d + c_2x_k^* - c_3z_k^* - u_{1,k}^*) + \lambda_{3,k+1}c_4z_k^*, \\
 \lambda_{3,k} = & -B_3 - \lambda_{2,k+1}c_3y_k^* + \lambda_{3,k+1}(1 - m + c_4y_k^* + u_{2,k}^*), \\
 \lambda_{1,T} = & \lambda_{2,T} = \lambda_{3,T} = 0.
 \end{aligned} \quad (3-4)$$

Furthermore, the characterizations of  $u_{1,k}^*, u_{2,k}^*$  are

$$\begin{aligned}
 u_{1,k}^* = & \min\{\max\{-\lambda_{2,k+1}y_k^*/A_1, 0\}, M\}, \\
 u_{2,k}^* = & \min\{\max\{\lambda_{3,k+1}z_k^*/A_2, 0\}, M\}.
 \end{aligned} \quad (3-5)$$

*Proof.* Using the extension of Pontryagin’s maximum principle for discrete systems [Lenhart and Workman 2007; Pontryagin et al. 1962; Sethi and Thompson 2006], we have

$$\begin{aligned}
 \lambda_{1,k} = & \frac{\partial H_k}{\partial x_k} = B_1 + \lambda_{1,k+1}(1 + r - 2rx_k^* - c_1y_k^*) + \lambda_{2,k+1}c_2y_k^*, \\
 \lambda_{2,k} = & \frac{\partial H_k}{\partial y_k} = -B_2 - \lambda_{1,k+1}c_1x_k^* + \lambda_{2,k+1}(d + c_2x_k^* - c_3z_k^* - u_{1,k}^*) + \lambda_{3,k+1}c_4z_k^*, \\
 \lambda_{3,k} = & \frac{\partial H_k}{\partial z_k} = -B_3 - \lambda_{2,k+1}c_3y_k^* + \lambda_{3,k+1}(1 - m + c_4y_k^* + u_{2,k}^*).
 \end{aligned}$$

In addition, the transversality conditions are

$$\lambda_{1,T} = \lambda_{2,T} = \lambda_{3,T} = 0. \quad (3-6)$$

Using

$$\frac{\partial H_k}{\partial u_{1,k}} = -A_1 u_{1,k} - \lambda_{2,k+1} y_k,$$

and  $\partial H_k / \partial u_{1,k} = 0$  at  $u_1^*$  on the interior of the control set, we have the control characterization

$$u_{1,k}^* = \min\{\max\{-\lambda_{2,k+1} y_k^* / A_1, 0\}, M\}. \tag{3-7}$$

And using

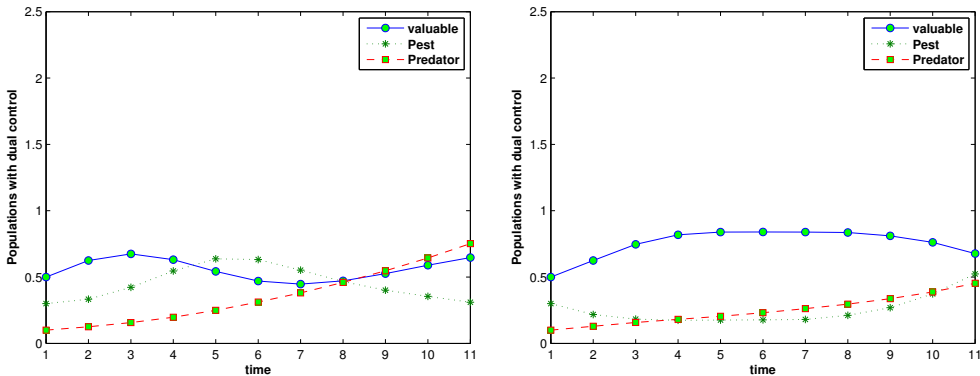
$$\frac{\partial H_k}{\partial u_{2,k}} = -A_2 u_{2,k} + \lambda_{3,k+1} z_k,$$

and  $\partial H_k / \partial u_{2,k} = 0$  at  $u_2^*$  on the interior of the control set, we have the control characterization

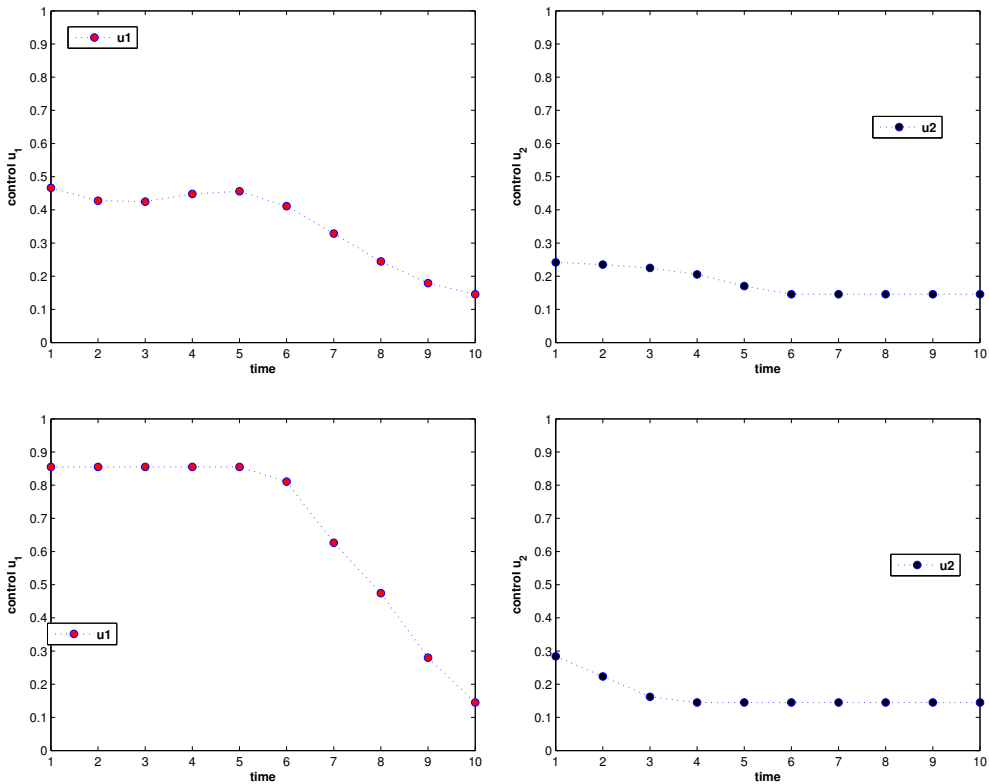
$$u_{2,k}^* = \min\{\max\{\lambda_{3,k+1} z_k^* / A_2, 0\}, M\}. \tag{3-8}$$

This concludes the proof. □

**3.2. Numerical results and conclusion.** In this section, we apply dual control, combining pesticide and biological control. Figure 3 shows the significant increase in valuable population and decrease in pest population after applying dual control, and maintaining the predator at a reasonable level. Figure 3 (left) gives the result for  $A_1 = A_2 = 5$  and Figure 3 (right) gives the result for  $A_1 = A_2 = 1$ , with all the other parameters kept the same with Section 2.2. We vary the cost coefficients  $A_1, A_2$  to see the effect on the populations and the optimal control strategy. With the lower cost coefficients  $A_i, i = 1, 2$ , we can apply more pesticide and biological control, and the pest population can be reduced to a lower level; see Figures 3 and



**Figure 3.** Valuable, pest, predator populations with dual control. Left:  $A_1 = 5, A_2 = 5$ . Right:  $A_1 = 1, A_2 = 1$ .



**Figure 4.** Optimal dual control, different  $A_1$  and  $A_2$ . Top: pesticide,  $A_1 = 5$ ,  $A_2 = 5$  (left); biological control,  $A_1 = 5$ ,  $A_2 = 5$  (right). Bottom: pesticide,  $A_1 = 1$ ,  $A_2 = 1$  (left); biological control,  $A_1 = 1$ ,  $A_2 = 1$  (right).

4. We note that lowering the cost of both controls provides a substantial increase in pesticide use and only a modest increase in the use of biological control.

We also compare the result of biological control and dual control. We see from Figures 1 (right) and 3 that dual control gives better results for maintaining the valuable population and reducing the pest population, while keeping the predator at a low level.

In summary, we give a theoretical framework using discrete time optimal control theory for pest control problems and provide the numerical results. We apply the biological control and the combination of the pesticide and biological control (dual control) to find the optimal strategy. The results provide suggestions in the design of appropriate control strategies and assist management decision-making.

We should note that in our models (2-1) and (3-1), the order of events is that population growth occurs first, then it is increased/decreased by interactions with

other species or through human intervention. We can explore other order of events since Bodine et al. [2012] point out for discrete models different order of events can lead to qualitatively different optimal control strategies.

### Acknowledgements

The work was supported by NSF Stepping Up Undergraduate Research Summer Program at Middle Tennessee State University.

### References

- [Bodine et al. 2012] E. N. Bodine, L. J. Gross, and S. Lenhart, “Order of events matter: comparing discrete models for optimal control of species augmentation”, *J. Biol. Dyn.* **6**:suppl. 2 (2012), 31–49. MR 2994278
- [Clark 1990] C. W. Clark, *Mathematical bioeconomics: the optimal management of renewable resources*, 2nd ed., Wiley, New York, 1990. MR 91c:90037
- [Dabbs 2010] K. Dabbs, *Optimal control in discrete pest control models*, Ph.D. thesis, University of Tennessee, 2010, Available at <http://goo.gl/2Cp8ul>.
- [Driesche 1994] R. G. V. Driesche, “Classical biological control of environmental pests”, *Florida Entomologist* **77** (1994), 20–33.
- [Driesche et al. 2010] R. G. V. Driesche, R. I. Carruthers, T. Center, M. S. Hoddle, J. Hough-Goldstein, L. Morin, L. Smith, D. L. Wagner, B. Blossey, V. Brancatini, R. Casagrande, C. E. Causton, J. A. Coetzee, J. Cuda, J. Ding, S. V. Fowler, J. H. Frank, R. Fuester, J. Goolsby, M. Grodowitz, T. A. Heard, M. P. Hill, J. H. Hoffmann, J. Hubert, M. Julien, M. T. K. Kairo, M. Kenis, P. Mason, J. Medal, R. Messing, R. Miller, A. Moore, P. Neuenschwander, R. Newman, H. Norambuena, W. A. Palmer, R. Pemberton, A. P. Panduro, P. D. Pratt, M. Rayamajhi, S. Salom, D. Sands, S. Schooler, M. Schwarzländer, A. Sheppard, R. Shaw, P. W. Tipping, and R. van Klinken, “Classical biological control for the protection of natural ecosystems”, *Biological Control* **54** (2010), S2–S33.
- [Eilenberg et al. 2001] J. Eilenberg, A. Hajek, and C. Lomer, “Suggestions for unifying the terminology in biological control”, *BioControl* **46** (2001), 387–400.
- [Hawkins and Cornell 1999] B. Hawkins and H. Cornell, *Theoretical approaches to biological control*, Cambridge University Press, 1999.
- [Jang and Yu 2012] S. Jang and J. Yu, “Discrete-time host-parasitoid models with pest control”, *Journal of Biological Dynamics* **6**:2 (2012), 718–739.
- [Lenhart and Workman 2007] S. Lenhart and J. T. Workman, *Optimal control applied to biological models*, Chapman & Hall/CRC, Boca Raton, FL, 2007. MR 2008f:49001
- [van Lenteren and Woets 1988] J. C. van Lenteren and J. Woets, “Biological and integrated pest control in greenhouses”, *Ann. Rev. Entomol* **33** (1988), 239–269.
- [Pontryagin et al. 1962] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The mathematical theory of optimal processes*, Wiley Interscience, New York, 1962. MR 29 #3316b
- [Sethi and Thompson 2006] S. P. Sethi and G. L. Thompson, *Optimal control theory: applications to management science and economics*, 2nd ed., Kluwer, Boston, 2006. MR 84g:49002
- [Tang and Cheke 2008] S. Tang and R. A. Cheke, “Models for integrated pest control and their biological implications”, *Math. Biosci.* **215**:1 (2008), 115–125. MR 2009i:92099
- [Whittle et al. 2007] A. J. Whittle, S. Lenhart, and L. J. Gross, “Optimal control for management of an invasive plant species”, *Math. Biosci. Eng.* **4**:1 (2007), 101–112. MR 2007h:92094

Received: 2012-10-05

Revised: 2013-11-11

Accepted: 2013-12-20

wandi.ding@mtsu.edu

*Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, TN 37132, United States*

rch3g@mtmail.mtsu.edu

*Computational Science Program, Middle Tennessee State University, Murfreesboro, TN 37132, United States*

bcathey1@utk.edu

*Department of Physics and Astronomy, University of Tennessee, Knoxville, Knoxville, TN 37996, United States*

evan.lancaster@mtsu.edu

*Department of University Studies, Middle Tennessee State University, Murfreesboro, TN 37132, United States*

rgermick@utk.edu

*Department of Electrical Engineering, University of Tennessee, Knoxville, TN 37996, United States*



# Distribution of genome rearrangement distance under double cut and join

Jackie Christy, Josh McHugh, Manda Riehl and Noah Williams

(Communicated by Anant Godbole)

Using the double-cut-and-join (DCJ) model for genome rearrangement we use combinatorial techniques to analyze the distribution of genomes under DCJ distance. We present an exponential generating function for the number of genomes that are maximally distant from a given genome and provide a formula for the number of genomes that are any given distance from an arbitrary starting genome.

## 1. Introduction

Many mathematical models have been developed to aid biologists and bioinformaticians in their study of the genome rearrangement problem, whose goal is to find the optimal sequence of mutations for the transformation of one genome into another. Using the double-cut-and-join (DCJ) model, Bergeron, Mixtacki, and Stoye [Bergeron et al. 2006] found that the distance between two genomes is completely determined by a bipartite graph created from the genomes. We utilize their data structure to find the distribution of genomes that are distance  $d$  from a given genome under DCJ. In Section 2, we introduce genome rearrangement, DCJ, and an important result of the same authors. In Section 3, we present a generating function for the number of maximally distant genomes from a given genome, and in Section 4, we obtain the distribution of all genomes by distance from a given genome.

## 2. Background

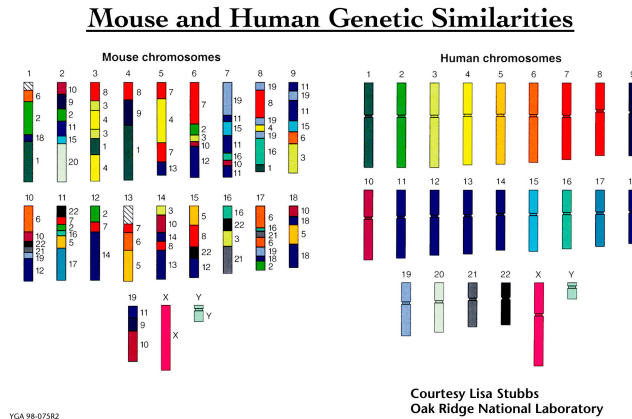
***A brief history of genome rearrangements.*** Deoxyribonucleic acid (DNA) contains instructions for the creation of the proteins necessary for the development and

---

*MSC2010:* 05E05, 68R15.

*Keywords:* genome rearrangement, double cut and join, generating function.

The authors received support for this work from UWEC's Office of Research and Sponsored Programs. Williams was also supported by the UWEC Foundation.



**Figure 1.** Preserved segments between mouse and human genomes showing long stretches of conserved DNA from their common ancestor. More than ninety percent of the mouse genome consists of shuffled pieces of the human genome [NHGRI 2002].

survival of living organisms. The entire collection of DNA in an organism is called the organism's genome, and this DNA is contained within chromosomes comprised of genes. When DNA is replicated, occasionally something goes awry and a mutation occurs, slightly changing an organism's genetic make-up. A sufficient number of mutations can result in death, disease, or the development of a new species.

In the genome rearrangement problem, the object is to find the optimal sequence of mutations that transforms one genome into another, where both genomes are defined on the same set of genes. The number of mutations in this most efficient scenario is defined to be the *distance* between the two genomes.

In the simplest case, genomes can be modeled by permutations under the assumptions that all genomes share the same set of genes, there are no duplicated genes, and only a single chromosome is considered [Fertin et al. 2009]. Most models now use objects that are more complicated than permutations by removing some or all of these assumptions [Yancopoulos et al. 2005]. For example, signed permutations are utilized to better model that DNA is oriented, and ordered set partitions can be used for multiple chromosomes.

**Double cut and join.** In the DCJ model, genes are numbered and oriented, as shown in the figures on the next page. Consequently, a gene may be represented as a numbered left or right arrow with labeled ends; for example,  $7h$  and  $7t$  denote the head and tail of the seventh gene. Chromosomes are collections of arrows that have



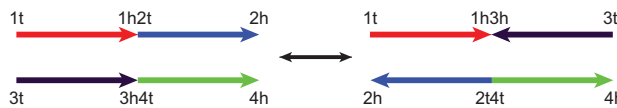
been joined head to head, tail to tail, or head to tail, and genomes constitute sets of chromosomes. Alternatively, a genome may be represented as a collection of vertices that correspond to the locations where genes meet. An *internal vertex*, or *adjacency*, occurs where two genes are joined in one of the three fashions mentioned above, and an *external vertex*, or *telomere*, occurs where the head or tail of a gene is not connected to other genes. Note that there are always an even number of telomeres in a genome, and that the number of genes in a genome is equivalent to the sum of the adjacencies and the number of pairs of telomeres present.

DCJ is a broad model that encompasses linear and circular chromosomes and incorporates the following mutations:

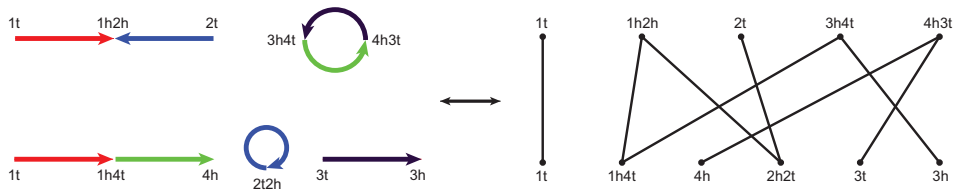
- Inversions: reverse the order of a chromosome or part of the genome
- Interchanges: switch two segments of the genome
- Translocations: swap the ends of two chromosomes
- Circularizations and linearizations: convert between linear and circular chromosomes.

A DCJ operation involves making two cuts in a genome and rejoining the pieces in one of the following ways:

- Two internal vertices  $\{a, b\}$  and  $\{c, d\}$  can be replaced with two new internal vertices  $\{a, d\}$  and  $\{b, c\}$  or  $\{a, c\}$  and  $\{b, d\}$ . See Figure 2.



**Figure 2.** An illustration of the first type of mutation allowed under DCJ. The genome at right is obtained by replacing internal vertices  $\{1h, 2t\}$  and  $\{3h, 4t\}$  in the leftmost genome with internal vertices  $\{1h, 3h\}$  and  $\{2t, 4t\}$ .



**Figure 3.** The bipartite adjacency graph constructed from two multi-chromosomal genomes.

- ii. An internal vertex  $\{a, b\}$  and an external vertex  $\{c\}$  can be replaced with new internal and external vertices  $\{a, c\}$  and  $\{b\}$  or  $\{b, c\}$  and  $\{a\}$ .
- iii. Two external vertices  $\{a\}$  and  $\{b\}$  can be replaced by an internal vertex  $\{a, b\}$ .
- iv. An internal vertex  $\{a, b\}$  can be replaced by two external vertices  $\{a\}$  and  $\{b\}$ .

Any genome can be represented by a distinct arrangement of adjacencies and telomeres. Bergeron, Mixtacki, and Stoye found that the DCJ distance between two genomes is completely determined by a bipartite graph whose vertices correspond to the sets of adjacencies and telomeres of the two genomes. In this graph, two vertices are connected with an edge for every head or tail that they share (see Figure 3). The DCJ distance between the genomes can be determined based on the number of cycles and odd-length paths in this graph.

**Theorem 1** [Bergeron et al. 2006]. *The DCJ distance between two genomes,  $A$  and  $B$ , defined on the same set of  $N$  genes, is given by*

$$d_{DCJ}(A, B) = N - (C + I/2),$$

where  $C$  is the number of cycles and  $I$  is the number of odd-length paths in the adjacency graph of  $A$  and  $B$ .

Consider Figure 3, which depicts two genomes and their adjacency graph. Notice that the adjacency  $\{3h, 4t\}$  in the first genome is connected to the adjacency  $\{1h, 4t\}$  and the telomere  $\{3h\}$  in the second genome. Using Theorem 1, we can calculate that the DCJ distance is  $4 - (0 + 2/2) = 3$  because there are no cycles and two odd-length paths in the adjacency graph. The following sequence of three DCJ operations demonstrates one way that the first genome may be transformed into the second genome using the fewest number of mutations.

Operation 1: Replace internal vertices  $\{1h, 2h\}$  and  $\{3h, 4t\}$  with internal vertices  $\{1h, 4t\}$  and  $\{2h, 3h\}$ ; see DCJ operation i. This is a linearization and an insertion.

Operation 2: Exchange internal vertex  $\{2h, 3h\}$  and external vertex  $\{2t\}$  for internal vertex  $\{2h, 2t\}$  and external vertex  $\{3h\}$ ; see DCJ operation ii. This constitutes a translocation and a circularization.

Operation 3: Replace internal vertex  $\{4h, 3t\}$  with external vertices  $\{4h\}$  and  $\{3t\}$ ; see DCJ operation iv. This models a translocation.

### 3. Counting maximally distant genomes

Building on the result of Theorem 1, we observed that the maximum distance between two genomes defined on  $N$  genes is  $N$  and occurs when  $C + I/2 = 0$ . This means that there are no cycles and no odd-length paths in the adjacency graph of two maximally distant genomes. We established the following result by considering

an arbitrary starting genome defined on  $N$  genes and counting the number of distinct adjacency graphs that could be created from it, where each adjacency graph contained only even-length paths.

**Theorem 2.** *The number of genomes that are the maximum DCJ distance away from a genome containing  $2m$  telomeres and  $n$  adjacencies is given by*

$$G_{\max}(m, n) = (2m - 1)!! \sum_{k=0}^n \binom{m+n-1}{k} \binom{n}{k} k! 2^k.$$

*Proof.* We count the number of distinct adjacency graphs that contain exclusively even-length paths, where the upper genome contains  $m$  pairs of telomeres and  $n$  adjacencies. In this proof and in subsequent proofs, we refer to upper and lower vertices as those adjacencies and telomeres located in the upper and lower genomes, respectively. Consider the following procedure:

1. Sum over the number of even-length paths  $j$ . The minimum number is  $m$  since each even-length path may contain no more than one pair of upper telomeres. The maximum number of even-length paths is  $m + n$  since each pair of upper telomeres and each upper adjacency may be in a path of length two. Note that if we let  $k = m + n - j$ , then the sum from  $j = m$  to  $m + n$  becomes the sum from  $k = 0$  to  $n$ .
2. Place the  $2m$  upper telomeres into pairs. Each pair will define the endpoints of an even-length path. There are  $(2m - 1)(2m - 3) \cdots 3 \cdot 1 = (2m - 1)!!$  ways to accomplish this.
3. Arrange the upper adjacencies into the order that they will appear in the even-length paths. This can be done in  $n!$  ways. The next step will involve partitioning these adjacencies into paths.
4. The even-length paths are constructed in the following way. We begin by partitioning the  $n$  upper adjacencies into  $j$  even-length paths. Since each path must be nonempty, one adjacency must be placed into each of the  $j - m$  paths without upper telomeres. The number of ways to do this is

$$\binom{(n-j+m)+(j-1)}{j-1} = \binom{m+n-1}{j-1}.$$

In addition, once all of the upper vertices have been arranged and assigned to paths, there are two choices for how to connect each upper adjacency with its neighbors on its path. (For instance, if a path contains ordered upper vertices  $\{1t\}, \{1h, 2t\}, \{2h\}$ , there are two possibilities for the lower adjacencies in between:  $\{1t, 1h\}, \{2t, 2h\}$  and  $\{1t, 2t\}, \{1h, 2h\}$ .) To accomplish this, we multiply by two for every upper adjacency except for those in paths of length two.

We have overcounted since the even-length paths we are creating are non-directed and the upper adjacencies in each non-upper-telomere-containing path can be in right-to-left or left-to right-order (except of course for paths of length two). For the upper-telomere-containing paths, this ordering of the upper adjacencies is significant because it determines which telomere is adjacent to which adjacency along the path. Hence, we divide by two for every non-upper-telomere-containing even-length path of length greater than two. Since the number of non-upper-telomere-containing paths of length two is equivalent to the number of upper adjacencies in even-length paths of length two, we multiply by  $2^{m+n-j}$ .

We have overcounted further since the paths that do not contain upper telomeres are not distinct. To resolve this situation, we divide by  $(j - m)!$ .

Combining these four steps yields

$$G_{\max}(m, n) = \sum_{j=m}^{m+n} (2m - 1)!! n! \binom{m+n-1}{j-1} \frac{2^{m+n-j}}{(j - m)!}. \tag{1}$$

By defining  $k = m + n - j$ , and rearranging the summation above, we obtain

$$G_{\max}(m, n) = (2m - 1)!! \sum_{k=0}^n \binom{m+n-1}{k} \binom{n}{k} k! 2^k.$$

Alternatively, if we define  $k = j - m$  in (1), we have

$$G_{\max}(m, n) = (2m - 1)!! \sum_{k=0}^n \binom{m+n-1}{n-k} \frac{n!}{k!} 2^{n-k}, \tag{2}$$

which is useful in simplifying the formula of Theorem 7 below. □

Next, fix  $m$ , and consider the collection of genomes having a  $2m$  telomeres and a variable number of adjacencies  $n$ . For such a collection, we obtain an infinite sequence  $\{g_m^n\}$  over  $n$ , where each term represents the number of maximally distant genomes from a genome having  $2m$  telomeres and  $n$  adjacencies. For example, the sequence associated with a genome containing two pairs of external vertices is

$$3, 15, 111, 1083, 13083, \dots, g_2^n, \dots$$

where  $g_2^n$  is given by  $G_{\max}(2, n)$  and represents the number of maximally distant genomes from a starting genome with two pairs of external vertices and  $n$  internal vertices. We now find the exponential generating function for this sequence.

**Lemma 3.** 
$$\left(\frac{1}{1 - 2x}\right)^{m+n} = \sum_{j=0}^{\infty} \binom{m+n+j-1}{j} (2x)^j.$$

*Proof.* We have  $\frac{1}{1-2x} = \sum_{i=0}^{\infty} (2x)^i$  and hence

$$\begin{aligned} \left(\frac{1}{1-2x}\right)^{m+n} &= \left(\sum_{i=0}^{\infty} (2x)^i\right)^{m+n} \\ &= \underbrace{(1+2x+\dots+(2x)^i+\dots)\cdots(1+2x+\dots+(2x)^i+\dots)}_{m+n \text{ terms}}. \end{aligned}$$

Next, consider this multiplication in a combinatorial sense where the resulting product, an infinite series, is formed term by term and where each term is the product of  $m+n$  elements, one coming from each initial series. When adding these terms, consider the coefficient of  $x^j$ . Using a bijection to a familiar problem of placing  $j$  balls into  $m+n$  bins, one can count the number of terms having degree  $j$ , and then,  $2^j$  can be factored from each term. Thus, the coefficient of  $(2x)^j$  is simply the number of terms having degree  $j$  and is expressed by

$$\binom{m+n+j-1}{j}.$$

Furthermore, the sum of all  $x^j$  and their coefficients is equivalent to the product of the  $m+n$  series. That is,

$$\begin{aligned} \underbrace{(1+2x+\dots+(2x)^i+\dots)\cdots(1+2x+\dots+(2x)^i+\dots)}_{m+n \text{ terms}} \\ = \sum_{j=0}^{\infty} \binom{m+n+j-1}{j} (2x)^j. \end{aligned}$$

Thus,

$$\left(\frac{1}{1-2x}\right)^{m+n} = \sum_{j=0}^{\infty} \binom{m+n+j-1}{j} (2x)^j. \quad \square$$

**Theorem 4.** *The exponential generating function for the sequence  $\{g_m^n\}$  is*

$$g_m(x) = \left(\frac{(2m-1)!}{2^{m-1}(m-1)!}\right) \frac{e^{\frac{x}{1-2x}}}{(1-2x)^m},$$

where the  $n$ -th term of the sequence  $\{g_m^n\}$ , or  $G_{\max}(m, n)$ , is given by  $g_m^{(n)}(0)$ .

*Proof.* We have

$$\frac{e^{\frac{x}{1-2x}}}{(1-2x)^m} = \sum_{n=0}^{\infty} \frac{\left(\frac{x}{1-2x}\right)^n}{n! (1-2x)^m} = \sum_{n=0}^{\infty} \frac{x^n}{n!} \left(\frac{1}{1-2x}\right)^{m+n}.$$

Using Lemma 3, we obtain

$$\frac{e^{\frac{x}{1-2x}}}{(1-2x)^m} = \sum_{n=0}^{\infty} \left( \frac{x^n}{n!} \sum_{j=0}^{\infty} \binom{m+n+j-1}{j} (2x)^j \right).$$

Expanding this series yields

$$\begin{aligned} & \frac{e^{\frac{x}{1-2x}}}{(1-2x)^m} \\ &= \binom{m-1}{0} + \binom{m}{1}(2x) + \dots + \binom{m+j-1}{j}(2x)^j + \dots \\ &+ \binom{m}{0}x + \binom{m+1}{1}(2x)x + \dots + \binom{m+j}{j}(2x)^jx + \dots \\ &+ \dots \\ &+ \binom{m+n-1}{0} \frac{x^n}{n!} + \binom{m+n}{1}(2x) \frac{x^n}{n!} + \dots + \binom{m+n+j-1}{j}(2x)^j \frac{x^n}{n!} + \dots \end{aligned}$$

By looking at the coefficient of each power of  $x$ , the following infinite series is created (consider a diagonal argument).

$$\begin{aligned} \frac{e^{\frac{x}{1-2x}}}{(1-2x)^m} &= \sum_{n=0}^{\infty} \left( x^n \sum_{k=0}^n \binom{m+n-1}{k} \frac{2^k}{(n-k)!} \right) \\ &= \sum_{n=0}^{\infty} \left( x^n \frac{n!}{n!} \sum_{k=0}^n \binom{m+n-1}{k} \frac{k!}{k!} \frac{2^k}{(n-k)!} \right) \\ &= \sum_{n=0}^{\infty} \left( x^n \frac{1}{n!} \sum_{k=0}^n \binom{m+n-1}{k} \frac{n! k! 2^k}{k! (n-k)!} \right) \\ &= \sum_{n=0}^{\infty} \left( \frac{x^n}{n!} \sum_{k=0}^n \binom{m+n-1}{k} \binom{n}{k} k! 2^k \right). \end{aligned}$$

Thus,

$$\begin{aligned} & \left( \frac{(2m-1)!}{2^{m-1}(m-1)!} \right) \frac{e^{\frac{x}{1-2x}}}{(1-2x)^m} \\ &= \sum_{n=0}^{\infty} \left( \frac{x^n}{n!} \left( \frac{(2m-1)!}{2^{m-1}(m-1)!} \right) \sum_{k=0}^n \binom{m+n-1}{k} \binom{n}{k} k! 2^k \right). \quad \square \end{aligned}$$

### 4. Distribution of DCJ distance

To understand the way in which distance from a given genome is distributed across all genomes, we count the total number of genomes that are each distance away from an arbitrary genome. Employing Theorem 1, we recognize that a destination

genome is distance  $d$  away from a starting genome precisely when there are a total of  $N - d$  cycles and pairs of odd-length paths in the adjacency graph between the two genomes. Consequently, we count the number of distinct adjacency graphs we can construct from a given starting genome that include exactly  $N - d$  cycles and pairs of odd-length paths.

**Lemma 5.** *The number of ways to arrange  $2p$  upper telomeres and  $k$  upper adjacencies into distinct adjacency graphs that contain exclusively odd-length paths is*

$$k! \binom{2p+k-1}{k} 2^k.$$

*Proof.* Consider the following counting procedure.

1. We begin by arranging the  $k$  upper adjacencies according to the order that they will appear in the odd-length paths. This can be accomplished in  $k!$  ways.
2. Next, partition the  $k$  upper adjacencies into  $2p$  odd-length paths. Each of these paths is distinct because it contains a distinct upper telomere. The number of ways to do this is

$$\binom{2p+k-1}{k}.$$

3. Once all of the upper vertices have been assigned to paths and have been arranged, there are two choices for how to connect each upper adjacency with its neighbors on its path. (For instance, if an odd-length path contains ordered upper vertices  $\{1t\}$  and  $\{1h, 2t\}$ , there are two possibilities for the lower adjacencies in between:  $\{1t, 1h\}$ ,  $\{2t\}$  and  $\{1t, 2t\}$ ,  $\{1h\}$ ). To accomplish this, we multiply by  $2^k$ .

Multiplying these terms yields

$$k! \binom{2p+k-1}{k} 2^k. \quad \square$$

**Lemma 6.** *The number of ways to arrange  $i$  upper internal vertices into distinct adjacency graphs that contain  $q$  cycles and no even-length or odd-length paths is*

$$s(i, q) 2^{i-q},$$

where  $s(a, b)$  are the unsigned Stirling numbers of the first kind.

*Proof.* The unsigned Stirling numbers of the first kind  $s(i, q)$  count the number of permutations of  $i$  elements (the upper adjacencies) into  $q$  disjoint cycles. Note that for this Stirling sequence, the clockwise or counterclockwise orientation of each cycle that contains more than two upper adjacencies is distinct. If we impose a lexicographic ordering of the upper adjacencies in each cycle, the clockwise or counterclockwise orientation of these adjacencies can represent the two ways

in which the adjacency with the smallest value in the lexicographic ordering can connect with its neighbors in the cycle. (Suppose the upper adjacency  $\{1h, 2t\}$  has the smallest value in its cycle with respect to the lexicographic ordering.  $\{1h, 2t\}$  can connect to its left neighbor through a lower adjacency that contains the end  $1h$  or through a lower adjacency that contains the end  $2t$ . The end that  $\{1h, 2t\}$  contributes to the lower adjacency to its right is determined by this choice.)

Once all of the upper adjacencies have been arranged into cycles, there are two choices for how to connect each upper adjacency with its neighbors on its path. (If a path contains ordered upper adjacencies  $\{1h, 2t\}, \{2h, 1t\}, \{3t, 3h\}$ , there are eight possibilities for the lower adjacencies in between. These include  $\{2t, 2h\}, \{1t, 3t\}, \{3h, 1h\}$  and  $\{2t, 1t\}, \{2h, 3t\}, \{3h, 1h\}$  for example.) We have already connected one upper adjacency to its neighbors in each cycle that has more than two upper adjacencies. To connect the others in these cycles, we multiply by two for each additional upper adjacency. For cycles containing exactly one upper adjacency, there is only one way to create the lower adjacency in the cycle. For cycles containing exactly two upper adjacencies, there are two ways to form the two lower adjacencies. (For  $\{1h, 3t\}$  and  $\{2t, 2h\}$  the lower adjacencies could be  $\{1h, 2t\}, \{3t, 2h\}$  or  $\{1h, 2h\}, \{3t, 2t\}$ .)

Hence, we multiply by two for every upper adjacency in a cycle beyond the first upper adjacency in that cycle. Since there are  $q$  cycles, we multiply by  $2^{i-q}$ .

Collecting everything together yields  $s(i, q)2^{i-q}$ . □

Combining Lemmas 6 and 5 and Theorem 2, we establish the following result that classifies all genomes according to their distance from a given genome.

**Theorem 7.** *The number of genomes that are a distance  $d$  away from a starting genome having  $2m$  telomeres and  $n$  adjacencies is*

$$\begin{aligned}
 &G(m, n, d) \\
 &= \sum_{c=\max\{0, n-d\}}^{\min\{n, m+n-d\}} \sum_{i=c}^n \sum_{j=0}^{n-i} \sum_{k=0}^{n-i-j} \frac{s(i, c)n! (2(d+c-n)-1)!!}{i!k!2^{c+k-n}} \binom{2m}{2(d+c-n)} \\
 &\quad \times \binom{2(m+n-d-c)+j-1}{j} \binom{d+c-i-j-1}{n-i-j-k},
 \end{aligned}$$

where  $s(a, b)$  are the unsigned Stirling numbers of the first kind.

*Proof.* We count the number of distinct adjacency graphs from a genome with  $2m$  telomeres and  $n$  adjacencies, where each graph contains a total of  $N - d$  cycles and pairs of odd-length paths (the remaining paths are of even length). Let  $i$  and  $j$  represent the number of upper adjacencies in cycles and in odd-length paths, respectively, and define  $c$  to be the number of cycles in the adjacency graph. Consider the following counting procedure.



1. We begin by summing over the number of cycles  $c$  in the adjacency graph. The minimum number is  $\max\{0, n - d\}$  because the number of cycles must be nonnegative, and we are restricted by the maximum number of odd-length paths that can be formed (recall that the adjacency graph must have  $N - d$  cycles and pairs of odd-length paths). Since each odd-length path must contain an upper telomere, the number of odd-length paths that can be formed is at most  $2m$ . Recall that  $N - d = I/2 + c$ , where  $I$  is the number of odd-length paths in the adjacency graph (Theorem 1). It follows that  $2m$  is the maximum value for  $I$ , and in this case,  $N - d = m + c$ . Substituting  $m + n$  for  $N$  and simplifying yields  $c = n - d$ .

The maximum number of cycles that can be in the adjacency graph is  $\min\{n, m + n - d\}$ . We are restricted by the number of upper adjacencies  $n$  since each cycle contains exclusively adjacencies. We are also restricted by  $N - d$  since the total number of cycles and pairs of odd-length paths must not exceed  $N - d = m + n - d$ .

2. Next, we sum over  $i$ , the number of upper adjacencies that are in cycles. This is at least  $c$  and at most  $n$ .
3. We then sum over  $j$ , the number of upper adjacencies that are in odd-length paths.  $j$  can be 0, but it must not exceed  $n - i$  since  $n - i$  is the number of upper adjacencies that remain after the first two steps.
4. Now, we choose  $2(m + n - d - c)$  upper telomeres to be in odd-length paths. Notice that after we have decided on the number of cycles  $c$ , we know from Theorem 1 that there are  $N - d - c = m + n - d - c$  pairs of odd-length paths in the adjacency graph. Thus, we multiply by

$$\binom{2m}{2(m+n-d-c)} = \binom{2m}{2(d+c-m)}.$$

5. Next, we pick the upper adjacencies that are in cycles and those that are in odd-length paths. This can be done in

$$\binom{n}{i} \binom{n-i}{j}$$

ways.

6. We now arrange into odd-length paths the  $n + m - d - c$  pairs of upper telomeres and  $j$  upper adjacencies that we have selected to be in odd-length paths. From Lemma 5, the number of ways to do this is

$$j! \binom{2(n+m-d-c)+j-1}{j} 2^j.$$

7. We proceed by arranging into  $c$  cycles, the  $i$  upper adjacencies that we have selected for this purpose. Lemma 6 establishes that there are

$$s(i, c)2^{i-c}$$

ways to accomplish this, where  $s(a, b)$  are the unsigned Stirling numbers of the first kind.

8. The remaining  $2(d + c - n)$  upper telomeres and  $n - i - j$  upper adjacencies are placed into paths of even length. There are

$$(2(d + c - n) - 1)!! \sum_{k=0}^{n-i-j} \binom{(d+c-n)+(n-i-j)-1}{n-i-j-k} \frac{(n-i-j)!}{k!} 2^{n-i-j-k}$$

ways to do this by (see Equation (2) in the proof of Theorem 2).

We now combine these eight steps and place the sums together, using the abbreviation

$$\sum' = \sum_{c=\max\{0, n-d\}}^{\min\{n, m+n-d\}} \sum_{i=c}^n \sum_{j=0}^{n-i} \sum_{k=0}^{n-i-j}$$

for simplicity. We obtain

$$\begin{aligned} G(m, n, d) &= \sum' \binom{2m}{2(d+c-m)} \binom{n}{i} \binom{n-i}{j} \\ &\quad \times j! \binom{2(n+m-d-c)+j-1}{j} 2^j s(i, c) 2^{i-c} \\ &\quad \times (2(d+c-n) - 1)!! \binom{d+c-i-j-1}{n-i-j-k} \frac{(n-i-j)!}{k!} 2^{n-i-j-k} \\ &= \sum' \frac{s(i, c) (2(d+c-n) - 1)!! j! (n-i-j)! \binom{n}{i} \binom{n-i}{j}}{k! 2^{c+k-n}} \\ &\quad \times \binom{2m}{2(d+c-n)} \binom{2(m+n-d-c)+j-1}{j} \binom{d+c-i-j-1}{n-i-j-k} \\ &= \sum' \frac{s(i, c) (2(d+c-n) - 1)!! j! (n-i-j)! n! (n-i)!}{k! 2^{c+k-n} i! (n-i)! (n-i-j)! j!} \\ &\quad \times \binom{2m}{2(d+c-n)} \binom{2(m+n-d-c)+j-1}{j} \binom{d+c-i-j-1}{n-i-j-k} \\ &= \sum' \frac{s(i, c) n! (2(d+c-n) - 1)!!}{i! k! 2^{c+k-n}} \binom{2m}{2(d+c-n)} \\ &\quad \times \binom{2(m+n-d-c)+j-1}{j} \binom{d+c-i-j-1}{n-i-j-k}. \quad \square \end{aligned}$$

**Remark.** Let  $G(m, n, d)$  be the number of genomes that are a distance  $d$  away from a starting genome having  $2m$  telomeres and  $n$  adjacencies, and let  $G_{\max}(m, n)$  be the number of genomes that are the maximum DCJ distance away from the same genome. Then  $G_{\max}(m, n) = G(m, n, m + n)$ .

In the case where  $d = m + n$  the inner two sums in  $G(m, n, m + n)$  collapse with  $c = i = 0$ , and we have,

$$G(m, n, m + n) = \sum_{j=0}^n \sum_{k=0}^{n-j} \frac{s(0, 0)n!(2m - 1)!!}{0! k! 2^{k-n}} \binom{2m}{2m} \binom{j-1}{j} \binom{m+n-j-1}{n-j-k}.$$

Since  $\binom{j-1}{j} = 0$  unless  $j = 0$ , the outer sum collapses, and we obtain

$$\begin{aligned} G(m, n, m + n) &= \sum_{k=0}^n \frac{n!(2m - 1)!!}{k!} 2^{n-k} \binom{m+n-1}{n-k} \\ &= (2m - 1)!! \sum_{k=0}^n \binom{m+n-1}{n-k} \frac{n!}{k!} 2^{n-k} \\ &= G_{\max}(m, n). \end{aligned}$$

**Theorem 8.** Let  $G(m, n, d)$  be the number of genomes that are a distance  $d$  away from a starting genome having  $2m$  telomeres and  $n$  adjacencies. Then,  $G(m, 0, d) = G(m - 1, 1, d)$ .

*Proof.* From Theorem 7 we have

$$\begin{aligned} G(m, 0, d) &= \sum_{c=0}^0 \sum_{i=0}^0 \sum_{j=0}^0 \sum_{k=0}^0 \frac{s(0, 0)0!(2(d+0)-1)!!}{0! 0! 2^0} \binom{2m}{2(d+0)} \binom{(2(m+0)+0-1)}{0} \binom{(d+0-1)}{0} \\ &= (2d-1)!! \binom{2m}{2d}, \end{aligned}$$

$$\begin{aligned} G(m-1, 1, d) &= \sum_{c=\max\{0, 1-d\}}^{\min\{1, m+1-d\}} \sum_{i=c}^1 \sum_{j=0}^{1-i} \sum_{k=0}^{1-i-j} \frac{s(i, c)1!(2(d+c-1)-1)!!}{i! k! 2^{c+k-1}} \binom{2(m-1)}{2(d+c-1)} \\ &\quad \times \binom{2(m-d-c)+j-1}{j} \binom{(d+c-i-j-1)}{1-i-j-k}. \end{aligned}$$

In this last equation, suppose  $m - d \neq 0$  and  $d \neq 0$ . Then, we have

$$G(m-1, 1, d) = \sum_{c=0}^1 \sum_{i=c}^1 \sum_{j=0}^{1-i} \sum_{k=0}^{1-i-j} \frac{s(i, c) (2(d+c-1)-1)!!}{i! k! 2^{c+k-1}} \binom{2(m-1)}{2(d+c-1)} \times \binom{2(m-d-c)+j-1}{j} \binom{d+c-i-j-1}{1-i-j-k}.$$

This sum becomes

$$\begin{aligned} G(m-1, 1, d) &= (2d-1)!! \binom{2m-2}{2d} + 0 + 2(2d-3)!! \binom{2m-2}{2d-d} 2^{m-d} \\ &\quad + (2d-3)!! \binom{2m-2}{2d-2} + 2(2d-3)!! \binom{2m-2}{2d-2} (d-1) \\ &= (2d-1)!! \binom{2m}{2d} \frac{(2m-2d)(2m-2d-1)}{(2m)(2m-1)} \\ &\quad + (2d-3)!! \binom{2m}{2d} \frac{(2d)(2d-1)}{(2m)(2m-1)} (4(m-d) + 1 + 2(d-1)) \\ &= (2d-1)!! \binom{2m}{2d} \left( \frac{(2m-2d)(2m-2d-1)}{(2m)(2m-1)} + \frac{2d(4m-4d+1+2d-2)}{(2m)(2m-1)} \right). \end{aligned}$$

Obtaining a common denominator and simplifying yields

$$G(m-1, 1, d) = (2d-1)!! \binom{2m}{2d} \left( \frac{4m^2-2m}{4m^2-2m} \right) = (2d-1)!! \binom{2m}{2d}.$$

Hence,  $G(m, 0, d) = G(m-1, 1, d)$  when  $m-d \neq 0$  and  $d \neq 0$ . A similar argument shows that  $G(m, 0, d) = G(m-1, 1, d)$  when  $m = d$ . Now, if  $d = 0$ , we have  $G(m, 0, 0) = G(m-1, 1, 0)$ , since there is only one genome that is a distance 0 away from a starting genome regardless of the starting genome. Thus, in all cases, we have established that

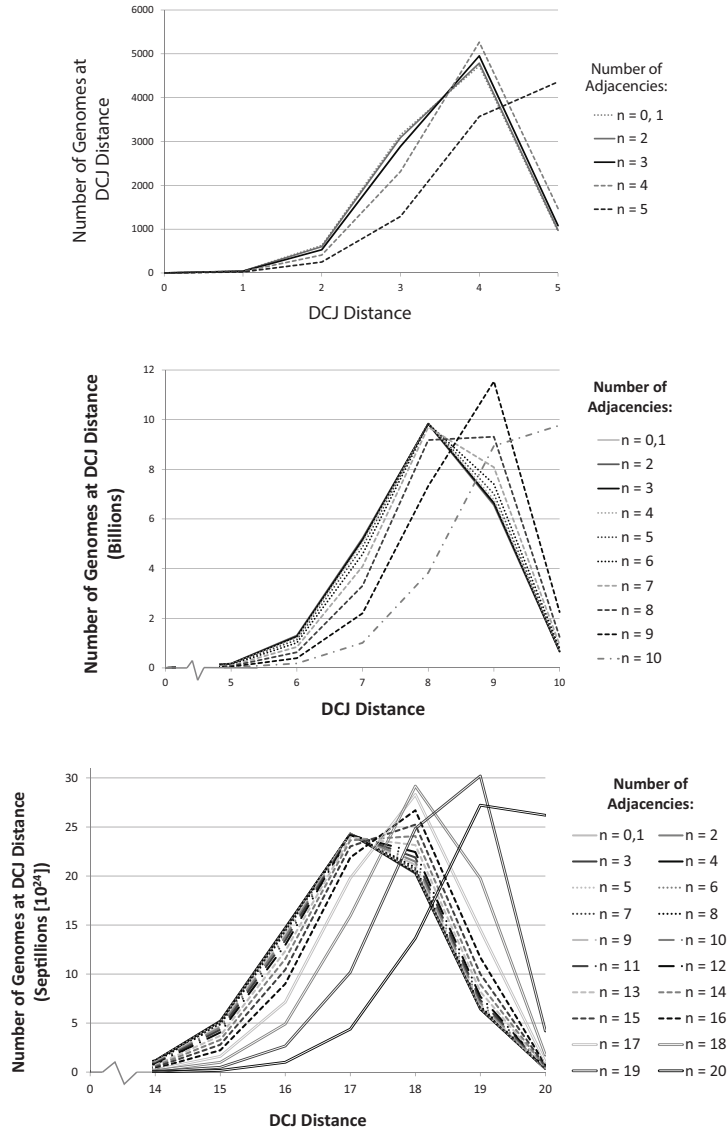
$$G(m, 0, d) = G(m-1, 1, d) = (2d-1)!! \binom{2m}{2d}. \quad \square$$

**Definition 9.** Consider all DCJ genomes defined the same set of  $N$  genes. Observe that for each of these genomes,  $N = m+n$ , where  $2m$  is the number of telomeres and  $n$  is the number of adjacencies in the genome. We define the *distance distribution* on  $N$  genes with respect to  $n$  to be the distribution of genomes according to their distance from a given genome containing  $n$  adjacencies and  $2(N-n)$  telomeres.

Figure 4 depicts the distance distribution on five and on ten genes for all possibilities of  $n$  adjacencies. These results contribute to the understanding of how DCJ distance is distributed over all genomes. The figure displays one property that

we have observed for every distance distribution we have considered thusfar. The following conjecture summarizes this feature.

**Conjecture 1.** *The distance distribution on  $N$  genes with respect to  $n$  is unimodal for  $n = 0, 1, \dots, N - 1$ .*



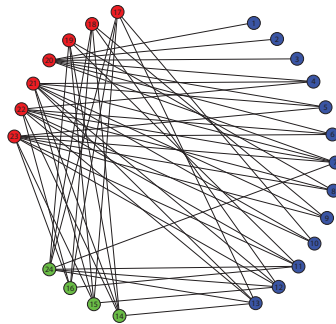
**Figure 4.** The distance distribution on  $n$  genes with respect to  $0, 1, \dots, n$  adjacencies, for  $n = 5$  (top),  $n = 10$  (middle) and  $n = 20$  (bottom).

Although we have yet to prove this claim, it has been verified for all distance distributions on  $N$  genes where  $1 \leq N \leq 10$ , and for the cases where  $N = 16$  and  $N = 20$ .

### 5. Concluding remarks

We would like to extend our results to an unsigned version of the DCJ model. Using a computer program that we created to simulate DCJ operations on unsigned genomes, we collected information about maximally distant genomes. For small values of  $N$ , we counted the total number of genomes that can be defined on a fixed number of genes.

In addition to examining maximally distant genomes, we investigated properties of the maximum distance graph  $M$ , whose vertices constitute all possible genomes of length  $N$  and whose edges link two vertices  $a$  and  $b$  whenever genome  $a$  is maximally distant from genome  $b$ . In Figure 5, we show such a graph for all genomes on three genes.



**Figure 5.** Maximum distance graph  $M$  for all genomes on 3 genes, with genomes labeled in lex order. It is a tripartite graph with maximally independent subsets of sizes 4, 7, and 11.

Ultimately, we would like to develop a formula for the distance between unsigned genomes and one that counts all of the possible unsigned genomes defined on a fixed set of genes. We could then extend our results from Sections 3 and 4 to the unsigned DCJ model.

### References

- [Bergeron et al. 2006] A. Bergeron, J. Mixtacki, and J. Stoye, “A unifying view of genome rearrangements”, pp. 163–173 in *Algorithms in Bioinformatics*, edited by P. Bücher and B. M. E. Moret, Lecture Notes in Computer Science **4175**, 2006.
- [Fertin et al. 2009] G. Fertin, A. Labarre, I. Rusu, É. Tannier, and S. Vialette, *Combinatorics of genome rearrangements*, MIT Press, Cambridge, MA, 2009. MR 2010e:92083

[NHGRI 2002] National Human Genome Research Institute, “The mouse genome and the measure of man”, technical report (NIH News Advisory), 2002.

[Yancopoulos et al. 2005] S. Yancopoulos, O. Attie, and R. Friedberg, “Efficient sorting of genomic permutations by translocation, inversion and block interchange”, *Bioinformatics* **21** (2005), 3340–3346.

Received: 2012-12-14      Revised: 2013-03-30      Accepted: 2013-04-01

christyj@uwec.edu                      *Department of Mathematics, University of Wisconsin,  
Eau Claire, WI 54702-4004, United States*

joshua.mchugh.129@gmail.com                      *Department of Mathematics, University of Wisconsin,  
Eau Claire, WI 54702-4004, United States*

riehlar@uwec.edu                      *Department of Mathematics, University of Wisconsin,  
Eau Claire, WI 54702-4004, United States*

noah.williams@colorado.edu                      *Department of Mathematics 340, University of Colorado,  
Campus Box 395, Boulder, CO 80309-0395, United States*





# Mathematical modeling of integrin dynamics in initial formation of focal adhesions

Aurora Blucher, Michelle Salas,  
Nicholas Williams and Hannah L. Callender

(Communicated by Michael Dorff)

Cellular motility is an important function in many cellular processes. Among the key players in cellular movement are transmembrane receptor proteins called integrins. Through the development of a mathematical model we investigate the dynamic relationship between integrins and other molecules known to contribute to initial cellular movement such as extracellular ligands and intracellular adhesion proteins called talin. Gillespie's stochastic simulation algorithm was used for numerical analysis of the model. From our stochastic simulation, we found that most activity in our system happens within the first five seconds. Additionally we found that while ligand-integrin-talin complexes form fairly early in the simulation, they soon disassociate into ligand-integrin or integrin-talin complexes, suggesting that the former tertiary complex is less stable than the latter two complexes. We also discuss our theoretical analysis of the model and share results from our sensitivity analysis, using standardized regression coefficients as measures of output sensitivity to input parameters.

## 1. Introduction

The processes of cellular movement and migration are vital to the performance and maintenance of an individual cell and in turn to the well-being of the larger organism. Embryonic development, the immune system response, and tissue regeneration all require cell motility to progress effectively. Cellular processes that are harmful to the body, such as cancer metastasis, also rely on cell motility [Lauffenburger and Horwitz 1996; Fletcher and Theriot 2004]. Due to the importance of cell motility for proper function of an organism, it is necessary to develop a deeper understanding of the mechanisms involved. One way to do so is through the use of mathematical models, which can provide insight beyond that garnered from traditional experimental methods and techniques.

---

*MSC2010:* primary 92C17, 92C37; secondary 90C31.

*Keywords:* cellular motility, mathematical modeling, focal adhesions, Gillespie's algorithm, integrin receptor, sensitivity analysis.

The process of cell motility can be broken down into four general steps. First, the cell protrudes a thin lamellipodium, a projection of the cell's cytoskeleton, from its leading edge in the desired direction of motion. Next, the lamellipodium attaches to the extracellular matrix through the use of focal adhesions, which are macromolecule assemblies containing integrin receptor proteins, actin, and other linking proteins. Then myosin-II, a motor protein in the cell, causes actin strands to converge, which pulls on the focal adhesions and generates traction. Finally, the traction causes weaker focal adhesions in the rear of the cell to detach and, through the contraction of the actin filaments, moves the cell body forward [Wehrle-Haller 2006; Ananthkrishnan and Ehrlicher 2007].

Our focus is on the development of the focal adhesions in the second step of the motility process. In particular, we seek to model the dynamics of integrins, transmembrane receptor proteins that play a major role in the development of these adhesions [Hynes 1992]. Within each focal adhesion, integrins form mechanical linkages to extracellular signaling molecules called ligands. The number of integrins bound to extracellular ligands as well as to intracellular adhesion proteins is a key factor in determining the strength and duration of the linkages, thus providing a deeper understanding of the overall motility properties of the cell.

Other approaches to modeling and investigating focal adhesions have varied. Some models have compared the strength of a given focal adhesion with the number of ligand-integrin bonds within the adhesion (see, for example, [Gallant and Garcia 2007; Gov 2006; Flaherty et al. 2007]). These models have investigated the forces required to detach the cell from its environment, given the number of ligand-integrin bonds. Other stress-based models, such as [Gallant and Garcia 2007; Cozens-Roberts et al. 1990; Ward and Hammer 1993], show the distribution and total stress in the focal adhesions in relation to time or strain of the integrins. Our model differs from these as it takes into account the initial formation of focal adhesions and the binding interactions between the primary molecules present within a nascent adhesion. This has allowed us to investigate which molecules contribute more and in what manner to the formation and fate of focal adhesions.

As our goal is to model the interactions between integrins and other molecules in a focal adhesion, it is necessary to take into account the likelihood that integrins bind to other molecules. Heterodimeric integrins exist in low-affinity ("inactive") and high-affinity ("active") states, and a variety of molecules from both inside and outside the cell are known to take part in the regulation of these integrin states. Among the most important of these molecules are talin and extracellular ligands [Small et al. 2002; Soll 1995]. Talin, which is an intracellular signaling molecule, can bind to and activate integrins from inside the cell. This activation is known to increase an integrin's affinity for ligands. A ligand binding to an integrin also activates the integrin from the outside of the cell, both providing linkage to the

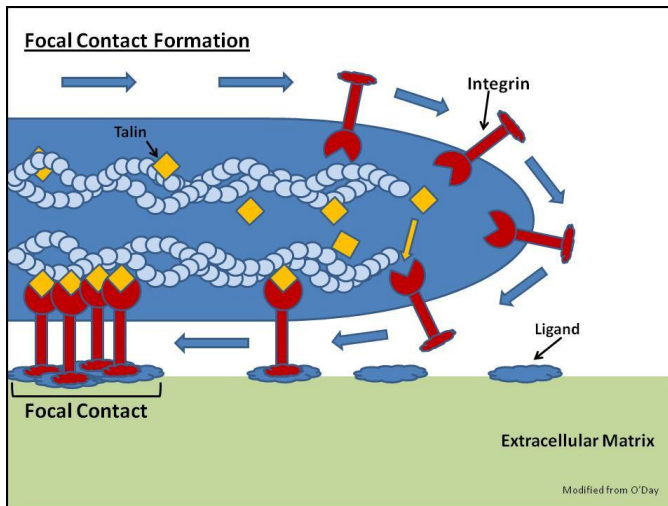
extracellular matrix and increasing the integrin’s affinity for intracellular molecules as well as for other ligands [Cluzel et al. 2005].

This paper is organized as follows. In Section 2 (stochastic simulation), we discuss how we use mass action kinetics and Gillespie’s algorithm to model a system involving key molecules participating in the early formation of focal contacts. In Section 4 (numerical results), we discuss our results from the stochastic simulation. In Section 5 (deterministic model), we conduct theoretical analysis on an ordinary differential equations model of our system to check the mathematical relevance of the system. Next, in Section 7 (sensitivity analysis), we discuss parameters which may cause more uncertainty in our model output and would therefore warrant further investigation. We conclude with a discussion and future directions.

### 2. Stochastic simulation

Experiments have indicated that the primary players in early focal adhesion formation include integrins, ligands, and talin [Cluzel et al. 2005]. At present, our model focuses on the initial dynamics between these molecules, as depicted in Figure 1.

Integrins bind with ligands to form integrin-ligand complexes, and conversely, ligand-integrin complexes can dissociate into free ligands and free integrins (see Equation (1)). Similarly, integrins can bind with talin to form integrin-talin complexes, which can then dissociate into free ligands and free talin (Equation (2)).



**Figure 1.** Depiction of formation of focal contacts (term commonly used for early focal adhesions) through lateral diffusion of integrins, binding of integrins to talin, and binding of integrins to extracellular ligands. Modified from Figure 12.2 of [O’Day 2012].

Free talin and free ligands can also each bond with complexes from Equations (1) and (2), respectively, to form integrin-ligand-talin complexes, which can then dissociate (Equations (3) and (4)) [Wehrle-Haller 2006]. As integrins are known to diffuse along the plasma membrane, a reaction was also included to allow for the diffusion of integrins into and out of our system, where the diffusing integrins come from an outside source (Equation (5)). Therefore the reactions included in our system are the following:



In the above reactions the following abbreviations are used: L for ligands, I for integrins, LI for ligand-integrin complexes, T for talin, IT for integrin-talin complexes, LIT for ligand-integrin-talin complexes, and S for an outside source of integrins.

Using these reactions, we seek to model the change in the number of each molecule over time. We begin by supposing that the initial number of molecules of each reactant is known. The state vector, denoted by  $X(t)$ , which changes each time one of the above reactions takes place, is used to record the amount of each reactant at any time  $t$ . This state vector evolves according to the propensity functions of our original reactions. The propensity function, denoted by  $a_j$ , is the likelihood that the  $j$ -th reaction will occur and is proportional to the product of the number of molecules of the reactants in the  $j$ -th reaction. For example, take the forward reaction of (1), where ligands bind to integrins with forward rate constant  $k_L^+$ . The propensity function for the forward reaction would be:

$$a_1 = [L] * [I] * k_L^+,$$

where  $[L]$  is the number of free ligand molecules and  $[I]$  of the number of free integrin molecules.

$$\begin{array}{c}
 \\
 v_1 \\
 v_2 \\
 v_3 \\
 v_4 \\
 v_5 \\
 v_6 \\
 v_7 \\
 v_8
 \end{array}
 \begin{bmatrix}
 & \text{L} & \text{I} & \text{IT} & \text{LI} & \text{LIT} & \text{T} \\
 -1 & -1 & 0 & 1 & 0 & 0 & \\
 1 & 1 & 0 & -1 & 0 & 0 & \\
 0 & -1 & 1 & 0 & 0 & -1 & \\
 0 & 1 & -1 & 0 & 0 & 1 & \\
 -1 & 0 & -1 & 0 & 1 & 0 & \\
 1 & 0 & 1 & 0 & -1 & 0 & \\
 0 & 0 & 0 & -1 & 1 & -1 & \\
 0 & 0 & 0 & 1 & -1 & 1 & 
 \end{bmatrix}$$

**Figure 2.** Stoichiometry matrix.

Next we create a stoichiometry matrix, which is used to track changes in the state vector and allows us to follow changes in the entire system rather than in one reactant. Figure 2 shows the entire stoichiometry matrix. Each row of the matrix tells us which molecules to add, take away, or keep fixed depending on which reaction occurred.

To illustrate how the stoichiometry matrix is used, take the forward reaction of (1), where one ligand and one integrin bind to form a ligand-integrin complex. As shown in Figure 2, the corresponding row for this reaction,  $v_1$ , has a “-1” in both the ligand and integrin columns to indicate the loss of one of each of these molecules and a “+1” in the ligand-integrin complex column to indicate the gain of one complex. Since the other reactants are unaffected by this reaction, all other entries in this row contain a zero.

In order to predict the future states of the whole system of molecules, we seek to model  $P(X, t)$ , the probability of the system being in a certain state,  $X(t)$ , at a certain point in time,  $t$ . This probability is equal to the probability of moving to that state from a neighboring one, given by  $a_j(X(t) - v_j) \cdot P(X(t) - v_j, t)$ , minus the probability of moving from that state to a neighboring one, given by  $a_j(X(t)) \cdot P(X(t), t)$  multiplied by the time step  $\Delta t$ . In general, for a system with  $M$  reactions, we sum all these probabilities for each of the  $M$  reactions, divide by  $\Delta t$ , and take the limit as  $\Delta t$  approaches zero to obtain the *Chemical Master equation*:

$$\frac{dP(X(t), t)}{dt} = \sum_{j=1}^M (a_j(X(t) - v_j) \cdot P(X(t) - v_j, t)) - a_j(X(t)) \cdot P(X(t), t)$$

which is a set of ordinary differential equations for the probability of the whole system being in a particular state  $X(t)$  at any time  $t$ .

### 3. Simulation method: Gillespie's algorithm

The chemical master equation has continuous time, but the state of the system is updated discretely. This makes it very difficult to obtain an analytic solution. Therefore, we approximate a solution using Gillespie's algorithm [1977], which can be summarized in the following steps:

- (1) Initialize the time  $t = t_0$  and the state of the system  $x = x_0$ .
- (2) Evaluate the propensities for each reaction,  $a_j$ , and the sum of the propensities,  $a_{\text{sum}}$ .
- (3) Randomly choose two numbers from a uniform distribution on  $[0, 1]$ , denoted  $\xi_1$  and  $\xi_2$ .
- (4) In order to obtain the next reaction that will take place, let  $j$  be the smallest integer satisfying  $\sum_{j=1}^M a_j(X(t)) > \xi_1 \cdot a_{\text{sum}}(X(t))$ , where  $a_{\text{sum}}$  is the sum of the propensities.
- (5) Let  $\tau = \ln(1/\xi_2)/a_{\text{sum}}(X(t))$ . This determines the next time a reaction will take place. For more details on the choice of  $\tau$ , see [Gillespie 2007].
- (6) Now that the next time step and reaction have been chosen, update the current time,  $t$ , by changing it to  $t + \tau$ . Similarly, update the current state of the system by setting  $X(t + \tau) = X(t) + v_j$ , where  $v_j$  is the  $j$ -th row of the stoichiometry matrix.
- (7) Repeat steps (1)–(5) until the desired time course has been reached.

The interested reader may see [Higham 2008] for a more detailed description of Gillespie's method. In the fifth step of the algorithm, the state of the system is updated with the stoichiometry matrix. To illustrate this step, assume the forward reaction of step (1) has been randomly chosen to occur at time  $t^*$ . This reaction corresponds to the first row of the stoichiometry matrix. Therefore, if the current state of our system is  $X(t)$ , then to get the new state of the system,  $v_1$  is added to  $X(t)$  to obtain  $X(t^*)$  as follows:

$$X(t) = \begin{array}{c} \text{L} \quad \text{I} \quad \text{IT} \quad \text{LI} \quad \text{LIT} \quad \text{T} \\ \left[ \begin{array}{cccccc} 15 & 10 & 0 & 0 & 0 & 0 \end{array} \right] \end{array}$$

$$v_1 = \left[ \begin{array}{cccccc} -1 & -1 & 0 & 1 & 0 & 0 \end{array} \right]$$

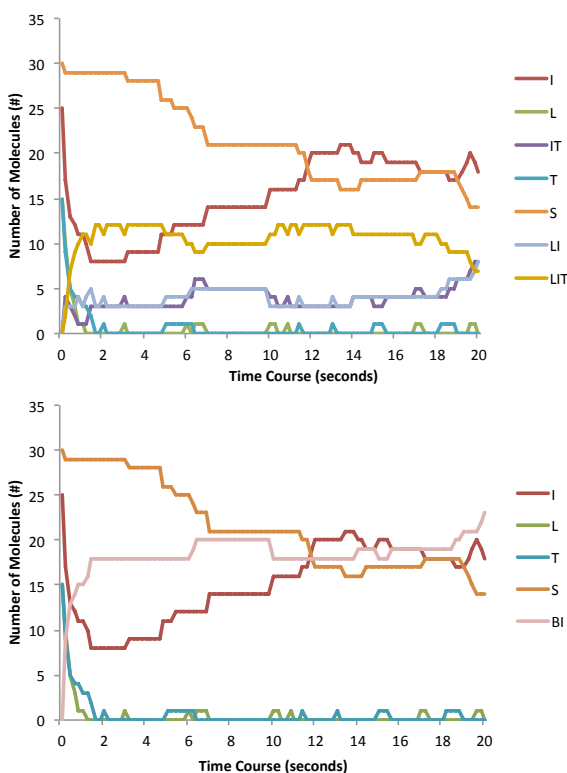
$$X(t^*) = \left[ \begin{array}{cccccc} 14 & 9 & 0 & 1 & 0 & 0 \end{array} \right]$$

These steps of the algorithm are repeated until the desired time course is achieved. In the next section, we describe the results obtained from simulations of our model for durations of both five and twenty seconds. We also provide an interpretation for the corresponding outputs in the context of our biological system.

### 4. Numerical results

Our stochastic simulations were run in COMplex PATHway SIMulator (CoPaSi), using Gillespie’s algorithm as described in Section 3 and the rate constants shown in Table 1 on the next page. Rate constants were chosen to reflect values of similar parameters from the literature [Lee et al. 2007; Calderwood et al. 2002] and represent an initial attempt to compare the stochastic and deterministic results. For more details on how CoPaSi implements Gillespie’s method, see [Gibson and Bruck 2000].

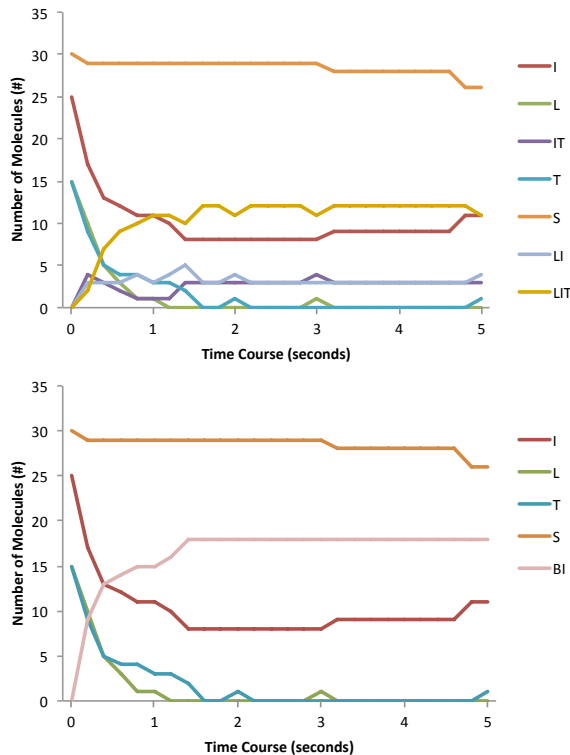
Figures 3 and 4 show several examples of CoPaSi output from one simulation. The two graphs in Figure 3 show reactant activity over twenty seconds: the upper graph shows activity for all reactants, while the lower combines all integrin complexes into *bound integrins* (BI). The two graphs in Figure 4 show the same reactant activity over the shorter time course of five seconds.



**Figure 3.** Representative output for each model variable, simulated stochastically through Gillespie’s algorithm (described in Section 3) in CoPaSi. Abbreviations are as follows: IT = integrin-talin complex; I = integrin; LIT = ligand-integrin-talin complex; LI = integrin-ligand complex; L = ligand; T = talin; BI = bound integrins (IT + LI + LIT).

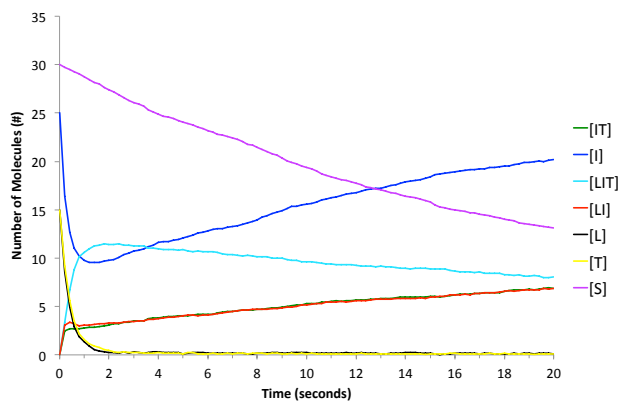
Reaction equation	Rate value	Parameter name
$I + L \longrightarrow LI$	0.1	$k_L^+$
$LI \longrightarrow I + L$	0.05	$k_L^-$
$I + T \longrightarrow IT$	0.1	$k_T^+$
$IT \longrightarrow I + T$	0.072	$k_T^-$
$IT + L \longrightarrow LIT$	0.5	$k_{LIT}^+$
$LIT \longrightarrow IT + L$	0.05	$k_{LIT}^-$
$LI + T \longrightarrow LIT$	0.3	$k_{LIT}^+$
$LIT \longrightarrow LI + T$	0.04	$k_{LIT}^-$
$S \longrightarrow I$	0.05	$k_D^+$
$I \longrightarrow S$	0.01	$k_D^-$

**Table 1.** Rate parameters for model reactions.



**Figure 4.** Representative output for each model variable, simulated stochastically through Gillespie’s algorithm (described in Section 3) in CoPaSi. Abbreviations are as follows: IT = integrin-talin complex; I = integrin; LIT = ligand-integrin-talin complex; LI = integrin-ligand complex; L = ligand; T = talin; BI = bound integrins (IT + LI + LIT).





**Figure 5.** The average of 100 stochastic simulations over 20 seconds using Gillespie’s algorithm in CoPaSi. Abbreviations are as follows: IT = integrin-talin complex; I = integrin; LIT = ligand-integrin-talin complex; LI = integrin-ligand complex; L = ligand; T = talin.

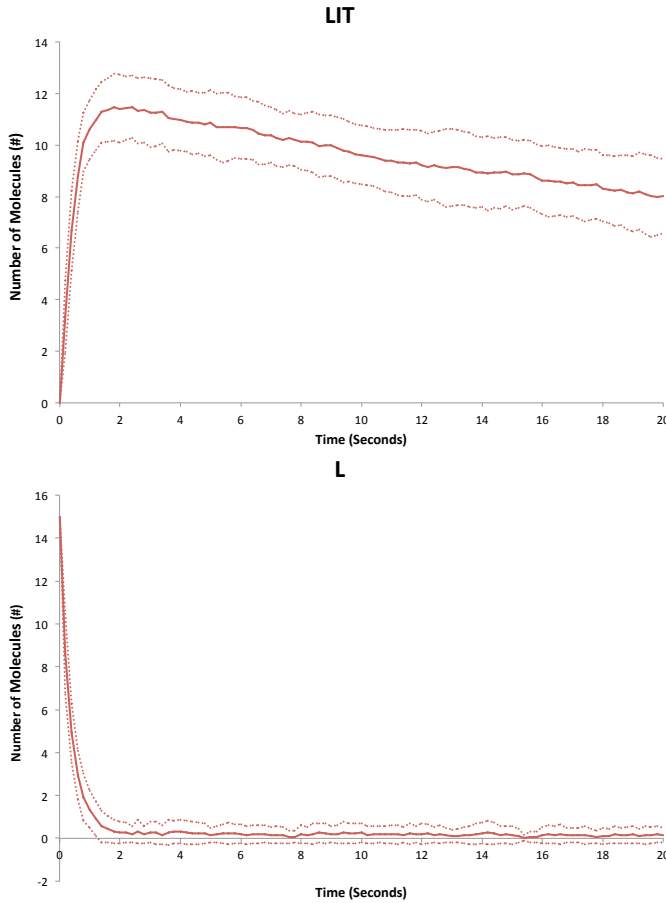
Of particular interest is the activity of talin and ligand molecules, which are depleted very early in the simulation (just before 2 seconds have elapsed) and remain very close to zero throughout the time course, with most activity occurring within the first five seconds. The corresponding increases in LI and IT complexes is expected, but the gradual increase in LI and IT and the gradual decrease in LIT complexes indicate that the former two complexes are more stable than the latter. This suggests that while LIT complexes form early on in focal adhesion development, the tertiary complex is less stable and therefore possibly less vital to continued stability of the overall focal adhesion.

While analyzing individual simulations of our stochastic model can provide information on the random interactions of individual species, the average of multiple simulations provides insight into overall trends in the dynamics of our system. As shown in Figure 5, the responses of the reactants averaged over 100 simulations are qualitatively similar to that of a sample individual simulation. Figure 5 also supports the observation from individual simulations that most of the system activity happens within the first five seconds.

Corresponding standard deviations were computed to examine the variability for each reactant in our system. Figure 6 shows the average of 100 simulations and the standard deviation for ligand-integrin-talin complexes and ligands, respectively.

## 5. Deterministic model

For a more thorough investigation of our model, we also looked at the reactions using ordinary differential equations. These equations are based on assumptions



**Figure 6.** The average (in solid lines) of 100 stochastic simulations with one standard deviation from the mean (in dotted lines) for ligand-integrin-talin complexes (top) and free ligands (bottom).

of mass-action kinetics which states that the rate of a chemical reaction is directly proportional to the molecular concentrations or number of molecules of the reacting substances. Take, for example, the forward reaction of (1), where ligands bind with integrins to form ligand-integrin complexes with forward rate constant  $k_L^+$ . Similarly, ligand-integrin complexes dissociate to form free ligands and integrins, with backwards rate constant  $k_L^-$ . However, this is only one of the reactions affecting the amount of free ligands at any time  $t$ . Taking into consideration the other reactions affecting the amount of free ligands, we form the following ordinary differential equation for the change in ligands:

$$\frac{d[L]}{dt} = -k_L^+[L] \cdot [I] + k_L^-[LI] - k_{LT}^+[L] \cdot [IT] + k_{LT}^-[LIT].$$

Note that  $k_L^+$  follows a negative sign since that is the rate at which we lose ligands and integrins, and  $k_L^-$  follows a positive sign because that is the rate at which we lose ligand-integrin complexes and thus gain ligands. We proceed in a similar manner for the rest of the reactions and form our system of ordinary differential equations:

$$\frac{d[L]}{dt} = -k_L^+[L] \cdot [I] + k_L^-[LI] - k_{LT}^+[L] \cdot [IT] + k_{LT}^-[LIT], \tag{6}$$

$$\frac{d[I]}{dt} = -k_L^+[L] \cdot [I] + k_L^-[LI] + k_D^+[S] - k_D^-[I] - k_T^+[T] \cdot [I] + k_T^-[IT], \tag{7}$$

$$\frac{d[LI]}{dt} = k_L^+[L] \cdot [I] - k_L^-[LI] - k_{LIT}^+[T] \cdot [LI] + k_{LIT}^-[LIT], \tag{8}$$

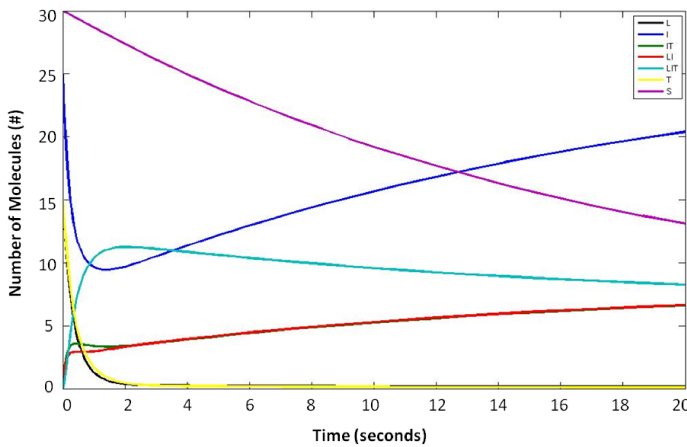
$$\frac{d[IT]}{dt} = k_T^+[T] \cdot [I] - k_T^-[IT] - k_{LT}^+[L] \cdot [IT] + k_{LT}^-[LIT], \tag{9}$$

$$\frac{d[LIT]}{dt} = k_{LT}^+[L] \cdot [IT] - k_{LT}^-[LIT] + k_{LIT}^+[T] \cdot [LI] - k_{LIT}^-[LIT], \tag{10}$$

$$\frac{d[S]}{dt} = k_D^-[I] - k_D^+[S], \tag{11}$$

$$\frac{d[T]}{dt} = -k_T^+[T] \cdot [I] + k_T^-[IT] - k_{LIT}^+[T] \cdot [LI] + k_{LIT}^-[LIT]. \tag{12}$$

Numerical solutions to this system of equations were obtained using the Matlab differential equation solver `ode15s`. The initial conditions were as follows:  $L = 15$ ,  $I = 25$ ,  $IT = 0$ ,  $LI = 0$ ,  $LIT = 0$ ,  $T = 15$ , and  $S = 30$ . The results of the deterministic simulation (Figure 7) are very similar to the average of the results of 100 stochastic simulations (Figure 5). Both graphs show similar behavior for the amounts of all



**Figure 7.** Results of deterministic simulation over 20 seconds.

reactants. Additionally, for both the stochastic and the deterministic simulation, the most dynamic behavior occurs within the first five seconds.

## 6. Qualitative analysis

While some systems of equations can be solved analytically, there are many systems for which this cannot be done. Often this is the case for systems of nonlinear equations representing real-world problems, such as complex biological systems. Rather than solving for an analytic solution, or approximating a solution numerically, certain aspects of the solution can be investigated to learn more about the overall qualitative behavior of the system. Among these qualities are existence, uniqueness, boundedness of solutions, and stability of steady state solutions. Due to the wide array of tools available for analysis of deterministic systems, the analysis we have used here is for a deterministic model, rather than a stochastic one.

The first aspect of our solution that we will check is existence. Existence of a solution to our system is important because it enables further model analysis. After verifying existence, we can check our solution for uniqueness. Since we are using deterministic analysis, we should find that our solution is unique, which means that there is exactly one solution for a given set of initial conditions. Biologically, this means that given the same initial amounts of reactants, our system will always behave in the same manner.

In order to show existence of a solution, we use the well-known theorem (stated below) that says that if the right-hand side and the partials of a system of the form  $\dot{x} = f(t, x)$  are continuous in some finite region  $B$ , and also bounded in region  $B$ , then the approximations given in (14) that satisfy the given initial condition in (13) converge uniformly on a given interval of time to a solution of the system.

**Theorem 1** [Brauer and Nohel 1969]. *Let  $f$  and  $\partial f/\partial x_j$  ( $j = 1, \dots, n$ ) be continuous on the box  $B = \{(t, x) : |t - t_0| \leq a, |x - \eta| \leq b\}$ , where  $a$  and  $b$  are positive numbers, and satisfying the bounds*

$$|f(t, x)| \leq N, \quad \left| \frac{\partial f(t, x)}{\partial x_j} \right| \leq K \quad (j = 1, \dots, n),$$

for  $(t, x)$  in  $B$ . Let  $\alpha$  be the smaller of the numbers  $a$  and  $b/N$  and define the successive approximations

$$\phi_0(t) = \eta, \tag{13}$$

$$\phi_n(t) = \eta + \int_{t_0}^t f(s, \phi_{n-1}(s)) ds. \tag{14}$$

Then the sequence  $\phi_j$  of successive approximations converges (uniformly) on the interval  $|t - t_0| \leq \alpha$  to a solution  $\phi(t)$  of the system that satisfies the initial condition  $\phi(t_0) = \eta$ .

We first note that our system is in the form  $\dot{x} = f(t, x)$ , where  $f(t, x)$  is a vector-valued function equal to the right-hand side of our system in (6)–(12). To check that our system satisfies the continuity requirements of the theorem, we check that the right-hand sides of (6)–(12) and all of their partials are continuous. For example, in (6) the right-hand side is composed of positive rate constants multiplied by variables, so it is continuous.

Next we check that the partial derivatives from this equation are continuous. The partial derivative with respect to L is

$$\frac{\partial f}{\partial [L]} = -k_L^+ [I] - k_{LT}^+ [IT].$$

The partial derivative does not contain any terms that could be discontinuous at any point. It can be shown that the partial derivatives with respect to the remaining variables in (6), as well as all partials for the remaining six equations of the system are continuous everywhere. Additionally, it can be seen that both the right-hand side and partial derivatives with respect to each variable in equations (6)–(12) are bounded in finite time. Therefore, the continuity and boundedness requirements have been met for Theorem 1.

The following well-known theorem can now be used to show uniqueness:

**Theorem 2** [Brauer and Nohel 1969]. *Suppose  $f$  and  $\partial f/\partial x_j$  ( $j = 1, \dots, n$ ) are continuous on the box  $B = \{(t, x) : |t - t_0| \leq a, |x - \eta| \leq b\}$ . Then there exists at most one solution of the system satisfying the initial condition  $\phi(t_0) = \eta$ .*

We have previously shown that our system meets both of these requirements, so we now know that our solution exists and is unique on a finite interval of time. In the final section we will discuss our ongoing qualitative analysis efforts.

### 7. Sensitivity analysis

Sensitivity analysis allows us to examine the strength of the relationship between the input parameters and output of our model. This can provide insight into which parameters of our system more strongly effect the model output and are therefore more of a priority when researching experimental values.

For our model, we used the sensitivity analysis method of standardized regression coefficients (SRCs) to determine which parameters have the greatest effect on the output. In this method, a model is represented as a linear model of the following form, shown here for ligands:

$$L_i(t) = b_0(t) + \sum_j b_j(t)m_{ij} + \epsilon_i(t),$$

where  $L_i$  is the linear fit for the  $i$ -th sample for ligands, each  $m_{ij}$  is a value of the sample matrix as described below, each  $b_j$  is a standardized regression coefficient,

and  $\epsilon_i(t)$  is the error. Here,  $i$  is the index for the number of samples taken, and  $j$  is the index for the number of parameters. The larger the absolute value of the  $b_j$ , the more sensitive the model is to the corresponding parameter.

To create our sample matrix, we used the sampling method of Latin Hypercube Sampling. For each parameter in our model, we create an interval containing the nominal value of the parameter, where the right endpoint is 10% higher than the nominal value and the left endpoint is 10% lower. We then divide the interval into subintervals of equal width and randomly choose a subinterval. From within this subinterval a value is randomly chosen and entered into the sample matrix. After this process has been repeated once for each parameter, the first row of the sample matrix has been created. Note that the sample matrix entry denoted by  $m_{ij}$  represents the value of the  $j$ -th parameter for the  $i$ -th sample. For the next sample, we exclude subintervals from which values have previously been chosen. We continue sampling until only one subinterval remains for each parameter, from which we pick the value for our final sample. The end result is an  $i$  by  $j$  matrix where each row is used to create a linear, time-dependent function for each variable being modeled. This method of sampling ensures that an accurate sampling of the entire interval is obtained, while also allowing for all parameters to change simultaneously through each run.

## 8. Sensitivity analysis results

Standardized regression coefficient values above zero indicate a positive relationship between a particular rate parameter and the amount of a given reactant, where an increase in the rate parameter results in an increase in the amount of the reactant. SRC values below zero indicate a negative relationship, where an increasing value of the rate parameter results in a decrease in the amount of the reactant. The  $R^2$  value is the correlation coefficient associated with using standardized regression coefficients to determine which parameters have a stronger effect on the output of our model. For all seven of our reactants, the  $R^2$  values for the SRCs were well above 0.95 for the time-course of interest and thus the results obtained from our sensitivity analysis are statistically valid. Table 2 provides an overview of the most influential parameters for each model reactant, while the discussion that follows provides an in-depth analysis of the sensitivity results for each reactant in the model.

**SRCs for ligands.** For ligands, all rate parameters have a greater effect for the first five seconds of the simulation, after which the value for each rate parameter levels off. This would seem to indicate that the amount of ligands in our system depends on the early activity of the simulation. In particular, the rate constant  $k_L^+$ , which is the rate at which ligands and integrins bind to form ligand-integrin complexes, has the largest negative effect in the first five seconds on the number of ligands. As

$k_L^+$  increases, it becomes more likely that ligands will bind with integrins to form ligand-integrin complexes, and therefore the number of free ligands will decrease.

**SRCs for integrins.** Two rate parameters in particular have an increasing effect on the amount of integrins:  $k_D^+$ , the rate at which integrins diffuse into the system, and  $k_D^-$ , the rate at which integrins diffuse out of the system. It is interesting to note that these parameters have more of an effect as the simulation continues, which suggests that once all the initial integrins are bound, diffusion into and out of the system will be a greater source for additional free integrins than other complexes disassociating to give free integrins. In the future, we would like to examine a more biologically relevant representation of a diffusing integrin source.

**SRCs for ligand-integrin complexes.** For ligand-integrin complexes, two parameters,  $k_L^+$  and  $k_{LIT}^+$  have a greater effect at the beginning of the simulation before leveling off, while the rest of the parameters have a growing effect on the system. In particular,  $k_L^+$ , the rate at which ligands and integrins bind, most likely has a greater effect on the number of ligand-integrin complexes because such complexes are formed from the free ligands and integrins available at the beginning of the simulation. As other reactions occur, however, there are additional ways to form ligand-integrin complexes (such as a ligand-integrin-talin complex disassociating to produce one ligand-integrin complex and one free talin molecule). Thus, the number of ligand-integrin complexes would depend less on  $k_L^+$  as the simulation progresses.

**SRCs for integrin-talin complexes.** The rate parameter  $k_T^+$  has a greater positive effect than any other parameter on integrin-talin complexes at the very beginning of the simulation (until approximately two seconds) but then decreases and levels

Reactant	Positive effect	Negative effect
L	$k_{LIT}^-, k_L^-$	$k_L^+, k_T^+$
I	$k_D^+, k_L^-$	$k_D^-, k_{LIT}^-$
LI	$k_{LIT}^-, k_{LIT}^-$	$k_L^-, k_T^-$
IT	$k_{LIT}^-, k_{LIT}^-$	$k_L^-, k_T^-$
LIT	$k_L^-, k_T^-$	$k_{LIT}^-, k_{LIT}^-$
T	$k_{LIT}^-, k_T^-$	$k_T^+, k_L^+$
S	$k_D^-$	$k_D^+$

**Table 2.** Summary of sensitivity analysis results. For each reactant, the two parameters with the strongest positive and negative effects are listed, as determined by standardized regression coefficient values.

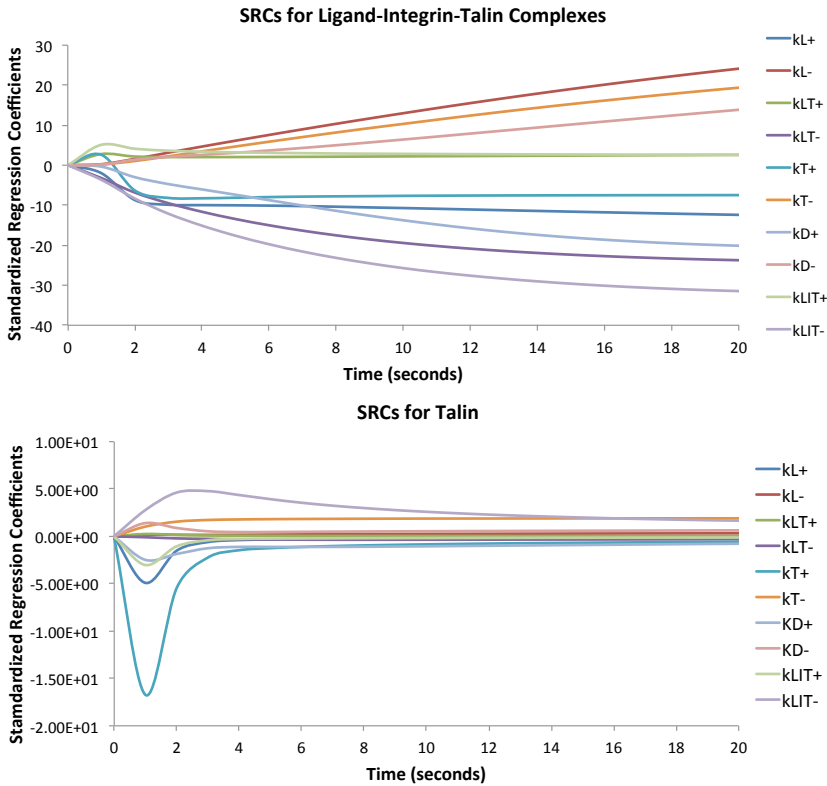
off. However, after about five seconds into the simulation,  $k_L^-$  has the greatest negative effect and  $k_{LIT}^-$  has the greatest positive effect on the number of integrin-talin complexes. This suggests that as  $k_L^-$  increases, the number of ligand-integrin complexes decreases, which in turn results in a decrease in the number of ligand-integrin-talin complexes. With fewer LIT complexes to disassociate into free ligands and integrin-talin complexes, there will be a decrease in the number integrin-talin complexes. This illustrates how the rate parameter  $k_L^-$  actually has an indirectly negative effect on the number of integrin-talin complexes.

**SRCs for ligand-integrin-talin complexes.** For ligand-integrin-talin complexes, the rate parameter that has the greatest effect is  $k_{LIT}^-$ , which is the rate at which ligand-integrin-talin complexes disassociate to produce talin molecules and ligand-integrin complexes (Figure 8). Uniquely, the parameter  $k_T^+$ , or the rate at which integrin and talin molecules bind to form integrin-talin complexes, has a positive effect until about two seconds into the simulation, at which point it assumes a negative effect. This could be explained by the fact that while many free talin molecules still exist in the beginning of the simulation,  $k_T^+$  will positively contribute to the number of integrin-talin complexes, which in turn will positively effect the number of ligand-integrin-talin complexes. After all the initial free talin molecules are gone, however, an increase in  $k_T^+$  will mean that any later free talin molecules will be more likely to bind with integrins than other reactants. This will leave very few free talin to bind with ligand-integrin complexes, resulting in an overall decrease in the number of ligand-integrin-talin complexes. Thus, after the point in the simulation where the initial free talin molecules are gone,  $k_T^+$  will indirectly negatively effect the number of ligand-integrin-talin complexes.

**SRCs for talin.** Most rate parameters have a stronger effect on the amount of talin within the first five seconds of the simulation, after which they level off (Figure 8). The rate parameter  $k_T^+$ , which is the rate at which integrins and talin molecules bind, has the greatest negative effect during the first five seconds. It is interesting to note here that we might expect the parameter  $k_{LIT}^+$ , or the rate at which ligand-integrin complexes and talin molecules bind to form ligand-integrin-talin complexes, to have a greater effect on the number of talin molecules. However, while  $k_{LIT}^+$  does have a reasonable negative effect, it is overshadowed by the negative effect of  $k_T^+$ . This reflects the earlier inference that as  $k_T^+$  is increased, there is a decrease in the amount of ligand-integrin-talin complexes.

**SRCs for diffusing integrins.** For the amount of diffusing integrins, the only rate parameters that have an effect are  $k_D^+$  and  $k_D^-$ . This is reasonable given that the only reaction these integrins are involved in is either diffusion into the system or diffusion out of the system. Thus, the number of integrins in the pool outside of the system is entirely dependent on the rate at which integrins both leave and return to the pool.





**Figure 8.** Standardized regression coefficients for ligand-integrin-talin complexes (top) and free talin molecules (bottom) in the system over a time period of 20 seconds.

### 9. Conclusion and future work

The averaged results of our stochastic simulations are similar to the results from our deterministic simulation, with both simulation types indicating that initial focal adhesion formation occurs rapidly. This makes sense biologically because a motile cell receiving outside signaling would likely be required to react quickly in response. However, the speed of focal adhesion formation still needs to be determined experimentally, and our model only offers a possible outcome of such experiments.

The first step for future qualitative analysis for our system will be to demonstrate boundedness of our solutions. Boundedness of solutions is an important aspect to check given the context of our model. For instance, it would not make sense if we discovered that a solution approached infinity in finite time or attained negative values, since our system is modeling finite numbers of molecules over time. Additionally, the steady-state solutions to the system can be solved for, and

investigation can be conducted as to local and global stability of these steady-state solutions.

While the sensitivity analysis is a good start in analyzing how sensitive the model is to different parameters, a more accurate assessment could be accomplished with additional rate values from the literature. Additional methods of sensitivity analysis could shed more light on how the rate parameters affect the simulation output. These methods include factors prioritization, which would pinpoint the most influential factors in our system, and the Method of Morris, which would identify factors with negligible effects and allow us to narrow our focus. While the sensitivity analysis conducted thus far is only for the deterministic model, methods for sensitivity analysis of stochastic models are currently being investigated and are an area for future work.

In the future, we would like to find more accurate values from the literature for the rate parameters in our model, beginning with the parameters to which our model output is most sensitive, as indicated by our sensitivity analysis. We would also like to include additional molecules involved in focal adhesion formation, such as  $\text{PIP}_2$ , which increases the affinity of talin for integrins. We could then see if we retain a longer period of dynamic behavior in our stochastic results compared to our deterministic results. Additionally, we would like to allow for the diffusion of molecules other than integrins in and out of our system. This could result in more activity within our system as molecules are replenished.

### Acknowledgements

We would like to thank the Center for Undergraduate Research in Mathematics for their financial support of this work through an NSF grant (DMS-063664).

### References

- [Ananthakrishnan and Ehrlicher 2007] R. Ananthakrishnan and A. Ehrlicher, “The forces behind cell movement”, *Int. J. Biol. Sci.* **3**:5 (2007), 303–317.
- [Brauer and Nohel 1969] F. Brauer and J. A. Nohel, *The qualitative theory of ordinary differential equations: an introduction*, W. A. Benjamin, New York, 1969. Reprinted Dover, New York, 1989. Zbl 0179.13202
- [Calderwood et al. 2002] D. A. Calderwood, B. Yan, J. M. de Pereda, B. G. Alvarez, Y. Fujioka, R. C. Liddington, and M. H. Ginsberg, “The phosphotyrosine binding-like domain of talin activates integrins”, *J. Biol. Chem.* **277**:24 (2002), 21749–21758.
- [Cluzel et al. 2005] C. Cluzel, F. Saltel, J. Lussi, F. Puhle, B. A. Imhof, and B. Wehrle-Haller, “The mechanisms and dynamics of avb3 integrin clustering in living cells”, *J. Cell Biol.* **171**:2 (2005), 383–392.
- [Cozens-Roberts et al. 1990] C. Cozens-Roberts, D. A. Lauffenburger, and J. A. Quinn, “Receptor-mediated cell attachment and detachment kinetics”, *J. Biophys.* **58** (1990), 841–856.

- [Flaherty et al. 2007] B. Flaherty, J. P. McGarry, and P. E. McHugh, “Mathematical models of cell motility”, *Cell Biochem. Biophys.* **49** (2007), 14–28.
- [Fletcher and Theriot 2004] D. A. Fletcher and J. A. Theriot, “An introduction to cell motility for the physical scientist”, *Phys. Biol.* **1**:1 (2004).
- [Gallant and Garcia 2007] N. D. Gallant and A. J. Garcia, “Model of integrin-mediated cell adhesion strengthening”, *J. Biomech.* **40** (2007), 1301–1309.
- [Gibson and Bruck 2000] M. A. Gibson and J. Bruck, “Efficient exact stochastic simulation of chemical systems with many species and many channels”, *J. Phys. Chem. A* **104**:9 (2000), 1876–1889.
- [Gillespie 1977] D. T. Gillespie, “Exact stochastic simulation of coupled chemical reactions”, *J. Phys. Chem.* **81**:25 (1977), 2340–2361.
- [Gillespie 2007] D. T. Gillespie, “Stochastic simulation of chemical kinetics”, *Annu. Rev. Phys. Chem.* **58**:1 (2007), 35–55.
- [Gov 2006] N. S. Gov, “Modeling the size distribution of focal adhesions”, *Biophys. J.* **91** (2006), 2844–2847.
- [Higham 2008] D. J. Higham, “Modeling and simulating chemical reactions”, *SIAM Rev.* **50**:2 (2008), 347–368. MR 2009f:80015 Zbl 1144.80011
- [Hynes 1992] R. O. Hynes, “Integrins : versatility, modulation, and signaling in cell adhesion”, *Cell* **69**:1 (1992), 11–25.
- [Lauffenburger and Horwitz 1996] D. A. Lauffenburger and A. F. Horwitz, “Cell migration: a physically integrated molecular process”, *Cell* **84**:3 (1996), 359–369.
- [Lee et al. 2007] C. K. Lee, Y. M. Wang, L. S. Huang, and S. Lin, “Atomic force microscopy: determination of unbinding force, off rate and energy barrier for protein–ligand interaction”, *Micron* **38**:5 (2007), 446–461.
- [O’Day 2012] D. O’Day, *Human cell biology*, eBookIt.com, 2012.
- [Small et al. 2002] J. V. Small, T. Stradal, E. Vignat, and K. Rottner, “The lamellipodium: where motility begins”, *Trends Cell Biol.* **12**:3 (2002), 112–120.
- [Soll 1995] D. R. Soll, “The use of computers in understanding how animal cells crawl”, *Int. Rev. Cytol.* **163** (1995), 43–104.
- [Ward and Hammer 1993] M. D. Ward and D. A. Hammer, “A theoretical analysis for the effect of focal contact formation on cell-substrate attachment strength”, *J. Biophys.* **64** (1993), 936–959.
- [Wehrle-Haller 2006] B. Wehrle-Haller, “The role of integrins in cell migration”, in *Integrins and development*, edited by E. H. J. Danen, Landes Bioscience, Austin, TX, 2006. Madame Curie Biosci. Database.

Received: 2013-01-08      Revised: 2013-08-25      Accepted: 2013-08-29

blucher@ohsu.edu      *Department of Mathematics, University of Portland, 5000  
N. Willamette Boulevard, Portland, OR 97203, United States*

salas11@up.edu      *Department of Biology, University of Portland, 5000  
N. Willamette Boulevard, Portland, OR 97203, United States*

williams13@up.edu      *School of Engineering, University of Portland, 5000  
N. Willamette Boulevard, Portland, OR 97203, United States*

callende@up.edu      *Department of Mathematics, University of Portland, 5000  
N. Willamette Boulevard, Portland, OR 97203, United States*



# Investigating root multiplicities in the indefinite Kac–Moody algebra $E_{10}$

Vicky Klima, Timothy Shatley, Kyle Thomas and Andrew Wilson

(Communicated by Jim Haglund)

Following a procedure outlined by Kang, we view the generalized eigenspaces, known as root spaces, of the infinite dimensional Kac–Moody algebra  $E_{10}$  as generalized eigenspaces for representations of the finite dimensional special linear algebra  $A_9$ . Then, using the combinatorial representation theory of the special linear Lie algebras, we determine the dimensions of certain root spaces in  $E_{10}$ .

## 1. Introduction

For the past forty years, Kac–Moody algebras have been a rich area of study due to their numerous applications to other areas of mathematics and physics. Kac–Moody algebras are of one of three types (i) finite, (ii) affine, or (iii) indefinite. While the root multiplicities of finite and affine Kac–Moody algebras are well known [Kac 1990], we still do not have a general knowledge of root multiplicities in indefinite type algebras.

Building on the work of Feingold and Frenkel [1983], Kac, Moody, and Wakimoto [Kac et al. 1988] studied root multiplicities of the indefinite algebra  $E_{10}$  by considering this algebra as an extension of the affine algebra  $E_9$ . Later, Kang [1993a] developed a general construction for a class of indefinite algebras in which he built the larger algebra from a related affine or finite algebra and certain modules over that algebra. Kang’s construction has been used to study the multiplicities of the indefinite algebras  $HA_n^{(1)}$  [Benkart et al. 1995; Kang 1993b; 1994b; 1994a; Kang and Melville 1994; Hontz and Misra 2002a],  $HC_n^{(1)}$  [Klima and Misra 2008],  $HG_2^{(1)}$  [Hontz and Misra 2002b] and  $HD_4^{(3)}$  [Hontz and Misra 2002b], as well as  $EHA_1^{(1)}$  and  $EHA_2^{(2)}$  [Sthanumoorthy and Uma Maheswari 2012]. In this paper, as in [Kac et al. 1988], we consider the multiplicities of the indefinite algebra  $E_{10}$ .

*MSC2010:* 17B67.

*Keywords:* Kac–Moody, representation theory, combinatorial representation theory, root multiplicity. This research was partially supported by a CURM mini-grant funded by the NSF grant DMS-063664. Vicky Klima would like to thank Kailash Misra for introducing her to this problem and Jennifer Hontz for her helpful introduction to the programming necessary to complete the project.

However, in applying Kang's construction we choose to build  $E_{10}$  not from the infinite-dimensional  $E_9$  but rather from the finite-dimensional simple algebra  $A_9$ . Using a multiplicity formula also due to Kang [1994b] along with the combinatorial representation theory for  $A_9$  we determine multiplicities for roots up to degree  $-5$ . We recover some of the results in [Kac et al. 1988] and develop a recursive procedure to extend these results.

## 2. Background

A Lie algebra over  $\mathbb{C}$  is a vector space  $\mathfrak{g}$  over  $\mathbb{C}$ , with an antisymmetric, bilinear operation  $[\cdot, \cdot] : \mathfrak{g} \times \mathfrak{g} \rightarrow \mathfrak{g}$ , called the bracket, such that the following property—the Jacobi identity—holds:  $[a, [b, c]] = [b, [a, c]] + [[a, b], c]$  for all  $a, b, c \in \mathfrak{g}$ .

**Example 1.** The Lie algebra of  $2 \times 2$  trace zero complex matrices with the commutator bracket,  $[A, B] = AB - BA$ , is known as  $sl(2, \mathbb{C})$ . This Lie algebra has basis

$$\left\{ e = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, f = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, h = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \right\}$$

and relations  $[e, f] = h$ ,  $[h, e] = 2e$ ,  $[h, f] = -2f$ .

Any subalgebra  $\mathfrak{s}$  of  $\mathfrak{g}$  is a vector space over  $\mathbb{C}$  and thus a linear functional  $\alpha$  on  $\mathfrak{s}$  is simply a linear function that assigns to each element of  $\mathfrak{s}$  a corresponding complex number. The set of all linear functionals on  $\mathfrak{s}$  is itself a vector space over  $\mathbb{C}$ , denoted  $\mathfrak{s}^*$ . A Lie algebra  $\mathfrak{g}$  is  $\mathfrak{s}$ -diagonalizable if  $\mathfrak{g}$  can be written as a direct sum of subspaces

$$\mathfrak{g}_\alpha = \{g \in \mathfrak{g} \mid [s, g] = \alpha(s)g \text{ for all } s \in \mathfrak{s}\}.$$

**Example 2.** The Lie algebra  $\mathfrak{g} = sl(2, \mathbb{C})$  introduced in Example 1 is diagonalizable over its subalgebra  $\mathfrak{h} = \{h\}$ . Let  $\alpha \in \mathfrak{h}^*$  be defined by  $\alpha(h) = 2$ . Then

$$\mathfrak{g}_\alpha = \{g \in \mathfrak{g} \mid [h, g] = 2g \text{ for all } h \in \mathfrak{h}\} = \text{span}\{e\}.$$

Similarly,  $\mathfrak{g}_0 = \text{span}\{h\} = \mathfrak{h}$  and  $\mathfrak{g}_{-\alpha} = \text{span}\{f\}$ . Therefore  $\mathfrak{g}$  is  $\mathfrak{h}$ -diagonalizable with decomposition  $\mathfrak{g} = \mathfrak{g}_\alpha \oplus \mathfrak{g}_0 \oplus \mathfrak{g}_{-\alpha}$ .

We say a subalgebra  $\mathfrak{s}$  of  $\mathfrak{g}$  is *abelian* if  $[a, b] = 0$  for all  $a, b \in \mathfrak{s}$ . A *Cartan subalgebra*,  $\mathfrak{h}$ , of  $\mathfrak{g}$  is a maximal abelian subalgebra of  $\mathfrak{g}$  such that  $\mathfrak{g}$  is diagonalizable over  $\mathfrak{h}$ . Given a Lie algebra  $\mathfrak{g}$  and a Cartan subalgebra  $\mathfrak{h}$  we define the *roots* of  $\mathfrak{g}$  as those nonzero  $\alpha \in \mathfrak{h}^*$  such that

$$\mathfrak{g}_\alpha = \{g \in \mathfrak{g} \mid [h, g] = \alpha(h)g \text{ for all } h \in \mathfrak{h}\} \neq 0.$$

In this case we call  $\mathfrak{g}_\alpha$  the *root-space* associated with the root  $\alpha$  and  $\dim(\mathfrak{g}_\alpha)$  the *multiplicity* of the root  $\alpha$ .

The decomposition of  $sl(2, \mathbb{C})$  given in Example 2 is a root-space decomposition for this algebra because the subalgebra  $\mathfrak{h} = \text{span}\{h\}$  is a Cartan subalgebra of  $sl(2, \mathbb{C})$ . The root spaces in this example are simply the eigenspaces for  $\mathfrak{g}$  relative to  $\mathfrak{h}$ . In general, when the Cartan subalgebra  $\mathfrak{h}$  is of dimension greater than one the root spaces of  $\mathfrak{g}$  are generalized eigenspaces for  $\mathfrak{g}$  relative to  $\mathfrak{h}$ .

**2.1. Cartan matrices and the Weyl group.** The algebra  $sl(2, \mathbb{C})$  is a member of the family of finite dimensional simple Lie algebras  $A_n = sl(n + 1, \mathbb{C})$ , where each algebra  $A_n$  consists of the  $(n + 1) \times (n + 1)$  trace-zero complex matrices under the commutator bracket. This family is one of four families of classical finite dimensional simple Lie algebras, each of which can be modeled as collections of familiar types of matrices.

In the late nineteenth century, William Killing and Élie Cartan showed that these four classical families along with five exceptional families were the only finite-dimensional simple Lie algebras. Given a finite-dimensional simple Lie algebra  $\mathfrak{g}$  with Cartan subalgebra  $\mathfrak{h} = \text{span}\{h_i\}_{i=1}^n$  of dimension  $n$ , they described the root-system of  $\mathfrak{g}$  using a linearly independent set of *simple roots*,

$$\Pi = \{\alpha_i\}_{i=1}^n \subseteq \mathfrak{h}^*,$$

and recorded defining information for the simple roots in a Cartan matrix

$$A_{n \times n} = (a_{ij}) \text{ with } a_{ij} = \alpha_j(h_i). \tag{1}$$

Let  $\mathfrak{g}(A)$  be the Lie algebra with Cartan matrix  $A$ , Cartan subalgebra  $\mathfrak{h}$  and simple roots  $\{\alpha_i\}_{i=1}^n$ . For each  $i \in \{1, 2, \dots, n\}$  define the *simple reflection*  $r_i$  on  $\mathfrak{h}^*$  by  $r_i(\lambda) = \lambda - \lambda(h_i)\alpha_i$ . The *Weyl group* associated with  $A$  is the group generated by all simple reflections and for finite Lie algebras this group is also finite. If

$$\omega = \prod_{k=1}^t r_{i_k},$$

where  $t$  is minimal amongst all such expressions we say  $\omega$  is a reduced expression and call  $t$  the length of  $\omega$ , denoted  $\ell(\omega)$ . We can recover the set of all roots of the finite-dimensional simple Lie algebra  $\mathfrak{g}(A)$  by letting the Weyl group associated with  $A$  act on the set of simple roots. Each root space in a finite-dimensional simple Lie algebra,  $\mathfrak{g}(A)$ , is one-dimensional and therefore understanding the root system is equivalent to understanding the algebra itself. Hence, classifying the finite-dimensional simple Lie algebras amounts to classifying their Cartan matrices. See [Berman and Parshall 2002] for an excellent source on the historical development of Kac–Moody algebras beginning with Killing and Cartan’s work on the classification of the finite-dimensional simple Lie algebras.

Each Cartan matrix as given in (1) has the following properties, where the indices

range from 1 through  $n$ :

$$a_{ii} = 2 \qquad \text{for all } i. \tag{2}$$

$$a_{ij} \in \mathbb{Z}_{\leq 0} \qquad \text{for } i \neq j. \tag{3}$$

$$a_{ij} = 0 \iff a_{ji} = 0 \qquad \text{for all } i, j. \tag{4}$$

$$\det A \neq 0. \tag{5}$$

$$\text{Each proper principal minor of } A \text{ is positive.} \tag{6}$$

Every indecomposable matrix — that is, every matrix whose rows or columns cannot be permuted to block diagonal form — with these properties is the Cartan matrix (in the sense of (1)) for a unique (up to isomorphism) simple Lie algebra. Therefore we call any indecomposable square matrix with properties (2)–(6) an indecomposable *Cartan matrix*. In 1966, Jean-Pierre Serre developed a presentation of the unique (up to isomorphism) simple Lie algebra corresponding to a given indecomposable Cartan matrix  $A_{n \times n} = (a_{ij})$  using generators  $\{e_i, h_i, f_i\}_{i=1}^n$  and the following relations:

$$\left. \begin{aligned} [h_i, h_j] &= 0 \\ [h_i, e_j] &= a_{ij}e_j \\ [h_i, f_j] &= -a_{ij}f_j \end{aligned} \right\} \qquad \text{for all } i, j,$$

$$[e_i, f_j] = \begin{cases} h_i & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases} \tag{7}$$

$$\underbrace{[e_i, \dots [e_i, [e_i, e_j]] \dots]}_{1-a_{ij} \text{ times}} = \underbrace{[f_i, \dots [f_i, [f_i, f_j]] \dots]}_{1-a_{ij} \text{ times}} = 0 \quad \text{if } i \neq j.$$

In fact, Serre’s presentation applied to decomposable Cartan matrices as well, in which case the corresponding semisimple Lie algebra is the direct sum of the simple algebras associated with the indecomposable blocks.

**2.2. Kac–Moody algebras.** Kac [1968] and Moody [1968] independently extended Serre’s construction to a larger class of algebras, now known as Kac–Moody algebras. These algebras are defined in terms of a generalized Cartan matrix (GCM) which must meet only the conditions (2), (3), and (4) associated with a Cartan matrix. The Kac–Moody algebra  $\mathfrak{g}(A)$  defined by GCM  $A_{n \times n}$  is the algebra with generators  $\{e_i, f_i, h_i\}_{i=1}^n$  subject to Serre’s relations. We define the Weyl group associated with a GCM in the same way as that associated with a Cartan matrix; however, the Weyl group associated with a GCM may be infinite dimensional. Additionally, while always finite-dimensional, root spaces in a Kac–Moody algebra may be of dimensional greater than one. Roots of the Kac–Moody algebra  $\mathfrak{g}(A)$  that are reflections of one another — that is roots  $\alpha$  and  $\alpha'$  such that  $\alpha' = \omega(\alpha)$  for some  $\omega$



in the Weyl group associated with  $A$  — are called *conjugate* and *conjugate roots* have identical multiplicities.

The Kac–Moody algebra associated with indecomposable GCM  $A$  is of one of three types: finite, affine, or indefinite. If  $A$  is nonsingular and each proper principal minor of  $A$  is positive,  $\mathfrak{g}(A)$  is of finite type. In this case  $A$  is not only a generalized Cartan Matrix, but also a Cartan matrix and thus  $\mathfrak{g}(A)$  is a finite-dimensional simple Lie algebra whose root spaces are necessarily one-dimensional. If  $A$  is singular but each proper principal minor of  $A$  is positive,  $A$  is of affine type. While affine algebras are infinite dimensional and contain some roots of multiplicity greater than one, root multiplicities in these algebras are well-understood; see [Kac 1990, Chapter 6] for details. If  $\mathfrak{g}(A)$  is neither finite nor affine, it is indefinite. In general root-multiplicities for these infinite dimensional algebras are not well-understood.

Each GCM  $A_{n \times n} = (a_{ij})$ , of rank  $\ell$  is associated with a realization  $(\mathfrak{h}, \Pi, \Pi^\vee)$  where the Cartan subalgebra  $\mathfrak{h}$  is a  $2n - \ell$  dimensional complex vector space and the simple roots  $\Pi = \{\alpha_1, \alpha_2, \dots, \alpha_n\} \subseteq \mathfrak{h}^*$  and simple coroots

$$\Pi^\vee = \{h_1, h_2, \dots, h_n\} \subseteq \mathfrak{h}$$

are such that  $\Pi$  and  $\Pi^\vee$  are both linearly independent with  $\alpha_j(h_i) = a_{ij}$  for all  $i, j \in \{1, 2, \dots, n\}$ . All roots of the Kac–Moody algebra  $\mathfrak{g}(A)$  are either positive, that is the root can be written as a nonnegative integral linear combination of the simple roots, or negative, that is the root can be written as a nonpositive integral linear combination of the simple roots. Let  $\Delta, \Delta_+$ , and  $\Delta_-$  represent the set of roots, positive roots, and negative roots respectively. Then,  $\mathfrak{g}_+ = \bigoplus_{\alpha \in \Delta_+} \mathfrak{g}$  (resp.  $\mathfrak{g}_- = \bigoplus_{\alpha \in \Delta_-} \mathfrak{g}$ ) and we have the triangular decomposition  $\mathfrak{g} = \mathfrak{g}_- \oplus \mathfrak{h} \oplus \mathfrak{g}_+$ .

**2.3. Dynkin diagrams.** The generalized Cartan matrix is often presented in an equivalent, graphical form known as a Dynkin diagram. In this paper we will only consider Lie algebras with symmetric GCM. The *Dynkin diagram* for the Lie algebra  $\mathfrak{g}(A)$  with symmetric GCM,  $A_{n \times n}$ , is a graph with  $n$  vertices, each associated with a simple root  $\alpha_i$ , in which vertex  $i$  is connected to vertex  $j$  using  $a_{ij}^2 = [\alpha_j(h_i)]^2$  edges for  $i \neq j$ .

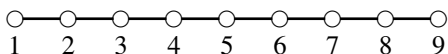
**Example 3.** The Lie algebra  $\mathfrak{g} = A_9 = sl(10, \mathbb{C})$  plays a key role in our work. Let  $E_{ij}$  be the  $10 \times 10$  matrix with  $(i, j)$  entry equal to one and all other entries zero. Then  $\mathfrak{g}$  is the 99-dimensional algebra generated by

$$\{e_i = E_{i,i+1}, f_i = E_{i+1,i}, h_i = E_{ii} - E_{i+1,i+1}\}_{i=1}^9,$$

with the nine-dimensional Cartan subalgebra  $\mathfrak{h} = \text{span}\{h_i\}_{i=1}^9$ . Define the linear functionals  $\epsilon_i$  on  $\mathfrak{h}$  by  $\epsilon_i(X) = X_{ii}$ . Then every root of  $\mathfrak{g}$  is a nonnegative or nonpositive integral linear combination of the simple roots  $\Pi = \{\alpha_i = \epsilon_i - \epsilon_{i+1}\}_{i=1}^9$ . Therefore,  $A_9$  has the Cartan matrix  $A = (a_{ij})$ , where, for  $i, j \in \{1, 2, \dots, 8\}$ ,

$$a_{ij} = \alpha_j(h_i) = \begin{cases} -1 & \text{if } |i - j| = 1, \\ 2 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

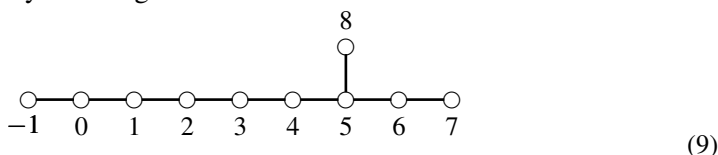
Here is the corresponding Dynkin diagram:



**Example 4.** We wish to explore root multiplicities in the Kac–Moody algebra  $E_{10}$ . Using the standard ordering of the simple roots,  $E_{10}$  is associated with the GCM  $A = (a_{ij})$ , where, for  $i, j \in \{-1, 0, 1, \dots, 8\}$ ,

$$a_{ij} = \alpha_j(h_i) = \begin{cases} -1 & \text{if } |i - j| = 1 \text{ and } i, j \in \{-1, 0, 1, \dots, 7\}, \\ -1 & \text{if } i = 5 \text{ and } j = 8, \\ 2 & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

or equivalently with Dynkin diagram



Note that the Dynkin diagram formed by removing vertex  $-1$  from (9) corresponds to GCM  $A' = (a_{ij})$  for  $i, j \in \{0, 1, \dots, 8\}$  with  $a_{ij}$  given in (8). Since  $\det A' = 0$  we see that  $E_{10}$  is of indefinite type.

**2.4. Lie algebra modules.** A vector space  $V$  over  $\mathbb{C}$  is a *module* over the Lie algebra  $\mathfrak{g}$  if there is a bilinear map from  $\mathfrak{g} \times V$  into  $V$  given by  $(g, v) \rightarrow g \cdot v$  such that

$$[x, y] \cdot v = x \cdot (y \cdot v) - y \cdot (x \cdot v) \quad \text{for all } x, y \in \mathfrak{g} \text{ and } v \in V. \tag{10}$$

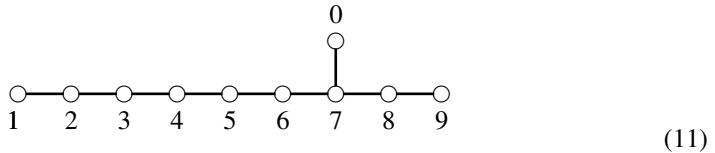
Every Lie algebra  $\mathfrak{g}$  is a module over itself via the *adjoint action*  $g \cdot v = [g, v]$ ; in this case (10) is simply the Jacobi identity. In our work we will only deal with modules over finite algebras and the definitions below pertain to such algebras. However, each of these ideas can be extended to all Kac–Moody algebras (see [Kac 1990, Chapter 9]).

Let  $\mathfrak{g}$  be a finite Lie algebra with Cartan subalgebra  $\mathfrak{h}$ . Given any  $\lambda \in \mathfrak{h}^*$  the  $\lambda$  *weight space*,  $V_\lambda$ , is defined as  $V_\lambda = \{v \in V \mid h \cdot v = \lambda(h) \cdot v \text{ for all } h \in \mathfrak{h}\}$ . If  $V_\lambda \neq 0$ , we call  $\lambda$  a *weight* of  $V$  and  $\dim(V_\lambda)$  the *weight multiplicity* of  $\lambda$  in  $V$ . The  $\mathfrak{g}$ -module  $V$  is a *highest weight module* if there exists a  $\lambda \in \mathfrak{h}^*$  and a  $v_\lambda \in V$ ,  $v_\lambda \neq 0$ , such that  $e_i \cdot v_\lambda = 0$  for all  $i \in \{1, 2, \dots, n\}$ ,  $h_i \cdot v_\lambda = \lambda(h_i)v_\lambda$  for all  $i \in \{1, 2, \dots, n\}$ , and  $V$  is generated by the images of  $v_\lambda$  under successive applications of the elements

$f_i \in \mathfrak{g}$  where the  $e_i, f_i,$  and  $h_i$  are the generators of  $\mathfrak{g}$  subject to Serre’s relations (7). In such a case, we call  $v_\lambda$  a *highest weight vector* and  $\lambda$  the *highest weight* of  $V$ .

### 3. The construction

If we remove vertex 8 from the Dynkin diagram for  $E_{10}$  given in (9) we have, up to the labeling of the vertices, the Dynkin diagram for  $\mathfrak{g} = A_9$  (with Cartan subalgebra  $\mathfrak{h}$ ). For convenience, we relabel the simple roots of  $\tilde{\mathfrak{g}} = E_{10}$  (with Cartan subalgebra  $\tilde{\mathfrak{h}}$ ) according to the following Dynkin diagram:



With this choice of ordering, restricting the domain of the simple roots  $\alpha_i$  ( $i = 1, \dots, 9$ ) of  $\tilde{\mathfrak{g}}$  to  $\mathfrak{h}$  gives the corresponding simple roots for  $\mathfrak{g}$ . Thus we use the same notation for the simple roots in both algebras.

Kang [1993a] introduced a construction for certain indefinite Kac–Moody algebras in which he builds the larger algebra,  $\tilde{\mathfrak{g}}$ , from a smaller algebra,  $\tilde{\mathfrak{g}}_0$ , a suitable  $\mathfrak{g}_0$ -module  $V$ , and  $V^*$ . In this section we specify Kang’s construction to the algebra  $\tilde{\mathfrak{g}} = E_{10}$ .

As one would expect,  $\mathfrak{g} = A_9$  plays an important role in our version of Kang’s construction. More specifically, we let  $\tilde{\mathfrak{g}}_0 = A_9 + \tilde{\mathfrak{h}}$ , and choose the highest weight  $\tilde{\mathfrak{g}}_0$ -module  $V = V(-\alpha_0)$  where  $\alpha_0 \in \tilde{\mathfrak{h}}^*$  is given by

$$\alpha_0(h_i) = \begin{cases} 2 & \text{if } i = 0, \\ -1 & \text{if } i = 7, \\ 0 & \text{if } i = 1, 2, 3, 4, 5, 6, 8, 9. \end{cases}
 \tag{12}$$

The Cartan matrix for  $\tilde{\mathfrak{g}}$  is nonsingular and thus the simple coroots  $\{h_i\}_{i=0}^9$  form a basis for  $\tilde{\mathfrak{h}}$ . Since  $\{h_i\}_{i=1}^9$  is a basis for the Cartan subalgebra of  $A_9$  we have  $\tilde{\mathfrak{g}}_0 = A_9 + \tilde{\mathfrak{h}} = A_9 \oplus \text{span}\{h_0\}$  and any  $\tilde{\mathfrak{g}}_0$ -module, specifically  $V(-\alpha_0)$ , will be an  $A_9$ -module as well via the restricted module action.

We can realize  $\tilde{\mathfrak{g}}$  as

$$\left( \bigoplus_{i \geq 1} \tilde{\mathfrak{g}}_{-i} \right) \oplus \tilde{\mathfrak{g}}_0 \oplus \left( \bigoplus_{i \geq 1} \tilde{\mathfrak{g}}_i \right),
 \tag{13}$$

where  $\tilde{\mathfrak{g}}_{-1} = V(-\alpha_0)$ ,  $\tilde{\mathfrak{g}}_1 = V(-\alpha_0)^*$ , and each subspace  $\tilde{\mathfrak{g}}_{-j}$  (resp.  $\tilde{\mathfrak{g}}_j$ ) with  $j > 1$  is a quotient of the space consisting of all brackets (in the free sense) of  $j$  vectors from  $V(-\alpha_0)$  (resp.  $V(-\alpha_0)^*$ ). Furthermore, each subspace  $\tilde{\mathfrak{g}}_{\pm j}$  with  $j > 1$  is completely reducible as a sum of highest-weight  $A_9$ -modules. See [Kang 1993a] for details regarding the construction including the bracket structure for (13).

### 4. Combinatorial representation theory of $A_9$

The construction of the previous section allows us to use the combinatorial representation theory of  $A_9$  to study root multiplicities in  $E_{10}$ . We say that an  $A_9$ -weight  $\mu_i = \sum_{i=1}^{10} k_i \epsilon_i$  is *dominant* if and only if  $k_1 \geq k_2 \geq \dots \geq k_{10} \geq 0$ . For example, restricting the domain of the weight  $-\alpha_0$  as defined in (12) to  $\mathfrak{h}$ , the Cartan subalgebra of  $A_9$ , we find  $-\alpha_0|_{\mathfrak{h}} = \epsilon_1 + \epsilon_2 + \dots + \epsilon_7$ , a dominant  $A_9$ -weight. We can express the dominant weights,  $\lambda$ , of  $A_9$  using certain ordered sets of positive integers, known as partitions and study weight multiplicities in  $V(\lambda)$  using related combinatorial objects, known as Young tableaux.

A *partition of the positive integer  $n$*  is a set  $\lambda = \{\lambda_1, \lambda_2, \dots, \lambda_t\}$  of positive integers written in weakly decreasing order such that  $\lambda_1 + \lambda_2 + \dots + \lambda_t = n$ . We call  $\ell(\lambda) = t$  the length of the partition  $\lambda$  and say that  $|\lambda| = n$ . We identify the dominant  $A_9$ -weight  $\lambda = \sum_{i=1}^{10} \lambda_i \epsilon_i$  with the partition  $\lambda = \{\lambda_1, \lambda_2, \dots, \lambda_t\}$  where  $t$  is the largest integer such that  $\lambda_t \neq 0$ . We compare partitions  $\lambda$  and  $\mu$  using the *dominance order on partitions*, in which we fill either  $\lambda$  or  $\mu$  with trailing zeros so that each partition is of the same length and say  $\lambda \geq \mu$  if and only if

$$\sum_{i=1}^m \lambda_i \geq \sum_{i=1}^m \mu_i \quad \text{for } m \text{ from } 1 \text{ to } \max\{\ell(\lambda), \ell(\mu)\}.$$

A *Young diagram* is a collection of boxes arranged in left-justified rows with a weakly decreasing number of boxes in each row and a *Young tableau* is a filling of the Young diagram with positive integers in such a way that the entries are weakly increasing across each row and strictly increasing down each column. For a given Young tableau,  $Y$ , let  $\lambda_i$  give the number of boxes in row  $i$  of the tableaux and  $\mu_i$  gives the number of  $i$ 's that appear in the filling of the tableaux. We say  $Y$  is of *shape*  $\lambda$  and *weight*  $\mu$ . The shape of a Young tableau is necessarily a partition while the weight of the tableau may or may not be.

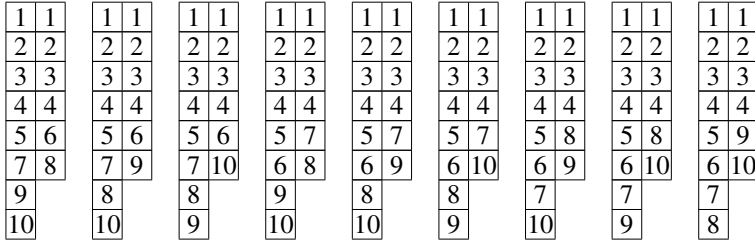
**Example 5.** The tableaux in Figure 1 are the only tableaux of shape

$$\{2, 2, 2, 2, 2, 1, 1\} = \{2^6, 1^2\}$$

and weight

$$\{2, 2, 2, 2, 1, 1, 1, 1, 1, 1\} = \{2^4, 1^6\}.$$

The basis vectors for the highest weight  $A_9$ -module with dominant highest-weight  $\lambda$  can be parameterized by the set of all Young tableaux of shape  $\lambda$ . If  $v \in V(\lambda)$  corresponds to the Young tableau  $Y$  then the weight of the vector  $v$  in  $V(\lambda)$  is the same as the weight of the Young tableau  $Y$ , leading to the following proposition.



**Figure 1.** Young diagrams for Example 5.

**Proposition 1** [Fulton 1997]. *Let  $\lambda$  be a dominant weight of  $A_9$ . The weight multiplicity of  $\mu$  in  $V(\lambda)$ , denoted  $\dim V(\lambda)_\mu$  is the number of Young tableaux of shape  $\lambda$  and weight  $\mu$ .*

Furthermore, every weight  $\mu = \sum_{i=1}^{10} \mu_i \epsilon_i$  of the highest-weight  $A_9$ -module with dominant highest-weight  $\lambda$  is conjugate to the dominant weight formed by rearranging the coefficients of the  $\epsilon_i$  in weakly decreasing order. If  $\mu$  is dominant then  $\mu$  may be identified with a partition and the number of Young tableaux of shape  $\lambda$  and weight  $\mu$  is known as the *Kostka number*  $\mathcal{K}_{\lambda, \mu}$ .

### 5. Root multiplicity calculations in $E_{10}$

Each positive root of  $E_{10}$  is conjugate to its negative which will be of the same multiplicity. Therefore we may restrict our studies to the negative roots of  $E_{10}$ ; let  $\alpha = -\sum_{i=0}^9 k_i \alpha_i$  be such a root. We call  $j = k_0$  the degree of the root  $\alpha$ . In this section we determine the multiplicities of roots of degree  $-5 \leq j \leq 0$ .

Viewing  $E_{10}$  as presented in (13), roots of degree zero appear as roots of  $\tilde{\mathfrak{g}}_0 = A_9 + \tilde{\mathfrak{h}}$  and hence as roots of the finite-dimensional simple Lie algebra  $A_9$ . These roots are of multiplicity one.

Roots of negative degree appear as weights of certain  $A_9$ -modules and each weight is conjugate to a dominant weight of the same multiplicity. Therefore, we will consider only dominant, negative roots in  $E_{10}$ . Let  $\alpha$  be any dominant  $E_{10}$  root of degree  $-j$  with  $j > 0$ . There exist positive integers  $k_i$  such that

$$\begin{aligned} \alpha &= -j\alpha_0 + \sum_{i=1}^9 k_i \alpha_i \\ &= j(\epsilon_1 + \epsilon_2 + \dots + \epsilon_7) + (j - k_1)\epsilon_1 + \sum_{i=2}^7 (j - k_i + k_{i+1})\epsilon_i + \sum_{i=8}^9 (k_{i-1} - k_i)\epsilon_i + k_9\epsilon_i \\ &= \{j - k_1, j - k_2 + k_1, j - k_3 + k_2, j - k_4 + k_3, j - k_5 + k_4, \\ &\quad j - k_6 + k_5, j - k_7 + k_6, k_7 - k_8, k_8 - k_9, k_9\}. \end{aligned}$$

The (necessarily positive) sums of the first  $t$  terms in  $\alpha$  for  $t = 1, 2, \dots, 10$  are

$j - k_1, 2j - k_2, 3j - k_3, 4j - k_4, 5j - k_5, 6j - k_6, 7j - k_7, 7j - k_8, 7j - k_9,$  and  $7j,$  respectively, with each  $k_i$  a positive integer, leading to the following proposition.

**Proposition 2.** *Every dominant degree  $-j$  root of  $E_{10}$  is a partition  $\alpha$  of  $7j$  such that  $\ell(\alpha) \leq 10$  and  $\alpha \leq \{j^7\}$  where  $\leq$  is the dominance order introduced on page 536.*

**5.1. Roots of degree negative one.** Using the construction of  $E_{10}$  presented in (13), roots of degree  $-1$  appear as weights of the  $A_9$ -module  $V(-\alpha_0) = V(\{1^7\})$ . By Proposition 2 any dominant weight of  $V(\{1^7\})$ ,  $\mu$  must be a partition of seven such that  $\mu < \{1^7\}$ . However any  $\mu < \{1^7\}$  can be a partition of at most six and thus  $\{1^7\}$  is the only dominant weight of  $V(-\alpha_0)$ . Therefore the roots of degree  $-1$  are of the form  $\alpha = k_1\epsilon_1 + k_2\epsilon_2 + \dots + k_{10}\epsilon_{10}$  where  $\{k_i\}_{i=1}^{10}$  is a permutation of  $\{1, 1, 1, 1, 1, 1, 0, 0, 0\}$ . By Proposition 1, each of these roots is of multiplicity

$$\mathcal{H}_{\{1^7\},\{1^7\}} = 1$$

**5.2. Roots of degree less than negative one—Kang’s formula.** Again viewing  $E_{10}$  as presented in (13), root spaces for roots of degree less than  $-1$  appear as weights of the  $A_9$ -module  $\bigoplus_{i \geq 1} \tilde{\mathfrak{g}}_{-i}$ . Kang [1994b] used the specific structure of  $\bigoplus_{i \geq 1} \tilde{\mathfrak{g}}_{-i}$  (see [Kang 1993a] for details regarding this structure), the Euler-Poincaré principle, and Kostant’s formula to develop both recursive and closed form multiplicity formulas for algebras with realizations such as the one given in (13). Theorem 3 gives Kang’s recursive formula as it is summarized in [Hontz and Misra 2002a] and as it pertains to  $E_{10}$ . One can find similar applications of this formula in [Benkart et al. 1995].

**Theorem 3.** *Let  $\alpha = \sum_{i=0}^9 k_i \alpha_i$  be a dominant root of  $E_{10}$  with  $\deg(\alpha) = k_0 = -j$ . Then for  $j \geq 2$ ,*

$$\text{mult}(\alpha) = \sum_{k=2}^j (-1)^k X_k(\alpha) - \sum_{k=2}^j (-1)^k Y_k(\alpha),$$

with

$$X_k(\alpha) = \sum_{\substack{\beta_1 > \dots > \beta_r \\ k_1 + \dots + k_r = k \\ k_1 \beta_1 + \dots + k_r \beta_r = \alpha}} \binom{\text{mult}(\beta_1)}{k_1} \dots \binom{\text{mult}(\beta_r)}{k_r}, \text{ and}$$

$$Y_k(\alpha) = \sum_{\substack{\omega \in W(S) \\ \ell(\omega) = k \\ \deg(\omega\rho - \rho) = -j}} \dim V(\omega\rho - \rho)_\alpha,$$

where  $\text{mult}(\beta_i)$  is the multiplicity of  $\beta_i$  as a root of  $E_{10}$ ,  $\rho \in \tilde{\mathfrak{h}}^*$  such that  $\rho(h_i) = 1$  for all  $i \in \{0, 1, \dots, 9\}$ ,  $V(\omega\rho - \rho)$  is the highest-weight  $A_9$ -module with highest

weight  $\omega\rho - \rho$ , and the reflections  $\omega \in W(S)$  can be built from simple reflections using the recursive procedure defined in Lemma 4.

For a more precise definition of the set  $W(S)$  along with a proof of Lemma 4 see [Kang 1993a].

**Lemma 4** [Kang 1993a]. *The only length one element of  $W(S)$  is  $\omega' = r_0$ . Suppose  $\omega = \omega' r_j$  and  $l(\omega) = l(\omega') + 1$ . Then  $\omega \in W(S)$  if and only if  $\omega' \in W(S)$  and  $\omega'(\alpha_j) = \sum_{i=0}^9 k_i \alpha_i$  where each  $k_i \geq 0$  and  $k_0 \neq 0$ .*

In the following examples we apply Theorem 3 to determine the multiplicities of specific degree  $-2$  and degree  $-3$  roots. Recall, we are using the ordering of the simple roots given in (11).

**Example 6.** In this example, we find the multiplicity in  $E_{10}$  of the degree  $-2$  root  $\alpha = -2\alpha_0 - \alpha_5 - 2\alpha_6 - 3\alpha_7 - 2\alpha_8 - \alpha_9$ . The root  $\alpha$  can be expressed by

$$\begin{aligned} \alpha &= 2\alpha_0 - \alpha_5 - 2\alpha_6 - 3\alpha_7 - 2\alpha_8 - \alpha_9 \\ &= 2(\epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4 + \epsilon_5 + \epsilon_6 + \epsilon_7) - (\epsilon_5 - \epsilon_6) - 2(\epsilon_6 - \epsilon_7) \\ &\quad - 3(\epsilon_7 - \epsilon_8) - 2(\epsilon_8 - \epsilon_9) - (\epsilon_9 - \epsilon_{10}) \\ &= 2\epsilon_1 + 2\epsilon_2 + 2\epsilon_3 + 2\epsilon_4 + \epsilon_5 + \epsilon_6 + \epsilon_7 + \epsilon_8 + \epsilon_9 + \epsilon_{10} \\ &= \{2^4, 1^6\} \end{aligned}$$

and  $\text{mult}(\alpha) = X_2(\alpha) - Y_2(\alpha)$ , where  $X_2$  and  $Y_2$  are defined in Theorem 3.

Given that all degree  $-1$  roots are of multiplicity one, we have

$$\begin{aligned} X_2(\alpha) &= \sum_{\substack{\beta_1 \\ 2\beta_1 = \alpha}} \binom{1}{2} + \sum_{\substack{\beta_1 > \beta_2 \\ \beta_1 + \beta_2 = \alpha}} \binom{1}{1} \binom{1}{1} \\ &= \text{the number of pairs } (\beta_1, \beta_2) \text{ of roots of degree } -1 \\ &\quad \text{such that } \beta_1 > \beta_2 \text{ and } \beta_1 + \beta_2 = \alpha. \quad (14) \end{aligned}$$

Let  $\beta_1$  and  $\beta_2$  be roots of degree  $-1$  such that  $\beta_1 + \beta_2 = \{2^4, 1^6\}$ . Then,  $\beta_1$  and  $\beta_2$  can each be viewed as ordered sets whose terms are permutations of  $\{1, 1, 1, 1, 1, 1, 1, 0, 0, 0\}$ . The first term in both  $\beta_1$  and  $\beta_2$  must be one as this is the only way for their sum to be two. The same statement holds for the second, third, and fourth terms of  $\beta_1$  and  $\beta_2$ . The remaining six terms of  $\beta_1$  could then be any of the  $C(6, 3)$  permutations of  $\{1, 1, 1, 0, 0, 0\}$ . Once we have determined  $\beta_1$ ,  $\beta_2 = \{2^4, 1^6\} - \beta_1$  is fixed. Therefore, we have  $C(6, 3)$  pairs  $(\beta_1, \beta_2)$  of distinct degree  $-2$  roots with  $\beta_1 + \beta_2 = \{2^4, 1^6\}$ , exactly half which will be such that  $\beta_1 > \beta_2$ . Hence,  $X_2(\{2^6, 1^4\}) = C(6, 3)/2 = 10$ .

Next we turn our attention to the calculation of  $Y_2(\alpha)$ . To do this we must first find all  $\omega \in W(S)$  of length two such that  $\text{deg}(\omega\rho - \rho) = -2$ . Lemma 4 implies

that any  $\omega \in W(S)$  of length two will be of the form  $\omega = r_0 r_j$  for  $j \in \{0, 1, \dots, 9\}$  where  $r_0(\alpha_j) = \sum_{i=0}^9 k_i \alpha_i$  for some  $k_i$  with each  $k_i \geq 0$  and  $k_0 \neq 0$ . For any simple root  $\alpha_j$ ,  $r_0(\alpha_j) = \alpha_j - \alpha_j(h_0)\alpha_0$ . Referring to the Dynkin diagram for  $E_{10}$  given in (11), we see

$$\alpha_j(h_0) = \begin{cases} 2 & \text{if } j = 0, \\ -1 & \text{if } j = 7, \\ 0 & \text{otherwise.} \end{cases}$$

and thus

$$r_0(\alpha_j) = \begin{cases} \alpha_0 - 2\alpha_0 = -\alpha_0 & \text{if } j = 0, \\ \alpha_7 + \alpha_0 & \text{if } j = 7, \\ \alpha_j & \text{otherwise.} \end{cases}$$

Observe that the only  $\omega \in W(S)$  of length two is  $\omega = r_0 r_7$  and for this choice of  $\omega$ ,  $\omega\rho - \rho$  is of degree  $-2$ . Specifically,

$$\begin{aligned} \omega\rho - \rho &= r_0 r_7(\rho) - \rho \\ &= r_0[\rho - \rho(h_7)\alpha_7] - \rho \\ &= r_0(\rho - \alpha_7) - \rho \quad (\text{since } \rho(h_i) = 1 \text{ for all } i) \\ &= -2\alpha_0 - \alpha_7 \\ &= 2\epsilon_1 + 2\epsilon_2 + 2\epsilon_3 + 2\epsilon_4 + 2\epsilon_5 + 2\epsilon_6 + \epsilon_7 + \epsilon_8. \end{aligned}$$

Therefore,

$$\begin{aligned} Y_2(\alpha) &= \sum_{\substack{\omega \in W(S) \\ \ell(\omega)=2 \\ \deg(\omega\rho-\rho)=-2}} \dim V(\omega\rho - \rho)_\alpha \\ &= \dim V(\{2^6, 1^2\})_\alpha \\ &= \mathcal{H}_{\{2^6, 1^2\}, \alpha} \quad (\text{by Proposition 1}) \\ &= \mathcal{H}_{\{2^6, 1^2\}, \{2^4, 1^6\}} \\ &= 9 \quad (\text{by Example 5}) \end{aligned} \tag{15}$$

and  $\text{mult}(\alpha) = \text{mult}(\{2^4, 1^6\}) = X_2(\{2^4, 1^6\}) - Y_2(\{2^4, 1^6\}) = 10 - 9 = 1$ .

**Example 7.** In this example we find the multiplicity of the  $E_{10}$  root

$$\alpha = -3\alpha_0 - \alpha_2 - 2\alpha_3 - 3\alpha_4 - 4\alpha_5 - 5\alpha_6 - 6\alpha_7 - 4\alpha_8 - 2\alpha_9.$$

The root  $\alpha$  is of degree  $-3$  and thus  $\text{mult}(\alpha) = X_2(\alpha) - X_3(\alpha) - Y_2(\alpha) + Y_3(\alpha)$  where  $X_2, X_3, Y_2$ , and  $Y_3$  are defined in Theorem 3.



Given that all degree  $-1$  and degree  $-2$  roots are of multiplicity one, we have

$$\begin{aligned}
 X_2(\alpha) &= \sum_{\substack{\beta_1 > \beta_2 \\ \beta_1 + \beta_2 = \alpha}} \underbrace{\binom{1}{1} \binom{1}{1}}_1 \\
 &= \text{The number of pairs } (\beta_1, \beta_2) \text{ with } \deg(\beta_1) = -2, \\
 &\quad \deg(\beta_2) = -1, \text{ and } \beta_1 + \beta_2 = \alpha. \quad (16)
 \end{aligned}$$

Let  $\beta_1$  be a root of degree  $-2$  and  $\beta_2$  be a root of degree  $-1$  such that

$$\begin{aligned}
 \beta_1 + \beta_2 &= \alpha \\
 &= -3\alpha_0 - \alpha_2 - 2\alpha_3 - 3\alpha_4 - 4\alpha_5 - 5\alpha_6 - 6\alpha_7 - 4\alpha_8 - 2\alpha_9 \\
 &= 3(\epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4 + \epsilon_5 + \epsilon_6 + \epsilon_7) - (\epsilon_2 - \epsilon_3) - 2(\epsilon_3 - \epsilon_4) - 3(\epsilon_4 - \epsilon_5) \\
 &\quad - 4(\epsilon_5 - \epsilon_6) - 5(\epsilon_6 - \epsilon_7) - 6(\epsilon_7 - \epsilon_8) - 4(\epsilon_8 - \epsilon_9) - 2(\epsilon_9 - \epsilon_{10}) \\
 &= 3\epsilon_1 + 2\epsilon_2 + 2\epsilon_3 + 2\epsilon_4 + 2\epsilon_5 + 2\epsilon_6 + 2\epsilon_7 + 2\epsilon_8 + 2\epsilon_9 + 2\epsilon_{10} \\
 &= \{3, 2^9\}.
 \end{aligned}$$

Then,  $\beta_1$  and  $\beta_2$  can be viewed as ordered sets whose terms are permutations of  $\{2, 2, 2, 2, 1, 1, 1, 1, 1, 1\}$  and  $\{1, 1, 1, 1, 1, 1, 1, 0, 0, 0\}$  respectively. The first term in  $\beta_1$  must be two with the first term of  $\beta_2$  being one, as this is the only way for their sum to be three. The remaining nine terms of  $\beta_1$  could then be any of the  $C(9, 3)$  permutations of  $\{2, 2, 2, 1, 1, 1, 1, 1, 1\}$ . Once we have determined  $\beta_1$ ,  $\beta_2 = \{3, 2^9\} - \beta_1$  is fixed. Therefore  $X_2(\{3, 2^9\}) = C(9, 3) = 84$ .

We can also simplify  $X_3(\alpha) = X_3(\{3, 2^9\})$  using the fact that all degree  $-1$  and degree  $-2$  roots are of multiplicity one.

$$\begin{aligned}
 X_3(\alpha) &= \sum_{\substack{\beta_1 \\ 3\beta_1 = \alpha}} \binom{1}{3} + \sum_{\substack{\beta_1 \neq \beta_2 \\ 2\beta_1 + \beta_2 = \alpha}} \binom{1}{2} \binom{1}{1} + \sum_{\substack{\beta_1 > \beta_2 > \beta_3 \\ \beta_1 + \beta_2 + \beta_3 = \alpha}} \binom{1}{1} \binom{1}{1} \binom{1}{1} \\
 &= \text{The number of triples } (\beta_1, \beta_2, \beta_3) \\
 &\quad \text{with } \beta_1 > \beta_2 > \beta_3 \text{ and } \beta_1 + \beta_2 + \beta_3 = \alpha. \quad (17)
 \end{aligned}$$

Let  $\beta_1, \beta_2$  and  $\beta_3$  be degree  $-1$  roots such that  $\beta_1 + \beta_2 + \beta_3 = \alpha = \{3, 2^9\}$ . Then,  $\beta_1, \beta_2$ , and  $\beta_3$  can each be viewed as ordered sets whose terms are permutations of  $\{1, 1, 1, 1, 1, 1, 1, 0, 0, 0\}$ . The first term in  $\beta_1, \beta_2$ , and  $\beta_3$  must each be one as this is the only way for their sum to be three. The remaining nine terms of  $\beta_1$  could then be any of the  $C(9, 3)$  permutations of  $\{1, 1, 1, 1, 1, 1, 0, 0, 0\}$ . Three of the nine remaining terms in  $\beta_1$  will be zero. The corresponding terms in  $\beta_2$  and  $\beta_3$  must both be one as this is the only way for the terms to sum to two. The remaining six terms of  $\beta_2$  could then be any of the  $C(6, 3)$  permutations of  $\{1, 1, 1, 0, 0, 0\}$ . Once we have determined  $\beta_1$  and  $\beta_2, \beta_3 = \{3, 2^9\} - \beta_1 - \beta_2$  is fixed. Therefore, we have

$C(9, 3) \cdot C(6, 3)$  pairs  $(\beta_1, \beta_2)$  of distinct degree  $-2$  roots with  $\beta_1 + \beta_2 + \beta_3 = \{3, 2^9\}$ , exactly  $1/3!$  of which will be such that  $\beta_1 > \beta_2 > \beta_3$ . Hence,

$$X_3(\{3, 2^9\}) = (C(9, 3) \cdot C(6, 2))/3! = 280.$$

Next we turn our attention to the calculation of  $Y_2(\alpha)$  and  $Y_3(\alpha)$ . Recall that

$$Y_2(\alpha) = \sum_{\substack{\omega \in W(S) \\ \ell(\omega)=2 \\ \deg(\omega\rho-\rho)=-3}} \dim V(\omega\rho - \rho)_\alpha.$$

However, in Example 6 we found all  $\omega \in W(S)$  of length two and none of these were of degree  $-3$ . Therefore,

$$Y_2(\alpha) = 0. \tag{18}$$

To evaluate  $Y_3(\alpha)$  we must first determine all  $\omega \in W(S)$  of length three such that  $\deg(\omega\rho - \rho) = -3$ . Since the only  $\omega \in W(S)$  of length two is  $r_0r_7$ , Lemma 4 implies that any  $\omega \in W(S)$  of length three will be of the form  $\omega = r_0r_7r_j$  for  $j \in \{0, 1, \dots, 9\}$  where  $r_0r_7(\alpha_j) = \sum_{i=0}^9 k_i\alpha_i$  for some  $k_i$  with each  $k_i \geq 0$  and  $k_0 \neq 0$ . But then,

$$r_0r_7(\alpha_j) = \begin{cases} -\alpha_7 & \text{if } j = 0, \\ \alpha_0 + \alpha_6 + \alpha_7 & \text{if } j = 6, \\ \alpha_0 + \alpha_7 + \alpha_8 & \text{if } j = 8, \\ \alpha_j & \text{otherwise.} \end{cases}$$

and so the only  $\omega \in W(S)$  of length 3 are  $\omega_1 = r_0r_7r_6$  and  $\omega_2 = r_0r_7r_8$ . Since

$$\omega_1\rho - \rho = 3\epsilon_1 + 3\epsilon_2 + 3\epsilon_3 + 3\epsilon_4 + 3\epsilon_5 + 2\epsilon_6 + 2\epsilon_7 + 2\epsilon_8$$

and

$$\omega_2\rho - \rho = 3\epsilon_1 + 3\epsilon_2 + 3\epsilon_3 + 3\epsilon_4 + 3\epsilon_5 + 3\epsilon_6 + \epsilon_7 + \epsilon_8 + \epsilon_9,$$

each of which is of degree  $-3$ , we have

$$\begin{aligned} Y_3(\alpha) &= \sum_{\substack{\omega \in W(S) \\ \ell(\omega)=3 \\ \deg(\omega\rho-\rho)=-3}} \dim V(\omega\rho - \rho)_\alpha \\ &= \dim V(3\epsilon_1 + 3\epsilon_2 + 3\epsilon_3 + 3\epsilon_4 + 3\epsilon_5 + 2\epsilon_6 + 2\epsilon_7 + 2\epsilon_8)_\alpha \\ &\quad + \dim V(3\epsilon_1 + 3\epsilon_2 + 3\epsilon_3 + 3\epsilon_4 + 3\epsilon_5 + 3\epsilon_6 + \epsilon_7 + \epsilon_8 + \epsilon_9)_\alpha \\ &= \mathcal{H}_{\{3^5, 2^3\}, \{3, 2^9\}} + \mathcal{H}_{\{3^6, 1^3\}, \{3, 2^9\}} \\ &= 120 + 84 = 204. \end{aligned} \tag{19}$$

Therefore,  $\text{mult}(\alpha) = \text{mult}(\{3, 2^9\}) = X_2 - X_3 - Y_2 + Y_3 = 84 - 280 - 0 + 204 = 8$ .

**Theorem 5.** *The only dominant degree  $-2$  root of  $E_{10}$  is  $\{2^4, 1^6\}$  and this root is of multiplicity one.*

*Proof.* If  $\alpha$  is a dominant degree  $-2$  root of  $E_{10}$  then  $\alpha$  meets the conditions given in Proposition 2 for  $j = 2$  with  $0 \neq \text{mult}(\alpha) = X_2(\alpha) - Y_2(\alpha)$ . Using the counting methods demonstrated in Example 6, we have found  $X_2(\alpha)$  as stated in (14),  $Y_2(\alpha)$  as stated in (15), and  $\text{mult}(\alpha)$  for each potential dominant degree  $-2$  root, as follows:

$\alpha$	$X_2$	$Y_2$	$\text{mult}(\alpha)$
$\{2^7\}$	0	0	0
$\{2^6, 1^2\}$	1	1	0
$\{2^5, 1^4\}$	3	3	0
$\{2^4, 1^6\}$	10	9	1

The table shows that  $\{2^4, 1^6\}$  is the only dominant  $E_{10}$  root of degree  $-2$ . □

**Theorem 6.** *The only dominant degree  $-3$  roots of  $E_{10}$  are  $\{3, 2^9\}$  and  $\{3^2, 2^7, 1\}$  which are of multiplicities eight and one respectively.*

*Proof.* If  $\alpha$  is a dominant degree  $-3$  root of  $E_{10}$  then  $\alpha$  meets the conditions given in Proposition 2 for  $j = 3$  and  $0 \neq \text{mult}(\alpha) = X_2(\alpha) - X_3(\alpha) - Y_2(\alpha) + Y_3(\alpha)$ . Using the counting methods demonstrated in Example 7, we have found  $X_2(\alpha)$  as stated in (16),  $X_3(\alpha)$  as stated in (17),  $Y_2(\alpha)$  as stated in (18),  $Y_3(\alpha)$  as stated in (19), and  $\text{mult}(\alpha)$  for each potential dominant degree  $-3$  root, as follows:

$\alpha$	$X_2$	$X_3$	$Y_2$	$Y_3 = \mathcal{H}_{\{3^5, 2^3\}, \alpha} + \mathcal{H}_{\{3^6, 1^3\}, \alpha}$	$\text{mult}(\alpha)$
$\{3^7\}$	0	*0	0	0	0
$\{3^6, 2, 1\}$	0	*0	0	0	0
$\{3^6, 1^3\}$	0	*1	0	0 + 1	0
$\{3^5, 2^3\}$	0	*1	0	1 + 0	0
$\{3^5, 2^2, 1^2\}$	0	*2	0	1 + 1	0
$\{3^5, 2, 1^4\}$	0	*6	0	2 + 4	0
$\{3^4, 2^4, 1\}$	0	6	0	4 + 2	0
$\{3^4, 2^3, 1^3\}$	1	15	0	7 + 7	0
$\{3^3, 2^6\}$	0	15	0	10 + 5	0
$\{3^3, 2^5, 1^2\}$	5	40	0	20 + 15	0
$\{3^2, 2^7, 1\}$	21	105	0	50 + 35	1
$\{3, 2^9\}$	82	280	0	120 + 84	8

The table shows that  $\{3, 2^9\}$  and  $\{3^2, 2^7, 1\}$  are the only dominant roots of degree  $-3$ . □

We developed Maple worksheets to automate the multiplicity calculations. (For examples see [mathsci.appstate.edu/~vlw/E10mult.html](http://mathsci.appstate.edu/~vlw/E10mult.html).) These worksheets apply Kang’s multiplicity formula using Maple packages by John Stembridge [2004; 2005] to do the combinatorial calculations. Using the worksheets we have found all dominant  $E_{10}$  roots of degree up to  $-5$ .

**Theorem 7.** *The dominant degree  $-4$  roots of  $\tilde{\mathfrak{g}} = E_{10}$  are  $\{4, 3^6, 2^3\}$ ,  $\{3^9, 1\}$ , and  $\{3^8, 2^2\}$ , with multiplicities of 1, 1, and 8 respectively.*

**Theorem 8.** *The dominant degree  $-5$  roots of  $\tilde{\mathfrak{g}} = E_{10}$  are  $\{4^6, 3^3, 2\}$ ,  $\{5, 4^3, 3^6\}$ , and  $\{4^5, 3^5\}$  with multiplicities 1, 1, and 8 respectively.*

### 6. Conclusions

As in [Kac et al. 1988] we have studied root multiplicities in  $E_{10}$ . We have worked in the basis  $\mathcal{B}_\epsilon = \{\epsilon_i\}_{i=1}^{10}$  whereas the authors of [Kac et al. 1988] use the basis  $\mathcal{B}_{\alpha'} = \{\alpha'_i\}_{i=-1}^8$  ordered according to the Dynkin diagram given in (9). Using the transition matrix from the basis  $B_\epsilon$  to the basis  $\mathcal{B}_{\alpha'}$ ,

$$\begin{bmatrix} 6/7 & -1/7 & -1/7 & -1/7 & -1/7 & -1/7 & -1/7 & -1/7 & -1/7 & -1/7 \\ 5/7 & 5/7 & -2/7 & -2/7 & -2/7 & -2/7 & -2/7 & -2/7 & -2/7 & -2/7 \\ 4/7 & 4/7 & 4/7 & -3/7 & -3/7 & -3/7 & -3/7 & -3/7 & -3/7 & -3/7 \\ 3/7 & 3/7 & 3/7 & 3/7 & -4/7 & -4/7 & -4/7 & -4/7 & -4/7 & -4/7 \\ 2/7 & 2/7 & 2/7 & 2/7 & 2/7 & -5/7 & -5/7 & -5/7 & -5/7 & -5/7 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & -6/7 & -6/7 & -6/7 & -6/7 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ -1/7 & -1/7 & -1/7 & -1/7 & -1/7 & -1/7 & -1/7 & -1/7 & -1/7 & -1/7 \end{bmatrix}$$

and remembering that any permutation of a dominant  $E_{10}$  root is again a root of the same multiplicity we were able to compare our results with those of [Kac et al. 1988].

We have found the multiplicity of 3527 roots in  $E_{10}$  with negative degree. The majority (3442) of these roots have coefficients for  $\alpha'_{-1}$  of either  $-1$  or  $-2$  and their multiplicities agree with those stated in [Kac et al. 1988]. The root

$$\{1, 3^9\}_{\mathcal{B}_\epsilon} = \{-3, -4, -5, -6, -7, -8, -9, -6, -3, -4\}_{\mathcal{B}_{\alpha'}}$$

was not addressed in [Kac et al. 1988]. This root is conjugate to the dominant root  $\{3^9, 1\}_{\mathcal{B}_\epsilon}$  and thus is of multiplicity one. The remaining 84 new  $E_{10}$  roots are conjugate to  $\{4^6, 3^3, 2\}_{\mathcal{B}_\epsilon}$  and are also of multiplicity one.

## References

- [Benkart et al. 1995] G. Benkart, S.-J. Kang, and K. C. Misra, “Indefinite Kac–Moody algebras of special linear type”, *Pacific J. Math.* **170**:2 (1995), 379–404. MR 96k:17037 Zbl 0857.17020
- [Berman and Parshall 2002] S. Berman and K. H. Parshall, “Victor Kac and Robert Moody: their paths to Kac–Moody Lie algebras”, *Math. Intelligencer* **24**:1 (2002), 50–60. MR 2003b:17029 Zbl 1048.17001
- [Feingold and Frenkel 1983] A. J. Feingold and I. B. Frenkel, “A hyperbolic Kac–Moody algebra and the theory of Siegel modular forms of genus 2”, *Math. Ann.* **263**:1 (1983), 87–144. MR 86a:17006 Zbl 0489.17008
- [Fulton 1997] W. Fulton, *Young tableaux*, London Mathematical Society Student Texts **35**, Cambridge University Press, 1997. MR 99f:05119 Zbl 0878.14034
- [Hontz and Misra 2002a] J. Hontz and K. C. Misra, “On root multiplicities of  $HA_n^{(1)}$ ”, *Internat. J. Algebra Comput.* **12**:3 (2002), 477–508. MR 2003d:17029 Zbl 1035.17035
- [Hontz and Misra 2002b] J. Hontz and K. C. Misra, “Root multiplicities of the indefinite Kac–Moody Lie algebras  $HD_4^{(3)}$  and  $HG_2^{(1)}$ ”, *Comm. Algebra* **30**:6 (2002), 2941–2959. MR 2003f:17028 Zbl 1058.17014
- [Kac 1968] V. G. Kac, “Simple irreducible graded Lie algebras of finite growth”, *Izv. Akad. Nauk SSSR Ser. Mat.* **32** (1968), 1323–1367. In Russian; translated in *Math. USSR, Isv.* **2**(1968), 1271–1311. MR 41 #4590 Zbl 0222.17007
- [Kac 1990] V. G. Kac, *Infinite-dimensional Lie algebras*, 3rd ed., Cambridge University Press, 1990. MR 92k:17038 Zbl 0716.17022
- [Kac et al. 1988] V. G. Kac, R. V. Moody, and M. Wakimoto, “On  $E_{10}$ ”, pp. 109–128 in *Differential geometrical methods in theoretical physics* (Como, 1987), edited by K. Bleuler and M. Werner, NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. **250**, Kluwer Acad. Publ., Dordrecht, 1988. MR 90e:17031 Zbl 0674.17007
- [Kang 1993a] S.-J. Kang, “Kac–Moody Lie algebras, spectral sequences, and the Witt formula”, *Trans. Amer. Math. Soc.* **339**:2 (1993), 463–493. MR 93m:17013 Zbl 0794.17014
- [Kang 1993b] S.-J. Kang, “Root multiplicities of the hyperbolic Kac–Moody Lie algebra  $HA_1^{(1)}$ ”, *J. Algebra* **160**:2 (1993), 492–523. MR 94i:17031 Zbl 0828.17027
- [Kang 1994a] S.-J. Kang, “On the hyperbolic Kac–Moody Lie algebra  $HA_1^{(1)}$ ”, *Trans. Amer. Math. Soc.* **341**:2 (1994), 623–638. MR 94d:17033 Zbl 0828.17026
- [Kang 1994b] S.-J. Kang, “Root multiplicities of Kac–Moody algebras”, *Duke Math. J.* **74**:3 (1994), 635–666. MR 95c:17036 Zbl 0823.17031
- [Kang and Melville 1994] S.-J. Kang and D. J. Melville, “Root multiplicities of the Kac–Moody algebras  $HA_n^{(1)}$ ”, *J. Algebra* **170**:1 (1994), 277–299. MR 95m:17018 Zbl 0828.17028
- [Klima and Misra 2008] V. W. Klima and K. C. Misra, “Root multiplicities of the indefinite Kac–Moody algebras of symplectic type”, *Comm. Algebra* **36**:2 (2008), 764–782. MR 2009c:17036 Zbl 1132.17014
- [Moody 1968] R. V. Moody, “A new class of Lie algebras”, *J. Algebra* **10** (1968), 211–230. MR 37 #5261
- [Stembridge 2004] J. Stembridge, “The coxeter package, a Maple package for working with root systems and finite coxeter groups”, website, 2004, <http://www.math.lsa.umich.edu/~jrs/maple.html>.
- [Stembridge 2005] J. Stembridge, “The SF package, a Maple package for computations symmetric functions”, website, 2005, <http://www.math.lsa.umich.edu/~jrs/maple.html>.

[Sthanumoorthy and Uma Maheswari 2012] N. Sthanumoorthy and A. Uma Maheswari, “Structure and root multiplicities for two classes of extended hyperbolic Kac–Moody algebras  $EHA_1^{(1)}$  and  $EHA_2^{(2)}$  for all cases”, *Comm. Algebra* **40**:2 (2012), 632–665. MR 2889486 Zbl 1267.17030

Received: 2013-01-17      Accepted: 2013-06-02

klimavw@appstate.edu	<i>Department of Mathematical Sciences, Appalachian State University, 121 Bodenheimer Drive, Boone, NC 28607, United States</i>
tshatl1@lsu.edu	<i>Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803-4918, United States</i>
thomaska1@appstate.edu	<i>Department of Mathematical Sciences, Appalachian State University, 121 Bodenheimer Drive, Boone, NC 28607, United States</i>
wilsonat@appstate.edu	<i>Department of Mathematical Sciences, Appalachian State University, 121 Bodenheimer Drive, Boone, NC 28607, United States</i>

# On a state model for the SO(2n) Kauffman polynomial

Carmen Caprau, David Heywood and Dionne Ibarra

(Communicated by Colin Adams)

François Jaeger presented the two-variable Kauffman polynomial of an unoriented link  $L$  as a weighted sum of HOMFLY-PT polynomials of oriented links associated with  $L$ . Murakami, Ohtsuki and Yamada (MOY) used planar graphs and a recursive evaluation of these graphs to construct a state model for the  $sl(n)$ -link invariant (a one-variable specialization of the HOMFLY-PT polynomial). We apply the MOY framework to Jaeger's work, and construct a state summation model for the SO(2n) Kauffman polynomial.

## 1. Introduction

The SO(2n) Kauffman polynomial  $\llbracket L \rrbracket$  of an unoriented link  $L$  is a Laurent polynomial in  $q$ , uniquely determined by the following axioms:

- (1)  $\llbracket L_1 \rrbracket = \llbracket L_2 \rrbracket$ , whenever  $L_1$  and  $L_2$  are regular isotopic links.
- (2)  $\llbracket \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \rrbracket - \llbracket \begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array} \rrbracket = (q - q^{-1}) \left( \llbracket \begin{array}{c} \frown \\ \smile \end{array} \rrbracket - \llbracket \begin{array}{c} \smile \\ \frown \end{array} \rrbracket \right)$ .
- (3)  $\llbracket \bigcirc \rrbracket = \frac{q^{2n-1} - q^{1-2n}}{q - q^{-1}} + 1$ .
- (4)  $\llbracket \begin{array}{c} \text{loop} \\ \diagdown \diagup \end{array} \rrbracket = q^{2n-1} \llbracket \begin{array}{c} \text{arc} \\ \diagdown \diagup \end{array} \rrbracket, \quad \llbracket \begin{array}{c} \text{loop} \\ \diagup \diagdown \end{array} \rrbracket = q^{1-2n} \llbracket \begin{array}{c} \text{arc} \\ \diagup \diagdown \end{array} \rrbracket$ .

The diagrams in both sides of the second or fourth equations represent parts of larger link diagrams that are identical except near a point where they look as indicated. For more details about this polynomial (and its two-variable extension, namely the Dubrovnik version of the two-variable Kauffman polynomial) we refer the reader to [Kauffman 1990; 2001].

---

*MSC2010:* primary 57M27; secondary 57M27, 57M15.

*Keywords:* graphs, invariants for knots and links, Kauffman polynomial.

Kauffman and Vogel [1992] extended the two-variable Dubrovnik polynomial to a three-variable rational function for knotted 4-valent graphs (4-valent graphs embedded in  $\mathbb{R}^3$ ) with rigid vertices. For the case of the  $SO(2n)$  Kauffman polynomial, this extension is obtained by defining

$$\begin{aligned} \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] &:= \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] - q \left[ \begin{array}{c} \diagup \diagdown \\ \diagup \diagdown \end{array} \right] - q^{-1} \left[ \begin{array}{c} \diagdown \diagup \\ \diagdown \diagup \end{array} \right] \\ &= \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] - q \left[ \begin{array}{c} \diagdown \diagup \\ \diagdown \diagup \end{array} \right] - q^{-1} \left[ \begin{array}{c} \diagup \diagdown \\ \diagup \diagdown \end{array} \right]. \end{aligned}$$

That is, the invariant for knotted 4-valent graphs with rigid vertices is defined in terms of the  $SO(2n)$  Kauffman polynomial. In [Kauffman and Vogel 1992], it was also shown that the resulting polynomial of a knotted 4-valent graph satisfies certain graphical relations, which determine values for each unoriented planar 4-valent graph by recursive formulas defined entirely in the category of planar graphs.

The results in [Kauffman and Vogel 1992] imply that there is a state model for the Kauffman polynomial of an unoriented link via planar 4-valent graphs. This model can also be deduced from Carpentier's work [2000] on the Kauffman–Vogel polynomial by changing one's perspective (the focus of Carpentier's paper is on invariants for graphs rather than on the Kauffman polynomial for links). A somewhat similar approach was used in [Caprau and Tipton 2011] to construct a rational function in three variables which is an invariant of regular isotopy of unoriented links, and provides a state summation model for the Dubrovnik version of the two-variable Kauffman polynomial. The corresponding state model makes use of a special type of planar trivalent graphs.

François Jaeger found a relationship between the two-variable Kauffman polynomial and the regular isotopy version of the HOMFLY-PT polynomial. He showed that the Kauffman polynomial of an unoriented link  $L$  can be obtained as a weighted sum of HOMFLY-PT polynomials of oriented links associated with  $L$ . For a brief description of Jaeger's construction we refer the reader to [Kauffman 2001]. Murakami, Ohtsuki and Yamada [1998] (MOY) used planar trivalent graphs to construct in a beautiful graphical calculus for the  $sl(n)$ -link polynomial (a one-variable specialization of the HOMFLY-PT polynomial).

The motivation for this paper has its source in the following, natural, questions: Is there a way to apply the MOY model to Jaeger's formula and derive a state summation model for the  $SO(2n)$  Kauffman polynomial? And if so, how is the resulting state model for the  $SO(2n)$  Kauffman polynomial related to the one implicitly given in [Kauffman and Vogel 1992]?

We slightly alter the MOY model for the  $sl(n)$ -link polynomial by working with (planar, cross-like oriented) 4-valent graphs instead of trivalent graphs. Implementing the MOY model into Jaeger's construction, we show that in order to construct a



state model for the Kauffman polynomial it is not sufficient to allow only cross-like oriented 4-valent graphs but also alternating oriented vertices. The skein formalism that we obtain is as follows:

$$\begin{aligned} \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] &= q \left[ \begin{array}{c} \frown \\ \smile \end{array} \right] + q^{-1} \left[ \begin{array}{c} \smile \\ \frown \end{array} \right] - \left[ \begin{array}{c} \times \\ \times \end{array} \right], \\ \left[ \bigcirc \right] &= [2n - 1] + 1, \\ \left[ \begin{array}{c} \text{loop} \end{array} \right] &= ([2n - 2] + [2]) \left[ \begin{array}{c} \text{arc} \end{array} \right], \\ \left[ \begin{array}{c} \text{two circles} \end{array} \right] &= ([2n - 3] + 1) \left[ \begin{array}{c} \text{two arcs} \end{array} \right] + [2] \left[ \begin{array}{c} \times \\ \times \end{array} \right], \\ \left[ \begin{array}{c} \text{crossing} \end{array} \right] + \left[ \begin{array}{c} \text{crossing with arc} \end{array} \right] - \left[ \begin{array}{c} \text{crossing with arc} \end{array} \right] - \left[ \begin{array}{c} \text{crossing with arc} \end{array} \right] - [2n - 4] \left[ \begin{array}{c} \text{crossing with arc} \end{array} \right] = \\ &= \left[ \begin{array}{c} \text{crossing} \end{array} \right] + \left[ \begin{array}{c} \text{crossing with arc} \end{array} \right] - \left[ \begin{array}{c} \text{crossing with arc} \end{array} \right] - \left[ \begin{array}{c} \text{crossing with arc} \end{array} \right] - [2n - 4] \left[ \begin{array}{c} \text{crossing with arc} \end{array} \right], \end{aligned}$$

where

$$[n] = \frac{q^n - q^{-n}}{q - q^{-1}},$$

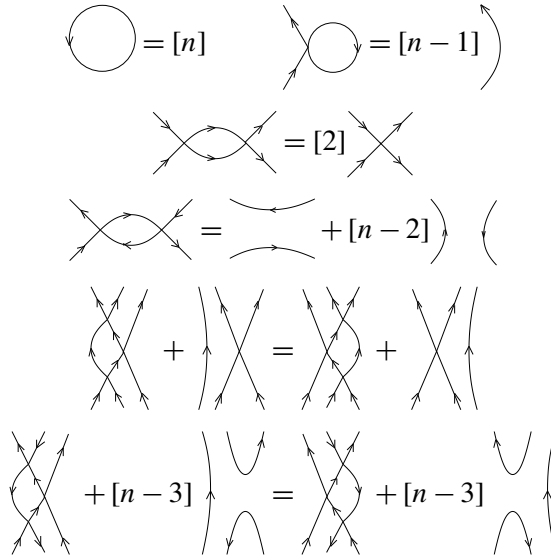
and  $n \in \mathbb{Z}$  with  $n \geq 2$ .

Comparing the graph skein relations above with the graphical relations derived by Kauffman and Vogel in [1992], it is not hard to see that the state model for the  $SO(2n)$  Kauffman polynomial that we arrive at is essentially the same as that implied by the work in [Kauffman and Vogel 1992] (up to a negative sign for the weight received by the “flat resolution” of a crossing), and that given in [Caprau and Tipton 2011, Subsection 5.1] (up to a change of variables). We would like to point out that Hao Wu [2012] used a different approach to write the Kauffman–Vogel graph polynomial as a state sum of the MOY graph polynomial.

The paper is organized as follows: In Section 2 we provide a version of the MOY state model for the  $sl(n)$ -link polynomial, and in Section 3 we review Jaeger’s formula for the Kauffman polynomial. The heart of the paper is Section 4, in which we derive the state model for the  $SO(2n)$  Kauffman polynomial.

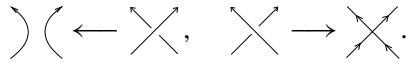
## 2. The MOY state model for the $sl(n)$ polynomial

In this section, we give the [Murakami et al. 1998] state model for the regular isotopy version of the  $sl(n)$  polynomial of an oriented link  $L$ . The  $sl(n)$  polynomial is a one-variable specialization of the well-known HOMFLY-PT polynomial (see [Freyd et al. 1985; Przytycki and Traczyk 1988]). Let  $D$  be a generic diagram of



**Figure 1.** Web skein relations.

$L$  containing  $c$  crossings. We resolve each crossing of  $D$  in the two ways shown below:



This process yields  $2^c$  resolutions (states) corresponding to the link diagram  $D$ . A resolution  $\Gamma$  of  $D$  is a 4-valent oriented planar graph in  $\mathbb{R}^2$ , possibly with loops with no vertices, such that each vertex is crossing-type oriented:  $\times$ . There is a well-defined Laurent polynomial  $R(\Gamma) \in \mathbb{Z}[q, q^{-1}]$  associated to a resolution  $\Gamma$ , such that it satisfies the skein relations depicted in Figure 1, where

$$[n] = \frac{q^n - q^{-n}}{q - q^{-1}} \quad \text{and} \quad n \in \mathbb{Z}, \quad \text{with } n \geq 2$$

(the symbol  $R$  is omitted in the graph skein relations to avoid clutter). We will refer to  $R(\Gamma)$  as the *MOY graph polynomial* (see [Murakami et al. 1998]).

Decompose each crossing in  $D$  as explained in Figure 2, and form the following linear combination of the MOY evaluations of all  $2^c$  resolutions  $\Gamma$  of  $D$ :

$$R(D) = \sum_{\Gamma} a_{\Gamma} R(\Gamma),$$

where the coefficients  $a_{\Gamma} \in \mathbb{Z}[q, q^{-1}]$  are given by the rules depicted in Figure 2.

It is an enjoyable exercise to verify that  $R(D_1) = R(D_2)$ , whenever diagrams  $D_1$  and  $D_2$  differ by a Reidemeister II or III move. Excluding rightmost terms from

$$\begin{aligned}
 R\left(\begin{array}{c} \nearrow \searrow \\ \nwarrow \nearrow \end{array}\right) &= q R\left(\begin{array}{c} \nearrow \\ \searrow \end{array}\right) \left(\begin{array}{c} \nwarrow \\ \nearrow \end{array}\right) - R\left(\begin{array}{c} \nwarrow \nearrow \\ \nearrow \nwarrow \end{array}\right) \\
 R\left(\begin{array}{c} \nwarrow \nearrow \\ \nearrow \nwarrow \end{array}\right) &= q^{-1} R\left(\begin{array}{c} \nwarrow \\ \nearrow \end{array}\right) \left(\begin{array}{c} \nearrow \\ \nwarrow \end{array}\right) - R\left(\begin{array}{c} \nearrow \nwarrow \\ \nwarrow \nearrow \end{array}\right)
 \end{aligned}$$

**Figure 2.** Decomposition of crossings.

the decomposition rules of crossings, we obtain Conway’s skein relation:

$$R\left(\begin{array}{c} \nearrow \searrow \\ \nwarrow \nearrow \end{array}\right) - R\left(\begin{array}{c} \nwarrow \nearrow \\ \nearrow \nwarrow \end{array}\right) = (q - q^{-1}) R\left(\begin{array}{c} \nearrow \\ \searrow \end{array}\right) \left(\begin{array}{c} \nwarrow \\ \nearrow \end{array}\right).$$

We note that  $R(L) := R(D)$  is the regular isotopy version of the  $sl(n)$  polynomial of the link  $L$ , and that it satisfies the following:

$$R\left(\begin{array}{c} \nearrow \\ \searrow \end{array}\right) = q^n R\left(\begin{array}{c} \nearrow \\ \searrow \end{array}\right) \quad \text{and} \quad R\left(\begin{array}{c} \nwarrow \\ \nearrow \end{array}\right) = q^{-n} R\left(\begin{array}{c} \nwarrow \\ \nearrow \end{array}\right).$$

### 3. Jaeger’s model for the Kauffman polynomial

In the late 80s, François Jaeger found a relationship between the two-variable Kauffman polynomial and the regular isotopy version of the HOMFLY-PT polynomial. He showed that the Kauffman polynomial of an unoriented link  $L$  can be obtained as a weighted sum of HOMFLY-PT polynomials of oriented links associated with  $L$ . Since this construction is only briefly described in [Kauffman 2001], we provide here a thorough exposition of it, which is necessary in order to understand our main Section 4. Moreover, we describe Jaeger’s model for the  $SO(2n)$  Kauffman polynomial by considering the  $sl(n)$ -link invariant instead of the HOMFLY-PT polynomial.

Given an unoriented link diagram  $L$ , splice some of the crossings of  $L$  and orient the resulting link. This results in a *state* for the expansion  $[[L]]$ . Each state receives a certain weight, according to the following skein relation:

$$\begin{aligned}
 & \left[ \begin{array}{c} \nearrow \searrow \\ \nwarrow \nearrow \end{array} \right] \\
 &= (q - q^{-1}) \left( \left[ \begin{array}{c} \nearrow \searrow \\ \nwarrow \nearrow \end{array} \right] - \left[ \begin{array}{c} \nearrow \\ \searrow \end{array} \right] \left[ \begin{array}{c} \nwarrow \\ \nearrow \end{array} \right] \right) + \left[ \begin{array}{c} \nwarrow \nearrow \\ \nearrow \nwarrow \end{array} \right] + \left[ \begin{array}{c} \nwarrow \nearrow \\ \nwarrow \nearrow \end{array} \right] + \left[ \begin{array}{c} \nwarrow \nearrow \\ \nearrow \nwarrow \end{array} \right] + \left[ \begin{array}{c} \nwarrow \nearrow \\ \nwarrow \nearrow \end{array} \right]. \quad (*)
 \end{aligned}$$

It is important to remark that the formula (\*) requires states that are oriented in a globally compatible way as oriented link diagrams. Moreover, observe that the orientation and the weight of a state are determined by how the crossings are spliced. When approaching a crossing by traveling along the understrand, a splicing

is obtained by either turning right or left at that crossing. In both cases, the strands of the splicing are oriented according to the direction of the traveling. If the crossing is spliced by turning right, then it receives the weight  $q - q^{-1}$ , and if it is spliced by turning left, it receives the weight  $-(q - q^{-1})$ . If a crossing is left unspliced, its weight (in the total weight of the state) is equal to 1.

The *weight*  $b_\sigma$  of a state  $\sigma$  is obtained by taking the product of the weights  $\pm(q - q^{-1})$  or 1 according to the skein relation (\*). Define the *evaluation* of a state  $\sigma$  by the formula

$$[\sigma] = (q^{1-n})^{\text{rot}(\sigma)} R(\sigma),$$

where  $\text{rot}(\sigma)$  is the *rotation number* of the oriented link diagram  $\sigma$ , and  $R(\sigma)$  is the regular isotopy version of the  $sl(n)$  polynomial of  $\sigma$ .

The rotation number (also called the Whitney degree) of an oriented link diagram is obtained by splicing every crossing according to its orientation, and then adding the rotation numbers of all of the resulting Seifert circles, where a counterclockwise oriented circle contributes a  $+1$ , and a clockwise oriented circle contributes a  $-1$ . It is well-known that the rotation number is a regular isotopy invariant for oriented links.

Equipped with the above definitions and conventions, we are ready to state Jaeger’s theorem.

**Theorem 1** (Jaeger). The Kauffman polynomial  $\llbracket L \rrbracket$  of an unoriented link diagram  $L$  can be obtained as follows:

$$\llbracket L \rrbracket = \sum_{\sigma} b_{\sigma} [\sigma],$$

where the sum is over all states  $\sigma$  associated with  $L$  that have globally compatible orientations.

*Proof.* First note that the Conway identity holds for  $[\cdot]$ :

$$\begin{aligned} \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] - \left[ \begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array} \right] &= q^{(1-n)\text{rot}(\times)} R\left(\begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array}\right) - q^{(1-n)\text{rot}(\times)} R\left(\begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array}\right) \\ &= q^{(1-n)\text{rot}(\smile)} \left( R\left(\begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array}\right) - R\left(\begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array}\right) \right) \\ &= (q - q^{-1}) q^{(1-n)\text{rot}(\smile)} R\left(\begin{array}{c} \smile \\ \smile \end{array}\right) \\ &= (q - q^{-1}) \left[ \begin{array}{c} \smile \\ \smile \end{array} \right]. \end{aligned}$$

Then,

$$\begin{aligned} & \left[ \begin{array}{c} \diagdown \\ \diagup \end{array} \right] - \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] \\ &= (q - q^{-1}) \left( \left[ \begin{array}{c} \smile \\ \smile \end{array} \right] - \left[ \begin{array}{c} \smile \\ \smile \end{array} \right] \right) + \left[ \begin{array}{c} \diagdown \\ \diagdown \end{array} \right] + \left[ \begin{array}{c} \diagup \\ \diagup \end{array} \right] + \left[ \begin{array}{c} \diagdown \\ \diagup \end{array} \right] + \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] \\ & \quad - (q - q^{-1}) \left( \left[ \begin{array}{c} \smile \\ \smile \end{array} \right] - \left[ \begin{array}{c} \smile \\ \smile \end{array} \right] \right) - \left[ \begin{array}{c} \diagup \\ \diagup \end{array} \right] - \left[ \begin{array}{c} \diagdown \\ \diagdown \end{array} \right] - \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] - \left[ \begin{array}{c} \diagdown \\ \diagup \end{array} \right], \end{aligned}$$

and by the Conway identity, we obtain

$$\begin{aligned} \left[ \begin{array}{c} \diagdown \\ \diagup \end{array} \right] - \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] &= (q - q^{-1}) \left( \left[ \begin{array}{c} \smile \\ \smile \end{array} \right] + \left[ \begin{array}{c} \smile \\ \smile \end{array} \right] + \left[ \begin{array}{c} \smile \\ \smile \end{array} \right] + \left[ \begin{array}{c} \smile \\ \smile \end{array} \right] \right) \\ & \quad - (q - q^{-1}) \left( \left[ \begin{array}{c} \smile \\ \smile \end{array} \right] + \left[ \begin{array}{c} \smile \\ \smile \end{array} \right] + \left[ \begin{array}{c} \smile \\ \smile \end{array} \right] + \left[ \begin{array}{c} \smile \\ \smile \end{array} \right] \right) \\ &= (q - q^{-1}) \left( \left[ \begin{array}{c} \smile \\ \smile \end{array} \right] - \left[ \begin{array}{c} \smile \\ \smile \end{array} \right] \right). \end{aligned}$$

Observe that

$$\begin{aligned} \left[ \begin{array}{c} \circ \\ \downarrow \end{array} \right] &= q^{(1-n)\text{rot}(\circ)} R \left( \begin{array}{c} \circ \\ \downarrow \end{array} \right) = q^{1-n} [n], \\ \left[ \begin{array}{c} \circ \\ \uparrow \end{array} \right] &= q^{(1-n)\text{rot}(\circ)} R \left( \begin{array}{c} \circ \\ \uparrow \end{array} \right) = q^{n-1} [n], \end{aligned}$$

and, therefore we have

$$\left[ \begin{array}{c} \circ \\ \circ \end{array} \right] = \left[ \begin{array}{c} \circ \\ \downarrow \end{array} \right] + \left[ \begin{array}{c} \circ \\ \uparrow \end{array} \right] = (q^{1-n} + q^{n-1}) [n] = \frac{q^{2n-1} - q^{1-2n}}{q - q^{-1}} + 1.$$

Moreover,

$$\begin{aligned} \left[ \begin{array}{c} \circlearrowright \\ \downarrow \end{array} \right] &= q^{(1-n)\text{rot}(\circlearrowright)} R \left( \begin{array}{c} \circlearrowright \\ \downarrow \end{array} \right) = q^{(1-n)(1+\text{rot}(\curvearrowright))} q^n R(\curvearrowright) \\ &= q [\curvearrowright], \end{aligned}$$

$$\begin{aligned} \left[ \begin{array}{c} \circlearrowleft \\ \downarrow \end{array} \right] &= q^{(1-n)\text{rot}(\circlearrowleft)} R \left( \begin{array}{c} \circlearrowleft \\ \downarrow \end{array} \right) = q^{(1-n)(-1+\text{rot}(\curvearrowleft))} q^n R(\curvearrowleft) \\ &= q^{2n-1} [\curvearrowleft]. \end{aligned}$$

Therefore,

$$\begin{aligned}
 \llbracket \text{link} \rrbracket &= (q - q^{-1}) (\llbracket \text{link} \rrbracket - \llbracket \text{link} \rrbracket) + \llbracket \text{link} \rrbracket + \llbracket \text{link} \rrbracket \\
 &= (q - q^{-1}) (q^{n-1} [n] \llbracket \text{link} \rrbracket - \llbracket \text{link} \rrbracket) \\
 &\quad + q \llbracket \text{link} \rrbracket + q^{2n-1} \llbracket \text{link} \rrbracket \\
 &= q^{2n-1} \llbracket \text{link} \rrbracket + q^{2n-1} \llbracket \text{link} \rrbracket = q^{2n-1} \llbracket \text{link} \rrbracket.
 \end{aligned}$$

Similarly, one can show that

$$\llbracket \text{link} \rrbracket = q^{1-2n} \llbracket \text{link} \rrbracket.$$

It remains to show that  $\llbracket \cdot \rrbracket$  is a regular isotopy invariant for unoriented links.

$$\begin{aligned}
 \llbracket \text{link} \rrbracket &= (q - q^{-1}) (\llbracket \text{link} \rrbracket - \llbracket \text{link} \rrbracket + \llbracket \text{link} \rrbracket - \llbracket \text{link} \rrbracket) \\
 &\quad + \llbracket \text{link} \rrbracket + \llbracket \text{link} \rrbracket + \llbracket \text{link} \rrbracket + \llbracket \text{link} \rrbracket \\
 &= (q - q^{-1}) (\llbracket \text{link} \rrbracket - q \llbracket \text{link} \rrbracket + q^{-1} \llbracket \text{link} \rrbracket - \llbracket \text{link} \rrbracket) \\
 &\quad + \llbracket \text{link} \rrbracket + \llbracket \text{link} \rrbracket + \llbracket \text{link} \rrbracket + \llbracket \text{link} \rrbracket \\
 &= (q - q^{-1}) (\llbracket \text{link} \rrbracket - \llbracket \text{link} \rrbracket - q \llbracket \text{link} \rrbracket + q^{-1} \llbracket \text{link} \rrbracket) + \llbracket \text{link} \rrbracket \\
 &= \llbracket \text{link} \rrbracket,
 \end{aligned}$$

by the Conway identity for  $\llbracket \cdot \rrbracket$ .

The invariance of  $\llbracket \cdot \rrbracket$  under the Reidemeister III move is verified in a similar fashion, and we leave the details to the reader. □

### 4. The $SO(2n)$ Kauffman polynomial via planar 4-valent graphs

We seek to construct a state summation model for the  $SO(2n)$  Kauffman polynomial, that works in much the same way as the MOY model works for the  $sl(n)$  polynomial. Moreover, we want to derive such a state model by implementing the MOY construction into Jaeger’s theorem. Therefore, the states corresponding to an unoriented link diagram  $L$  will be unoriented 4-valent graphs obtained by resolving a crossing of  $L$  in one of the following ways:



and we want to find some  $A, B, C \in \mathbb{Z}[q, q^{-1}]$ , such that

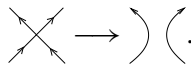
$$\left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] = A \left[ \begin{array}{c} \smile \\ \frown \end{array} \right] + B \left[ \begin{array}{c} \smile \\ \smile \end{array} \right] \left[ \begin{array}{c} \frown \\ \frown \end{array} \right] + C \left[ \begin{array}{c} \diagup \diagdown \\ \diagup \diagdown \end{array} \right]. \tag{4-1}$$

The state model that we wish to construct requires a consistent method to evaluate closed, unoriented 4-valent graphs (the states associated with  $L$ ).

To this end, we note that implementing the MOY state summation into Jaeger’s model requires the bracket evaluation  $[\Gamma]$ , where  $\Gamma$  is an oriented 4-valent planar graph whose vertices are crossing-type oriented. We define

$$[\Gamma] := (q^{1-n})^{\text{rot}(\Gamma)} R(\Gamma), \tag{4-2}$$

where  $\text{rot}(\Gamma)$ , the *rotation number* of such a graph  $\Gamma$ , is the sum of the rotation numbers of the disjoint oriented circles obtained by splicing each vertex of  $\Gamma$  according to the orientation of its edges:



We will regard the Equation (4-2) as a skein relation, as explained below:

$$\left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] = (q^{1-n})^{\text{rot}(\times)} R\left( \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right) = (q^{1-n})^{\text{rot}(\circ \cup \circ)} R\left( \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right). \tag{4-3}$$

Jaeger’s theorem implies that

$$\left[ \begin{array}{c} \smile \\ \frown \end{array} \right] = \left[ \begin{array}{c} \smile \\ \frown \end{array} \right] + \left[ \begin{array}{c} \smile \\ \smile \end{array} \right] + \left[ \begin{array}{c} \smile \\ \frown \end{array} \right] + \left[ \begin{array}{c} \smile \\ \smile \end{array} \right],$$

and to have a consistent construction, the evaluation



will contain the bracket evaluations

$$\left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right],$$

for all such orientations of the vertex.

To determine what the coefficients  $A$ ,  $B$ , and  $C$  must be, we compute

$$\left[ \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] \right]$$

via Jaeger’s model, and throughout the process, we evaluate the resulting oriented link diagrams using the MOY construction for the  $sl(n)$  polynomial,  $R$ .

$$\begin{aligned} \left[ \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] \right] &= (q - q^{-1}) \left( \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] - \left[ \begin{array}{c} \curvearrowright \\ \curvearrowleft \end{array} \right] \right) + \left[ \begin{array}{c} \diagup \diagdown \\ \diagup \diagdown \end{array} \right] + \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] + \left[ \begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array} \right] + \left[ \begin{array}{c} \diagdown \diagup \\ \diagdown \diagup \end{array} \right] \\ &= (q - q^{-1}) \left( \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] - \left[ \begin{array}{c} \curvearrowright \\ \curvearrowleft \end{array} \right] \right) \\ &\quad + (q^{1-n})^{\text{rot}(\curvearrowright)} R \left( \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right) + (q^{1-n})^{\text{rot}(\curvearrowleft)} R \left( \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right) \\ &\quad + (q^{1-n})^{\text{rot}(\curvearrowright)} R \left( \begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array} \right) + (q^{1-n})^{\text{rot}(\curvearrowleft)} R \left( \begin{array}{c} \diagdown \diagup \\ \diagdown \diagup \end{array} \right). \end{aligned}$$

Employing the skein relations in Figure 2, we have

$$\begin{aligned} \left[ \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] \right] &= (q - q^{-1}) \left( \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] - \left[ \begin{array}{c} \curvearrowright \\ \curvearrowleft \end{array} \right] \right) \\ &\quad + (q^{1-n})^{\text{rot}(\curvearrowright)} \left( q R \left( \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right) - R \left( \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right) \right) \\ &\quad + (q^{1-n})^{\text{rot}(\curvearrowleft)} \left( q^{-1} R \left( \begin{array}{c} \curvearrowright \\ \curvearrowleft \end{array} \right) - R \left( \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right) \right) \\ &\quad + (q^{1-n})^{\text{rot}(\curvearrowright)} \left( q R \left( \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right) - R \left( \begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array} \right) \right) \\ &\quad + (q^{1-n})^{\text{rot}(\curvearrowleft)} \left( q^{-1} R \left( \begin{array}{c} \curvearrowright \\ \curvearrowleft \end{array} \right) - R \left( \begin{array}{c} \diagdown \diagup \\ \diagdown \diagup \end{array} \right) \right). \end{aligned}$$

Making use of the skein relation (4-3), we obtain

$$\begin{aligned} \left[ \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] \right] &= (q - q^{-1}) \left( \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] - \left[ \begin{array}{c} \curvearrowright \\ \curvearrowleft \end{array} \right] \right) + q \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] - \left[ \begin{array}{c} \diagup \diagdown \\ \diagup \diagdown \end{array} \right] + q^{-1} \left[ \begin{array}{c} \curvearrowright \\ \curvearrowleft \end{array} \right] - \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] \\ &\quad + q \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] - \left[ \begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array} \right] + q^{-1} \left[ \begin{array}{c} \curvearrowright \\ \curvearrowleft \end{array} \right] - \left[ \begin{array}{c} \diagdown \diagup \\ \diagdown \diagup \end{array} \right] \\ &= q \left( \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] + \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] + \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] + \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] \right) \\ &\quad + q^{-1} \left( \left[ \begin{array}{c} \curvearrowright \\ \curvearrowleft \end{array} \right] + \left[ \begin{array}{c} \curvearrowright \\ \curvearrowleft \end{array} \right] + \left[ \begin{array}{c} \curvearrowright \\ \curvearrowleft \end{array} \right] + \left[ \begin{array}{c} \curvearrowright \\ \curvearrowleft \end{array} \right] \right) \end{aligned}$$



$$-q \left[ \begin{array}{c} \curvearrowright \\ \curvearrowleft \end{array} \right] - q \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] \left[ \begin{array}{c} \curvearrowleft \\ \curvearrowleft \end{array} \right] - q^{-1} \left[ \begin{array}{c} \curvearrowright \\ \curvearrowleft \end{array} \right] \left[ \begin{array}{c} \curvearrowright \\ \curvearrowleft \end{array} \right] - q^{-1} \left[ \begin{array}{c} \curvearrowleft \\ \curvearrowleft \end{array} \right] \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] \\ - \left( \left[ \begin{array}{c} \times \\ \times \end{array} \right] + \left[ \begin{array}{c} \times \\ \times \end{array} \right] + \left[ \begin{array}{c} \times \\ \times \end{array} \right] + \left[ \begin{array}{c} \times \\ \times \end{array} \right] \right).$$

Therefore, we have

$$\left[ \begin{array}{c} \times \\ \times \end{array} \right] = q \left[ \begin{array}{c} \curvearrowright \\ \curvearrowleft \end{array} \right] + q^{-1} \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] \left[ \begin{array}{c} \curvearrowleft \\ \curvearrowleft \end{array} \right] - q \left[ \begin{array}{c} \curvearrowright \\ \curvearrowleft \end{array} \right] - q \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] \left[ \begin{array}{c} \curvearrowleft \\ \curvearrowleft \end{array} \right] \\ - q^{-1} \left[ \begin{array}{c} \curvearrowright \\ \curvearrowleft \end{array} \right] \left[ \begin{array}{c} \curvearrowright \\ \curvearrowleft \end{array} \right] - q^{-1} \left[ \begin{array}{c} \curvearrowleft \\ \curvearrowleft \end{array} \right] \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] - \left( \left[ \begin{array}{c} \times \\ \times \end{array} \right] + \left[ \begin{array}{c} \times \\ \times \end{array} \right] + \left[ \begin{array}{c} \times \\ \times \end{array} \right] + \left[ \begin{array}{c} \times \\ \times \end{array} \right] \right).$$

Comparing the last equality with (4-1), we see that in order to work with a certain evaluation

$$\left[ \begin{array}{c} \times \\ \times \end{array} \right]$$

for an unoriented vertex, we must also take in consideration *alternating orientations* for edges meeting at a vertex, and define the *bracket* of an *alternating oriented vertex* as follows:

$$\left[ \begin{array}{c} \times \\ \times \end{array} \right] := q \left[ \begin{array}{c} \curvearrowright \\ \curvearrowleft \end{array} \right] \left[ \begin{array}{c} \curvearrowleft \\ \curvearrowleft \end{array} \right] + q^{-1} \left[ \begin{array}{c} \curvearrowleft \\ \curvearrowleft \end{array} \right] \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right]. \tag{4-4}$$

The above computations also imply the need of the following definition:

$$\left[ \begin{array}{c} \times \\ \times \end{array} \right] := \left[ \begin{array}{c} \times \\ \times \end{array} \right] + \left[ \begin{array}{c} \times \\ \times \end{array} \right] + \left[ \begin{array}{c} \times \\ \times \end{array} \right] + \left[ \begin{array}{c} \times \\ \times \end{array} \right] + \left[ \begin{array}{c} \times \\ \times \end{array} \right] + \left[ \begin{array}{c} \times \\ \times \end{array} \right].$$

Implementing the above definitions into our previous computations, we obtain

$$\left[ \begin{array}{c} \times \\ \times \end{array} \right] = q \left[ \begin{array}{c} \curvearrowright \\ \curvearrowleft \end{array} \right] + q^{-1} \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] \left[ \begin{array}{c} \curvearrowleft \\ \curvearrowleft \end{array} \right] - \left[ \begin{array}{c} \times \\ \times \end{array} \right]. \tag{4-5}$$

Therefore,  $A = q$ ,  $B = q^{-1}$ , and  $C = -1$ .

We have seen that the implementation of the MOY state model into Jaeger’s state summation requires *balanced* oriented 4-valent graphs (in the sense that the total degree of a vertex is zero), with vertices being either crossing-type oriented or alternating oriented.

**Proposition 1.** *The following identity holds:*

$$\left[ \begin{array}{c} \bigcirc \end{array} \right] = [2n - 1] + 1.$$

*Proof.* This identity holds by Jaeger’s theorem. □

**Proposition 2.** *The following graph skein relation holds:*

$$\left[ \begin{array}{c} \text{loop} \\ \downarrow \end{array} \right] = ([2n - 2] + [2]) \left[ \begin{array}{c} \text{arc} \\ \downarrow \end{array} \right].$$

*Proof.*

$$\left[ \begin{array}{c} \text{loop} \\ \downarrow \end{array} \right] = \left[ \begin{array}{c} \text{loop} \\ \downarrow \text{right} \end{array} \right] + \left[ \begin{array}{c} \text{loop} \\ \downarrow \text{left} \end{array} \right] + \left[ \begin{array}{c} \text{loop} \\ \downarrow \text{right} \end{array} \right] + \left[ \begin{array}{c} \text{loop} \\ \downarrow \text{left} \end{array} \right].$$

Now, for the first oriented diagram, we have

$$\begin{aligned} \left[ \begin{array}{c} \text{loop} \\ \downarrow \end{array} \right] &= q^{(1-n)\text{rot}(\text{loop})} R \left( \begin{array}{c} \text{loop} \\ \downarrow \end{array} \right) = q^{(1-n)\text{rot}(\text{loop})} [n - 1] R \left( \begin{array}{c} \text{arc} \\ \downarrow \end{array} \right) \\ &= q^{1-n} [n - 1] q^{(1-n)\text{rot}(\text{arc})} R \left( \begin{array}{c} \text{arc} \\ \downarrow \end{array} \right) = q^{1-n} [n - 1] \left[ \begin{array}{c} \text{arc} \\ \downarrow \end{array} \right], \end{aligned}$$

and for the third oriented diagram, we have

$$\begin{aligned} \left[ \begin{array}{c} \text{loop} \\ \downarrow \text{right} \end{array} \right] &= q \left[ \begin{array}{c} \text{loop} \\ \downarrow \end{array} \right] + q^{-1} \left[ \begin{array}{c} \text{arc} \\ \downarrow \end{array} \right] \\ &= q \cdot q^{(1-n)\text{rot}(\text{loop})} R \left( \begin{array}{c} \text{loop} \\ \downarrow \end{array} \right) + q^{-1} \left[ \begin{array}{c} \text{arc} \\ \downarrow \end{array} \right] \\ &= q \cdot q^{1-n} \cdot q^{(1-n)\text{rot}(\text{arc})} [n] R \left( \begin{array}{c} \text{arc} \\ \downarrow \end{array} \right) + q^{-1} \left[ \begin{array}{c} \text{arc} \\ \downarrow \end{array} \right] \\ &= q^{2-n} [n] \left[ \begin{array}{c} \text{arc} \\ \downarrow \end{array} \right] + q^{-1} \left[ \begin{array}{c} \text{arc} \\ \downarrow \end{array} \right] = (q^{2-n} [n] + q^{-1}) \left[ \begin{array}{c} \text{arc} \\ \downarrow \end{array} \right]. \end{aligned}$$

Similarly, we obtain

$$\left[ \begin{array}{c} \text{loop} \\ \downarrow \text{left} \end{array} \right] = q^{n-1} [n - 1] \left[ \begin{array}{c} \text{arc} \\ \downarrow \end{array} \right] \quad \text{and} \quad \left[ \begin{array}{c} \text{loop} \\ \downarrow \text{right} \end{array} \right] = (q^{n-2} [n] + q) \left[ \begin{array}{c} \text{arc} \\ \downarrow \end{array} \right].$$

Using these evaluations for each of the oriented states, we arrive at

$$\left[ \begin{array}{c} \text{loop} \\ \downarrow \end{array} \right] = ([2n - 2] + [2]) \left( \left[ \begin{array}{c} \text{arc} \\ \downarrow \end{array} \right] + \left[ \begin{array}{c} \text{arc} \\ \downarrow \end{array} \right] \right) = ([2n - 2] + [2]) \left[ \begin{array}{c} \text{arc} \\ \downarrow \end{array} \right]. \quad \square$$

**Proposition 3.** *The following skein relation holds:*

$$\left[ \begin{array}{c} \text{crossing} \\ \downarrow \end{array} \right] = ([2n - 3] + 1) \left[ \begin{array}{c} \text{cup} \\ \downarrow \end{array} \right] + [2] \left[ \begin{array}{c} \text{crossing} \\ \downarrow \end{array} \right].$$

*Proof.* We know that

$$\begin{aligned} \left[ \begin{array}{c} \text{crossing} \\ \downarrow \end{array} \right] &= \left[ \begin{array}{c} \text{crossing} \\ \downarrow \text{right} \end{array} \right] + \left[ \begin{array}{c} \text{crossing} \\ \downarrow \text{left} \end{array} \right] + \left[ \begin{array}{c} \text{crossing} \\ \downarrow \text{right} \end{array} \right] + \left[ \begin{array}{c} \text{crossing} \\ \downarrow \text{left} \end{array} \right] + \left[ \begin{array}{c} \text{crossing} \\ \downarrow \text{right} \end{array} \right] \\ &\quad + \left[ \begin{array}{c} \text{crossing} \\ \downarrow \text{left} \end{array} \right] + \left[ \begin{array}{c} \text{crossing} \\ \downarrow \text{right} \end{array} \right] + \left[ \begin{array}{c} \text{crossing} \\ \downarrow \text{left} \end{array} \right] + \left[ \begin{array}{c} \text{crossing} \\ \downarrow \text{right} \end{array} \right] + \left[ \begin{array}{c} \text{crossing} \\ \downarrow \text{left} \end{array} \right]. \end{aligned}$$

Now,

$$\left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] = (q^{1-n})^{\text{rot}(\curvearrowright)} R(\curvearrowright) = (q^{1-n})^{\text{rot}(\curvearrowright)} [2] R(\curvearrowright) = [2] \left[ \begin{array}{c} \times \\ \times \end{array} \right],$$

and

$$\begin{aligned} \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] &= (q^{1-n})^{\text{rot}(\curvearrowright)} R(\curvearrowright) = (q^{1-n})^{\text{rot}(\curvearrowright)} \left( R(\curvearrowright) + [n-2] R(\curvearrowright) \right) \\ &= \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] + q^{n-1} [n-2] \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right], \end{aligned}$$

where we used the fact that

$$\text{rot}(\curvearrowright) = \text{rot}(\curvearrowright) - 1.$$

We also have that

$$\begin{aligned} \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] &= q \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] + q^{-1} \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] = q \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] + q^{-1} (q^{1-n})^{\text{rot}(\curvearrowright)} R(\curvearrowright) \\ &= q \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] + q^{-1} q^{n-1} (q^{1-n})^{\text{rot}(\curvearrowright)} [n-1] R(\curvearrowright) \\ &= q \left[ \begin{array}{c} \times \\ \times \end{array} \right] + q^{n-2} [n-1] \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right]. \end{aligned}$$

Similarly, for the bigon with alternating oriented vertices, we have

$$\begin{aligned} \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] &= q^{-1} \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] + q \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] \\ &= q^{-1} \left( q \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] + q^{-1} \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] \right) + q \left( q \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] + q^{-1} \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] \right) \\ &= \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] + q^{-2} \cdot q^{n-1} [n] \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] + q^2 \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] + \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] \\ &= q^2 \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] + (q^{n-3} [n] + 2) \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right]. \end{aligned}$$

The remaining diagrams can be evaluated similarly. Thus, we have

$$\begin{aligned} \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] &= \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] + \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] + \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] + \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] + \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] \\ &\quad + \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] + \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] + \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] + \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right] + \left[ \begin{array}{c} \curvearrowright \\ \curvearrowright \end{array} \right], \end{aligned}$$

and using the above computations yields

$$\begin{aligned}
 \llbracket \text{Diagram 1} \rrbracket &= [2] \llbracket \text{Diagram 2} \rrbracket + \left( \llbracket \text{Diagram 3} \rrbracket + q^{n-1}[n-2] \llbracket \text{Diagram 4} \rrbracket \right) \\
 &+ [2] \llbracket \text{Diagram 5} \rrbracket + \left( \llbracket \text{Diagram 6} \rrbracket + q^{1-n}[n-2] \llbracket \text{Diagram 7} \rrbracket \right) \\
 &+ \left( q \llbracket \text{Diagram 8} \rrbracket + q^{n-2}[n-1] \llbracket \text{Diagram 9} \rrbracket \right) + \left( q \llbracket \text{Diagram 10} \rrbracket + q^{n-2}[n-1] \llbracket \text{Diagram 11} \rrbracket \right) \\
 &+ \left( q^{-1} \llbracket \text{Diagram 12} \rrbracket + q^{2-n}[n-1] \llbracket \text{Diagram 13} \rrbracket \right) + \left( q^{-1} \llbracket \text{Diagram 14} \rrbracket + q^{2-n}[n-1] \llbracket \text{Diagram 15} \rrbracket \right) \\
 &+ \left( q^{-2} \llbracket \text{Diagram 16} \rrbracket + (q^{3-n}[n]+2) \llbracket \text{Diagram 17} \rrbracket \right) + \left( q^2 \llbracket \text{Diagram 18} \rrbracket + (q^{n-3}[n]+2) \llbracket \text{Diagram 19} \rrbracket \right).
 \end{aligned}$$

Combining like terms, we have

$$\begin{aligned}
 \llbracket \text{Diagram 1} \rrbracket &= (q + q^{-1}) \llbracket \text{Diagram 2} \rrbracket + (q + q^{-1}) \llbracket \text{Diagram 5} \rrbracket + (q + q^{-1}) \llbracket \text{Diagram 8} \rrbracket + (q + q^{-1}) \llbracket \text{Diagram 10} \rrbracket \\
 &+ [2] \llbracket \text{Diagram 2} \rrbracket + [2] \llbracket \text{Diagram 5} \rrbracket + ([2n-3] + 1) \llbracket \text{Diagram 9} \rrbracket + ([2n-3] + 1) \llbracket \text{Diagram 11} \rrbracket \\
 &+ ([2n-3] + 1) \llbracket \text{Diagram 13} \rrbracket + ([2n-3] + 1) \llbracket \text{Diagram 15} \rrbracket \\
 &= [2] \llbracket \text{Diagram 2} \rrbracket + ([2n-3] + 1) \llbracket \text{Diagram 9} \rrbracket,
 \end{aligned}$$

which completes the proof. □

**Proposition 4.** *The following graph skein relation holds:*

$$\begin{aligned}
 \llbracket \text{Diagram 20} \rrbracket + \llbracket \text{Diagram 21} \rrbracket - \llbracket \text{Diagram 22} \rrbracket - \llbracket \text{Diagram 23} \rrbracket - [2n-4] \llbracket \text{Diagram 24} \rrbracket &= \\
 \llbracket \text{Diagram 25} \rrbracket + \llbracket \text{Diagram 26} \rrbracket - \llbracket \text{Diagram 27} \rrbracket - \llbracket \text{Diagram 28} \rrbracket - [2n-4] \llbracket \text{Diagram 29} \rrbracket.
 \end{aligned}$$

*Proof.* To prove the statement, one can use the same approach as in the previous propositions, namely evaluating

$$\llbracket \text{Diagram 20} \rrbracket \quad \text{and} \quad \llbracket \text{Diagram 21} \rrbracket$$

by summing over all bracket evaluations for all the associated oriented diagrams. To avoid cumbersome computations, we use instead the fact that  $\llbracket \cdot \rrbracket$  is invariant under the Reidemeister III move. That is,

$$\llbracket \text{Diagram 30} \rrbracket = \llbracket \text{Diagram 31} \rrbracket.$$

Using the skein relation (4-5), we have

$$\begin{aligned} \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] &= q \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] + q^{-1} \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] - \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right], \\ \left[ \begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array} \right] &= q \left[ \begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array} \right] + q^{-1} \left[ \begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array} \right] - \left[ \begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array} \right]. \end{aligned}$$

Since  $[\cdot]$  is invariant under the Reidemeister II move, we have

$$\left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] = \left[ \begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array} \right],$$

and we obtain that

$$\left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] = \left[ \begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array} \right].$$

Using again the skein relation (4-5), we have

$$\begin{aligned} 0 &= \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] - \left[ \begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array} \right] \\ &= q \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] + q^{-1} \left[ \begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array} \right] - \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] - \left( q \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] + q^{-1} \left[ \begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array} \right] - \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] \right) \\ &= q \left( q \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] + q^{-1} \left[ \begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array} \right] - \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] \right) + q^{-1} \left( q \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] + q^{-1} \left[ \begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array} \right] - \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] \right) \\ &\quad - \left( q \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] + q^{-1} \left[ \begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array} \right] - \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] \right) - q \left( q \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] + q^{-1} \left[ \begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array} \right] - \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] \right) \\ &\quad - q^{-1} \left( q \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] + q^{-1} \left[ \begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array} \right] - \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] \right) + q \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] + q^{-1} \left[ \begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array} \right] - \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right]. \end{aligned}$$

Applying Proposition 2 and canceling terms, we arrive at

$$\begin{aligned} 0 &= \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] - \left[ \begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array} \right] \\ &= q^2 \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] + ([2n - 2] + [2]) \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] - ([2n - 2] + [2]) \left[ \begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array} \right] \\ &\quad - q \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] + \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] + q^{-2} \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] - q^{-1} \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] + \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] - q^2 \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] \\ &\quad - \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] - q^{-2} \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] + q^{-1} \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] + q \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right] - \left[ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \right]. \end{aligned}$$

Now, from Proposition 3, we have

$$\begin{aligned}
 \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] &= [2] \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] + ([2n - 3] + 1) \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right], \\
 \left[ \begin{array}{c} \diagdown \\ \diagup \end{array} \right] &= [2] \left[ \begin{array}{c} \diagdown \\ \diagup \end{array} \right] + ([2n - 3] + 1) \left[ \begin{array}{c} \diagdown \\ \diagup \end{array} \right], \\
 \left[ \begin{array}{c} \diagup \\ \diagup \end{array} \right] &= [2] \left[ \begin{array}{c} \diagup \\ \diagup \end{array} \right] + ([2n - 3] + 1) \left[ \begin{array}{c} \diagup \\ \diagup \end{array} \right], \\
 \left[ \begin{array}{c} \diagdown \\ \diagdown \end{array} \right] &= [2] \left[ \begin{array}{c} \diagdown \\ \diagdown \end{array} \right] + ([2n - 3] + 1) \left[ \begin{array}{c} \diagdown \\ \diagdown \end{array} \right].
 \end{aligned}$$

Making the above replacements and combining like terms gives us

$$\begin{aligned}
 0 &= (q^2 - q[2]) \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] + ([2n - 2] + [2] - q[2n - 3] - q - q^{-1}[2n - 3] - q^{-1}) \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] \\
 &\quad + \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] \times \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] + (q^{-2} - q^{-1}[2]) \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] + \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] \times \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] + (q[2] - q^2) \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] \\
 &\quad - \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] \times \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] + (q^{-1}[2n - 3] + q^{-1} + q[2n - 3] + q - [2n - 2] - [2]) \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] \\
 &\quad \quad \quad + (q^{-1}[2] - q^{-2}) \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] - \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] \times \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] \\
 &= \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] \times \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] + \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] \times \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] - \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] - \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] - [2n - 4] \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] \\
 &\quad - \left( \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] \times \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] + \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] \times \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] - \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] - \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] - [2n - 4] \left[ \begin{array}{c} \diagup \\ \diagdown \end{array} \right] \right),
 \end{aligned}$$

and the statement follows. □

Propositions 1–4 provide consistent and sufficient skein relations to evaluate any planar unoriented 4-valent graph. In addition, the skein relation (4-5) together with these propositions yield a state summation model for the  $SO(2n)$  Kauffman polynomial.

### Acknowledgements

Caprau would like to thank Lorenzo Traldi for his useful comment and question via e-mail after the paper [Caprau and Tipton 2011] appeared on the arXiv, which motivated this work. Heywood’s contribution was partially supported by an Undergraduate Research Grant from the California State University, Fresno.

### References

[Caprau and Tipton 2011] C. Caprau and J. Tipton, “The Kauffman polynomial and trivalent graphs”, 2011. To appear in *Kyungpook Mathematical Journal*. arXiv math.GT/1107.1210

- [Carpentier 2000] R. P. Carpentier, “From planar graphs to embedded graphs—a new approach to Kauffman and Vogel’s polynomial”, *J. Knot Theory Ramifications* **9**:8 (2000), 975–986. MR 2011m:57006
- [Freyd et al. 1985] P. Freyd, D. Yetter, J. Hoste, W. B. R. Lickorish, K. Millett, and A. Ocneanu, “A new polynomial invariant of knots and links”, *Bull. Amer. Math. Soc. (N.S.)* **12**:2 (1985), 239–246. MR 86e:57007
- [Kauffman 1990] L. H. Kauffman, “An invariant of regular isotopy”, *Trans. Amer. Math. Soc.* **318**:2 (1990), 417–471. MR 90g:57007
- [Kauffman 2001] L. H. Kauffman, *Knots and physics*, 3rd ed., Series on Knots and Everything **1**, World Scientific Publishing Co., River Edge, NJ, 2001. MR 2002h:57012
- [Kauffman and Vogel 1992] L. H. Kauffman and P. Vogel, “Link polynomials and a graphical calculus”, *J. Knot Theory Ramifications* **1**:1 (1992), 59–104. MR 92m:57012
- [Murakami et al. 1998] H. Murakami, T. Ohtsuki, and S. Yamada, “Homfly polynomial via an invariant of colored plane graphs”, *Enseign. Math. (2)* **44**:3-4 (1998), 325–360. MR 2000a:57023
- [Przytycki and Traczyk 1988] J. H. Przytycki and P. Traczyk, “Invariants of links of Conway type”, *Kobe J. Math.* **4**:2 (1988), 115–139. MR 89h:57006
- [Wu 2012] H. Wu, “On the Kauffman–Vogel and the Murakami–Ohtsuki–Yamada graph polynomials”, *J. Knot Theory Ramifications* **21**:10 (2012), 1250098, 40. MR 2949230

Received: 2013-04-10

Revised: 2013-10-24

Accepted: 2013-10-27

ccaprau@csubfresno.edu

*Department of Mathematics, California State University,  
Fresno, 5245 N. Backer Avenue M/S PB108,  
Fresno, CA 93740-8001, United States*

davaudoo@gmail.com

*Department of Mathematics, California State University,  
Fresno, 5245 N. Backer Avenue M/S PB108,  
Fresno, CA 93740-8001, United States*

luxchasehidknd@yahoo.com

*Department of Mathematics, California State University,  
Fresno, 5245 N. Backer Avenue M/S PB108,  
Fresno, CA 93740-8001, United States*





# Invariant measures for hybrid stochastic systems

Xavier Garcia, Jennifer Kunze, Thomas Rudelius,  
Anthony Sanchez, Sijing Shao, Emily Speranza and Chad Vidden

(Communicated by David Royal Larson)

In this paper, we seek to understand the behavior of dynamical systems that are perturbed by a parameter that changes discretely in time. If we impose certain conditions, we can study certain embedded systems within a hybrid system as time-homogeneous Markov processes. In particular, we prove the existence of invariant measures for each embedded system and relate the invariant measures for the various systems through the flow. We calculate these invariant measures explicitly in several illustrative examples.

## 1. Introduction

An understanding of dynamical systems allows one to analyze the way processes evolve through time. Usually, such systems are given by differential equations that model real world phenomena. Unfortunately, these models are limited in that they cannot account for random events that may occur in application. These stochastic developments, however, may sometimes be modeled with Markov processes, and in particular with Markov chains. We can unite the two models in order to see how these dynamical systems behave with the perturbation induced by the Markov processes, creating a hybrid system consisting of the two components. Complicating matters, these hybrid systems can be described in either continuous or discrete time.

The focus of this paper is studying the way these hybrid systems behave as they evolve. We begin by defining limit sets for a dynamical system and stochastic processes. We next examine the limit sets of these hybrid systems and what happens as they approach the limit sets. Concurrently, we define invariant measures and prove their existence for hybrid systems while relating these measures to the flow. In addition, we supply examples with visuals that provide insight to the behavior of hybrid systems.

---

*MSC2010:* 34F05, 60J20, 37N20.

*Keywords:* dynamical systems, Markov processes, Markov chains, stochastic modeling.

Research supported by NSF grant DMS 0750986 (Kunze, Rudelius, Speranza); DMS 0750986 and DMS 0502354 (Garcia and Sanchez); and Iowa State University (Shao and Vidden).

### 2. The stochastic hybrid system

In this section, we define a hybrid system.

**Definition 1.** A Markov process  $X_t$  is called *time-homogeneous* on  $T$  if, for all  $t_1, t_2, k \in T$  and for any sets  $A_1, A_2 \in S$ ,

$$P(X_{t_1+k} \in A_1 \mid X_{t_1} \in A_2) = P(X_{t_2+k} \in A_1 \mid X_{t_2} \in A_2).$$

Otherwise, it is called *time-inhomogeneous*.

**Definition 2.** A Markov chain  $X_n$  is a Markov process for which perturbations occur on a discrete time set  $T$  and finite state space  $S$ .

For a Markov chain on the finite state space  $S$  with cardinality  $|S|$ , it is useful to describe the probabilities of transitioning from one state to another with a transition matrix

$$Q \equiv \begin{pmatrix} P_{1 \rightarrow 1} & \dots & P_{1 \rightarrow |S|} \\ \vdots & & \vdots \\ P_{|S| \rightarrow 1} & \dots & P_{|S| \rightarrow |S|} \end{pmatrix},$$

where  $P_{i \rightarrow j}$  is the probability of transitioning from state  $s_i \in S$  to state  $s_j \in S$ .

Also, for the purposes of this paper, we suppose that our Markov chain transitions occur regularly at times  $t = nh$  for some length of time  $h \in \mathbb{R}^+$  and for all  $n \in \mathbb{N}$ .

**Definition 3.** Let  $\{X_n\}$ , for  $X_n \in S$  and  $n \in \mathbb{N}$ , be a sequence of states determined by a Markov chain.

For  $t \in \mathbb{R}^+$ , define the Markov chain perturbation  $Z_t = X_{\lfloor t/h \rfloor}$ , where  $\lfloor t/h \rfloor$  is the greatest integer less than or equal to  $t/h$ .

Note that  $Z_t$ , instead of being defined only on discrete time values like a Markov chain, is instead a stepwise function defined on continuous time.

**Definition 4.** Given a metric space  $M$  and state space  $S$  as above, define a dynamical system  $\varphi$  with random perturbation function  $Z_t$ , as given in Definition 3, by

$$\varphi : \mathbb{R}^+ \times M \times S \rightarrow M,$$

with

$$\varphi(t, x_0, Z_0) = \varphi_{Z_t}(t - nh, \varphi_{Z_{nh}}(h, \dots \varphi_{Z_{2h}}(h, \varphi_{Z_h}(h, \varphi_{Z_0}(h, x_0))))),$$

where  $\varphi_{Z_k}$  represents the deterministic dynamical system  $\varphi$  evaluated in state  $Z_k$  and  $nh$  is the largest multiple of  $h$  less than  $t$ .

For ease of notation, let

$$x_t = \varphi(t, x_0, Z_0) \in M$$

represent the position of the system at time  $t$ .

**Definition 5.** Let

$$Y_t = \begin{pmatrix} x_t \\ Z_t \end{pmatrix}$$

define the hybrid system at time  $t$ . In other words, the hybrid system consists of both a position  $x_t = \varphi(t, x_0, Z_0) \in M$  and a state  $Z_t \in S$ .

The  $\omega$ -limit set has the following generalization in a hybrid system.

**Definition 6.** The stochastic limit set  $C(x)$  for an element of our state space  $x \in M$  and the hybrid system given above is the subset of  $M$  with the following three properties:

- (1) Given  $y \in M$  and  $t_k \rightarrow \infty$  such that  $x_{t_k} \rightarrow y$ ,  $P(y \in C(x)) = 1$ .
- (2)  $C(x)$  is closed.
- (3)  $C(x)$  is minimal: if some set  $C'(x)$  has properties 1 and 2, then  $C \subseteq C'$ .

### 3. The hybrid system as a Markov process

**Lemma 7.** *Each of the following is a Markov process:*

- (i) *Any deterministic dynamical system  $\varphi(t, x_0)$ .*
- (ii) *Any Markov chain perturbation  $Z_t$ , as in Definition 2.*
- (iii) *The corresponding hybrid system  $Y_t$ , as in Definition 5.*

*Proof.* (i) Any deterministic system is trivially a Markov process, since  $\varphi(t, x_0)$  is uniquely determined by  $\varphi(\tau, x_0)$  at any single past time  $\tau \in \mathbb{R}^+$ .

(ii) By definition, a Markov chain is a Markov process. However, the Markov chain perturbation  $Z_t$  is not exactly a Markov chain. A Markov chain exists on a discrete time set, in our case given by  $T = \{t \in \mathbb{R}^+ \mid t = nh \text{ for some } n \in \mathbb{N}\}$ ; conversely, the time set of  $Z_t$  is  $\mathbb{R}^+$ , with transitions between states occurring on the previous time set (that is, at  $t \equiv 0 \pmod{h}$ ). Despite this difference,  $Z_t$  maintains the Markov property: we can compute  $P(Z_t \in A)$  for any set  $A$  based solely on  $Z_{\tau_1}$  and the values of the times  $t$  and  $\tau_1$ . Explicitly, the probability that  $Z_t$  will be in state  $s_i$  at time  $t$  is given by

$$P(Z_t = s_i) = ((Q^T)^n)_{ij},$$

where  $n$  is the number of integer multiples of  $h$  (i.e., the number of transitions that occur) between  $t$  and  $\tau_1$ . Clearly, this is independent of the states  $Z_{\tau_i}$  for  $i > 1$ , so that the random perturbation is indeed a Markov process.

(iii) Now, keeping in mind that the hybrid system  $Y_t$  consists of both a location  $x_t \in M$  in the state space and a value  $Z_t \in S$  of the random component, we can combine (i) and (ii) to see that the entire system is also a Markov process. We see from (ii) that  $Z_t$  follows a Markov process. Furthermore,  $P(x_t \in A_x)$  at time  $t$  depends solely on the location  $x_{\tau_1}$  at any time  $\tau_1 < t$  and the states of the random perturbation sequence  $Z$  between  $t$  and  $\tau_1$ , regardless of any past behavior of the system. Hence, for any collection of sets  $A_\alpha$ ,  $\alpha \in \mathbb{N}$ ,

$$P(Z_t \in A_z \mid Z_{\tau_1} \in A_{z_1}, Z_{\tau_2} \in A_{z_2}, \dots, Z_{\tau_n} \in A_{z_n}) = P(Z_t \in A_z \mid Z_{\tau_1} \in A_{z_1}),$$

$$P(x_t \in A_x \mid x_{\tau_1} \in A_{x_1}, x_{\tau_2} \in A_{x_2}, \dots, x_{\tau_n} \in A_{x_n}) = P(x_t \in A_x \mid x_{\tau_1} \in A_{x_1}).$$

So,

$$P(Y_t \in A_y \mid Y_{\tau_1} \in A_{y_1}, Y_{\tau_2} \in A_{y_2}, \dots, Y_{\tau_n} \in A_{y_n}) = P(Y_t \in A_y \mid Y_{\tau_1} \in A_{y_1}).$$

Thus, the hybrid system is a Markov process.  $\square$

Unfortunately, the hybrid system is not time-homogeneous. Recall that state transitions of  $Z_t$  occur at times  $t = nh$  for  $n \in \mathbb{N}$ . So, the state of the system at time  $h/4$  uniquely determines the system at  $3h/4$ , since there is no transition in this interval. However, the system at time  $5h/4$  is not determined uniquely by the system at  $3h/4$ , since a stochastic transition occurs at  $t = h \in [\frac{3}{4}h, \frac{5}{4}h]$ . Therefore, with  $t_1 = h/4$ ,  $t_2 = 3h/4$ , and  $k = \frac{1}{2}$ ,

$$P(Y_{\frac{h}{4} + \frac{1}{2}} \in A \mid Y_{\frac{h}{4}} \in A_0) \neq P(Y_{\frac{3h}{4} + \frac{1}{2}} \in A \mid Y_{\frac{3h}{4}} \in A_0),$$

violating Definition 1. However, in order to satisfy the hypotheses of the Krylov–Bogolyubov theorem [Hairer 2010; 2006] found in Theorem 14, the hybrid system must be time-homogeneous.

To create a time-homogeneous system, we restrict the time set on which our Markov process is defined. Instead of allowing our time set

$$\{t, \tau_1, \tau_2, \tau_3, \dots, \tau_n\} \subset \mathbb{R}^+$$

to be any decreasing sequence of real numbers, we create time sets  $t_0 + nh$  for each  $t_0 \in [0, h)$  and  $n \in \mathbb{N}$ . In other words, we define a different time set for each value  $t_0 < h$  as

$$\{t \in \mathbb{R}^+ \mid t = t_0 + nh \text{ for some } n \in \mathbb{N}\}.$$

We call the hybrid system on these multiple, restricted time sets the *discrete system*.

**Proposition 8.** *The discrete hybrid system above is a time-homogeneous Markov process.*

*Proof.* First, we must show that the discrete hybrid system is a Markov process at all. This follows immediately from the proof that our original hybrid system is a Markov process. Since the Markov property holds for all  $t, \tau_1, \tau_2, \dots, \tau_n \in \mathbb{R}^+$ , it must necessarily hold for the specific time set

$$\{t \in \mathbb{R}^+ \mid \text{there exists } n \in \mathbb{N} \text{ such that } t = t_0 + nh\}$$

for each  $t_0 < h$ .

Now, it remains to show that this system is time-homogeneous. Recall that the time-continuous hybrid system failed to be time-homogeneous because its  $Z_t$  component was not time-homogeneous. Although transitions occurred only at regular, discrete time values, a test interval could be of any length; an interval of size  $h/2$ , for example, might contain either 0 or 1 transitions. However, because our discrete system creates separate time sets, any time interval — starting and ending within the same time set — must be of length  $nh$  for some  $n \in \mathbb{N}$ , and thus will contain precisely  $n$  potential transitions. So, taking  $t_1, t_2 \in \mathbb{R}^+$ , we know that

$$P(Y_{t_1+nh} \in A \mid Y_{t_1} \in A_0) = P(Y_{t_2+nh} \in A \mid Y_{t_2} \in A_0).$$

Note that the first component of the hybrid system,  $x_t$ , is also time-homogeneous under the discrete time system. Given  $Z_t$ , it can be treated as a deterministic system, and therefore time-homogeneous. Thus, the discrete hybrid system is time-homogeneous. □

#### 4. Invariant measures for the hybrid system

We now introduce several definitions that will lead to the main results of this paper.

**Definition 9.** Consider a hybrid system  $Y_t$  and a  $\sigma$ -algebra  $\Sigma$  on the space  $M$ . A measure  $\mu$  on  $M$  is invariant if, for all sets  $A \in \Sigma$  and all times  $t \in \mathbb{R}^+$ ,

$$\mu(A) = \int_{x_0 \in M} P(x_t \in A) \mu(dx).$$

**Definition 10.** Let  $(M, \mathcal{T})$  be a topological space, and let  $\Sigma$  be a  $\sigma$ -algebra on  $M$  that contains the topology  $\mathcal{T}$ . Let  $\mathcal{M}$  be a collection of probability measures defined on  $\Sigma$ . The collection  $\mathcal{M}$  is called tight if, for any  $\epsilon > 0$ , there is a compact subset  $K_\epsilon$  of  $M$  such that, for any measure  $\mu$  in  $\mathcal{M}$ ,

$$\mu(M \setminus K_\epsilon) < \epsilon.$$

Note that since  $\mu$  is a probability measure, it is equivalent to, say,  $\mu(K_\epsilon) > 1 - \epsilon$ .

The following definitions are from [Hairer 2010].

**Definition 11.** Let  $(M, \rho)$  be a separable metric space. Let  $\{\mathcal{P}(M)\}$  denote the collection of all probability measures defined on  $M$  (with its Borel  $\sigma$ -algebra).

A collection  $K \subset \{\mathcal{P}(M)\}$  of probability measures is tight if and only if  $K$  is sequentially compact in the space equipped with the topology of weak convergence.

**Definition 12.** Consider  $M$  with  $\sigma$ -algebra  $\Sigma$ . Let  $C^0(M, \mathbb{R})$  denote the set of continuous functions from  $M$  to  $\mathbb{R}$ . The probability measure  $\mathcal{P}(t, x, \cdot)$  on  $\Sigma$  induces a map

$$\mathcal{P}_t(x) : C^0(M, \mathbb{R}) \rightarrow \mathbb{R}, \quad \text{with} \quad \mathcal{P}_t(x)(f) = \int_{y \in M} f(y) \mathcal{P}(t, x, dy).$$

$\mathcal{P}_t$  is called a *Markov operator*.

**Definition 13.** A Markov operator  $\mathcal{P}$  is Feller if  $\mathcal{P}\varphi$  is continuous for every continuous bounded function  $\varphi : X \rightarrow \mathbb{R}$ . In other words, it is Feller if and only if the map  $x \mapsto \mathcal{P}(x, \cdot)$  is continuous in the topology of weak convergence.

We state the Krylov–Bogolyubov theorem without proof.

**Theorem 14 (Krylov–Bogolyubov).** *Let  $\mathcal{P}$  be a Feller Markov operator over a complete and separable space  $X$ . Assume that there exists  $x_0 \in X$  such that the sequence  $\mathcal{P}^n(x_0, \cdot)$  is tight. Then, there exists at least one invariant probability measure for  $\mathcal{P}$ .*

We now show that the conditions of the theorem are satisfied by the discrete hybrid system, yielding the existence of invariant measures as a corollary.

**Lemma 15.** *Given  $t_0 \in [0, h)$ , the discrete hybrid system Markov operators  $\mathcal{P}_n$  for  $n \in \mathbb{N}$  given by*

$$\mathcal{P}_n f(Y) \equiv \int_{M \times S} f(Y_1) \mathcal{P}(nh, Y, dY_1)$$

*are Feller.*

*Proof.* We begin by showing that  $\mathcal{P}_1$  is Feller. By induction, it follows that  $\mathcal{P}_n$  is Feller for all  $n \in \mathbb{N}$ . It is clear that there are only finitely many possible outcomes of running the hybrid system for time  $h$ . Namely, there are at most  $|S|$  possible outcomes, where  $|S|$  denotes the cardinality of  $S$ . Given

$$Y_0 = \begin{pmatrix} x_0 \\ Z_0 = s_i \end{pmatrix} \in M \times S,$$

the only possible outcomes at time  $t = 1$  are

$$Y_1^j = \begin{pmatrix} \varphi_j(t_0, \varphi_i(h - t_0, x)) \\ s_j \end{pmatrix}$$

for  $j \in \{1, \dots, |S|\}$ , where  $\varphi_i, \varphi_j$  are the flows of the dynamical systems corresponding to states  $s_i$  and  $s_j$ , respectively. The probability of the  $j$ -th outcome is

given by  $P_{i \rightarrow j}$ , the probability of transitioning from state  $s_i$  to state  $s_j$ . Therefore,

$$\mathcal{P}_1 f(Y) = \int_{M \times S} f(Y_1) \mathcal{P}(h, Y, dY_1) = \sum_{j=1}^{|S|} P_{i \rightarrow j} f(Y_1^j).$$

Each  $\varphi_i$  is continuous under the assumption that each flow is continuous with respect to its initial conditions. The map from  $s_i$  to  $s_j$  is continuous since  $S$  is finite, so every set is open and hence the inverse image of any open set is open. The function  $f$  is continuous by hypothesis, and any finite sum of continuous functions is also continuous. Therefore  $\mathcal{P}_1 f$  is also continuous, and hence  $\mathcal{P}_1$  is Feller.  $\square$

We see now that the conditions of Theorem 14 (Krylov–Bogolyubov) hold. Namely, because  $M$  and  $S$  are compact (the former by assumption, the latter since it is finite),  $M \times S$  is compact. Thus, any collection of measures is automatically tight, since we can take  $K_\epsilon = X$ . It is well known that any compact metric space is also complete and separable. Applying Theorem 14, then, gives the following corollary, which is one of the primary results of the paper.

**Corollary 16.** *The discrete hybrid system has an invariant measure for each  $t_0 \in [0, h)$ .*

So, rather than speaking of an invariant measure for the time-continuous hybrid system, we can instead imagine a periodic invariant measure cycling continuously through  $h$ . That is, for each time  $t_0 \in [0, h)$ , there exists a measure  $\mu_{t_0}$  such that for  $t \equiv 0 \pmod{h}$ ,

$$\mu_{t_0}(A) = \int_{Y \in M \times S} \mathcal{P}(t, Y, A) d\mu_{t_0}.$$

The measure  $\mu_{t_0}$  above is a measure on the product space  $M \times S$ , since this is where the hybrid system lives. However, what we are really after is an invariant measure on just  $M$ , the space where the dynamical system part of the hybrid system lives. Fortunately, we can define a measure on  $M$  by the following construction.

**Proposition 17.** *Given  $\mu_t$ , an invariant probability measure on  $M \times S$ , the function*

$$\tilde{\mu}_t(A) \equiv \mu_t(A, S),$$

where  $A \subseteq M$  is an invariant probability measure on  $M$ .

*Proof.* The fact that  $\tilde{\mu}_t$  is a probability measure follows almost immediately from the fact that  $\mu_t$  is a probability measure. The probability that  $x_t \in \emptyset$  is 0, so  $\tilde{\mu}_t(\emptyset) = 0$ . The probability that  $x_t \in M$  is 1, so  $\tilde{\mu}_t(M) = 1$ . Countable additivity of  $\tilde{\mu}_t$  follows from countable additivity of  $\mu_t$ . Therefore,  $\tilde{\mu}_t$  is a probability measure on  $M$ .  $\square$

Thus far, we have proven the existence of a measure  $\mu_{t_0}$  for  $t_0 \in [0, h)$  such that for  $t \equiv 0 \pmod{h}$ ,

$$\mu_{t_0}(A) = \int_{x_0 \in M, s \in S} P(\varphi(t, x_0, s) \in A) d\mu_{t_0}.$$

The following theorem relates the collection of invariant measures  $\{\tilde{\mu}_{t_0}\}$  using the flow  $\varphi$ . This is the main result of the paper.

**Theorem 18.** *Given invariant measure  $\mu_0$ , the measure  $\mu_t$  defined by*

$$\mu_t(A) = \sum_{s \in S} \int_{x_0 \in M} P(\varphi(t, x_0, s) \in A) d\mu_0$$

*is also invariant in the sense that  $\mu_t = \mu_{t+nh}$  for  $n \in \mathbb{N}$ .*

*Proof.* We will show that  $\mu_t = \mu_{t+h}$ . By induction, this implies that  $\mu_t = \mu_{t+nh}$  for all  $n \in \mathbb{N}$ . We have

$$\mu_{t+h}(A) = \sum_{s \in S} \int_{x_0 \in M} P(\varphi(t+h, x_0, s) \in A) d\mu_0.$$

Applying the definition of conditional probability,

$$\begin{aligned} & \sum_{s \in S} \int_{x_0 \in M} P(\varphi(t+h, x_0, s) \in A) d\mu_0 \\ &= \sum_{r \in S} \int_{y \in M} \left[ P(\varphi(t, y, r) \in A) \sum_{s \in S} \int_{x_0 \in M} P(\varphi(h, x_0, s) \in dy \times \{r\}) d\mu_0 \right]. \end{aligned}$$

Loosely speaking, the probability that a trajectory beginning at  $(x, s)$  will end in a set  $A$  after a time  $t+h$  is the product of the probability that a trajectory beginning at  $(y, r)$  will end in  $A$  after a time  $t$  multiplied by the probability that a trajectory beginning at  $(x, s)$  will end at  $(y, r)$  after a time  $h$ , integrating over all possible pairs  $(y, r)$ . Here, we have implicitly used the fact that the hybrid system is a Markov process to ensure that the state of the system at time  $t+h$  given the state at time  $h$  is independent of the initial state, and we have avoided the problem of time-inhomogeneity by considering trajectories that only begin at times congruent to 0 (mod  $h$ ).

Furthermore, we have

$$\mu_h(dy \times \{r\}) = \sum_{s \in S} \int_{x_0 \in M} P(\varphi(h, x_0, s) \in dy \times \{r\}) d\mu_0$$

and

$$\mu_h(dy \times \{r\}) = d\mu_h(y, r);$$



so,

$$\mu_{t+h}(A) = \sum_{r \in S} \int_{y \in M} P(\varphi(t, y, r) \in A) d\mu_h.$$

Since  $\mu_0$  is invariant by assumption,  $\mu_0 = \mu_h$ . Therefore,

$$\mu_{t+h}(A) = \sum_{r \in S} \int_{y \in M} P(\varphi(t, y, r) \in A) d\mu_0 = \mu_t(A). \quad \square$$

### 5. Examples

Some examples of hybrid systems can be found in [Ayers 2010; Baldwin 2007]. Here, we will examine two simple cases to illustrate the theory developed above.

**5.1. A one-dimensional hybrid system.** We begin with a one-dimensional linear dynamical system with a stochastic perturbation:

$$\dot{x} = -x + Z_t,$$

where  $Z_t \in \{-1, 1\}$ . Both components of this system have a single, attractive equilibrium point: for  $Z_t = 1$ , this is  $x = 1$ , and for  $Z_t = -1$ ,  $x = -1$ . At timesteps of length  $h = 1$ ,  $Z_t$  is perturbed by a Markov chain given by the transition matrix  $Q$ .  $Q$  is therefore a  $2 \times 2$  matrix of nonnegative entries,

$$Q = \begin{pmatrix} P_{1 \rightarrow 1} & P_{1 \rightarrow -1} \\ P_{-1 \rightarrow 1} & P_{-1 \rightarrow -1} \end{pmatrix},$$

where  $P_{i \rightarrow j}$  gives the probability of the equilibrium point transitioning from  $i$  to  $j$  at each integer timestep. Since the total probability measure must equal 1,

$$\sum_{j \in \{1, -1\}} P_{i \rightarrow j} = 1, \quad i \in \{1, -1\}.$$

Furthermore, to avoid the deterministic case, we take  $P_{i \rightarrow j} \neq 0$  for all  $i, j$ .

**Proposition 19.** *The stochastic limit set  $C(x_0) = [-1, 1]$  for all  $x_0 \in \mathbb{R}$ .*

*Proof.* We begin by showing that  $C(x) \subset [-1, 1]$ : that is, that every possible trajectory in our system will eventually enter and never leave  $[-1, 1]$ , meaning that no it is only possible to have  $t^* \rightarrow \infty$  such that  $x^* = y$  for  $y \in [-1, 1]$ . First, consider  $x_0 \in [-1, 1]$ . If we are in state  $Z_t = 1$ , then the trajectory is attracted upwards and bounded above by  $x = 1$ ; in state  $Z_t = -1$ , the trajectory is attracted downwards and bounded below by  $x = -1$ . In both cases, the trajectory cannot move above 1 or below  $-1$ , and so will remain in  $[-1, 1]$  for all time.

Now, consider  $x_0 \notin [-1, 1]$ . If the trajectory ever enters  $[-1, 1]$ , by similar argument as above, it will remain in that region for all time. So, it remains to show that  $\varphi(t, x_0, Z_0) \in [-1, 1]$  for some  $t \in \mathbb{R}$ . First, take  $x_0 > 1$ . In either state,

the trajectory will be attracted downward, and will eventually enter  $[1, 2]$  at time  $t_2$ . Once there, at the first timestep in which  $Z_t = -1$  it will cross  $x = 1$  and enter  $[-1, 1]$ . And since we have taken all entries of the transition probability matrix  $Q$  to be nonzero, there almost surely exists a time  $t_3 > t_2$  for which the state is  $Z_t = -1$ ; then, the trajectory will enter  $[-1, 1]$  and never leave. By similar argument, any trajectory starting at  $x_0 < -1$  will enter and never leave  $[-1, 1]$ . Thus,  $C(x) \subset [-1, 1]$ .

Now, we must show that  $[-1, 1] \in C(x)$ : that is, that for every trajectory  $\varphi(t, x_0, Z_0)$  and every point  $y \in [-1, 1]$  there is  $t^* \rightarrow \infty$  such that  $\varphi(t^*, x_0, Z_0) \rightarrow y$ . To do this, we really only need to show that given any point  $x_0 \in [-1, 1]$  and any transition matrix  $Q$ , there almost surely exists some time  $t^*$  with  $\varphi(t^*, x_0, Z_0) = x^*$ . If one such time  $t^*$  is guaranteed to exist, then we can iterate the process for a solution beginning at  $(t^*, x^*)$  to produce an infinite sequence of times. To show that  $t^*$  exists, we calculate a lower bound on the probability that  $\varphi(t_n, x_0, Z_0) = x^*$ .

Without loss of generality, suppose that  $x_0 > x^*$ . We have already shown that any solution will enter  $[-1, 1]$ , so take  $\sup(x_0) = 1$ . From here, we can calculate the minimum number of necessary consecutive periods,  $k$ , for which  $Z_n = -1$  in order for a solution with  $x_0 = 1$  to decay to  $x^*$ . The probability of this sequence of  $k$  consecutive periods occurring is given by

$$P_1(k) = (P_{1 \rightarrow -1})(P_{-1 \rightarrow -1})^{k-1}$$

if  $Z_0 = 1$  and

$$P_{-1}(k) = (P_{-1 \rightarrow -1})^k$$

if  $Z_0 = -1$ . Thus, for some  $t^* \in [0, k]$ ,

$$P(\varphi(t^*, x_0, Z_0) = x^*) \geq \min(P_1(k), P_{-1}(k)) > 0,$$

since  $P_{i \rightarrow j} > 0$ . So,

$$P(t^* \notin [0, k]) \leq 1 - P(x^*) < 1 \quad \text{and} \quad P(t^* \notin [0, mk]) \leq (1 - P(x^*))^m.$$

As  $m \rightarrow \infty$ ,  $(1 - P(x^*))^m \rightarrow 0$ . So, with probability 1, there exists  $t^*$  with  $\varphi(t^*, x_0, Z_0) = x^*$ .

By similar argument, for  $x_0 < x^*$  and all  $x^* \in (-1, 1)$ , we can find a time sequence  $\{t_n\}$  such that  $\varphi(t_n, x_0, Z_0) = x^*$ . So, we know that for all  $x^* \in (-1, 1)$ ,  $x^* \in C(x)$ .

So, we have proven that  $[-1, 1] \subseteq C(x)$  and  $(-1, 1) \subseteq C(x)$ . Since  $C(x)$  must by definition be closed,  $C(x) = [-1, 1]$ .  $\square$

We can study the behavior of this system numerically. Figure 1 (left) depicts a solution calculated for the transition matrix

$$Q_1 = \begin{pmatrix} 0.4 & 0.6 \\ 0.5 & 0.5 \end{pmatrix},$$

with initial values  $x_0 = 2, Z_0 = 1$ .

As expected, the trajectory enters the interval  $(-1, 1)$  and stays there for all time, oscillating between  $x = -1$  and  $x = 1$ . Intuitively, it seems that the trajectory will cross any  $x^*$  in this interval repeatedly, so that indeed  $C(x) = [-1, 1]$ . This is not quite so clear for the transition matrix

$$Q_2 = \begin{pmatrix} 0.1 & 0.9 \\ 0.1 & 0.9 \end{pmatrix},$$

which yields the trajectory shown in Figure 1 (right) for  $x_0 = 2, Z_0 = 1$ .

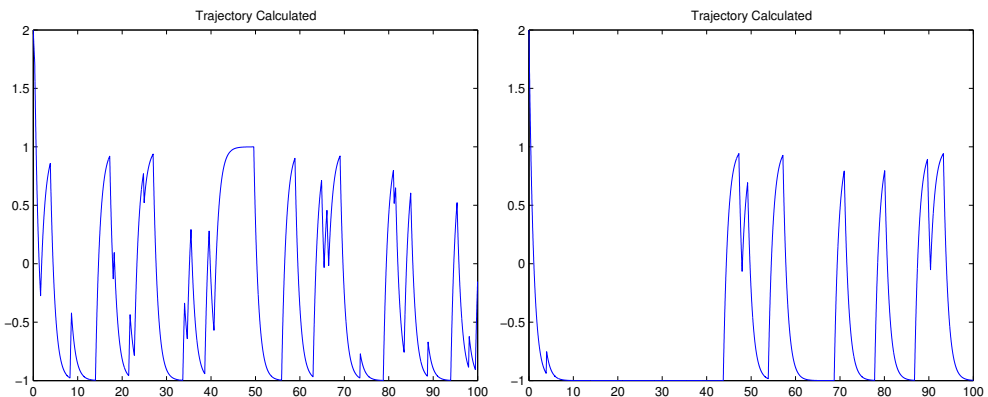
It may appear that some set of points near  $x = 1$  might be crossed by our path only a finite number of times. But, as proven above, any point in  $(-1, 1)$  will almost surely be reached infinitely many times as  $t \rightarrow \infty$ , so  $C(x) = [-1, 1]$ .

Now, we consider the eigenvalues and eigenvectors of the transition matrices. The eigenvector of  $Q_1^T$  with eigenvalue 1 is

$$\vec{v} = \begin{pmatrix} \frac{5}{11} \\ \frac{6}{11} \end{pmatrix},$$

and the eigenvector of  $Q_2^T$  with eigenvalue 1 is

$$\vec{v}' = \begin{pmatrix} \frac{9}{10} \\ \frac{1}{10} \end{pmatrix}.$$



**Figure 1.** A sample trajectory for a hybrid system with transition matrix  $Q_1$  (left) and  $Q_2$  (right).

These eigenvectors give the invariant measures on the state space  $S$ . We know from Proposition 17 that there also exists an invariant measure on  $M$ . Here, since any trajectory in  $M$  will almost surely enter  $C(x) = [-1, 1]$ , the support of the invariant measure must be contained in  $C(x)$ . It is not difficult to see that this invariant measure cannot be constant for all  $t \in \mathbb{R}^+$ . Given any point  $x_0 \in [-1, 1]$ , we know that at  $t = 1$ , one of two things will have happened to the trajectory:

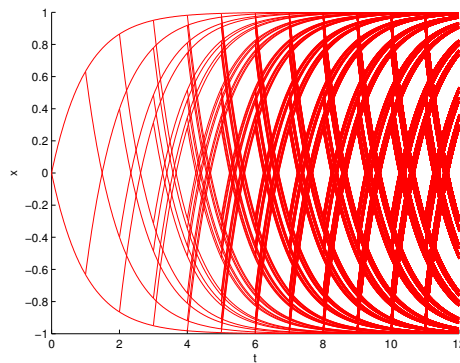
- (i) it will have decayed exponentially toward  $x = 1$ , if  $Z_1 = 1$ , or
- (ii) it will have decayed exponentially toward  $x = -1$ , if  $Z_1 = -1$ .

In case (i), if a solution begins at  $x_0 = -1$  for  $t = 0$ , the solution will have decayed to a value of  $1 - 2e^{-1} \approx 0.264$  by  $t = 1$ . In case (ii) a solution beginning at  $x_0 = 1$  for  $t = 0$  will decay to a value of  $-1 + 2e^{-1} \approx -0.264$ . Thus, if we are in case (i), all trajectories in  $[-1, 1]$  at  $t = n$  will be located in  $[0.264, 1]$  at  $t = n + 1$ . If we are in case (ii), all will be in  $[-1, -0.264]$ . It is not possible for any trajectory to be located in  $[-0.264, 0.264]$  at an integer time value. But, clearly, some solutions will cross into this region, as depicted in Figure 2. Therefore, no probability distribution will remain constant for all  $t$  in the time set  $\mathbb{R}^+$ .

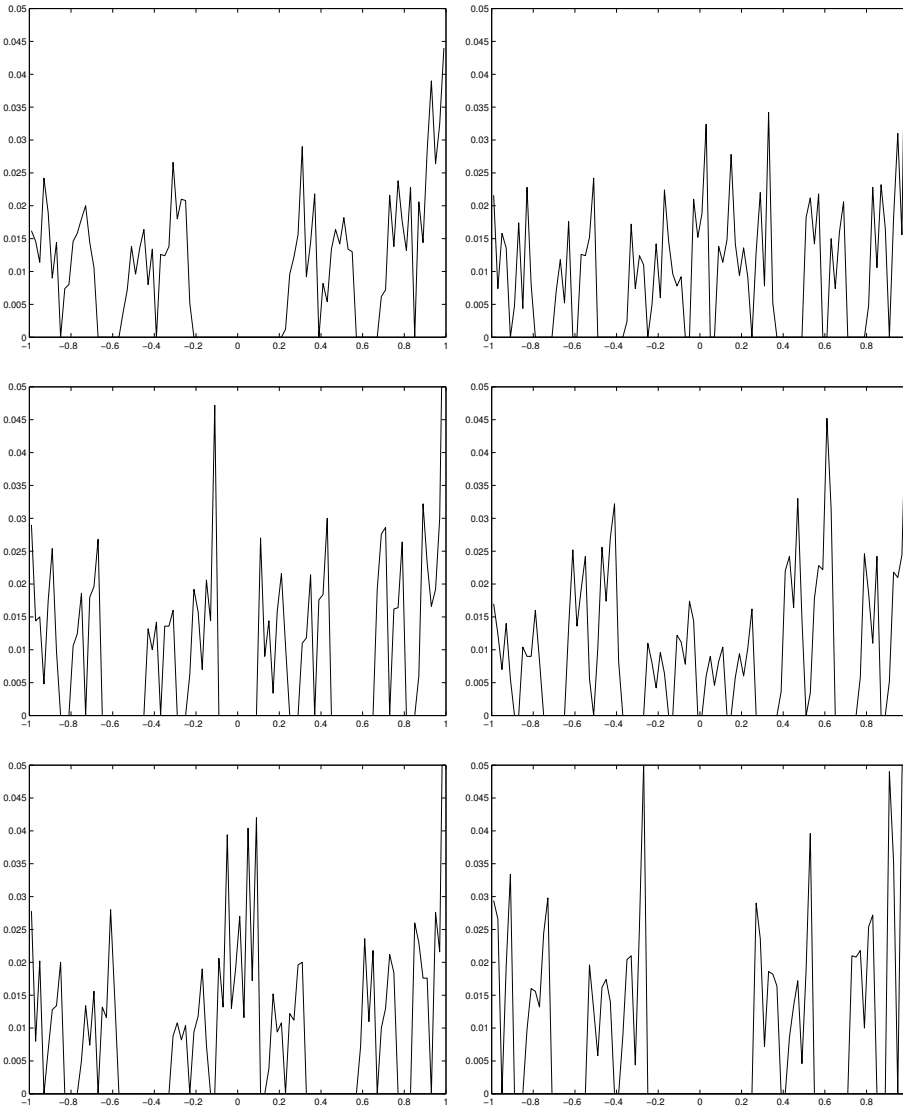
However, as Figure 2 suggests, there is some distribution that is invariant under  $t \rightarrow t + n$  for  $n \in \mathbb{N}$ . Approximations of the invariant measures at  $t \in [0, 1]$  for transition matrix  $Q_1$  are shown in Figure 3.

**5.2. A two-dimensional hybrid system.** Our second example is a two-dimensional system used to model the kinetics of chemical reactors. The general system  $f(x_1, x_2)$  is given by

$$\begin{aligned}\dot{x}_1 &= -\lambda x_1 - \beta(x_1 - x_c) + B Da f(x_1, x_2), \\ \dot{x}_2 &= -\lambda x_2 + Da f(x_1, x_2),\end{aligned}$$



**Figure 2.** A spider plot showing all possible trajectories starting at  $x_0 = 0$ .

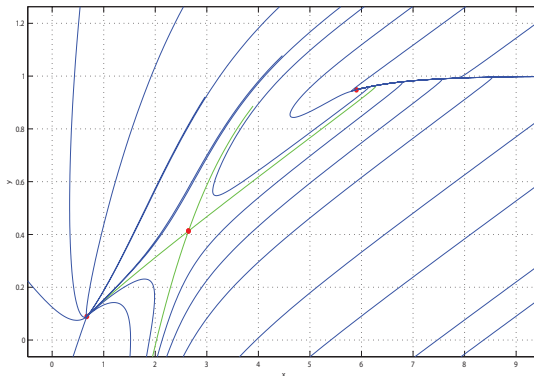


**Figure 3.** The invariant measure  $\tilde{\mu}_0$  for a hybrid system with transition matrix  $Q_1$ .

where  $\lambda$ ,  $\beta$ ,  $x_c$ ,  $Da$ , and  $B$  are physical parameters (see [Poore 1973]). Here, we use a simplified application of the system:

$$\begin{aligned} \dot{x}_1 &= -x_1 - 0.15(x_1 - 1) + 0.35(1 - x_2)e^{x_1} + Z_t(1 - x_1), \\ \dot{x}_2 &= -x_2 + 0.05(1 - x_2)e^{x_1}. \end{aligned}$$

This system is used to describe a continuous stirred tank reactor (CSTR). This type



**Figure 4.** Phase plane of the deterministic system,  $Z_n = 0$ .

of reactor is used to control chemical reactions that require a continuous flow of reactants and products and are easy to control the temperature with. They are also useful for reactions that require working with two phases of chemicals.

To understand the behavior of this system mathematically, we set our stochastic variable  $Z_t = 0$  and treat it as a deterministic system. This system has three fixed points, approximately at  $(0.67, 0.09)$ ,  $(2.64, 0.41)$ , and  $(5.90, 0.95)$ ; the former and latter are attractor points, while the middle is a saddle point, as shown in Figure 4. The saddle point  $(2.64, 0.41)$  creates a separatrix, a repelling equilibrium line between the two attracting fixed points. These points,  $(0.67, 0.09)$  and  $(5.90, 0.95)$ , comprise the  $\omega$ -limit set of our state space.

With this information, we proceed to analyze the stochastic system. As discussed above, the random variable here is  $Z_t$ , which in applications can take values between  $-0.15$  and  $0.15$ . To understand the full variability of this system, we take

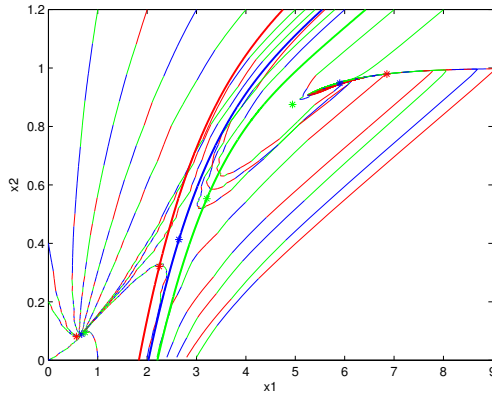
$$Z_t \in \{-0.15, 0, 0.15\}$$

with the transition matrix

$$\begin{pmatrix} 0.3 & 0.3 & 0.4 \\ 0.3 & 0.3 & 0.4 \\ 0.3 & 0.3 & 0.4 \end{pmatrix},$$

yielding the phase plane in Figure 5.

We see that, for  $x_0$  away from the separatrices,  $\varphi(t, x_0, Z_0)$  behaves similarly to  $\varphi(t, x_0)$ . Although state changes create some variability in a given trajectory, these paths move toward the groups of associated attracting fixed points, which define the stochastic limit sets for this system. However,  $\varphi(t, x_0, Z_0)$  for  $x_0$  between the red and green separatrices is unpredictable; depending on the sequence of state changes for a given trajectory, it might move either to the right or the left of the region



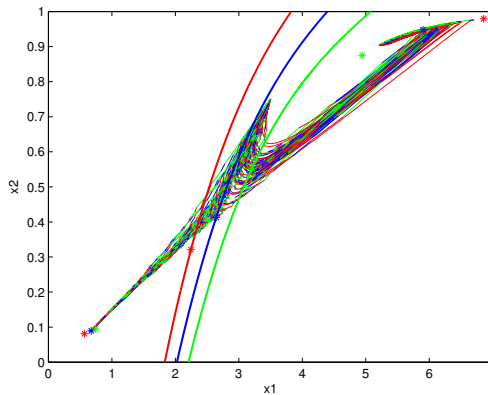
**Figure 5.** Phase plane with randomness, showing fixed points, separatrices, and portions of trajectories. Red, blue and green indicate states 1 ( $Z_t = -0.15$ ), 2 ( $Z_t = 0$ ) and 3 ( $Z_t = 0.15$ ).

defined by the separatrices. This area is the *bistable region*, because a trajectory beginning within it has two separate stochastic limit sets.

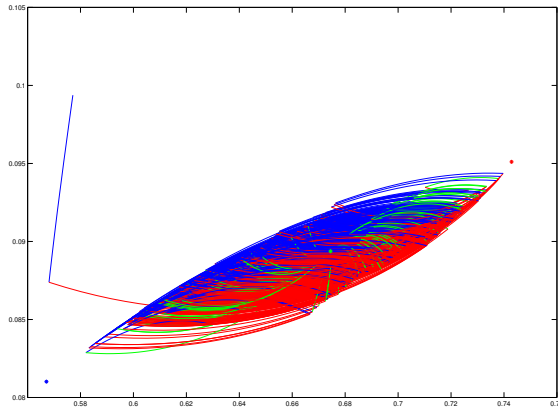
For example, we have in Figure 6 a spider plot beginning in the bistable region at (3.5, 0.75). A spider plot shows all possible trajectories starting from a single point in a hybrid system by, at each timestep, taking every possible state.

Thus, we see that the introduction of a stochastic element to a deterministic system can grossly affect the outcome of the system, as a trajectory can now cross any of the separatrices by being in a different state.

The stochastic element also affects the behavior of the hybrid system around the invariant region. In Figure 7, we show the path of a single trajectory in the invariant region defined by the fixed points near (0.67, 0.9). Plotting this trajectory for a long period of time approximates the invariant region that would appear if we ran a spider plot from the same point, but much more clearly.

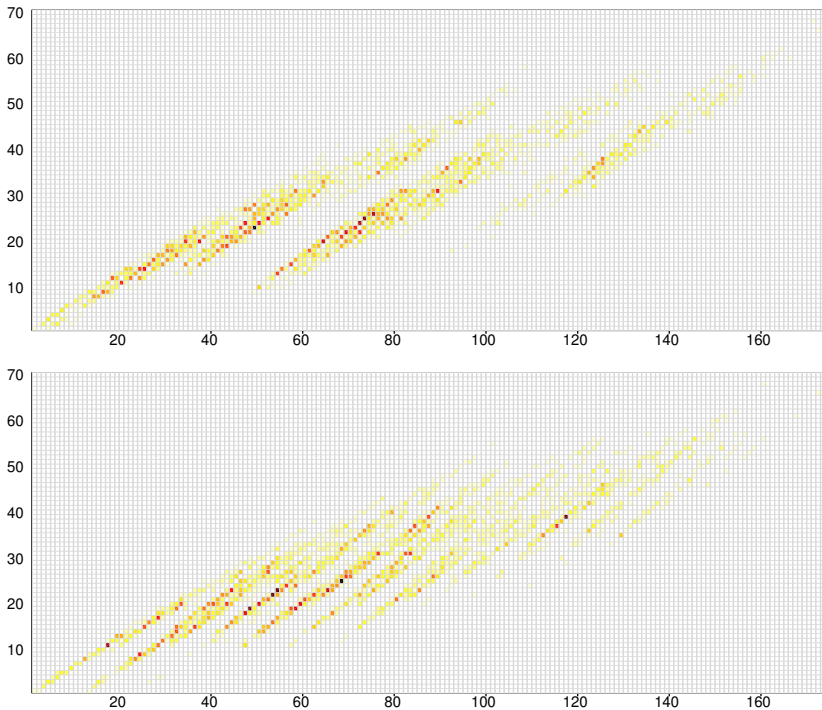


**Figure 6.** Spider plot. Color scheme as in Figure 5.



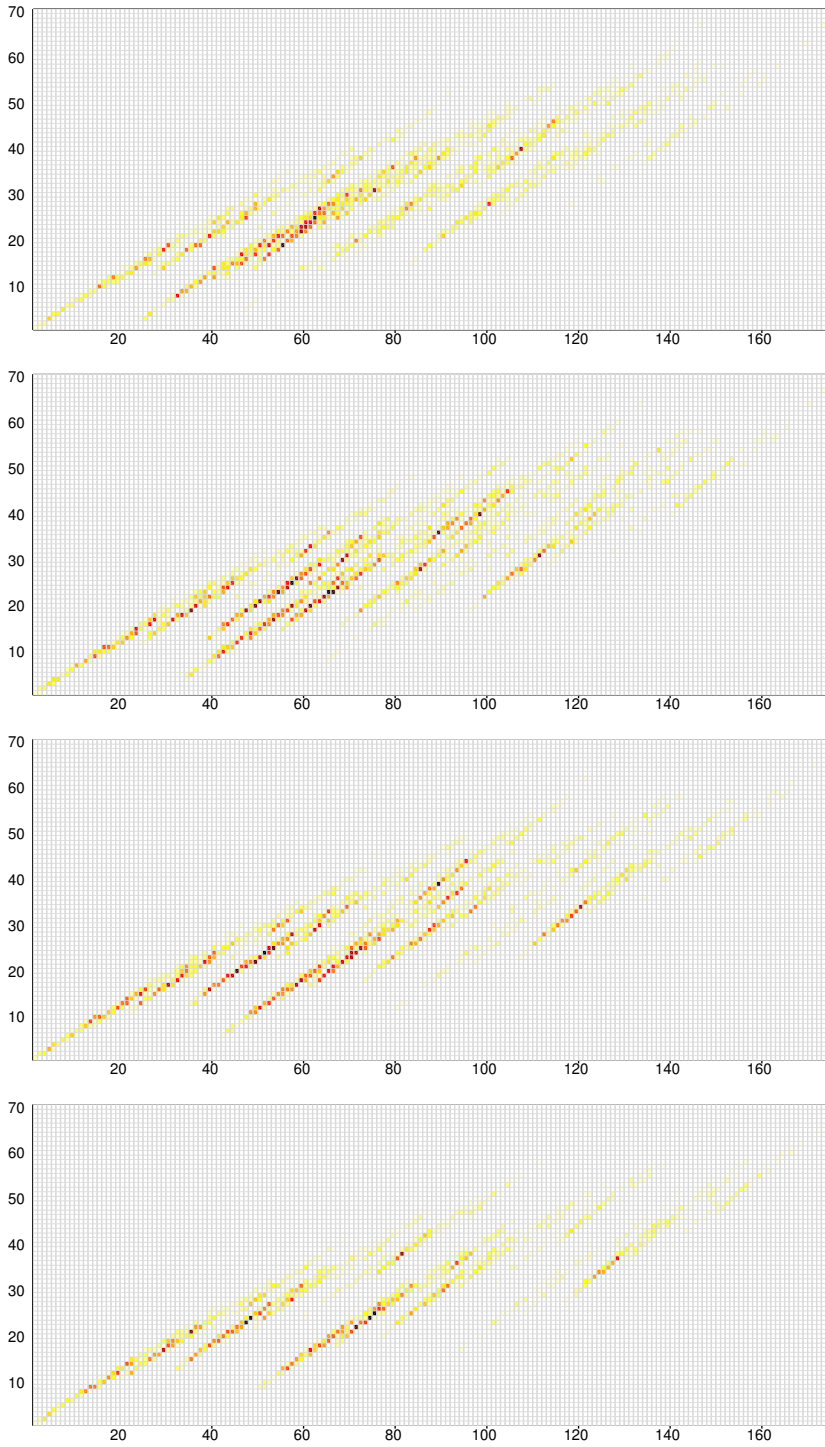
**Figure 7.** Random trajectory.

As we saw in the one-dimensional system, considering the counts taken at specific times in the interval between two state changes,  $h = 1$  (since our state transitions occur on  $\mathbb{N}$ ), yields a periodic set of invariant measures. Similarly to Figure 3, Figure 8 shows the positions of our random trajectory in the invariant region at time  $t, \text{ mod } h$ .



**Figure 8.** Count of trajectory paths within one timestep.





**Figure 8.** Count of trajectory paths within one timestep (continued).

A denser series of count images would show more clearly that the invariant measure at  $t \bmod h$  cycles continuously.

## 6. Conclusion

We have studied hybrid systems consisting of a finite set  $S$  of dynamical systems over a compact space  $M$  with a Markov chain on  $S$  acting at discrete time intervals. Such a hybrid system is a Markov process, which can be made time-homogeneous by discretizing the system. Then, there exists a family of invariant measures on the product space  $M \times S$ , which can be projected onto a family of measures on  $M$ . We have demonstrated a relation between the members of this family.

We have studied both a one-dimensional and a two-dimensional example of a hybrid system. These examples provide insight into the stochastic equivalent of  $\omega$ -limit sets and yield graphical representations of the invariant measures on these sets.

## Acknowledgements

We wish to recognize Kimberly Ayers for her helpful discussions and Professor Wolfgang Kliemann for his instruction and guidance. We would like to thank the Department of Mathematics at Iowa State University for their hospitality during the completion of this work. In addition, we would like to thank Iowa State University, Alliance, and the National Science Foundation for their support of this research. Figure 4 was drawn using the “pplane8.m” Matlab program.

## References

- [Ayers 2010] K. D. Ayers, “Stochastic perturbations of the Fitzhugh–Nagumo equations”, Undergraduate honors thesis, Bowdoin College, 2010.
- [Baldwin 2007] M. C. Baldwin, “Stochastic analysis of Marotzke and Stone climate model”, Master’s thesis, Iowa State University, 2007.
- [Hairer 2006] M. Hairer, “Ergodic properties of Markov processes”, lecture notes, 2006, available at <http://www.hairer.org/notes/Markov.pdf>.
- [Hairer 2010] M. Hairer, “Convergence of Markov processes”, lecture notes, 2010, available at <http://www.hairer.org/notes/Convergence.pdf>.
- [Poore 1973] A. B. Poore, “A model equation arising from chemical reactor theory”, *Arch. Rational Mech. Anal.* **52** (1973), 358–388. MR 49 #3272

Received: 2013-07-16      Accepted: 2013-10-05

garci363@umn.edu

*Department of Mathematics, University of Minnesota,  
Minneapolis, MN 55455, United States*

jckunze@smcm.edu

*Mathematics and Computer Science Department, St. Mary’s  
College of Maryland, St. Mary’s City, MD 20686, United States*

twr27@cornell.edu	<i>Department of Mathematics, Cornell University, Ithaca, NY 14850, United States</i>
anthony.sanchez.1@asu.edu	<i>School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ 85287-1804, United States</i>
sshao@iastate.edu	<i>Department of Mathematics, Iowa State University, Ames, IA 50011, United States</i>
esperanza@carroll.edu	<i>Department of Mathematics, Engineering, and Computer Science, Carroll College, Helena, MT 59625, United States</i>
cvidden@iastate.edu	<i>Department of Mathematics, Iowa State University, Ames, IA 50011, United States</i>



## 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 L<sup>A</sup>T<sub>E</sub>X but submissions in other varieties of T<sub>E</sub>X, 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 BibT<sub>E</sub>X 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

2014

vol. 7

no. 4

Whitehead graphs and separability in rank two	431
MATT CLAY, JOHN CONANT AND NIVETHA RAMASUBRAMANIAN	
Perimeter-minimizing pentagonal tilings	453
PING NGAI CHUNG, MIGUEL A. FERNANDEZ, NIRALEE SHAH, LUIS SORDO VIEIRA AND ELENA WIKNER	
Discrete time optimal control applied to pest control problems	479
WANDI DING, RAYMOND HENDON, BRANDON CATHEY, EVAN LANCASTER AND ROBERT GERMICK	
Distribution of genome rearrangement distance under double cut and join	491
JACKIE CHRISTY, JOSH MCHUGH, MANDA RIEHL AND NOAH WILLIAMS	
Mathematical modeling of integrin dynamics in initial formation of focal adhesions	509
AURORA BLUCHER, MICHELLE SALAS, NICHOLAS WILLIAMS AND HANNAH L. CALLENDER	
Investigating root multiplicities in the indefinite Kac–Moody algebra $E_{10}$	529
VICKY KLIMA, TIMOTHY SHATLEY, KYLE THOMAS AND ANDREW WILSON	
On a state model for the $SO(2n)$ Kauffman polynomial	547
CARMEN CAPRAU, DAVID HEYWOOD AND DIONNE IBARRA	
Invariant measures for hybrid stochastic systems	565
XAVIER GARCIA, JENNIFER KUNZE, THOMAS RUDELIUS, ANTHONY SANCHEZ, SIJING SHAO, EMILY SPERANZA AND CHAD VIDDEN	

