involve

a journal of mathematics

Editorial Board

Kenneth S. Berenhaut, *Managing Editor*

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

mathematical sciences publishers

2009 vol. 2, no. 5

involve

pjm.math.berkeley.edu/involve

EDITORS

MANAGING EDITOR

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

BOARD OF EDITORS

John V. Baxley	Wake Forest University, NC, USA baxley@wfu.edu	Chi-Kwong Li	College of William and Mary, USA ckli@math.wm.edu
Arthur T. Benjamin	Harvey Mudd College, USA benjamin@hmc.edu	Robert B. Lund	Clemson University, USA lund@clemson.edu
Martin Bohner	Missouri U of Science and Technology, USA bohner@mst.edu	Gaven J. Martin	Massey University, New Zealand g.j.martin@massey.ac.nz
Nigel Boston	University of Wisconsin, USA boston@math.wisc.edu	Mary Meyer	Colorado State University, USA meyer@stat.colostate.edu
Amarjit S. Budhiraja	U of North Carolina, Chapel Hill, USA budhiraj@email.unc.edu	Emil Minchev	Ruse, Bulgaria eminchev@hotmail.com
Pietro Cerone	Victoria University, Australia pietro.cerone@vu.edu.au	Frank Morgan	Williams College, USA frank.morgan@williams.edu
Scott Chapman	Trinity University, USA schapman@trinity.edu	Mohammad Sal Moslehian	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
Michael Dorff	Brigham Young University, USA mdorff@math.byu.edu	Ken Ono	University of Wisconsin, USA ono@math.wisc.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
Errin W. Fulp	Wake Forest University, USA fulp@wfu.edu	Y.-F. S. Pétermann	Université de Genève, Switzerland petermann@math.unige.ch
Andrew Granville	Université Montréal, Canada andrew@dms.umontreal.ca	Robert J. Plemmons	Wake Forest University, USA plemmons@wfu.edu
Jerrold Griggs	University of South Carolina, USA griggs@math.sc.edu	Carl B. Pomerance	Dartmouth College, USA carl.pomerance@dartmouth.edu
Ron Gould	Emory University, USA rg@mathcs.emory.edu	Bjorn Poonen	UC Berkeley, USA poonen@math.berkeley.edu
Sat Gupta	U of North Carolina, Greensboro, USA sngupta@uncg.edu	James Propp	U Mass Lowell, USA jpropp@cs.uml.edu
Jim Haglund	University of Pennsylvania, USA jhaglund@math.upenn.edu	Józeph H. Przytycki	George Washington University, USA przytyck@gwu.edu
Johnny Henderson	Baylor University, USA johnny_henderson@baylor.edu	Richard Rebarber	University of Nebraska, USA rrebarbe@math.unl.edu
Natalia Hritonenko	Prairie View A&M University, USA nahritonenko@pvamu.edu	Robert W. Robinson	University of Georgia, USA rwr@cs.uga.edu
Charles R. Johnson	College of William and Mary, USA crjohnso@math.wm.edu	Filip Saidak	U of North Carolina, Greensboro, USA f_saidak@uncg.edu
Karen Kafadar	University of Colorado, USA karen.kafadar@cudenver.edu	Andrew J. Sterge	Honorary Editor andy@ajsterge.com
K. B. Kulasekera	Clemson University, USA kk@ces.clemson.edu	Ann Trenk	Wellesley College, USA atrenk@wellesley.edu
Gerry Ladas	University of Rhode Island, USA gladas@math.uri.edu	Ravi Vakil	Stanford University, USA vakil@math.stanford.edu
David Larson	Texas A&M University, USA larson@math.tamu.edu	Ram U. Verma	University of Toledo, USA verma99@msn.com
Suzanne Lenhart	University of Tennessee, USA lenhart@math.utk.edu	John C. Wierman	Johns Hopkins University, USA wierman@jhu.edu

PRODUCTION

Production Manager: Paulo Ney de Souza Production Editors: Silvio Levy, Sheila Newberry Cover design: ©2008 Alex Scorpan

See inside back cover or http://pjm.math.berkeley.edu/involve for submission instructions and subscription prices. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to Mathematical Sciences Publishers, Department of Mathematics, University of California, Berkeley, CA 94704-3840, USA.

Involve, at Mathematical Sciences Publisher, Department of Mathematics, University of California, Berkeley, CA 94720-3840 is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

PUBLISHED BY

mathematical sciences publishers

http://www.mathscipub.org

A NON-PROFIT CORPORATION

Typeset in LaTeX

Copyright ©2010 by Mathematical Sciences Publishers

INVOLVE 2:5(2009)

On the orbits of an orthogonal group action

Kyle Czarnecki, R. Michael Howe and Aaron McTavish

(Communicated by Józef H. Przytycki)

Let G be the Lie group $\mathrm{SO}(n,\mathbb{R})\times\mathrm{SO}(n,\mathbb{R})$ and let V be the vector space of $n\times n$ real matrices. An action of G on V is given by

$$(g,h).v := g^{-1}vh, \quad (g,h)\in G, \quad v\in V.$$

We consider the orbits of this group action and demonstrate a cross-section to the orbits. We then determine the stabilizer for a typical element in this cross-section and completely describe the fundamental group of an orbit of maximal dimension.

1. Introduction

Let G be the Lie group $\mathrm{SO}(n,\mathbb{R})\times\mathrm{SO}(n,\mathbb{R})$ and let V be the vector space of $n\times n$ real matrices. An action of G on V is given by

$$(g,h).v := g^t vh = g^{-1}vh, \quad (g,h)\in G, \quad v\in V,$$

where g^t denotes the matrix transpose of g and where the operation on the right is matrix multiplication. This action is obviously smooth (having continuous derivatives of all orders) since the matrix entries in $(g,h).v$ are polynomial functions of the matrix entries of g, h and v.

For each $v\in V$ we define the *orbit of* v, denoted by $G.v\subseteq V$, as the set

$$G.v := \{(g,h).v \mid (g,h)\in G\}.$$

For $v,w\in V$ the relation

$$v\sim w \text{ if } v \text{ and } w \text{ are in the same } G\text{-orbit}$$

MSC2000: primary 22C05, 57S15; secondary 55Q52.
Keywords: representation theory, orbit, Lie group, homotopy group, Clifford algebra.
This research was conducted in part during the 2007 summer Undergraduate Research Experience in Pure and Applied Mathematics at the University of Wisconsin – Eau Claire, supported by NSF-REU grant DMS-0552350 and the Office of Research and Sponsored Programs (UW – Eau Claire).

is an equivalence relation and so V is partitioned into G-orbits. We also define G_v, the *stabilizer of* v, to be the those elements in G that fix v:

$$G_v =: \{(g, h) \in G \mid (g, h).v = v\}.$$

For each $v \in V$, G_v is a closed (usually not normal) subgroup of G, and so is a Lie group.

Let G/G_v denote the set of left cosets of G_v in G. Since G_v is a closed subgroup of G, G/G_v is a differentiable manifold and $dim G/G_v = \dim G - \dim G_v$, where dim indicates the dimension. Furthermore, G/G_v is diffeomorphic to the orbit $G.v$. If G_v is normal in G, then G/G_v is a Lie group [Bröcker and tom Dieck 1985, Section1.4].

A subset D of V is a *cross-section* to the orbits if every G-orbit intersects D. That is, for each $v \in V$ there is an element $(g, h) \in G$ and an element $d \in D$ such that $(g, h).v = d$. Some definitions of a cross-section are more restrictive, requiring that each orbit intersect the cross-section exactly once.

In this paper we consider the orbits of this group action. In Section 2 we demonstrate a cross-section of the orbits, and in Section 3 we determine the stabilizer for a typical element in this cross-section. In Section 4 we discuss the orbits for the case $n = 2$ and introduce generic orbits — those of maximal dimension — for arbitrary n. Section 5 reviews some useful information about fundamental groups, covering spaces, and the covering group $\mathrm{Spin}(n)$. Our main result is in Section 6 where we connect these ideas in order to completely describe the fundamental group of a generic orbit, and in Section 7 we work through an example that further exposes the anatomy. We close with a few remarks in Section 8 regarding those orbits that do not have maximal dimension.

2. Cross section to the orbits

In this section we show that the diagonal matrices with non-negative entries constitute a cross-section to the group action.

Proposition 2.1. *Let $G = \mathrm{SO}(n) \times \mathrm{SO}(n)$ and let V be the vector space of $n \times n$ real matrices. Let G act on V via $(g, h).v = g^t v h$. Then for each $v \in V$ there is a $(k_1, k) \in G$ such that $(k_1, k).v = \mathrm{diagonal}(d_1, \ldots, d_n)$, with $d_1 \geq d_2 \geq \cdots \geq d_n \geq 0$.*

Proof. Let $v \in GL(n, \mathbb{R})$ where $GL(n, \mathbb{R})$ is the (dense, open) subset of invertible $n \times n$ matrices in V. Then $v^t v$ is a symmetric matrix with positive eigenvalues, and hence is diagonalizable via conjugation by an element in $\mathrm{SO}(n, \mathbb{R})$. That is, there is a k in $\mathrm{SO}(n, \mathbb{R})$ such that

$$k^t v^t v k = a,$$

where $a = \mathrm{diagonal}(a_1, \ldots, a_n)$ with $a_1 \geq a_2 \geq \cdots \geq a_n > 0$.

Now let $a^{-1/2} = \text{diagonal}(1/\sqrt{a_1}, \ldots, 1/\sqrt{a_n})$. If $\mathcal{I}_n$ is the $n \times n$ identity matrix we have

$$\mathcal{I}_n = a^{-1/2}\, a\, a^{-1/2} = a^{-1/2}\,[k^t v^t v k]\, a^{-1/2} = (vka^{-1/2})^t vka^{-1/2}.$$

It follows that $vka^{-1/2}$ is in $O(n, \mathbb{R})$. Let $a^{1/2} = \text{diagonal}(\sqrt{a_1}, \ldots, \sqrt{a_n})$. Then

$$a^{1/2} = \mathcal{I}_n a^{1/2} = [a^{-1/2} k^t v^t v k a^{-1/2}]\, a^{1/2} = a^{-1/2} k^t v^t v k.$$

Thus, if $k_1 = vka^{-1/2}$, we can write this as

$$(k_1)^t vk = (k_1, k).v = a^{1/2},$$

where $k_1 \in O(n, \mathbb{R})$ and $k \in \text{SO}(n, \mathbb{R})$. If k_1 happens to be in $\text{SO}(n, \mathbb{R})$ we are done. If not, we can change the sign of one of the entries in $a^{-1/2}$ so that k_1 *is* in $\text{SO}(n, \mathbb{R})$, proving the result for any V in the dense subset of invertible $n \times n$ matrices. Since our group action is continuous, the result holds for all $v \in V$. We could also modify the above proof slightly to account for those eigenvalues of $v^t v$ that are equal to zero. □

3. The stabilizer of a representative element

Let Γ be an arbitrary group acting on a set X. If x and y are in the same Γ-orbit, then $x = \gamma.y$ for some $\gamma \in \Gamma$. It is a standard result that $\gamma^{-1}\Gamma_x \gamma = \Gamma_y$, that is, the stabilizers are isomorphic via conjugation. Therefore, it is sufficient to determine the stabilizers of those elements that are in the cross section.

We start with a simple example that demonstrates the general idea for the situation that we are considering. Let $d \in V$ and $(g, h) \in G$ be given by

$$d = \begin{pmatrix} d_1 & 0 & 0 \\ 0 & d_1 & 0 \\ 0 & 0 & d_2 \end{pmatrix}, \quad \text{where } d_1 > d_2 > 0,$$

$$g = \begin{pmatrix} g_{1,1} & g_{1,2} & g_{1,3} \\ g_{2,1} & g_{2,2} & g_{2,3} \\ g_{3,1} & g_{3,2} & g_{3,3} \end{pmatrix}, \quad h = \begin{pmatrix} h_{1,1} & h_{1,2} & h_{1,3} \\ h_{2,1} & h_{2,2} & h_{2,3} \\ h_{3,1} & h_{3,2} & h_{3,3} \end{pmatrix}.$$

We may assume $d_1 > d_2$ since conjugation by a matrix such as

$$\begin{pmatrix} 0 & 0 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \in \text{SO}(3)$$

will reorder the entries in d.

If (g, h) stabilizes d then $g^t dh = d$ or equivalently, $dh = gd$, so we have

$$\begin{pmatrix} d_1h_{1,1} & d_1h_{1,2} & d_1h_{1,3} \\ d_1h_{2,1} & d_1h_{2,2} & d_1h_{2,3} \\ d_2h_{3,1} & d_2h_{3,2} & d_2h_{3,3} \end{pmatrix} = \begin{pmatrix} d_1g_{1,1} & d_1g_{1,2} & d_2g_{1,3} \\ d_1g_{2,1} & d_1g_{2,2} & d_2g_{2,3} \\ d_1g_{3,1} & d_1g_{3,2} & d_2g_{3,3} \end{pmatrix}. \tag{3-1}$$

That is, the first entry in d acts on the first row of h, but acts on the first column of g, etc. The rows of g and h are orthonormal (considered as vectors in $\mathbb{R}^3$ with the usual dot product), and we compare the squared length of the first row of dh with the first row of gd in (3-1):

$$(d_1h_{1,1})^2 + (d_1h_{1,2})^2 + (d_1h_{1,3})^2 = (d_1g_{1,1})^2 + (d_1g_{1,2})^2 + (d_2g_{1,3})^2.$$

Since first rows of both h and g have length 1, we have

$$\begin{aligned} \Rightarrow (d_1)^2 &= (d_1)^2[(h_{1,1})^2 + (h_{1,2})^2 + (h_{1,3})^2] \\ &= (d_1g_{1,1})^2 + (d_1g_{1,2})^2 + (d_2g_{1,3})^2 < (d_1)^2, \end{aligned}$$

since $d_1 > d_2$. But this is impossible unless $g_{1,3} = 0$, and hence $h_{1,3} = 0$. Comparing the lengths of the second rows shows that $g_{2,3} = h_{2,3} = 0$, and applying this same reasoning to the columns gives $h_{3,1} = g_{3,1} = 0$ and $h_{3,2} = g_{3,2} = 0$.

We now have

$$\begin{pmatrix} d_1h_{1,1} & d_1h_{1,2} & 0 \\ d_1h_{2,1} & d_1h_{2,2} & 0 \\ 0 & 0 & d_2h_{3,3} \end{pmatrix} = \begin{pmatrix} d_1g_{1,1} & d_1g_{1,2} & 0 \\ d_1g_{2,1} & d_1g_{2,2} & 0 \\ 0 & 0 & d_2g_{3,3} \end{pmatrix},$$

which immediately implies that $h = g$. The condition that $g^t g = I$ gives us that each of the block submatrices must be orthogonal, and of course g must have determinant 1. Note that if we were to allow $d_2 = 0$ then $g_{3,3}$ and $h_{3,3}$ need not be equal.

An inductive argument on the different eigenvalues of d proves the general case and is not particularly enlightening, so we state the following result.

Proposition 3.1. *Let $G = \mathrm{SO}(n) \times \mathrm{SO}(n)$ and let V be the vector space of $n \times n$ real matrices. Let G act on V via $(g, h).v = g^t v h$. Let*

$$d = \mathrm{diagonal}(\underbrace{d_1, \ldots, d_1}_{s_1}, \ldots, \underbrace{d_k, \ldots, d_k}_{s_k}) \in V$$

with $d_1 > d_2 > \ldots > d_k \geq 0$, and let G_d be the stabilizer of d in G. If $d_k > 0$, then $G_d = \{(g, g) : g \in S(O(s_1) \times \cdots \times O(s_k))\}$.

That is, each g consists of block-diagonal matrices where each block is an $s_i \times s_i$ orthogonal matrix and where s_i is the multiplicity of the eigenvalue d_i in d. The "S" indicates that the product of the determinants of the blocks is 1. If $d_k = 0$ then $G_d = (g, h)$ where g and h consist of block-diagonal matrices with each i-th block in $O(s_i)$, and where $g = h$ except for the k-th block.

4. Orbits

A natural question is "What are these orbits like?" From the introduction we know that, for any element $v \in V$, the orbit $G.v$ is diffeomorphic to the coset space G/G_v, with $\dim G.v = \dim G - \dim G_v$. Since any two elements in the same G-orbit have isomorphic stabilizers, it will be sufficient to consider the orbits of those representative elements d in the cross-section D. In particular, the dimension of these orbits is completely determined by the multiplicity of the distinct eigenvalues of d and is independent of their actual values.

Example: n = 2. In low-dimensional cases we can use computer graphics to get an idea about the nature of these orbits, and we now illustrate this for the two-dimensional Lie group $G = \mathrm{SO}(2) \times \mathrm{SO}(2)$. Figure 1 shows the orbit of $d = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, with a cut-away view on the right. Note that, for $n = 2$, the orbit lies in $\mathrm{Mat}(2, \mathbb{R}) \cong \mathbb{R}^4$, and each figure is a projection of this orbit onto $\mathbb{R}^2$. Since G is abelian, G_d is normal in G and so G/G_d is an abelian Lie group which is compact since the quotient map is continuous. Since $G_d = \{(\mathcal{I}_2, \mathcal{I}_2), (-\mathcal{I}_2, -\mathcal{I}_2)\}$ which is discrete, the orbit $G.d$ has dimension 2. We conclude that this orbit is diffeomorphic to the 2-torus embedded in $\mathbb{R}^4$, since this is the only two-dimensional compact abelian Lie group. Notice that the graphics could be misleading, since we usually picture the 2-torus in $\mathbb{R}^3$ as resembling the surface of a donut.

Note that if an element d in the cross-section D has only one eigenvalue, then the stabilizer G_d is isomorphic to SO(2) and so the orbit $G.d$ is one-dimensional and is diffeomorphic to SO(2), that is, a circle.

Generic orbits. We now move on to consider the following special case of *generic* orbits — those with maximal dimension — for arbitrary n. We will reserve the symbol δ for a diagonal matrix in the cross-section D with n distinct eigenvalues.

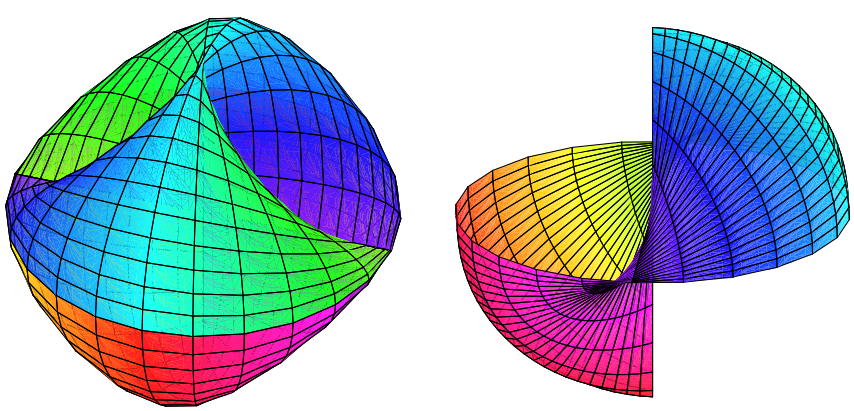


Figure 1. An orbit for $n = 2$ projected onto $\mathbb{R}^2$. Right: cut-away view of same orbit.

That is, $\delta = \text{diagonal}(d_1, \ldots, d_n)$ with $d_1 > d_2 > \cdots > d_n \geq 0$. From Proposition 3.1 we have $G_\delta = (g, g)$, where $g = \text{diagonal}(\pm 1, \ldots, \pm 1)$ has an even number of entries equal to -1. Since the stabilizer of δ is discrete, the dimension of the G-orbit of δ is equal to the dimension of G.

Proposition 4.1. *Let $G = \mathrm{SO}(n) \times \mathrm{SO}(n)$ and let V be the vector space of $n \times n$ real matrices. Let G act on V via $(g, h).v = g^t v h$. Let*

$$\delta = \text{diagonal}(d_1, d_2, \ldots, d_n) \in V$$

with $d_1 > d_2 > \cdots > d_n \geq 0$, and let G_δ be the stabilizer of δ in G. Then $|G_\delta|$, the order of G_δ, is 2^{n-1}.

Proof. From Proposition 3.1, G_δ consists of n copies of $O(1) = \pm 1$ lying in $\mathrm{SO}(n)$, so there must be an even number of entries equal to -1. Thus

$$|G_\delta| = \binom{n}{0} + \binom{n}{2} + \binom{n}{4} + \cdots + \binom{n}{k},$$

where $k = n$ if n is even and $k = n - 1$ if n is odd. From the binomial theorem,

$$\begin{aligned} 2^n &= (1+1)^n + (1-1)^n \\ &= \left[\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n}\right] + \left[\binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \cdots \pm \binom{n}{n}\right] \\ &= 2\left[\binom{n}{0} + \binom{n}{2} + \binom{n}{4} + \cdots + \binom{n}{k}\right] = 2\,|G_\delta|. \end{aligned}$$

□

Again, what are these orbits like? Figure 2 shows a (projection of a) two-dimensional slice of the orbit of $\delta = \text{diagonal}(2, 1, 0)$ for the case $n = 3$. Could this be just a torus in disguise, as was the case $n = 2$? One way to determine how interesting the orbits are is to consider their fundamental groups.

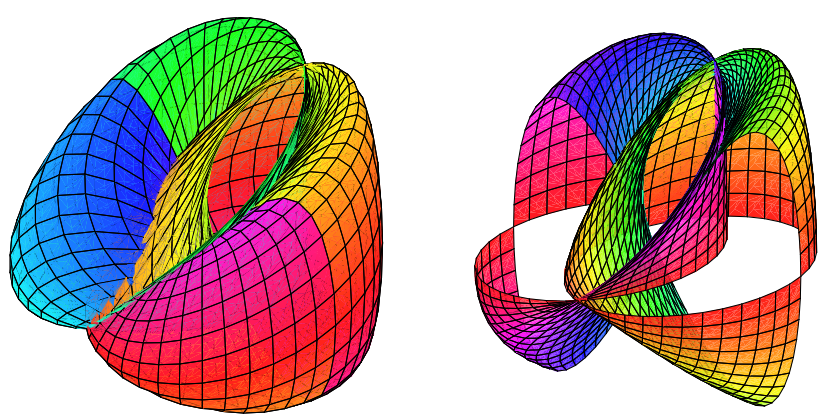


Figure 2. A section of an orbit for $n = 3$. Right: cut-away view of same section.

5. Fundamental groups, covering spaces and spin(n)

In order to make this exposition self-contained and to fix notation we review some background material that will be familiar to many readers.

Review of the fundamental group and covering spaces. Let X be a topological space and let $[0, 1] \subset \mathbb{R}$ be the closed unit interval. A *path in* X is a continuous map $f : [0, 1] \to X$. Two paths f and g from x_1 to x_2 are said to be *homotopic* if one can be continuously deformed into the other. This is obviously an equivalence relation, and we denote the equivalence class of f by $[f]$. Of special interest will be loops, or closed paths that start and end at a distinguished base point $x \in X$, and we can define a multiplication of loops by concatenation. That is, $f \cdot g$ means *first go around* f *and then go around* g. This operation is associative and is well defined when taking equivalence classes: $[f] \cdot [g] = [f \cdot g]$. The constant loop $e_x : [0, 1] \to X$ given by $e_x(t) = x$ serves as the identity element for this operation and the loop f^{-1} is the loop f traversed in the opposite direction. We can then define the *first homotopy group* or the *fundamental group*, denoted $\pi_1(X, x)$, as the group of (equivalence classes of) loops in X that start and end at x, along with this multiplication. If x_1 and x_2 are connected by a path in X, then $\pi_1(X, x_1)$ and $\pi_1(X, x_2)$ are isomorphic. Homeomorphic topological spaces have isomorphic fundamental groups, but the converse need not be true.

We will also require the notion of a covering. Let (X, x), (Y, y) be topological spaces with base points x and y respectively. A map $p : (Y, y) \to (X, x)$ is a *covering map* if

(i) $p(y) = x$;

(ii) p is continuous and surjective;

(iii) for every $x_0 \in X$ there is an open neighborhood $U_{x_0} \subset X$ so that $p^{-1}(U_{x_0})$ is a disjoint union of open sets $\{V_\alpha\}$ and so that for each α, the map p restricted to V_α is a homeomorphism of V_α onto U_{x_0}.

We then say that (Y, y) is a *covering space* of (X, x) and refer the the covering space along with the covering map as a *cover* of (X, x). We will also use the standard results, roughly stated, that the composition of covers is a cover, and that the cover of a product is the product of the respective covers.

Remark 5.1. A topological space with trivial fundamental group is called *simply connected*. A covering space that is simply connected is called a *universal covering space*. It is unique up to homeomorphism.

We will need the notion of *lifting* a path from a space to a covering space.

Let $p : (Y, y) \to (X, x)$ be a covering map. Let $f : [0, 1] \to X$ be a path starting at x. A lifting of f is a path $\tilde{f} : [0, 1] \to Y$ such that $p \circ \tilde{f} = f$. For the cases we

are considering, these lifts are unique up to homotopy. That is, let f be a path in X beginning at x, and let $\tilde{f}$ and $\tilde{g}$ be two lifts of f both beginning at y. Then $\tilde{f}$ is homotopic to $\tilde{g}$. In particular, $\tilde{f}$ and $\tilde{g}$ must end at the same point in Y.

Let $p : (Y, y) \to (X, x)$ be a covering map. A homeomorphism $h : Y \to Y$ is called a *deck transformation* or *covering transformation* if $p \circ h = p$. Clearly the collection of all such deck transformations is a group with the operation being composition of maps.

We will use the following fact to determine $\pi_1(G.\delta, \delta)$.

Theorem 5.2. [Massey 1991, Corollary 7.5] *If (Y, y) is a universal covering space of (X, x), the group of deck transformations of (Y, y) is isomorphic to $\pi_1(X, x)$. If $p : (Y, y) :\to (X, x)$ is a covering map, then the order of $\pi_1(X, x)$ is equal to the cardinality of the set $p^{-1}(x)$.*

Now consider the map $p_1 : G \to G.\delta$ given by $g \mapsto g.\delta$. Since $p_1^{-1}(\delta) = \{\gamma \in G \mid \gamma.\delta = \delta\} = G_\delta$ is discrete, Theorem E4 of [Hall 2003] has the following consequence.

Proposition 5.3. *Let $G = \mathrm{SO}(n) \times \mathrm{SO}(n)$ and let $\mathbf{1}$ denote the identity element in G. Let V be the vector space of $n \times n$ real matrices and let G act on V by*

$$(g, h).v := g^t v h, \quad (g, h) \in G, \quad v \in V.$$

If $\delta \in V$ is a diagonal matrix with n distinct eigenvalues, and if $G.\delta$ is the G-orbit of δ, then the map $p_1 : (G, \mathbf{1}) \to (G.\delta, \delta)$ given by $g \mapsto g.\delta$ is a covering map.

Said another way, G is a fiber bundle over the orbit $G.\delta$ with projection map $(g, h) \mapsto (g, h).\delta$ and discrete fiber G_δ.

Spin(n). We now provide a brief review of the construction of the Lie group $\mathrm{Spin}(n)$ and the covering map from $\mathrm{Spin}(n)$ to $\mathrm{SO}(n)$. This abridged description should be sufficient for our purposes, but for a more complete discussion, see [Bröcker and tom Dieck 1985]. The presentation below borrows extensively from the excellent exposition in [Simon 1996].

Consider the vector space $\mathbb{R}^n$ with standard basis $\{e_1, \dots, e_n\}$. We form $C(n)$, the *Clifford algebra* on $\mathbb{R}^n$, by declaring that multiplication is associative, distributive over addition, and obeys the relations $e_i e_j + e_j e_i = -2\delta_{ij}$. This is just a fancy way of saying that the basis elements anti-commute and $e_i^2 = -1$. If $I = i_1 i_2 \dots i_k$ is a multiindex with $1 \le i_1 < \cdots < i_k \le n$ we set $e_0 = 1$, we set $e_I = e_{i_1} e_{i_2} \dots e_{i_k}$ and we set $|I| = k$. Then $C(n)$ is an algebra with basis $\{e_I\}$ and it follows that the dimension of $C(n)$ is 2^n. We also have the subalgebra of even elements

$$C(n)_{\text{even}} = \{A \in C(n) \mid A \text{ is a linear combination of } e_I \text{ with } |I| \text{ even}\}.$$

Examples. We have canonical isomorphisms:

- $C(0) \cong \mathbb{R}$;
- $C(1) \cong \mathbb{C}$ via the map $e_1 \mapsto i = \sqrt{-1}$;
- $C(2) \cong \mathbb{H}$ (the quaternion algebra) via the map $e_1 \mapsto i$, $e_2 \mapsto j$ and so $e_1 e_2 \mapsto k$. Here, i, j, and k are those elements in $\mathbb{H}$ with $i^2 = j^2 = k^2 = -1$ and $ij = k$;
- we also have $C(3)_{\text{even}} \cong \mathbb{H}$ via the map $e_1 e_2 \mapsto i$, $e_1 e_3 \mapsto j$, so

$$(e_1 e_2)(e_1 e_3) = e_2 e_3 \mapsto k.$$

We can define $\mathrm{Spin}(n)$ to be the invertible elements S of $C(n)_{\text{even}}$ that (among other things) leave the vector space $W = \mathbb{R}^n$ invariant under conjugation:

$$SWS^{-1} \subseteq W.$$

Now consider the quadratic elements

$$q_{ij} = \tfrac{1}{2} e_i e_j,$$

for $1 \le i < j \le n$, and observe that they obey the same commutation relations as the generators L_{ij} of the Lie algebra $\mathfrak{so}(n)$. Therefore these quadratic elements form a Lie algebra isomorphic to $\mathfrak{so}(n)$, and so to get the group $\mathrm{Spin}(n)$ we exponentiate these quadratic elements:

$$\begin{aligned} S_{ij}(t) := \exp(t\, q_{ij}) &= 1 + (tq_{ij}) + \frac{1}{2!}(tq_{ij})^2 + \frac{1}{3!}(tq_{ij})^3 + \cdots \\ &= \cos(t/2) + \sin(t/2)(2q_{ij}), \end{aligned}$$

since $q_{ij}^2 = -1$. As t goes from 0 to 4π, $S_{ij}(t)$ gives a copy of $U(1)$ in $\mathrm{Spin}(n)$ which is homeomorphic to a circle in the plane spanned by 1 and $2q_{ij}$.

Now the elements A in $\mathrm{Spin}(n)$ act on $\mathbb{R}^n$ by conjugation and this gives a representation of $\mathrm{Spin}(n)$ on $\mathbb{R}^n$. Consequently, we have a map

$$R : \mathrm{Spin}(n) \to \mathrm{SO}(n, \mathbb{R})$$

defined by

$$Ae_i A^{-1} = \sum_{i=1}^{n} R_{ji}(A) e_j. \tag{5-1}$$

We now determine the matrix representation of the group elements

$$S_{ij}(t) := \exp(tq_{ij}) = \cos(t/2) + \sin(t/2)(e_i e_j) \tag{5-2}$$

by determining the action on the basis vectors. First observe that $e_i e_j$ commutes with e_k when k is equal to neither i nor j, so in this case

$$S_{ij}(t) e_k S_{ij}^{-1}(t) = \big(\cos(t/2) + \sin(t/2)(e_i e_j)\big) e_k \big(\cos(t/2) - \sin(t/2)(e_i e_j)\big) = e_k.$$

Now conjugating e_i by $S_{ij}(t)$ we have

$$\begin{aligned} S_{ij}(t)e_i S_{ij}^{-1}(t) &= \big(\cos(t/2)+\sin(t/2)(e_ie_j)\big)e_i\big(\cos(t/2)-\sin(t/2)(e_ie_j)\big) \\ &= \big(\cos(t/2)+\sin(t/2)(e_ie_j)]^2 e_i \\ &= \big(\cos^2(t/2)-\sin^2(t/2)\big)e_i - 2\cos(t/2)\sin(t/2)e_j \\ &= \cos(t)e_i - \sin(t)e_j. \end{aligned}$$

A similar computation applied to e_j gives

$$S_{ij}(t)\, e_j\, S_{ij}^{-1}(t) = \sin(t)e_i + \cos(t)e_j.$$

Therefore, conjugation by $S_{ij}(t) = \exp(tq_{ij})$ induces a rotation by an angle t in the e_i, e_j plane. Since these rotations generate $\mathrm{SO}(n)$, this map is surjective.

The following result is well known (see [Simon 1996, Sections VII.6–VII.7] or [Bröcker and tom Dieck 1985, Section 1.6].

Proposition 5.4. $\mathrm{Spin}(n)$ *is simply connected. If* $A \in \mathrm{Spin}(n)$ *and if* $R(A)$ *is the* $n \times n$ *matrix with entries* $R_{ji}(A)$ *described in* (5-1) *above, then the map* $R : (\mathrm{Spin}(n), \mathbf{1}) \to (\mathrm{SO}(n, \mathbb{R}), \mathbf{1})$ *is a twofold universal covering map and a homomorphism of Lie groups. The symbol* $\mathbf{1}$ *denotes the unit elements in the respective groups.*

6. The fundamental group of a generic orbit

We are now ready to determine the fundamental group for a generic orbit of maximum dimension. We will proceed by elaborating on some previously introduced ideas and connecting them together in order to invoke Theorem 5.2.

As before, $\delta \in D$ denotes an element in the cross-section with n distinct eigenvalues. By Proposition 3.1, a typical element in its stabilizer G_δ can be represented by a diagonal matrix with each entry equal to ± 1, and where an even number of entries are equal to -1. From now on, let $I = i_1 i_2 \cdots i_k$ be a multiindex with $1 \le i_1 < \cdots < i_k \le n$, k even and set $l = k/2$. Let ST_I be the element in G_δ with those entries that are equal to -1 indexed by I. For example, if $n = 6$,

$$ST_{1,2,3,5} = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Using this notation, $G_\delta = \{(ST_I, ST_I) : |I| \text{ is even}\}$.

Let $\tau = (t_1, \dots, t_l)$ and let $\mathrm{SO}_I(\tau)$ be the matrix consisting of rotations by an angle t_j in the planes indexed pairwise by I. These pairs are of the form i_{2m-1}, i_{2m}.

For example, if $I = 1, 2, 3, 5$ and $\tau = (t_1, t_2)$ then $\mathrm{SO}_I(\tau)$ rotates by an angle t_1 in the $1, 2$ plane and by an angle t_2 in the $3, 5$ plane. For instance, if $n = 6$,

$$\mathrm{SO}_{1,2,3,5}(\tau) = \begin{pmatrix} \cos t_1 & \sin t_1 & 0 & 0 & 0 & 0 \\ -\sin t_1 & \cos t_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \cos t_2 & 0 & \sin t_2 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -\sin t_2 & 0 & \cos t_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Notice that $\mathrm{SO}_{1,2,3,5}(\tau)$ is equal to the matrix product $\mathrm{SO}_{1,2}(t_1)\,\mathrm{SO}_{3,5}(t_2)$. It should be easy to see that

Lemma 6.1. $ST_I = \mathrm{SO}_I(\pm\pi, \dots, \pm\pi)$.

We next consider product of elements $S_{ij}(t) \in \mathrm{Spin}(n)$ and relate them to the corresponding elements in $\mathrm{SO}(n)$.

Lemma 6.2. *Let $I = i_1 i_2 \cdots i_k$ be a multiindex with k even and where*

$$i_1 < i_2 < \cdots < i_k.$$

Set $l = k/2$. Let $\tau = (t_1, \dots, t_l)$ and let $\mathrm{SO}_I(\tau)$ be the matrix consisting of rotations by an angle t_j in the planes indexed pairwise by I. Let $S_{i,j}(t)$ be defined as in (5-2), *and let $S_I(\tau)$ designate the product $S_I(\tau) = S_{i_1 i_2}(t_1) S_{i_3 i_4}(t_2) \cdots S_{i_{k-1} i_k}(t_l)$. Let $R : \mathrm{Spin}(n) \to \mathrm{SO}(n)$ be the covering map given by Proposition 5.4. Then $R(S_I(\tau)) = \mathrm{SO}_I(\tau)$.*

Further, $e_I := e_{i_1} e_{i_2} \cdots e_{i_k}$, we have $e_I = S_I(\pi, \dots, \pi)$.

Proof. Since the entries in the multiindex I are distinct, the designation $\mathrm{SO}_I(\tau) = \mathrm{SO}_{i_1 i_2 \cdots i_k}(\tau) = \mathrm{SO}_{i_1 i_2}(t_1)\,\mathrm{SO}_{i_3 i_4}(t_2) \cdots \mathrm{SO}_{i_{k-1} i_k}(t_l)$ is unambiguous. Since the map R is a representation, we have

$$\begin{aligned} R[S_I(\tau)] &= R[S_{i_1 i_2}(t_1)]\, R[S_{i_3 i_4}(t_2)] \cdots R[S_{i_{k-1} i_k}(t_l)] \\ &= \mathrm{SO}_{i_1 i_2}(t_1)\,\mathrm{SO}_{i_3 i_4}(t_2) \cdots \mathrm{SO}_{i_{k-1} i_k}(t_l) = \mathrm{SO}_I(\tau). \end{aligned}$$

For the last assertion, note that (5-2) gives $e_i e_j = S_{ij}(\pi)$ for any i, j, since $\cos(\pi/2) = 0$ and $\sin(\pi/2) = 1$. Hence

$$e_I = [e_{i_1} e_{i_2}][e_{i_3} e_{i_4}] \cdots [e_{i_{k-1}} e_{i_k}] = S_{i_1 i_2}(\pi) S_{i_3 i_4}(\pi) \cdots S_{i_{k-1} i_k}(\pi) = S_I(\pi, \dots, \pi),$$

as required. □

This next result is proven similarly.

Lemma 6.3. *Denote by π^+ an l-tuple $\pi^+ = (\pm\pi, \dots, \pm\pi)$ with an even number of entries equal to $-\pi$ and denote by π^- an l-tuple $\pi^- = (\pm\pi, \dots, \pm\pi)$ with an odd number of entries equal to $-\pi$. Let $S_I(\tau)$ and e_I be as in the previous lemma. Then $S_I(\pi^+) = e_I$ and $S_I(\pi^-) = -e_I$.*

Finally, let $\widetilde{\mathbf{1}}$ denote the unit element in $\widetilde{G} = \mathrm{Spin}(n) \times \mathrm{Spin}(n)$ and let $\mathbf{1}$ denote the unit element in $G = \mathrm{SO}(2, \mathbb{R}) \times \mathrm{SO}(2, \mathbb{R})$. Then $(\widetilde{G}, \widetilde{\mathbf{1}})$ is the universal covering space (in fact, a covering group) of $(G, \mathbf{1})$ and the map

$$\rho = R \times R : (\widetilde{G}, \widetilde{\mathbf{1}}) \to (G, \mathbf{1})$$

is a fourfold covering map. Now recall the covering map $p_1 : (G, \mathbf{1}) \to (G.\delta, \delta)$ from Proposition 5.3. It follows that the composition

$$P = \rho \circ p_1 : (\widetilde{G}, \widetilde{\mathbf{1}}) \to (G.\delta, \delta)$$

is a covering map and that $\widetilde{G}$ is the universal covering space of the orbit $G.\delta$.

Definition 6.4. $E(n) = \{\pm e_I : |I| \text{ is even}\}$.

Observe that $E(n)$ is closed under multiplication since, if $e_I e_J = e_K$ then $|K| = |I| + |J|$ when I and J are distinct indices, and the entries of K contract in pairs when I and J have repeated entries. For example, $e_{1,2}\, e_{2,3} = -e_{1,3}$. Since $(e_I)^{-1} = \pm e_I$, $E(n)$ is a group under multiplication. A computation very similar to that in Proposition 4.1 shows that $|E(n)| = 2^n$.

Definition 6.5. Consider the set $\widetilde{E(n)} = \{(\nu, \pm\nu) \mid \nu \in E(n)\}$. This is a subgroup of $\widetilde{G}$ which is isomorphic to the group $E(n) \times \mathbb{Z}_2$ via the identifications $(\nu, 1) \mapsto (\nu, \nu)$ and $(\nu, -1) \mapsto (\nu, -\nu)$ for $\nu \in E(n)$.

Proposition 6.6. $P^{-1}(\delta) = \widetilde{E(n)}$.

Proof.

$$\begin{aligned}
P[(e_I, e_I)] &= p_1 \circ [R(e_I), R(e_I)], \\
\text{Lemma 6.3} \Rightarrow &= p_1 \circ [R(S_I(\pi^+), R(S_I(\pi^+)], \\
\text{Lemma 6.2} \Rightarrow &= p_1 \circ [\mathrm{SO}_I(\pi^+), \mathrm{SO}_I(\pi^+)], \\
\text{Lemma 6.1} \Rightarrow &= p_1 \circ [ST_I, ST_I], \\
&= \delta.
\end{aligned}$$

The proofs of the other cases such as $P[(e_I, -e_I)] = \delta$ are similar and hence $\widetilde{E(n)} \subseteq P^{-1}(\delta)$.

Now $p_1^{-1}(\delta) = \{(ST_I, ST_I) : |I| \text{ is even}\} \subseteq G$ has order 2^{n-1} (Proposition 4.1) and ρ is a fourfold covering map $\widetilde{G} \to G$. Therefore the set $P^{-1}(\delta)$ has order 2^{n+1} which is equal to the order of $\widetilde{E(n)}$. $\square$

The main result of this paper completely describes the fundamental group of a generic orbit.

Theorem 6.7. *Let $G = \mathrm{SO}(n) \times \mathrm{SO}(n)$ and let V be the vector space of $n \times n$ real matrices. Let G act on V via $(g, h).v = g^t v h$. Let $\delta = \mathrm{diagonal}(d_1, d_2, \ldots, d_n) \in V$ with $d_1 > d_2 > \ldots > d_n \geq 0$, and let $G.\delta$ be the G-orbit of δ in V. Let $e_1, \ldots, e_n$ be the standard basis vectors in $\mathbb{R}^n$ and let $E(n) = \{\pm e_{i_1} \ldots e_{i_k} \mid k \text{ is even}\}$ be the group generated by the quadratic units $e_i e_j$, $i < j$ in the Clifford algebra on $\mathbb{R}^n$. Then the fundamental group $\pi_1(G.\delta, \delta)$ is isomorphic to $E(n) \times \mathbb{Z}_2$.*

Proof. We will show that the group of deck transformations $\mathrm{Aut}(\widetilde{G}, P)$ on the universal covering $(\widetilde{G}, \mathbf{1})$ is isomorphic to $\widetilde{E(n)}$ which is isomorphic to $E(n) \times \mathbb{Z}_2$.

For each $\widetilde{\omega} \in \widetilde{E(n)}$ and $\widetilde{s} \in \widetilde{G}$ define the left translation map $\mathcal{L}_{\widetilde{\omega}} : \widetilde{G} \to \widetilde{G}$ by $\mathcal{L}_{\widetilde{\omega}}(\widetilde{s}) = \widetilde{\omega}\widetilde{s}$, the operation on the right-hand side being multiplication in $\widetilde{G}$. It is a standard exercise that the set of all such translations $\mathbb{L} = \{\mathcal{L}_{\widetilde{\omega}} \mid \widetilde{\omega} \in \widetilde{E(n)}\}$ is a group that is isomorphic to $\widetilde{E(n)}$ via the map $\widetilde{\omega} \mapsto \mathcal{L}_{\widetilde{\omega}}$. Since $\widetilde{G}$ is a Lie group, each translation is continuous with a continuous inverse, hence a homeomorphism from $\widetilde{G}$ to $\widetilde{G}$. Furthermore, for each $\widetilde{\nu} \in \widetilde{E(n)}$, the composition $P \circ \mathcal{L}_{\widetilde{\omega}}(\widetilde{\nu}) = P(\widetilde{\omega}\widetilde{\nu}) = \delta$ so each $\mathcal{L}_{\widetilde{\omega}}$ is a deck transformation and therefore $\mathbb{L}$ is a subgroup of $\mathrm{Aut}(\widetilde{G}, P)$. But $\mathrm{Aut}(\widetilde{G}, P)$ has order 2^{n+1} by Theorem 5.2, and since both these groups have the same order, they must be equal. By Theorem 5.2 again we have $\pi_1(G.\delta, \delta) \cong \mathrm{Aut}(\widetilde{G}, P) = \mathbb{L} \cong \widetilde{E(n)} \cong E(n) \times \mathbb{Z}_2$. □

7. An illustration

We conclude with an example for $n = 6$ that further illustrates the previous constructions. The element

$$S_{3,5}(t) = \exp[(t/2)e_3e_5] = \cos(t/2) + \sin(t/2)e_3e_5$$

in Spin(6) defined in (5-2) is homeomorphic to a circle lying in the plane spanned by 1 and e_3e_5 in the Clifford algebra $C(6)$, and which projects onto the rotation $\mathrm{SO}_{3,5}(t)$ in SO(6) via the representation R. Consider the path $\widetilde{f} : [0, 4\pi] \to \widetilde{G}$ given by $t \mapsto (S_{35}(t), S_{35}(t))$.

Since $\widetilde{f}$ is homeomorphic to a circle and $\widetilde{G}$ is a simply connected covering group, $[\widetilde{f}]$ is trivial in $\pi_1(\widetilde{G}, \mathbf{1})$. Now as t goes from 0 to π, we get a path $\widetilde{f}_{[0,\pi]}$ from $(1, 1)$ to (e_3e_5, e_3e_5) in $\widetilde{G}$ that projects down via P to a loop $f : [0, \pi] \to G.\delta$ given by $f(t) = (\mathrm{SO}_{3,5}(t), \mathrm{SO}_{3,5}(t)).\delta$. By uniqueness of path lifting, f cannot be homotopic to the trivial loop since $\widetilde{f}_{[0,\pi]}$ is not trivial in $\widetilde{G}$. Similarly, as t goes from π to 2π, we get a path $\widetilde{f}_{[\pi,2\pi]}$ from (e_3e_5, e_3e_5) to $(-1, -1)$ in $\widetilde{G}$ that also projects down to the loop f in the orbit $G.\delta$. Not until t travels the entire distance $[0, 4\pi]$ do we obtain the product f^4 in $G.\delta$ that lifts to the (trivial) loop $\widetilde{f}$ in $\widetilde{G}$.

Thus, $[f]^4$ is trivial in $\pi_1(G.\delta, \delta)$. We chart here the information as the path $\widetilde{f}$ is projected onto G and then $G.\delta$ for the successive landmark values of t.

t	$\widetilde{f}(t))$	$\rho((S_{3,5}(t), S_{3,5}(t)))$	$P(S_{3,5}(t), S_{3,5}(t))$
0	$(1, 1)$	$(\mathcal{I}_6, \mathcal{I}_6)$	δ
π	(e_3e_5, e_3e_5)	$(ST_{3,5}, ST_{3,5})$	δ
2π	$(-1, -1)$	$(\mathcal{I}_6, \mathcal{I}_6)$	δ
3π	$(-e_3e_5, -e_3e_5)$	$(ST_{3,5}, ST_{3,5})$	δ
4π	$(1, 1)$	$(\mathcal{I}_n, \mathcal{I}_n)$	δ

As in the previous discussion regarding deck transformations in the proof of Theorem 6.7, we can translate the loop $\widetilde{f}$ via left multiplication by the element $(e_1e_2, e_1e_2) \in \widetilde{E(n)}$. This gives us the loop $\widetilde{g} : [0, 4\pi] \to \widetilde{G}$ given by $t \mapsto (\nu(t), \nu(t))$ where

$$\nu(t) = e_1e_2[\cos(t/2) + \sin((t/2))e_3e_5] = \cos(t/2)e_1e_2 + \sin(t/2)e_1e_2e_3e_5.$$

This is a loop starting at e_1e_2 which lies in the plane spanned by e_1e_2 and $e_1e_2e_3e_5$ in the Clifford algebra $C(6)$.

We check that

$$\nu^{-1}(t) = [-\cos(t/2)e_1e_2 + \sin(t/2)e_1e_2e_3e_5]$$

and that conjugating the basis vectors $e_i \in \mathbb{R}^6$ by $\nu(t)$ produces the map R which takes $\nu(t)$ to the rotation

$$R(\nu(t)) = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \cos t & 0 & \sin t & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -\sin t & 0 & \cos t & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \in \mathrm{SO}(6).$$

As above, the projection P maps $\widetilde{g}_{[0,\pi]}$ to the loop $g(t) = R(\nu(t)).\delta$ in the orbit $G.\delta$ and $[g]^4$ is trivial. Here is part of this information for the path $\widetilde{g}$:

t	$\widetilde{g}(t)$	$\rho(\widetilde{g}(t)))$	$P(\widetilde{g}(t)))$
0	(e_1e_2, e_1e_2)	$(ST_{1,2}, ST_{1,2})$	δ
π	$(e_1e_2e_3e_5, e_1e_2e_3e_5)$	$(ST_{1,2,3,5}, ST_{1,2,3,5})$	δ
2π	$(-e_1e_2, -e_1e_2)$	$(ST_{1,2}, ST_{1,2})$	δ
3π	$(-e_1e_2e_3e_5, -e_1e_2e_3e_5)$	$(ST_{1,2,3,5}, ST_{1,2,3,5})$	δ

By considering the loops in the orbit $G.\delta$ that lift to the path from

$$(1, 1) \to (e_1e_2, e_1e_2) \to (e_1e_2e_3e_5, e_1e_2e_3e_5)$$

in $\widetilde{G}$ we see that g and f cannot be homotopic, so $[g]$ and $[f]$ are distinct elements in $\pi_1(G.\delta, \delta)$.

8. Final remarks on the general case

Determining the first homotopy group for the orbits in the more general case, when the representative element d in the cross-section contains eigenvalues with multiplicities greater than 1, does not lend itself to such direct construction since the map $G \to G.d$ is not a covering map.

Acknowledgments

We would like to thank Professors C. Benson, G. Ratcliff and A. Smith for many helpful conversations.

References

[Bröcker and tom Dieck 1985] T. Bröcker and T. tom Dieck, *Representations of compact Lie groups*, Grad. Texts in Math. **98**, Springer, New York, 1985. MR 86i:22023 Zbl 0581.22009

[Hall 2003] B. C. Hall, *Lie groups, Lie algebras, and representations*, Grad. Texts in Math. **222**, Springer, New York, 2003. MR 2004i:22001 Zbl 1026.22001

[Massey 1991] W. S. Massey, *A basic course in algebraic topology*, Grad. Texts in Math. **127**, Springer, New York, 1991. MR 92c:55001 Zbl 0725.55001

[Simon 1996] B. Simon, *Representations of finite and compact groups*, Grad. Studies in Math. **10**, American Mathematical Society, Providence, RI, 1996. MR 97c:22001 Zbl 0840.22001

Received: 2008-04-08 Revised: Accepted: 2009-09-28

czarn005@rangers.uwp.edu *Department of Mathematics, University of Wisconsin – Parkside, 900 Wood Rd., P.O. Box 2000, Kenosha, WI 53141-2000, United States*

hower@uwec.edu *Department of Mathematics, University of Wisconsin – Eau Claire, 508 Hibbard Humanities Hall, Eau Claire, WI 54702-4004, United States* http://www.uwec.edu/math/Faculty/howe.htm

Aaron.D.McTavish@uwsp.edu *Department of Mathematical Sciences, University of Wisconsin – Stevens Point, Stevens Point, WI 54481-3897, United States*

INVOLVE 2:5(2009)

Symbolic computation of degree-three covariants for a binary form

Thomas R. Hagedorn and Glen M. Wilson

(Communicated by Scott Chapman)

We use elementary linear algebra to explicitly calculate a basis for, and the dimension of, the space of degree-three covariants for a binary form of arbitrary degree. We also give an explicit basis for the subspace of covariants complementary to the space of degree-three reducible covariants.

The study of invariant functions was one of the main influences on the development of modern algebra. Consider the following simple example. The group $G = \mathbb{Z}$ acts on $\mathbb{R}$ by addition: $g \cdot x = g + x$. We define a G-invariant function to be a real-valued function $f(x)$ on $\mathbb{R}$ such that $f \circ g = f$ for all $g \in G$. In other words, $f(x) = f(g + x)$ for all $g \in \mathbb{Z}$, $x \in \mathbb{R}$. The invariant functions are precisely the real-valued functions with period one. Hence, geometric information, such as periodicity, can be recovered by studying functions with certain algebraic properties.

In Section 1, we introduce the concepts of an invariant and covariant function for a binary form $Q(x, y)$. The problem of determining the complete set of these functions was widely worked on during the late nineteenth century. Gordan [1868] proved that the ring of invariants (and the ring of covariants) for a degree-n binary form is finitely generated. A milestone in the history of modern algebra was Hilbert's nonconstructive proof [1890] of the following fundamental theorem.

Theorem [Hilbert 1890]. *The ring of invariants* (*and the ring of covariants*) *for a degree-n homogeneous polynomial in m variables is finitely generated.*

Hilbert's theorem says that all invariants (resp. covariants) for a homogeneous polynomial can be expressed as polynomials in a certain finite set of invariants (resp. covariants). Hilbert [1893] subsequently gave a constructive proof of this theorem. The minimal size of the generating set is only known for a few values of (m, n). When $m = 2$, this number has been determined for $n \leq 8$ [Bedratyuk 2009; Bedratyuk and Bedratyuk 2008].

MSC2000: 13A50, 15A72, 16W22.
Keywords: theory of covariants, invariant theory, symbolic method, binary forms.

Let $\mathscr{C}^n_{d,h}$ denote the complex vector space of covariants of degree d, order h for a degree-n binary form (see Section 1 for a definition). Cayley and Sylvester proved a classical combinatorial formula for $\dim \mathscr{C}^n_{d,h}$ [Sturmfels 2008, p. 153]. Algorithms for calculating a basis for $\mathscr{C}^n_{d,h}$ are known, but in principle they have only been carried out in a few cases. In general, for a degree-n form in m variables, the most comprehensive treatment is due to Howe [1994], who has given an algorithm for calculating the set of invariants of degree $d \leq 6$.

Here we study the case when $d = 3$ and use an elementary argument involving matrix algebra to give an explicit basis for $\mathscr{C}^n_{3,h}$ in Theorem 6.1. While our result may not be new, we do not find it in the literature and it corrects the incorrect description of $\mathscr{C}^n_{3,h}$ in [Hilbert 1993, p. 62] (see the Historical remark in Section 6). As a corollary, we obtain an explicit form for the Cayley–Sylvester formula in this case. Finally, let $\mathrm{Red}^n_{3,h}$ denote the subspace of $\mathscr{C}^n_{3,h}$ consisting of reducible covariants (those that are polynomials in lower-degree covariants). In Corollary 6.4, we provide an explicit basis for the subspace in $\mathscr{C}^n_{3,h}$ complementary to $\mathrm{Red}^n_{3,h}$. Our argument is a variant of the classical straightening algorithm in invariant theory.

In the first two sections of this paper, we define the invariants and covariants of a binary form and review the classical symbolic method. There has been a wealth of excellent introductions to invariant theory recently written [Dolgachev 2003; Kraft and Weyman 1999; Olver 1999; Procesi 2007; Sturmfels 2008] and we refer the reader to them for a more comprehensive introduction to the subject. In the paper's next two sections, we introduce the combinatorial concepts of $\mathscr{H}$-and $\mathscr{U}$-matrices, and establish the relationship with $\mathscr{C}^n_{3,h}$. Finally, in Sections 5 and 6, we carry out calculations to determine an explicit basis for $\mathscr{C}^n_{3,h}$.

1. Basic notions

We review the basic definitions of invariants and covariants found, for example, in [Dolgachev 2003; Kraft and Weyman 1999; Olver 1999]. A binary form $Q(x, y)$ of degree n is a homogeneous polynomial

$$Q(x, y) = a_0 x^n + \binom{n}{1} a_1 x^{n-1} y + \ldots + \binom{n}{n-1} a_{n-1} x y^{n-1} + a_n y^n. \tag{1}$$

We let V_n denote the complex vector space of all binary forms of degree n with complex coefficients. The matrix group $\mathrm{SL}_2(\mathbb{C})$ acts on $v \in \mathbb{C}^2$ by matrix multiplication $g \cdot v = gv$ and induces an action on

$$(\mathbb{C}^2)^* = \{\text{Linear functions } h : \mathbb{C}^2 \to \mathbb{C}\}$$

by $(g \cdot h)(v) = h(g^{-1}v)$. In this context, we regard x, y as the coordinate functions on $\mathbb{C}^2$. Thus $x, y \in (\mathbb{C}^2)^*$ and $(\mathbb{C}^2)^* = \mathbb{C}x \oplus \mathbb{C}y$. If

$$g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{C}),$$

the explicit action of g on x, y is given by $g \cdot x = dx - by$, $g \cdot y = -cx + ay$. Defining $g \cdot (x^a y^b) = (g \cdot x)^a (g \cdot y)^b$, the $\mathrm{SL}_2(\mathbb{C})$-action naturally extends to a $\mathrm{SL}_2(\mathbb{C})$-action on V_n. Equivalently, the $\mathrm{SL}_2(\mathbb{C})$-action on $(\mathbb{C}^2)^*$ extends to the tensor product $\bigotimes_{i=1}^n (\mathbb{C}^2)^*$, and preserves the subspace $\mathrm{Sym}^n \mathbb{C}^2 \cong V_n$.

Remark 1.1. V_n is the unique (up to isomorphism) irreducible representation of $\mathrm{SL}_2(\mathbb{C})$ of dimension n.

Polynomial functions. Invariants and covariants for a binary form of degree n are specific examples of polynomial maps.

Definition 1.2. *Let $W = \bigoplus_{i=1}^k V_{n_i}$. A function $f : W \to \mathbb{C}$ is a* polynomial map *of degree d if there is a degree-d homogeneous polynomial $\hat{f} \in \mathbb{C}[x_{ij}]_{1 \le i \le k, 0 \le j \le n_i}$ such that for all binary forms $Q_i \in V_{n_i}$, expressed as $Q_i(x, y) = \sum_{j=0}^{n_i} a_{ij} x^{n-i} y^i$ as in* (1), *we have*

$$f(Q_1, \dots, Q_k) = \hat{f}(a_{ij}).$$

The polynomial $\hat{f}$ is uniquely determined and we identify f with $\hat{f}$. Let $P(W)_d$ denote the set of all degree-d polynomial maps on W. We say f has multidegree $\mathbf{d} = (d_1, \dots, d_k)$ if

$$f(t_1 Q_1, \dots, t_k Q_k) = t_1^{d_1} \dots t_k^{d_k} f(Q_1, \dots, Q_k) \quad \text{for all } t_i \in \mathbb{C},\ Q_i \in V_{n_i},$$

and we let $P(W)_{\mathbf{d}}$ denote the set of all such functions.

Example 1.3. Let $k = 1$, $n_1 = 1$, and $\hat{f}(x_0, x_1) = x_0 x_1$. Then f defined by $f(a_0 x + a_1 y) = \hat{f}(a_0, a_1) = a_0 a_1$ is a polynomial map of degree 2 on $W = V_1$. The function $f(a_0 x + a_1 y) = |a_0|$ is not a polynomial map.

More generally, consider a function $f : W \to V_h$. Since $\{x^h, x^{h-1}y, \dots, y^h\}$ is a basis for V_h, there are functions $f_i : W \to \mathbb{C}$ such that

$$f = f_0 x^h + f_1 x^{h-1} y + \dots + f_{h-1} x y^{h-1} + f_h y^h. \tag{2}$$

Definition 1.4. *A map $f : W \to V_h$ is a* polynomial map *of degree d if $f_i \in P(W)_d$ for each of the functions f_i in* (2). *We let $P(W)_{d,h}$ denote the set of all polynomial maps on $W \to V_h$ of degree d. An analogous definition applies if "degree d" is replaced by "multidegree $\mathbf{d}$".*

Invariants and covariants.

Definition 1.5. *A* $\mathrm{SL}_2(\mathbb{C})$-invariant *$f : V_n \to \mathbb{C}$ of degree d for a binary form of degree n is a polynomial map $f \in P(V_n)_d$ such that*

$$f(g \cdot Q) = f(Q) \qquad \text{for all } g \in \mathrm{SL}_2(\mathbb{C}) \text{ and } Q \in V_n.$$

Example 1.6. The first example of an invariant was discovered by Gauss in his study of binary quadratic forms. Let $Q(x, y) = a_0x^2 + 2a_1xy + a_2y^2$ and define the discriminant $\Delta(Q) = a_1^2 - a_0a_2$. If $g = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, we have $(g \cdot Q) = b_0x^2 + 2b_1xy + b_2y^2$, where

$$\begin{aligned} b_0 &= a_0d^2 - 2a_1cd + a_2c^2, \\ b_1 &= -a_0bd + a_1(ad + bc) - a_2ac, \\ b_2 &= a_0b^2 - 2a_1ab + a_2a^2. \end{aligned}$$

Δ is a degree-two $\mathrm{SL}_2(\mathbb{C})$-invariant of a binary quadratic form, as a straightforward calculation shows $\Delta(g \cdot Q) = \Delta(Q)$. It can be shown that if f is a degree-d invariant of a binary quadratic form, then d is even and f is a multiple of $\Delta^{d/2}$.

Classically, the interest in invariants was to use them to identify geometric properties of projective curves preserved under $\mathrm{SL}_2(\mathbb{C})$ transformations. However, invariants are not general enough to specify all such properties. The more general notion of covariants is needed to identify these properties.

Definition 1.7. *A* $\mathrm{SL}_2(\mathbb{C})$-covariant f *of degree d and order h for a form of degree n is a polynomial function $f \in P(V_n)_{d,h}$ such that*

$$f(g \cdot Q) = g \cdot f(Q) \qquad \textit{for all } g \in \mathrm{SL}_2(\mathbb{C}),\ Q \in V_n.$$

Let $\mathscr{C}(V_n)$ denote the vector space of all covariants for a form of degree n, and let $\mathscr{C}^n_{d,h} = \mathscr{C}(V_n)_{d,h}$ denote the space of those of degree d and order h.

Example 1.8. (i) An invariant is a covariant of order $h = 0$.

(ii) The simplest example of a covariant is the function $f : V_n \to V_n$ of degree 1, order n, defined by $f(Q) = Q$. $f \in \mathscr{C}^n_{1,n}$ is a covariant as the condition $f(g \cdot Q) = g \cdot f(Q)$ is trivially satisfied.

(iii) A more important covariant is the Hessian function. For $Q(x, y) \in V_n$, define the Hessian $H : V_n \to V_{2n-4}$ by

$$H(Q) = \frac{\partial^2 Q}{\partial x^2}\frac{\partial^2 Q}{\partial y^2} - \left(\frac{\partial^2 Q}{\partial x \partial y}\right)^2.$$

The Hessian has the property that $H(Q) = 0$ precisely when Q is the nth power of a linear form. It is a covariant of degree 2 and order $2n - 4$.

More generally, we have the following definition of a covariant. This definition is only used in the next section.

Definition 1.9. *Let $W = \bigoplus_{i=1}^k V_{n_i}$. A covariant of degree d, order h for W is a function $f \in P(W)_{d,h}$ satisfying*

$$f(g \cdot Q) = g \cdot f(Q) \qquad \textit{for all } g \in \mathrm{SL}_2(\mathbb{C}) \textit{ and } Q \in W.$$

We let $\mathscr{C}(W)$ denote the set of all covariants for W. A covariant $f \in \mathscr{C}(W)$ of order h is said to have multidegree $\mathbf{d}$ if $f \in P(W)_{\mathbf{d},h}$. We let $\mathscr{C}(W)_{d,h}$ (resp. $\mathscr{C}(W)_{\mathbf{d},h}$) denote the sets of covariants in $\mathscr{C}$ of order h and degree d (resp. multidegree $\mathbf{d}$).

2. The symbolic method

To prove our results, we will use the classical symbolic method. We review it here. Classically, the legitimacy of the symbolic method was questioned, but it has since been justified [Kung and Rota 1984; Dolgachev 2003]. We now introduce the symbolic method, following the presentation in [Kraft and Weyman 1999] and adopting their notations.

Symbolic method for $\mathscr{C}(L^J)$. Let $J = \{1, \dots, k\}$. Let x_{ab}, x_c for $a, b, c \in J$, $a \neq b$, denote independent variables and define the polynomial ring

$$\mathrm{Sym}_k = \mathbb{C}[x_{ab}, x_c | a, b, c \in J, a \neq b].$$

Let $P \in \mathrm{Sym}_k$ be a monomial and write

$$P = \prod_{a,b \in J} x_{ab}^{\lambda_{ab}} \prod_{c \in J} x_c^{\sigma_c}.$$

The *order* $\operatorname{ord} P$ and *weight* $\operatorname{wt} P = (\operatorname{wt}_a P)_{a \in J}$ are defined by

$$\operatorname{ord} P = \sum_{c \in J} \sigma_c, \qquad \operatorname{wt}_a P = \sum_{b \in J} (\lambda_{ab} + \lambda_{ba}) + \sigma_a.$$

We note that the symmetric group S_k naturally acts on Sym_k by

$$\sigma(x_{ab}) = x_{\sigma(a)\sigma(b)}, \qquad \sigma(x_c) = x_{\sigma(c)}.$$

Definition 2.1. *Let $l_1, l_2 \in L = V_1$ and let $l_i = a_{i0}x + a_{i1}y$. Define*

$$[l_1\, l_2] = \begin{vmatrix} a_{10} & a_{11} \\ a_{20} & a_{21} \end{vmatrix}.$$

Let $L^J = \bigoplus_{a \in J} L$ and let $l = (l_a) \in L^J$. L^J can be identified with L^k, where $k = |J|$. We denote the maps $l \mapsto [l_a\, l_b]$ and $l \mapsto l_c$ by the classical notation (ab) and c_x, respectively. We have $(ab), c_x \in \mathscr{C}(L^J)$.

Theorem 2.2 (First fundamental theorem). *The ring $C(L^J)$ is generated by all elements of the forms (ab) and c_x, for $a, b, c \in J$, and $a \neq b$.*

Define the map $\chi : \mathrm{Sym}_k \to \mathscr{C}(L^J)$ by

$$\chi(x_{ab}) = (ab), \qquad \chi(x_c) = c_x,$$

and extend it in the natural way to monomials and all of Sym_k.

Corollary 2.3. *$\chi : \mathrm{Sym}_k \to \mathscr{C}(L^J)$ is a surjective homomorphism of algebras. χ sends a monomial P of order h and weight w to a covariant $\chi(P)$ of order h and multidegree w.*

Let $\mathfrak{a} = \ker \chi$. Then

$$\mathrm{Sym}_k/\mathfrak{a} \cong \mathscr{C}(L^J). \tag{3}$$

We now describe the elements of $\mathfrak{a}$. For all distinct $a, b, c, d \in J$, we have

$$(ab) + (ba) = 0, \quad (ab)c_x + (ca)b_x + (bc)a_x = 0,$$
$$(ab)(cd) + (ad)(bc) - (ac)(bd) = 0.$$

Traditionally, these equations are called syzygies of the first, second, and third type. These syzygies motivate the definition of three subideals of $\mathfrak{a}$.

$$\mathfrak{a}_1 = \langle x_{ab} + x_{ba} \mid a \neq b \in J \rangle,$$
$$\mathfrak{a}_2 = \langle x_{ab}x_c + x_{ca}x_b + x_{bc}x_a \mid \text{distinct } a, b, c \in J \rangle,$$
$$\mathfrak{a}_3 = \langle x_{ab}x_{cd} + x_{ad}x_{bc} - x_{ac}x_{bd} \mid \text{distinct } a, b, c, d \in J \rangle.$$

The second fundamental theorem (or invariant theorem) for SL_2 states that these three syzygies generate all the relationships among the covariants $\mathscr{C}(L^J)$.

Theorem 2.4 (Second fundamental theorem). *Let $\mathfrak{a} = \ker \chi$. Then $\mathfrak{a} = \mathfrak{a}_1 + \mathfrak{a}_2 + \mathfrak{a}_3$.*

Notation. (i) Sym_k is bigraded by weight and order. Let $\mathrm{Sym}_{k,w,h} \subset \mathrm{Sym}_k$ be those elements with weight w and order h. Then $\mathrm{Sym}_k = \bigoplus_{w,h} \mathrm{Sym}_{k,w,h}$. The ideal $\mathfrak{a}$ is also bigraded and we let $\mathfrak{a}_{w,h} = \mathfrak{a} \cap \mathrm{Sym}_{k,w,h}$.

(ii) When the weight is $w = (n, \ldots, n)$, we write $w = (n)^k$ as shorthand.

Classical symbolic description of $\mathscr{C}^n_{k,h}$. Equation (3) describes $\mathscr{C}(L^J)$ in terms of the symbolic algebra. Classically, these same symbols were also used to denote covariants in $\mathscr{C}^n_{k,h}$ $(= \mathscr{C}(V_n)_{k,h})$. We now present this alternate symbolic method and relate the two notations.

Let $Q(x, y)$ be a degree n binary form with coefficients a_i as in (1). For each $a \in J$, let α_{a0}, α_{a1} be indeterminates with the property that $\alpha_{a0}^{n-i}\alpha_{a1}^i = a_i$, for $i = 0, \ldots, n$. Let $a_x = \alpha_{a0}x + \alpha_{a1}y$ and define

$$(ab) = \begin{vmatrix} \alpha_{a0} & \alpha_{a1} \\ \alpha_{b0} & \alpha_{b1} \end{vmatrix} = \alpha_{a0}\alpha_{b1} - \alpha_{a1}\alpha_{b0}.$$

Using these definitions, for a monomial $P = \prod_{a \neq b} x_{ab}^{\lambda_{ab}} \prod_c x_c^{\sigma_c} \in \mathrm{Sym}_{k,(n)^k,h}$, the expression

$$\psi(P) = \prod_{a \neq b} (ab)^{\lambda_{ab}} \prod_c c_x^{\sigma_c} \tag{4}$$

is a well-defined degree-h binary form whose coefficients are homogeneous degree k polynomials in the a_i. Moreover, the function $Q \mapsto \psi(P)$ can be seen to be a covariant in $\mathscr{C}^n_{k,h}$. Let $\psi(P)$ denote this covariant. We note that $\chi(P) \in \mathscr{C}(L^J)$ and $\psi(P) \in \mathscr{C}^n_{k,h}$ are different types of covariants, but the symbolic representation of $\chi(P)$ from the previous subsection equals the symbolic representation $\psi(P)$.

Example 2.5. Let $P = x^2_{ab} \in \mathrm{Sym}_{2,(2)^2,0}$. In the notation just introduced, the symbol $(ab)^2$ for the binary quadratic form Q of (1) represents

$$(ab)^2 = (\alpha_{a0}\alpha_{b1} - \alpha_{a1}\alpha_{b0})^2 = \alpha^2_{a0}\alpha^2_{b1} - 2\alpha_{a0}\alpha_{a1}\alpha_{b0}\alpha_{b1} + \alpha^2_{a1}\alpha^2_{b0} = 2(a_0a_2 - a_1^2).$$

Thus $\psi(P)$ is the covariant $Q \mapsto 2(a_0a_2 - a_1^2)$. Its symbolic representation $(ab)^2$ is the same as that of $\chi(P) = (ab)^2 \in \mathscr{C}(L^2)$.

The two symbolic methods are linked by the following proposition.

Proposition 2.6 [Kraft and Weyman 1999]. *There is a surjective homomorphism of vector spaces*

$$\Lambda : \mathscr{C}(L^J)_{(n)^k,h} \to \mathscr{C}^n_{k,h}$$

such that the composition $\Lambda \circ \chi : \mathrm{Sym}_{k,(n)^k,h} \to \mathscr{C}^n_{k,h}$ *is surjective with kernel*

$$I = \mathfrak{a}_{(n)^k,h} + \langle P - \sigma \cdot P : P \in \mathrm{Sym}_{k,(n)^k,h} \rangle.$$

If $P \in \mathrm{Sym}_{k,(n)^k,h}$ *is a monomial, then* $\Lambda(\chi(P)) = \psi(P)$. *Moreover,* Λ *sends a symbolic covariant in* $\mathscr{C}(L^J)$ *to the covariant in* $\mathscr{C}^n_{k,h}$ *with the same symbolic representation.*

Corollary 2.7. *The map* $P \mapsto \psi(P)$ *induces an isomorphism* $\mathrm{Sym}_{k,(n)^k,h}/I \cong \mathscr{C}^n_{k,h}$.

This result enables the easy classification of covariants with small degree.

Example 2.8 (Covariants of degree 1). Let $J = \{a\}$. By the corollary, $\mathscr{C}^n_{1,h} = 0$ when $h \neq n$. When $h = n$, $\mathscr{C}^n_{1,n}$ is generated by $\psi(x^n_a) = a^n_x$. As $I = \mathfrak{a}_{(n)^1,n} = 0$, $\mathscr{C}^n_{1,n}$ is the one-dimensional space generated by $g = a^n_x$. g is the trivial covariant with the property $g(Q) = Q$ for any degree n binary form Q.

Example 2.9 (Covariants of degree 2). Let $J = \{a, b\}$. We will show that

$$\dim \mathscr{C}^n_{2,h} = \begin{cases} 1 & \text{if } h \text{ is even, } h \leq 2n, \text{ and } h \equiv 2n \mod 4 \\ 0 & \text{otherwise.} \end{cases}$$

In the former case, $\mathscr{C}^n_{2,h}$ is generated by $g = (ab)^{n-h/2} a_x^{h/2} b_x^{h/2}$. When $h = 2n$, this is the covariant g such that $g(Q) = Q^2$. By the corollary, the vector space $\mathscr{C}^n_{2,h}$ is generated by the images $\psi(m)$ of the monomials $m = x^a_{12} x^b_{21} x^c_1 x^{h-c}_2$. Since $x^b_{21} - (-1)^b x^b_{12} \in \mathfrak{a}$, we only need consider $m = x^a_{12} x^c_1 x^{h-c}_2$. Since m has weight $(n)^2$, $a + c = n = a + (h - c)$ and $h = 2c$. Thus if h is odd, $\mathscr{C}^n_{2,h} = \{0\}$. If h is even, let $h = 2c$. Then $\mathscr{C}^n_{2,h}$ is at most a one-dimensional space generated by the

image of $m = x_{12}^{n-c} x_1^c x_2^c$. For this space to be nontrivial, we must also have $c \le n$. If $\sigma = (12)$, $m - \sigma(m) = (x_{12}^{n-c} - x_{21}^{n-c}) x_1^c x_2^c \in I$. Substituting for x_{21},

$$(1 - (-1)^{n-c}) x_{12}^{n-c} x_1^c x_2^c \in I.$$

Hence, when $n-c$ is odd, $m \in I$ and $\psi(m) = 0$. Finally, we show that $m \notin I$ when $h = 2c$, $h \le 2n$ and $n - c$ is even. Suppose that $m \in I$. Then the definition of I shows that

$$x_{12}^{n-c} x_1^c x_2^c = (x_{12} + x_{21}) f + [P - \sigma(P)], \tag{5}$$

where $f \in \mathrm{Sym}_2$, $\sigma = (12)$, and we can assume P is a linear combination of monomials of the form $x_{12}^i x_{21}^{n-c-i} x_1^c x_2^c$. Then

$$P - \sigma(P) = x_1^c x_2^c \sum_i a_i (x_{12} x_{21})^i \left[x_{21}^{n-c-2i} - x_{12}^{n-c-2i} \right].$$

Now (5) is an identity in Sym_2, so letting $x_{21} = -x_{12}$, we obtain an identity in $\mathbb{C}[x_1, x_2, x_{12}]$. But since $n - c$ is even, we obtain $x_{12}^{n-c} x_1^c x_2^c = 0$, in Sym_2, which gives a contradiction. Hence $m \notin I$ and $\mathscr{C}_{2,h}^n$ is one-dimensional in this case.

We now define the map Observe that the proof of Theorem 6.1 uses only the statements of Proposition 2.6 and Corollary 2.7. Consider $f \in \mathscr{C}(L^J)_{(n)^k,h}$. By Theorem 2.2, $f = \sum_P c_P P$, where

$$P = \prod_{a \ne b} (ab)^{\lambda_{ab}} \prod_{c \in J} c_x^{\sigma_c},$$

and $\mathrm{wt}\, P = n$, $\mathrm{ord}\, P = h$. We will define $\Lambda(P)$ for each monomial P. Then we can extend the domain of Λ to all of $\mathscr{C}(L^J)_{(n)^k,h}$ by defining

$$\Lambda(f) = \sum_P c_P \Lambda(P).$$

It remains to define $\Lambda(P) \in \mathscr{C}_{k,h}^n$. We will do so by defining $\Lambda(P)(g) \in V_h$, for $g \in V_n$. Among the many possibilities, fix a choice of integers $\lambda_{a_i b_j}, \sigma_{c_l} \in \{0, 1\}$, for $a, b, c \in J$, $i, j, l \in \{1, \dots, n\}$ such that

$$\sum_i \lambda_{a_i b_j} + \sigma_{b_j} = 1, \quad \sum_j \lambda_{a_i b_j} + \sigma_{a_i} = 1, \quad \sum_{i,j} \lambda_{a_i b_j} = \lambda_{ab}, \quad \sum_l \sigma_{c_l} = \sigma_c. \tag{6}$$

Since g is a complex polynomial of degree n in two variables, it factors as $g = g_1 \dots g_n$. Define

$$\Lambda(P)(g) = \frac{1}{(n!)^k} \sum_{\substack{(\tau_a)_{a \in J} \\ \tau_a \in S_n}} \left(\prod_{\substack{a \ne b \in J \\ 1 \le i, j \le n}} [g_{\tau_a(i)}\, g_{\tau_b(j)}]^{\lambda_{a_i b_j}} \prod_{\substack{c \in J \\ 1 \le l \le n}} g_{\tau_c(l)}^{\sigma_{c_l}} \right).$$

Because of the summation over $\tau_a \in S_n$, $\Lambda(P)(g)$ is independent of the choices of $\lambda_{a_i b_j}$, σ_{c_l}, and g_i. It is clear that $\Lambda(P) \in P(V_n)_{k,h}$. More work shows $\Lambda(P) \in \mathscr{C}^n_{k,h}$.

Example 2.10. We illustrate the map Λ. Consider $P = (ab)^2 \in \mathscr{C}(L^2)$. We have $k = 2$ and $\operatorname{ord} P = 0$. Then $\lambda_{ab} = 2$ is the only nonzero exponent among the λ_{cd}, σ_c. Choose $\lambda_{a_1 b_1} = \lambda_{a_2 b_2} = 1$ as the only nonzero exponents in (6). Let $g = a_0 x^2 + 2a_1 xy + a_2 y^2 = a_0(x - \alpha_1 y)(x - \alpha_2 y)$ and let $g_1 = a_0(x - \alpha_1 y)$, $g_2 = x - \alpha_2 y$. Then

$$\begin{aligned}\Lambda(P)(g) &= \tfrac{1}{4}\left([g_1\, g_1][g_2\, g_2] + [g_1\, g_2][g_2\, g_1] + [g_2\, g_1][g_1\, g_2] + [g_2\, g_2][g_1\, g_1]\right)\\ &= \tfrac{1}{4}\left([g_1\, g_2][g_2\, g_1] + [g_2\, g_1][g_1\, g_2]\right) = -\tfrac{1}{2}a_0^2(\alpha_1 - \alpha_2)^2\\ &= 2(a_0 a_2 - a_1^2),\end{aligned}$$

and $\Lambda(P) \in \mathscr{C}^2_{2,0}$. This calculation also shows that $\Lambda(\chi(x_{ab}^2)) = \psi(x_{ab}^2)$ by Example 2.5, which illustrates the second half of Proposition 2.6.

3. $\mathscr{H}$-matrices and $\mathscr{U}$-matrices

In the previous section, we used Corollary 2.7 to classify the covariants of degrees one and two by computing $\mathscr{C}^n_{k,h} \cong \mathrm{Sym}_{k,(n)^k,h}/I$. For larger degrees, the combinatorics in analyzing I are more difficult. In this section, we introduce $\mathscr{H}$-matrices and $\mathscr{U}$-matrices to simplify the analysis and then use them in the next section to classify the covariants of degree 3. Our goal is to define easily computed maps

$$\mathrm{Sym}_{3,(n)^3,h} \xrightarrow{\theta} \mathbf{H}^n_{3,h} \xrightarrow{\Delta} \mathbf{U}^n_{3,h},$$

with $\ker(\Delta \circ \theta) = I$. Then we will be able to explicitly compute $\mathscr{C}^n_{3,h}$.

$\mathscr{H}$-matrices. To a monomial $P = \prod_{a \neq b} x_{ab}^{\lambda_{ab}} \prod_{c \in J} x_c^{\sigma_c}$ in Sym_k we associate a $k \times k$ integral matrix $\theta(P) = (\theta_{ij})$ by

$$\theta_{ij} = \begin{cases} \lambda_{ij} & \text{if } i \neq j,\\ \sigma_i & \text{if } i = j.\end{cases}$$

When $P \in \mathrm{Sym}_{k,(n)^k,h}$, $\theta(P)$ will be an $\mathscr{H}$-matrix of type (n, k, h).

Definition 3.1. *Fix integers $k, n > 0$, $h \geq 0$. A $k \times k$-matrix $B = (b_{ij})$ is an $\mathscr{H}$-matrix of type (n, k, h) if the coefficients b_{ij} are nonnegative integers satisfying $\sum_{j=1}^k b_{jj} = h$ and for each i,*

$$\sum_{j=1}^k b_{ij} + \sum_{j=1}^k b_{ji} - b_{ii} = n.$$

We define $\mathscr{H}^n_{k,h}$ to be the set of all matrices B of type (n, k, h).

Example 3.2. Let $P = (ab)(bc)a_x c_x \in \mathrm{Sym}_3$. Then $\theta(P) = \left(\begin{smallmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{smallmatrix}\right) \in \mathcal{H}^2_{3,2}$.

Definition 3.3. *Let $\mathbf{H}^n_{k,h}$ be the complex vector space generated by the basis elements $[H]$, for $H \in \mathcal{H}^n_{k,h}$.*

If $H_1, H_2 \in \mathcal{H}^n_{k,h}$, then $2[H_1] - 3[H_2] \in \mathbf{H}^n_{k,h}$. We note that $2[H_1]$ is not the matrix obtained by multiplying the entries of H_1 by 2. We can extend the map θ to a map $\theta : \mathrm{Sym}_{k,(n)^k,h} \to \mathbf{H}^n_{k,h}$ by

$$\theta\Big(\sum_P c_P P\Big) = \sum_P c_P[\theta(P)],$$

where the sum is over monomials P.

The symmetric group S_k has a natural action on $\mathcal{H}^n_{k,h}$, and thus $\mathbf{H}^n_{k,h}$, defined by $\sigma \cdot A = (a_{\sigma^{-1}(i)\sigma^{-1}(j)})$, for $A \in \mathcal{H}^n_{k,h}$. It follows formally that

Proposition 3.4. *The map $\theta : \mathrm{Sym}_{k,(n)^k,h} \to \mathbf{H}^n_{k,h}$ is an S_k-equivariant isomorphism of $\mathbb{C}$-vector spaces.*

$\mathcal{U}$-matrices. The following subset of $\mathcal{H}$-matrices will be very useful.

Definition 3.5 ($\mathcal{U}$-matrices). *A $\mathcal{U}$-matrix is an upper-triangular $\mathcal{H}$-matrix (b_{ij}) whose diagonal elements form a nonincreasing sequence $b_{11} \geq b_{22} \geq \dots$. Let $\mathcal{U}^n_{k,h}$ be the set of all $\mathcal{U}$-matrices in $\mathcal{H}^n_{k,h}$. Let $\mathbf{U}^n_{k,h}$ be the subspace generated by formal complex linear combinations of $\mathcal{U}^n_{k,h}$-matrices.*

Example 3.6. $\left(\begin{smallmatrix} 3 & 2 & 3 \\ 0 & 2 & 4 \\ 0 & 0 & 1 \end{smallmatrix}\right)$ is a $\mathcal{U}$-matrix but $\left(\begin{smallmatrix} 2 & 1 & 3 \\ 1 & 3 & 3 \\ 1 & 0 & 1 \end{smallmatrix}\right)$ is not.

The $\mathcal{U}^n_{3,h}$-matrices can be easily parametrized. When $n-h$ is odd, $\mathcal{U}^n_{3,h} = \varnothing$. When $n \equiv h \mod 2$, we define

$$\mathcal{M}_{s,r} = \mathcal{M}_{s,r,h,n} = \begin{pmatrix} s & r+(n-h)/2 & h-s-r+(n-h)/2 \\ 0 & h-s-r & s+(n-h)/2 \\ 0 & 0 & r \end{pmatrix}, \tag{7}$$

for integers r, s. We usually drop the h, n indices as they are clear from the context. We also define

$$S_{n,h} = \big\{(s,r) \in \mathbb{Z}^2 \mid \max\big(0, \tfrac{1}{2}(h-n)\big) \leq r \leq \tfrac{1}{3}h,\ \tfrac{1}{2}(h-r) \leq s \leq h-2r\big\}. \tag{8}$$

Lemma 3.7. *Let n, h be nonnegative integers. If $n \equiv h \mod 2$, then $\mathcal{U}^n_{3,h} = \{\mathcal{M}_{s,r} \mid (s,r) \in S_{n,h}\}$. Otherwise $\mathcal{U}^n_{3,h} = \varnothing$.*

Proof. Assume $n \equiv h \mod 2$. If $(s,r) \in S_{n,h}$, then $s \geq h-s-r \geq r$ and $r + \frac{1}{2}(n-h) \geq 0$. Hence $\mathcal{M}_{s,r} \in \mathcal{U}^n_{3,h}$ for all $(s,r) \in S_{n,h}$. Now suppose $M = (b_{ij}) \in \mathcal{U}^n_{3,h}$. Let $s = b_{11}$, $r = b_{33}$. The order condition then shows $b_{22} = h - r - s$. Now $s + b_{12} + b_{13} = n$, $r + b_{13} + b_{23} = n$, and $h - r - s + b_{12} + b_{23} = n$ since $M \in \mathcal{U}^n_{3,h}$.

Solving these equations gives the formulas for b_{12}, b_{13}, b_{23} found in $\mathcal{M}_{s,r}$. Hence, $M = \mathcal{M}_{s,r} \in \mathcal{U}^n_{s,r}$, for some $(s, r) \in S_{n,h}$. The bounds follow from $s \geq h-s-r \geq r \geq 0$ and $b_{12} \geq 0$, proving the first assertion. When $n \not\equiv h \bmod 2$, the formulas for b_{12}, b_{21} show that these terms cannot be integers if $r, s \in \mathbb{Z}$. Hence $\mathcal{U}^n_{3,h} = \varnothing$. $\square$

Remark 3.8. If $n < h/3$, then $\mathcal{U}^n_{3,h} = \varnothing$.

4. The map Δ

We,will now construct the maps θ, Δ described at the beginning of Section 3. $\Delta : \mathcal{H}^n_{3,h} \to \mathbf{U}^n_{3,h}$ will be a S_3-invariant map and it will be used to construct a S_3-equivariant map $(\Delta \circ \theta) : \mathrm{Sym}_{3,(n)^3,h} \to \mathbf{U}^n_{3,h}$. Recalling that $I \subset \mathrm{Sym}_{3,(n)^3,h}$ is the kernel of $\Lambda \circ \chi$, we can then compute the image of I in $\mathbf{U}^n_{3,h}$ under $\Delta \circ \theta$. In Section 6, we will be able to use this result to explicitly classify the covariants in $\mathcal{C}^n_{3,h}$ using Corollary 2.7. We begin with the following lemma that follows from the calculation of the S_3-orbits in $\mathcal{H}^n_{3,h}$.

Lemma 4.1. *Let $M \in \mathcal{H}^n_{3,h}$. Among the matrices $A = (a_{ij})$ in the S_3-orbit of M in $\mathcal{H}^n_{3,h}$, there is a unique representative $\widetilde{M}$ satisfying three properties*:

(a) $a_{11} \geq a_{22} \geq a_{33}$.

(b) *If $a_{11} = a_{22}$, then (a_{12}, a_{13}) is the largest choice in the lexicographical ordering among the possible choices for A.*

(c) *If $a_{22} = a_{33} \neq a_{11}$, then (a_{23}, a_{13}) is the largest choice in the lexicographical ordering among the possible A.*

Example 4.2. If $M = \left(\begin{smallmatrix} 1 & 2 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 4 \end{smallmatrix}\right)$, then $\widetilde{M} = \left(\begin{smallmatrix} 4 & 0 & 1 \\ 0 & 2 & 1 \\ 0 & 2 & 1 \end{smallmatrix}\right)$. If $M = \left(\begin{smallmatrix} 2 & 1 & 1 \\ 2 & 2 & 0 \\ 0 & 1 & 4 \end{smallmatrix}\right)$, then $\widetilde{M} = \left(\begin{smallmatrix} 4 & 1 & 0 \\ 0 & 2 & 2 \\ 1 & 1 & 2 \end{smallmatrix}\right)$.

Definition 4.3. *Let $M = (m_{ij}) \in \mathcal{H}^n_{3,h}$. We define $\epsilon_M = (-1)^{m_{21}+m_{31}+m_{32}}$,*

$$M^* = \begin{pmatrix} m_{11} & m_{12}+m_{21} & m_{13}+m_{31} \\ 0 & m_{22} & m_{23}+m_{32} \\ 0 & 0 & m_{33} \end{pmatrix}, \quad \textit{and} \quad \overline{M} = \epsilon_M[M^*] \in \mathbf{U}^n_{3,h},$$

where $[M]$ represents the basis element represented by M. We define $\Delta(M) = \overline{\widetilde{M}}$ and extend it to a map $\Delta : \mathbf{H}^n_{3,h} \to \mathbf{U}^n_{3,h}$ by

$$\Delta\Big(\sum_{H \in \mathcal{H}^n_{3,h}} c_H[H]\Big) = \sum_{H \in \mathcal{H}^n_{3,h}} c_H \Delta(H).$$

Example 4.4. Let $M = \left(\begin{smallmatrix} 1 & 1 & 1 \\ 4 & 1 & 3 \\ 3 & 1 & 2 \end{smallmatrix}\right) \in \mathcal{H}^{10}_{3,4}$. Then $\Delta(M) = -[\mathcal{M}_{2,1}] \in \mathbf{U}^{10}_{3,4}$.

Letting S_3 act trivially on $\mathbf{U}^n_{3,h}$, we have:

Lemma 4.5. *$\Delta : \mathbf{H}^n_{3,h} \to \mathbf{U}^n_{3,h}$ is a surjective S_3-equivariant homomorphism of vector spaces.*

Proof. If $M \in \mathcal{U}^n_{3,h}$, then $\Delta(M) = [M]$. So the subspace in $\mathbf{H}^n_{3,h}$ generated by $[M]$, for $M \in \mathcal{U}^n_{3,h}$, surjectively maps onto $\mathbf{U}^n_{3,h}$. □

The composition map $\Delta \circ \theta : \mathrm{Sym}_{3,(n)^3,h} \to \mathbf{U}^n_{3,h}$ is a surjective S_3-equivariant homomorphism of vector spaces. To analyze its kernel, we define the vector spaces

$$\begin{aligned}
&\mathfrak{a}_{1;(n)^3,h} := \mathfrak{a}_1 \cap \mathrm{Sym}_{3,(n)^3,h},\\
&\mathfrak{a}_{2;(n)^3,h} := \mathfrak{a}_2 \cap \mathrm{Sym}_{3,(n)^3,h},\\
&I_1 := \mathfrak{a}_{1;(n)^3,h} + \langle P - \sigma \cdot P \mid P \in \mathrm{Sym}_{3,(n)^3,h}\rangle \subset \mathrm{Sym}_{3,(n)^3,h},\\
&\mathfrak{g} := (\Delta \circ \theta)(I_1) \subset \mathbf{U}^n_{3,h},\\
&\mathfrak{h} := (\Delta \circ \theta)(\mathfrak{a}_{2;(n)^3,h}) \subset \mathbf{U}^n_{3,h}.
\end{aligned}$$

We begin by determining $\mathfrak{g}$.

Proposition 4.6. (a) *A basis of $\mathfrak{g}$ is given by $\{[B]\}$, with $B = (b_{ij}) \in \mathcal{U}^n_{3,h}$ and*

(i) *$b_{11} = b_{22} \neq b_{33}$ and $b_{12} \equiv 1 \bmod 2$, or*
(ii) *$b_{22} = b_{33} \neq b_{11}$ and $b_{23} \equiv 1 \bmod 2$, or*
(iii) *$b_{11} = b_{22} = b_{33}$ and $b_{12} \equiv 1 \bmod 2$.*

(b) *The induced map $\overline{\Delta \circ \theta} : \mathrm{Sym}_{3,(n)^3,h}/I_1 \to \mathbf{U}^n_{3,h}/\mathfrak{g}$ is an isomorphism of vector spaces.*

Proof of 4.6(a). Since $\Delta \circ \theta$ is S_3-equivariant, $\mathfrak{g}$ is generated by $(\theta \circ \Delta)(\mathfrak{a}_{1;(n)^3,h})$. Now $\mathfrak{a}_{1;(n)^3,h}$ is generated by the images $(x_{ab}+x_{ba})P$, where $P = \prod_{a\neq b} x_{ab}^{\lambda_{ab}} \prod_c x_c^{\sigma_c}$ is an order h monomial in Sym_3 with weight $(n-1, n-1, n)$, $(n-1, n, n-1)$, or $(n, n-1, n-1)$. Let $M = \theta(P)$. By the S_3-equivariancy, we can assume $M = (m_{ab})$ with $m_{11} \geq m_{22} \geq m_{33}$. If $m_{11} > m_{22} > m_{33}$, then $(\Delta\circ\theta)(x_{ab}P) = -(\Delta\circ\theta)(x_{ba}P)$ and

$$(\Delta \circ \theta)((x_{ab} + x_{ba})P) = 0.$$

Now suppose $m_{11} = m_{22} > m_{23}$ and $(a, b) = (1, 2)$. By working through the various cases, one finds that the only case when $(\Delta\circ\theta)(x_{12}P) + (\Delta\circ\theta)(x_{21}P) \neq 0$ occurs when $m_{12} = m_{21}$. In this case, calculation shows $(\Delta \circ \theta)(x_{12}P) = (\Delta \circ \theta)(x_{21}P)$ and

$$(\Delta \circ \theta)((x_{12} + x_{21})P) = 2[B],$$

where $B = (b_{ij})$ is a matrix described in case (i). In particular, $b_{12} = 2m_{12} + 1$ is odd. Similar calculations in the other cases establish the rest of (a). □

Proof of 4.6(b). By the definition of $\mathfrak{g}$, the map $\overline{\Delta \circ \theta}$ is well-defined. Moreover it is surjective as $\Delta \circ \theta$ is surjective. To show it is injective, we need only

show that $\ker\Delta \subset \theta(I_1)$ as θ is an isomorphism. Let $\alpha \in \ker\Delta$ and express $\alpha = \sum_{M\in\mathcal{H}^n_{3,h}} c_M[M]$. We can rewrite α as $\alpha = \sum_{N\in\mathcal{U}^n_{3,h}} \alpha_N$, where

$$\alpha_N = \sum_{\substack{M\in\mathcal{H}^n_{3,h}\\ \Delta(M)=\pm[N]}} c_M[M].$$

Let $\Delta(\alpha_N) = b_N[N]$. Then $0 = \Delta(\alpha) = \sum_N b_N[N]$. Since the elements $[N]$, for $N \in \mathcal{U}^n_{3,h}$, form a basis of $\mathbf{U}^n_{3,h}$, we have $b_N = 0$, $\Delta(\alpha_N) = 0$ and $\alpha_N \in \ker\Delta$. We now fix $N \in \mathcal{U}^n_{3,h}$ and show $\alpha_N \in \theta(I_1)$.

For each S_3-orbit C in the set of $M \in \mathcal{H}^n_{3,n}$ such that $\Delta(M) = \pm[N]$, let $d_C = \sum_{M\in C} c_M$. Let $M_C \in \mathcal{U}^n_{3,h}$ to be the unique representative of C specified by Lemma 4.1. Then

$$\alpha_N = \sum_{\Delta(M)=\pm[N]} c_M[M] = \sum_C d_C[M_C] + \sum_C \sum_{M\in C} c_M([M]-[M_C]),$$

where the sum is over the finite number of orbits C. Since Δ is constant on the orbit C, we have

$$0 = \Delta(\alpha_N) = \sum_C d_C\,\Delta([M_C]).$$

Taking the difference of the two equations, we have

$$\alpha_N = \sum_C d_C\,([M_C] - \Delta([M_C])) + \sum_C \sum_{M\in C} c_M([M]-[M_C]).$$

The definition of Δ shows that $[M_C] - \Delta([M_C]) \in \theta(\mathfrak{a}_{1;(n)^3,h}) \subset \theta(I_1)$. As the second summand is formally in $\theta(I_1)$, $\alpha_N \in \theta(I_1)$ and (b) is proved. □

Define $\phi : \mathbf{U}^n_{3,h} \to \mathcal{C}^n_{3,h}$ by $\phi = \Lambda\circ\chi\circ\theta^{-1}$, where we restrict the domain of the isomorphism $\theta^{-1} : \mathbf{H}^n_{3,h} \to \mathrm{Sym}_{3,(n)^3,h}$ to $\mathbf{U}^n_{3,h}$.

Corollary 4.7. *ϕ induces an isomorphism* $\mathbf{U}^n_{3,h}/(\mathfrak{g}+\mathfrak{h}) \cong \mathcal{C}^n_{3,h}$.

Proof. By Corollary 2.7, we have an isomorphism $\Lambda\circ\chi : \mathrm{Sym}_{3,(n)^3,h}/I \cong \mathcal{C}^n_{3,h}$. Since $k = 3$, the syzygies of the first and second kind generate all the syzygies in $\mathfrak{a}$. Hence $I = I_1 + \mathfrak{a}_{2;(n)^3,h}$ and $(\Delta\circ\theta)(I) = \mathfrak{g}+\mathfrak{h}$. Using Proposition 4.6(b), it follows that $\Delta\circ\theta$ induces an isomorphism $\mathrm{Sym}_{3,(n)^3,h}/I \to \mathbf{U}^n_{3,h}/(\mathfrak{g}+\mathfrak{h})$. Putting these isomorphisms together and noting that Δ is the identity on $\mathbf{U}^n_{3,h}$, gives the claimed isomorphism ϕ. □

5. Calculation of $\mathfrak{g}+\mathfrak{h}$

In this section, we explicitly calculate $\mathfrak{g}+\mathfrak{h}$. This calculation is then used in the next section to calculate $\mathcal{C}^n_{3,h}$. We begin by calculating $\mathfrak{g}$.

Proposition 5.1. (a) *If $n+h \equiv 0 \bmod 4$, then $\mathfrak{g} = \{0\}$.*

(b) *If* $n+h \equiv 2 \bmod 4$ *and* $\alpha = \max(0, (h-n)/2)$, *then* $\mathfrak{g}$ *is generated by*

$$\{[\mathcal{M}_{h-2r,r}] : \alpha \le r \le h/3\} \cup \{[\mathcal{M}_{(h-r)/2,r}] : \alpha \le r < h/3,\ r \equiv h \pmod 2\}.$$

Proof of 5.1(a). Let $n+h=4t$ for some t. Now by Proposition 4.6, $\mathfrak{g}$ is generated by $[M]$, with matrices M of three types: (i), (ii), and (iii). Suppose M has type (i). Then $r = h-2s$ and

$$r + \frac{n-h}{2} = h - 2s + \frac{n-h}{2} = \frac{n+h}{2} - 2s = 2(t-s)$$

is even, contradicting the definition of M. Similarly, assuming M has types (ii), (iii) leads to contradictions. Hence, $\mathfrak{g} = \{0\}$.

Proof of 5.1(b). By Proposition 4.6, $\mathfrak{g}$ is generated by $[\mathcal{M}_{s,r}]$, where (i) $s = h-r-s$ or (ii) $h-r-s = r$. Thus $s = (h-r)/2$ or $h-2r$. Each of the matrices $\mathcal{M}_{h-2r,r}$, for $\max(0, (h-n)/2) \le r \le h/3$, and $\mathcal{M}_{(h-r)/2,r}$, for $0 \le r \le h/3$ and $r \equiv h \pmod 2$, is in $\mathcal{U}^n_{3,h}$. These matrices are distinct except when $s = r = h-r-s$, $h \equiv 0 \pmod 3$, and $r = s = h/3$. □

Unfortunately, it is not as simple to describe the generators of $\mathfrak{h}$. Instead, we can determine the generators of $(\mathfrak{g}+\mathfrak{h})/\mathfrak{g}$. Then by combining them with the generators of $\mathfrak{g}$, we will have a set of generators for $\mathfrak{g}+\mathfrak{h}$. Recalling the definition of $\mathcal{M}_{s,r}$ in (7), we define

$$m_{s,r} = [\mathcal{M}_{s,r}] - [\mathcal{M}_{s-1,r}] + [\mathcal{M}_{s-1,r+1}] \in \mathbf{U}^n_{3,h},$$

for nonnegative integers r, s satisfying $r+s \le h$.

Proposition 5.2. *Let* $S_{n,h}$ *be defined as in* (8). *Then* $(\mathfrak{g}+\mathfrak{h})/\mathfrak{g}$ *is generated by* $\Delta(m_{s+1,r})$ *for all* $(s,r) \in S_{n-1,h-1}$.

Before proving Proposition 5.2, we need to establish some lemmas about the functions ϵ_M, M^*, $\overline{M}$, $\widetilde{M}$, and $\Delta(M)$, defined in Definition 4.3.

Lemma 5.3. *For* $\sigma \in S_3$, $M \in \mathcal{H}^n_{3,h}$, $\epsilon_{\sigma(M)} = \epsilon_M \epsilon_{\sigma(M^*)}$.

Proof. Let $M = (m_{ij})$. The lemma follows from straightforward calculation for each $\sigma \in S_3$. For example, when $\sigma = (12)$, one has

$$\epsilon_{\sigma(M)} = (-1)^{m_{12}+m_{31}+m_{32}} = (-1)^{m_{21}+m_{31}+m_{32}}(-1)^{m_{12}+m_{21}} = \epsilon_M \epsilon_{\sigma(M^*)}. \qquad \square$$

Lemma 5.4. *Suppose we are given* $\sigma \in S_3$, $M \in \mathcal{H}^n_{3,h}$ *with* $\sigma(M)^* = M^*$. *Then* $\overline{\sigma(M)} = \operatorname{sgn}(\sigma)^{(n+h)/2}\overline{M}$, *where* sgn *is the nontrivial homomorphism* $\operatorname{sgn} : S_3 \to \{\pm 1\}$.

Proof. We have $\overline{M} = \epsilon_M M^*$. Then by Lemma 5.3,

$$\overline{\sigma(M)} = \epsilon_M \epsilon_{\sigma(M^*)} \sigma(M)^* = \epsilon_{\sigma(M^*)} \overline{M}.$$

If $\sigma = 1$, then the lemma trivially holds true. If $\sigma \neq 1$, then at least two of the diagonal elements of M are equal. If $\sigma = (12)$, then $m_{11} = m_{22}$. Since $(\sigma(M^*))_{21} = (n+h)/2 - m_{11} - m_{22} \equiv (n+h)/2 \bmod 2$, $\epsilon_{\sigma(M^*)} = (-1)^{(n+h)/2} = \operatorname{sgn}(\sigma)^{(n+h)/2}$ and the lemma is established. Similar calculations establish the lemma for the other choices of σ. □

Lemma 5.5. *Assume $n+h$ is even. If $n+h \equiv 0 \bmod 4$ or the diagonal entries of $M \in \mathcal{H}^n_{3,h}$ are distinct, then $\Delta(M) = \epsilon_M \Delta(M^*)$. Otherwise, $\Delta(M) = \pm\Delta(M^*)$.*

Proof. Let $\sigma \in S_3$ be such that $\sigma(M) = \widetilde{M}$. We then have

$$\begin{aligned}\overline{\sigma(M^*)} &= \epsilon_{\sigma(M^*)}\sigma(M^*)^* \\ &= \epsilon_M \epsilon_{\sigma(M)}\sigma(M)^* \qquad \text{(by Lemma 5.3)} \\ &= \epsilon_M \Delta(M).\end{aligned}$$

Now if the diagonal elements of M are distinct, $\widetilde{(M^*)} = \sigma(M^*)$ and thus

$$\Delta(M^*) = \overline{\widetilde{M^*}} = \overline{\sigma(M^*)} = \epsilon_M \Delta(M).$$

Now suppose the diagonal entries of M are not distinct. Then $\widetilde{(M^*)} = (\sigma_1\sigma)(M^*)$, where $\sigma_1 \in S_3$ has the property that $(\sigma_1(\sigma(M)))^* = \sigma(M)^*$. By Lemma 5.4,

$$\Delta(M^*) = \overline{(\sigma_1\sigma)(M^*)} = \operatorname{sgn}(\sigma_1)^{(n+h)/2}\overline{\sigma(M)} = \operatorname{sgn}(\sigma_1)^{(n+h)/2}\epsilon_M \Delta(M).$$

When $n+h \equiv 0 \bmod 4$, $\operatorname{sgn}(\sigma_1)^{(n+h)/2} = 1$; otherwise, it equals ± 1. Hence, the lemma is established. □

Proof of Proposition 5.2. By definition, $\mathfrak{h}$ is generated by

$$(\Delta \circ \theta)((x_{ij}x_k + x_{jk}x_i + x_{ki}x_j)P), \tag{9}$$

where $P \in \mathrm{Sym}_{3,d-1,h-1}$ is a monomial. Since Δ is invariant under the action of S_3 on $\mathcal{H}^n_{3,h}$, one can assume that $\theta(P) = \widetilde{\theta(P)}$. Then $\theta(P)^* = \mathcal{M}_{s,r}$, with $(s,r) \in S_{n-1,h-1}$. It is also enough to consider the cases $(i,j,k) = (1,2,3), (1,3,2)$. Assume $(i,j,k) = (1,2,3)$. Then $\theta(x_{23}x_1P)^* = \mathcal{M}_{s+1,r}$ and $(\theta(\widetilde{x_{23}x_1P}))^* = \mathcal{M}_{s+1,r}$ by the assumption on P, so

$$(\Delta \circ \theta)(x_{23}x_1P) = \epsilon_P \Delta(\mathcal{M}_{s+1,r}).$$

When $n+h \equiv 0 \bmod 4$, by Lemma 5.5, $\theta(x_{31}x_2P)^* = \mathcal{M}_{s,r}$, $\theta(x_{12}x_3P)^* = \mathcal{M}_{s,r+1}$, and thus

$$(\Delta \circ \theta)(x_{31}x_2P) = -\epsilon_P \Delta(\mathcal{M}_{s,r}), \qquad (\Delta \circ \theta)(x_{12}x_3P) = \epsilon_P \Delta(\mathcal{M}_{s,r+1}).$$

Thus (9) equals $\epsilon_P \Delta(m_{s+1,r})$. When $(i,j,k) = (1,3,2)$, one gets $-\epsilon_P \Delta(m_{s+1,r})$, proving the lemma when $n+h \equiv 0 \bmod 4$. Now suppose $n+h \equiv 2 \bmod 4$ and consider $x_{31}x_2P$. If the diagonal elements of $\theta(x_{31}x_2P)$ are distinct, we have

$(\Delta \circ \theta)(x_{31}x_2P) = -\epsilon_P \Delta(\mathcal{M}_{s,r})$ by Lemma 5.5. If they are not distinct, then $(\Delta \circ \theta)(x_{31}x_2P) = -\epsilon_P \Delta(\mathcal{M}_{s,r})$ in $\mathfrak{g}/(\mathfrak{g}+\mathfrak{h})$ as both sides are 0 by Proposition 5.1. The same logic shows $(\Delta \circ \theta)(x_{12}x_3P) = \epsilon_P \Delta(\mathcal{M}_{s,r+1})$, establishing (9) and proving the proposition when $n+h \equiv 2 \bmod 4$. □

To make the generators of $(\mathfrak{g}+\mathfrak{h})/\mathfrak{g}$ explicit, we define the following elements of $\mathbf{H}^n_{3,h}$:

$$n_{s,r} = [\mathcal{M}_{s,r}] - (1+(-1)^{(n+h)/2})[\mathcal{M}_{s-1,r}],$$
$$p_{s,r} = (1+(-1)^{(n+h)/2})[\mathcal{M}_{s,r}] + [\mathcal{M}_{s-1,r+1}],$$

for integers $s, r \geq 0$. In general, $n_{s,r}$, $p_{s,r}$ will not be elements of $\mathbf{U}^n_{3,h}$.

Proposition 5.6. $(\mathfrak{g}+\mathfrak{h})/\mathfrak{g}$ *is generated by the following elements of* $\mathbf{U}^n_{3,h}$:

(a) $m_{s,r}$, *where* $\max(0, (h-n)/2) \leq r \leq (h-4)/3$, $(h-r)/2+1 \leq s \leq h-2r-1$.

(b) $n_{h-2r,r}$, *where* $\max(0, (h-n)/2) \leq r \leq (h-2)/3$.

(c) $p_{(h-r+1)/2,r}$, *where* $\max(0, (h-n)/2) \leq r \leq (h-3)/3$ *and* $r \equiv h+1 \bmod 2$.

(d) $[\mathcal{M}_{(h+2)/3,(h-1)/3}]$, *if* $h \equiv 1 \bmod 3$.

Proof. By Proposition 5.2, $\mathfrak{h}$ is generated by $\Delta(m_{s,r})$, where $(s-1,r) \in S_{n-1,h-1}$. It follows immediately that $(s,r) \in S_{n,h}$ and $\mathcal{M}_{s,r} \in \mathcal{U}^n_{3,h}$. However, the terms $\mathcal{M}_{s-1,r}$, $\mathcal{M}_{s-1,r+1}$ might not be in $\mathcal{U}^n_{3,h}$, so we need to do a case-by-case analysis. Since $(s-1,r) \in S_{n-1,h-1}$, we have $s-1 \geq h-s-r \geq r$. We separately analyze the cases when we have equality or strict inequality.

Case (a): Suppose $s-1 > h-s-r > r$. Then $(s-1,r)$, $(s-1,r+1) \in S_{n,h}$, and $\mathcal{M}_{s-1,r}$, $\mathcal{M}_{s-1,r+1}$ in $\mathcal{U}^n_{3,h}$. Thus $\Delta(m_{s,r}) = m_{s,r}$. We now prove the claimed inequalities for s, r. From the assumptions, we have $2s \geq h-r+2$ and $h \geq s+2r+1$, giving the claimed conditions on s. Combining these equations, we obtain $h \geq (h-r)/2+2r+2$ and $h \geq 3r+4$. Thus $r \leq (h-4)/3$. The lower bound on r follows from $(s-1,r) \in S_{n-1,h-1}$. Conversely, if s, r satisfies the bounds in (a), then one can show that $s-1 > h-s-r > r$ and $(s-1,r) \in S_{n-1,h-1}$.

Case (b): Suppose $s-1 > h-s-r$ and $h-s-r = r$. Then $(s-1,r)$ is an element of $S_{n,h}$ but $(s-1,r+1)$ is not. Hence $\mathcal{M}_{s-1,r} \in \mathcal{U}^n_{3,h}$, but $\mathcal{M}_{s-1,r+1} \notin \mathcal{U}^n_{3,h}$. Since $s = h-2r$, we have

$$\Delta(\mathcal{M}_{s-1,r+1}) = (-1)^{s-1+(n-h)/2}[\mathcal{M}_{s-1,r}] = (-1)^{1+h+(n-h)/2}[\mathcal{M}_{s-1,r}],$$

and $\Delta(m_{s,r}) = n_{s,r} \in \mathbf{U}^n_{3,h}$. To establish the bounds, we see that $s-1 \geq r+1$. Then

$$3r+1 = r+r+(r+1) \leq r+(h-s-r)+(s-1) = h-1$$

and $r \leq (h-2)/3$. Conversely, if r satisfies the stated bounds in (b), and $s = h-2r$, one can show that $(s-1,r) \in S_{n-1,h-1}$ and $s-1 > h-s-r = r$.

Case (c): Suppose $s-1=h-s-r$ and $h-s-r>r$. Then $(s-1,r+1)$ is an element of $S_{n,h}$, but $(s-1,r)$ is not. Then $\mathcal{M}_{s-1,r+1}\in\mathcal{U}^n_{3,h}$, but $\mathcal{M}_{s-1,r}\notin\mathcal{U}^n_{3,h}$. As $r=h+1-2s$, r and h have different parities. Then

$$\Delta(\mathcal{M}_{s-1,r})=(-1)^{r+(n-h)/2}[\mathcal{M}_{s,r}]=(-1)^{1+h+(n-h)/2}[\mathcal{M}_{s,r}]$$

and $\Delta(m_{s,r})=[1+(-1)^{(n+h)/2}][\mathcal{M}_{s,r}]+[\mathcal{M}_{s-1,r+1}]=p_{s,r}$. Now $h-2r-1\geq a$, so $b=h+1-2s\geq h+1-2(h-2r-1)$ and $r\leq(h-3)/3$. Conversely, when the stated conditions on r hold, one can show that $(s,r)\in S_{n-1,h-1}$, with $a=(h-r+1)/2$, and $s-1=h-s-r>b$.

Case (d): Suppose $s-1=h-s-r=r$. Then $h-1=3r$ and $h\equiv 1 \bmod 3$. In this case, neither $(s-1,r)$ nor $(s-1,r+1)$ is in $S_{n,h}$. Then

$$\Delta(\mathcal{M}_{s-1,r})=(-1)^{r+(n-h)/2}[\mathcal{M}_{s,r}]=-(-1)^{(n+h)/2}[\mathcal{M}_{s,r}],$$

and $\Delta(\mathcal{M}_{s-1,r+1})=(-1)^{(n+h)/2}[\mathcal{M}_{s,r}]$. Thus

$$\Delta(m_{s,r})=\Delta([\mathcal{M}_{s,r}]-[\mathcal{M}_{s-1,r}]+[\mathcal{M}_{s-1,r+1}])=(1+2(-1)^{(n+h)/2})[\mathcal{M}_{s,r}].$$

Regardless of whether $n+h\equiv 0,2 \bmod 4$, $[\mathcal{M}_{s,r}]\in\mathfrak{h}$. Since $s=(h+2)/3$, $r=(h-1)/3$, we obtain the result in (d) in the proposition. □

6. Calculation of degree-three covariants

In this section, we determine an explicit basis for $\mathcal{C}^n_{3,h}$ in Theorem 6.1 and derive the formulas for $\dim\mathcal{C}^n_{3,h}$ in Table 1 as a corollary. We establish these results by combining the calculation of $\mathfrak{g}+\mathfrak{h}$ from the previous section with Corollary 4.7 to calculate $\mathcal{C}^n_{3,h}$. To simplify the statement of the theorem, we use the map $\phi:\mathbf{U}^n_{3,h}\to\mathcal{C}^n_{3,h}$ defined before Corollary 4.7. ϕ has the property that if $M=(m_{ij})\in\mathcal{U}^n_{3,h}$, then

$$\phi([M])=\prod_{a\neq b}(ab)^{m_{ab}}\prod_c c_x^{m_{cc}},$$

where we use the classical symbolic notation for $\mathcal{C}^n_{3,h}$ (see page 516).

Theorem 6.1. *Let n, h have the same parity. Then a basis of $\mathcal{C}^n_{3,h}$ is given as follows.*

(a) *If $n+h\equiv 1 \bmod 2$, then $\mathcal{C}^n_{3,h}=\{0\}$.*

(b) *When $n+h\equiv 0 \bmod 4$, the elements $\phi([\mathcal{M}_{(h-r)/2,r}])$, where $\max(0,\frac{h-n}{2})\leq r\leq h/3$ and $r\equiv h \bmod 2$, form a basis.*

(c) *If $n+h\equiv 2 \bmod 4$, the elements $\phi([\mathcal{M}_{(h-r+1)/2,r}])$, where $\max(0,\frac{h-n}{2})\leq r<(h-1)/3$ and $r\equiv h+1 \bmod 2$ form a basis.*

Example 6.2. When $n = 10$, $h = 12$, Theorem 6.1(c) states that $\{\phi([\mathcal{M}_{6,1}]), \phi([\mathcal{M}_{5,3}])\}$ is a basis of $\mathscr{C}^{10}_{3,12}$. Using the classical symbolic notation (compare Example 2.5), these covariants are $(ac)^4(bc)^5a_x^6b_x^5c_x$ and $(ab)^2(ac)^3(bc)^5a_x^5b_x^4c_x^3$.

Proof of 6.1. By Corollary 4.7,

$$\mathscr{C}^n_{3,h} \cong \mathbf{U}^n_{3,h}/\,(\mathfrak{g}+\mathfrak{h}) \cong (\mathbf{U}^n_{3,h}/\mathfrak{g})/\,(\mathfrak{g}+\mathfrak{h}/\mathfrak{g})\,.$$

By Lemma 3.7, (a) is established. We now consider cases (b) and (c). By the same lemma, the set

$$\{[\mathcal{M}_{s,r}] \mid (s,r) \in S_{n,h}\}$$

is a basis for $\mathbf{U}^n_{3,h}$. We define an order on the basis elements $\mathcal{M}_{s,r}$ by defining $[\mathcal{M}_{s,r}] \geq [\mathcal{M}_{s',r'}]$ if $(s, h-r-s, r) \geq (s', h-r', s', r')$ in the lexicographic order. We will prove the proposition by ordering the generators of $\mathfrak{g}+\mathfrak{h}$, from largest to smallest, by their largest terms (which will be distinct). Let $a = \dim \mathbf{U}^n_{3,h}$ and $b = \dim(\mathfrak{g}+\mathfrak{h})$. By expressing the generators of $\mathfrak{g}+\mathfrak{h}$ in terms of the $[\mathcal{M}_{s,r}]$, we obtain a $b \times a$ upper-triangular matrix $\mathscr{A}$ of relations. A basis of the quotient space $\mathbf{U}^n_{3,h}/(\mathfrak{g}+\mathfrak{h})$ is then given by the cosets of $[\mathcal{M}_{s,r}]$, for (s,r) corresponding to nonpivot columns of $\mathscr{A}$. We now separately analyze the details of parts (b) and (c).

(b) When $n+h \equiv 0 \bmod 4$, $\mathfrak{g} = 0$ and we are reduced to calculating the quotient space $\mathbf{U}^n_{3,h}/\mathfrak{h}$. Proposition 5.6 gives a basis for the vector space $\mathfrak{h}$ and the leading terms of each of the $m_{s,r}$, $n_{s,r}$, $p_{s,r}$ specified in Proposition 5.6 are $[\mathcal{M}_{s,r}]$. We note that by the proof of Proposition 5.6, the specified pairs (s,r) are distinct and comprise all $(s,r) \in S_{n,h}$ with $(s-1,r) \in S_{n-1,h-1}$. Hence the generators of $\mathbf{U}^n_{3,h}/\mathfrak{h}$ are the $[\mathcal{M}_{s,r}]$ for those pairs $(s,r) \in S_{n,h}$ with $(s-1,r) \notin S_{n-1,h-1}$. The definition of $S_{n,h}$ shows that such a pair (s,r) occurs precisely when $3 \mid h$ and $r = h/3$ or when $h-r$ is even and $s = (h-r)/2$. In the first case, let $h = 3t$. Then $r = t$ and the condition on s shows $s = t$. As such, this situation is a subcase of the second case. $\mathscr{C}^n_{3,h}$ is thus generated by the $[\mathcal{M}_{s,r}]$, for (s,r) in the second case, and this is what the proposition states.

(c) By Propositions 5.1 and 5.6, $\mathfrak{g}+\mathfrak{h}$ is generated by $[\mathcal{M}_{s,r}]$ where $2s = h-r$ or $2r = h-s$, and by $m_{s,r}$, $n_{h-2r,r}$, $p_{(h-r+1)/2,r}$. Since $n+h \equiv 2 \bmod 4$, $n_{h-2r,r} = [\mathcal{M}_{h-2r,r}]$, and $p_{(h-r+1)/2,r} = [\mathcal{M}_{(h-r-1)/2,r+1}]$ and both of them are included in the former set. The leading term of $m_{s,r}$ is $[\mathcal{M}_{s,r}]$ and the corresponding (s,r) specified in Proposition 5.6 are all those pairs $(s,r) \in S_{n,h}$ with $s-1 > h-s-r > r$. Since $(s,r) \in S_{n,h}$ implies $s \geq h-s-r \geq r$, the generators of $\mathbf{U}^n_{3,h}/(\mathfrak{g}+\mathfrak{h})$ are the $[\mathcal{M}_{s,r}]$ for which $(s,r) \in S_{n,h}$, $s-1 = h-s-r$, and $h-s-r > r$, which is what the proposition claims. $\square$

When $n \geq h$:

	$n+h \equiv 0 \bmod 4$	$n+h \equiv 2 \bmod 4$
h even	$\left[\frac{h}{6}\right]+1$	$\left[\frac{h}{6}\right]$
h odd	$\left[\frac{h+3}{6}\right]$	$\left[\frac{h+3}{6}\right]$

When $\frac{1}{3}h \leq n \leq h$:

$h \bmod 3$	$n+h \equiv 0 \pmod 4$	$n+h \equiv 2 \pmod 4$
0	$\left[\frac{3n-h}{12}\right]+1$	$\left[\frac{3n-h-2}{12}\right]+1$
1	$\left[\frac{3n-h}{12}\right]+1$	$\left[\frac{3n-h-2}{12}\right]$
2	$\left[\frac{3n-h}{12}\right]+1$	$\left[\frac{3n-h-2}{12}\right]+1$

Table 1. Formulas for $\dim \mathscr{C}^n_{3,h}$.

Corollary 6.3. *The dimension of the vector space $\mathscr{C}^n_{3,h}$ is given by the formulas in Table 1 when n, h have the same parity and $n \geq h/3$. Otherwise* $\dim \mathscr{C}^n_{3,h} = 0$.

Corollary 6.3 can also be derived via an explicit calculation using the classical Cayley–Sylvester formula for calculating $\dim \mathscr{C}^n_{d,h}$ [Sturmfels 2008, p. 153]. We note that the case $n = h$ appears twice in Table 1 with seemingly different formulas, but calculation shows that the formulas agree.

Theorem 6.1 parametrizes all the covariants in $\mathscr{C}^n_{3,h}$. We say $f \in \mathscr{C}^n_{3,h}$ is a reducible covariant if f can be written as a linear combination of products gh of covariants g, h with degree less than 3, $\deg g + \deg h = 3$, and $\operatorname{ord} g + \operatorname{ord} h = h$ Let $\mathrm{Red}^n_{3,h}$ be the subspace of $\mathscr{C}^n_{3,h}$ generated by the reducible covariants. An irreducible covariant $f \in \mathscr{C}^n_{3,h}$ is a covariant that is not reducible.

By Example 2.8, the only nonzero covariants of degree one are multiples of $(\Lambda \circ \chi)(x_1^n)$. By Example 2.9, the only nonzero degree-two covariants of order h occur when $0 \leq h \leq 2n$ and $h \equiv 2n \bmod 4$. In this case, they are given by multiples of $(\Lambda \circ \chi)(x_{12}^{n-(h/2)} x_1^{h/2} x_2^{h/2})$. Thus, $\mathrm{Red}^n_{3,h}$ is a one-dimensional space precisely when $h \geq n$ and $h + n \equiv 0 \bmod 4$, in which case it is generated by $\mathcal{M}_{n,(h-n)/2}$. Otherwise $\mathrm{Red}^n_{3,h} = \{0\}$.

Corollary 6.4. *If $h < n$ or $h + n \not\equiv 0 \bmod 4$, then $\mathrm{Red}^n_{3,h} = \{0\}$ and $\mathscr{C}^n_{3,h}$ contains no reducible covariants. If $h \geq n$ and $h + n \equiv 0 \bmod 4$, then the covariants $\phi([\mathcal{M}_{(h-r)/2,r}])$, where $r \equiv h \bmod 2$ and $(h-n)/2 < r \leq h/3$, form a basis for the subspace of $\mathscr{C}^n_{3,h}$ complementary to $\mathrm{Red}^n_{3,h}$.*

Proof. Only the second part remains to be shown. In this case, $\mathrm{Red}^n_{3,h}$ is generated by $[\mathcal{M}_{n,(h-n)/2}]$. We can assume $h \leq 3n$ as otherwise there are no nonzero

covariants. If $h = 3n$, then $(n, (h-n)/2) = (n, n)$, and $[\mathcal{M}_{n,(h-n)/2}]$ is one of the basis elements of $\mathcal{C}^n_{3,h}$ given by Theorem 6.1. By deleting this basis element, the remaining basis elements give a basis for the subspace of $\mathcal{C}^n_{3,h}$ complementary to $\mathrm{Red}^n_{3,h}$. Now let $h = 3n - 4k$, where $k \geq 1$. By using a syzygy of the second type, we have

$$[\mathcal{M}_{n,(h-n)/2}] = [\mathcal{M}_{n,n-2k}] = 2[\mathcal{M}_{n-1,n-2k}]$$

in $\mathbf{U}^n_{3,h}/(\mathfrak{g}+\mathfrak{h})$. If $k = 1$, $(n-1, n-2k) = ((h-r)/2, r)$ for $n = h - 2k$, and $[\mathcal{M}_{n,(h-n)/2}] = [\mathcal{M}_{(h-r)/2,r}]$ is again one of the basis elements of $\mathcal{C}^n_{3,h}$ given by Theorem 6.1. By excluding this basis element, we obtain the desired basis for the complementary subspace. Now suppose $k > 1$. Then by applying the second syzygy $k-1$ additional times, we obtain in $\mathbf{U}^n_{3,h}/(\mathfrak{g}+\mathfrak{h})$

$$[\mathcal{M}_{n,n-2k}] = 2[\mathcal{M}_{n-k,n-2k}] + \sum_{r=1}^{k-1} c_r [\mathcal{M}_{n-k,n-2k+r}],$$

for some constants c_r. Now by the ordering introduced in the proof of Theorem 6.1, $[\mathcal{M}_{n-k,n-2k}] > [\mathcal{M}_{n-k,n-2k+r}]$ for $1 \leq r \leq k-1$, and thus each $[\mathcal{M}_{n-k,n-2k+r}]$ in the summand can be expressed as a linear combination of $[\mathcal{M}_{s,r}]$ given by Theorem 6.1 with $[\mathcal{M}_{s,r}] < [\mathcal{M}_{n-k,n-2k}]$. Let $\mathcal{I}$ be the space generated by the $[\mathcal{M}_{(h-r)/2,r}]$ specified in the Corollary. Each $[\mathcal{M}_{n-k,n-2k+r}]$ in the summand can then be expressed as a linear combination of elements in $\mathcal{I}$. Thus the space generated by $[\mathcal{M}_{n,n-2k}]$ and $\mathcal{I}$ is also the space generated by $[\mathcal{M}_{n-k,n-2k}]$ and $\mathcal{I}$. Since $(n-k, n-2k) = (\frac{h-r}{2}, r)$ for $r = n - 2k$, this is $\mathcal{C}^n_{3,h}$, by Theorem 6.1. Thus $\mathcal{I}$ is the subspace of $\mathcal{C}^n_{3,h}$ complementary to $\mathrm{Red}^n_{3,h}$. □

Historical remark. We would like to note a correction to a claim in Hilbert's fundamental book on covariants [1993]. First, we define the *weight* w of a covariant of degree d, order h, and degree-n form to be $w = (dn - h)/2$. When $d = 3$,

$$w = (3n - h)/2 = 2n - (n + h)/2.$$

On [Hilbert 1993, p. 62], there appears the statement:

> "Regarding the covariants of degree three, they all have odd weight $p = 2\pi + 1$ and are those which occur in the following expression, where $p = 3, 5, 7, \ldots, n$, respectively $n - 1$:"

Hilbert then gives an explicit formula for a covariant f_p of weight p. In total, the claim is that all degree-three covariants have odd weight and that there is exactly one nonreducible covariant of each odd weight $3 \leq p \leq n$. It is clear, both from the Cayley–Sylvester formula and from Theorem 6.1, that this is incorrect. From Theorem 6.1, one sees that in general there are many covariants with a given even

weight when $n+h \equiv 0 \bmod 4$. Similarly with $n+h \equiv 2 \bmod 4$, there are generally multiple covariants of a given odd weight.

Comparison with result of Kraft and Weyman. Kraft and Weyman [1999, Theorem 6.7] establish a generating set for the covariants in $\mathscr{C}^n_{3,h}$ with similarities to our Theorem 6.1. We now briefly discuss the differences between these results. Using the classical symbolic notation (see page 516), let

$$P = (ab)^\alpha (bc)^\beta (ca)^\gamma a_x^{n-\alpha-\gamma} b_x^{n-\alpha-\beta} c_x^{n-\beta-\gamma} \in \mathscr{C}^n_{3,h}$$

be a covariant of order $h = 3n - 2m$, where $m = \alpha + \beta + \gamma$. We assume $n \geq \max(\alpha+\beta, \alpha+\gamma, \beta+\gamma)$. We define $\operatorname{cat} P := \max(\alpha, \beta, \gamma)$. Kraft and Weyman prove:

Theorem 6.5 (Kraft and Weyman's abc-Theorem). *Assume $n, h \geq 0$.*

(a) *If $n \leq h$, then P is a linear combination of covariants*

$$Q = (ab)^\mu (bc)^\eta a_x^{n-\mu} b_x^{n-\mu-\eta} c_x^{n-\eta}$$

with $\mu + \eta = m$, $\mu \geq 2\eta$ and $\mu \geq \operatorname{cat} P$.

(b) *If $h = n$, then P belongs to the ideal $\mathscr{I} \subset P(V_n)$ generated by all covariants of degree $k \leq 2$ and order $h \leq \frac{3}{4}n$.*

(c) *If $n \geq h$, then P belongs to the ideal $\mathscr{J} \subset P(V_n)$ generated by all covariants with degree $k \leq 3$ and order $h \leq \frac{3}{4}n$.*

Since $\mathscr{C}^n_{3,h}$ is generated by P, as α, β, γ vary, part (a) of Theorem 6.5 gives a spanning set for the vector space $\mathscr{C}^n_{3,h}$ when $n \leq h$. However, this spanning set is almost always linearly dependent and doesn't give a basis for $\mathscr{C}^n_{3,h}$. For example, when $n = 10$, $h = 12$, part (a) shows that $\mathscr{C}^{10}_{3,12}$ is generated by the three symbolic monomials Q corresponding to $(\mu, \eta) = (6, 3)$, $(7, 2)$, and $(8, 1)$ (the covariant corresponding to $(\mu, \eta) = (9, 0)$ is zero). However, $\dim \mathscr{C}^{10}_{3,12} = 2$ by Table 1, and Example 6.2 shows that the two monomials $(ac)^4(bc)^5 a_x^6 b_x^5 c_x$ and $(ab)^2(ac)^3(bc)^5 a_x^5 b_x^4 c_x^3$ form a basis for $\mathscr{C}^{10}_{3,12}$.

Similarly, parts (b), (c) of Theorem 6.5 show that the vector space of covariants $\mathscr{C}^n_{3,h}$ is contained in the respective ideals $\mathscr{I}, \mathscr{J}$ of the ring $P(V_n)$, but do not establish a basis for $\mathscr{C}^n_{3,h}$. To see the difference, we consider the case when $n = h = 3$. Then part (b) states that $\mathscr{C}^3_{3,3}$ is contained in the ideal $\mathscr{I}$ of $P(V_3)$ (a ring containing both covariant and noncovariant functions) generated by reducible covariants. However, by Corollary 6.4, we know that $\mathrm{Red}^3_{3,3} = \{0\}$ and $\mathscr{C}^3_{3,3}$ is generated as a vector space by a single irreducible covariant.

Acknowledgements

It is a pleasure to thank The College of New Jersey for its support of this work through its Summer Undergraduate Research Program. We would like to thank both J. Hatley for his computational help that aided our calculations, and the referee for many helpful suggestions.

References

[Bedratyuk 2009] L. Bedratyuk, "A complete minimal system of covariants for the binary form of degree 7", *J. Symbolic Comput.* **44**:2 (2009), 211–220. MR 2009j:13006 Zbl 05494972

[Bedratyuk and Bedratyuk 2008] L. Bedratyuk and S. Bedratyuk, "A complete system of covariants for the binary form of the eighth degree", *Mat. Visn. Nauk. Tov. Im. Shevchenka* **5** (2008), 11–22. In Ukrainian; English version in http://www.arxiv.org/abs/math/0612113. Zbl 05592395

[Dolgachev 2003] I. Dolgachev, *Lectures on invariant theory*, London Mathematical Society Lecture Note Series **296**, Cambridge University Press, 2003. MR 2004g:14051 Zbl 1023.13006

[Gordan 1868] P. Gordan, "Beweis, dass jede Covariante und Invariante einer binären Form eine ganze Funktion mit numerischen Coefficienten einer endlichen Anzahl solcher Formen ist", *J. Reine und Angewandte Mathematik* **69** (1868), 323–354. JFM 01.0060.01

[Hilbert 1890] D. Hilbert, "Ueber die Theorie der algebraischen Formen", *Math. Ann.* **36**:4 (1890), 473–534. MR 1510634

[Hilbert 1893] D. Hilbert, "Ueber die vollen Invariantensysteme", *Math. Ann.* **42**:3 (1893), 313–373. MR 1510781 JFM 25.0173.01

[Hilbert 1993] D. Hilbert, *Theory of algebraic invariants*, Cambridge University Press, 1993. MR 97j.01049 Zbl 0801.13001

[Howe 1994] R. Howe, "The invariants of degree up to 6 of all n-ary m-ics", pp. 335–348 in *Lie theory and geometry*, edited by J.-L. Brylinski et al., Progr. Math. **123**, Birkhäuser, Boston, MA, 1994. MR 96f:13007 Zbl 0901.15018

[Kraft and Weyman 1999] H. Kraft and J. Weyman, "Degree bounds for invariants and covariants of binary forms", preprint, 1999, Available at www.math.unibas.ch/~kraft/Papers/KWJordan.pdf.

[Kung and Rota 1984] J. P. S. Kung and G.-C. Rota, "The invariant theory of binary forms", *Bull. Amer. Math. Soc.* (*N.S.*) **10**:1 (1984), 27–85. MR 85g:05002 Zbl 0577.15020

[Olver 1999] P. J. Olver, *Classical invariant theory*, London Mathematical Society Student Texts **44**, Cambridge University Press, 1999. MR 2001g:13009 Zbl 0971.13004

[Procesi 2007] C. Procesi, *Lie groups*, Springer, New York, 2007. MR 2007j:22016 Zbl 1154.22001

[Sturmfels 2008] B. Sturmfels, *Algorithms in invariant theory*, 2nd ed., Springer, Vienna, 2008. MR 94m:13004 Zbl 1154.13003

Received: 2008-10-16 Revised: 2009-12-08 Accepted: 2009-12-21

hagedorn@tcnj.edu *Department of Mathematics and Statistics, The College of New Jersey, P.O. Box 7718, Ewing, NJ 08628-0718, United States*

glenmatthewwilson@gmail.com *Department of Mathematics and Statistics, The College of New Jersey, P.O. Box 7718, Ewing, NJ 08628-0718, United States*

INVOLVE 2:5(2009)

Isometric composition operators acting on the Chebyshev space

Thomas E. Goebeler, Jr. and Ashley L. Potter

(Communicated by David Larson)

Norms of certain composition operators are given in terms of their symbols in some finite-dimensional setting. Then a family of isometric composition operators acting on certain vector spaces is identified.

1. Introduction

Much research has been done concerning composition operators over the last five decades. However, this research has primarily focused on a host of standard questions about composition operators in the realm of complex functions. Recently the first author has proposed a number of questions regarding composition operators in the real function area for undergraduate student research supported by the Ursinus College Summer Fellows Program [Doperak 2006; Gareau 2005; Kunaszuk 2006; Potter 2007]. This topic offers accessibility for undergraduate research while providing fertile ground for genuinely new results.

This paper focuses on one particular real function space, the Chebyshev space, T, where we explore norm-related ideas for composition operators, specifically norms and isometries. The ultimate goal of this research is to find the norm of a composition operator C_g acting on the Chebyshev space in terms of its *symbol* g. We begin by considering norms of composition operators in the infinite-dimensional space, and move on to examine the topic in the finite-dimensional space. After looking at the various finite-dimensional subspaces, we begin to look at qualities that would lead to the symbol inducing an isometry.

2. Definitions

We set down some terminology and basic facts here. A *composition operator* C_g acts on functions f according to the rule $C_g(f) = f \circ g$. The function g is called the symbol of the operator. Necessarily g must have range contained in the domain of

MSC2000: 47B33, 47B38.
Keywords: composition operator, norm, isometry.

f for this to make sense. The typical assumptions are that C_g has a vector space $\mathcal{F}$ of functions as its domain, that the functions $f \in \mathcal{F}$ and g have a common domain D, and that $g(D) \subseteq D$.

To speak of an operator's norm, it must first be known that the operator is bounded on its domain space. The following definition is operational for any linear operator, not just composition operators.

Definition 1. A linear operator A is *bounded* from a vector space V to another vector space W provided there exists an $M \in \mathbb{R}$ such that $\|Af\|_W \leq M \cdot \|f\|_V$ for all $f \in V$. This constant M in the inequality is referred to as a bound.

If an operator $A : \mathcal{F} \to \mathcal{G}$ is known to be bounded, we can give it a norm (the operator norm), defined as $\|A\| = \sup_{f \neq 0} \|Af\|_{\mathcal{G}} / \|f\|_{\mathcal{F}}$.

The setting for the present work is a real function space. This is atypical of composition operator research, which sees the majority of work done on complex function spaces. We find this setting surprisingly rich, once we make some modifications to the questions we ask. The reader interested in more information regarding composition operators on spaces of complex functions can consult [Cowen and MacCluer 1995], which is widely regarded as the best resource for beginners in the field.

For the purposes of this paper, we make the following real-function definition of the Chebyshev space.

Definition 2. The *Chebyshev space* T is herein the completion of the set of all continuous functions defined on the interval $[-1, 1]$, taking on real values, and obeying the following integral convergence condition:

$$\|f\|^2 = \int_{-1}^{1} |f(x)|^2 \frac{1}{\sqrt{1-x^2}}\, dx < \infty.$$

There are many functions in the vector space T, including polynomials and many other elementary functions. In fact, the monomials form a basis for T.

Orthogonalizing the basis vectors $\{1, x, x^2, \ldots\}$ with respect to the inner product

$$\langle u, v \rangle = \int_{-1}^{1} u(x)\overline{v(x)} \frac{1}{\sqrt{1-x^2}} dx$$

leads to $1,\ x,\ x^2 - \frac{1}{2}, \ldots,$ which are the *Chebyshev polynomials* (*of the first kind*). However, the polynomials are most commonly normalized so that if v_n is the n-th Chebyshev polynomial, $v_n(1) = 1$. Doing so shows the first four are $v_0 = 1,\ v_1 = x,\ v_2 = 2x^2 - 1,\ v_3 = 4x^3 - 3x$.

More information about the Chebyshev polynomials can be found in [Lebedev 1972]. Initially we seek a formula for the norm of C_g in terms of a calculation involving the symbol g, but as it will be shown, it seems likely that C_g is unbounded

on the Chebyshev space for all but a few symbols. We thus reconsider the question of norm for the restriction of the composition operator to finite-dimensional subspaces.

Definition 3. The finite dimensional subspace T_n of T is the subspace of T that contains all polynomials of degree at most n. More concisely, we can say that $T_n = \mathrm{Span}\{1, x, x^2, \dots, x^n\}$.

To say $f \in T_n$ means that f is a linear combination of the basis vectors, in other words, $f = c_0 \cdot 1 + c_1 \cdot x + c_2 \cdot x^2 + \cdots + c_n \cdot x^n$.

Since T_n is finite-dimensional we know that C_g is automatically bounded, simply by virtue of being linear. See [Horn and Johnson 1985]. This turns out to be a crucial restriction that leads to both norm formulas and identification of isometric composition operators. For more information regarding norms, see [Akhiezer and Glazman 1993; Dunford and Schwartz 1958; Reed and Simon 1980].

3. Preliminary investigation

As is always the case when the domain and range spaces are the same and use the same norm, the identity function $g(x) = x$ induces a composition operator with norm 1:

$$\|C_g\| = \sup_{f \neq 0} \frac{\|C_g(f)\|}{\|f\|} = \frac{\|f \circ g\|}{\|f\|} = \frac{\|f\|}{\|f\|} = 1.$$

The next example suggests most composition operators will fail to be bounded on T. Consider the symbol $g(x) = ax$ for $|a| < 1$. In the following, suppose $0 < a < 1$. This condition is imposed to guarantee that range$(g) \subset [-1, 1]$. A straightforward substitution (valid since g is an increasing absolutely continuous function) leads us to

$$\begin{aligned}\|C_{ax}(f)\|^2 &= \int_{-1}^{1} |C_{ax}(f)(x)|^2 \frac{1}{\sqrt{1-x^2}} dx = \int_{-1}^{1} [f(ax)]^2 \frac{1}{\sqrt{1-x^2}} dx \\ &= \frac{1}{a} \int_{-a}^{a} [f(u)]^2 \frac{1}{\sqrt{1-\left(\frac{u}{a}\right)^2}} dx.\end{aligned}$$

But $\dfrac{u}{a} \geq u$ for $a \in (0, 1)$ so

$$\frac{1}{\sqrt{1-\left(\frac{u}{a}\right)^2}} \geq \frac{1}{\sqrt{1-u^2}}.$$

Thus,

$$\|C_{ax}(f)\|^2 \geq \frac{1}{a} \int_{-a}^{a} [f(u)]^2 \frac{1}{\sqrt{1-u^2}} du.$$

If this lower bound is not convergent as an improper integral, the operator is unbounded; therefore this case will be investigated further. Let

$$f_a(x) = \frac{1}{\sqrt[4]{|a-x|}}$$

and observe that $f_a \in T$, as

$$\begin{aligned}\int_{-1}^{1} |f_a|^2 \frac{1}{\sqrt{1-x^2}}\,dx &= \int_{-1}^{1} \left|\frac{1}{\sqrt[4]{|a-x|}}\right|^2 \frac{1}{\sqrt{1-x^2}}\,dx \\ &= \int_{-1}^{1} \frac{1}{\sqrt{1-x}}\frac{1}{\sqrt{1+x}}\frac{1}{\sqrt{|a-x|}}\,dx < \infty\end{aligned}$$

is an improper integral with singularities as both endpoints of the interval of integration and at $x = a$. Note $(C_{ax}(f_a))(x) = f_a(ax) = |a|^{-1/4}\,|1-x|^{-1/4}$. Therefore,

$$\begin{aligned}\|C_{ax}(f_a)\|^2 &= \int_{-1}^{1} |f_a(ax)|^2 \frac{1}{\sqrt{1-x^2}}dx = \int_{-1}^{1} \left|\frac{1}{\sqrt[4]{|a|}} \cdot \frac{1}{\sqrt[4]{1-x}}\right|^2 \frac{1}{\sqrt{1-x^2}}dx \\ &= \frac{1}{\sqrt{|a|}} \int_{-1}^{1} \frac{1}{\sqrt{1-x}} \cdot \frac{1}{\sqrt{1-x}} \cdot \frac{1}{\sqrt{1+x}}dx = \frac{1}{\sqrt{|a|}} \int_{-1}^{1} \frac{1}{1-x} \cdot \frac{1}{\sqrt{1+x}}dx = \infty.\end{aligned}$$

The divergence is driven by the factor $1/(1-x)$. We conclude that C_{ax} is unbounded for $a \in (0, 1)$, and, by symmetry, for $a \in (-1, 0)$. Note that the operator C_{-x} is bounded since C_x is bounded and the integral and weight are symmetric about 0. (For the reader with a background in Lebesgue spaces, notice also that when $a = 0$, C_{ax} amounts to being the operator of point-evaluation at 0. Since our space is really the space $L^2([-1, 1],\ dx/\sqrt{1-x^2})$, that is, the completion of the polynomials in the Chebyshev norm, C_0 fails to make sense. Indeed, all point-evaluation operators on such an L^2-space are unbounded.)

A similar argument shows that when $g(x) = ax + b$, with range$(g) \subseteq [-1, 1]$, C_g is again unbounded. Algebra shows the conditions on a and b are $|a| \le 1$ and $|b| \le 1 - |a|$. This means graph$(g) \subseteq [-1, 1] \times [-1, 1]$; this Cartesian product of intervals will be called the "box." We include some examples to illustrate the phrases "in the box" and "out of the box" (Figure 1).

Again, a test function shows C_g is unbounded: let $f_{a,b}(x) = 1/\sqrt[4]{|a+b-x|}$ and perform the norm calculation of $C_{ax+b}(f_{a,b})$.

4. A change of venue

The unboundedness of operators with such elementary symbols leads us to restrict the operator C_g to finite-dimensional subspaces of T.

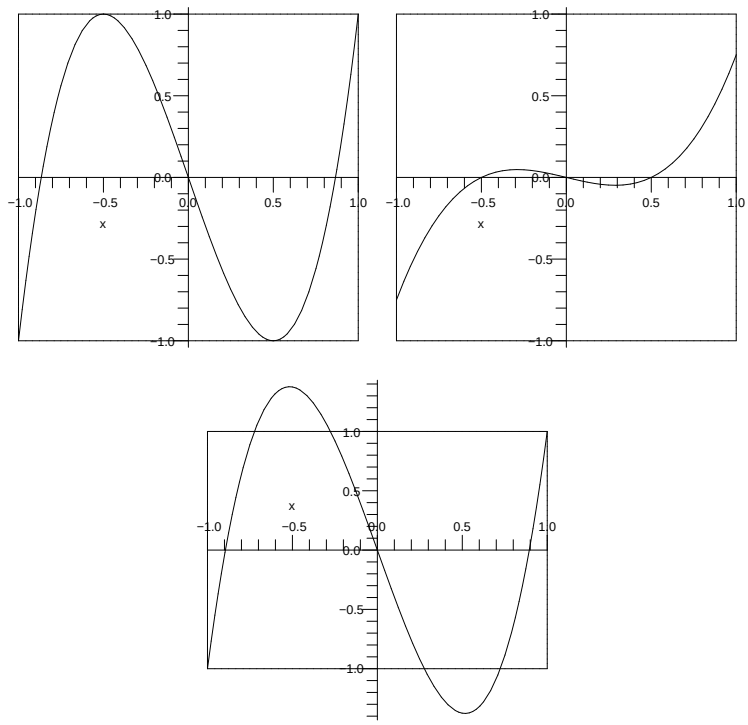


Figure 1. $g(x)$ in the box (top line); $g(x)$ outside the box (bottom).

4.1. $\boldsymbol{T_1 \to T_1}$**.** In order for C_g to map T_n back to T_n, g must be at most degree one and thus $g(x) = ax + b$ for some $a, b \in \mathbb{R}$. A concrete subcase would be $T_1 \to T_1$, and thus input functions are of the form $c_0 + c_1 x$ while $g(x) = ax + b$. In order for $C_g : T_n \to T_{2n}$, g must be at most degree two, $g(x) = ax^2 + bx + c$ for $a, b, c \in \mathbb{R}$. A concrete subcase would be $T_1 \to T_2$, and thus again $f \in T_1$ has the form $c_0 + c_1 x$ while $g(x) = ax^2 + bx + c$.

Employing the definition of the norm of an operator we have

$$\begin{aligned}
\|C_g\|^2_{T_1 \to T_1} &= \sup_{f \neq 0} \frac{\|C_g(f)\|^2}{\|f\|^2} = \max_{f \neq 0} \frac{\|C_g(f)\|^2}{\|f\|^2} \\
&= \max_{c_0^2 + c_1^2 \neq 0} \frac{\int_{-1}^{1} (c_0 + c_1(ax+b))^2 \frac{1}{\sqrt{1-x^2}}\, dx}{\int_{-1}^{1} (c_0 + c_1 x)^2 \frac{1}{\sqrt{1-x^2}}\, dx} \\
&= \max_{c_0^2 + c_1^2 \neq 0} \frac{a^2 c_1^2 + 2b^2 c_1^2 + 4bc_1c_0 + 2c_0^2}{c_1^2 + 2c_0^2}.
\end{aligned}$$

It is proper to replace the supremum with a maximum since we are working on a finite dimensional space; details of this thought can be found in [Horn and

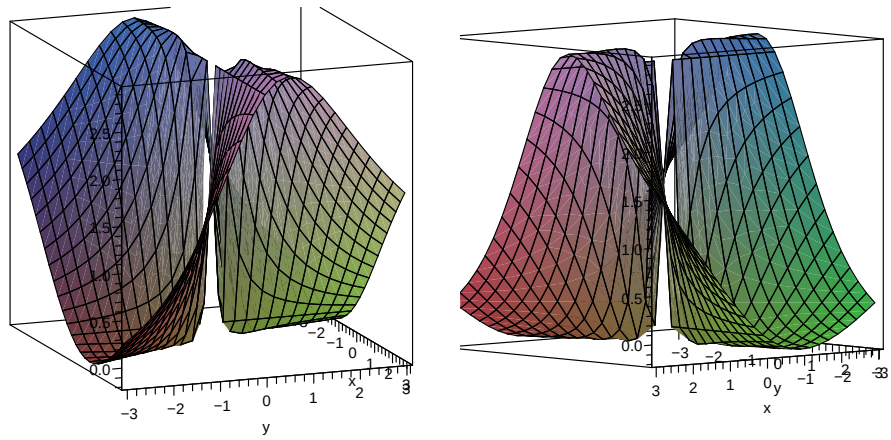


Figure 2. Two views of the "ridge".

Johnson 1985]. The quantity being maximized in the last line will be called the norm quotient.

After finding partial derivatives, and solving for c_1, critical lines are found. These critical lines correspond to "ridges" of the norm quotient, see Figure 2, rather than isolated maxima or minima.

The following critical lines were found:

$$c_1 = \frac{c_0(a^2+2b^2-1) \pm c_0\sqrt{(a^2+2b^2-1)^2+8b^2}}{2b}.$$

After expanding and simplifying, the following formula is obtained:

$$\|C_g\|^2_{T_1\to T_1} = \frac{\sqrt{(a^2+2b^2-1)^2+8b^2}+a^2+2b^2+1}{2}.$$

At this point an interesting question arises: how do these norms change when viewing the operators successively $T_2 \to T_2$, $T_3 \to T_3, \dots, T_n \to T_n, \dots$, as $n \to \infty$? In other words, can we determine for all symbols g the limit $\lim_{n\to\infty} \|C_g\|_{T_n\to T_n}$?

4.2. $T_1 \to T_2$. We restrict the domain of C_g to polynomials of the form $c_0 + c_1 x$ and restrict the form of the symbol to $ax^2 + bx + c$. Thus we are viewing the operator as $C_g : T_1 \to T_2$. Throughout this section, we assume $a \neq 0$, or else this reduces to the situation in Section 4.1, $T_1 \to T_1$.

Before proceeding with the calculation of the norm, we will address the constraints on a, b, and c for $g(x)$ to stay within the $[-1, 1]$ interval. By inspection, $|c| \leq 1$ because otherwise $g(x)$ would be outside the box, for then it would have an intercept outside the box. Using the fact that $|c| \leq 1$, it becomes clear that $|a + c| \leq 1$. To see this, consider the endpoints. Evaluation of g at -1 yields

$-1 \le a-b+c \le 1$. Evaluation of g at 1 yields $-1 \le a+b+c \le 1$. Adding these inequalities gives us $-2 \le 2a+2c \le 2$ which can be reduced to $-1 \le a+c \le 1$ and thus $|a+c| \le 1$.

By combining the previous constraints, it must be true that $|a| \le 2$. To put constraints on b, in terms of a and c, there are two conditions:

- The values of ax^2+bx+c at the endpoints of the interval $[-1, 1]$ cannot be above or below the corners of the $[-1, 1] \times [-1, 1]$ box.
- If the vertex of ax^2+bx+c occurs for $x \in [-1, 1]$, it cannot be above the upper boundary or below the lower boundary of the $[-1, 1] \times [-1, 1]$ box.

The endpoints concerning b will be addressed first:

$$\begin{aligned} g(-1) &= a-b+c, & g(1) &= a+b+c, \\ -1 &\le a-b+c \le 1, & -1 &\le a+b+c \le 1, \\ -1-a-c &\le -b \le 1-a-c, & -1-a-c &\le b \le 1-a-c, \\ -1+(a+c) &\le b \le 1+(a+c), & -1-(a+c) &\le b \le 1-(a+c). \end{aligned}$$

For $a+c>0$, it is clear that

$$-1-(a+c) \le -1+(a+c) \le b \le 1-(a+c) \le 1+(a+c).$$

Thus, $-1+(a+c) \le b \le 1-(a+c)$, or, $|b| \le |1-(a+c)|$. For $a+c<0$, following a similar approach, we arrive at $|b| \le |1+(a+c)|$. Combining both cases, the inequality limiting the values of b is

$$|b| \le \big|1-|a+c|\big|.$$

The next condition for the bounds of b concerns the vertex location within the $[-1, 1] \times [-1, 1]$ box. The vertex is where $x = -\frac{b}{2a}$. (Recall that we have assumed $a \neq 0$.) If the vertex is outside the box ($|b| > 2|a|$), we need only consider the values of the symbol at $x = \pm 1$, as above. However, if the vertex is in the box ($|b| \le 2|a|$), further constraints must be imposed for graph(g) to be in the box.

To find the bounds on b when the vertex is inside the box, calculate:

$$g\Big(-\frac{b}{2a}\Big) = a\Big(-\frac{b}{2a}\Big)^2 + b\Big(-\frac{b}{2a}\Big) + c.$$

We require $|a\,(-b/2a)^2 + b\,(-b/2a) + c| \le 1$, which simplifies to $|c - b^2/4a| \le 1$. Algebra shows we therefore want $-4(1-c) \le b^2/a \le 4(1+c)$.

This leads to two cases, depending upon the sign of a. If $a>0$, the necessary condition is $-4a(1-c) \le b^2 \le 4a(1+c)$, or, $|b| \le 2\sqrt{a(1+c)}$. If $a<0$, the necessary condition is $4a(1+c) \le b^2 \le -4a(1-c)$, or, $|b| \le 2\sqrt{-a(1-c)}$.

Summarizing, for graph(g) to be in the box, we need

$$\begin{cases} |c| \le 1, \\ |a+c| \le 1, \\ |b| \le \big|1-|a+c|\big|, \end{cases}$$

and if $|b| \le 2|a|$, also

$$\begin{cases} |b| \le \sqrt{a(1+c)}\,, & a > 0, \\ |b| \le \sqrt{-a(1-c)}\,, & a < 0. \end{cases}$$

Now that the constraints on a, b, and c for $g(x) = ax^2 + bx + c$ to stay within the $[-1, 1] \times [-1, 1]$ box have been established, computing the actual norm of the composition operator will be the next step.

With $0 \ne f \in T_1$,

$$\begin{aligned} \|C_g\|^2_{T_1 \to T_2} &= \sup \frac{\|C_g(f)\|^2}{\|f\|^2} = \max \frac{\displaystyle\int_{-1}^{1} [c_0 + c_1(ax^2+bx+c)]^2 \frac{1}{\sqrt{1-x^2}}\,dx}{\displaystyle\int_{-1}^{1} [c_0 + x_1 x]^2 \frac{1}{\sqrt{1-x^2}}\,dx} \\ &= \max \frac{3a^2c_1^2 + 8ac_1(cc_1+c_0) + 4(b^2c_1 + 2(c^2c_1^2 + 2cc_0c_1 + c_0^2))}{4(2c_0^2+c_1^2)}. \end{aligned}$$

The quantity being maximized in the last line will be called the norm quotient. After finding partial derivatives the following critical points are found:

$$c_1 = \frac{c_0(3a^4+8ac+4b^2+8c-4)}{4a+8c} \pm \frac{c_0\sqrt{\begin{array}{r} 9a^4+48a^3c+8a^2(3b^2+14c+1)+64ac(b^2+2c^2+1) \\ +16b^4+32b^2(2c^2-1)+64c^4+64c^2+16 \end{array}}}{4a+8c}.$$

To attain a formula, the value for c_1 is substituted into the norm quotient. After expanding and simplifying using a computer algebra system, the following formula is obtained:

$$\|C_g\|^2_{T_1 \to T_2} = \frac{\sqrt{\begin{array}{r} 9a^4+48a^3c+8a^2(3b^2+14c+1)+64ac(b^2+2c^2+1) \\ +16b^4+32b^2(2c^2-1)+64c^4+64c^2+16 \end{array}}}{8} + \frac{b^2}{2} + c^2 + ac + \frac{3a^2+4}{8}.$$

We can raise the companion question as to the norm of the operator when viewing it successively $T_2 \to T_4$, $T_3 \to T_6$, $\ldots$, $T_n \to T_{2n}$, $\ldots$, as $n \to \infty$.

4.3. $T_1 \to T_3$. Following the template of the previous section, we discover the norm of a composition operator in the $T_1 \to T_3$ subspace:

$$\begin{aligned}\|C_g\|^2_{T_1\to T_3} = {} & \frac{5a^2+12ac+2(3b^2+8bd+4(c^2+2d^2+1))}{16} \\ & +\frac{1}{16}\Big(25a^4+120a^3c+4a\big(15b^2+40bd+4(14c^2+5(2d^2-1))\big) \\ & \quad +48ac(3b^2+8bd+4(c^2+2d^2-1)) \\ & \quad +4\big(9b^4+48b^3d+8b^2(3c^2+14d^2+1)+64bd(c^2+2d^2+1)\big) \\ & \qquad +64(c^4+2c^2(2d^2-1)+4d^4+4d^2+1)\Big)^{1/2}.\end{aligned}$$

Likewise we can ask about norms viewing the operator $T_2 \to T_6$, $T_3 \to T_9, \dots,$ $T_n \to T_{3n}, \dots,$ as $n \to \infty$.

There are more potentially interesting questions open to us as the number of coefficients increases. Finding a norm formula for general symbols is the ultimate goal.

5. Isometries

Definition 4. An *isometry* is a bijective map $f : X \to Y$ between two normed spaces that preserves lengths, that is, $\|f(x)\|_Y = \|x\|_X$, where $\|\cdot\|_X$ and $\|\cdot\|_Y$ are the norms associated with the spaces X and Y.

For our purposes, the isometry will be viewed as acting between two finite-dimensional subspaces T_n and T_m of T. More precisely, C_g will act as an isometry when the norm of the input $\|f\|$ equals the norm of the output $\|C_g(f)\|$, and thus automatically $\|C_g\| = 1$.

Theorem 5. *When the symbol $g(x)$ is a normalized Chebyshev polynomial in the subspace T_n, $n > 0$, the induced operator $C_g : T_1 \to T_n$ is an isometry.*

Proof. Let $v_n(x)$ be the Chebyshev polynomial (so the symbol $g(x)$ is $v_n(x)$, the polynomial of degree n) and $f(x)$ any nonzero linear polynomial. Then

$$\begin{aligned}\|C_{v_n}\|^2 &= \sup\frac{\|C_{v_n}(f)\|^2}{\|f\|^2} = \max\frac{\|f\circ v_n\|^2}{\|f\|^2} = \max\frac{\int_{-1}^{1}[c_0+c_1(v_n(x))]^2\frac{1}{\sqrt{1-x^2}}\,dx}{\int_{-1}^{1}[c_0+c_1x]^2\frac{1}{\sqrt{1-x^2}}\,dx} \\ &= \max\frac{\int_{-1}^{1}[c_0^2+2c_0c_1v_n(x)+c_1^2(v_n(x))^2]\frac{1}{\sqrt{1-x^2}}\,dx}{\int_{-1}^{1}[c_0^2+2c_0c_1x+c_1^2x^2]\frac{1}{\sqrt{1-x^2}}\,dx} \\ &= \max\frac{\int_{-1}^{1}c_0^2\frac{1}{\sqrt{1-x^2}}\,dx+2c_0c_1\int_{-1}^{1}v_n(x)\frac{1}{\sqrt{1-x^2}}\,dx+c_1^2\int_{-1}^{1}(v_n(x))^2\frac{1}{\sqrt{1-x^2}}\,dx}{\frac{\pi}{2}[c_1^2+2c_0^2]}.\end{aligned}$$

But $\int_{-1}^{1} v_n(x)\frac{1}{\sqrt{1-x^2}}\,dx = 0$ when $0 \neq n$. So,

$$\|C_{v_n}\|^2 = \max \frac{\pi c_0^2 + \frac{1}{2}\pi c_1^2}{\frac{1}{2}\pi(c_1^2 + 2c_0^2)} = 1.$$

This calculation shows that the norm quotient is independent of the choice of c_0 and c_1, and thus $\|C_{v_n}(f)\| = \|f\|$. □

We were encouraged to look for more isometric composition operators and began with those acting from $T_1 \to T_3$. With the previous theorem in mind, the first general form of the symbol considered was $g(x) = ax^3 + cx$, since that is the form of the Chebyshev polynomial of degree three. With $0 \neq f \in T_1$,

$$\frac{\|C_{ax^3+cx}(f)\|^2}{\|f\|^2} = \frac{\|f \circ (ax^3+cx)\|^2}{\|f\|^2} = \frac{\int_{-1}^{1}[c_0 + c_1(ax^3+cx)]^2 \frac{1}{\sqrt{1-x^2}}\,dx}{\int_{-1}^{1}[c_0 + c_1 x]^2 \frac{1}{\sqrt{1-x^2}}\,dx}$$
$$= \frac{5a^2c_1^2 + 12acc_1^2 + 8(c^2c_1^2 + 2c_0^2)}{8(c_1^2 + 2c_0^2)}.$$

To find what the values of a and c must be for C_g to act as an isometry, we forced this norm quotient to be 1. This leads to $25a^2 + 40c^2 = 40$, the equation of an ellipse, which amounts to requiring

$$c = \frac{-3a \pm \sqrt{16 - a^2}}{4}. \tag{5-1}$$

We can conclude that any symbol of form $g(x) = ax^3 + cx$, with c as in (5-1), will act as an isometry in $T_1 \to T_3$, assuming the symbol is admissible.

We continued this technique with the symbols $g(x) = ax^3 + d$, $g(x) = ax^3 + bx^2 + cx$, $g(x) = ax^3 + cx + d$, and $g(x) = ax^3 + bx^2 + d$, as well as other examples. However, each time we found dependence on c_0 and c_1, and thus C_g could not act isometrically on the whole subspace.

With the success of finding the isometry family for the symbol $g(x) = ax^3 + cx$ in comparison to the lack of success with the examples explored above, we wondered if this was the only form of g that could act as an isometry. We backtracked to the general form $g(x) = ax^3 + bx^2 + cx + d$ and constructed some test functions.

From Section 4.3 where we found $\|C_g\|_{T_1 \to T_3}$, we began testing with various functions $f \in T_1$ to find necessary conditions on the form of g for C_g to be an isometry. The first such functions are $f_1(x) = x + 1$ and $f_2(x) = x - 1$. Recall that

$$\frac{\|C_g(f)\|^2}{\|f\|^2} = \frac{5a^2c_1^2 + 12acc_1^2 + 2(3b^2c_1^2 + 8bc_1(dc_1 + c_0) + 4(c^2c_1^2 + 2(d^2c_1^2 + 2dc_0c_1 + c_0^2)))}{8(2c_0^2 + c_1^2)}.$$

We see that

$$\frac{\|C_g(x+1)\|_{T_3}^2}{\|x+1\|_{T_1}^2} = \frac{5a^2+12ac+2(3b^2+8b(d+1)+4(c^2+2(d^2+2d+1)))}{24},$$
$$\frac{\|C_g(x-1)\|_{T_3}^2}{\|x-1\|_{T_1}^2} = \frac{5a^2+12ac+2(3b^2+8b(d-1)+4(c^2+2(d^2-2d+1)))}{24}.$$

For C_g to qualify as an isometry, each norm quotient must be 1 and thus the difference of the two norm quotients must be 0:

$$0 = \|C_g(x+1)\|_{T_3}^2 - \|C_g(x+1)\|_{T_3}^2 = \frac{4b-8d}{3}.$$

Thus $\frac{4b-8d}{3} = 0$, or $b = 2d$. Next we used $f_3(x) = x + \frac{1}{2}$, and $f_4(x) = x - \frac{1}{4}$. The norm quotients for $f_3(x)$ and $f_4(x)$ with $b = 2d$ are

$$\frac{\left\|C_g\left(x+\frac{1}{2}\right)\right\|_{T_3}^2}{\|x+\frac{1}{2}\|_{T_1}^2} = \frac{5a^2+12ac+4(2c^2+18d^2+8d+1)}{12},$$
$$\frac{\left\|C_g\left(x-\frac{1}{4}\right)\right\|_{T_3}^2}{\|x-\frac{1}{4}\|_{T_1}^2} = \frac{5a^2+12ac+8c^2+72d^2-16d}{9}.$$

We set each new norm quotient equal to 1 and solved for d. In solving

$$\|C_g\left(x+\tfrac{1}{2}\right)\|_{T_3}^2 = \|x+\tfrac{1}{2}\|_{T_1}^2$$

for d, we find

$$d_1^{\pm} = -\frac{2}{9} \pm \frac{\sqrt{2}\sqrt{-45a^2-4(27ac+2(9c^2-13))}}{36}. \tag{5-2}$$

While for $\left\|C_g\left(x-\frac{1}{4}\right)\right\|_{T_3}^2 = \left\|x-\frac{1}{4}\right\|_{T_1}^2$ we have

$$d_2^{\pm} = \frac{1}{9} \pm \frac{\sqrt{2}\sqrt{-45a^2-4(27ac+2(9c^2-10))}}{36}. \tag{5-3}$$

We are seeking coefficients for g that will induce an isometry. Thus a single choice for d must serve for all test functions. The expressions for d above must therefore be equal and their difference 0. We examined each of the four pairings and present the most illuminating one, which turns out to be $d_1^+ = d_2^-$. Rearranging terms gives

$$\frac{\sqrt{2}\sqrt{-45a^2-4(27ac+2(9c^2-13))}}{36} + \frac{\sqrt{2}\sqrt{-45a^2-4(27ac+2(9c^2-10))}}{36} = \frac{1}{3},$$

which can be solved for c by squaring both sides, isolating the product of radicals on one side of the equation, squaring again, and simplifying. The solution is

$$c = \frac{-3a \pm \sqrt{16 - a^2}}{4},$$

which is (5-1). When we substitute either expression for c into (5-2), we find $d_1 = 0,\ -\frac{4}{9}$. Likewise, using (5-3), we get $d_2 = 0,\ \frac{2}{9}$.

Thus, for C_g to be an isometry d must be 0, and since we already discovered $b = 2d$, necessarily $b = 0$. Therefore for a symbol g to induce an isometry is for g to be of the form $g(x) = ax^3 + cx$. This was the form first investigated at the start of this subsection where we discovered the only family of the form $g(x) = ax^3 + cx$ is when c is as in (5-1). Therefore, the *only* family that enables g to induce as an isometry from T_1 to T_3 is the one described above.

Now that we have specified the form of the symbol g, we must determine the constraints on g to be in the $[-1, 1] \times [-1, 1]$ box. That is, we must make sure there are symbols satisfying the condition for C_g to be an isometry and which are themselves admissible.

There are two basic criteria for $g(x) = ax^3 + cx$ to stay in the box:

- The values of $ax^3 + cx$ at the endpoints of the interval $[-1, 1]$ cannot be above or below the corners of the $[-1, 1] \times [-1, 1]$ box (that is, the graph enters the left side of the box and exits on the right, not the top or the bottom).
- If the local extrema of $ax^3 + cx$ (if any) occur for $x \in [-1, 1]$, they cannot be above the upper boundary or below the lower boundary of the $[-1, 1] \times [-1, 1]$ box (that is, the graph does not penetrate the top or bottom of the box).

The endpoints will be considered first; we require at the left endpoint $-1 \leq -(a + c) \leq 1$, and at the right endpoint $-1 \leq a + c \leq 1$. These requirements can be summarized by $|a + c| \leq 1$. Graphically (with a as the horizontal axis and c as the vertical axis), this is represented by a "strip" in the ac plane, between the lines $c = 1 - a$ and $c = -1 - a$. This is seen in Figure 3.

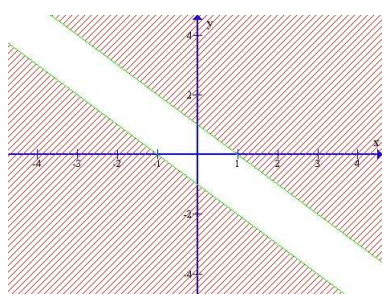


Figure 3. Strip of admissibility. The slanted shading indicates regions of inadmissibility.

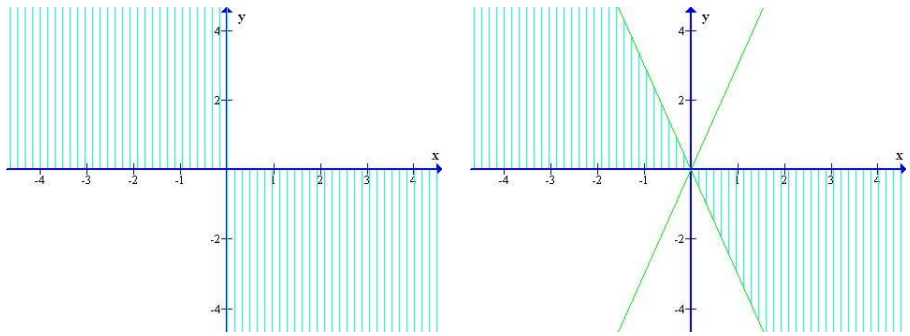


Figure 4. Regions of admissibility, when a and c have the same sign (left), and when a and c have opposite signs, $|c| > 3|a|$ (right). The vertical shading indicates further investigation is needed.

Next, we consider the local extrema of $g(x) = ax^3 + cx$ which occur when $x^2 = -c/3a$. Two cases arise: when a and c have the same sign, and when a and c have opposite signs. We consider the case when they have the same sign first.

If a and c have the same sign, $g'(x)$ will have no real roots and thus g has no local extrema. Thus, in addition to the strip described above, quadrants I ($a, c \geq 0$) and III ($a, c \leq 0$) are part of the admissible region; see Figure 4, left.

If a and c have opposite signs, g has local extrema. The roots of g' are outside the box if $-c/3a > 1$. This corresponds to the cones bounded by $c = \pm 3a$ in the ac plane; see Figure 4, right.

On the other hand, if $-c/3a \leq 1$, or $|c| \leq 3|a|$, then $\left|g\left(\pm\sqrt{-c/3a}\right)\right| \leq 1$ must be true. Evaluation at the critical points requires that

$$-1 \leq g\left(\pm\sqrt{-c/3a}\,\right) \leq 1.$$

Without loss of generality, consider the case of c negative. Now we are considering

$$-1 \leq a\left(\sqrt{\frac{-c}{3a}}\right)^3 + c\sqrt{\frac{-c}{3a}} \leq 1 \quad \Rightarrow \quad -1 \leq a\left(\frac{-c}{3a}\right)^{3/2} + c\sqrt{\frac{-c}{3a}} \leq 1 \quad \Rightarrow$$

$$-1 \leq a\frac{\sqrt{-c}}{3a}\sqrt{\frac{-c}{3a}} + c\sqrt{\frac{-c}{3a}} \leq 1 \quad \Rightarrow \quad -1 \leq \frac{1}{3}|c|\frac{\sqrt{-c}}{\sqrt{3a}} + c\frac{\sqrt{-c}}{\sqrt{3a}} \leq 1 \quad \Rightarrow$$

$$-\sqrt{3a} \leq \sqrt{-c}\left(-\tfrac{1}{3}c + c\right) \leq \sqrt{3a} \quad \Rightarrow \quad -\sqrt{3a} \leq \tfrac{2}{3}c\sqrt{-c} \leq \sqrt{3a}.$$

Next, we use this relation to solve for c:

$$-\sqrt{3a} \leq \tfrac{2}{3}(-c)^{3/2} \leq \sqrt{3a} \quad \Rightarrow \quad -\tfrac{3}{2}\sqrt{3a} \leq (-c)^{3/2} \leq \tfrac{3}{2}\sqrt{3a} \quad \Rightarrow$$

$$-c \leq (\tfrac{3}{2}\sqrt{3a})^{2/3} \quad \Rightarrow \quad -c \leq 3\left(\tfrac{1}{4}\right)^{1/3}(\sqrt{a})^{2/3} \quad \Rightarrow \quad c \geq -3(a/4)^{1/3}.$$

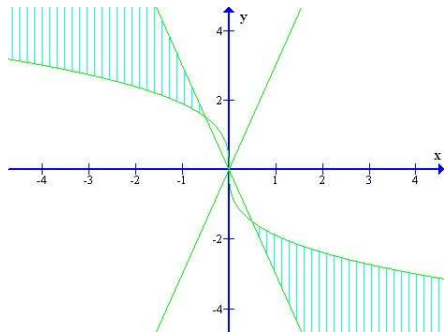


Figure 5. Cubic with cones, a and c of opposite signs, $|c| \leq 3|a|$. The vertical shading indicates regions of inadmissibility.

Similarly, when $c \geq 0$, we have

$$c \leq 3(-a/4)^{1/3}.$$

The corresponding region is shown in Figure 5. Keep in mind that the cubic is relevant only outside the cones.

The combinations of the previous four regions is seen in Figure 6, which represents the total admissible area for $g(x)$.

Lastly, we consider the graph of c in terms of a from the condition to be an isometry (5-1). Graphing both the positive and negative roots reveals an ellipse, seen in Figure 7. This ellipse represents all the possibilities of $g(x)$ within the family. The parts that lie in the admissible region are shown with heavy printing. The positive and negative Chebyshev polynomials are at the extreme ends the major axis of the ellipse, sitting as isolated points. This means there exists a nontrivial family of isometries acting from T_1 to T_3, represented by the continuous arc of the ellipse in the admissible region, along with the isometries identified by Theorem 5, represented by isolated points at the major vertices of the ellipse. Not only have these been identified, but these are the totality of all possible composition isometries between these subspaces. This is a very satisfying answer to the question.

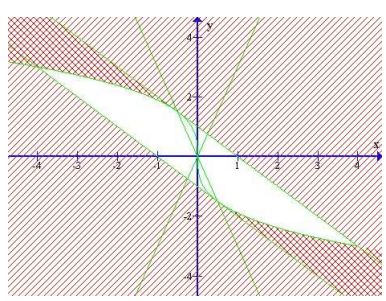


Figure 6. Total region of admissibility for $g(x)$ (unshaded).

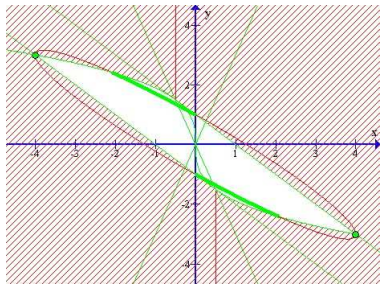


Figure 7. Admissible isometry-inducing symbols.

6. Conclusion

Some simple formulas for computing norms of composition operators on finite-dimensional subspaces of the Chebyshev space show some direction for future lines of investigation. We raised the more specific question of isometric composition operators, especially in the case of $T_1 \to T_3$. This revealed a family of operators whose symbols' coefficients vary over a continuum and a pair of isolated symbols corresponding to the Chebyshev polynomials of order 3. The geometric connection to a norm question was surprising and pleasing.

References

[Akhiezer and Glazman 1993] N. I. Akhiezer and I. M. Glazman, *Theory of linear operators in Hilbert space*, Dover, New York, 1993. MR 94i:47001 Zbl 0874.47001

[Cowen and MacCluer 1995] C. C. Cowen and B. D. MacCluer, *Composition operators on spaces of analytic functions*, CRC Press, Boca Raton, FL, 1995. MR 97i:47056 Zbl 0873.47017

[Doperak 2006] J. Doperak, *Norms of composition operators on the space of hermite polynomials*, Summer Fellows undergraduate thesis, Ursinus College, 2006.

[Dunford and Schwartz 1958] N. Dunford and J. T. Schwartz, *Linear operators*, vol. 1, Wiley, New York, 1958. MR 90g:47001a Zbl 0084.10402

[Gareau 2005] M. Gareau, *Composition operators acting on real-valued functions*, Summer Fellows undergraduate thesis, Ursinus College, 2005.

[Horn and Johnson 1985] R. A. Horn and C. R. Johnson, *Matrix analysis*, Cambridge University Press, Cambridge, 1985. MR 87e:15001 Zbl 0576.15001

[Kunaszuk 2006] J. Kunaszuk, *Norms of composition operators acting on the legendre polynomials*, Summer Fellows undergraduate thesis, Ursinus College, 2006.

[Lebedev 1972] N. N. Lebedev, *Special functions and their applications*, revised ed., Dover, New York, 1972. MR 50 #2568 Zbl 0271.33001

[Potter 2007] A. Potter, *Composition operators acting on the Chebyshev space*, Summer Fellows undergraduate thesis, Ursinus College, 2007.

[Reed and Simon 1980] M. Reed and B. Simon, *Methods of modern mathematical physics, I: Functional analysis*, 2nd ed., Academic Press, New York, 1980. MR 85e:46002 Zbl 0459.46001

Received: 2008-11-10 Revised: 2009-06-25 Accepted: 2009-08-12

tgoebeler@ursinus.edu *Department of Mathematics and Computer Science, Ursinus College, Collegeville, PA 19426, United States*

ashpot17@gmail.com *Department of Mathematics, Perkiomen Valley High School, 509 Gravel Pike, Collegeville, PA 19426, United States*

INVOLVE 2:5(2009)

Markov partitions for hyperbolic sets

Todd Fisher and Himal Rathnakumara

(Communicated by Kenneth S. Berenhaut)

We show that if f is a diffeomorphism of a manifold to itself, Λ is a mixing (or transitive) hyperbolic set, and V is a neighborhood of Λ, then there exists a mixing (or transitive) hyperbolic set $\tilde{\Lambda}$ with a Markov partition such that $\Lambda \subset \tilde{\Lambda} \subset V$. We also show that in the topologically mixing case the set $\tilde{\Lambda}$ will have a unique measure of maximal entropy.

1. Introduction

A dynamical system consists of a space and a rule to dictate the evolution of the points in the space. In particular, a discrete dynamical system (X, f) consists of a topological space X and a map $f : X \to X$. The nth iterate of f, denoted f^n, is defined as the map f composed n times, where $n \in \mathbb{N}$. If f is a bijection, then its inverse f^{-1} exists and we can form the nth iterate of f^{-1} by composition, $f^{-n} : X \to X$.

We assume in this paper that the maps associated with dynamical systems are homeomorphisms so that f^{-1} exists and f^{-n} is well-defined. In the study of dynamical systems it is important to look at the overall effect of the rule for individual points in the space. In this analysis we look at orbits of points in the space where the *orbit of a point* $x \in X$ is defined as

$$\mathcal{O}(x) = \{f^n(x) \in X : n \in \mathbb{Z}\}.$$

Throughout the paper we let M be a compact, smooth, boundaryless manifold and denote the set of diffeomorphisms from M to itself by $\mathrm{Diff}(M)$. A set X is *invariant* under f if $f(X) = X$. Invariant sets play an important role in dynamical systems and often allow one to decompose a space into invariant "indecomposable" sets. A compact set $\Lambda \subset M$ that is invariant under $f \in \mathrm{Diff}(M)$ is a *hyperbolic set*

MSC2000: 37A35, 37D05, 37D15.
Keywords: Markov partitions, hyperbolic, entropy, specification, finitely presented.

if there exists a splitting of the tangent space $T_\Lambda f = \mathbb{E}^u \oplus \mathbb{E}^s$ and positive constants $C \geq 1$ and $\lambda < 1$ such that for any point $x \in \Lambda$ and any $n \in \mathbb{N}$ we have

$$\|Df_x^n v\| \leq C\lambda^n \|v\| \quad \text{for } v \in E_x^s,$$
$$\|Df_x^{-n} v\| \leq C\lambda^n \|v\| \quad \text{for } v \in E_x^u.$$

Hyperbolic sets were introduced by Smale and Anosov in the 1960s. The compactness of the manifold together with the expansion and contraction in the tangent bundle allows for complicated and interesting orbit structures. Additionally, hyperbolic sets are structurally stable, or in other words, the dynamics of a hyperbolic set are preserved under perturbations.

One of the main tools in studying hyperbolic sets is the use of a Markov partition introduced by Adler and Weiss for hyperbolic toral automorphisms of the 2-torus [Adler and Weiss 1967]. Markov partitions are defined in Section 2. It was shown in [Fisher 2006] that if $f \in \mathrm{Diff}(M)$, Λ is a hyperbolic set for f, and V is a neighborhood of f, then there exists a hyperbolic set $\tilde{\Lambda}$ for f such that $\Lambda \subset \tilde{\Lambda} \subset V$ and $\tilde{\Lambda}$ has a Markov partition. For a Markov partition there is a canonically associated symbolic space called a subshift of finite type. (For the definition of a subshift of finite type see Section 2.)

Often one is interested in studying hyperbolic sets that satisfy additional properties. Two such properties are topological mixing and transitivity. A dynamical system (X, f) is *topologically mixing* if for any open sets U and V there exists some $N \in \mathbb{N}$ such that $f^n(U) \cap V \neq \varnothing$ for all $n \geq N$. A dynamical system (X, f) is *transitive* if there exists a point $x \in X$ such that the forward orbit of x,

$$\mathcal{O}^+(x) = \{f^n(x) : n \in \mathbb{N}\},$$

is dense in X. A standard result about transitivity is the following: if X is a locally compact Hausdorff space, then (X, f) is *topologically transitive* if and only if for any open sets U and V in X there exists some $n \in \mathbb{N}$ such that $f^n(U) \cap V \neq \varnothing$ [Brin and Stuck 2002, page 31].

The main result of the present work is that we can strengthen the result on Markov partitions in [Fisher 2006] with respect to topological mixing and transitivity.

Theorem 1.1. *If Λ is a topologically mixing hyperbolic set for $f \in \mathrm{Diff}(M)$ and V is a neighborhood of Λ, then there exists a hyperbolic set $\tilde{\Lambda}$ for f containing Λ and contained in V such that $(\tilde{\Lambda}, f)$ has a Markov partition coming from an associated mixing subshift of finite type. Furthermore, if Λ is transitive, then $(\tilde{\Lambda}, f)$ has a Markov partition coming from an associated transitive subshift of finite type.*

We note that a standard result is that if the subshift of finite type is mixing (or transitive) and associated to a Markov partition for a hyperbolic set, then the

hyperbolic set is mixing (or transitive). Bowen [1974] provided a nice connection between mixing hyperbolic sets and the entropy for the system. The topological entropy of a dynamical system, denoted $h_{\text{top}}(f)$, is a number that, in a certain manner, measures the topological complexity of the system. Whereas, the measure theoretic entropy, denoted $h_\mu(f)$, of a dynamical system is a number that, in some manner, measures the complexity of the system as seen by the measure μ.

A measure μ is *invariant* for the dynamical system (X, f) if

$$\mu(f^{-1}(A)) = \mu(A)$$

for all measurable sets A. We denote the set of invariant Borel probability measures as $\mathcal{M}(f)$. If X is a compact metrizable space and f is continuous, then we know that $\mathcal{M}(f) \neq \varnothing$ [Katok and Hasselblatt 1995, page 135]. The variational principle says that if f is a homeomorphism of a compact metrizable space, then $h_{\text{top}}(f) = \sup_{\mu\in\mathcal{M}(f)} h_\mu(f)$ [Katok and Hasselblatt 1995, page 181]. A measure $\mu \in \mathcal{M}(f)$ such that $h_{\text{top}}(f) = h_\mu(f)$ is a *measure of maximal entropy*. If there is a unique measure of maximal entropy, then f is called *intrinsically ergodic*. From Theorem 1.1 and Bowen's results we are then able to show the following.

Corollary 1.2. *If Λ is a topologically mixing hyperbolic set and V is a neighborhood of Λ, then there exists a hyperbolic set $\tilde{\Lambda}$ containing Λ and contained in V such that $\tilde{\Lambda}$ is intrinsically ergodic with respect to f.*

2. Background

As we will be looking at subshifts of finite type we first review some definitions and facts about subshifts of finite type. Let $A = [a_{ij}]$ be an $n \times n$ matrix with entries of zeros and ones such that there is one or more one in each row and column. Such a matrix is called an *adjacency matrix*. Let $\mathcal{A}_n = \{1, ..., n\}$ and call a transition from i to j to be *admissible* for A if $a_{ij} = 1$. Define

$$\Sigma_A = \big\{\omega = (\omega_k)_{k\in\mathbb{Z}} \mid \omega_k \in \mathcal{A}_n \text{ and } \omega_k\omega_{k+1} \text{ is admissible for all } k \in \mathbb{Z}\big\}.$$

The map on Σ_A defined by $\sigma(\omega) = \omega'$ where $\omega'_j = \omega_{j+1}$ is called the *shift map*. The *subshift of finite type* is the space (Σ_A, σ) together with the product metric on Σ_A. A matrix A is *positive* if each entry is positive. A matrix A is *primitive* if there is some power $N \in \mathbb{N}$ such that A^N is positive.

If a matrix A is primitive, then the subshift of finite type associated with A is topologically mixing. Furthermore, a subshift of finite type associated with an $M \times M$ matrix A is transitive if and only if for each i, j ($1 \le i, j \le M$) there exists some $n \in \mathbb{N}$ such that $a^n_{ij} > 0$ [Robinson 1999, page 80].

A *topological semiconjugacy* between a pair of dynamical systems (X, f) and (Y, g) exists if there is a continuous surjective map $h : X \to Y$ such that

$$h \circ f = g \circ h.$$

The space (Y, g) is called a *factor* of (X, f), and (X, f) is called an *extension* of (Y, g).

A dynamical system (X, f) where X is a compact metric space and f is a homeomorphism is *expansive* if there exists a constant $c > 0$ such that for all $x, y \in X$ if $d(f^n(x), f^n(y)) < c$ for all $n \in \mathbb{Z}$, then $x = y$.

We now review some facts about expansive and finitely presented dynamical systems. For $\varepsilon > 0$ and $x \in X$ the *ε-stable set* is

$$W^s_\varepsilon(x) = \big\{y \in X \mid d(f^n(x), f^n(y)) < \varepsilon \text{ for all } n \geq 0\big\},$$

and the *ε-unstable set* is

$$W^u_\varepsilon(x) = \big\{y \in X \mid d(f^{-n}(x), f^{-n}(y)) < \varepsilon \text{ for all } n \geq 0\big\}.$$

For $x \in X$ and $f : X \to X$, an expansive homeomorphism, the *stable set* is

$$W^s(x) = \big\{y \in X \mid \lim_{n\to\infty} d(f^n(x), f^n(y)) = 0\big\}$$

and the *unstable set* is

$$W^u(x) = \big\{y \in X \mid \lim_{n\to\infty} d(f^{-n}(x), f^{-n}(y)) = 0\big\}.$$

Let (Y, f) be expansive and fix $\varepsilon < c/2$, where c is an expansive constant of (Y, f). Following [Fried 1987] we define

$$D_\varepsilon = \big\{(x, y) \in Y \times Y \mid W^s_\varepsilon(x) \text{ meets } W^u_\varepsilon(y)\big\}$$

and $[\cdot, \cdot] : D_\varepsilon \to Y$ so that $[x, y] = W^s_\varepsilon(x) \cap W^u_\varepsilon(y)$. It follows that $[\cdot, \cdot]$ is continuous.

Definition 2.1. A *rectangle* is a closed set $R \subset Y$ such that $R \times R \subset D_\varepsilon$.

For R a rectangle and $x \in R$, denote the stable and unstable sets of x in R, respectively, as

$$W^s(x, R) = R \cap W^s_\varepsilon(x), \quad W^u(x, R) = R \cap W^u_\varepsilon(x).$$

A rectangle R is *proper* if $R = \overline{\mathring{R}}$, where $\mathring{R}$ denotes the interior of R.

Definition 2.2. Let (Y, f) be expansive with constant $c > 0$ and $0 < \varepsilon < c/2$. A finite cover $\mathcal{R}$ of Y by proper rectangles with diameter$(R) < \varepsilon$ for any $R \in \mathcal{R}$ is a *Markov partition* if $R_i, R_j \in \mathcal{R}$, $x \in \mathring{R}_i$, and $f(x) \in \mathring{R}_j$, then

$$f(W^s(x, R_i)) \subset R_j, \quad f^{-1}\big(W^u(f(x), R_j)\big) \subset R_i, \quad \text{and} \quad \mathring{R}_i \cap \mathring{R}_j = \varnothing \text{ if } i \neq j.$$

For a Markov partition $\mathcal{R}$ of a system (X, f) we define the adjacency matrix A such that $a_{ij} = 1$ if $f(\mathring{R}_i) \cap \mathring{R}_j \neq \varnothing$. The subshift of finite type (Σ_A, σ) is said to be *associated* with $\mathcal{R}$ and there is a canonical semiconjugacy h from (Σ_A, σ) to (X, f).

Fried [1987] defined *finitely presented systems* as expansive homeomorphisms of a compact space that are factors of a subshift of finite type. In the same paper he shows that any finitely presented dynamical system has a Markov partition.

Remark 2.3. For $f \in \mathrm{Diff}(M)$ and Λ a hyperbolic set for f, the system $(\Lambda, f|_\Lambda)$ is expansive. Furthermore, any subshift of finite type is expansive. Also, for a hyperbolic set Λ for a diffeomorphism and $x \in \Lambda$, the sets $W^s(x)$ and $W^u(x)$ are injectively immersed submanifolds of Euclidean spaces.

3. Results

Proof of Theorem 1.1. Before proceeding to the proof of Theorem 1.1 we first review some facts about shadowing for hyperbolic sets. A sequence $\{x_k\}_a^b$ is an *ε-chain* if $d(f(x_k), x_{k+1}) < \varepsilon$ for all k where $-\infty \le a < b \le \infty$. A point y *δ-shadows* an ε-chain $\{x_k\}$ if $d(f^k(y), x_k) < \delta$ for all k. We next state the Shadowing Theorem [Brin and Stuck 2002, page 113].

Theorem 3.1 (Shadowing Theorem). *Let M be a Riemannian manifold, d the natural distance function, f a diffeomorphism of M to itself, and Λ a hyperbolic set for f. Then for every $\delta > 0$ there exists an $\varepsilon > 0$ such that if $\{x_n\}$ is an ε-chain of f and $d(x_k, \Lambda) < \varepsilon$ for all k, then there is some $y \in \bigcup_{x\in\Lambda} B_\varepsilon(x)$ that δ-shadows the ε-chain $\{x_k\}$.*

Proof of Theorem 1.1. We first assume that Λ is topologically mixing. To prove the theorem it will be sufficient to show that the subshift of finite type constructed in [Fisher 2006] giving the hyperbolic set $\tilde{\Lambda}$ will be topologically mixing.

Let U be a neighborhood of Λ. A standard result for hyperbolic sets states that there is a neighborhood V of Λ such that $\overline{V} \subset U$ and $\Lambda_V = \bigcap_{n\in\mathbb{Z}} f^n(\overline{V})$ is hyperbolic [Katok and Hasselblatt 1995, page 271]. Let $d(\cdot,\cdot)$ be an adapted metric on Λ_V. Note that this can be extended continuously to a neighborhood $V' \subset U$ of Λ_V.

Fix $\eta > 0$ and $\delta \le \eta$ such that for any two points $x, y \in \Lambda_V$, if $d(x, y) < \delta$ then

$$f^{-1}\big(W^s_\eta(f(x))\big) \cap f\big(W^u_\eta(f^{-1}(y))\big) = W^s_\eta(x) \cap W^u_\eta(y)$$

consists of one point, and the set $\bigcup_{x\in\Lambda} B_{2\eta}(x)$ is contained in $V \cap V'$; see [Fisher 2006] for an argument explaining the existence of η and δ. Fix $0 < \varepsilon \le \delta/2$ as in the conclusion of the Shadowing Theorem so that every ε-orbit is $\delta/2$-shadowed and contained in $V \cap V'$.

Let $\nu < \varepsilon/2$ such that $d(f(x), f(y)) < \varepsilon/2$ and $d(f^{-1}(x), f^{-1}(y)) < \varepsilon/2$ when $d(x, y) < \nu$ for any $x, y \in \Lambda_V$. Let $\{p_i\}_{i=1}^N$ be a ν-dense set of points in Λ and let the adjacency matrix A be defined by

$$a_{ij} = \begin{cases} 1 & \text{if } d(f(p_i), p_j) < \varepsilon, \\ 0 & \text{if } d(f(p_i), p_j) \geq \varepsilon. \end{cases}$$

Let (Σ_A, σ) be the subshift of finite type associated with A. Then we know there exists a hyperbolic set $\tilde{\Lambda}$ contained in $\bar{V}$ [Fisher 2006] such that $\tilde{\Lambda} \subset \Lambda_V$, that contains Λ and there exists a semiconjugacy $\beta : \Sigma_A \to \tilde{\Lambda}$. To see that $\tilde{\Lambda}$ is topologically mixing it is sufficient to see that Σ_A is topologically mixing.

We now show that Σ_A is topologically mixing by showing that A is primitive. Given sets $B_\nu(p_i)$ and $B_\nu(p_j)$ there exists some N_{ij} such that for all $n \geq N_{ij}$ we have

$$f^n(B_\nu(p_i)) \cap B_\nu(p_j) \neq \varnothing,$$

since Λ is topologically mixing for f. We let $M = \max\{N_{ij}\}$. Then

$$f^n(B_\nu(p_i)) \cap B_\nu(p_j) \neq \varnothing$$

for all $n \geq M$. We now show that this implies that $a_{ij}^n > 0$ for all $n \geq M$. This is equivalent to showing there is a sequence of $(n+1)$-symbols coming from $\mathscr{A}_N$ such that each transition is allowed and the sequence starts with i and ends with j [Robinson 1999, page 76].

Indeed, let $n \geq M$ and $x \in f^n(B_\nu(p_i)) \cap B_\nu(p_j)$. Since

$$\bigcup_{k=1}^{N} f(B_\nu(p_k)) = \Lambda,$$

we know that there exists some p_{i_1} such that $x \in f(B_\nu(p_{i_1}))$. By the definition of ν we know that $d(f(p_{i_1}), p_j) < \varepsilon$ and i_1 to j is an allowed transition in Σ_A. Inductively, let $1 \leq k \leq n-2$ and assume that for each l such that $1 \leq l \leq k$ there is some p_{i_l} such that $f^{-l}(x) \in f(B_\nu(p_{i_l}))$ and

$$\begin{aligned} d(f(p_{i_l}), p_j) &< \varepsilon \quad \text{if } l = 1, \\ d(f(p_{i_l}), p_{i_{l-1}}) &< \varepsilon \quad \text{else.} \end{aligned}$$

Then $i_k i_{k-1} \cdots i_1 j$ is a sequence of $k+1$ symbols in $\mathscr{A}_N$ with allowed transitions and $f^{-l}(x) \in B_\nu(p_{i_l})$ for all $1 \leq l \leq k$. We know that $f^{-(k+1)}(x) \in f(B_\nu(p_{i_{k+1}}))$ for some $i_{k+1} \in \mathscr{A}_N$ and

$$d(f(p_{i_{k+1}}), p_{i_k}) < \varepsilon.$$

Hence, i_{k+1} to i_k is an allowed transition in Σ_A and $i_{k+1} i_k \cdots i_1 j$ is a sequence of $k+2$ symbols in $\mathscr{A}_N$ with allowed transitions. Therefore, there is a sequence

$i_{n-1}i_{n-2}\cdots i_1 j$ of n-terms in $\mathscr{A}_N$ with allowed transitions. Finally, we know that $f^{-n}(x) \in B_\nu(p_i)$ and

$$f^{-(n-1)}(x) \in f(B_\nu(p_i)) \cap B_\nu(p_{i_{n-1}}).$$

So i to i_{n-1} is an allowed transition. Hence, $ii_{n-1}\cdots i_1 j$ is an allowed word in Σ and $a_{ij}^n > 0$. Therefore, A is primitive and Σ_A is topologically mixing.

The proof of the transitive case is similar. Indeed, given sets $B_\nu(p_i)$ and $B_\nu(p_j)$ there exists some N_{ij} such that

$$f^{N_{ij}}(B_\nu(p_i)) \cap B_\nu(p_j) \neq \varnothing.$$

Hence, a similar argument as above shows that $a_{ij}^{N_{ij}} > 0$ and Σ_A is transitive. □

Intrinsic ergodicity for mixing hyperbolic sets. The proof of Corollary 1.2 will use the property of specification. A *specification*, $S = (\tau, P)$, for a dynamical system consists of

(1) a finite collection $\tau = \{I_1, \dots, I_n\}$ of finite intervals $I_i = [a_i, b_i] \subset \mathbb{Z}$, and

(2) a map $P : \bigcup_{i=1}^m I_i \longrightarrow X$ such that $f^{t_2-t_1}(P(t_1)) = P(t_2)$. for all $t_1, t_2 \in I_i \in \tau$.

A specification S is said to be *r-spaced*, where $r \in \mathbb{N}$, if $a_{i+1} > b_i + r$ for all $i \in \{1, \dots, n-1\}$ and the minimal such r is called the spacing of S. A specification $S = (\tau, P)$ provides a way of parametrizing a collection of orbit segments τ of f. We say that S is *ε-shadowed* by $x \in X$ if $d(f^n(x), P(n)) < \varepsilon$ for all $n \in \bigcup_{i=1}^m I_i$.

Definition 3.2. Let X be a compact metric space and $f : X \to X$ a homeomorphism. The dynamical system (X, f) is said to have the *specification property* if for all $\varepsilon > 0$ there exists an $M_\varepsilon \in N$ such that any M_ε-spaced specification S is ε-shadowed by a point of X.

The next result is stated without proof in [Sigmund 1974]. We provide a proof for completeness.

Lemma 3.3. *If (X, f) has the specification property and (Y, g) is a factor of (X, f), then (Y, g) has the specification property.*

Proof. Fix $\varepsilon > 0$ and let d_X and d_Y denote metrics for X and Y, respectively. Let $\varepsilon' > 0$ such that if $d_X(x_1, x_2) < \varepsilon'$, then

$$d_Y(h(x_1), h(x_2)) < \varepsilon,$$

where $x_1, x_2 \in X$. Such an $\varepsilon' > 0$ can always be chosen since h is continuous. Fix $M_{\varepsilon'} \in \mathbb{N}$ such that any $M_{\varepsilon'}$-spaced specification is ε'-shadowed by a point of X and let $M_\varepsilon = M_{\varepsilon'}$.

Let $S = (\tau, P)$ be an M_ε-spaced specification in (Y, g) where $\tau = \{I_1, \dots, I_m\}$ is a collection of M_ε-spaced intervals. Let

$$B = \{y_1, y_2, \dots, y_m\} \subset Y,$$

where $y_i = P(a_i)$ for all $1 \le i \le m$.

Fix $A = \{x_1, \dots, x_m\} \subset X$ such that h restricted to A is a bijection onto B and $h(x_i) = y_i$ for $1 \le i \le m$. The orbit segment for x_i in I_i is given by

$$\left\{f^{a_i}(x_i), \dots, f^{b_i}(x_i)\right\} \quad \text{for } 1 \le i \le m.$$

Define $P_X : \bigcup_{i=1}^m I_i \to X$ such that $P(a_i) = x_i$ for all i such that $1 \le i \le m$.

Now, given that (X, f) has the specification property, we know there exists an ε'-shadowing point x for the specification (τ, P_X) and $h(x) \in Y$. Furthermore,

$$h(f^{a_i}(x_i)) = g^{a_i}(y_i)$$

since h is a semiconjugacy. Hence, $d\big(h(x), h(P(n))\big) < \varepsilon$ for all $n \in \bigcup_{i=1}^m I_i$ and $h(x)$ is an ε-shadowing point for the specification S. □

Theorem 3.4 [Bowen 1974]. *Let X be a compact metric space and f be an expansive homeomorphism with the specification property. Then f is intrinsically ergodic.*

Weiss [1973] showed that a mixing subshift of finite type has the specification property. Since subshifts of finite type are expansive, we know from Theorem 3.4 that a topologically mixing subshift of finite type is intrinsically ergodic.

From Lemma 3.3 we know that a factor of a mixing subshift of finite type is intrinsically ergodic.

Corollary 3.5. *Any topologically mixing finitely presented system is intrinsically ergodic.*

Proof. Let (X, f) be a topologically mixing finitely presented system. To prove the corollary we show there is a topologically mixing subshift of finite type that is an extension of (X, f). Let $\mathcal{R}$ be a Markov partition for (X, f) and A be the adjacency matrix associated with $\mathcal{R}$. Let R_i and R_j be rectangles in $\mathcal{R}$. Since (X, f) is topologically mixing and the rectangles are proper, we know there exists some $N_{ij} \in \mathbb{N}$ such that $f^n(\mathring{R}_i) \cap \mathring{R}_j \neq \varnothing$ for all $n \ge N_{ij}$. Using arguments as in the proof of Theorem 1.1, we know that $a_{ij}^n > 0$ for all $n \ge N_{ij}$. Define $N = \max(N_{ij})$. Then A^n is positive for all $n \ge N$ and the subshift of finite type associated with the Markov partition $\mathcal{R}$ is topologically mixing. □

Proof of Corollary 1.2. Let $f \in \mathrm{Diff}(M)$ for some manifold M, let Λ be a topologically mixing hyperbolic set for f, and V be a neighborhood of Λ. From Theorem 1.1 we know that there exists a topologically mixing hyperbolic set $\tilde{\Lambda}$ contained

in V and containing Λ with a Markov partition. Therefore, $\tilde{\Lambda}$ is finitely presented and from Corollary 3.5 we know that $\tilde{\Lambda}$ is intrinsically ergodic. $\square$

References

[Adler and Weiss 1967] R. L. Adler and B. Weiss, "Entropy, a complete metric invariant for automorphisms of the torus", *Proc. Nat. Acad. Sci. U.S.A.* **57** (1967), 1573–1576. MR 35 #3031 Zbl 0177.08002

[Bowen 1974] R. Bowen, "Some systems with unique equilibrium states", *Math. Systems Theory* **8**:3 (1974), 193–202. MR 53 #3257 Zbl 0299.54031

[Brin and Stuck 2002] M. Brin and G. Stuck, *Introduction to dynamical systems*, Cambridge University Press, Cambridge, 2002. MR 2003m:37001 Zbl 01849967

[Fisher 2006] T. Fisher, "Hyperbolic sets that are not locally maximal", *Ergodic Theory Dynam. Systems* **26**:5 (2006), 1491–1509. MR 2008a:37031 Zbl 1122.37022

[Fried 1987] D. Fried, "Finitely presented dynamical systems", *Ergodic Theory Dynam. Systems* **7**:4 (1987), 489–507. MR 89h:58157 Zbl 0652.54028

[Katok and Hasselblatt 1995] A. Katok and B. Hasselblatt, *Introduction to the modern theory of dynamical systems*, Encyclopedia of Mathematics and its Applications **54**, Cambridge University Press, Cambridge, 1995. MR 96c:58055 Zbl 0878.58020

[Robinson 1999] C. Robinson, *Dynamical systems: stability, symbolic dynamics, and chaos*, 2nd ed., CRC Press, Boca Raton, FL, 1999. MR 2001k:37003 Zbl 0914.58021

[Sigmund 1974] K. Sigmund, "On dynamical systems with the specification property", *Trans. Amer. Math. Soc.* **190** (1974), 285–299. MR 50 #4898 Zbl 0286.28010

[Weiss 1973] B. Weiss, "Subshifts of finite type and sofic systems", *Monatsh. Math.* **77** (1973), 462–474. MR 49 #5308 Zbl 0285.28021

Received: 2009-01-13 Revised: 2009-09-01 Accepted: 2009-10-28

tfisher@math.byu.edu *Department of Mathematics, Brigham Young University, Provo, UT 84602, United States*
http://math.byu.edu/~tfisher/

himal46@gmail.com *Department of Mathematics, Brigham Young University, Provo, UT 84602, United States*

INVOLVE 2:5(2009)

Ineffective perturbations in a planar elastica

Kaitlyn Peterson and Robert Manning

(Communicated by Natalia Hritonenko)

An elastica is a bendable one-dimensional continuum, or idealized elastic rod. If such a rod is subjected to compression while its ends are constrained to remain tangent to a single straight line, buckling can occur: the elastic material gives way at a certain point, snapping to a lower-energy configuration.

The bifurcation diagram for the buckling of a planar elastica under a load λ is made up of a trivial branch of unbuckled configurations for all λ and a sequence of branches of buckled configurations that are connected to the trivial branch at pitchfork bifurcation points. We use several perturbation expansions to determine how this diagram perturbs with the addition of a small intrinsic shape in the elastica, focusing in particular on the effect near the bifurcation points.

We find that for almost all intrinsic shapes $\varepsilon f(s)$, the difference between the buckled solution and the trivial solution is $O(\varepsilon^{1/3})$, but for some ineffective f, this difference is $O(\varepsilon)$, and we find functions $u_j(s)$ so that f is ineffective at bifurcation point number j when $\langle f, u_j\rangle = 0$. These ineffective perturbations have important consequences in numerical simulations, in that the perturbed bifurcation diagram has sharper corners near the former bifurcation points, and there is a higher risk of a numerical simulation inadvertently hopping between branches near these corners.

1. Introduction

Consider a common scenario for symmetry breaking in bifurcation theory. A problem exhibiting some symmetry has a bifurcation diagram with a number of bifurcation points (BPs). The addition of a perturbation breaks this symmetry and removes the BPs, splitting the diagram into separate components. An example of this scenario is shown in Figure 1, in which a pitchfork bifurcation is perturbed to yield two separate branches.

MSC2000: 34B15, 34E10, 34G99, 74K10.

Keywords: elastic rod, intrinsic shape, undetermined-gauges perturbation expansion, pitchfork bifurcations.

This work was supported by NSF grant DMS-0384739.

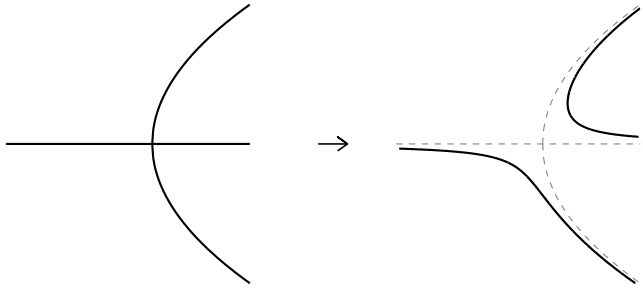


Figure 1. Standard perturbation of a pitchfork bifurcation into two separate branches.

To give a specific example, the buckling of a uniform, isotropic, intrinsically straight rod in three dimensions has a bifurcation diagram containing pitchfork BPs corresponding to the classic Euler buckling modes. There is a natural perturbation to consider for this three-dimensional buckling problem: the presence of intrinsic curvature. Even rods designed to be straight are likely to have small curvature *imperfections*, and these can break the qualitative nature of the bifurcation diagram from the pitchfork structure seen in the intrinsically straight case.

Such elastic rod models have been used to represent the bending and twisting of DNA. For many DNA sequences, the intrinsic shape is nearly straight, but the minimum-energy stacking configurations of consecutive base-pairs do introduce small intrinsic bends that depend on the specific sequence of the DNA. Multiple studies have sought to determine these stacking configurations as a function of sequence [De Santis et al. 1992; Bolshoy et al. 1991; Olson et al. 1998; Dixit et al. 2005], and then derive from these stacking configurations the corresponding intrinsic curvature for a continuum elastic rod [Manning et al. 1996].

These DNA models have seen increasing use in studying a phenomenon called *DNA looping*: the bending and twisting of DNA a few hundreds of base-pairs long in response to prescribed relative positions and orientations of the two ends (these boundary conditions coming from, for example, a bound protein of known structure [Swigon et al. 2006; Goyal et al. 2007; Kahn and Crothers 1998] or laser tweezer experiments [Seol et al. 2007; Marko and Siggia 1995]). Given the wide variety of DNA sequences, and the almost-as-wide variety of parameters for determining local bending from sequence, it would be beneficial to have an automated algorithm to compute the lowest-energy components of the bifurcation diagram given specified choices of DNA sequence, stacking parameters, and boundary conditions. A strong understanding of the splitting of the unperturbed pitchfork diagram for base cases such as buckling or periodic boundary conditions is an important precursor to ensuring that such an automated algorithm finds all relevant components of the diagram.

How typical is the "standard" splitting shown in Figure 1? We analyzed this question for one of the simplest bifurcation problems—the buckling of a planar elastica—under the family of perturbations of (infinitesimal) intrinsic curvature. Like the three-dimensional buckling problem described above, this problem exhibits a sequence of pitchfork BPs. This two-dimensional problem is useful as a base case for the more general three-dimensional DNA looping problem. The formulation of two-dimensional bending is simple enough that closed-form analysis is not too unwieldy; at the same time, it seems reasonable to suppose that the three-dimensional results would be analogous, since, in some sense, three-dimensional bending is composed of two orthogonal directions of two-dimensional bending. The particular buckling boundary conditions we choose yield a classic problem in mechanical engineering, but are not the typical boundary conditions for DNA looping. Still, a likely computational approach to determining configurations obeying arbitrary DNA looping boundary conditions would involve beginning from one of a small number of simple configurations, one of which would be the straight-rod configuration studied here (in addition, to, perhaps, a circle and a semicircle).

Thus, this choice of a simple model problem should allow us to focus on the fundamental questions of which perturbations are atypical and how the bifurcation diagrams of atypical perturbations differ from the typical case. Our main finding is that for these atypical cases, the perturbation of the BP is significantly smaller than in the typical case, leading us to label these perturbations as *ineffective*.

This question is related to the analysis of *unfolding* a pitchfork bifurcation diagram in dynamical systems; see [Glendinning 1994; Iooss and Joseph 1980]. The standard example is to consider the algebraic equation $\lambda x - x^3 = 0$, which exhibits a pitchfork BP at $\lambda = 0$. The addition of a second parameter, for example, $\alpha + \lambda x - x^3 = 0$, splits the diagram as in Figure 1 if $\alpha \neq 0$. The boundary case $\alpha = 0$ is the analogue of the ineffectiveness condition we derive for the elastica, although the analysis and final result are more involved since our mathematical setting is an ODE (plus boundary conditions and an integral constraint) rather than an algebraic equation, and the perturbations considered are a space of functions rather than a single parameter. Furthermore, in the standard unfolding study, the focus is generally on $\alpha = 0$ as the transition between qualitative behaviors, whereas we focus on determining the leading-order behaviors of solutions both away from, and directly on, this boundary case.

In addition to this theoretical analysis, we present some computational results motivated by the idea of deriving an automated algorithm to determine DNA looping configurations. One approach to performing such computations is to use exactly the symmetry-breaking path considered here: begin with the symmetric problem for which solutions are known, and proceed via a continuation algorithm to numerical solutions for the perturbed problem. We explore how these continuation

algorithms can fail if the perturbation is ineffective (or even nearly ineffective), due to the presence of sharp corners in a branch of solutions. Thus, in designing a system for automatically computing such bifurcation diagrams for a wide range of intrinsic shapes, the ineffectivity conditions we derive serve as an important red flag that numerical difficulties may arise in a specific subset of computations.

Our analysis proceeds as follows. First, we formulate the planar buckling problem in Section 2, including an $O(\varepsilon)$ intrinsic-curvature term. Next, in Section 3, we apply a standard perturbation expansion to the *trivial branch* of $\varepsilon = 0$ unbuckled configurations. Away from the BPs, this expansion gives an $O(\varepsilon)$ approximation to the perturbation of the trivial branch. At the BPs, this analysis breaks down for most intrinsic curvature profiles, but for certain special profiles, it does still yield an $O(\varepsilon)$ solution. These cases are exactly the ineffective perturbations, and we derive conditions for when they occur. In Section 4, we apply an alternative perturbation technique called undetermined gauges to the effective perturbations at the BPs and find an $O(\varepsilon^{1/3})$ leading-order term for the perturbation of the unbuckled configuration. Finally, in Section 5, we present several examples illustrating the theory and a computational study verifying the numerical difficulties created by ineffective or nearly ineffective perturbations.

2. The planar buckling problem

We consider an inextensible and unshearable elastic rod in the plane, assumed for simplicity to have total arc-length 1. We parametrize the rod by arc-length s, and denote the configuration of the rod at arc-length-value s by $(x(s), y(s))$. We choose coordinates and boundary conditions as shown in Figure 2: the $s = 0$ end of the rod is at the origin, we impose clamped boundary conditions at each end requiring the

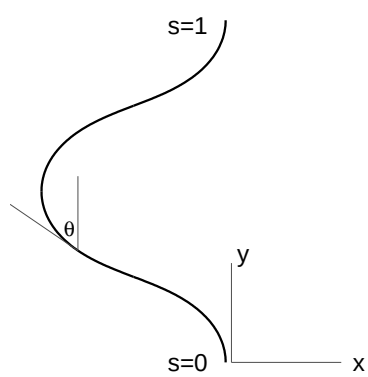


Figure 2. Boundary conditions on the planar elastica. The $s = 0$ end of the rod is held at the origin with vertical tangent, while the $s = 1$ end of the rod is held at $x = 0$, also with vertical tangent.

tangent vectors to be vertical, and further require $x(1) = 0$. The inextensibility–unshearability constraint implies that $(x'(s), y'(s))$ is a unit vector; this allows us to describe the rod by a single unknown function $\theta(s)$, where $(x'(s), y'(s)) = (\cos\theta(s), \sin\theta(s))$. The clamped boundary conditions are given by $\theta(0) = \theta(1) = 0$, and the additional constraint $x(1) = 0$ can be rewritten as $\int_0^1 \sin\theta(s)ds = 0$.

We place a mass $m > 0$ at the $s = 1$ end of the rod, and assume the following functional for the energy of the rod-plus-mass system:

$$E[\theta] \equiv \int_0^1 \Big(\frac{K}{2}(\theta'(s) - \varepsilon f(s))^2 + mg\cos\theta(s)\Big)\, ds.$$

The first term represents the bending energy of the rod, and the second term the potential energy of the load. The term $\varepsilon f(s)$ is used to model intrinsic curvature: the minimum-energy configuration of the rod has $\theta'(s) = \varepsilon f(s)$, and deviations from $\varepsilon f(s)$ involve a quadratic energy cost, weighted by the stiffness parameter K. For simplicity, we express energy in units of K, and define $\lambda = mg/K > 0$, so that the energy functional becomes

$$E[\theta] = \int_0^1 \Big(\frac{1}{2}(\theta'(s) - \varepsilon f(s))^2 + \lambda\cos\theta(s)\Big)\, ds.$$

We thus consider the calculus of variations problem to find critical points of E subject to the boundary conditions $\theta(0) = \theta(1) = 0$ and the isoperimetric constraint $\int_0^1 \sin\theta(s)ds = 0$. These critical points are found by solving the ordinary differential equation (ODE) defined by the Euler–Lagrange equation (with Lagrange multiplier μ included because of the isoperimetric constraint):

$$\theta''(s) = \varepsilon f'(s) - \lambda\sin\theta(s) + \mu\cos\theta(s).$$

Thus, the mathematical problem we considered was, given known values for the load λ, perturbation parameter ε, and intrinsic curvature profile $f(s)$ (with f' not identically zero, since otherwise ε has no effect), find solutions $(\theta(s), \mu)$ of

$$\begin{aligned} &\theta''(s) = \varepsilon f'(s) - \lambda\sin\theta(s) + \mu\cos\theta(s), \\ &\theta(0) = \theta(1) = 0, \quad \int_0^1 \sin\theta(s)ds = 0. \end{aligned} \tag{1}$$

For $\varepsilon = 0$, the solutions to (1) as λ varies yield the familiar force-length diagram seen in Figure 3. (All bifurcation diagrams in this article were computed using the parameter-continuation package AUTO97 [Doedel et al. 1991a; 1991b].) One solution (for each value of λ) is $\theta(s) \equiv 0$, $\mu = 0$ (a straight rod), and this corresponds to the horizontal line at the top of Figure 3. We call this the *trivial branch*.

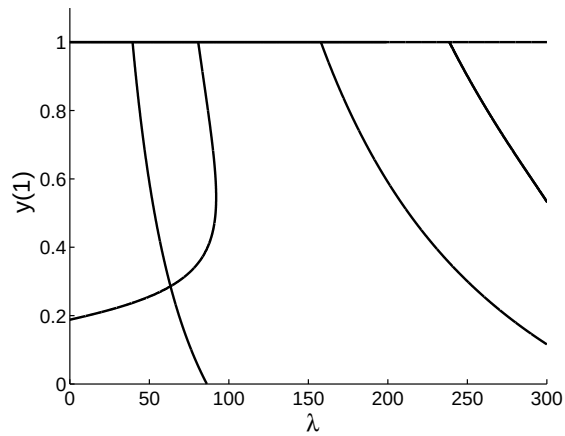


Figure 3. Bifurcation diagram for a planar elastica with no intrinsic curvature ($\varepsilon = 0$). The height of the top of the rod ($y(1)$) is plotted against the imposed load λ.

There are pitchfork bifurcation points[1] at all values of λ satisfying

$$2 - 2\cos\sqrt{\lambda} - \sqrt{\lambda}\sin\sqrt{\lambda} = 0 \tag{2}$$

(see Section 3 for a derivation of this equation). This equation has a countable sequence of solutions that we will label as $0 < \lambda_1 < \lambda_2 < \cdots$. For n odd,

$$\lambda_n = (n+1)^2\pi^2,$$

whereas for n even,

$$(n+0.5)^2\pi^2 < \lambda_n < (n+1)^2\pi^2, \quad \text{with } (n+1)^2\pi^2 - \lambda_n \to 0 \quad \text{as } n \to \infty.$$

Two properties of λ_n (for n even) will be useful to us:

Lemma 1. *For n even*,

$$\sin\sqrt{\lambda_n} = \frac{4\sqrt{\lambda_n}}{\lambda_n + 4}, \quad \cos\sqrt{\lambda_n} = \frac{4 - \lambda_n}{\lambda_n + 4}.$$

Proof. By (2), we have

$$1 - \cos\sqrt{\lambda_n} = \frac{\sqrt{\lambda_n}\sin\sqrt{\lambda_n}}{2}, \quad 1 + \cos\sqrt{\lambda_n} = 2 - \frac{\sqrt{\lambda_n}\sin\sqrt{\lambda_n}}{2}.$$

[1]The appearance of the pitchfork bifurcation points in Figure 3 differs slightly from the standard picture from Figure 1 since the two outer prongs of the pitchfork are folded on top of each other due to the choice of $y(1)$ as ordinate.

Multiplying these two equations, we find

$$\sin^2\sqrt{\lambda_n} = \sqrt{\lambda_n}\sin\sqrt{\lambda_n} - \frac{\lambda_n \sin^2\sqrt{\lambda_n}}{4}.$$

Collecting terms,

$$\sin^2\sqrt{\lambda_n}\left(1+\frac{\lambda_n}{4}\right) = \sqrt{\lambda_n}\sin\sqrt{\lambda_n},$$

and since $\sin\sqrt{\lambda_n} \neq 0$ for n even, we may divide both sides by it and solve for $\sin\sqrt{\lambda_n}$ to find the desired result.

The formula for $\cos\sqrt{\lambda_n}$ follows from the formula

$$\cos\sqrt{\lambda_n} = -\sqrt{1-\sin^2\sqrt{\lambda_n}}$$

(note the minus sign due to the fact that $\sqrt{\lambda_n}$ is just below an odd multiple of π). □

Lemma 2. *For n even*, $-\lambda_n - \lambda_n\cos\sqrt{\lambda_n} + 2\sqrt{\lambda_n}\sin\sqrt{\lambda_n} = 0$.

Proof. Since $(n+0.5)^2\pi^2 < \lambda_n < (n+1)^2\pi^2$, we have

$$\sin\sqrt{\lambda_n} \neq 0 \quad \text{and} \quad 1-\cos(\sqrt{\lambda_n}) \neq 0.$$

Thus

$$\frac{\sin\sqrt{\lambda_n}}{1-\cos\sqrt{\lambda_n}} = \frac{1+\cos\sqrt{\lambda_n}}{\sin\sqrt{\lambda_n}}, \tag{3}$$

since cross-multiplying yields the identity $1-\cos^2\sqrt{\lambda_n} = \sin^2\sqrt{\lambda_n}$. By (2), the left side of (3) equals $2/\sqrt{\lambda_n}$, and therefore,

$$\frac{1+\cos\sqrt{\lambda_n}}{\sin\sqrt{\lambda_n}} = \frac{2}{\sqrt{\lambda_n}}.$$

Cross-multiplying, and multiplying both sides by $\sqrt{\lambda_n}$, yields the desired equality. □

3. Perturbation of trivial branch for small ε

For fixed $\lambda > 0$, $f(s)$, and $\varepsilon > 0$, we now seek a solution to (1) that will be close to the solution on the trivial branch for $\varepsilon = 0$. We use a standard perturbation analysis, writing

$$\theta(s) = \theta_0(s) + \varepsilon\theta_1(s) + \cdots, \qquad \mu = \mu_0 + \varepsilon\mu_1 + \cdots.$$

The $O(1)$ terms $\theta_0(s)$ and μ_0 vanish since $\mu = 0$ and $\theta(s) \equiv 0$ on the trivial branch. Therefore, we seek the $O(\varepsilon)$ terms $\theta_1(s)$ and μ_1, by plugging these expansions into (1). After Taylor expanding $\sin\theta$ and $\cos\theta$ and isolating the $O(\varepsilon)$ terms we find

$$\theta_1'' = f'(s) - \lambda\theta_1 + \mu_1, \quad \theta_1(0) = \theta_1(1) = 0, \quad \int_0^1 \theta_1(s)ds = 0. \tag{4}$$

By a standard integrating factor approach, we find the general solution to the ODE in (4):

$$\theta_1(s) = \frac{1}{\sqrt{\lambda}} \int_0^s f'(t) \sin(\sqrt{\lambda}(s-t))\, dt + \frac{\mu_1}{\lambda} + C_1 \cos(\sqrt{\lambda} s) + C_2 \sin(\sqrt{\lambda} s)$$

When we require that this general solution satisfy the remaining conditions in (4), we find three linear equations in μ_1, C_1, and C_2:

$$\begin{bmatrix} 1/\lambda & 1 & 0 \\ 1/\lambda & \cos\sqrt{\lambda} & \sin\sqrt{\lambda} \\ 1/\lambda & (\sin\sqrt{\lambda})/\sqrt{\lambda} & (1-\cos\sqrt{\lambda})/\sqrt{\lambda} \end{bmatrix} \begin{bmatrix} \mu_1 \\ C_1 \\ C_2 \end{bmatrix} = \begin{bmatrix} 0 \\ (1/\sqrt{\lambda}) \int_0^1 f'(t) \sin(\sqrt{\lambda}(1-t))\, dt \\ (1/\sqrt{\lambda}) \int_0^1 \int_0^s f'(t) \sin(\sqrt{\lambda}(s-t))\, dt\, ds \end{bmatrix} \tag{5}$$

The bottom term on the right side can be simplified by switching the order of integration:

$$\int_0^1 \int_t^1 \frac{f'(t)}{\sqrt{\lambda}} \sin\bigl(\sqrt{\lambda}(s-t)\bigr)\, ds\, dt,$$

and then computing the inner integral:

$$\frac{1}{\lambda} \int_0^1 f'(t)\bigl[1 - \cos\bigl(\sqrt{\lambda}(1-t)\bigr)\bigr]\, dt.$$

Inserting this into (5), and multiplying both sides by λ, gives

$$\begin{bmatrix} 1 & \lambda & 0 \\ 1 & \lambda\cos\sqrt{\lambda} & \lambda\sin\sqrt{\lambda} \\ 1 & \sqrt{\lambda}\sin\sqrt{\lambda} & \sqrt{\lambda}(1-\cos\sqrt{\lambda}) \end{bmatrix} \begin{bmatrix} \mu_1 \\ C_1 \\ C_2 \end{bmatrix} = \begin{bmatrix} 0 \\ \sqrt{\lambda} \int_0^1 f'(t) \sin(\sqrt{\lambda}(1-t))\, dt \\ \int_0^1 f'(t)[1-\cos(\sqrt{\lambda}(1-t))]\, dt \end{bmatrix} \tag{6}$$

The matrix on the left has determinant $\lambda^{3/2}(2\cos\sqrt{\lambda} - 2 + \sqrt{\lambda}\sin\sqrt{\lambda})$, so (6) has a unique solution if $2\cos\sqrt{\lambda} - 2 + \sqrt{\lambda}\sin\sqrt{\lambda} \neq 0$. In other words, referring back to (2), we have shown that away from the bifurcation points, that is, for $\lambda \neq \lambda_n$, the standard perturbation expansion yields a solution, that is, a $O(\varepsilon)$ approximation to $\theta(s)$ near the trivial branch.

(We note that the BP condition (2) can be derived by a computation much like the one just completed; in that instance, we would be looking for solutions near the trivial branch to the problem without the f' term. We would use the same perturbation expansion, resulting in (6) but with a zero vector as the right side, and so we would have a nontrivial solution (a BP) exactly when the determinant of the matrix vanishes.)

What is particularly interesting for our study is whether (6) might have a solution even at a bifurcation point. Here we can use a fact from linear algebra (see [Shifrin and Adams 2002, Section 3.4], for example):

Theorem 1. *For an $n \times n$ matrix A, the column space of A is equal to the orthogonal complement of the null-space of A^T.*

Denote by $\boldsymbol{b}$ the right-hand side of (6). We want to know whether $\boldsymbol{b}$ is in the column space of the matrix $\boldsymbol{A}$ on the left side, which by Theorem 1 is true if and only if $\langle \boldsymbol{b}, \boldsymbol{u} \rangle = 0$ for all null-vectors $\boldsymbol{u}$ of $\boldsymbol{A}^T$. The null-vector for $\lambda = \lambda_n$ has a different form depending on whether n is odd or even.

If n is odd, then $\lambda_n = (n+1)^2\pi^2$ and

$$\boldsymbol{A}^T = \begin{bmatrix} 1 & 1 & 1 \\ (n{+}1)^2\pi^2 & (n{+}1)^2\pi^2 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

so the null space of $\boldsymbol{A}^T$ is span $\{(-1, 1, 0)\}$.

If n is even, no obvious simplification can be made to the form of the matrix:

$$\boldsymbol{A}^T = \begin{bmatrix} 1 & 1 & 1 \\ \lambda & \cos\sqrt{\lambda} & \sqrt{\lambda}\sin\sqrt{\lambda} \\ 0 & \lambda\sin\sqrt{\lambda} & \sqrt{\lambda}(1-\cos\sqrt{\lambda}) \end{bmatrix},$$

but, using (2) and Lemma 2, one sees that the null-space of $\boldsymbol{A}^T$ is span $\{(-1,-1,2)\}$.

Combining these null-vectors of $\boldsymbol{A}^T$ with Theorem 1 shows that (6) has a solution at λ_n if and only if

$$\begin{aligned} \int_0^1 f'(t)\sin(\sqrt{\lambda_n}(1-t))\,dt &= 0 \quad \text{for } n \text{ odd}, \\ \int_0^1 f'(t)\big[2 - 2\cos\big(\sqrt{\lambda_n}(1-t)\big) - \sqrt{\lambda_n}\sin\big(\sqrt{\lambda_n}(1-t)\big)\big]\,dt &= 0 \quad \text{for } n \text{ even}. \end{aligned} \tag{7}$$

Intrinsic curvature profiles f satisfying (7) for some n are the *ineffective* perturbations for bifurcation point λ_n.

4. Undetermined-gauges analysis of bifurcation points

We now turn to the analysis of the bifurcation points $\lambda = \lambda_n$. In Section 3 we showed that for the ineffective perturbations defined by (7), the standard perturbation analysis predicts an $O(\varepsilon)$ lowest-order term for θ_1 and μ_1, but that this analysis fails for the remaining perturbations. Here we investigate those cases by applying a more general technique, the methods of undetermined gauges [Murdock 1999], which is used to derive the leading-order behavior when it does not follow the standard $O(\varepsilon)$ pattern.

We are, as before, considering (1), but now we formulate a more general perturbation expansion for θ and μ. This expansion is computed one term at a time, so we begin by writing

$$\theta(s) = \delta_1(\varepsilon)\theta_1(s), \quad \mu = \delta_1(\varepsilon)\mu_1,$$

where $\delta_1(\varepsilon)$ is an unknown function of ε that our analysis will determine. We will restrict attention to the family $\delta_1(\varepsilon) = \varepsilon^a$ for a real. We insert these expressions into (1) and expand the sin and cos as Taylor series:

$$\begin{aligned} &\delta_1\theta_1'' = \varepsilon f'(s) - \lambda_n(\delta_1\theta_1 + \cdots) + (\delta_1\mu_1)(1 + \cdots), \\ &\delta_1\theta_1(0) = \delta_1\theta_1(1) = 0, \qquad \int_0^1 (\delta_1\theta_1 + \cdots)\, ds = 0. \end{aligned}$$

We want to look at the leading order terms, but in the ODE, there is a question of whether δ_1 or ε is dominant, or if they could be of the same order. If ε were dominant, then the leading-order terms of the ODE would be the nonsensical $0 = \varepsilon f'(s)$. If δ_1 and ε were of the same order, that is, $\delta_1 = \varepsilon$, then the leading-order terms of the ODE would give the same equation as in the standard perturbation expansion, which we know from Section 3 has no solution. Hence, the dominant term must be δ_1, so we keep the $O(\delta_1)$ terms to find

$$\theta_1'' = -\lambda_n\theta_1 + \mu_1, \quad \theta_1(0) = \theta_1(1) = 0, \quad \int_0^1 \theta_1(s)\, ds = 0.$$

The general solution of the ODE is $\theta_1(s) = C_1 \cos(\sqrt{\lambda_n}s) + C_2 \sin(\sqrt{\lambda_n}s) + \mu_1/\lambda_n$. Imposing $\theta_1(0) = 0$, we find $\mu_1 = -\lambda_n C_1$, and then the other two conditions give the linear system

$$\begin{bmatrix} \cos\sqrt{\lambda_n} - 1 & \sin\sqrt{\lambda_n} \\ \dfrac{\sin\sqrt{\lambda_n}}{\sqrt{\lambda_n}} - 1 & \dfrac{1 - \cos\sqrt{\lambda_n}}{\sqrt{\lambda_n}} \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Since the determinant of the matrix in this system is zero by (2), we have nontrivial solutions (C_1, C_2), namely any nonzero null-vector of the matrix. If n is odd, the matrix simplifies to $\left[\begin{smallmatrix} 0 & 0 \\ -1 & 0 \end{smallmatrix}\right]$, so $(0, 1)$ is a null-vector and the solution coming from this first gauge — still with n odd — is

$$\begin{gathered} \theta_1 = C \sin(\sqrt{\lambda_n}s) = C \sin(\pi(n+1)s) \quad (C \neq 0 \text{ to be determined}), \\ \mu_1 = 0. \end{gathered} \tag{8}$$

On the other hand, if n is even, then $(\sin\sqrt{\lambda_n}, 1-\cos\sqrt{\lambda_n})$ is a null-vector of the matrix, which means that the solution coming from this first gauge is

$$\begin{aligned}\theta_1 &= k\big[\sin\sqrt{\lambda_n}\big(\cos(\sqrt{\lambda_n}s)-1\big)+(1-\cos\sqrt{\lambda_n})\sin(\sqrt{\lambda_n}s)\big],\\ \mu_1 &= -k\lambda_n\sin\sqrt{\lambda_n}.\end{aligned}$$

By (2), $1-\cos\sqrt{\lambda_n}=\frac{1}{2}\sqrt{\lambda_n}\sin\sqrt{\lambda_n}$, so that this solution can be rewritten as

$$\theta_1 = k\sin\sqrt{\lambda_n}\big(\cos(\sqrt{\lambda_n}s)-1+\tfrac{1}{2}\sqrt{\lambda_n}\sin(\sqrt{\lambda_n}s)\big),\quad \mu_1=-k\lambda_n\sin\sqrt{\lambda_n}.$$

For simplicity, we write $k\sin\sqrt{\lambda_n}/2$ as a single constant C for the final form of the solution for n even:

$$\begin{gathered}\theta_1 = C\big(2\cos(\sqrt{\lambda_n}s)-2+\sqrt{\lambda_n}\sin(\sqrt{\lambda_n}s)\big)\quad (C\neq 0 \text{ to be determined}),\\ \mu_1=-2C\lambda_n.\end{gathered}\tag{9}$$

In order to determine C and δ_1, we add another gauge.

$$\theta=\delta_1(\varepsilon)\theta_1(s)+\delta_2(\varepsilon)\theta_2(s),\quad \mu=\delta_1(\varepsilon)\mu_1+\delta_2(\varepsilon)\mu_2,$$

for $\delta_2(\varepsilon)$ an unknown function of ε (again in the family ε^a), by definition of lower order in ε than δ_1. As before, we insert these expressions into (1) and Taylor-expand the sin and cos terms, looking for the next-lowest-order terms after $O(\delta_1)$.

For the boundary conditions, these next-lowest-order terms give $\delta_2(\varepsilon)\theta_2(0)=\delta_2(\varepsilon)\theta_2(1)=0$, or $\theta_2(0)=\theta_2(1)=0$. As for the integral condition, since $\sin\theta=\theta-\theta^3/6+\cdots$, there are two next-lowest-order candidates: δ_2 from the θ term, and δ_1^3 from the θ^3 term. We list them both for the time being:

$$\int_0^1\big(\delta_2\theta_2(s)-\tfrac{1}{6}(\delta_1)^3\theta_1(s)^3+\cdots\big)\,ds=0,\tag{10}$$

Finally, we look at the ODE. The sin term yields the same two possible second-lowest-order terms δ_2 or $(\delta_1)^3$ as in the integral condition, and so, in fact, does the cos term (δ_2 from the $\delta_2\mu_2$ term in μ times the 1 in the cos expansion, or $(\delta_1)^3$ from the $\delta_1\mu_1$ term in μ times the $-(\delta_1\theta_1)^2/2$ from the cos expansion):

$$\begin{aligned}\delta_2\theta_2'' = \varepsilon f'(s)-\lambda_n\big[\delta_2\theta_2(s)-\tfrac{1}{6}(\delta_1)^3(\theta_1(s))^3+\cdots\big]\\ +\big[\delta_2\mu_2-\tfrac{1}{2}(\delta_1)^3\mu_1(\theta_1(s))^2+\cdots\big].\end{aligned}\tag{11}$$

Overall, there are three candidates for next-lowest-order term: ε, δ_2, and $(\delta_1)^3$. We have to consider all possibilities for the relative rankings of these terms, *including ties*. The arguments below rule out all possibilities except having all three of the same order.

Case 1: ε lowest-order. This cannot be true, since the dominant terms in (11) would give $0 = f'(s)$, but by assumption f' is not identically zero.

Case 2: $(\delta_1)^3$ lowest-order. This cannot be true, since the dominant terms in (11) would give

$$0 = -\tfrac{1}{6}\big(\theta_1(s)\big)^2(\theta_1(s) + 3\mu_1),$$

and neither θ_1 nor $\theta_1 + 3\mu_1$ is identically zero.

Case 3: δ_2 lowest-order. The dominant terms in (11) would give the same equation we solved for θ_1 and μ_1 (including the boundary and integral conditions), so that $\theta_2 = \theta_1$ and $\mu_2 = \mu_1$. Thus, our gauge expansions would reduce to $\theta = (\delta_1 + \delta_2)\theta_1$ and $\mu = (\delta_1 + \delta_2)\mu_1$, and we would essentially be back where we began this step, having replaced the unknown $\delta_1(\varepsilon)$ by another unknown $\delta_1(\varepsilon) + \delta_2(\varepsilon)$, without having learned anything about the connection of δ_1 to ε. Thus, we reject this case.

Case 4: ε and $(\delta_1)^3$ tied for lowest-order. Since $(\delta_1)^3 = \varepsilon$, the dominant terms in (11) would give $0 = f'(s) - \frac{1}{6}(\theta_1(s))^3$. This requires the perturbation f' to take the very particular form of the cubes of the functions (8) or (9). Since this case does not yield a solution for a general perturbation, we reject this case.

Case 5: ε and δ_2 tied for lowest-order. This cannot be true: we would have $\delta_2 = \varepsilon$ and the dominant terms in (11) would give the same equation (4) from the standard perturbation expansion, as well as the same integral and boundary conditions, and we know that this system has no solution for $\lambda = \lambda_n$ and the effective perturbations we are considering in this section.

Case 6: δ_2 and $(\delta_1)^3$ tied for lowest order. We have $(\delta_1)^3 = \delta_2$ and the dominant terms in (11) would give

$$\theta_2'' = -\lambda_n\theta_2 + \tfrac{1}{6}\lambda_n(\theta_1)^3 + \mu_2 - \tfrac{1}{2}\mu_1(\theta_1)^2.$$

This equation can be solved in closed form using the forms of μ_1 and θ_1 from (8) and (9), as follows.

For n odd, we have $\mu_1 = 0$ and $\theta_1 = C\sin(\sqrt{\lambda_n}s)$, and the ODE has solution

$$\begin{aligned}\theta_2(s) = {} & C_1\cos(\sqrt{\lambda_n}s) + C_2\sin(\sqrt{\lambda_n}s) \\ & + \mu_2/\lambda_n - \tfrac{1}{16}C^3 s\sqrt{\lambda_n}\cos(\sqrt{\lambda_n}s) + \tfrac{1}{192}C^3\sin\big(3\sqrt{\lambda_n}s\big).\end{aligned}$$

Applying $\theta_2(0) = \theta_2(1) = 0$ leads to the impossible conclusion that $C = 0$.

For n even, the solution of the ODE is

$$\begin{aligned}\theta_2(s) = {} & C_1\cos(\sqrt{\lambda_n}s) + C_2\sin(\sqrt{\lambda_n}s) + \mu_2/\lambda_n + \tfrac{8}{3}C^3 \\ & + \tfrac{1}{16}C^3(\lambda_n - 12)(2 - \lambda_n s)\cos(\sqrt{\lambda_n}s) + \tfrac{1}{96}C^3(3\lambda_n - 4)\cos(3\sqrt{\lambda_n}s) \\ & + \tfrac{1}{192}C^3\sqrt{\lambda_n}(\lambda_n - 12)\big[24s\sin(\sqrt{\lambda_n}s) + \sin(3\sqrt{\lambda_n}s)\big].\end{aligned}$$

The condition $\theta_2(0) = 0$ allows us to solve for μ_2, leaving

$$\begin{aligned}\theta_2(s) = C_1\big[\cos\big(\sqrt{\lambda_n}s\big)-1\big] + C_2\sin(\sqrt{\lambda_n}s) + \tfrac{1}{96}C^3(148-15\lambda_n)\\ + \tfrac{1}{16}C^3(\lambda_n-12)(2-\lambda_n s)\cos(\sqrt{\lambda_n}s) + \tfrac{1}{96}C^3(3\lambda_n-4)\cos(3\sqrt{\lambda_n}s)\\ + \tfrac{1}{192}C^3\sqrt{\lambda_n}(\lambda_n-12)\big[24s\sin(\sqrt{\lambda_n}s) + \sin(3\sqrt{\lambda_n}s)\big].\end{aligned} \tag{12}$$

Next we impose the boundary condition $\theta_2(1) = 0$, and using Lemma 1, we find

$$\theta_2(1) = -\frac{2\lambda_n}{4+\lambda_n}C_1 + \frac{4\sqrt{\lambda_n}}{4+\lambda_n}C_2 + \frac{C^3\lambda_n^2(\lambda_n-12)}{16(4+\lambda_n)} = 0. \tag{13}$$

Recalling (10), since $\delta_2 = (\delta_1)^3$, the integral condition is

$$\int_0^1 \big(\theta_2(s) - \tfrac{1}{6}(\theta_1(s))^3\big)\,ds = 0.$$

Again using Lemma 1, we can simplify this to

$$-\frac{\lambda_n}{4+\lambda_n}C_1 + \frac{2\sqrt{\lambda_n}}{4+\lambda_n}C_2 + \frac{C^3\lambda_n(20+\lambda_n)}{8(4+\lambda_n)} = 0. \tag{14}$$

Subtracting two times (14) from (13) gives

$$\frac{C^3\lambda_n^2(\lambda_n-12)}{16(4+\lambda_n)} - \frac{C^3\lambda_n(20+\lambda_n)}{4(4+\lambda_n)} = \frac{C^3\lambda_n(\lambda_n-20)}{16} = 0,$$

which implies either $C = 0$, $\lambda_n = 0$, or $\lambda_n = 20$, none of which is true.

Having ruled out all other cases, we can conclude that δ_2, $(\delta_1)^3$ and ε are all of the same order, that is, $\delta_1 = \varepsilon^{1/3}$ and $\delta_2 = \varepsilon$. In particular, we have shown that $\theta = O(\varepsilon^{1/3})$.

In fact, this second gauge allows us to completely determine the leading-order behavior of $\theta(s)$ and μ, as summarized by the following theorem:

Theorem 2. *For n odd, we have $\theta(s) = C\varepsilon^{1/3}\sin(\sqrt{\lambda_n}s) + \cdots$ and $\mu = \varepsilon\mu_2 + \cdots$, with $\lambda_n = \pi(n+1)$,*

$$C = \Big(\frac{16}{\lambda_n}\int_0^1 f'(t)\sin(\sqrt{\lambda_n}(1-t))\,dt\Big)^{1/3},\quad \mu_2 = -\int_0^1 f'(t)\big[1-\cos(\sqrt{\lambda_n}(1-t))\big]\,dt.$$

For n even, we have $\theta(s) = C\varepsilon^{1/3}(2\cos(\sqrt{\lambda_n}s) - 2 + \sqrt{\lambda_n}\sin(\sqrt{\lambda_n}s)) + \cdots$ and $\mu = -2\varepsilon^{1/3}C\lambda_n + \cdots$, with

$$C = \Big(\frac{16}{(\lambda_n)^2(\lambda_n-20)}\int_0^1 f'(t)\Big[2-2\cos(\sqrt{\lambda_n}(1-t)) - \sqrt{\lambda_n}\sin(\sqrt{\lambda_n}(1-t))\Big]\,dt\Big)^{1/3}.$$

Proof. For n even, all that remains is to determine C, while for n odd, we must compute C and μ_2. The derivation largely follows the computations in Case 6. The ODE to be solved is

$$\theta_2'' = f' - \lambda_n\theta_2 + \tfrac{1}{6}\lambda_n(\theta_1)^3 + \mu_2 - \tfrac{1}{2}\mu_1(\theta_1)^2, \tag{15}$$

identical to Case 6 except for the addition of f' to the right side. Therefore, the solution of the ODE will be the expression found in Case 6 plus the term

$$\frac{1}{\sqrt{\lambda_n}}\int_0^s f'(t)\sin(\sqrt{\lambda_n}(s-t))\,dt, \tag{16}$$

the solution to the equation $\theta_2'' = f' - \lambda_n\theta_2$ seen in Section 3. Since this new term vanishes at $s = 0$, it will have no effect on the first step from Case 6 (in which we get an expression for μ_2 using the condition $\theta_2(0) = 0$).

Thus, for n odd, the solution to (15) plus $\theta_2(0) = 0$ is

$$\begin{aligned}\theta_2(s) = C_1\big(\cos(\sqrt{\lambda_n}s) - 1\big) + C_2\sin(\sqrt{\lambda_n}s) - \tfrac{1}{16}C^3 s\sqrt{\lambda_n}\cos(\sqrt{\lambda_n}s)\\ + \tfrac{1}{192}C^3\sin(3\sqrt{\lambda_n}s) + \frac{1}{\sqrt{\lambda_n}}\int_0^s f'(t)\sin\big(\sqrt{\lambda_n}(s-t)\big)\,dt.\end{aligned}$$

Next we impose the condition $\theta_2(1) = 0$ to find the given formula for C. Note that the integral does not vanish (and hence $C \neq 0$ as required) since that is our definition of what makes a perturbation f' effective.

Finally, we impose the condition

$$\int_0^1 \theta_2(s)\,ds = 0$$

(for n odd, the quantity $(\theta_1)^3 = \sin^3(\pi(n+1)s)$ has zero integral, so this term drops out of the integral condition) to find

$$C_1 = \frac{1}{\sqrt{\lambda_n}}\int_0^1\int_0^s f'(t)\sin(\sqrt{\lambda_n}(s-t))\,dt\,ds = \frac{1}{\lambda_n}\int_0^1 f'(t)\big[1-\cos(\sqrt{\lambda_n}(1-t))\big]\,dt,$$

where the second equality comes from switching the order of integration as in Section 3. Since $\mu_2 = -C_1\lambda_n$ (from the $\theta_2(0)$ condition), we find the given formula for μ_2.

For n even, the solution to (15) plus $\theta_2(0) = 0$ is (12) plus the term (16). Using the same steps as in Case 6, the boundary condition $\theta_2(1) = 0$ and the integral

condition yield the equations

$$-\frac{2\lambda_n}{4+\lambda_n}C_1+\frac{4\sqrt{\lambda_n}}{4+\lambda_n}C_2+\frac{C^3\lambda_n^2(\lambda_n-12)}{16(4+\lambda_n)}+\frac{1}{\sqrt{\lambda_n}}\int_0^1 f'(t)\sin(\sqrt{\lambda_n}(1-t))\,dt=0,$$

$$-\frac{\lambda_n}{4+\lambda_n}C_1+\frac{2\sqrt{\lambda_n}}{4+\lambda_n}C_2+\frac{C^3\lambda_n(20+\lambda_n)}{8(4+\lambda_n)}+\frac{1}{\lambda_n}\int_0^1 f'(t)(1-\cos(\sqrt{\lambda_n}(1-t)))\,dt=0.$$

Subtracting twice the second equation from the first yields the formula for C. □

5. Examples

Bifurcation diagrams with effective and ineffective perturbations. We consider four perturbations f':

f_1', effective for both BPs,

f_2', ineffective for the first BP and effective for the second,

f_3', effective for the first BP and ineffective for the second,

f_4', ineffective for both BPs.

Specifically, we first define

$$u_1(s)=\sqrt{2}\sin(\sqrt{\lambda_1}(1-s)),$$
$$u_2(s)=\frac{1}{\sqrt{\lambda_2}}\big[2-2\cos(\sqrt{\lambda_2}(1-s))-\sqrt{\lambda_2}\sin(\sqrt{\lambda_2}(1-s))\big].$$

These are length-1 elements in L^2 in the directions of the functions that define ineffectiveness for the BPs, in the sense that f' is ineffective at the n-th BP if $\langle f', u_n\rangle = 0$. We note that u_1 and u_2 are orthogonal.

We define our four perturbations by

$$f_1'=(u_1+u_2)/\sqrt{2},\quad f_2'=u_2,\quad f_3'=u_1,$$
$$f_4'(s)=\sqrt{2\pi^2-3}(s+\sin(2\pi s)/\pi)/(\pi\sqrt{6}).$$

By design, all the f_j' have length 1 (to allow comparisons); f_2' is orthogonal to u_1, f_3' is orthogonal to u_2, f_4' is orthogonal to both u_1 and u_2; and all other pairings of f_j' with u_k are not close to orthogonal.

The bifurcation diagrams for these perturbations, with $\varepsilon = 1$, are shown in the four panels of Figure 4. The difference between effective and ineffective perturbations is clear: in each case where a perturbation is effective, the diagram is relatively smooth near the former BP, whereas in the ineffective cases, the diagram has a sharp corner. Indeed, to numerically compute some of these corners required a significant amount of care, for example, the use of a very small step size, or a temporary increase in ε by an order of magnitude just to get onto the perturbation of the bifurcating branch.

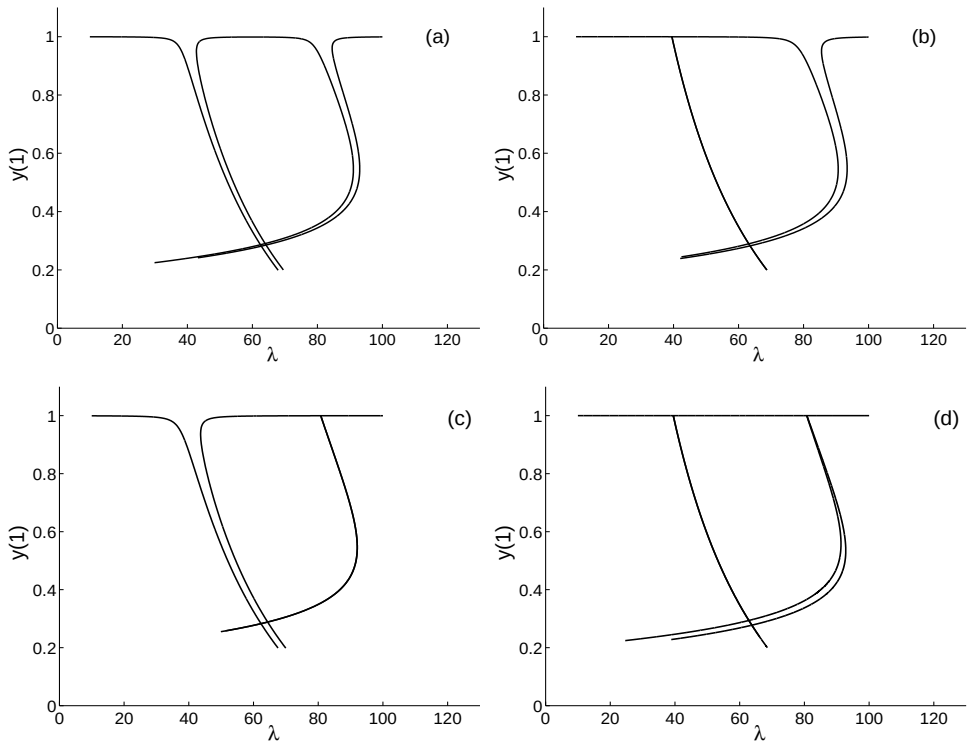


Figure 4. Bifurcation diagrams for elastica perturbed by intrinsic curvature profiles f' (see previous page for the specific functions used): (a) effective perturbation at both BPs; (b) ineffective at the first BP, effective at the second; (c) effective at the first BP, ineffective at the second; (d) ineffective at both BPs.

Leading-order behaviors in ε. Finally, we show two examples illustrating the lowest-order expressions for $\theta(s)$ and μ found in Section 4. In our first example, We take $f'(s) = s$, a perturbation that is effective at $\lambda = \lambda_1$ ($\langle f', u_1\rangle = 0.39$) and ineffective at $\lambda = \lambda_2$ ($\langle f', u_2\rangle = 0$). In Figure 5, we show the graphs of $\theta(s)$ for $\lambda = \lambda_1, \lambda_2$, and for $\varepsilon = \frac{1}{4}, \frac{1}{2}$, and 1.

Our theoretical prediction for the behavior at $\lambda = \lambda_1$ comes from Theorem 2 (since f' is effective). We compute $\lambda_1 = 2\pi$,

$$C = \Big(\frac{16}{4\pi^2}\int_0^1 t\,\sin(2\pi\,(1-t))\,dt\Big)^{1/3} = 0.401,$$

and

$$\mu_2 = -\int_0^1 t\big(1-\cos(2\pi\,(1-t))\big)\,dt = -0.5.$$

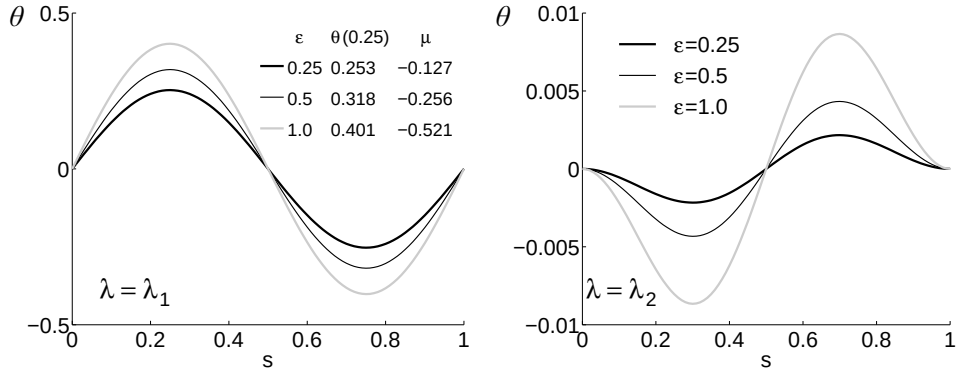


Figure 5. Graphs of $\theta(s)$ for $f'(s) = s$, $\lambda = \lambda_1, \lambda_2$ and $\varepsilon = \frac{1}{4}$, $\frac{1}{2}$, 1. For $\lambda = \lambda_1$, the dependence of θ on ε seems to be larger than $O(\varepsilon)$, in line with the $O(\varepsilon^{1/3})$ prediction of the theory. The values of $\theta(\frac{1}{4})$ (the maxima) and μ are also reported. For $\lambda = \lambda_2$, the dependence of θ on ε appears to be approximately $O(\varepsilon)$, in line with the fact that f' is ineffective at λ_2.

Thus, our predicted behavior at $\lambda = \lambda_1$ is

$$\theta(s) \approx 0.401\varepsilon^{1/3} \sin(2\pi s), \quad \mu \approx -0.5\varepsilon.$$

The shape of the actual solution $\theta(s)$ in Figure 5 matches the predicted $\sin(2\pi s)$, and the scaling with ε is clearly larger than $O(\varepsilon)$. Furthermore, from the table inset in the figure, we see that both $\theta(\frac{1}{4})$ (the heights of the peaks) and μ are close matches with our predicted formulas.

As for $\lambda = \lambda_2$, since f' satisfies the ineffectivity condition, we expect $\theta(s)$ to have $O(\varepsilon)$ behavior rather than $O(\varepsilon^{1/3})$. Indeed, we see in Figure 5 that this appears to be the case, as $\theta(s)$ appears to be roughly halved when ε is halved. In this case, our theory does not give a predictive formula for the leading-order behavior of $\theta(s)$ or μ; the system (6) has an infinite number of solutions, and one would have to proceed to higher-order terms in the standard perturbation expansion in order to determine which of these solutions is relevant.

In our second example, we take $f'(s) = \sin(3\pi s)$, a perturbation that is ineffective at $\lambda = \lambda_1$ ($\langle f', u_1\rangle = 0$) and is effective at $\lambda = \lambda_2$ ($\langle f', u_2\rangle = -0.67$). In Figure 6, we show the graphs of $\theta(s)$ for $\lambda = \lambda_1, \lambda_2$ and $\varepsilon = \frac{1}{4}, \frac{1}{2}$, and 1.

Our theoretical prediction for the behavior at $\lambda = \lambda_2$ comes from Theorem 2 (since f' is effective). Using the formulas in that theorem, we compute $\lambda_2 = 80.7629$ and $C = 0.05557$. Thus, our predicted behavior at $\lambda = \lambda_2$ is

$$\theta(s) \approx -0.05557\varepsilon^{1/3}(2\cos(8.987s) - 2 + 8.987\sin(8.987s)), \quad \mu \approx 8.976\varepsilon^{1/3}.$$

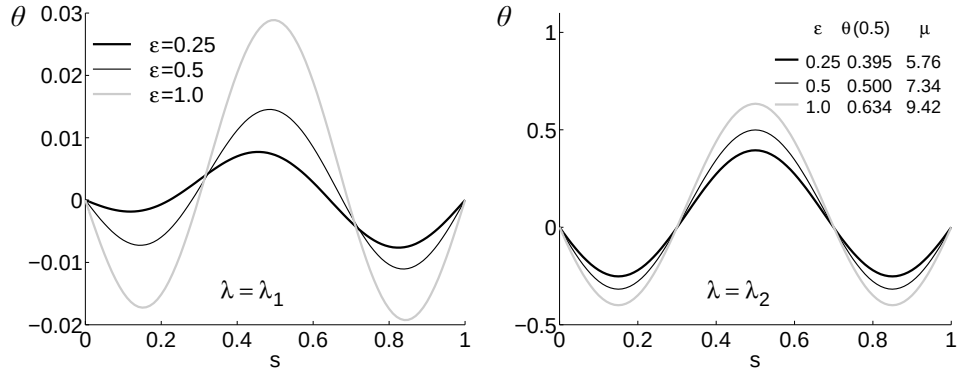


Figure 6. Graphs of $\theta(s)$ for $f'(s) = \sin(3\pi s)$, $\lambda = \lambda_1$, λ_2 and $\varepsilon = \frac{1}{4}, \frac{1}{2}, 1$. For $\lambda = \lambda_1$, the dependence of θ on ε appears to be approximately $O(\varepsilon)$, in line with the fact that f' is ineffective at λ_1. For $\lambda = \lambda_2$, the dependence of θ on ε appears to be larger than $O(\varepsilon)$, in line with the $O(\varepsilon^{1/3})$ prediction of the theory. The values of $\theta(\frac{1}{2})$ and μ are also reported.

The shape of the actual solution $\theta(s)$ in Figure 6 matches the predicted functional form, and the scaling with ε is clearly larger than $O(\varepsilon)$. Furthermore, from the table inset in the figure, we see that both $\theta(\frac{1}{2})$ (which from our theoretical formula should equal $0.623\varepsilon^{1/3}$) and μ are close matches with our predicted formulas.

Computational impact. Apart from an interest in understanding on a theoretical level how a shape perturbation affects the bifurcation diagram for buckling, we were also motivated by a pragmatic concern: to what extent ineffective perturbations would interfere with the design of an automated algorithm to compute bifurcation diagrams for a given intrinsic shape. The sharp corners in parts (b), (c) and (d) of Figure 4 suggest potential computational challenges, and we explored that question more concretely with the following numerical study.

We generated intrinsic shapes in three different categories (*random*, *nearly ineffective*, and *ineffective*) as follows. Let $f_1(s)$, $f_2(s)$, $f_3(s)$, $f_4(s)$ be the Gram–Schmidt orthonormal basis (in $L^2([0, 1])$) generated by the functions $\{s, s^2, s^3, s^4\}$:

$$f_1(s) = \sqrt{3}s, \qquad f_3(s) = 15\sqrt{7}s^3 - 20\sqrt{7}s^2 + 6\sqrt{7}s,$$
$$f_2(s) = 4\sqrt{5}s^2 - 3\sqrt{5}s, \qquad f_4(s) = 168s^4 - 315s^3 + 180s^2 - 30s.$$

The intrinsic shape function $f'(s)$ is defined as a linear combination

$$c_1 f_1(s) + c_2 f_2(s) + c_3 f_3(s) + c_4 f_4(s)$$

of these basis functions, the coefficients c_i being chosen according to different rules for the three cases. For a *random* perturbation, we choose four independent random numbers x_1, x_2, x_3, x_4 from a normal distribution with mean 0 and standard deviation 1, and then define

$$c_j = x_j / \sqrt{(x_1)^2 + (x_2)^2 + (x_3)^2 + (x_4)^2}.$$

For an *ineffective* perturbation, we take x_1, x_2, x_3 as above, but then choose x_4 such that $f'(s)$ is ineffective, according to the $n = 1$ case of (7), and then normalize to define the c_j as in the random case. For a *nearly ineffective* perturbation, we generate x_1, x_2, x_3, x_4 as in the ineffective case, but then add to x_4 a random number with normal distribution with mean 0 and standard deviation 0.01, before normalizing to define the c_j as in the random case.

Given each intrinsic shape, we performed a parameter continuation computation in AUTO97 to attempt to compute the first branch of buckled configurations for that intrinsic shape. We began the computation with $\lambda = 28$, zero intrinsic shape, and a straight rod configuration. Then we turned on the intrinsic shape via parameter continuation, by multiplying the intrinsic shape function $f'(s)$ by a parameter μ that was slowly increased from 0 (no intrinsic shape) to 1 (intrinsic shape determined by f'). Then we increased λ to a target maximum value of 50. A successful computation would follow the bend of the branch of solutions, with significant rod buckling occurring around $\lambda = 4\pi^2$ and continuing until λ reaches 50; see part (a) of Figure 4. However, in cases where a sharp corner exists near $\lambda = 4\pi^2$, as in the other parts of Figure 4, the computation could jump branches and end at a nearly straight configuration at $\lambda = 50$.

To assess in an automated way the success of this computation, we did a third parameter continuation step that decreased μ from 1 back down to 0. Thus, successful runs end at the $\lambda = 50$ point on the first bifurcating branch for the intrinsically-straight rod (that is, Figure 3), while unsuccessful runs end with the rod completely extended. Inspection of the value of $y(1)$ at the end of the third parameter continuation step allowed easy distinction of these two cases.

Results of these computations are shown in Table 1. The body of the table shows the percentage of successful runs out of a total of 300 attempts. To give a sense of variability of these results, we report in parentheses the corresponding standard deviation for a binomial random variable with $N = 300$ and p taken as the observed percentage of successes: $\sigma = \sqrt{p(1-p)/N}$. (For results reported as 100% (or 0%), all (or none) of the 300 computations were successful, and thus no meaningful estimate of σ can be provided).

AUTO97 allows a variety of parameter-stepping algorithms, and Table 1 shows the results for six different approaches: three with fixed step size and three with variable step size. The step sizes $\Delta\tau$ are in terms of a *pseudoarc-length* that is a

continuation method	random shape	nearly ineffective shape	ineffective shape
$\Delta\tau = 0.02$	100%	41% ($\pm$3%)	0%
$\Delta\tau = 0.04$	100%	16% ($\pm$2%)	0%
$\Delta\tau = 0.1$	100%	0.7% ($\pm$0.5%)	0%
$0.002 \le \Delta\tau \le 0.2$	99.7% ($\pm$0.3%)	7% ($\pm$2%)	5% ($\pm$1%)
$0.004 \le \Delta\tau \le 0.4$	99.7% ($\pm$0.3%)	21% ($\pm$2%)	5% ($\pm$1%)
$0.01 \le \Delta\tau \le 1$	98.7% ($\pm$0.7%)	3% ($\pm$0.9%)	0%

Table 1. Percentage of successful computations of the first branch of buckled configurations for different intrinsic shapes: those that are ineffective (in the sense of Section 3), those that are nearly ineffective (small perturbations of exactly ineffective shapes), and randomly chosen shapes (see text for detailed descriptions of the three cases). Each row represents one step-size strategy within the AUTO97 parameter continuation algorithm.

combination of the change in the parameter value λ and the change in the solution vector of the discretized Euler–Lagrange equations (1); this approach allows the traversing of "folds" in the bifurcation diagram where $\Delta\lambda = 0$. Thus, one can informally think of the change in the parameter λ in each step as being some fraction of $\Delta\tau$ (though what that fraction is will vary according to the change in the solution vector at that point on the branch). For the variable step-size computations, the step size $\Delta\tau$ is allowed to varying over two orders of magnitude, with an initial value in the middle (for example, $(\Delta\tau)_{\text{init}} = 0.02$ with $0.002 \le \Delta\tau \le 0.2$ throughout the computation in row 4 of Table 1). AUTO97 adjusts the step size with each step according to the convergence properties of the previous step, striving to take smaller steps when the convergence is more difficult.

Random and ineffective shapes behave radically differently for the range of step sizes shown here, and even the nearly ineffective shapes show a relatively high rate of computational failure, suggesting that this phenomenon will be met in practice for some shapes (despite the fact that the set of precisely ineffective shapes is measure zero). As would be expected, for fixed step sizes, smaller ones are more successful, though of course at the cost of computation time. (Even for ineffective perturbations, a sufficiently small step size will yield successful branch tracking, though $\Delta\tau$ needs to be significantly smaller than 0.02). For variable step sizes, the data suggests a more complicated situation, including some behavior in the nearly ineffective column that is not monotonic with step-size bounds. This might be explained by the fact that the automated adjustment in step size presumably

increases the step size consistently in the early part of the computation when the rod is barely changing, and this increased step size might increase the probability of jumping over a corner in the branch (though the likelihood of this jump might also depend sensitively on the initial point chosen, since that could determine whether the jump happens to straddle the corner). Further study would be needed to fully understand this behavior, but it seems clear at least that the ineffectivity condition derived here is a useful flag for intrinsic shapes that call for a strong decrease in the step size.

References

[Bolshoy et al. 1991] A. Bolshoy, P. McNamara, R. E. Harrington, and E. N. Trifonov, "Curved DNA without A-A: Experimental estimation of all 16 wedge angles", *Proc. Natl. Acad. Sci. USA* **88**:6 (1991), 2312–2316.

[De Santis et al. 1992] P. De Santis, A. Palleschi, M. Savino, and A. Scipioni, "Theoretical prediction of the gel electrophoretic retardation changers due to point mutations in a tract of Sv40 DNA", *Biophys. Chem.* **42**:2 (1992), 147–152.

[Dixit et al. 2005] S. B. Dixit, D. L. Beveridge, D. A. Case, T. E. C. III, E. Giudice, F. Lankas, R. Lavery, J. H. Maddocks, R. Osman, H. Sklenar, K. M. Thayer, and P. Varnai, "Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides, II: Sequence context effects on the dynamical structures of the 10 unique dinucleotide steps", *Biophys. J.* **89**:6 (2005), 3721–3740.

[Doedel et al. 1991a] E. Doedel, H. B. Keller, and J.-P. Kernévez, "Numerical analysis and control of bifurcation problems, I: Bifurcation in finite dimensions", *Internat. J. Bifur. Chaos Appl. Sci. Engrg.* **1**:3 (1991), 493–520. MR 93c:34001a Zbl 0876.65032

[Doedel et al. 1991b] E. Doedel, H. B. Keller, and J.-P. Kernévez, "Numerical analysis and control of bifurcation problems, II: Bifurcation in infinite dimensions", *Internat. J. Bifur. Chaos Appl. Sci. Engrg.* **1**:4 (1991), 745–772. MR 93c:34001a Zbl 0876.65060

[Glendinning 1994] P. Glendinning, *Stability, instability and chaos: an introduction to the theory of nonlinear differential equations*, Cambridge University Press, Cambridge, 1994. MR 96e:34017 Zbl 0808.34001

[Goyal et al. 2007] S. Goyal, T. Lillian, S. Blumberg, J.-C. Meiners, E. Meyhöfer, and N. C. Perkins, "Intrinsic curvature of DNA influences LacR-mediated looping", *Biophys. J.* **93**:12 (2007), 4342–4359.

[Iooss and Joseph 1980] G. Iooss and D. D. Joseph, *Elementary stability and bifurcation theory*, Springer, New York, 1980. MR 83e:34003 Zbl 0525.34001

[Kahn and Crothers 1998] J. D. Kahn and D. M. Crothers, "Measurement of the DNA bend angle induced by the catabolite activator protein using Monte Carlo simulation of cyclization kinetics", *J. Mol. Biol.* **276**:1 (1998), 287–309.

[Manning et al. 1996] R. S. Manning, J. H. Maddocks, and J. D. Kahn, "A continuum rod model of sequence-dependent DNA structure", *J. Chem. Phys.* **105**:13 (1996), 5626–5646.

[Marko and Siggia 1995] J. Marko and E. D. Siggia, "Stretching DNA", *Macromolecules* **28**:26 (1995), 8759–8770.

[Murdock 1999] J. A. Murdock, *Perturbations: Theory and methods*, Classics in Applied Mathematics **27**, Soc. Industrial Appl. Math., Philadelphia, 1999. MR 2000h:34088 Zbl 0810.34047

[Olson et al. 1998] W. K. Olson, A. A. Gorin, X. J. Lu, L. M. Hock, and V. B. Zhurkin, "DNA sequence-dependent deformability deduced from protein-DNA crystal complexes", *Proc. Natl. Acad. Sci. USA* **95**:19 (1998), 11163–11168.

[Seol et al. 2007] Y. Seol, J. Li, P. C. Nelson, T. T. Perkins, and M. D. Betterton, "Elasticity of short DNA molecules: Theory and experiment for contour lengths of 0.6–7 μm", *Biophys. J.* **93**:12 (2007), 4360–4359.

[Shifrin and Adams 2002] T. Shifrin and M. R. Adams, *Linear algebra: a geometric approach*, W. H. Freeman, New York, 2002.

[Swigon et al. 2006] D. Swigon, B. D. Coleman, and W. K. Olson, "Modeling the Lac repressor-operator assembly, I: The influence of DNA looping on Lac repressor conformation", *Proc. Natl. Acad. Sci. USA* **103**:26 (2006), 9879–9884.

Received: 2009-02-12 Accepted: 2009-05-02

kaitpeterson@gmail.com *Mathematics Department, Haverford College, 370 Lancaster Ave., Haverford, PA 19041, United States*

rmanning@haverford.edu *Mathematics Department, Haverford College, 370 Lancaster Ave., Haverford, PA 19041, United States*

INVOLVE 2:5(2009)

A tiling approach to Fibonacci product identities

Jacob Artz and Michael Rowell

(Communicated by Arthur T. Benjamin)

In 1998 Filipponi and Hart introduced a number of Fibonacci product identities. This paper provides a combinatorial proof for such identities via tilings. The methods used in the proof are then further used to produce some new Zeckendorf representations and a known Fibonacci identity.

1. Introduction

The discovery of the Fibonacci sequence is credited to Leonardo of Pisa (c. 1170–1250), posthumously nicknamed Fibonacci ("son of Bonaccio"). He is said to have come upon it while considering the breeding of rabbits [Russel 2008]. We define the Fibonacci sequence recursively.

Definition 1.1. Let $f_0 = 1$, $f_1 = 1$. For $n \geq 2$, set $f_n = f_{n-1} + f_{n-2}$. We say that f_n is the n-th *Fibonacci number*.

The first few terms of the Fibonacci sequence are $1, 1, 2, 3, 5, 8, 13, 21, \ldots$. Fibonacci numbers can also be explicitly defined, although the statement of each term is somewhat less elegant than its recursive counterpart:

$$F_n = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right].$$

For the purpose of this paper, we introduce an interpretation of the Fibonacci sequence which allows us to combinatorially prove Fibonacci identities. We will interpret the n-th Fibonacci number as the number of tilings of a $1 \times n$ board using 1×1 squares and 1×2 dominoes. Thus, $(f_0, f_1, f_2, f_3, f_4, \ldots) = (1, 1, 2, 3, 5, \ldots)$. (For more on this interpretation, see [Benjamin and Quinn 2003].)

Zeckendorf's Theorem says that any positive integer n can be represented as a sum of distinct, nonconsecutive Fibonacci numbers, excluding f_0. This was first published in [Lekkerkerker 1952], though it had been proved by Zeckendorf many years before. Finding the Zeckendorf representation of a particular number n is in

MSC2000: 05A19, 11B39.
Keywords: Fibonacci products, tiling.

fact easy: starting from n, successively subtract the largest Fibonacci number that will fit. (With some thinking this also justifies the theorem: basically, repeats are impossible because $f_n \le 2f_{n-1}$, so what's less at each stage is strictly less than what was just subtracted; and consecutive Fibonacci numbers don't occur, because if f_n and f_{n+1} occurred one would instead have used f_{n+2}. See also [Brown 1964].)

It is an interesting problem to find explicitly the Zeckendorf representation of various numbers. For example, the following identities are found in [Filipponi and Hart 1998]:

Theorem 1.2. *For $k \ge 0$ and $n \ge 2k$,*

$$f_{2k+1} f_n = f_{n+2k} + f_{n+2k-4} + \cdots + f_{n-2k+4} + f_{n-2k} = \sum_{i=0}^{k} f_{n+2k-4i}.$$

Theorem 1.3. *For $k \ge 1$ and $n \ge 2k$,*

$$f_{2k} f_n = f_{n+2k-1} + f_{n+2k-5} + \cdots + f_{n-2k+3} + f_{n-2k} = \sum_{i=0}^{k-1} f_{n+2k-1-4i} + f_{n-2k}.$$

Wood [2007] presents combinatorial proofs for the expansion of $f_4 f_n$ and $f_5 f_n$, but no unifying counting argument is shown. Gerdemann [2009] provides a combinatorial proof of the existence of Zeckendorf representations but, due to the algorithmic nature of the proof, is unable to produce closed form identities like the Fibonacci products provided above.

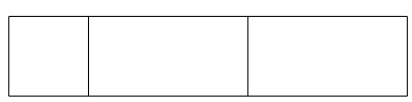
In the sequel we give examples of how tilings of one-row boards can help prove Fibonacci identities, starting with simple examples in Section 2 and continuing with provide combinatorial proofs for Theorems 1.2 and 1.3 in Section 3. In Section 4, we use the methods from Section 3 to find new Zeckendorf representations and a known Fibonacci identity. Directions for future research are given in Section 5.

2. Basic methods of tiling

A *cell* is a space of length one appearing on a $1 \times n$ board. While the meaning of this term may seem obvious, we wish to eradicate any confusion with the term *tile*, defined as a one- or two-cell piece used to create a tiling of the board.

We now illustrate two basic methods of showing Fibonacci identities using tilings; see Benjamin and Quinn [2003].

Considering the location of a fault. A *fault* in a tiling is the coordinate of any boundary between tiles. For example, in this tiling of length 5 with 3 tiles, the faults are at positions 1 and 3:



Theorem 2.1. *For $n, m \geq 1$,*

$$f_{m+n} = f_m f_n + f_{m-1} f_{n-1}.$$

Proof. The left side of our equation, f_{m+n}, is the number of ways of tiling a board of length $m+n$.

To interpret the right side, we consider the cell at position m of a board of length $m+n$. If the board has a fault at m, then there are f_m ways to tile the first m tiles, and f_n ways to tile the remaining n tiles. Therefore, there are $f_m f_n$ ways to tile the entire board. If the board does not have a fault at m, then there must be a domino covering cells m and $m+1$. Using similar reasoning as in our previous case, there are $f_{m-1} f_{n-1}$ ways to tile the entire board. Therefore, there are a total of $f_m f_n + f_{m-1} f_{n-1}$ ways to tile a board of length $m+n$. □

Finding correspondences. In the second method, we show identities by interpreting a side as multiple copies of board. We again give illustrate the method by giving the proof of another basic identity.

Theorem 2.2. *For $n \geq 2$,*

$$2 f_n = f_{n+1} + f_{n-2}.$$

Proof. We will show that two copies of each tiling of length n can be mapped to the set counted by tilings of length $n+1$ and those of length $n-2$. Begin by examining our first copies of tilings of length n. We append a square to each of our length-n tilings to form the set of all boards length of $n+1$ that end in a square tile. We now consider our second copies of our length-n tiling. If the last tile is a square, remove the last square and add a domino, resulting in all $n+1$ tilings that end in a domino. If the last tile is a domino, then remove the domino to form all tilings of length $n-2$. □

3. Fibonacci products and tiling

Before tackling Theorems 1.2 and 1.3, we begin with proving a simple, yet very useful identity,

Theorem 3.1. *For $k, n \geq 2$,*

$$f_k f_n = f_{n+k-1} + f_{k-2} f_{n-2}.$$

Proof. We begin by noting that the left side counts the number of tilings of length $n+k$ that have a fault at n.

We then note that $f_{n-2} f_{k-2}$ counts the number of tilings of length $n+k-4$ that have a fault at $n-2$. Inserting two dominoes between cells $n-2$ and $n-1$ yields all tilings of length $n+k$ with a fault at n and with dominoes on both sides of the fault.

We now inspect f_{n+k-1}, all tilings of a board length $n+k-1$. For all tilings that do *not* have a fault at n (implying there is a fault at $n-1$), we insert a square between cells $n-1$ and n, creating all tilings of length $n+k$ with a fault at n and with a square to the left of the fault and a domino to the right. For all tilings that *do* have a fault at n, we insert a square between cells n and $n+1$, creating all tilings of length $n+k$ with a fault at n and a square to the right of the fault.

The set of all tilings of length $n+k$ with a fault at n is exactly equal to the three cases defined above. □

We can now see that inductively applying this theorem to the term $f_k f_n$ will allow us to combinatorially prove Theorems 1.2 and 1.3. For example ($n \geq 6$),

$$\begin{aligned} f_7 f_n &= f_{n+6} + f_5 f_{n-2} = f_{n+6} + f_{n+2} + f_3 f_{n-4} \\ &= f_{n+6} + f_{n+2} + f_{n-2} + f_1 f_{n-6} = f_{n+6} + f_{n+2} + f_{n-2} + f_{n-6}. \end{aligned}$$

Proof of Theorem 1.2. The theorem is trivial for $k = 0$. Assuming the theorem is true for $k \leq m$, we see that

$$\begin{aligned} f_{2(m+1)+1} f_n &= f_{n+2m+2} + f_{2m+1} f_{n-2} && \text{by Theorem 3.1} \\ &= f_{n+2m+2} + \sum_{i=0}^{m} f_{n+2m-2-4i} && \text{by the inductive hypothesis} \\ &= \sum_{i=0}^{m+1} f_{n+2m+2-4i}. \end{aligned}$$ □

Proof of Theorem 1.3. In the case $k = 1$, our theorem reduces to Theorem 2.2. Assuming that the theorem is true for $k \leq m$, we see that

$$\begin{aligned} f_{2(m+1)} f_n &= f_{n+2m+1} + f_{2m} f_{n-2} && \text{by Theorem 3.1} \\ &= f_{n+2m+1} + \sum_{i=0}^{m-1} f_{n+2m+1-4i} + f_{n-2m} && \text{by inductive hypothesis} \\ &= \sum_{i=0}^{m} f_{n+2m+1-4i} + f_{n-2m}. \end{aligned}$$ □

We now have a unifying combinatorial proof for the following identities:

$$\begin{aligned} f_2 f_n &= f_{n+1} + f_{n-2}, \\ f_3 f_n &= f_{n+2} + f_{n-2}, \\ f_4 f_n &= f_{n+3} + f_{n-1} + f_{n-4}, \\ f_5 f_n &= f_{n+4} + f_n + f_{n-4}, \\ f_6 f_n &= f_{n+5} + f_{n+1} + f_{n-3} + f_{n-6}, \\ f_7 f_n &= f_{n+6} + f_{n+2} + f_{n-2} + f_{n-6}, \\ &\cdots \end{aligned}$$

We note that the proofs of Theorems 1.2 and 1.3 need not rely on induction. In Theorem 1.2, the term $f_{n+2k-4i}$ can be interpreted as all tilings of length $n+2k+1$ that have a fault at n and whose nearest square tile to the fault is exactly $2i$ cells away. Because $2k+1$ is odd, we are guaranteed to have a square tile within at most $2k$ of the fault, thus i ranges from 0 to k. Theorem 1.3 is slightly different in that we are not guaranteed to have a square tile within $2k-2$ of our fault. Thus, we must add the term f_{n-2k} to account for all tilings of length $n+2k$ with a square no closer than $2k$ cells from the fault at n.

4. Further observations

Using Theorem 3.1, we can determine other closed form Zeckendorf representations and a known Fibonacci identity.

Lemma 4.1. *For $k \geq 1$ and $n \geq 2k$,*

$$(f_{2k} + f_{2k-2}) f_n = f_{n+2k} + f_{n-2k}.$$

Proof. In the case $k = 1$, our lemma reduces to the case $k = 1$ in Theorem 1.2. Assuming our lemma is true for $k \leq m$ and applying Theorem 3.1 we see that

$$\begin{aligned}(f_{2m+2} + f_{2m}) f_n &= f_{n+2m+1} + f_{n+2m-1} + (f_{2m} + f_{2m-2}) f_{n-2} \\ &= f_{n+2m+1} + f_{n+2m-1} + f_{n+2m-2} + f_{n-2m-2} \\ &= f_{n+2m+2} + f_{n-2m-2},\end{aligned}$$

where our second to last line follows from the inductive hypothesis, and the last line follows from the recursive definition of the Fibonacci sequence. □

As with our two main theorems, we can prove Lemma 4.1 without induction. Using Theorem 1.3 we see that

$$(f_{2k} + f_{2k-2}) f_n = \sum_{i=0}^{2k-2} f_{2k-1+n-2i} + f_{n-2k+2} + f_{n-2k}.$$

It is left to show that

$$f_{n+2k} = \sum_{i=0}^{2k-2} f_{2k-1+n-2i} + f_{n-2k+2}.$$

This can be done by considering the position of the last square. Note that $f_{2k-1+n-2i}$ counts the number of tilings of length $n+2k$ with the last square in the $(n+2k+1-2i)$ cell followed by dominoes. Our sum accounts for all tilings of length $n+2k$ with the last square appearing somewhere past the $(n-2k+2)$ cell. We then see that f_{n-2k+2} counts the remaining tilings of length $n+2k$, namely those which end in $k-1$ dominoes.

Lemma 4.2. *For $k \geq 1$ and $n > 2k$,*

$$(f_{2k-1} + f_{2k+1}) f_n = f_{n+2k+1} - f_{n-2k-1}.$$

Proof. In the case $k = 1$, we repeatedly use our recursive definition of the Fibonacci sequence,

$$\begin{aligned} f_{n+3} - 4f_n &= (f_{n+2} + f_{n+1}) - 4f_n = (f_{n+1} + f_n) + (f_n + f_{n-1}) - 4f_n \\ &= \big((f_n + f_{n-1}) + f_n\big) + (f_n + f_{n-1}) - 4f_n = -f_n + 2f_{n-1} = f_{n-3}, \end{aligned}$$

where our last line follows from Theorem 2.2. Assuming our lemma is true for $k \leq m$ and applying Theorem 3.1 we see that

$$\begin{aligned} (f_{2m+1} + f_{2m+3}) f_n &= f_{2m+n} + f_{2m-1} f_{n-2} + f_{2m+2+n} + f_{2m+1} f_{n-2} \\ &= f_{n+2m} + f_{n+2m+2} + f_{n+2m-1} - f_{n-2m-3} \\ &= f_{n+2m+3} - f_{n-2m-3}, \end{aligned}$$

where our second to last line follows from using our inductive hypothesis and our last line follows from the recursive definition of the Fibonacci sequence. □

Using Theorem 1.2 and Lemma 4.1 we can now give the Zeckendorf representation of the following family of identities:

$$\begin{aligned} 3f_n &= (f_0 + f_2) f_n = f_{n+2} + f_{n-2}, \\ 4f_n &= (f_1 + f_3) f_n = f_{n+2} + f_n + f_{n-2}, \\ 7f_n &= (f_2 + f_4) f_n = f_{n+4} + f_{n-4}, \\ 11f_n &= (f_3 + f_5) f_n = f_{n+4} + f_{n+2} + f_n + f_{n-2} + f_{n-4}, \\ 18f_n &= (f_4 + f_6) f_n = f_{n+6} + f_{n-6}, \\ &\cdots \end{aligned}$$

Combining Lemmas 4.1 and 4.2 we obtain our Fibonacci identity, which appears in a more general form as Identity 48 in [Benjamin and Quinn 2003].

Theorem 4.3. *For $k \geq 1$ and $n > k$,*

$$(f_{k+1} + f_{k-1}) f_n = f_{n+k+1} - (-1)^k f_{n-k-1}. \tag{4-1}$$

5. Future work

As previously mentioned, Theorems 1.2 and 1.3 first appeared in [Filipponi and Hart 1998] with some other Fibonacci products. The authors also presented Zeckendorf representations for $2f_k f_n, 3f_k f_n, 4f_k f_n$ and $5f_k f_n$. It is not difficult to see that these formulas can also be obtained using our counting method presented above. One only needs to determine the appropriate recurrence relation.

Also, while the proofs in Section 4 relied on the use of Theorem 3.1, it is not entirely clear how to construct a counting proof for Theorem 4.3. It would be interesting to see a combinatorial proof of the identity.

Acknowledgment

The authors would like to thank the reviewers for their comments and suggestions. In particular, it was quite helpful to consider combinatorial proofs without the use of induction.

References

[Benjamin and Quinn 2003] A. T. Benjamin and J. J. Quinn, *Proofs that really count: The art of combinatorial proof*, The Dolciani Mathematical Expositions **27**, Mathematical Association of America, Washington, DC, 2003. MR 2004f:05001 Zbl 1044.11001

[Brown 1964] J. L. Brown, "A new characterization of the Fibonacci numbers", *Fibonacci Quart.* **2** (1964), 163–168.

[Filipponi and Hart 1998] P. Filipponi and E. L. Hart, "The Zeckendorf decomposition of certain Fibonacci–Lucas products", *Fibonacci Quart.* **36**:3 (1998), 240–247. MR 99d:11006 Zbl 0942.11012

[Gerdemann 2009] D. Gerdemann, "Combinatorial proofs of Zeckendorf family identities", *Fibonacci Quart.* **46/47**:3 (2009), 249–261.

[Lekkerkerker 1952] C. G. Lekkerkerker, "Voorstelling van natuurlijke getallen door een som van getallen van Fibonacci", *Simon Stevin* **29** (1952), 190–195. MR 15,401c

[Russel 2008] D. Russel, "A short biography of Leonardo Pisano Fibonacci", 2008, Available at http://math.about.com/od/mathematicians/a/fibonacci.htm.

[Wood 2007] P. M. Wood, "Bijective proofs for Fibonacci identities related to Zeckendorf's theorem", *Fibonacci Quart.* **45**:2 (2007), 138–145 (2008). MR 2009b:05032 Zbl 1162.11014

Received: 2009-08-04 Revised: 2009-09-14 Accepted: 2009-09-19

artz8028@pacificu.edu *Mathematics Department, Pacific University, 2043 College Way, Forest Grove, OR 97116, United States*

rowell@pacificu.edu *Mathematics Department, Pacific University, 2043 College Way, Forest Grove, OR 97116, United States*
www.math.pacificu.edu/~rowell

INVOLVE 2:5(2009)

Frame theory for binary vector spaces

Bernhard G. Bodmann, My Le, Letty Reza,
Matthew Tobin and Mark Tomforde

(Communicated by David Larson)

We develop the theory of frames and Parseval frames for finite-dimensional vector spaces over the binary numbers. This includes characterizations which are similar to frames and Parseval frames for real or complex Hilbert spaces, and the discussion of conceptual differences caused by the lack of a proper inner product on binary vector spaces. We also define switching equivalence for binary frames, and list all equivalence classes of binary Parseval frames in lowest dimensions, excluding cases of trivial redundancy.

1. Introduction

There are many conceptual similarities between frames and error-correcting linear codes. Frame theory is concerned with stable linear embeddings of Hilbert spaces obtained from mapping a vector to its frame coefficients [Duffin and Schaeffer 1952; Christensen 2003; Han et al. 2007]. The linear dependencies incorporated in the frame coefficients of a vector help recover from errors such as noise, quantization and data loss [Goyal et al. 1998; 2001; Rath and Guillemot 2003; 2004; Püschel and Kovačević 2006], just as linear codes help recover from symbol decoding errors and erasures [MacWilliams and Sloane 1977]. Frame design for specific purposes has been related to optimization problems of a geometric nature [Casazza and Kovačević 2003; Strohmer and Heath 2003; Holmes and Paulsen 2004] or even a discrete one [Bodmann and Paulsen 2005; Xia et al. 2005; Kalra 2006], including combinatorial considerations that are more commonly associated with error-correcting codes. On the other hand, one may ask whether concepts from frame theory yield insights in the binary setting. This is the motivation of the present paper.

MSC2000: 15A03, 15A33, 42C15.

Keywords: frames, binary numbers, Parseval frames, finite-dimensional vector spaces, binary numbers, binary vector spaces.

Parts of this research were supported by NSF Grant DMS-0807399, and by DMS-0604429 (REU supplement).

We translate many of the essential results on frames for finite-dimensional real or complex Hilbert spaces to analogous statements for vector spaces over the binary numbers. In the first part, we show that in the binary case, the spanning property of a family of vectors is equivalent to having a reconstruction identity with a dual family. This means, both properties can be used interchangeably as a definition of frames, as on finite dimensional real or complex Hilbert spaces. On the other hand, we demonstrate that an attempt to define binary frames similarly to the real or complex case via norm inequalities fails in binary vector spaces, because they lack an inner product and a polarization identity. In the main part of this paper, we focus on Parseval frames, which have a particularly simple reconstruction identity. We characterize binary Parseval frames in terms of their frame operator and develop a notion of switching equivalence for binary frames, similar to the concept for real or complex frames [Goyal et al. 2001; Holmes and Paulsen 2004; Bodmann and Paulsen 2005]. Moreover, we introduce the notion of trivial redundancy, caused by repeated vectors or the inclusion of the zero vector in the frame. Ignoring cases of trivial redundancy and choosing representatives from each switching equivalence class simplifies the enumeration of binary Parseval frames. By an exhaustive search, we have found that if $k \in \{4, 5, \dots, 11\}$, then all frames that are not trivially redundant in $\mathbb{Z}_2^4$ with k vectors belong to one switching equivalence class. Further simplifications for the search of all binary Parseval frames are obtained from a combinatorial consideration, which might be useful for a future effort to catalogue binary Parseval frames in larger dimensions.

The remainder of this paper is organized as follows. In Section 2, we define frames for finite-dimensional binary vector spaces. Section 3 specializes the discussion to Parseval frames. Finally, in Section 4, we define switching equivalence for binary frames and give a catalogue of representatives from each equivalence class of Parseval frames in lowest dimensions, excluding trivially redundant ones.

2. Preliminaries

In this section we first revisit the essentials of frames over the fields $\mathbb{R}$ or $\mathbb{C}$, the real or complex numbers. We then proceed to develop the concept of frames over the field $\mathbb{Z}_2$, that is, the field with two elements $\{0, 1\}$, where 0 is the neutral element with respect to addition, and 1 is the neutral element with respect to multiplication. The main insight of this section is that while there are equivalent characterizations of certain types of frames when the ground field is $\mathbb{R}$ or $\mathbb{C}$, this is not true over $\mathbb{Z}_2$, because the polarization identity is no longer available due to the lack of an inner product.

If $\mathcal{H}$ is a finite-dimensional Hilbert space over $\mathbb{R}$ or $\mathbb{C}$ with inner product $\langle \cdot, \cdot \rangle$, then a family of vectors $\mathcal{F} := \{f_1, f_2, \dots, f_k\}$ in $\mathcal{H}$ is called a *frame* if there exist

real numbers A and B such that $0 < A \leq B < \infty$ and

$$A\|x\|^2 \leq \sum_{j=1}^{k} |\langle x, f_j\rangle|^2 \leq B\|x\|^2 \qquad \text{for all } x \in \mathcal{H}. \tag{2-1}$$

The inequalities displayed in (2-1) are known as the *frame condition*, and it can be shown that when $\mathcal{H}$ is finite dimensional, then the set $\mathcal{F}$ satisfies the frame condition if and only if $\operatorname{span} \mathcal{F} = \mathcal{H}$ [Han et al. 2007, Proposition 3.18]. In this case, there exist vectors $\{g_1, g_2, \dots, g_k\}$ which provide the *reconstruction identity*

$$x = \sum_{j=1}^{k} \langle x, f_j\rangle g_j \qquad \text{for all } x \in \mathcal{H}.$$

While the family $\{g_j\}_{j=1}^k$ may not be unique, there is a canonical choice. If we define the so-called *frame operator* S on $\mathcal{H}$ by $Sx = \sum_{j=1}^{k} \langle x, f_j\rangle f_j$, then setting $g_j = S^{-1} f_j$ for $j \in \{1, 2, \dots, k\}$ yields the reconstruction identity [Christensen 2003]. The family $\{g_j\}_{j=1}^k$ is also called the *canonical dual frame*.

A frame $\mathcal{F} = \{f_1, \dots, f_k\}$ is called a *Parseval frame* (or sometimes a *normalized tight frame*) if we can choose $A = B = 1$ in the frame condition, so that

$$\sum_{j=1}^{k} |\langle x, f_j\rangle|^2 = \|x\|^2 \qquad \text{for all } x \in \mathcal{H}. \tag{2-2}$$

Using the polarization identity, it can be shown (see [Han et al. 2007, Proposition 3.11]) that $\mathcal{F}$ is a Parseval frame if and only if

$$x = \sum_{j=1}^{k} \langle x, f_j\rangle f_j \qquad \text{for all } x \in \mathcal{H}. \tag{2-3}$$

The simple form of the reconstruction formula for Parseval frames has many practical uses in engineering and computer science [Goyal et al. 1998; 2001; Kovačević and Chebira 2008].

We now turn to frames over the binary numbers.

The first two goals in this paper are to develop the notion of frames and of Parseval frames for finite-dimensional vector spaces over the field $\mathbb{Z}_2$. Any such vector space has the form $\mathbb{Z}_2^n = \mathbb{Z}_2 \oplus \cdots \oplus \mathbb{Z}_2$ for some $n \in \mathbb{N}$.

Definition 2.1. A family of vectors $\mathcal{F} = \{f_1, f_2, \dots, f_k\}$ in $\mathbb{Z}_2^n$ is a *frame* if it spans $\mathbb{Z}_2^n$.

We have chosen this form of the definition because the field $\mathbb{Z}_2$ has no notion of positive elements, so that it is impossible to find a properly defined inner product, let alone a norm on $\mathbb{Z}_2^n$, which would be needed to formulate a direct analogue of the frame condition (2-1).

Nevertheless, we want to show that an analogue of the reconstruction identity can be deduced with the help of a $\mathbb{Z}_2$-valued "dot product" in place of an inner product.

Definition 2.2. We define a bilinear map $(\cdot,\cdot) : \mathbb{Z}_2^n \times \mathbb{Z}_2^n \to \mathbb{Z}_2$, called the *dot product* on $\mathbb{Z}_2^n$, by

$$\left(\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}, \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}\right) := \sum_{i=1}^{n} a_i b_i .$$

We see that the dot product $(\cdot,\cdot)$ is symmetric and $\mathbb{Z}_2$-linear in each component, but it is degenerate: It is possible to have $x \in \mathbb{Z}_2^n$ with $(x,x) = 0$ but $x \neq 0$. Furthermore, because the dot product is degenerate, it does not provide a norm on $\mathbb{Z}_2^n$. Nonetheless, we will use the dot product as an analogue of the inner products on $\mathbb{R}^n$ and $\mathbb{C}^n$, and for expressions in $\mathbb{R}^n$ or $\mathbb{C}^n$ involving $\langle x, y\rangle$ or $\|x\|^2$, we shall consider analogous expressions in $\mathbb{Z}_2^n$ involving (x, y) or (x, x), respectively.

To establish the equivalence between the spanning property and the reconstruction identity for frames, we unfortunately cannot simply use the same strategy as in the real or complex case. If we take the dot product instead of an inner product to define the frame operator, then the spanning property of the frame does not guarantee that the frame operator is invertible. To see this, we note that the family $\{1, 1\}$ is spanning for $\mathbb{Z}_2$, but the analogue of the frame operator maps every $x \in \mathbb{Z}_2$ to $x + x = 0$. A similar family can be obtained for any $\mathbb{Z}_2^n$, $n \geq 1$, by repeating vectors of an arbitrary spanning set.

To build an alternative strategy that relates the spanning property with the existence of a reconstruction identity, we first recall that the dot product mediates a canonical mapping between vectors and linear functionals.

Lemma 2.3. *If $\phi : \mathbb{Z}_2^n \to \mathbb{Z}_2$ is a linear functional then there exists a unique $z \in \mathbb{Z}_2^n$ such that $\phi(x) = (x, z)$ for all $x \in \mathbb{Z}_2^n$.*

Proof. Let ϕ be a linear functional. Let $\{e_1,\ldots,e_n\}$ be the canonical basis for $\mathbb{Z}_2^n$, and let $z = \phi(e_1)e_1 + \cdots + \phi(e_n)e_n$. We now observe that if $x \in \mathbb{Z}_2^n$, with $x = \sum_{i=1}^n a_i e_i$ for $a_i \in \mathbb{Z}_2$, then $\phi(x) = \sum_{i=1}^n a_i \phi(e_i) = (x, z)$.

To verify the uniqueness, assume there is z' such that $\phi(x) = (x, z') = (x, z)$. Choosing x among the canonical basis vectors gives $\phi(e_i) = (e_i, z') = (e_i, z)$ and thus z and z' are identical. □

Theorem 2.4. *Given a family $\mathcal{F} = \{f_j\}_{j=1}^k$ in $\mathbb{Z}_2^n$, then $\mathcal{F}$ is a frame if and only if there exist vectors $\{g_j\}_{j=1}^k$ such that for all $y \in \mathbb{Z}_2^n$*

$$y = \sum_{j=1}^{k} (y, g_j) f_j. \tag{2-4}$$

Proof. We note that if (2-4) is true, then necessarily $\{f_j\}_{j=1}^k$ is spanning.

Conversely, assume that $\{f_j\}_{j=1}^k$ is a frame for $\mathbb{Z}_2^n$. In a first step, we prove that there are linear functionals $\{\gamma_1, \gamma_2, \dots, \gamma_k\}$ such that $y = \sum_{j=1}^k \gamma_j(y) f_j$ for all $y \in \mathbb{Z}_2^n$. For any family of linear functionals $\gamma_1, \gamma_2, \dots, \gamma_k$, we note that the expression $\sum_{j=1}^k \gamma_j(y) f_j$ is linear in y, so it is enough to show that there exist linear functionals giving

$$w_i = \sum_{j=1}^k \gamma_j(w_i) f_j \quad \text{for all vectors in some basis } w_1, \dots, w_n \text{ of } \mathbb{Z}_2^n.$$

To establish this, we choose a subset of $\{f_1, \dots, f_k\}$ which is spanning and linearly independent, that is, a basis. Without loss of generality, assume that this set is $\{f_1, \dots, f_n\}$. Choosing the dual basis $\{\gamma_1, \dots, \gamma_n\}$ to $\{f_1, \dots, f_n\}$, characterized by

$$\gamma_j(f_i) = \delta_{ij}, \quad \text{for all } i, j \in \{1, 2, \dots n\},$$

we obtain

$$\sum_{j=1}^n \gamma_j(f_i) f_j = f_i.$$

Thus if we enlarge the set $\{\gamma_j\}_{j=1}^n$ by setting $\gamma_j = 0$ if $j > n$, then

$$f_i = \sum_{j=1}^k \gamma_j(f_i) f_j$$

and by linearity

$$y = \sum_{j=1}^k \gamma_j(y) f_j \quad \text{for any } y \in \mathbb{Z}_2^n.$$

In the final step of the proof, we apply the preceding lemma which yields for each γ_j a corresponding vector g_j satisfying $\gamma_j(y) = (y, g_j)$ for all $y \in \mathbb{Z}_2^n$. □

3. Parseval frames for $\mathbb{Z}_2^n$

In this section we present the definition of Parseval frames for $\mathbb{Z}_2^n$ and illustrate the conceptual differences between such frames in the real or complex case and in the binary case.

Definition 3.1. A family of vectors $\mathcal{F} = \{f_1, \dots, f_k\}$ in $\mathbb{Z}_2^n$ is a *Parseval frame* if

$$x = \sum_{j=1}^k (x, f_j) f_j \qquad \text{for all } x \in \mathbb{Z}_2^n. \tag{3-1}$$

Observe that a binary Parseval frame necessarily spans $\mathbb{Z}_2^n$, and moreover if $\mathcal{F}$ is a Parseval frame, we must have $k \geq n$.

It is natural to ask if, in analogy with the real and complex cases, being a Parseval frame in $\mathbb{Z}_2^n$ is equivalent to having a Parseval identity as in (2-2). It turns out that this is not the case.

Proposition 3.2. *If $\mathcal{F} = \{f_1, \dots, f_k\}$ is a Parseval frame for $\mathbb{Z}_2^n$, then*

$$\sum_{j=1}^{k} (x, f_j)^2 = (x, x) \qquad \text{for all } x \in \mathbb{Z}_2^n. \tag{3-2}$$

However, in general, the converse does not hold.

Proof. If $\mathcal{F}$ is a Parseval frame, then using the $\mathbb{Z}_2$-linearity of the first component of the dot product, for any $x \in \mathbb{Z}_2^n$ we have

$$(x, x) = \left(\sum_{j=1}^{k} (x, f_j) f_j, x \right) = \sum_{j=1}^{k} (x, f_j)(f_j, x) = \sum_{j=1}^{k} (x, f_j)^2.$$

To see that the converse does not hold in general, consider $\binom{1}{1} \in \mathbb{Z}_2^2$, then for any $x = \binom{a_1}{a_2} \in \mathbb{Z}_2^2$ we have

$$\left(x, \binom{1}{1}\right) = a_1 + a_2 = a_1^2 + a_2^2 = (x, x).$$

Hence $\mathcal{F} = \left\{\binom{1}{1}\right\}$ satisfies (3-2). However, $\mathcal{F}$ contains one element, so $\mathcal{F}$ does not span $\mathbb{Z}_2^2$, and $\mathcal{F}$ is not a Parseval frame. □

Remark 3.3. More generally, we can produce counterexamples for any $n \geq 2$, meaning sets which give the Parseval property without spanning $\mathbb{Z}_2^n$. First we consider even n. Let $\{f_1, \dots, f_k\}$ be the family of all vectors which contain exactly two 1's. Thus, there are $k = \binom{n}{2}$ such vectors. If the first vector is chosen as $f_1 = (1, 1, 0, \dots, 0)^t$ and $x = (a_1, a_2, \dots, a_n)^t$, then over $\mathbb{Z}_2$,

$$(x, f_1)^2 = (a_1 + a_2)^2 = a_1^2 + a_2^2 .$$

Evaluating other dot products similarly gives

$$\sum_{j=1}^{k} (x, f_j)^2 = \sum_{i=1}^{n} a_i^2$$

because each a_i^2 appears in $n - 1$ terms in the sum, and $n - 1 \bmod 2 = 1$ by the assumption that n is even.

However, the vectors $\{f_j\}_{j=1}^{k}$ are not spanning for $\mathbb{Z}_2^n$, because they contain an even number of 1's and so does any linear combination of them.

If n is odd, then we split $\mathbb{Z}_2^n = \mathbb{Z}_2 \oplus \mathbb{Z}_2^{n-1}$ and construct the above family $\{f_1, f_2, \dots, f_k\}$ for the second summand. Now this family can be enlarged by

the first canonical basis vector e_1 to $\{e_1, f_1, f_2, \dots, f_k\}$ which has the Parseval property but is not spanning, because $\{f_1, f_2, \dots, f_k\}$ does not span $\mathbb{Z}_2^{n-1}$.

4. Towards a catalogue of binary Parseval frames

In principle, all Parseval frames for $\mathbb{Z}_2^n$ could be catalogued individually, but even for relatively small n this is already an extensive list. In order to obtain a more efficient way of enumerating Parseval frames, we use an equivalence relation which has been called switching equivalence for real or complex frames [Goyal et al. 2001; Holmes and Paulsen 2004; Bodmann and Paulsen 2005]. It is most easily formulated in terms of the Grammian of a Parseval frame, as defined below. The catalogue of frames can then be reduced to representatives of each equivalence class. To prepare the definition of the equivalence relation, we discuss certain matrices related to frames.

We write $A \in M_{m,n}(\mathbb{Z}_2)$ when A an $m \times n$ matrix with entries in $\mathbb{Z}_2$. We often view A as a linear map from $\mathbb{Z}_2^m$ to $\mathbb{Z}_2^n$ by left multiplication. In particular, $A \in M_n$ denotes an $n \times n$ matrix which is associated with a map from $\mathbb{Z}_2^n$ to itself. We write $A_{i,j}$ for the (i, j)th entry of A, and we let A^* denote the transpose of A; that is, $A^* \in M_{n,m}(\mathbb{Z}_2)$ with $A^*_{i,j} := A_{j,i}$. By the rules of matrix multiplication, we have $(Ax, y) = (x, A^*y)$ for all $A \in M_n(\mathbb{Z}_2)$.

Definition 4.1. If $U \in M_n(\mathbb{Z}_2)$, then we say U is a *unitary* if U is invertible and $U^{-1} = U^*$.

Lemma 4.2. *If $x \in \mathbb{Z}_2^n$ and $(x, y) = 0$ for all $y \in \mathbb{Z}_2^n$, then $x = 0$.*

Proof. Write

$$x = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}.$$

If $\{e_1, \dots, e_n\}$ is the standard basis for $\mathbb{Z}_2^n$, then for all $1 \le i \le n$ we have $a_i = (x, e_i) = 0$. Thus $x = 0$. □

Proposition 4.3. *Let $U \in M_n(\mathbb{Z}_2)$, then U is a unitary if and only if for all $x, y \in \mathbb{Z}_2^n$ we have $(Ux, Uy) = (x, y)$.*

Proof. If U is a unitary, then $U^* = U^{-1}$ and for all $x, y \in \mathbb{Z}_2^n$ we have

$$(Ux, Uy) = (x, U^*Uy) = (x, Iy) = (x, y).$$

Conversely, if $(Ux, Uy) = (x, y)$ for all $x, y \in \mathbb{Z}_2^n$, then for a given $x \in \mathbb{Z}_2^n$ we see that $(U^*Ux, y) = (Ux, Uy) = (x, y)$ for all $y \in \mathbb{Z}_2^n$, and Lemma 4.2 implies that $U^*Ux = x$. Since x was arbitrary, this shows that $U^*U = I$, and because U is square, we have that U is invertible and $U^{-1} = U^*$. □

In contrast the case of Hilbert spaces over $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$, the condition $\langle Ux, Ux\rangle = \langle x, x\rangle$ for all $x \in \mathbb{F}^n$ is not equivalent to unitarity when the field $\mathbb{F}$ is $\mathbb{Z}_2$.

We have the following counterexamples for $n \geq 2$.

Proposition 4.4. *For any $n \geq 2$, there exist $A \in M_n(\mathbb{Z}_2)$ such that $(Ax, Ax) = (x, x)$ for all $x \in \mathbb{Z}_2^n$ but A is not invertible, and thus not unitary.*

Proof. We define the matrix A by

$$A_{i,j} = \begin{cases} 1 & \text{if } i = j = 1 \text{ or } j - i = 1\,, \\ 0 & \text{else.} \end{cases}$$

This means, the last row of A contains only zeros and thus A does not have full rank and is not invertible.

However, given $x = (a_1, a_2, \ldots, a_n)^t$ we have

$$Ax = \begin{pmatrix} a_1 + a_2 \\ a_3 \\ a_4 \\ \vdots \\ a_n \\ 0 \end{pmatrix},$$

and thus

$$(Ax, Ax) = (a_1 + a_2)^2 + a_3^2 + \cdots + a_n^2 = \sum_{i=1}^{n} a_i^2 = (x, x). \qquad \square$$

Definition 4.5. Let $\mathcal{F} = \{f_1, \ldots, f_k\} \subseteq \mathbb{Z}_2^n$. The *analysis operator* for $\mathcal{F}$ is the $k \times n$ matrix containing the frame vectors as rows,

$$\Theta_{\mathcal{F}} = \begin{pmatrix} \leftarrow & f_1 & \rightarrow \\ & \vdots & \\ \leftarrow & f_k & \rightarrow \end{pmatrix}.$$

The *synthesis operator* for $\mathcal{F}$ is the $n \times k$ matrix

$$\Theta_{\mathcal{F}}^* = \begin{pmatrix} \uparrow & & \uparrow \\ f_1 & \cdots & f_k \\ \downarrow & & \downarrow \end{pmatrix},$$

with the elements of $\mathcal{F}$ as columns. The *frame operator* for $\mathcal{F}$ is the $n \times n$ matrix

$$S_{\mathcal{F}} := \Theta_{\mathcal{F}}^* \Theta_{\mathcal{F}},$$

and the *Grammian operator* for $\mathcal{F}$ is the $k \times k$ matrix

$$G_{\mathcal{F}} := \Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^*.$$

Note that $(G_{\mathcal{F}})_{i,j} = (f_j, f_i)$ for all $1 \le i, j \le k$. When there is no ambiguity in the choice of $\mathcal{F}$, we shall omit the $\mathcal{F}$ subscript on these matrices and simply write Θ, Θ^*, S, and G.

Theorem 4.6. *Let $\mathcal{F} = \{f_1, \dots, f_k\} \subseteq \mathbb{Z}_2^n$, then $\mathcal{F}$ is a Parseval frame if and only if $S_{\mathcal{F}}$ is equal to the identity matrix.*

Proof. Let $\{e_1, \dots, e_n\}$ be the standard basis for $\mathbb{Z}_2^k$. Observe that for any $x \in \mathbb{Z}_2^n$ we have $\Theta_{\mathcal{F}} x = \sum_{i=1}^k (x, f_i) e_i$. Also, for any $1 \le i \le n$ we have $\Theta_{\mathcal{F}}^* e_i = f_i$. Thus we have

$$S_{\mathcal{F}} x = \Theta_{\mathcal{F}}^* \Theta_{\mathcal{F}} x = \Theta_{\mathcal{F}}^* \left(\sum_{i=1}^k (x, f_i) e_i \right) = \sum_{i=1}^k (x, f_i) \Theta_{\mathcal{F}}^* e_i = \sum_{i=1}^k (x, f_i) f_i.$$

It follows that $\sum_{i=1}^k (x, f_i) f_i = x$ for all $x \in \mathbb{Z}_2^n$ if and only if $S_{\mathcal{F}} x = x$ for all $x \in \mathbb{Z}_2^n$. Thus $\mathcal{F}$ is a Parseval frame if and only if $S_{\mathcal{F}}$ is the identity matrix. □

Definition 4.7. Given two families $\mathcal{F} = \{f_1, \dots, f_k\}$ and $\mathcal{G} = \{g_1, \dots, g_k\}$ in $\mathbb{Z}_2^n$, then we say $\mathcal{F}$ *is unitarily equivalent to* $\mathcal{G}$ if there exists a unitary $U \in M_n(\mathbb{Z}_2)$ such that $U f_i = g_i$ for all $1 \le i \le k$.

It is easy to show that unitary equivalence is an equivalence relation.

Proposition 4.8. *Let $\mathcal{F} = \{f_1, \dots, f_k\} \subseteq \mathbb{Z}_2^n$ and $\mathcal{G} = \{g_1, \dots, g_k\} \subseteq \mathbb{Z}_2^n$ be Parseval frames, then $\mathcal{F}$ is unitarily equivalent to $\mathcal{G}$ if and only if $G_{\mathcal{F}} = G_{\mathcal{G}}$.*

Proof. Since $\mathcal{F}$ and $\mathcal{G}$ are Parseval frames, it follows from Theorem 4.6 that $S_{\mathcal{F}}$ and $S_{\mathcal{G}}$ are the identity matrices. Suppose that $G_{\mathcal{F}} = G_{\mathcal{G}}$. Define U to be the $n \times n$ matrix $U := \Theta_{\mathcal{G}}^* \Theta_{\mathcal{F}}$, then

$$\begin{aligned} U^* U &= (\Theta_{\mathcal{G}}^* \Theta_{\mathcal{F}})^* \Theta_{\mathcal{G}}^* \Theta_{\mathcal{F}} = \Theta_{\mathcal{F}}^* \Theta_{\mathcal{G}} \Theta_{\mathcal{G}}^* \Theta_{\mathcal{F}} = \Theta_{\mathcal{F}}^* G_{\mathcal{G}} \Theta_{\mathcal{F}} \\ &= \Theta_{\mathcal{F}}^* G_{\mathcal{F}} \Theta_{\mathcal{F}} = \Theta_{\mathcal{F}}^* \Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^* \Theta_{\mathcal{F}} = S_{\mathcal{F}} S_{\mathcal{F}} = I. \end{aligned}$$

Since U is square, it follows that U is invertible and $U^{-1} = U^*$, so that U is a unitary. Furthermore,

$$U \Theta_{\mathcal{F}}^* = \Theta_{\mathcal{G}}^* \Theta_{\mathcal{F}} \Theta_{\mathcal{F}}^* = \Theta_{\mathcal{G}}^* G_{\mathcal{F}} = \Theta_{\mathcal{G}} G_{\mathcal{G}} = \Theta_{\mathcal{G}}^* \Theta_{\mathcal{G}} \Theta_{\mathcal{G}}^* = S_{\mathcal{G}} \Theta_{\mathcal{G}}^* = \Theta_{\mathcal{G}}^*.$$

Thus U times the ith column of $\Theta_{\mathcal{F}}^*$ is equal to the ith column of $\Theta_{\mathcal{G}}^*$. Thus for all $1 \le i \le k$ we have $U f_i = g_i$, so that $\mathcal{F}$ and $\mathcal{G}$ are unitarily equivalent.

Conversely, if $\mathcal{F}$ and $\mathcal{G}$ are unitarily equivalent, then there exists a unitary $U \in M_n(\mathbb{Z}_2)$ such that $U f_i = g_i$ for all $1 \le i \le k$. Thus Proposition 4.3 implies that

$$(G_{\mathcal{F}})_{i,j} = (f_j, f_i) = (U f_j, U f_i) = (g_j, g_i) = (G_{\mathcal{G}})_{i,j}.$$

Hence $G_{\mathcal{F}} = G_{\mathcal{G}}$. □

Example 4.9. We present two examples of unitary equivalence. First set

$$\mathcal{F} = \left\{ \begin{pmatrix} 0\\1\\0\\0\\1 \end{pmatrix}, \begin{pmatrix} 0\\1\\0\\1\\1 \end{pmatrix}, \begin{pmatrix} 0\\1\\1\\0\\1 \end{pmatrix}, \begin{pmatrix} 1\\0\\1\\1\\1 \end{pmatrix}, \begin{pmatrix} 1\\1\\0\\0\\1 \end{pmatrix}, \begin{pmatrix} 1\\1\\1\\1\\0 \end{pmatrix} \right\},$$

$$\mathcal{H} = \left\{ \begin{pmatrix} 0\\1\\0\\1\\0 \end{pmatrix}, \begin{pmatrix} 0\\1\\0\\1\\1 \end{pmatrix}, \begin{pmatrix} 0\\1\\1\\1\\0 \end{pmatrix}, \begin{pmatrix} 1\\0\\1\\1\\1 \end{pmatrix}, \begin{pmatrix} 1\\1\\0\\1\\0 \end{pmatrix}, \begin{pmatrix} 1\\1\\1\\0\\1 \end{pmatrix} \right\}.$$

Computing the Grammian for both $\mathcal{F}$ and $\mathcal{H}$ we find

$$G_{\mathcal{F}} = G_{\mathcal{H}} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

First notice the structure of Θ^* created by $\mathcal{F}$ and $\mathcal{H}$:

$$\Theta^*_{\mathcal{F}} = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}, \quad \Theta^*_{\mathcal{H}} = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}.$$

The fourth and fifth rows have swapped places, so naturally one would expect the unitary operator to reflect that. In fact, the proof gives a direct way to compute U, namely if $f_i = Uh_i$ then $U = \Theta^*_{\mathcal{F}}\Theta_{\mathcal{H}}$. Computing U confirms this:

$$\Theta^*_{\mathcal{F}}\Theta_{\mathcal{H}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Next, take for $\mathcal{F}$ and $\mathcal{H}$ two Parseval frames found in $\mathbb{Z}_2^5$ with six elements:

$$\mathcal{F} = \left\{ \begin{pmatrix} 0\\0\\0\\1\\1 \end{pmatrix}, \begin{pmatrix} 0\\0\\1\\0\\1 \end{pmatrix}, \begin{pmatrix} 0\\1\\0\\0\\1 \end{pmatrix}, \begin{pmatrix} 1\\0\\0\\0\\1 \end{pmatrix}, \begin{pmatrix} 1\\1\\1\\1\\0 \end{pmatrix}, \begin{pmatrix} 1\\1\\1\\1\\1 \end{pmatrix} \right\},$$

$$\mathcal{H} = \left\{ \begin{pmatrix} 0\\1\\1\\1\\1 \end{pmatrix}, \begin{pmatrix} 1\\0\\0\\0\\1 \end{pmatrix}, \begin{pmatrix} 1\\0\\0\\1\\0 \end{pmatrix}, \begin{pmatrix} 1\\0\\1\\0\\0 \end{pmatrix}, \begin{pmatrix} 1\\1\\0\\0\\0 \end{pmatrix}, \begin{pmatrix} 1\\1\\1\\1\\1 \end{pmatrix} \right\}.$$

Here, while not quite as obvious, differences in the two Parseval frames can be expressed in terms of row manipulations of the synthesis operator, which amount to left multiplication with a unitary U, $\Theta^*_{\mathcal{F}} = U\Theta^*_{\mathcal{H}}$.

We introduce an additional way to identify frames which coarsens the equivalence relation.

Definition 4.10. Two families $\mathcal{F} = \{f_1, f_2, \dots, f_k\}$ and $\mathcal{G} = \{g_1, g_2, \dots, g_k\}$ in $\mathbb{Z}^n$ are called *switching equivalent* if there is a unitary U and a permutation π of the set $\{1, 2, \dots, k\}$ such that

$$f_j = U g_{\pi(j)} \quad \text{for all } j \in \{1, 2, \dots\}.$$

Theorem 4.11. *Two Parseval frames $\mathcal{F}$ and $\mathcal{H}$ are switching equivalent if and only if there exists a permutation π of the index set such that $(G_{\mathcal{F}})_{i,j} = (G_{\mathcal{H}})_{\pi(i),\pi(j)}$.*

Proof. The condition on the Grammians amounts to the identity

$$G_{\mathcal{F}} = M G_{\mathcal{H}} M^*$$

for a permutation matrix with entries

$$M_{i,j} = \begin{cases} 1 & \text{if } \pi(i) = j, \\ 0 & \text{else}. \end{cases}$$

Being identical up to conjugation by permutation matrices defines an equivalence relation for Grammians, and thus for frames, which is coarser than unitary equivalence.

Moreover, with a similar proof as in the preceding proposition, we see that the two Grammians are related by conjugation with a permutation matrix M if and only if there exists a unitary U such that

$$\Theta^*_{\mathcal{F}} = U\Theta^*_{\mathcal{H}} M^*. \qquad \square$$

Apart from switching equivalence, there are other simple ways in which two Parseval frames can be related to each other. For example, adding zero vectors to a Parseval frame gives another Parseval frame. Moreover, adding pairs of arbitrary vectors to a Parseval frame preserves the Parseval property. In both cases, we have artificially increased the redundancy by enlarging the frame. In our catalogue, we discard Parseval frames with such a trivial source of redundancy.

Definition 4.12. A Parseval frame $\mathcal{F} = \{f_1, f_2, \dots, f_k\}$ for $\mathbb{Z}_2^n$ is called *trivially redundant* if there is $j \in \{1, 2, \dots k\}$ with $f_j = 0$, or if there are two indices $i \neq j$ with $f_i = f_j$.

After repeated vectors are removed, Parseval frames can be interpreted as sets of vectors. We consider the set-theoretic complement of such a Parseval frame.

Theorem 4.13. *Let $n \geq 3$. Let $\mathcal{F} = \{f_i\}_{i=1}^k$ be a family without repeated vectors in $\mathbb{Z}_2^n$ and $\mathcal{G} = \mathbb{Z}_2^n \setminus \mathcal{F}$. If $\mathcal{F}$ is a Parseval frame for $\mathbb{Z}_2^n$, then $\mathcal{G}$ is also a Parseval frame.*

Proof. Let $\mathcal{X} = \left\{x \in \mathbb{Z}_2^n\right\}$, then we count $2^n - 1$ nonzero elements in $\mathcal{X}$. Thinking of $\mathcal{X}$ as a sequence $\{f_i\}_{i=0}^m$ where $m = 2^n - 1$ and the entries of f_i are given by the binary expansion of i, let $\Theta_{\mathcal{X}}^*$ be the matrix with f_i as the ith column.

By simple counting, there are 2^{n-1} elements with 1 in the ith position. This means, in each row of $\Theta_{\mathcal{X}}^*$ the number 1 appears exactly 2^{n-1} times. Furthermore there are 2^{n-2} elements with 1 in the ith and jth position, similarly for any fixed choice of 1 or 0 in the ith and jth position. If $n \geq 3$, then 2^{n-2} is even and consequently, the dot product of any row of $\Theta_{\mathcal{X}}^*$ with itself or any other row is equal to 0, i.e. $\Theta_{\mathcal{X}}^*\Theta_{\mathcal{X}} = 0$.

If $\mathcal{F}$ is a Parseval frame, then $\Theta_{\mathcal{F}}^*\Theta_{\mathcal{F}} = I$ which implies via the matrix product that there is an odd number of elements in $\mathcal{F}$ with 1 in the ith position, and that among the elements with a 1 in the ith position there is an even number of elements with a 1 in the jth position.

As remarked above, in the entire space there is an even number of elements with 1 in the ith position and an even number of elements with 1 in the ith and jth position. Thus there is an odd number of elements in the complement $\mathcal{G} = \mathcal{X} \setminus \mathcal{F}$ with 1 in the ith position and an even number of such elements with 1 in the jth position, that is $\Theta_{\mathcal{G}}^*\Theta_{\mathcal{G}} = I$. Hence $\mathcal{G}$ is a Parseval Frame. □

Corollary 4.14. *If $\mathcal{F}$ is a Parseval frame for $\mathbb{Z}_2^n$ which is not trivially redundant, and $\mathcal{G}$ is its set-theoretic complement, then both $\mathcal{F}$ and $\mathcal{G} \setminus \{0\}$ are Parseval frames and one of them contains at most $2^{n-1} - 1$ vectors.*

Proof. After removing the zero vector from $\mathcal{G}$, the union of both Parseval frames $\mathcal{F}$ and $\mathcal{G} \setminus \{0\}$ contains $2^n - 1$ vectors. This implies that one of the two frames contains at most half this number, meaning at most $2^{n-1} - 1$ vectors. □

To complete the catalogue of binary Parseval frames for $\mathbb{Z}_2^n$, it is only necessary to find one representative from each switching equivalence class of Parseval frames with at most $2^{n-1} - 1$ vectors. Once these Parseval frames have been found, their complements complete the list, because the switching equivalence of a pair of frames is equivalent to that of their complements.

Proposition 4.15. *Two frames that are not trivially redundant are switching equivalent if and only if their set-theoretic complements are.*

Proof. This is a consequence of the fact that unitaries are one-to-one maps on the set $\mathbb{Z}_2^n$. Thus, if a unitary U maps a frame $\mathcal{F}$ to a frame $\mathcal{G}$, then it also maps the complement of $\mathcal{F}$ to the complement of $\mathcal{G}$, and vice versa. □

n	k	vectors
3	3	1 2 4
	4	3 5 6 7
4	4	1 2 4 8
	5	1 6 10 12 14
	6	1 3 5 9 14 15
	7	1 2 3 7 11 12 15
	8	4 5 6 8 9 10 13 14
	9	2 4 6 7 8 10 11 12 13
	10	2 3 4 5 7 8 9 11 13 15
	11	3 5 6 7 9 10 11 12 13 14 15

Table 1. Representatives of all switching-equivalence classes, excluding trivially redundant Parseval frames, for $\mathbb{Z}_2^3$ and $\mathbb{Z}_2^4$.

We conclude with Table 1, a complete list of representatives of switching-equivalence classes of Parseval frames for $n = 3$ and $n = 4$, excluding ones that are trivially redundant. Each frame vector in our list is recorded by the integer obtained from the binary expansion with the entries of the vector. For example, if a frame vector in $\mathbb{Z}_2^4$ is $f_1 = (1, 0, 1, 1)$, then it is represented by the integer $2^0+2^2+2^3 = 13$. Accordingly, in $\mathbb{Z}_2^4$, the standard basis is recorded as the sequence of numbers $1, 2, 4, 8$.

As explained above, the part of the table with $k > 2^{n-1} - 1$ vectors has been obtained by taking complements of Parseval frames with $k \leq 2^{n-1} - 1$ vectors, according to Corollary 4.14 and Proposition 4.15. An exhaustive search shows that there is only one switching equivalence class for $n = 3$ and $k \in \{3, 4\}$ and for $n = 4$ and each $k \in \{4, 5, 6, 7\}$, consequently also for $k \in \{8, 9, 10, 11\}$.

References

[Bodmann and Paulsen 2005] B. G. Bodmann and V. I. Paulsen, "Frames, graphs and erasures", *Linear Algebra Appl.* **404** (2005), 118–146. MR 2006a:42047 Zbl 1088.46009

[Casazza and Kovačević 2003] P. G. Casazza and J. Kovačević, "Equal-norm tight frames with erasures", *Adv. Comput. Math.* **18**:2-4 (2003), 387–430. MR 2004e:42046 Zbl 1035.42029

[Christensen 2003] O. Christensen, *An introduction to frames and Riesz bases*, Applied and Numerical Harmonic Analysis, Birkhäuser, Boston, MA, 2003. MR 2003k:42001 Zbl 1017.42022

[Duffin and Schaeffer 1952] R. J. Duffin and A. C. Schaeffer, "A class of nonharmonic Fourier series", *Trans. Amer. Math. Soc.* **72** (1952), 341–366. MR 13,839a Zbl 0049.32401

[Goyal et al. 1998] V. K. Goyal, M. Vetterli, and N. T. Thao, "Quantized overcomplete expansions in $\mathbb{R}^N$: analysis, synthesis, and algorithms", *IEEE Trans. Inform. Theory* **44**:1 (1998), 16–31. MR 99a:94004 Zbl 0905.94007

[Goyal et al. 2001] V. K. Goyal, J. Kovačević, and J. A. Kelner, "Quantized frame expansions with erasures", *Appl. Comput. Harmon. Anal.* **10**:3 (2001), 203–233. MR 2002h:94012 Zbl 0992.94009

[Han et al. 2007] D. Han, K. Kornelson, D. Larson, and E. Weber, *Frames for undergraduates*, Student Math. Library **40**, American Mathematical Society, Providence, RI, 2007. MR 2367342 Zbl 1143.42001

[Holmes and Paulsen 2004] R. B. Holmes and V. I. Paulsen, "Optimal frames for erasures", *Linear Algebra Appl.* **377** (2004), 31–51. MR 2004j:42028 Zbl 1042.46009

[Kalra 2006] D. Kalra, "Complex equiangular cyclic frames and erasures", *Linear Algebra Appl.* **419**:2-3 (2006), 373–399. MR 2007j:42024 Zbl 1119.42013

[Kovačević and Chebira 2008] J. Kovačević and A. Chebira, "An introduction to frames", *Found. Trends Signal Processing* **2**:1 (2008), 1–94.

[MacWilliams and Sloane 1977] F. J. MacWilliams and N. J. Sloane, *The theory of error-correcting codes*, North Holland, Amsterdam, 1977.

[Püschel and Kovačević 2006] M. Püschel and J. Kovačević, "Real, tight frames with maximal robustness to erasures", pp. 63–72 in *Proc. Data Compr. Conf.* (Snowbird, UT, 2005), edited by J. A. Storer and M. Cohn, IEEE Press, 2006.

[Rath and Guillemot 2003] G. Rath and C. Guillemot, "Performance analysis and recursive syndrome decoding of DFT codes for Bursty erasure recovery", *IEEE Trans. Signal Process.* **51**:5 (2003), 1335–1350. MR 2049911

[Rath and Guillemot 2004] G. Rath and C. Guillemot, "Frame-theoretic analysis of DFT codes with erasures", *IEEE Trans. Signal Process.* **52**:2 (2004), 447–460. MR 2005d:94067

[Strohmer and Heath 2003] T. Strohmer and R. W. Heath, Jr., "Grassmannian frames with applications to coding and communication", *Appl. Comput. Harmon. Anal.* **14**:3 (2003), 257–275. MR 2004d:42053 Zbl 1028.42020

[Xia et al. 2005] P. Xia, S. Zhou, and G. B. Giannakis, "Achieving the Welch bound with difference sets", *IEEE Trans. Inform. Theory* **51**:5 (2005), 1900–1907. MR 2007b:94148a

Received: 2009-08-06 Accepted: 2009-08-12

bgb@math.uh.edu *Department of Mathematics, University of Houston, Houston, TX 77204-3008, United States*

myle_64@yahoo.com *Department of Mathematics, University of Houston, Houston, TX 77204-3008, United States*

letty@math.uh.edu *Department of Mathematics, University of Houston, Houston, TX 77204-3008, United States*

matt_tobin67@hotmail.com *Department of Mathematics, University of Houston, Houston, TX 77204-3008, United States*

tomforde@math.uh.edu *Department of Mathematics, University of Houston, Houston, TX 77204-3008, United States*

INVOLVE 2:5(2009)

Some results on the size of sum and product sets of finite sets of real numbers

Derrick Hart and Alexander Niziolek

(Communicated by Andrew Granville)

Let A and B be finite subsets of positive real numbers. Solymosi gave the sum-product estimate $\max(|A+A|, |A\cdot A|) \geq (4\lceil \log |A| \rceil)^{-1/3}|A|^{4/3}$, where $\lceil\ \rceil$ is the ceiling function. We use a variant of his argument to give the bound

$$\max(|A+B|, |A\cdot B|) \geq (4\lceil \log |A| \rceil \lceil \log |B| \rceil)^{-1/3}\, |A|^{2/3}\, |B|^{2/3}.$$

(This isn't quite a generalization since the logarithmic losses are worse here than in Solymosi's bound.)

Suppose that A is a finite subset of real numbers. We show that there exists an $a \in A$ such that $|aA + A| \geq c|A|^{4/3}$ for some absolute constant c.

1. Introduction

Given finite subsets A and B of an additive group, the *sum set* of A and B is

$$A + B = \{a + b : a \in A,\ b \in B\}.$$

Similarly, define the *product set* by

$$A \cdot B = \{ab : a \in A,\ b \in B\}.$$

If M and N are numbers (depending on A and B) we write $M \gtrsim N$ to mean that $M \geq cN$ for some constant $c > 0$ (independent of A and B). We write $M \approx N$ to mean that $cN \leq M \leq c'N$ for $c, c' > 0$.

Suppose that $A = B$ is an arithmetic progression. Then

$$|A + A| \lesssim |A| \quad \text{and} \quad |A \cdot A| \gtrsim |A|^{2-\delta},$$

where here and throughout $\delta \to 0$ as $|A| \to \infty$ and $|\cdot|$ denotes the size of the set. In contrast, if $A = B$ is a geometric progression then

$$|A + A| \gtrsim |A|^2 \quad \text{and} \quad |A \cdot A| \lesssim |A|.$$

MSC2000: 11B13, 11B75.
Keywords: sum-product estimate, multiplicative energy, Solymosi bound.

These examples led Erdős and Szemerédi [1983] to ask whether both the product set and sum set can be small at the same time. They conjectured that it is not possible in the following sense.

Conjecture 1. *Let A be a finite subset of $\mathbb{Z}$. Then*

$$\max(|A+A|, |A\cdot A|) \geq |A|^{2-\delta}.$$

They showed that

$$\max(|A+A|, |A\cdot A|) \geq |A|^{1+\varepsilon},$$

for a positive ε.

The explicit bound $\varepsilon \geq \frac{1}{31}$ was given by Nathanson [1997], and $\varepsilon \geq \frac{1}{15}$ by Ford [1998]. A breakthrough was achieved by Elekes [1997], who connected the problem to incidence geometry and applied the Szemerédi–Trotter incidence theorem to obtain $\varepsilon \geq \frac{1}{4}$. This was improved by Solymosi [2005] to $\varepsilon \geq \frac{3}{14} - \delta$. These bounds hold in the more general context of finite subsets of $\mathbb{R}$.

Recently, by a short and ingenious argument it was shown by Solymosi [2009] that $\varepsilon \geq \frac{1}{3} - \delta$. In Section 3 we mimic Solymosi's argument with a few changes to give an analogous estimate for sums and products of different sets.

Given the strong relationship between sums and products one may ask a related question: how large is the set $A\cdot B + C$ guaranteed to be? Elekes (see [Alon and Spencer 2000]) showed that $|A\cdot B+C| \gtrsim \sqrt{|A|\,|B|\,|C|}$ with certain size restrictions on the three sets. His argument relied on the aforementioned Szeremedi–Trotter incidence theorem and is short enough to present in the next few lines.

Let P be a set of points in $\mathbb{R}^2$ and L a set of lines. We say a point $p \in P$ is incident to a line $l \in L$ if p lies on l. In this case, we denote this incidence by $(p, l) \in P \times L$.

Theorem 2 [Szemerédi and Trotter 1983]. *Let $I_{P,L}$ denote the number of incidences between P and L. Then bound*

$$I_{P,L} \lesssim |P|^{2/3}\,|L|^{2/3} + |P| + |L|.$$

Let $L = \{y = ax + c : a \in A,\ c \in C\}$ and $P = B \times A\cdot B + C$. Clearly, given any $a \in A$, $b \in B$, $c \in C$, the point $(b, ab + c)$ is incident to the line $y = ax + c$. Therefore, by Szemeredi–Trotter, $|A|\,|B|\,|C| \lesssim |A|^{2/3}\,|B|^{2/3}\,|C|^{2/3}\,|A\cdot B + C|^{2/3}$.

In the context of $\mathbb{F}_q$, the finite field containing q elements, similar questions have been explored as well. Bourgain [2005] showed that for $A \subseteq \mathbb{F}_q$ such that $|A| \gtrsim q^{3/4}$, one has $A\cdot A + A\cdot A + A\cdot A = \mathbb{F}_q$; in particular, if $|A| \approx q^{3/4}$, then $|A\cdot A + A\cdot A + A\cdot A| \gtrsim |A|^{4/3}$. In [Hart and Iosevich 2008] it was shown that if $|A| \gtrsim q^{3/4}$, then $A\cdot A + A\cdot A = \mathbb{F}_q^*$; in particular, if $|A| \approx q^{3/4}$, then $|A\cdot A + A\cdot A| \gtrsim |A|^{4/3}$. Due to the misbehavior of the zero element, it is not possible to guarantee that $A\cdot A + A\cdot A = \mathbb{F}_q$ unless A is a positive proportion of

the elements of $\mathbb{F}_q$. Under the weaker conclusion that $|A \cdot A + A \cdot A| \gtrsim q$ it is shown in the same paper that one may take $|A| \gtrsim q^{2/3}$. Shparlinski [2008] applied multiplicative character sums to show that if $|A| \gtrsim q^{2/3}$, then $|A \cdot A + A| \gtrsim q$, implying that if $|A| \approx q^{2/3}$, then $|A \cdot A + A| \gtrsim |A|^{3/2}$.

Theorem 3 [Chapman et al. 2009, Theorem 2.10]. *Let A be a subset of $\mathbb{F}_q^*$. Then*

$$|A|^{-1} \sum_{a \in A} |aA + A| \gtrsim \min(q, |A|^3 q^{-1}).$$

In particular, if $|A| \approx q^{2/3}$, there exists a subset A' of A with $|A'| \gtrsim |A|$ such that

$$|aA + A| \gtrsim |A|^{3/2} \approx q,$$

for all $a \in A'$.

It is natural to ask whether a similar statement holds in the case that A is a finite subset of the real numbers. We show that this is in fact the case in Section 4.

2. Statement of results

Define the multiplicative energy of the finite subsets A, B, C, D of real numbers by

$$\mathrm{E}(A, B, C, D) = \big|\big\{(x_1, x_2, y_1, y_2) \in A \times B \times C \times D : x_1 y_2 = x_2 y_1\big\}\big|.$$

For A, B finite subsets of positive real numbers with $|A| \le |B|$, the argument of [Solymosi 2009] gives the bound

$$\mathrm{E}(A, B, A, B) \le 4\lceil \log |A| \rceil \, |A + A| \, |B + B|. \tag{2-1}$$

A short Cauchy–Schwarz argument gives that $\mathrm{E}(A, B, A, B) \ge |A|^2 |B|^2 / |A \cdot B|$, which in turn gives the sum-product inequality

$$|A|^2 |B|^2 \le 4\lceil \log |A| \rceil |A + A| \, |B + B| \, |A \cdot B|. \tag{2-2}$$

In the case that $A = B$, this immediately implies the Solymosi sum-product bound discussed in the introduction:

$$\max(|A + A|, |A \cdot A|) \ge (4\lceil \log |A| \rceil)^{-1/3} |A|^{4/3}. \tag{2-3}$$

We will use a slight variant of the argument of Solymosi to give a different bound on the multiplicative energy:

Theorem 4. *Let A, B, C, D be finite subsets of positive real numbers. Then*

$$\mathrm{E}(A, B, C, D) \le 4\lceil \log(\min(|A|, |C|)) \rceil \lceil \log(\min(|B|, |D|)) \rceil |A + B| \, |C + D|.$$

(Notice that the logarithmic loss is worse than what was obtained by Solymosi.)

Using the fact that $\mathrm{E}(A, B, A, B) \geq |A|^2|B|^2/|A \cdot B|$, we obtain the following sum-product estimate.

Corollary 5. *Let A, B be finite subsets of positive real numbers. Then*

$$\max(|A+B|, |A \cdot B|) \geq (4\lceil \log |A| \rceil \lceil \log |B| \rceil)^{-1/3} |A|^{2/3} |B|^{2/3}. \tag{2-4}$$

One may compare this to the result of applying Plünnecke's inequality to (2-2):

$$\max(|A+B|, |A \cdot B|) \geq (4\lceil \log |A| \rceil)^{-1/5} |A|^{3/5} |B|^{3/5}. \tag{2-5}$$

We will also show this:

Theorem 6. *Let A, B, C be finite subsets of $\mathbb{R}$ such that $|B|^{1/2}|C|^{-1/2} \lesssim |A| \lesssim |B|^2|C|$. Then*

$$|A|^{-1} \sum_{a \in A} |aB + C| \gtrsim |A|^{1/3} |B|^{1/3} |C|^{2/3}. \tag{2-6}$$

In particular, there exists an $a \in A$ such that

$$|aB + C| \gtrsim |A|^{1/3} |B|^{1/3} |C|^{2/3}. \tag{2-7}$$

3. Proof of Theorem 4

We begin by writing

$$\begin{aligned} \mathrm{E}(A, B, C, D) &= \sum_{x_1 y_2 = x_2 y_1} A(x_1)B(x_2)C(y_1)D(y_2) \\ &= \sum_{t \neq 0} \sum_{\substack{x_1 = t x_2 \\ y_1 = t y_2}} (A \times C)(x_1, y_1)(B \times D)(x_2, y_2), \end{aligned}$$

where $A(\cdot)$ denotes the characteristic function of the set A and $\times$ denotes the Cartesian product. Summing in t we have

$$\mathrm{E}(A, B, C, D) = \sum_{y \in (B \times D)} |(A \times C) \cap l_{m_y}|,$$

where l_{m_y} is the line through the origin and the point y with slope m_y. Each $y \in (B \times D)$ lies on some line l_{m_y} with $m_y \in D \cdot B^{-1} = \{db^{-1} : d \in D, b \in B\}$. Since the quantity $|(A \times C) \cap l_{m_y}|$ is constant and nonzero for each y on l_{m_y} with slope m_y in $C \cdot A^{-1}$, we have

$$\mathrm{E}(A, B, C, D) = \sum_{m \in M} |(A \times C) \cap l_m| \, |(B \times D) \cap l_m|,$$

where $M = C \cdot A^{-1} \cap D \cdot B^{-1}$. We then take a dyadic decomposition

$$\mathrm{E}(A, B, C, D) = \sum_{\substack{0 \le i \le \lceil \log(\min(|A|,|C|)) \rceil \\ 0 \le j \le \lceil \log(\min(|B|,|D|)) \rceil}} \sum \sum_{m \in M_{i,j}} |(A \times C) \cap l_m| \, |(B \times D) \cap l_m|,$$

where $M_{i,j} = \{m \in M : 2^i \le |(A \times C) \cap l_m| < 2^{i+1}, 2^j \le |(B \times D) \cap l_m| < 2^{j+1}\}$. Therefore, for some i' and j',

$$\frac{\mathrm{E}(A, B, C, D)}{\lceil \log(\min(|A|, |C|)) \rceil \lceil \log(\min(|B|, |D|)) \rceil} \le \sum_{m \in M_{i',j'}} |(A \times C) \cap l_m| \, |(B \times D) \cap l_m|.$$

Set $n = |M_{i',j'}|$ and order the elements of $M_{i',j'}$: $m_1 < m_2 < \ldots < m_n$. This gives

$$\frac{\mathrm{E}(A, B, C, D)}{\lceil \log(\min(|A|, |C|)) \rceil \lceil \log(\min(|B|, |D|)) \rceil} \le 4n2^{i'+j'}.$$

Given that $|(A \times C) \cap l_{m_l} + (B \times D) \cap l_{m_{l+1}}| = |(A \times C) \cap l_{m_l}| \, |(B \times D) \cap l_{m_{l+1}}|$, noting that any two sum sets $(A \times C) \cap l_{m_l} + (B \times D) \cap l_{m_{l+1}}$ and $(A \times C) \cap l_{m_k} + (B \times D) \cap l_{m_{k+1}}$ are disjoint for any $l \ne k$ gives

$$n2^{i'+j'} \le \left| \bigcup_{l=1}^{n} ((A \times C) \cap l_{m_l} + (B \times D) \cap l_{m_{l+1}}) \right| \le |A + B| \, |C + D|.$$

Here, in an abuse of notation, $(B \times D) \cap l_{m_{n+1}}$ is the orthogonal projection of $(B \times D) \cap l_{m_n}$ onto the vertical line running through the minimal element of B. We may without loss of generality assume that the minimal element of B is also the minimal element of $A \cup B$.

4. Proof of Theorem 6

We will need a lemma, whose proof we will delay until the end of the section.

Lemma 7. *Let A, B, C be finite subsets of $\mathbb{R}$ such that $|B|^{1/2} |C|^{-1/2} \lesssim |A| \lesssim |B|^2 |C|$. Then*

$$\left| \left\{ (a, b, c, d, e) \in A \times B \times C \times B \times C : ab + c = ad + e \right\} \right| \lesssim |A|^{2/3} |B|^{5/3} |C|^{4/3}.$$

With this lemma in hand one may then apply the Cauchy–Schwarz inequality:

$$|A| \, |B|^2 \, |C|^2 = |A|^{-1} \Bigg(\sum_{\substack{t \in aB + C \\ a \in A}} \sum_{ab+c=t} B(b) C(c) \Bigg)^2$$
$$\le \Bigg(|A|^{-1} \sum_{a \in A} |aB + C| \Bigg) \sum_{\substack{t \in aB + C \\ a \in A}} \Bigg(\sum_{ab+c=t} B(b) C(c) \Bigg)^2.$$

Noting that

$$\sum_{\substack{t\in aB+C\\ a\in A}}\Big(\sum_{ab+c=t} B(b)C(c)\Big)^2 = \big|\big\{(a,b,c,d,e)\in A\times B\times C\times B\times C : ab+c=ad+e\big\}\big|$$

completes the proof of Theorem 6.

Proof of Lemma 7. We will apply the Szemerédi–Trotter incidence theorem. For a fixed $b\in B$, consider the set of lines $L_b=\{y=(b-d)x+c : c\in C, d\in B\}$. Also consider the set of points $P=\{(a,e)\in(A\times C)\}$. Then $|\{(a,b,c,d,e)\in A\times B\times C\times B\times C : ab+c=ad+e\}|\le |B|\max_{b\in B} I_{P,L_b}$. Noting that $|L_b|=|B|\,|C|$ and $|P|=|A|\,|C|$ and applying the Szemerédi–Trotter theorem gives

$$\big|\big\{(a,b,c,d,e)\in A\times B\times C\times B\times C : ab+c=ad+e\big\}\big| \lesssim |A|^{2/3}\,|B|^{5/3}\,|C|^{4/3},$$

as long as $|B|^{1/2}\,|C|^{-1/2}\lesssim |A|\lesssim |B|^2|C|$. □

5. Remarks

The argument of Elekes [1997] actually gives a more general bound for finite subsets A, B, C of positive real numbers:

$$\max(|A+B|,|A\cdot C|)\gtrsim |A|^{3/4}\,|B|^{1/4}\,|C|^{1/4}.$$

A direct application of Plünnecke's inequality [Tao and Vu 2006, Corollary 6.26] to (2-3) gives

$$\max(|A+B|,|A\cdot C|)\ge (4\lceil\log|A|\rceil)^{-1/6}|A|^{2/3}\,|B|^{1/3}\,|C|^{1/6}.$$

This bound is preferable if $|B|$ is much larger than $|A|\,|C|$. We do not currently know of a way to use Solymosi's argument to obtain an improved bound for the case that the three sets are close together in size.

References

[Alon and Spencer 2000] N. Alon and J. H. Spencer, *The probabilistic method*, 2nd ed., Wiley-Interscience, New York, 2000. MR 2003f:60003 Zbl 0996.05001

[Bourgain 2005] J. Bourgain, "Mordell's exponential sum estimate revisited", *J. Amer. Math. Soc.* **18**:2 (2005), 477–499. MR 2006b:11099 Zbl 1072.11063

[Chapman et al. 2009] J. Chapman, M. B. Erdoğan, D. Hart, A. Iosevich, and D. Koh, "Pinned distance sets, k-simplices, Wolff's exponent in finite fields and sum-product estimates", preprint, 2009. arXiv 0903.4218

[Elekes 1997] G. Elekes, "On the number of sums and products", *Acta Arith.* **81**:4 (1997), 365–367. MR 98h:11026 Zbl 0887.11012

[Erdős and Szemerédi 1983] P. Erdős and E. Szemerédi, "On sums and products of integers", pp. 213–218 in *Studies in pure mathematics*, Birkhäuser, Basel, 1983. MR 86m:11011 Zbl 0526.10011

[Ford 1998] K. Ford, "Sums and products from a finite set of real numbers", *Ramanujan J.* **2**:1-2 (1998), 59–66. MR 99i:11014 Zbl 0908.11008

[Hart and Iosevich 2008] D. Hart and A. Iosevich, "Sums and products in finite fields: an integral geometric viewpoint", pp. 129–135 in *Radon transforms, geometry, and wavelets*, Contemp. Math. **464**, Amer. Math. Soc., Providence, RI, 2008. MR 2009m:11032

[Nathanson 1997] M. B. Nathanson, "On sums and products of integers", *Proc. Amer. Math. Soc.* **125**:1 (1997), 9–16. MR 97c:11010 Zbl 0869.11010

[Shparlinski 2008] I. E. Shparlinski, "On the solvability of bilinear equations in finite fields", *Glasg. Math. J.* **50**:3 (2008), 523–529. MR 2009j:11189 Zbl 05350503

[Solymosi 2005] J. Solymosi, "On the number of sums and products", *Bull. London Math. Soc.* **37**:4 (2005), 491–494. MR 2006c:11021 Zbl 1092.11018

[Solymosi 2009] J. Solymosi, "Bounding multiplicative energy by the sumset", *Adv. Math.* **222**:2 (2009), 402–408. MR 2538014 Zbl 05597906

[Szemerédi and Trotter 1983] E. Szemerédi and W. T. Trotter, Jr., "Extremal problems in discrete geometry", *Combinatorica* **3**:3-4 (1983), 381–392. MR 85j:52014 Zbl 0541.05012

[Tao and Vu 2006] T. Tao and V. Vu, *Additive combinatorics*, vol. 105, Cambridge University Press, Cambridge, 2006. MR 2008a:11002 Zbl 1127.11002

Received: 2009-09-07 Accepted: 2009-11-12

dnhart@math.rutgers.edu *Department of Mathematics, Rutgers University, Piscataway, NJ 08854-8019, United States*
http://www.math.rutgers.edu/~dnhart/

alexnizi@eden.rutgers.edu *School of Engineering, Rutgers University, Piscataway, NJ 08854-8019, United States*

INVOLVE 2:5(2009)

Proof of the planar double bubble conjecture using metacalibration methods

Rebecca Dorff, Gary Lawlor, Donald Sampson and Brandon Wilson

(Communicated by Frank Morgan)

We prove the double bubble conjecture in $\mathbb{R}^2$: that the standard double bubble in $\mathbb{R}^2$ is boundary length-minimizing among all figures that separately enclose the same areas. Our independent proof is given using the new method of *metacalibration*, a generalization of traditional calibration methods useful in minimization problems with fixed volume constraints.

1. Introduction

Isoperimetric problems have had a long history. The earliest proofs that the circle maximizes area for a figure of given perimeter can be traced to the ancient Greeks. The first results are attributed to the second century mathematician Zenodorus. After more than a millennium, Steiner was the first to realize that the ancient Greek proofs were insufficient by modern standards and worked to complete them. Weierstrass, however, was the first to give a rigorous proof of the isoperimetric inequality, and furthered the development of analysis and calculus in order to do so [Siegel 2003].

While many different proofs exist for the isoperimetric inequality in two dimensions, few of these methods can be applied to generalizations of the problem, including having multiple enclosed areas or higher dimensional analogs. The traditional approach, and so far most successful, has been to use the calculus of variations to isolate properties of the boundary-minimizing figure and compare all possible figures of this type. Some advancements in "multiple bubble" problems were made this way by Frederick Almgren [1976], Jean Taylor [1976], and Frank Morgan [1994], who proved regularity results in $\mathbb{R}^n$ for $n \geq 4$, $n = 3$, and $n = 2$, respectively. Morgan realized that a careful analysis of minimizers in the plane was absent from the literature, and providing this, showed that perimeter-minimizing planar figures must consist of circular arcs meeting at vertices of degree three,

MSC2000: 49Q05, 49Q10, 53A10.
Keywords: calibration, metacalibration, double bubble, isoperimetric, optimization.

forming 120-degree angles. This reduced the argument to listing all the combinatorial types meeting these requirements, possibly subject to some bounds on the number of components that each area may be broken up into. Using this result, students of the 1990 SMALL group under Frank Morgan proved that the standard double bubble was perimeter-minimizing among all figures separately enclosing two fixed areas [Foisy et al. 1993]. This method was also employed by Wacharin Wichiramala, whose doctoral dissertation proves the corresponding result for three separated areas. Unfortunately, this approach is marked by an ever-increasing combinatorial complexity. For example, Wichiramala's dissertation [2004] considered some fifty-four possible configurations in order to prove minimization of the standard triple bubble. This complexity proves to be a significant barrier to further results.

We present new proofs of the isoperimetric inequality in the plane, for one and two areas, using *metacalibration*, a new method of proof developed by Gary Lawlor. Metacalibration is an extension of previous work in the field of calibration that has been modified to handle a new class of minimization problems. Section 2 discusses metacalibration in further detail. A reformulation of ideas by Gromov [Milman and Schechtman 1986], Lawlor's proof of the two-dimensional isoperimetric inequality is given in Section 3. Lawlor also showed using metacalibration that the standard double bubble is perimeter-minimizing among all figures enclosing two equal areas. Section 4 contains the authors' generalization of this proof to any two fixed areas. Metacalibration has a strong potential for other applications in which the standard variational approach fails. Further work is currently being made in extending our results to other multiple bubble problems, including the as-yet-unproven triple bubble conjecture in $\mathbb{R}^3$.

2. Metacalibration

Each minimization problem[1] can be expressed in the following terms. Let a set of constraints C be given. Let S be the set of all competitors σ that satisfy C. For each competitor $\sigma \in S$, we define the function $P(\sigma)$ which expresses the quantity to be minimized (perimeter, energy, etc.). The minimizing property of a conjectured minimizer $\mu \in S$ is shown by proving $P(\mu) \leq P(\sigma)$ for all $\sigma \in S$.

Metacalibration solves minimization problems by comparing $P(\sigma)$ to an intermediary function $G(\sigma)$ which simplifies the conditions for comparison. This function is called a *calibration function* and is defined as follows.

[1] As the methods of metacalibration work identically for minimization as well as maximization problems, we will present these methods in the context of a minimization problem. Maximization problems are solved identically with the obvious inequalities reversed.

Definition 2.1. A function G *calibrates* a conjectured minimizer, $\mu \in S$, if for every competitor $\sigma \in S$,

(1) $P(\mu) = G(\mu)$,

(2) $G(\mu) \leq G(\sigma)$, and

(3) $G(\sigma) \leq P(\sigma)$.

The following theorem encapsulates this idea.

Theorem 2.2. Calibration Theorem. *If a function G calibrates a conjectured minimizer μ with respect to P, then μ minimizes the function P.*

Proof. Take any competitor $\sigma \in S$. By the definition of a calibration, $P(\mu) = G(\mu) \leq G(\sigma) \leq P(\sigma)$. Thus for any $\sigma \in S$, $P(\mu) \leq P(\sigma)$ and μ minimizes P. □

One can see from the above arguments that this method yields simple and elegant results if we are able to identify a suitable calibration function $G(\sigma)$. Finding such a function is one of the difficulties of this method. Following is a short description of some characteristics of calibration functions.

In order to establish the necessary identities, calibration functions typically have the form of an integral. For example, if we parametrize σ by some variable t, the necessary relations may be established by comparing the rate of change of $G(\sigma)$ and $P(\sigma)$ with respect to t. Suppose we find a function $df_\sigma/dt = f'_\sigma$ such that $f'_\sigma(t) \leq P'_\sigma(t)$, with equality on the minimizer. Letting σ be parametrized by $t \in [t_0, t_1]$, we define the function $G(\sigma) = \int_{t_o}^{t_1} f'_\sigma dt$. We find that $G(\sigma)$ will be a calibration function if $f_\sigma(t_1) - f_\sigma(t_0)$ is constant for all competitors, or at least minimal on μ. If this is the case, f'_σ is called a *calibration*. In many cases these conditions allow us to explicitly determine the function f'_σ. In any case however, solving these conditions will often give insight into the form or character of the necessary function f'_σ.

In the past, calibrations have been functions of spacial variables, such as position or tangent vectors. In metacalibration, we also allow the calibration function $G(\sigma)$ to depend on characteristics of σ itself. These may include variables such as arc length or enclosed area. Another useful tool of metacalibration is that it allows other variables to be defined abstractly, such as characteristics of the competitor under mapping or projection. These additional allowances of metacalibration allow calibration functions to consider a wider range of competitors, enabling metacalibration to take on various problems beyond the reach of standard methods.

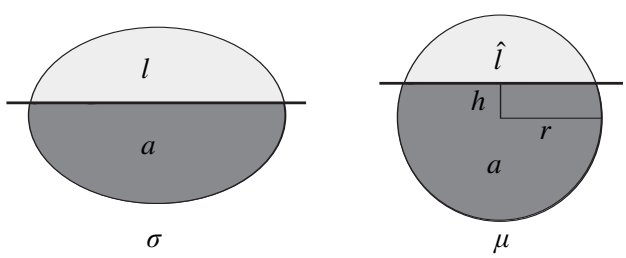
In the following sections we show how two classical geometric optimization problems may be solved using the methods of metacalibration. While each is a previously solved problem, they demonstrate the usefulness of this method and its future extension into yet unsolved problems.

3. The circle is perimeter-minimizing in $\mathbb{R}^2$

We begin our application of metacalibration with the classic example of the isoperimetric inequality in $\mathbb{R}^2$. Of all figures that enclose the same area in the plane, we wish to show that the circle of that area minimizes total perimeter. The minimizing property of the circle is shown using the calibration theorem of the previous section. To do this, we first define the function $G(\sigma)$, and subsequently prove that it satisfies each of the conditions for calibration.

Take an arbitrary competing figure σ in S, the set of competitor figures (C^1 manifolds) that enclose a fixed area A_0. We assume that the boundary of the figure is equal to the boundary of its interior. This ensures that σ will have no obviously unnecessary perimeter elements that do not enclose area. Let an arbitrary axis for the parameter t be given. We will parametrize σ by a set of slicing lines perpendicular to the t-axis, whose position is determined by the variable t. In our depictions, this is an upward sweeping line with a vertical t-axis. We let t_0 be the bottom of the figure and t_1 the top, so that σ is fully contained in the sweeping interval $[t_0, t_1]$. Let $\mathcal{R}_\sigma$ be the projection of σ in the t-axis. Note that $\mathcal{R}_\sigma \subseteq [t_0, t_1]$.

For any slicing line, given by the parameter t, we define the following functions for use in the calibration function. Let $a(t)$ be the area enclosed by the figure and contained in the sweeping interval $[t_0, t]$ (that is, below the slicing line t). Thus $a(t_0) = 0$ and $a(t_1) = A_0$ for any competitor. Let $l(t)$ be the total length of the intersection of the slicing line t with the interior of σ:



To complete our definition of the calibration function we define a map F from a slicing line on the competitor to a slicing line on the conjectured minimizer, the circle enclosing area A_0. Let r denote the radius of this circle. We also parametrize the circle with a set of parallel slicing lines. The position of these lines will be defined by the variable h, the y-coordinate of the lines, where $h = 0$ passes through the center of the circle. Let $F(t) = h$ match the slicing line of the circle such that the area enclosed by the circle under the line h is equal to $a(t)$. Let $\hat{l}$ be the length of the intersection of the slicing line h with the interior of the circle. These functions are taken as C^1, since any competitor for which they are not may be shown to be nonminimal by a standard variational argument. While the functions defined above are all functions of the parameter t, we typically suppress the notation.

Using the above functions we are now able to define the calibration function $G(\sigma)$.

Theorem 3.1. *The function* $G(\sigma) = \int_{\mathcal{R}_\sigma} f'_\sigma dt$ *calibrates the circle, where* $f_\sigma(t) = \big(2a(t) - h(t)l(t)\big)/r$.

Proof of each condition for calibration will be given separately in Lemmas 3.2, 3.3, 3.4.

Lemma 3.2. *For* μ, *the circle enclosing area* A_0, $P(\mu) = G(\mu)$.

Proof. Note that $G(\mu) = \int_{\mathcal{R}_\mu} f' dt = g(t_1) - g(t_0) = 2A_0/r + 0$. Noting that in a circle $A_0 = \pi r^2$, we find $2A_0/r = 2\pi r^2/r = 2\pi r$, which is the perimeter of the circle. Thus $P(\mu) = G(\mu)$. □

Lemma 3.3. *For any competitor* $\sigma \in S$ *and the circle* μ, $G(\mu) = G(\sigma)$.

Proof. Note that since σ may be disconnected with several components, $\mathcal{R}_\sigma = \bigcup_1^m [x_i, y_i]$ where $x_1 = t_0$, $y_m = t_1$, and $y_i < x_{i+1}$. Thus

$$G(\sigma) = \int_{\mathcal{R}_\sigma} f' dt = \sum_1^m \int_{x_i}^{y_i} f' dt = \sum_1^m f(y_i) - f(x_i).$$

Note that $l(x_i) = l(y_i) = 0$ and $a(y_i) = a(x_{i+1})$ for all i. Thus

$$G(\sigma) = \sum_1^m f(y_i) - f(x_i) = \sum_1^m \frac{2a(y_i)}{r} - \frac{2a(x_i)}{r} = \frac{2a(y_m)}{r} - \frac{2a(x_1)}{r}.$$

But $a(y_m) = a(t_1) = A_0$ and $a(x_1) = a(t_0) = 0$. This implies that $G(\sigma) = 2A_0/r$, which is constant among all competitors in S, including μ. Thus $G(\mu) = G(\sigma)$ for all $\sigma \in S$. □

Lemma 3.4. *For any competitor* σ, $G(\sigma) \le P(\sigma)$.

Proof. Differentiating f by t we find $f' = (2A' - h'l - hl')/r$. Now from the definition of A and of h we can see that $A' = l$ in the competitor σ and $A' = h'\hat{l}$ in the conjectured minimizer μ. Substitution reveals

$$f' = \frac{(2-h')l - hl'}{r} = \frac{(2-h')h'\hat{l} - hl'}{r}.$$

Noting that $(2-h')h'$ has a maximum at 1 for $h' = 1$, we find

$$f' \le \frac{\hat{l} - hl'}{r} = \frac{1}{r}(\frac{\hat{l}}{2}, -h) \cdot (2, l').$$

Now $\frac{1}{r}(\frac{\hat{l}}{2}, -h)$ is a unit vector in the circle μ, so by the Cauchy–Schwartz inequality,

$$f' \le \|(2, l')\| = 2\sqrt{1 + (l'/2)^2}.$$

Since the slicing line at height t will intersect the figure at least twice if $l \neq 0$, $2\sqrt{1+(l'/2)^2}$ realizes the minimum of P' via a symmetrization argument. Thus $f' \leq P'(\sigma)$, and since $f_\sigma(0) = P_\sigma(0) = 0$, this implies $G(\sigma) = \int_{\mathcal{R}_\sigma} g' dt \leq \int_{\mathcal{R}_\sigma} P'(\sigma)\, dt = P(\sigma)$. □

Lemmas 3.2, 3.3, and 3.4 show that the function $G(\sigma)$ defined in Theorem 3.1 calibrates the circle. Thus by the calibration theorem, the circle minimizes perimeter of all C^1 manifolds enclosing a fixed area. Similar proofs that the sphere and n-sphere are boundary-minimizing in their respective dimensions have been found by the authors. In this paper, however, we investigate generalizations to multiple bubbles in two dimensions. The next section uses the methods of metacalibration to prove perimeter minimization of the standard double bubble in $\mathbb{R}^2$.

4. The standard double bubble is perimeter-minimizing in $\mathbb{R}^2$

The double bubble conjecture in $\mathbb{R}^2$ was first proved by students of Frank Morgan in an NSF funded REU in 1990 [Foisy et al. 1993]. They showed that the way to separately enclose two given areas with the least perimeter is the "standard double bubble," a figure with three circular arcs meeting two vertices at 120 angles. Here we present a new independent proof of the double bubble conjecture using the method of metacalibration.

Theorem 4.1. *The standard double bubble in $\mathbb{R}^2$ minimizes total perimeter of figures (unions of C^1 manifolds) enclosing two separate fixed areas.*

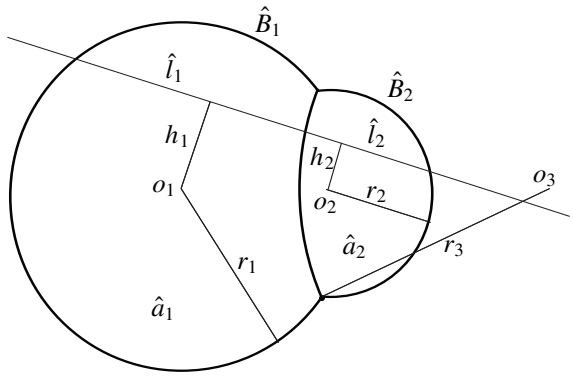
This will result from the calibration theorem using the calibration function defined below.

As with the calibration for the circle, we use a map from slices of the competitor to slices of our conjectured minimizer, the standard double bubble.

Each competitor double bubble $\sigma \in S$ will consist of two enclosed regions B_1 and B_2, with fixed areas A_1 and A_2. For a given competitor σ, parametrize parallel lines traversing the figure with the variable t. We let $a_i(t)$ be the area of B_i below the line t, and let $l_i(t)$ be the length of the intersection of the line t with B_i.

There is a unique standard double bubble that separately encloses areas A_1 and A_2 [Foisy et al. 1993]. Let $\hat{B}_1$ and $\hat{B}_2$ be the regions of this standard double bubble (of areas A_1 and A_2, respectively), and let r_1 and r_2 be the radii of the outer arcs bordering $\hat{B}_1$ and $\hat{B}_2$. Without loss of generality we assume that $r_1 \geq r_2$, or equivalently $A_1 \geq A_2$. We denote the radius of the arc separating $\hat{B}_1$ and $\hat{B}_2$ by r_3. We also denote the centers of each of these three arcs as o_1, o_2, and o_3 respectively. It is known that $1/r_3 = 1/r_2 - 1/r_1$ for all standard double bubbles [Isenberg 1978, pp. 88–95]. We also use the parameter h to define slices of the standard double bubble. However, in order to map slicing lines t in the competitor to slicing lines

in the standard bubble, matching both areas as before, slicing lines in the standard bubble must be allowed to tilt. Thus we parametrize slicing lines in the standard double bubble with two variables: h_1 and h_2, the signed distance from the slicing line to o_1 and o_2, respectively ($h_i < 0$ if the line passes below o_i). As with the competitor, we let $\hat{a}_i(t)$ be the area of $\hat{B}_i$ below the line (h_1, h_2), and let $\hat{l}_i(t)$ be the length of the intersection of the line (h_1, h_2) with $\hat{B}_i$:



Each slice of a competitor defines a point $(a_1, a_2) \in [0, A_1] \times [0, A_2]$. In this sense, each competitor σ, when parametrized by t, describes a path $\sigma : [0, 1] \to [0, A_1] \times [0, A_2]$ such that $\sigma(0) = (0, 0)$, $\sigma(1) = (A_1, A_2)$ and $\sigma(t) = (a_1(t), a_2(t))$.

We will define a map $F : [0, A_1] \times [0, A_2] \to [-r_1, r_1] \times [-r_2, r_2]$ between slices of the competitor and slices of the standard double bubble where $F(a_1, a_2) = (h_1, h_2)$. We will also define F such that for all $(a_1, a_2) \in [0, A_1] \times [0, A_2]$,

$$a_i = \hat{a}_i(F(a_1, a_2)).$$

There are, however, some points in $[0, A_1] \times [0, A_2]$ that do not have such a map into the double bubble. For example, if $A_1 \neq A_2$, no slice (h_1, h_2) of the standard double bubble gives rise to $(\hat{a}_1, \hat{a}_2) = (A_1, 0)$. Thus we will need to restrict F to some $K \subseteq [0, A_1] \times [0, A_2]$ such that $F|_K$ exists and is one-to-one.

To define K, we first define $\mathbb{H} \subset [-r_1, r_1] \times [-r_2, r_2]$ which will be the image of K under F. Let

$$\mathbb{H} = \{(h_1, h_2) \in [-r_1, r_1] \times [-r_2, r_2] \text{ such that } |h_1 - h_2| \leq |o_1 - o_2|\},$$

where $|o_1 - o_2|$ is the distance between the centers o_1 and o_2. This limits, by standard geometric properties, the parametrization (h_1, h_2) to slices that are realizable on the standard double bubble. Let $E : [-r_1, r_1] \times [-r_2, r_2] \to [0, A_1] \times [0, A_2]$ be the function that takes a slicing line in the standard double bubble and returns the area in each bubble under the slice. Thus we let $K = E(\mathbb{H})$.

Lemma 4.2. *The map $F = E^{-1} : K \to \mathbb{H}$ exists, is one-to-one, and is continuously differentiable.*

Proof. Proof follows by application of the inverse function theorem on F^{-1}.

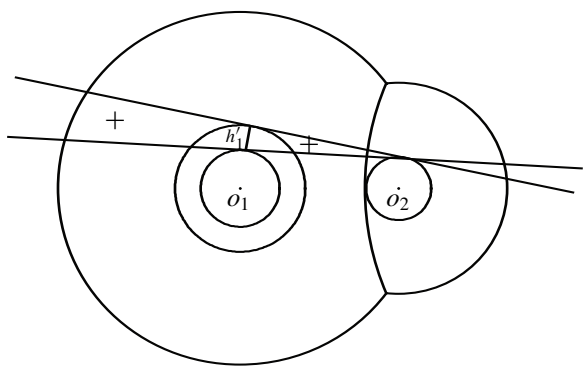
Consider $E : \mathrm{int}(\mathbb{H}) \to \mathrm{int}(K)$ where $E(h_1, h_2) = (\hat{a}_1, \hat{a}_2)$. Note that the map E is continuously differentiable. Note also that

$$DE = \begin{bmatrix} \dfrac{\partial a_1}{\partial h_1} & \dfrac{\partial a_1}{\partial h_2} \\ \dfrac{\partial a_2}{\partial h_1} & \dfrac{\partial a_2}{\partial h_2} \end{bmatrix}.$$

To show that DE is invertible, we merely need to show that $\det(DE) \neq 0$, or that

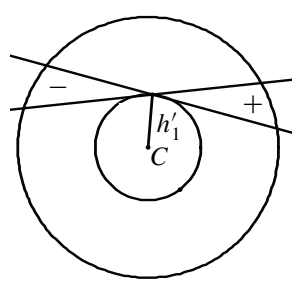
$$\frac{\partial a_1}{\partial h_1}\frac{\partial a_2}{\partial h_2} \neq \frac{\partial a_2}{\partial h_1}\frac{\partial a_1}{\partial h_2}.$$

It is easy to see from the figure below that, as h_i increases, so must a_i; therefore $\partial a_i / \partial h_i > 0$.



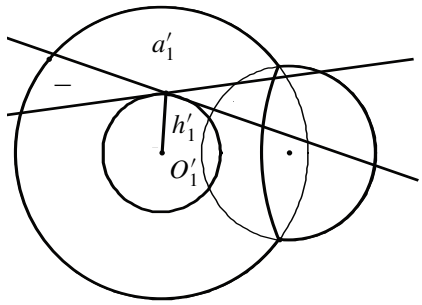
Claim. *When h_2 increases, $\partial a_1/\partial h_2 \leq 0$ and $|\partial a_1/\partial h_2| < |\partial a_1/\partial h_1|$.*

Proof. Let B_1 be a single bubble, as in the figure:



Note that h_1 is constant here. Since the slicing line must be tangent to the circle centered at o_1 with a radius of h_1, both slicing lines must intersect at a point, losing area on the left and gaining area on the right. By basic trigonometry, these

two areas will be additive inverses. Now if we add B_2, some of the area on the right is lost to B_2, like this:



So as h_2 increases, a_1 will lose more area than it gains, making $\partial a_1/\partial h_2 \leq 0$. Note also that $|\partial a_1/\partial h_2|$, the area lost by adding B_2, represents a subset of the area gained by a_2 as h_2 increases ($|\partial a_2/\partial h_2|$). So $|\partial a_1/\partial h_2| < |\partial a_1/\partial h_1|$. □

We claim by the same method that $\partial a_2/\partial h_1 \leq 0$ and $|\partial a_2/\partial h_1| < |\partial a_2/\partial h_2|$.

It follows that $(\partial a_1/\partial h_1)(\partial a_2/\partial h_2)$ and $(\partial a_1/\partial h_2)(\partial a_2/\partial h_1)$ are always non-negative,

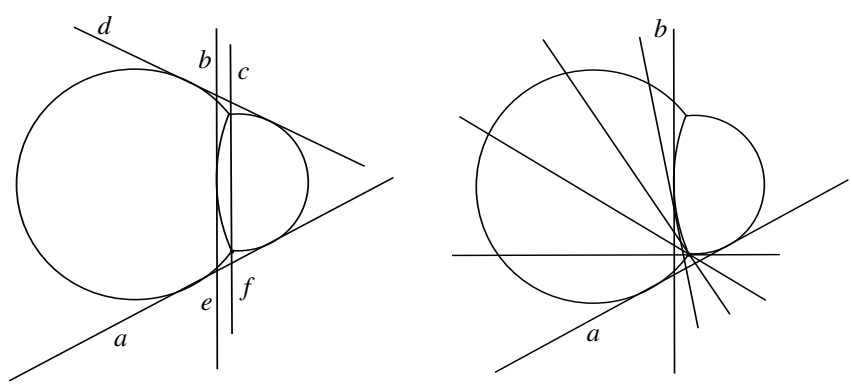
$$(\partial a_1/\partial h_2)(\partial a_2/\partial h_2) < (\partial a_1/\partial h_1)(\partial a_2/\partial h_2),$$

and $\det(DE) \neq 0$. Hence DE is invertible everywhere on $\operatorname{int}(\mathbb{H})$.

Thus the inverse function theorem implies that E is locally a bijection ($F = E^{-1}$ exists) and that F is continuously differentiable. The map F is also one-to-one since $F(a_1, a_2) = F(b_1, b_2) = (h_1, h_2)$ implies $(a_1, a_2) = (b_1, b_2) = E(h_1, h_2)$.

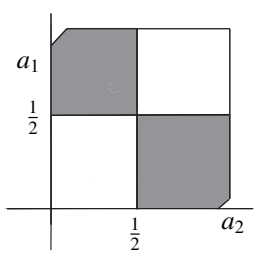
We complete the proof of Lemma 4.2 by showing that $F = E^{-1} : \partial K \to \partial\mathbb{H}$ exists and is one-to-one. We do this by describing a smooth bijection from ∂K onto $\partial\mathbb{H}$.

For the slicing line to be on the boundary of K, it must be tangent to the boundary of at least one of the bubbles (save on the extremes where $|h_1 - h_2| = |o_1 - o_2|$). Imagine the line tangent to both bubbles on the bottom, marked a in the figure:



We'll move the line upward in bubble 1, keeping the line tangent to the bottom of bubble 2, until we reach line b, where it is vertical, touching the left side of bubble 2. It then slides horizontally to the right to touch the vertices of the bubbles (line c). The line will rotate upwards in bubble 2 while touching the boundary of bubble 1 until it is tangent to the top of both bubbles (line d). Then the line will stay touching the boundary of bubble 2 and rotate down through bubble 1 until it is once again vertical touching the boundary (line e). It will shift right until it touches the intersections (line f). Finally, the line will touch bubble 1's boundary and move down through bubble 2 until reaching the bottom of both bubbles and the starting point (line a). In this way, we've smoothly and bijectively traversed all of the slicing lines that are on the boundary of K. (Note that slicing lines are of necessity *oriented*, as each slicing line has a defined region *below* the line.) $\square$

This implies that if $\sigma(t) \in K$ for all $t \in [0, 1]$, $F(\sigma(t))$ describes a continuously differentiable path in $\mathbb{H}$ where $(\hat{a}_1, \hat{a}_2)|_{F(\sigma(t))} = \sigma(t)$. To ensure that $\sigma(t)$ is always in the domain of F, we place the following restriction on the parametrizations of a competitor σ. Let t_0 denote the line that passes through the centroids of B_1 and B_2. (If these coincide, any line passing through them is sufficient.) Parametrize the competitor σ such that all slicing lines are parallel to t_0. Since t_0 passes through the centroids of B_1 and B_2, it cuts their areas exactly in half, and $\sigma(t_0) = (A_1/2, A_2/2)$ for all competitors σ. Now $\sigma(t)$ is nondecreasing in both a_1 and a_2, so $\sigma(t) \in [0, A_1/2] \times [0, A_2/2] \cup [A_1/2, A_1] \times [A_2/2, A_2] \subset K$ for all t. confining it to the white part of the figure. Hence the mapping F may be applied to any competitor σ given this orientation of slicing lines.



This map allows the following definition of $G(\sigma)$.

Theorem 4.3. *Let* $f_\sigma = \sum(2A_i - h_i l_i)/r_i$ *for* $i = 1, 2$. *The function* $G(\sigma) = \int_{t_0}^{t_1} f'_\sigma dt$ *calibrates the standard double bubble.*

We prove the three conditions for a calibration function separately in Lemmas 4.4, 4.5, and 4.10.

Lemma 4.4. $G(\mu) = P(\mu)$ *for* μ *a standard double bubble.*

Proof. Consider the vector fields $\boldsymbol{V} = (x, y)/r_1$ and $\boldsymbol{W} = (x-m, y)/r_2$, where $(0, 0) = o_1$ and $(m, 0) = o_2$. We claim the total perimeter of a standard double bubble, $P(\mu)$, is equal to

$$\int_{\partial \hat{B}_1} \boldsymbol{V} \cdot \boldsymbol{N} ds + \int_{\partial \hat{B}_2} \boldsymbol{W} \cdot \boldsymbol{N} ds, \tag{1}$$

where $\boldsymbol{N}$ is the unit normal to the surface at that point. To see this, note that on the outer arcs of radius r_1 and r_2, $\boldsymbol{V} = \boldsymbol{N}$ and $\boldsymbol{W} = \boldsymbol{N}$, and the integral over these arcs reduces to $\int 1 ds$. The integral over the center arc reduces to $\int (\boldsymbol{W} - \boldsymbol{V}) \cdot \boldsymbol{N} ds$, with the appropriate normal vector. However,

$$\begin{aligned} \boldsymbol{W} - \boldsymbol{V} &= \frac{1}{r_2}(x - m, y) - \frac{1}{r_1}(x, y) = \Big(\frac{1}{r_2} - \frac{1}{r_1}\Big)(x, y) - \frac{1}{r_2}(m, 0) \\ &= \frac{1}{r_3}(x, y) - \frac{1}{r_2}(m, 0) = \frac{1}{r_3}\Big(x - \frac{r_3}{r_2} m, y\Big). \end{aligned}$$

If we denote the intersection of the three arcs in a standard double bubble by p, we see by the fact that these arcs meet at 120° angles that $m\angle o_1 p o_2 = 60°$ and $m\angle o_1 p o_3 = 120°$. By application of the law of sines we find that

$$\frac{|o_1 - o_2|}{\sin 60°} = \frac{r_2}{\sin(m\angle o_2 o_1 p)} \quad \text{and} \quad \frac{|o_1 - o_3|}{\sin 120°} = \frac{r_3}{\sin(m\angle o_2 o_1 p)},$$

which reduces to

$$\frac{|o_1 - o_2|}{r_2 \sin 60°} = \frac{|o_1 - o_3|}{r_3 \sin 120°} \frac{r_3 |o_1 - o_2|}{r_2} = |o_1 - o_3| \frac{r_3}{r_2} m = |o_1 - o_3|.$$

Thus $o_3 = (\frac{r_3}{r_2} m, 0)$, and the vector $\boldsymbol{W} - \boldsymbol{V} = \frac{1}{r_3}(x - \frac{r_3}{r_2} m, y)$ is equal to the unit normal on the remaining arc. Thus the line integral on this arc also evaluates to $\int 1 ds$. Since the line integrals over all three arcs evaluate to $\int 1 ds$, or the length of each arc, the expression in (1) above is equal to $P(\mu)$. As a consequence of the divergence theorem, however, we have

$$\begin{aligned} P(\mu) &= \int_{\partial \hat{B}_1} \boldsymbol{V} \cdot \boldsymbol{N}\, ds + \int_{\partial \hat{B}_2} \boldsymbol{W} \cdot \boldsymbol{N}\, ds = \int_{\hat{B}_1} \operatorname{div} \boldsymbol{V}\, dA + \int_{\hat{B}_2} \operatorname{div} \boldsymbol{W}\, dA \\ &= \int_{\hat{B}_1} \frac{\partial}{\partial x}\Big(\frac{x}{r_1}\Big) + \frac{\partial}{\partial y}\Big(\frac{y}{r_1}\Big)\, dA + \int_{\hat{B}_2} \frac{\partial}{\partial x}\Big(\frac{x-m}{r_2}\Big) + \frac{\partial}{\partial y}\Big(\frac{y}{r_2}\Big)\, dA \\ &= \int_{\hat{B}_1} \frac{1}{r_1} + \frac{1}{r_1}\, dA + \int_{\hat{B}_2} \frac{1}{r_2} + \frac{1}{r_2}\, dA = \int_{\hat{B}_1} \frac{2}{r_1}\, dA + \int_{\hat{B}_2} \frac{2}{r_2}\, dA = \frac{2A_1}{r_1} + \frac{2A_2}{r_2}. \end{aligned}$$

Now consider $G(\sigma)$:

$$G(\sigma) = \int_0^1 f'_\sigma \, dt = f_\sigma(1) - f_\sigma(0) = \sum \frac{2a_i - h_i l_i}{r_i}|_{t=1} - \sum \frac{2a_i - h_i l_i}{r_i}|_{t=0}$$
$$= \sum \frac{2A_i - 1\cdot 0}{r_i} - \sum \frac{2\cdot 0 - 1\cdot 0}{r_i} = \frac{2A_1}{r_1} + \frac{2A_2}{r_2}.$$

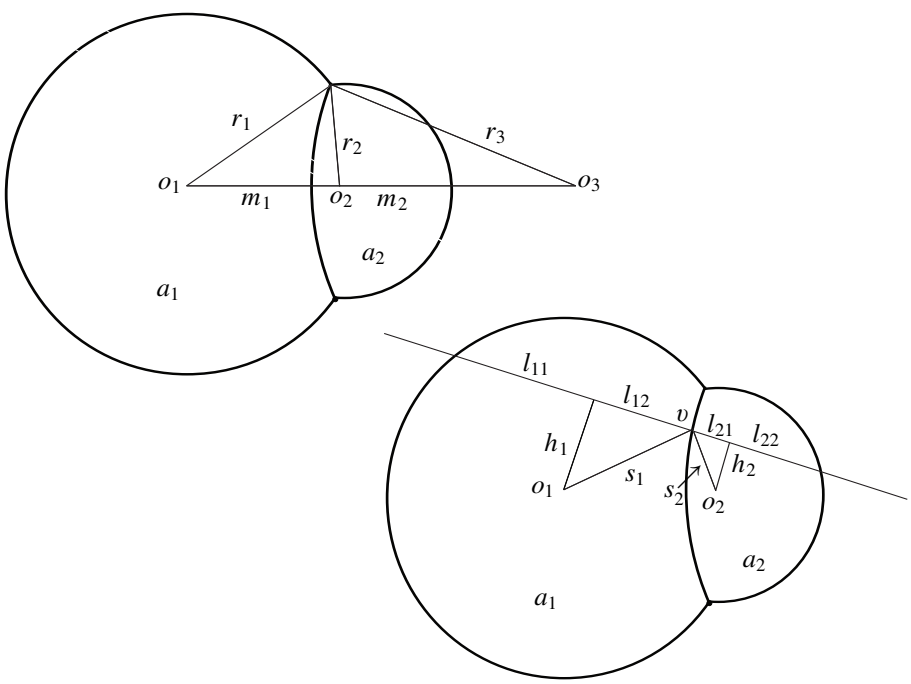
This is true for all competitors σ, including the standard double bubble μ; thus $P(\mu) = G(\mu)$. □

Lemma 4.5. *$G(\sigma) = G(\mu)$ for all competitors σ, where μ is the standard double bubble.*

Proof. It was shown in Lemma 4.4 that $G(\sigma) = 2A_1/r_1 + 2A_2/r_2$ for all competitors σ, including μ. Thus $G(\sigma) = G(\mu)$ for all competitors σ. □

The final condition for calibration, that $G(\sigma) \le P(\sigma)$ for all competitors σ, will be proved in Lemma 4.10 as a result of the following propositions. For this section we introduce the notation $G_\sigma(t) = \int_{t_0}^t f'_\sigma dt$ and $P_\sigma(t) =$ the total perimeter of σ lying below the slicing line t. We will show that $G'_\sigma(t) \le P'_\sigma(t)$ for all $t \in [t_0, t_1]$, and $G(\sigma) \le P(\sigma)$ will result by integration.

In the following propositions we will use the notation shown in these figures:



Let $m_1 = |o_1 - o_2|$ (the m of Lemma 4.4) and $m_2 = |o_2 - o_3|$. Also let v be the intersection of the slicing line (h_1, h_2) with the center arc between $\hat{B}_1$ and $\hat{B}_2$.

Define s_1, s_2, as the distances between v and o_1, o_2 respectively. The lengths l_{11}, l_{12}, l_{21}, and l_{22} are defined as shown, where $l_{11} + l_{12} = \hat{l}_1$ and $l_{21} + l_{22} = \hat{l}_2$.

Proposition 4.6. *For all slices* (h_1, h_2) *of the standard double bubble*,

$$\frac{l_{11}^2 - l_{12}^2}{r_1} = \frac{l_{21}^2 - l_{22}^2}{r_2}.$$

Proof. It was shown in Lemma 4.4 that $o_1o_3 = m_1 + m_2 = (r_3/r_2)m_1$. Given that $r_3 = r_1r_2/(r_1 - r_2)$, we find that $m_1/m_2 = (r_1 - r_2)/r_2$. Noting that $\cos(\angle o_1o_2v) = -\cos(\angle vo_2o_3)$, by application of the law of cosines we find

$$\frac{m_1^2 + s_2^2 - s_1^2}{2s_2m_1} = -\frac{s_2^2 + m_2^2 - r_3^2}{2s_2m_2},$$

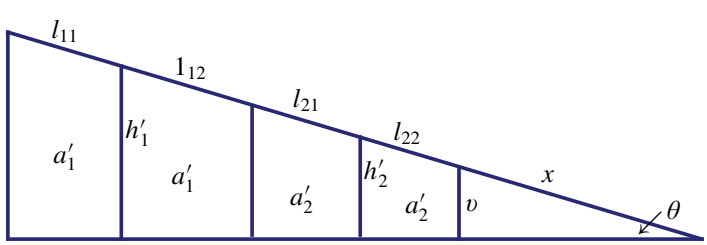
from which we obtain, successively,

$$\begin{aligned}
-s_2^2\left(1 + \frac{m_1}{m_2}\right) &= m_1^2 + \frac{m_1}{m_2}(m_2^2 - r_3^2) - s_1^2,\\
-s_2^2\left(1 + \frac{r_1 - r_2}{r_2}\right) &= r_1^2 + r_2^2 - r_1r_2 + \frac{r_1 - r_2}{r_2}(r_2^2 + r_3^2 - r_2r_3 - r_3^2) - s_1^2,\\
-s_2^2\,\frac{r_1}{r_2} &= r_1^2 + r_2^2 - r_1r_2 + \frac{r_1 - r_2}{r_2}\left(r_2^2 - r_2\,\frac{r_1r_2}{r_1 - r_2}\right) - s_1^2,\\
r_1r_2 - s_2^2\,\frac{r_1}{r_2} &= r_1^2 + r_2^2 + \frac{r_1 - r_2}{r_2}\,\frac{-r_2^3}{r_1 - r_2} - s_1^2 = r_1^2 - s_1^2,\\
\frac{r_2^2 - s_2^2}{r_2} &= \frac{r_1^2 - s_1^2}{r_1}.
\end{aligned}$$

Now by the Pythagorean theorem we find $l_{21}^2 - l_{22}^2 = r_2^2 - h_2^2 - (s_2^2 - h_2^2) = r_2^2 - s_2^2$ and $l_{11}^2 - l_{12}^2 = r_1^2 - h_1^2 - (s_1^2 - h_1^2) = r_1^2 - s_1^2$. Substitution shows that $(l_{11}^2 - l_{12}^2)/r_1 = (l_{21}^2 - l_{22}^2)/r_2$. □

Proposition 4.7. $\sum((2 - h_i')a_i')/r_i \le \sum \hat{l}_i/r_i$.

Proof. Let $l_{11}, l_{12}, l_{21}, l_{22}, h_1$, and h_2 be defined as above. Consider the diagram



which describes the instantaneous change in (h_1, h_2). Unlike the parametrizations of slicing lines in the circle, in the standard double bubble slicing lines may rotate.

This rotation is described by the relative angle θ' and the distance x between the axis of rotation and the double bubble.

We can compute the total change in area in both $\hat{B}_1$ and $\hat{B}_2$ as follows:

$$A_1' = \tfrac{1}{2}\theta'\big((x+\hat{l}_2+\hat{l}_1)^2-(x+\hat{l}_2)^2\big) = \tfrac{1}{2}\theta'(2x+2\hat{l}_2+\hat{l}_1)(\hat{l}_1),$$
$$A_2' = \tfrac{1}{2}\theta'\big((x+\hat{l}_2)^2-(x)^2\big) = \tfrac{1}{2}\theta'(2x+\hat{l}_2)(\hat{l}_2).$$

We also note that $h_1' = \theta'(x+\hat{l}_2+l_{12})$ and $h_2' = \theta'(x+l_{22})$. Thus

$$\begin{aligned}\sum \frac{(2-h_i')A_i'}{r_i} &= \frac{(2-h_1')(\theta'/2)(2x+2\hat{l}_2+\hat{l}_1)(\hat{l}_1)}{r_1} + \frac{(2-h_2')(\theta'/2)(2x+\hat{l}_2)(\hat{l}_2)}{r_2}\\ &= \frac{(2-h_1')\dfrac{h_1'}{2(x+\hat{l}_2+l_{12})}(2x+2\hat{l}_2+\hat{l}_1)(\hat{l}_1)}{r_1} + \frac{(2-h_2')\dfrac{h_2'}{2(x+l_{22})}(2x+\hat{l}_2)(\hat{l}_2)}{r_2}.\end{aligned}$$

Both of these components are maximized when $h_1' = h_2' = 1$, so we have:

$$\begin{aligned}\sum \frac{(2-h_i')A_i'}{r_i} &\le \frac{(2x+2\hat{l}_2+\hat{l}_1)(\hat{l}_1)}{2(x+\hat{l}_2+l_{12})r_1} + \frac{(2x+\hat{l}_2)(\hat{l}_2)}{2(x+l_{22})r_2}\\ &= \frac{\hat{l}_1}{r_1} + \frac{\hat{l}_2}{r_2} + \frac{(l_{11}-l_{12})(l_{12}+l_{11})}{2(x+\hat{l}_2+l_{12})r_1} + \frac{(l_{21}-l_{22})(l_{22}+l_{21})}{2(x+l_{22})r_2}\\ &= \frac{\hat{l}_1}{r_1} + \frac{\hat{l}_2}{r_2} + \frac{l_{11}^2-l_{12}^2}{2(x+\hat{l}_2+l_{12})r_1} + \frac{l_{21}^2-l_{22}^2}{2(x+l_{22})r_2}.\end{aligned}$$

By Proposition 4.6 we find $\frac{l_{11}^2-l_{12}^2}{r_1} = \dfrac{l_{21}^2-l_{22}^2}{r_2}$. Using this substitution we find

$$\begin{aligned}\frac{l_{11}^2-l_{12}^2}{2(x+\hat{l}_2+l_{12})r_1} + \frac{l_{21}^2-l_{22}^2}{2(x+l_{22})r_2} &= \frac{l_{21}^2-l_{22}^2}{2r_2}\left(\frac{-1}{x+\hat{l}_2+l_{12}} + \frac{1}{x+l_{22}}\right)\\ &= \frac{l_{21}^2-l_{22}^2}{2r_2}\,\frac{l_{12}+l_{21}}{(x+\hat{l}_2+l_{12})(x+l_{22})}.\end{aligned}$$

Now since $x+l_{22} > 0$ and $l_{22} > l_{21}$, this term is always negative, and maximized at zero as $a \to \infty$. Thus

$$\begin{aligned}\sum \frac{(2-h_i')A_i'}{r_i} &\le \frac{\hat{l}_1}{r_1} + \frac{\hat{l}_2}{r_2} + \frac{l_{21}^2-l_{22}^2}{2r_2}\,\frac{l_{12}+l_{21}}{(x+\hat{l}_2+l_{12})(x+l_{22})}\\ &\le \frac{\hat{l}_1}{r_1} + \frac{\hat{l}_2}{r_2} = \sum \frac{\hat{l}_i}{r_i}.\end{aligned}$$

□

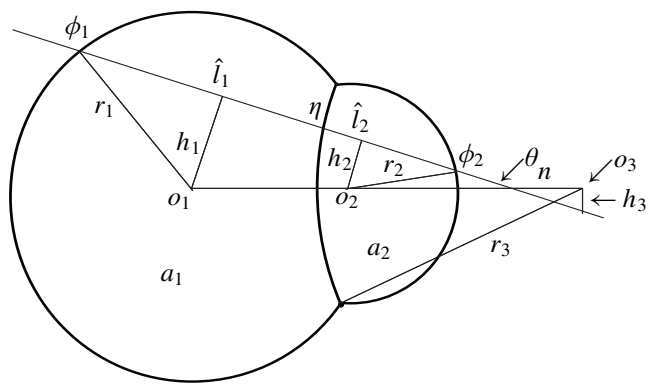
Proposition 4.8. *For all slices of a standard double bubble, we have*

$$\left(\frac{l_{12}}{r_1}+\frac{l_{21}}{r_2}\right)^2+\left(\frac{h_1}{r_1}-\frac{h_2}{r_2}\right)^2=1.$$

Proof. We will show that

$$\frac{h_1}{r_1}-\frac{h_2}{r_2}=-\cos\eta \quad \text{and} \quad \frac{l_{12}}{r_1}+\frac{l_{21}}{r_2}=-\sin\eta, \tag{2}$$

where the angle η is defined in the figure:



The proposition results from the equalities in (2). To show them, we prove that

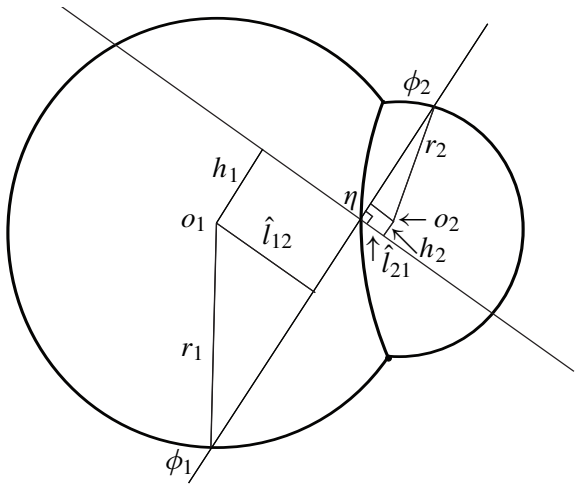
$$\cos\phi_1-\cos\phi_2+\cos\eta=0.$$

for any slice of a standard double bubble. Note that $\cos\phi_1=h_1/r_1$, $\cos\phi_2=h_2/r_2$, and $\cos\eta=h_3/r_3$. (Here h_3 is the signed distance from the slicing line to o_3, measured similarly to h_1 and h_2.) Let n be the signed distance from o_3 to the intersection of the slicing line and the line through the centers, and θ the angle formed at the intersection, as in the figure above.

We find that $h_1=(m_1+m_2+n)\sin\theta$, $h_2=(m_2+n)\sin\theta$, and $h_3=n\sin\theta$. Hence

$$\begin{aligned}
&\cos\phi_1-\cos\phi_2+\cos\eta\\
&\quad=\frac{h_1}{r_1}-\frac{h_2}{r_2}+\frac{h_3}{r_3}=\frac{(m_1+m_2+n)\sin\theta}{r_1}-\frac{(m_2+n)\sin\theta}{r_2}+\frac{n\sin\theta}{r_3}\\
&\quad=\big(r_2(m_1+m_2+n)-r_1(m_2+n)+(r_1-r_2)n\big)\frac{\sin\theta}{r_1r_2}\\
&\quad=\left(r_2\Big(\frac{m_1}{m_2}+1\Big)-r_1\right)\frac{m_2\sin\theta}{r_1r_2}=\left(r_2\Big(\frac{r_1-r_2}{r_2}+1\Big)-r_1\right)\frac{m_2\sin\theta}{r_1r_2}=0.
\end{aligned}$$

This implies immediately that $h_1/r_1-h_2/r_2=-\cos\eta$. Now consider a slicing line perpendicular to the original at v, as in the figure on the top of the next page. In this slicing, we see that $l_{12}/r_1=\cos\phi_1$ and $l_{21}/r_2=-\cos\phi_2$. By the property just



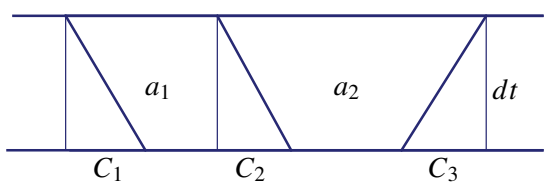
Towards the proof of Proposition 4.8.

proven, $\cos\phi_1 - \cos\phi_2 = -\cos(90^\circ + \eta)$, or equivalently $l_{12}/r_1 + l_{21}/r_2 = -\sin\eta$. These two relations together finish the proof. □

Proposition 4.9. *The following identity holds for all slices t of any competitor σ:*

$$\sqrt{1-\left(\frac{h_1}{r_1}\right)^2} + \sqrt{1-\left(\frac{h_1}{r_1}-\frac{h_2}{r_2}\right)^2} + \sqrt{1-\left(\frac{h_2}{r_2}\right)^2} + \sum \frac{-h_i l_i'}{r_i} \le P_\sigma'(t).$$

Proof. Suppose the slicing line t intersects σ in three locations. Consider $P_\sigma'(t)$, shown in the figure:



Note that $\|(c_1, dt)\| + \|(c_2, dt)\| + \|(c_3, dt)\| = P_\sigma'(t)\,dt$. Now consider the unit vectors

$$\left(-h_1, \sqrt{1-h_1^2}\right),\ \left(-h_1+h_2, \sqrt{1-(h_1-h_2)^2}\right),\ \left(-h_2, \sqrt{1-h_2^2}\right).$$

By the Cauchy–Schwartz inequality,

$$\left(-h_1,\sqrt{1-h_1^2}\right)\cdot(c_1,dt)+\left(h_2-h_1,\sqrt{1-(h_1-h_2)^2}\right)\cdot(c_2,dt)+\left(-h_2,\sqrt{1-h_2^2}\right)\cdot(c_3,dt)$$
$$\le \|(c_1,dt)\|+\|(c_2,dt)\|+\|(c_3,dt)\| = P_\sigma'(t)\,dt.$$

Noting that $(c_1+c_2)/dt = l_1'$ and $(c_3-c_2)/dt = l_2'$, this reduces to

$$\sqrt{1-\left(\frac{h_1}{r_1}\right)^2} + \sqrt{1-\left(\frac{h_1}{r_1}-\frac{h_2}{r_2}\right)^2} + \sqrt{1-\left(\frac{h_2}{r_2}\right)^2} + \sum \frac{-h_i l_i'}{r_i} \le P_\sigma'(t).$$

Now this inequality still holds even if t intersects σ in more than three locations by ignoring any additional intersections. There are however additional cases that are not covered by the above proof (such as when there are only two intersections), but for the sake of brevity we simply assert that proof of all other cases continues in much the same fashion. □

Lemma 4.10. *$G(\sigma) \le P(\sigma)$ for all competitors σ.*

Proof. We use the preceding propositions to show that $G'_\sigma(t) \le P'_\sigma(t)$ for all $t \in [0, 1]$:

$$\begin{aligned}
G'_\sigma(t) &= \sum \frac{2A'_i - h'_i l_i - h_i l'_i}{r_i} = \sum \frac{(2-h'_i)A'_i - h_i l'_i}{r_i} \\
&\le \sum \frac{\hat{l}_i - h_i l'_i}{r_i} \qquad \text{(by Proposition 4.7)} \\
&= \frac{l_{11}}{r_1} + \frac{l_{12}}{r_1} + \frac{l_{21}}{r_2} + \frac{l_{22}}{r_2} + \sum \frac{-h_i l'_i}{r_i} \\
&= \sqrt{1 - \left(\frac{h_1}{r_1}\right)^2} + \sqrt{1 - \left(\frac{h_1}{r_1} - \frac{h_2}{r_2}\right)^2} + \sqrt{1 - \left(\frac{h_2}{r_2}\right)^2} + \sum \frac{-h_i l'_i}{r_i} \\
&\qquad \text{(by the Pythagorean theorem and Proposition 4.8)} \\
&\le P'_\sigma(t) \qquad \text{(by Proposition 4.9).}
\end{aligned}$$

Noting that $P_\sigma(0) = 0$, we complete the proof by integration: $G'_\sigma(t) \le P'_\sigma(t)$, so

$$\int_0^1 G'_\sigma(t)\,dt \le \int_0^1 P'_\sigma(t)\,dt,$$

and therefore $G(\sigma) \le P_\sigma(1) - P_\sigma(0)$ and $G(\sigma) \le P(\sigma)$. □

This completes the proof of Theorem 4.3, namely that the standard double bubble minimizes perimeter among all figures (unions of C^1 manifolds) that separately enclose two fixed areas.

5. Further research

Extending the work that we have here presented seems well within the grasp of many undergraduate researchers. We are currently working on extending it to include many other such problems including soap films on wire frames, both with and without trapped bubbles. These problems are uniquely suited to metacalibrations because they include both fixed volume and fixed boundary constraints. The planar problem for three or more bubbles has proven somewhat more difficult to tackle, in large part because we no longer have the topological property of being able to match areas under a given slicing line, which was possible with two areas. Alternative slicing methods are being investigated to attack this problem.

Another interesting problem under investigation is that of generalizing the above double bubble proof to n dimensions, hopefully providing a compelling alternative to current proofs for the double bubble in $\mathbb{R}^n$. It is hoped that metacalibrations will become a useful tool to solve problems in geometric optimization.

Acknowledgements

The authors respectfully thank coresearchers James Dilts and Drew Johnson for their assistance and contributions to the proof. This research was funded by the College of Physical and Mathematical Sciences at Brigham Young University.

References

[Almgren 1976] F. J. Almgren, Jr., "Existence and regularity almost everywhere of solutions to elliptic variational problems with constraints", *Mem. Amer. Math. Soc.* **4**:165 (1976), viii+199. MR 54 #8420 Zbl 0327.49043

[Foisy et al. 1993] J. Foisy, M. Alfaro, J. Brock, N. Hodges, and J. Zimba, "The standard double soap bubble in $\mathbf{R}^2$ uniquely minimizes perimeter", *Pacific J. Math.* **159**:1 (1993), 47–59. MR 94b:53019 Zbl 0738.49023

[Isenberg 1978] C. Isenberg, *The science of soap films and soap bubbles*, Tieto Ltd., Clevedon, 1978. MR 83b:00001 Zbl 0447.76001

[Milman and Schechtman 1986] V. D. Milman and G. Schechtman, *Asymptotic theory of finite-dimensional normed spaces*, Lecture Notes in Math. **1200**, Springer, Berlin, 1986. MR 87m:46038 Zbl 0606.46013

[Morgan 1994] F. Morgan, "Soap bubbles in $\mathbf{R}^2$ and in surfaces", *Pacific J. Math.* **165**:2 (1994), 347–361. MR 96a:58064 Zbl 0820.53002

[Siegel 2003] A. Siegel, "A historical review of the isoperimetric theorem in 2-d, and its place in elementary plane geometry", preprint, Courant Institute, 2003, Available at http://www.cs.nyu.edu/faculty/siegel/SCIAM.pdf.

[Taylor 1976] J. E. Taylor, "The structure of singularities in soap-bubble-like and soap-film-like minimal surfaces", *Ann. of Math.* (2) **103**:3 (1976), 489–539. MR 55 #1208a Zbl 0335.49032

[Wichiramala 2004] W. Wichiramala, "Proof of the planar triple bubble conjecture", *J. Reine Angew. Math.* **567** (2004), 1–49. MR 2005a:53013 Zbl 1078.53010

Received: 2009-10-22 Accepted: 2009-10-23

beccadorff@gmail.com *Mathematics Department, Brigham Young University, Provo, UT 84602, United States*

lawlor@mathed.byu.edu *Department of Mathematics Education, Brigham Young University, 185 TMCB, Provo, UT 84602, United States*

dsampson@byu.net *Mathematics Department, Brigham Young University, Provo, UT 84602, United States*
http://sites.google.com/site/sampsondcs/

mechanicaltrombone@hotmail.com
Mathematics Department, Duke University, Box 90320, Durham, NC 27708-0320, United States

Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the Submission page at the Involve website.

Originality. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles in *Involve* are usually in English, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

Format. Authors are encouraged to use LaTeX but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@mathscipub.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

2009 vol. 2 no. 5

On the orbits of an orthogonal group action 495
KYLE CZARNECKI, R. MICHAEL HOWE AND AARON MCTAVISH

Symbolic computation of degree-three covariants for a binary form 511
THOMAS R. HAGEDORN AND GLEN M. WILSON

Isometric composition operators acting on the Chebyshev space 533
THOMAS E. GOEBELER, JR. AND ASHLEY L. POTTER

Markov partitions for hyperbolic sets 549
TODD FISHER AND HIMAL RATHNAKUMARA

Ineffective perturbations in a planar elastica 559
KAITLYN PETERSON AND ROBERT MANNING

A tiling approach to Fibonacci product identities 581
JACOB ARTZ AND MICHAEL ROWELL

Frame theory for binary vector spaces 589
BERNHARD G. BODMANN, MY LE, LETTY REZA, MATTHEW TOBIN AND MARK TOMFORDE

Some results on the size of sum and product sets of finite sets of real numbers 603
DERRICK HART AND ALEXANDER NIZIOLEK

Proof of the planar double bubble conjecture using metacalibration methods 611
REBECCA DORFF, GARY LAWLOR, DONALD SAMPSON AND BRANDON WILSON

1944-4176(2009)2:5;1-7