

# involve

a journal of mathematics

## Editorial Board

Kenneth S. Berenhaut, *Managing Editor*

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



## MANAGING EDITOR

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

## BOARD OF EDITORS

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

## PRODUCTION

Silvio Levy, Scientific Editor


Cover: Alex Scorpan

See inside back cover or [msp.org/involve](http://msp.org/involve) for submission instructions. The subscription price for 2015 is US \$140/year for the electronic version, and \$190/year (+\$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to MSP.

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

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

PUBLISHED BY

 **mathematical sciences publishers**  
nonprofit scientific publishing

<http://msp.org/>

© 2015 Mathematical Sciences Publishers

# The $\Delta^2$ conjecture holds for graphs of small order

Cole Franks

(Communicated by Ronald Gould)

An  $L(2, 1)$ -labeling of a simple graph  $G$  is a function  $f : V(G) \rightarrow \mathbb{Z}$  such that if  $xy \in E(G)$ , then  $|f(x) - f(y)| \geq 2$ , and if the distance between  $x$  and  $y$  is two, then  $|f(x) - f(y)| \geq 1$ .  $L(2, 1)$ -labelings are motivated by radio channel assignment problems. Denote by  $\lambda_{2,1}(G)$  the smallest integer such that there exists an  $L(2, 1)$ -labeling of  $G$  using the integers  $\{0, \dots, \lambda_{2,1}(G)\}$ . We prove that  $\lambda_{2,1}(G) \leq \Delta^2$ , where  $\Delta = \Delta(G)$ , if the order of  $G$  is no greater than  $(\lfloor \Delta/2 \rfloor + 1)(\Delta^2 - \Delta + 1) - 1$ . This shows that for graphs no larger than the given order, the 1992 “ $\Delta^2$  conjecture” of Griggs and Yeh holds. In fact, we prove more generally that if  $L \geq \Delta^2 + 1$ ,  $\Delta \geq 1$ , and

$$|V(G)| \leq (L - \Delta) \left( \left\lfloor \frac{L-1}{2\Delta} \right\rfloor + 1 \right) - 1,$$

then  $\lambda_{2,1}(G) \leq L - 1$ . In addition, we exhibit an infinite family of graphs with  $\lambda_{2,1}(G) = \Delta^2 - \Delta + 1$ .

## 1. Introduction

The *channel assignment problem* is the determination of assignments of channels (integers) to stations in such a way that those stations close enough to interfere receive distant enough channels. Hale [1980] formulated the problem in terms of  $T$ -colorings, which are integer colorings in which adjacent vertices' colors cannot differ by a member of a set of integers  $T$  with  $\{0\} \subset T$ . Roberts [1988] proposed a generalization in which closer transmitters would be required to have channels that differed by more than those of the slightly more distant transmitters, adding a condition for nonadjacent vertices as well. The  $L(2, 1)$ -labeling problem was first studied by Griggs and Yeh [1992] in response to Roberts' proposal. An  $L(2, 1)$ -labeling of a graph  $G$  is an integer labeling of  $G$  in which two vertices at distance one from each other must have labels differing by at least 2, and those at distance two must differ by at least 1. Denote by  $\lambda_{2,1}(G)$  the smallest number such that there exists an  $L(2, 1)$ -labeling of  $G$  with the difference  $\lambda_{2,1}(G)$  between the highest and

MSC2010: 97K30.

Keywords:  $L(2,1)$ -labeling, graph labeling, channel assignment.

lowest label. If there is no possibility for confusion,  $\lambda_{2,1}(G)$  is sometimes written  $\lambda_{2,1}$ . The  $L(2, 1)$ -labeling problem has been studied extensively with the central goal of finding bounds on  $\lambda_{2,1}$ . Griggs and Yeh bounded the  $\lambda_{2,1}$  number for cycles, paths, trees, and the  $n$ -cube. They also proved the bound  $\lambda_{2,1} \leq \Delta(G)^2 + 2\Delta(G)$ , where  $\Delta(G)$  is the maximum degree over the set of degrees of vertices in  $V(G)$ . In this paper, we will write  $\Delta$  when the meaning is clear from context. Chang and Kuo [1996] improved the bound to  $\Delta^2 + \Delta$ , and by modifying their algorithm, Gonçalves [2007] reduced the bound to  $\Delta^2 + \Delta - 2$ . Bounds on the  $\lambda_{2,1}$  number have been found for many subclasses of graphs, such as Sakai's bound [1991] of  $(\Delta + 3)^2/4$  for *chordal* graphs — graphs containing no induced cycle of length four. All examples tested have corroborated the conjecture Griggs and Yeh made in their 1992 paper:

$\Delta^2$  **conjecture.** If  $\Delta(G) \geq 2$ , then  $\lambda_{2,1} \leq \Delta^2$ .

However, the conjecture remains unproven, and it is difficult to test the bound for graphs of any significant size. The largest step towards the proof of the conjecture was made by Havet, Reed, and Sereni [2012] who proved that the conjecture holds for all graphs with  $\Delta$  larger than some  $\Delta_0$ , but  $\Delta_0 \approx 10^{69}$ . Consequently,  $\lambda_{2,1}(G) \leq \Delta^2 + C$  for some absolute constant  $C$ . The upper bound set by the conjecture, if proven, would be tight — the Moore graphs are known to satisfy  $\lambda_{2,1} = \Delta^2$  [Griggs and Yeh 1992].

## 2. Preliminaries

The proof of [Theorem 3](#) involves a classic result of Pósa about the existence of Hamilton cycles and paths in graphs of high degree (see [Kronk 1969]). In this respect, our argument has a similar flavor to the proof in [Griggs and Yeh 1992] that  $\lambda_{2,1} \leq \Delta^2$  for graphs of order less than  $\Delta^2 + 1$ . In addition, we will use the powerful result of Szemerédi and Hajnal [1970] on equitable colorings.

**Theorem** (Pósa). *Let  $G$  have  $n \geq 3$  vertices. If for every  $k$ ,  $1 \leq k \leq (n - 1)/2$  and  $|\{v : d(v) \leq k\}| < k$ , then  $G$  is Hamiltonian.*

**Corollary 1.** *Let  $G$  have  $n \geq 2$  vertices. If for every  $k$ ,  $0 \leq k \leq (n - 2)/2$  and  $|\{v : d(v) \leq k\}| \leq k$ , then  $G$  has a Hamilton path.*

*Proof.* The corollary follows easily by adding a dominating vertex to  $G$  and observing that by Pósa's theorem the new graph is Hamiltonian.  $\square$

**Theorem** (Szemerédi, Hajnal). *If  $\Delta(G) \leq r$ , then  $G$  can be equitably colored with  $r + 1$  colors; that is, the sizes of the color classes differ by at most one.*

See also [Kierstead et al. 2010; Kierstead and Kostochka 2008].

### 3. Main result

The following lemma is the key ingredient in the proof of the main result. The lemma requires a concept which we will call the square color graph. Let  $G$  be a graph. Let  $C_0, \dots, C_{l-1}$  be the color classes of a proper coloring  $C$  of  $G^2$  with  $l$  colors, where  $G^2$  is the graph with  $V(G^2) = V(G)$  and  $E(G^2) = \{xy \mid d(x, y) \leq 2\}$ . The *square color graph* of  $C$ , denoted  $\mathcal{G}$ , is the graph with

$$V(\mathcal{G}) = \{C_0, \dots, C_{l-1}\} \quad \text{and} \quad E(\mathcal{G}) = \{C_i C_j \mid G[C_i \cup C_j] \text{ contains an edge of } G\}.$$

Here  $G[C_i \cup C_j]$  denotes the induced subgraph formed by the vertices in  $C_i \cup C_j$ .

**Lemma 2.** *Let  $G$  be a graph, and let  $C$  be a proper coloring of  $G^2$  with  $l$  colors. If the complement  $\mathcal{G}^c$  of the square color graph of  $C$  has a Hamilton path, then  $\lambda_{2,1}(G) \leq l - 1$ .*

*Proof.* By assumption,  $\mathcal{G}^c$  has a Hamiltonian path  $P = \{p_0, p_1, \dots, p_{l-1}\}$ . Recall that the vertices of  $P$  are color classes partitioning  $G$ . Let  $f : V(G) \rightarrow \mathbb{Z}$  be defined as  $f : v \mapsto i$ , where  $i$  is the unique index such that  $v \in p_i$ . We now check that  $f$  is an  $L(2, 1)$ -labeling of  $G$ . If  $d(x, y) = 2$ , then  $x$  and  $y$  are given two different labels because  $C$  is a coloring of  $G^2$ . If  $d(x, y) = 1$ , then  $x$  and  $y$  are in two distinct color classes  $p_i$  and  $p_j$  such that  $p_i p_j \in E(\mathcal{G})$ . Then  $p_i p_j \notin E(\mathcal{G}^c)$ , so  $i \neq j \pm 1$  because otherwise  $p_i p_j \in E(P)$ . Therefore  $|f(x) - f(y)| \geq 2$ , and  $f$  is an  $L(2, 1)$ -labeling for  $G$ .  $\square$

**Theorem 3.** *Let  $G$  be a graph with  $\Delta = \Delta(G) \geq 1$ , and let  $L$  be an integer with  $L \geq \Delta^2 + 1$ . Then  $\lambda_{2,1}(G) \leq L - 1$  if*

$$|V(G)| \leq (L - \Delta) \left( \left\lfloor \frac{L-1}{2\Delta} \right\rfloor + 1 \right) - 1.$$

Before the proof of [Theorem 3](#), we will discuss two corollaries that have implications for the  $\Delta^2$  conjecture.

**Corollary 4.** *Let  $G$  be a graph of with  $\Delta = \Delta(G) \geq 1$ . Then  $\lambda_{2,1}(G) \leq \Delta^2$  if*

$$|V(G)| \leq \left( \left\lfloor \frac{\Delta}{2} \right\rfloor + 1 \right) (\Delta^2 - \Delta + 1) - 1.$$

*Proof.* Using [Theorem 3](#) with  $L = \Delta^2 + 1$  gives the desired result.  $\square$

[Corollary 4](#) significantly expands the known orders of graphs that satisfy the  $\Delta^2$  conjecture; it does so more dramatically as  $\Delta(G)$  increases. For  $\Delta(G) = 3$ ,  $|V(G)| \leq 13$  suffices as opposed to the previously known  $|V(G)| \leq 10$  [[Griggs and Yeh 1992](#)]. For  $\Delta(G) = 4$ , we have  $|V(G)| \leq 38$  as opposed to  $|V(G)| \leq 17$  [[loc. cit.](#)]. If  $G$  is the Hoffman–Singleton graph, then  $\Delta(G) = 7$ ,  $|V(G)| = 50 = \Delta^2 + 1$ , and, in fact,  $\lambda_{2,1}(G) = 49 = \Delta^2$  [[loc. cit.](#)]. It might seem productive to look

among minor variations of the Hoffman–Singleton graph for counterexamples to the  $\Delta^2$  conjecture, but [Corollary 4](#) suggests otherwise—the conjecture holds if  $\Delta(G) = 7$  and  $|V(G)| \leq 169$ . The bounds on  $|V(G)|$  established in [Corollary 4](#) grow quickly with  $\Delta$ , as they are cubic in  $\Delta$  rather than quadratic as in [[loc. cit.](#)].

For some  $|V(G)|$ , we can also use [Theorem 3](#) to find upper bounds on  $\lambda_{2,1}(G)$  that are stronger than the best known bound of Gonçalves [[loc. cit.](#)]. The bound on  $|V(G)|$  in the following corollary is larger than the bound in [Theorem 3](#).

**Corollary 5.** *Let  $G$  be a graph with  $\Delta = \Delta(G) \geq 3$ . Then  $\lambda_{2,1}(G) < \Delta^2 + \Delta - 2$  if*

$$|V(G)| \leq \left( \left\lfloor \frac{\Delta}{2} \right\rfloor + 1 \right) (\Delta^2 - 2) - 1.$$

*Proof.* Apply [Theorem 3](#) with  $L = \Delta^2 + \Delta - 2$ . This gives

$$|V(G)| \leq \left( \left\lfloor \frac{\Delta}{2} + \frac{1}{2} - \frac{3}{2\Delta} \right\rfloor + 1 \right) (\Delta^2 - 2) - 1.$$

Since we have assumed  $\Delta \geq 3$ , we have  $0 \leq 1/2 - 3/(2\Delta) < 1/2$ , so

$$\left\lfloor \frac{\Delta}{2} + \frac{1}{2} - \frac{3}{2\Delta} \right\rfloor = \left\lfloor \frac{\Delta}{2} \right\rfloor. \quad \square$$

We now proceed to the proof of [Theorem 3](#).

*Proof.* Let  $L$  be as in [Theorem 3](#). We will show that for any integers  $q, r$  with  $q \geq 0, 0 \leq r \leq L - 1$ , and

$$Lq + r \leq M = (L - \Delta) \left( \left\lfloor \frac{L-1}{2\Delta} \right\rfloor + 1 \right) - 1,$$

if  $|V(G)| = Lq + r$  and  $\Delta(G) = \Delta$ , then  $G$  has an  $L(2, 1)$ -labeling with span at most  $L - 1$ . This is sufficient to prove [Theorem 3](#), as for any integer  $n$ , there exist unique integers  $q \geq 0$  and  $r \in \{0, \dots, L - 1\}$  with  $Lq + r = n$ . Suppose  $|V(G)| = Lq + r$ . Recall that  $L \geq \Delta^2 + 1 \geq \Delta(G^2) + 1$ . By the Szemerédi–Hajnal theorem,  $G^2$  has an equitable coloring  $C$  with  $L$  color classes. For convenience, we will use all  $L$  color classes even if several are empty. This means  $L - r$  classes have  $q$  vertices and  $r$  classes have  $q + 1$  vertices. Our goal is to prove that the complement of the square color graph of  $C$ , or  $\mathcal{G}^c$ , has a Hamiltonian path. Note that  $d_{\mathcal{G}}(V) \leq \Delta|V|$  for all  $V \in V(\mathcal{G})$ . Write the degree of  $V$  in  $\mathcal{G}^c$  as  $d_c(V)$ .

If  $q \leq \lfloor (L - 1)/2\Delta \rfloor - 1$ , then

$$\Delta(q + 1) \leq \Delta \left\lfloor \frac{L - 1}{2\Delta} \right\rfloor \leq \left\lfloor \frac{L - 1}{2} \right\rfloor,$$

so that  $\delta(\mathcal{G}^c) \geq L - 1 - \lfloor (L - 1)/2 \rfloor \geq (L - 1)/2$ , and the conditions of [Corollary 1](#) are satisfied. Therefore  $\mathcal{G}^c$  has a Hamiltonian path.

Otherwise,  $q = \lfloor (L - 1)/2\Delta \rfloor$  and

$$r \leq L - 1 - \Delta \left( \left\lfloor \frac{L-1}{2\Delta} \right\rfloor + 1 \right) \leq L - 1$$

because otherwise  $Lq + r > M$ .

Now suppose  $k$  is an integer with  $0 \leq k \leq (L - 2)/2$  as in [Corollary 1](#). If  $d_c(V) \leq k$ , then

$$\frac{L-2}{2} \geq L - 1 - d_G(V) \geq L - 1 - \Delta|V|,$$

so that  $|V| \geq (1/\Delta)(L - 1 - (L - 2)/2) = (L - 1)/2\Delta + 1/2\Delta > q$ . Therefore  $|V| = q + 1$ , so we know there are at most  $r$  vertices with  $d_c(V) \leq k$ . For any such vertex  $V$ ,

$$d_c(V) \geq L - 1 - (q + 1)\Delta = L - 1 - \Delta \left( \left\lfloor \frac{L-1}{2\Delta} \right\rfloor + 1 \right) \geq r \geq 0.$$

Now the conditions of [Corollary 1](#) are satisfied, so  $\mathcal{G}^c$  still has a Hamiltonian path. From [Lemma 2](#),  $\mathcal{G}^c$  having a Hamiltonian path implies that  $\lambda_{2,1}G \leq L - 1$ . As

$$Lq + L - 1 - \Delta \left( \left\lfloor \frac{L-1}{2\Delta} \right\rfloor + 1 \right) = (L - \Delta) \left( \left\lfloor \frac{L-1}{2\Delta} \right\rfloor + 1 \right) - 1 = M,$$

this argument works for any  $|V(G)| \leq M$ . □

**Corollary 6.** *Let  $G$  be a graph of order  $n$  with  $\Delta = \Delta(G) \geq 1$ , and let  $L$  be an integer with  $L \geq \Delta^2 + 1$ . If*

$$n \leq (L - \Delta) \left( \left\lfloor \frac{L-1}{2\Delta} \right\rfloor + 1 \right) - 1,$$

*then there is an  $L(2, 1)$ -labeling of  $G$  with a span at most  $L - 1$  that is equitable. If  $n \geq L$ , the labeling is no-hole.*

*Proof.* The proof follows immediately from the proof of [Theorem 3](#). □

The next corollary concerns algorithms involved in finding these labelings. In general, determining if  $\lambda_{2,1}(G) \leq k$  for positive integers  $k \geq 4$  is NP-complete [[Fiala et al. 2001](#)].

**Corollary 7.** *Let  $G$  be a graph of order  $n$  with  $\Delta = \Delta(G) \geq 1$  and  $L \geq \Delta^2 + 1$ . There is an algorithm with polynomial running time in  $n$  to compute an  $L(2, 1)$ -labeling of  $G$  with span at most  $L - 1$  for all  $n$  and  $L$  such that*

$$n \leq (L - \Delta) \left( \left\lfloor \frac{L-1}{2\Delta} \right\rfloor + 1 \right) - 1.$$

*Proof.* If  $L \geq 2n + 1$ , the appropriate labeling can be obtained by labeling the vertices  $0, 2, \dots, 2n$  in any order [Griggs and Yeh 1992]. This can clearly be done in polynomial time. Otherwise, in [Kierstead et al. 2010] there is shown to be an algorithm, polynomial in  $n$ , to equitably color  $G^2$  with  $L$  colors. Degree sequences satisfying the conditions of Pósa's theorem also satisfy those of Chvátal's theorem [Bondy and Chvátal 1976], and the paper's authors exhibit an algorithm, polynomial in  $p$ , to find Hamilton cycles in graphs of order  $p$  which satisfy the conditions of Chvátal's theorem. From the proofs of Lemma 2 and Corollary 1, we see that to find the labeling, it is enough find a Hamilton cycle in a certain graph, namely  $\mathcal{G}^c$  with a dominating vertex added, of order  $L + 1 \leq 2n + 2$  that satisfies the conditions of Pósa's theorem. From [Bondy and Chvátal 1976], we can do this with an algorithm that is polynomial in  $L + 1$ , which must also be polynomial in  $n$ . These two algorithms in succession yield the desired algorithm.  $\square$

#### 4. Comments on diameter-2 graphs

It was previously known that diameter-2 graphs satisfy the  $\Delta^2$  conjecture, and for other than a few exceptional graphs,  $\Delta^2 - 1$  suffices to label diameter-2 graphs [Griggs and Yeh 1992]. In this section, we knock this bound down by one, showing that  $\Delta^2 - 2$  suffices to label all but a finite handful of diameter-2 graphs.

**Theorem 8** [Griggs and Yeh 1992]. *The  $\Delta^2$  conjecture holds for diameter-2 graphs. In addition,  $\lambda_{2,1} \leq \Delta^2 - 1$  for diameter-2 graphs with  $\Delta \geq 2$  except for  $C_3, C_4$ , and the Moore graphs. For these exceptional graphs,  $\lambda_{2,1} = \Delta^2$ .*

The proof of these facts rely on Brooks' theorem and several results from Griggs and Yeh:

**Theorem 9** (Brooks [Lovász 1975]). *If  $G$  is an odd cycle or a complete graph,  $\chi(G) \leq \Delta + 1$ ; otherwise,  $\chi(G) \leq \Delta$ .*

**Lemma 10** [Griggs and Yeh 1992].  $\lambda_{2,1}(G) \leq |V(G)| + \chi(G) - 2$ .

**Lemma 11** [Griggs and Yeh 1992]. *There exists an injective  $L(2, 1)$ -labeling of a graph  $G$  with span  $|V(G)| - 1$  if and only if the complement of  $G$  has a Hamilton path.*

**Theorem 12** [Griggs and Yeh 1992]. *Let  $C_n$  be a cycle on  $n$  vertices. Then  $\lambda_{2,1}(C_n) = 4$ .*

We now proceed to prove Theorem 8.

*Proof.* If  $\Delta = 2$ , one can verify the theorem readily using Theorem 12. Suppose  $\Delta \geq 3$ . We now split into cases.

In the first case, suppose  $\Delta \geq (|V(G)|)/2$ . Lemma 10 implies

$$\lambda_{2,1}(G) \leq 2\Delta + \chi(G) - 2.$$



If  $G$  is a complete graph, then clearly  $\lambda_{2,1}(G) = 2\Delta(G)$ . As  $\Delta \geq 3$ ,  $G$  is not an odd cycle. Otherwise,  $2\Delta + \chi(G) - 2 \leq 3\Delta - 2$  by Brooks' theorem. Note that in both cases,  $\Delta(G) \geq 3$  implies that  $\lambda_{2,1}(G) \leq \Delta^2 - 2$ .

In the second case, suppose  $\Delta \leq (|V(G)| - 1)/2$ . Then  $\delta(G^c) \geq (|V(G)| - 1)/2$ . Also, we have assumed  $G$  has  $\Delta \geq 3$ , so  $|V(G)| \geq 7$ . By [Corollary 1](#),  $G^c$  has a Hamilton path. By [Lemma 11](#), there is an  $L(2, 1)$ -labeling of  $G$  with span  $|V(G)| - 1$ . As the Moore graphs are the only diameter-2 graphs with  $|V(G)| = \Delta^2 + 1$ , [Theorem 8](#) holds.  $\square$

In fact, we can do better by the following result:

**Theorem 13** [[Erdős et al. 1980](#)]. *Except  $C_4$ , there is no diameter-2 graph of order  $\Delta^2$ .*

This and the proof of [Theorem 8](#) imply the following theorem.

**Theorem 14.** *With the exception of  $C_3, C_4, C_5$ , and the Moore graphs, any diameter-2 graph with  $\Delta(G) \geq 2$  has  $\lambda_{2,1}(G) \leq \Delta^2 - 2$ .*

We also have some comments on a special family of diameter-2 graphs that have large  $\lambda_{2,1}$  number. In order to do this, we must define the points of the *Galois plane*, denoted  $PG_2(n)$ . Let  $F$  be a finite field of order  $n$ . Let  $P = F^3 \setminus \{(0, 0, 0)\}$ . Define an equivalence relation  $\equiv$  on  $P$  by  $(x_1, x_2, x_3) \equiv (y_1, y_2, y_3) \iff (x_1, x_2, x_3) = (cy_1, cy_2, cy_3)$  for some  $c \in F$ . The *points* of  $PG_2(n)$  are the equivalence classes.

**Definition 15.** The *polarity graph* of  $PG_2(n)$ , denoted  $H$ , is the graph with the points of  $PG_2(n)$  as vertices and with two vertices  $(x_1, x_2, x_3)$  and  $(y_1, y_2, y_3)$  adjacent if and only if  $y_1x_1 + y_2x_2 + y_3x_3 = 0$ .

By the properties of  $PG_2(n)$ , we know that the diameter of  $H$  is two,  $\Delta(H) = n + 1$ , and its order is  $n^2 + n + 1 = \Delta^2 - \Delta + 1$  [[Kárteszi 1976](#)]. This implies that  $\lambda_{2,1}(H) \geq \Delta^2 - \Delta$ . In fact, Yeh showed that  $\lambda_{2,1}(H) = \Delta^2 - \Delta$  [[Griggs and Yeh 1992](#)]. This is an infinite family of graphs, as finite fields exist for  $n = p^k$  with  $p$  prime.

However, we can improve this by one. This construction follows that of Erdős, Fajtlowicz and Hoffman [[Erdős et al. 1980](#)]. A vertex  $(x, y, z)$  in  $H$  has degree  $n$  if and only if the norm  $x^2 + y^2 + z^2$  is equal to 0. Suppose  $F$  has characteristic 2 and the order of  $F$  is  $n$ . If  $(a, b, c)$  is in  $H$  then it is adjacent to the point  $(b+c, a+c, a+b)$ , which has norm equal to 0 and is also in  $H$ . In other words, every vertex in  $H$  is adjacent to a vertex of degree  $n$ . We proceed to find the number of points of degree  $n$  in  $H$ . Since  $F$  has characteristic 2,  $f(x) = x^2$  is injective and hence surjective on  $F$ . This means we can choose  $x^2$  and  $y^2$  freely as long as one of them is nonzero, and then  $z^2$  is determined. We must also eliminate proportional pairs, so in total this leaves  $(n^2 - 1)/(n - 1) = n + 1$  vertices of degree  $n$ .

Now we can make an  $(n + 1)$ -regular, diameter-2 graph  $\tilde{H}(n)$  by adding a vertex that is adjacent to all vertices of degree  $n$ . This graph is of order  $n^2 + n + 2 = \Delta^2 - \Delta + 2$ .

**Theorem 16.** *The graph  $\tilde{H}(n)$  has  $\lambda_{2,1}(\tilde{H}) = \Delta^2 - \Delta + 1$ .*

*Proof.* Because  $\tilde{H}$  has diameter 2,  $\lambda_{2,1}(\tilde{H}) \geq \Delta^2 - \Delta + 1$ . As  $\Delta \geq 3$ , we have  $\Delta \leq (\Delta^2 - \Delta + 1)/2 = (|V(H)| - 1)/2$ . By the proof of [Theorem 8](#),  $\lambda_{2,1}(\tilde{H}) \leq |V(G)| - 1 = \Delta^2 - \Delta + 1$ .  $\square$

Since  $\tilde{H}(n)$  exists for all  $n = 2^k$ , this is an infinite family of graphs.

### Acknowledgements

The author would like to thank J. R. Griggs for helpful suggestions and interesting discussions, as well as J.-S. Sereni for his astute observations which helped resolve an error. The author also would like to thank the University of South Carolina for its support.

### References

- [Bondy and Chvátal 1976] J. A. Bondy and V. Chvátal, “A method in graph theory”, *Discrete Math.* **15**:2 (1976), 111–135. [MR 54 #2531](#)
- [Chang and Kuo 1996] G. J. Chang and D. Kuo, “The  $L(2, 1)$ -labeling problem on graphs”, *SIAM J. Discrete Math.* **9**:2 (1996), 309–316. [MR 97b:05132](#)
- [Erdős et al. 1980] P. Erdős, S. Fajtlowicz, and A. J. Hoffman, “Maximum degree in graphs of diameter 2”, *Networks* **10**:1 (1980), 87–90. [MR 81b:05061](#) [Zbl 0427.05042](#)
- [Fiala et al. 2001] J. Fiala, T. Kloks, and J. Kratochvíl, “Fixed-parameter complexity of  $\lambda$ -labelings”, *Discrete Appl. Math.* **113**:1 (2001), 59–72. [MR 2002h:68075](#)
- [Gonçalves 2007] D. Gonçalves, “On the  $L(p, 1)$ -labelling of graphs”, pp. 81–86 in *EUROCOMB '05: combinatorics, graph theory and applications* (Berlin, 2005), vol. 5, Elsevier, Amsterdam, 2007.
- [Griggs and Yeh 1992] J. R. Griggs and R. K. Yeh, “Labelling graphs with a condition at distance 2”, *SIAM J. Discrete Math.* **5**:4 (1992), 586–595. [MR 93h:05141](#)
- [Hajnal and Szemerédi 1970] A. Hajnal and E. Szemerédi, “Proof of a conjecture of P. Erdős”, pp. 601–623 in *Combinatorial theory and its applications, II* (Balatonfüred, 1969), North-Holland, Amsterdam, 1970. [MR 45 #6661](#)
- [Hale 1980] W. Hale, “Frequency assignment: Theory and applications”, pp. 1497–1514 in *Proceedings of the IEEE*, vol. 68, IEEE, 1980.
- [Havet et al. 2012] F. Havet, B. Reed, and J.-S. Sereni, “Griggs and Yeh’s conjecture and  $L(p, 1)$ -labelings”, *SIAM J. Discrete Math.* **26**:1 (2012), 145–168. [MR 2902638](#)
- [Kártészzi 1976] F. Kártészzi, *Introduction to finite geometries*, Texts in Advanced Mathematics **2**, North-Holland, Amsterdam, 1976. [MR 54 #11156](#) [Zbl 0325.50001](#)
- [Kierstead and Kostochka 2008] H. A. Kierstead and A. V. Kostochka, “A short proof of the Hajnal–Szemerédi theorem on equitable colouring”, *Combin. Probab. Comput.* **17**:2 (2008), 265–270. [MR 2009a:05071](#)

- [Kierstead et al. 2010] H. A. Kierstead, A. V. Kostochka, M. Mydlarz, and E. Szemerédi, “A fast algorithm for equitable coloring”, *Combinatorica* **30**:2 (2010), 217–224. [MR 2011h:05097](#)
- [Kronk 1969] H. V. Kronk, “Variations on a theorem of Pósa”, pp. 193–197 in *The many facets of graph theory*, edited by G. Chartrand and S. F. Kapoor, Lecture Notes in Math. **110**, Springer, Berlin, 1969. [MR 41 #99](#)
- [Lovász 1975] L. Lovász, “Three short proofs in graph theory”, *J. Combinatorial Theory Ser. B* **19**:3 (1975), 269–271. [MR 53 #211](#) [Zbl 0322.05142](#)
- [Roberts 1988] F. S. Roberts, 1988. private communication to J. R. Griggs.
- [Sakai 1991] D. Sakai, 1991. private communication to J. R. Griggs.

Received: 2013-02-15

Revised: 2013-04-15

Accepted: 2013-10-02

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

*Department of Mathematics, Rutgers University,  
Hill Center - Busch Campus, 110 Frelinghuysen Road,  
Piscataway, NJ 08854, United States*



# Linear symplectomorphisms as $R$ -Lagrangian subspaces

Chris Hellmann, Brennan Langenbach and Michael VanValkenburgh

(Communicated by Ravi Vakil)

The graph of a real linear symplectomorphism is an  $R$ -Lagrangian subspace of a complex symplectic vector space. The restriction of the complex symplectic form is thus purely imaginary and may be expressed in terms of the generating function of the transformation. We provide explicit formulas; moreover, as an application, we give an explicit general formula for the metaplectic representation of the real symplectic group.

## 1. Introduction

**1.1. Overview.** As part of our symplectic upbringing, our ancestors impressed upon us the Symplectic Creed:

*Everything is a Lagrangian submanifold* [Weinstein 1981].

Obviously false if taken literally, rather than a “creed” it might be called the Maslow–Weinstein hammer, or, in French, *la déformation professionnelle symplectique*, saying that “if all you have is a [symplectic form], everything looks like a [Lagrangian submanifold],” or, in other words, to a symplectic geometer, everything should be expressed in terms of Lagrangian submanifolds. In this paper we consider a vector space endowed with *two* symplectic forms, namely the real and imaginary parts  $\operatorname{Re} \omega^{\mathbb{C}}$  and  $\operatorname{Im} \omega^{\mathbb{C}}$  of a complex symplectic form  $\omega^{\mathbb{C}}$ , and begin with the simple observation that

*Not every Lagrangian submanifold* [with respect to  $\operatorname{Re} \omega^{\mathbb{C}}$ ] *is a Lagrangian submanifold* [with respect to  $\operatorname{Im} \omega^{\mathbb{C}}$ ].

We study its implications for the classification of real linear symplectomorphisms  $\mathcal{H}$ , as the graph of  $\mathcal{H}$  is essentially by definition a Lagrangian subspace with respect to  $\operatorname{Re} \omega^{\mathbb{C}}$ ; we ask, with some abuse of language:

---

*MSC2010:* 37J10, 51A50, 70H15, 81S10.

*Keywords:* complex symplectic linear algebra, linear symplectomorphisms, Lagrangian submanifolds, the metaplectic representation.

**Open problem.** Is every  $2n \times 2n$  skew-symmetric matrix of the form  $\text{Im } \omega^{\mathbb{C}}|_{\text{graph } \mathcal{H}}$  for some  $\mathcal{H}$ ?

We believe that an answer would shed some light on the structure of linear symplectomorphisms. While our primary reason for writing this article is to precisely formulate the above open problem, which we do in [Section 1.2](#), our primary technical result is to rewrite it in terms of generating functions; after all, if one guiding principle is the Symplectic Creed, another is that “symplectic topology is the geometry of generating functions” [[Viterbo 1992](#)]. Or, to go further back, while Sir William Rowan Hamilton first conceived of generating functions (or as he called them, *characteristic functions*) as mathematical tools in his symplectic formulation of optics, he later found, in his symplectic formulation of classical mechanics, that the generating function for a physical system is the least action function, in a sense that we will not make precise [[Abraham and Marsden 1978](#); [Hamilton 1834](#)]; this gives a striking connection with the calculus of variations. Moreover, in Fresnel optics and quantum mechanics, the generating function is used as the phase function of an oscillatory integral operator; the integral operator is said to “quantize” the corresponding symplectomorphism [[Grigis and Sjöstrand 1994](#); [Guillemin and Sternberg 1984](#)]. (Loosely speaking, when differentiating the integral, one finds that the phase function must satisfy the Hamilton–Jacobi equation.) This topic will be touched upon in [Section 3](#). For us, the generating function corresponding to the linear symplectomorphism  $\mathcal{H}$  is the scalar-valued function  $\Phi$  in our main theorem:

**Theorem 1.** *For each  $\mathcal{H} \in \text{Sp}(2n, \mathbb{R})$  there exists a quadratic form  $\Phi : \mathbb{C}^n \times \mathbb{R}^{2n} \rightarrow \mathbb{R}$  such that*

$$\text{graph}_{\mathbb{C}} \mathcal{H} = \left\{ \left( z, -2 \frac{\partial \Phi}{\partial z}(z, \theta) \right) : \frac{\partial \Phi}{\partial \theta}(z, \theta) = 0 \right\},$$

and the restriction of  $\omega^{\mathbb{C}}$  to  $\text{graph}_{\mathbb{C}} \mathcal{H}$  is given by

$$\begin{aligned} & \omega^{\mathbb{C}} \left( \left( z, -2 \frac{\partial \Phi}{\partial z}(z, \theta) \right), \left( w, -2 \frac{\partial \Phi}{\partial z}(w, \eta) \right) \right) \\ &= 2 \sum_{j=1}^n \sum_{\ell=1}^{2n} \frac{\partial^2 \Phi}{\partial z_j \partial \theta_{\ell}} (z_j \eta_{\ell} - w_j \theta_{\ell}) + 2 \sum_{j,m=1}^n \frac{\partial^2 \Phi}{\partial z_j \partial \bar{z}_m} (z_j \bar{w}_m - w_j \bar{z}_m). \quad (1) \end{aligned}$$

Moreover, our construction provides an explicit general formula for  $\Phi$ .

Our notation will be explained in the following subsection, along with the necessary background and a restatement of the open problem. We prove the theorem in [Section 2](#), and in [Section 3](#) we show how our construction seems to adequately answer a question of Folland [[1989](#)] regarding the metaplectic representation. We

conclude with a broad indication of future work. In the [Appendix](#) we give additional linear-algebraic background and some new elementary results relevant to our problem, and also give an additional restatement of our open problem.

**1.2. Background and restatement of the problem.** In a real symplectic vector space there is already a natural complex structure; the model example is  $\mathbb{R}^{2n}$  with the  $2n \times 2n$  matrix  $\mathcal{J} = \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix}$ , where of course  $\mathcal{J}^2 = -I$ . What we mean by “complex symplectic linear algebra” is something else; we instead consider  $\mathbb{C}^{2n}$  with the above matrix  $\mathcal{J}$ , that is, we consider

$$\omega^{\mathbb{C}} = \sum_{j=1}^n d\zeta_j \wedge dz_j \quad \text{on } \mathbb{C}_z^n \times \mathbb{C}_\zeta^n$$

(a nondegenerate alternating bilinear form over  $\mathbb{C}$ ). The basic formalism of complex symplectic linear algebra is not new; indeed, complex symplectic structures naturally appear in the theory of differential equations and have been studied through that lens (see, for example, [Schapira 1981] and [Sjöstrand 1982], or [Everitt and Markus 2004] for another perspective). The point of view of this paper is that elementary linear-algebraic aspects remain unexplored in the complex case and may help us better understand the real case.

A *symplectic vector space* over a field<sup>1</sup>  $K$  is by definition a pair  $(V, \omega)$ , where  $V$  is a finite-dimensional vector space over  $K$  and  $\omega$  is a nondegenerate alternating bilinear form on  $V$ . The basic example is  $\mathbb{R}_x^n \times \mathbb{R}_\xi^n$  with the symplectic form  $\omega = \sum_{j=1}^n d\xi_j \wedge dx_j$ :

$$\omega((x, \xi), (x', \xi')) = \sum_{j=1}^n (\xi_j x'_j - x_j \xi'_j). \tag{2}$$

In fact, this is essentially the only example: for a general symplectic vector space  $(V, \omega)$  over a field  $K$  one can find a basis  $\{e_1, \dots, e_n, f_1, \dots, f_n\}$  for  $V$  such that

$$\omega(e_j, e_k) = 0, \quad \omega(f_j, f_k) = 0, \quad \omega(f_j, e_k) = \delta_{jk} \quad \text{for all } j, k.$$

Such a basis is called a *symplectic basis*, and  $\omega$  is of the form (2) in these coordinates. (In particular, a symplectic vector space is necessarily even-dimensional.) Note that  $\omega$  vanishes on the span of the  $e_j$ , and it vanishes on the span of the  $f_j$ ; such a subspace is called a *Lagrangian subspace*: a maximal subspace on which  $\omega$  vanishes. (A Lagrangian subspace of  $V$  is necessarily of dimension  $n$ .)

The symplectic formalism is fundamental in Hamiltonian mechanics: the symplectic form provides an isomorphism between tangent space and cotangent space,

---

<sup>1</sup>Duistermaat’s book [1996] on Fourier integral operators contains a brief treatment of symplectic vector spaces over a general field.

mapping the Hamiltonian vector field of a function  $f$  to the differential of  $f$ :

$$df = \omega(\cdot, H_f).$$

A linear symplectomorphism  $T$  on  $(V, \omega)$  is a linear isomorphism on  $V$  such that  $T^*\omega = \omega$ , that is,

$$\omega(Tv, Tv') = \omega(v, v') \quad \text{for all } v, v' \in V.$$

This is equivalent to the property that a symplectic basis is mapped to a symplectic basis.

We now let  $(V, \omega)$  be a real symplectic vector space. Then

$$(V \times V, \omega \oplus -\omega)$$

is a real symplectic vector space. We write  $\omega_0 = \omega \oplus -\omega$  so that, by definition,

$$\omega_0((v, w), (v', w')) = \omega(v, v') - \omega(w, w').$$

The following classical result (see [Tao 2012] for a broad perspective) justifies this choice of the symplectic form:

*A map  $\mathcal{H} : V \rightarrow V$  is a linear symplectomorphism if and only if its graph  $\{(v, \mathcal{H}(v)) : v \in V\}$  is a Lagrangian subspace of  $(V \times V, \omega_0)$ .*

For a basic example, let

$$\mathcal{H} : \mathbb{R}_x^n \times \mathbb{R}_\xi^n \rightarrow \mathbb{R}_y^n \times \mathbb{R}_\eta^n, \quad (x, \xi) \mapsto (y, \eta),$$

be a linear symplectomorphism. Then graph  $\mathcal{H}$  is a Lagrangian subspace for

$$\omega^{\mathbb{R}} = \sum_{j=1}^n d\xi_j \wedge dx_j - d\eta_j \wedge dy_j.$$

The point of view of this paper is to consider graph  $\mathcal{H}$  as an  $\mathbb{R}$ -linear subspace of a complex symplectic vector space. After all, with  $z_j = x_j + iy_j$  and  $\zeta_j = \xi_j + i\eta_j$ , we have the complex symplectic form

$$\omega^{\mathbb{C}} = \sum_{j=1}^n d\zeta_j \wedge dz_j \quad \text{on } \mathbb{C}_z^n \times \mathbb{C}_\zeta^n,$$

which induces the two *real* symplectic forms

$$\operatorname{Re} \omega^{\mathbb{C}} = \sum_{j=1}^n d\xi_j \wedge dx_j - d\eta_j \wedge dy_j, \quad \operatorname{Im} \omega^{\mathbb{C}} = \sum_{j=1}^n d\xi_j \wedge dy_j + d\eta_j \wedge dx_j$$

on  $\mathbb{R}_{x,\xi}^{2n} \times \mathbb{R}_{y,\eta}^{2n}$ . We then say that an  $\mathbb{R}$ -linear  $2n$ -dimensional subspace of  $\mathbb{R}_{x,\xi}^{2n} \times \mathbb{R}_{y,\eta}^{2n}$  is an *R-Lagrangian subspace* if it is Lagrangian with respect to  $\operatorname{Re} \omega^{\mathbb{C}}$ , and an



$I$ -Lagrangian subspace if it is Lagrangian with respect to  $\text{Im } \omega^{\mathbb{C}}$ . Thus the graph of  $\mathcal{H} : \mathbb{R}_{x,\xi}^{2n} \rightarrow \mathbb{R}_{y,\eta}^{2n}$  may be considered as an  $R$ -Lagrangian subspace of  $(\mathbb{C}_z^n \times \mathbb{C}_\zeta^n, \omega^{\mathbb{C}})$ .

Writing a symplectic matrix  $\mathcal{H} \in \text{Sp}(2n, \mathbb{R})$  as  $\mathcal{H} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$ , we have

$$\text{graph } \mathcal{H} = \{((x, \xi), (Ax + B\xi, Cx + D\xi)) : (x, \xi) \in \mathbb{R}^{2n}\};$$

or, in terms of  $(z, \zeta)$ , we have

$$\text{graph}_{\mathbb{C}} \mathcal{H} = \{(x + i(Ax + B\xi), \xi + i(Cx + D\xi)) : (x, \xi) \in \mathbb{R}^{2n}\}.$$

Thus

$$\omega^{\mathbb{C}}|_{\text{graph}_{\mathbb{C}} \mathcal{H}} = i \text{Im } \omega^{\mathbb{C}}|_{\text{graph } \mathcal{H}}$$

is given by

$$\begin{aligned} & \omega^{\mathbb{C}}((x + i(Ax + B\xi), \xi + i(Cx + D\xi)), (x' + i(Ax' + B\xi'), \xi' + i(Cx' + D\xi'))) \\ &= i \begin{pmatrix} x^T & \xi^T \end{pmatrix} \begin{pmatrix} C^T - C & -A^T - D \\ A + D^T & B - B^T \end{pmatrix} \begin{pmatrix} x' \\ \xi' \end{pmatrix}. \end{aligned}$$

The symplectic form  $\text{Re } \omega^{\mathbb{C}}$  vanishes, but the symplectic form  $\text{Im } \omega^{\mathbb{C}}$  might *not* vanish; that is,  $\text{graph}_{\mathbb{C}} \mathcal{H}$  is  $R$ -Lagrangian but not necessarily  $I$ -Lagrangian.

We have thus defined a map from the group of symplectic matrices to the space of skew-symmetric matrices

$$\mathfrak{X} : \text{Sp}(2n, \mathbb{R}) \rightarrow \mathfrak{so}(2n, \mathbb{R}), \quad \begin{pmatrix} A & B \\ C & D \end{pmatrix} \mapsto \begin{pmatrix} C^T - C & -A^T - D \\ A + D^T & B - B^T \end{pmatrix}.$$

We can thus restate our open problem:

**Open problem.** Is the map  $\mathfrak{X} : \text{Sp}(2n, \mathbb{R}) \rightarrow \mathfrak{so}(2n, \mathbb{R})$  a surjection?

While we do not solve this problem, the main result of the paper is [Theorem 1](#); we can explicitly construct a generating function  $\Phi$  for  $\mathcal{H}$  and thus give an alternate characterization of  $\omega^{\mathbb{C}}|_{\text{graph}_{\mathbb{C}} \mathcal{H}}$  and hence of  $\mathfrak{X}$ .

## 2. In terms of generating functions: the proof of the theorem

Generating functions (in the sense of symplectic geometry) were discovered by Sir William Rowan Hamilton in his extensive work on optics. In modern language (and in the linear case), light rays are specified by the following data:  $\mathbb{R}_x^2$  is a plane of initial positions perpendicular to the optical axis of the system,  $\xi \in \mathbb{R}^2$  are the initial “directions” (multiplied by the index of refraction),  $\mathbb{R}_y^2$  is a plane of terminal positions, and  $\eta \in \mathbb{R}^2$  are the terminal directions. The spaces  $\mathbb{R}_{x,\xi}^4$  and  $\mathbb{R}_{y,\eta}^4$  are given the standard symplectic structures. Taken piece by piece, the optical system consists of a sequence of reflections and refractions for each light ray, the laws of which were long known; Hamilton’s discovery was that, taken as a whole, the optical

system is determined by a single function, the *generating function*, or, as Hamilton called it, the *characteristic function*, of the optical system. The transformation from initial conditions to terminal conditions is a symplectomorphism expressible in terms of a single scalar-valued function, “by which means optics acquires, as it seems to me, an uniformity and simplicity in which it has been hitherto deficient” [Hamilton 1828, Section IV, Paragraph 20].<sup>2</sup>

The optical framework gives an intuitive reason why, in the symplectic matrix  $\mathcal{H} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$ , the rank of  $B$  plays a special role in characterizing  $\mathcal{H}$  and thus its generating function. Again,  $\mathcal{H}$  maps the initial (position, direction)-pair  $(x, \xi)$  to the terminal (position, direction)-pair

$$\begin{pmatrix} y \\ \eta \end{pmatrix} = \begin{pmatrix} Ax + B\xi \\ Cx + D\xi \end{pmatrix}.$$

The case  $B = 0$  corresponds to perfect focusing: all the rays from a given position  $x$  arrive at the same position  $y$ , resulting in a perfect image. And the case  $\det B \neq 0$  corresponds to *no* such focusing: two rays with initial position  $x$  but different initial directions must arrive at different positions  $y$ . (See [Guillemin and Sternberg 1984] for an exposition of symplectic techniques in optics.)

**2.1. When  $B$  is invertible.** We recall that

$$\text{graph}_{\mathbb{C}} \mathcal{H} = \{(x + i(Ax + B\xi), \xi + i(Cx + D\xi)) : (x, \xi) \in \mathbb{R}^{2n}\},$$

taken over the reals, is an  $R$ -Lagrangian subspace of  $(\mathbb{C}_z^n \times \mathbb{C}_\zeta^n, \omega^{\mathbb{C}})$ , and we note that

$$\pi : \text{graph}_{\mathbb{C}} \mathcal{H} \rightarrow \mathbb{C}^n, \quad (z, \zeta) \mapsto z,$$

is an  $\mathbb{R}$ -linear transformation whose kernel is given by  $(x, \xi) \in \{0\} \times \ker B$ . Thus it is an  $\mathbb{R}$ -linear isomorphism if and only if  $B$  is invertible. In this case, the general theory of symplectic geometry gives the existence of a real  $C^\infty$  function  $\Phi$  defined on  $\text{graph}_{\mathbb{C}} \mathcal{H}$  such that

$$\text{graph}_{\mathbb{C}} \mathcal{H} = \left\{ \left( z, -2 \frac{\partial \Phi}{\partial z}(z) \right) : z \in \mathbb{C}^n \right\}.$$

---

<sup>2</sup>There are different types of generating functions in symplectic geometry, and, as Arnold writes, “[the apparatus of generating functions] is unfortunately noninvariant and it uses, in an essential way, the coordinate structure in phase space” [Arnold 1978, Section 47]. For our purposes, we may take the term “generating function” to broadly mean a scalar-valued function which generates a symplectomorphism (or, more generally, a Lagrangian submanifold) in the same sense that a potential function generates a conservative vector field. Our generating functions are denoted by the symbol  $\Phi$  below.

Hence if  $\det B \neq 0$ , then

$$\begin{aligned} \text{graph}_{\mathbb{C}} \mathcal{H} &= \{(x + i(Ax + B\xi), \xi + i(Cx + D\xi)) : (x, \xi) \in \mathbb{R}^{2n}\} \\ &= \{(z, -2(\partial\Phi/\partial z)(z)) : z \in \mathbb{C}^n\} \\ &= \{(p + iq, B^{-1}(q - Ap) + i(Cp + DB^{-1}(q - Ap))) : p + iq \in \mathbb{C}^n\}, \end{aligned}$$

where we write  $z = p + iq$ , so that

$$\Phi(p, q) = \frac{1}{2}p^T B^{-1}Ap - p^T B^{-1}q + \frac{1}{2}q^T DB^{-1}q. \quad (3)$$

This function appears in [Folland 1989, Equation (4.54)] and in [Guillemin and Sternberg 1984, Section 11]. (Note that  $B^{-1}A$  and  $DB^{-1}$  are symmetric since  $\mathcal{H}$  is symplectic.) Substituting  $p = x$  and  $q = Ax + B\xi$ , we arrive at the following expression, with the obvious abuse of notation:

$$\Phi(x, \xi) = \frac{1}{2}x^T A^T Cx + x^T C^T B\xi + \frac{1}{2}\xi^T B^T D\xi. \quad (4)$$

Or, writing  $\Phi$  with respect to  $z$  and  $\bar{z}$ , we have

$$\begin{aligned} \Phi(z) &= \frac{1}{8}z^T (B^{-1}A + 2iB^{-1} - DB^{-1})z \\ &\quad + \frac{1}{4}\bar{z}^T (B^{-1}A - i(B^T)^{-1} + iB^{-1} + DB^{-1})z \\ &\quad + \frac{1}{8}\bar{z}^T (B^{-1}A - 2iB^{-1} - DB^{-1})\bar{z}. \end{aligned}$$

Thus

$$\left( \frac{\partial^2 \Phi}{\partial z_j \partial \bar{z}_k} \right) = \frac{1}{4}(B^{-1}A - iB^{-1} + i(B^T)^{-1} + DB^{-1}).$$

We can directly compute  $\omega^{\mathbb{C}}$  restricted to  $\text{graph}_{\mathbb{C}} \mathcal{H}$  in terms of  $z$  and  $\bar{z}$ :

$$\begin{aligned} \omega^{\mathbb{C}} \left( \left( z, -2 \frac{\partial \Phi}{\partial z}(z) \right), \left( z', -2 \frac{\partial \Phi}{\partial z}(z') \right) \right) &= 4i \operatorname{Im} \left( \sum_{j,k} z_j \left( \frac{\partial^2 \Phi}{\partial z_j \partial \bar{z}_k} \right) \bar{z}'_k \right) \\ &= 2 \sum_{j,k} \frac{\partial^2 \Phi}{\partial z_j \partial \bar{z}_k} (z_j \bar{z}'_k - z'_j \bar{z}_k). \end{aligned}$$

If we substitute

$$\begin{aligned} z &= x + i(Ax + B\xi), \\ z' &= x' + i(Ax' + B\xi'), \end{aligned}$$

then after a lengthy mechanical calculation we recover the expression

$$\begin{aligned} \omega^{\mathbb{C}} \left( \left( z, -2 \frac{\partial \Phi}{\partial z}(z) \right), \left( z', -2 \frac{\partial \Phi}{\partial z}(z') \right) \right) \\ = 2 \sum_{j,k} \frac{\partial^2 \Phi}{\partial z_j \partial \bar{z}_k} (z_j \bar{z}'_k - z'_j \bar{z}_k) = i \begin{pmatrix} x^T & \xi^T \end{pmatrix} \mathfrak{X}(\mathcal{H}) \begin{pmatrix} x' \\ \xi' \end{pmatrix}. \end{aligned}$$

**2.2. When  $B$  is not invertible.** When  $B$  is *not* invertible, we seek

$$\Phi = \Phi(z, \theta) \in C^\infty(\mathbb{C}^n \times \mathbb{R}^N)$$

such that

$$\text{graph}_{\mathbb{C}} \mathcal{H} = \left\{ \left( z, -2 \frac{\partial \Phi}{\partial z}(z, \theta) \right) : \frac{\partial \Phi}{\partial \theta}(z, \theta) = 0 \right\}. \quad (5)$$

We follow the general method outlined by Guillemin and Sternberg [1977].

Let

$$W = \text{graph}_{\mathbb{C}} \mathcal{H}, \quad X = \{(z, 0); z \in \mathbb{C}^n\}, \quad Y = \{(0, \zeta); \zeta \in \mathbb{C}^n\}.$$

Since  $W$  is an  $R$ -Lagrangian subspace, we know that  $W \cap Y$  and  $PW \subset X$  are orthogonal with respect to  $\text{Re } \omega^{\mathbb{C}}$ , where  $P$  is the projection onto  $X$  along  $Y$ . Indeed,

$$W \cap Y = \{(0, \xi + iD\xi) : \xi \in \ker B\}, \quad PW = \{(x + i(Ax + B\xi), 0) : (x, \xi) \in \mathbb{R}^{2n}\},$$

and we can check directly that, with  $\xi \in \ker B$ ,

$$\omega^{\mathbb{C}}((0, \xi + iD\xi), (x' + i(Ax' + B\xi'), 0)) = i[\xi^T(A + D^T)x' + \xi^T B\xi'].$$

Since  $\text{graph}_{\mathbb{C}} \mathcal{H}$  is not a  $\mathbb{C}$ -linear subspace but an  $\mathbb{R}$ -linear subspace, for now we prefer to write

$$\begin{aligned} W \cap Y &= \{(0, \xi; 0, D\xi) : \xi \in \ker B\}, \\ PW &= \{(x, 0; Ax + B\xi, 0) : (x, \xi) \in \mathbb{R}^{2n}\}. \end{aligned}$$

We note that  $PW \oplus (W \cap Y)$  has real dimension  $2n$ , hence is a Lagrangian subspace of  $(\mathbb{R}^{4n}, \text{Re } \omega^{\mathbb{C}})$ .

We seek to write  $\text{graph } \mathcal{H}$  as the graph of a function from  $PW \oplus (W \cap Y)$  to a complementary Lagrangian subspace; as a first step, we choose a convenient symplectic basis. We let  $\{b_1, \dots, b_k\}$  be an orthonormal basis for  $\ker B$  and extend to an orthonormal basis  $\{b_1, \dots, b_n\}$  for  $\mathbb{R}^n$ , so that

$$\{(0, b_j; 0, Db_j) : j = 1, \dots, k\}$$

is a basis for  $W \cap Y$ , and

$$\{(0, 0; Bb_j, 0) : j = k + 1, \dots, n\} \cup \{(b_j, 0; Ab_j, 0) : j = 1, \dots, n\}$$

is a basis for  $PW$ . We then extend to the following symplectic basis for  $(\mathbb{R}^{4n}, \text{Re } \omega^{\mathbb{C}})$ :

$$\begin{aligned} \{(0, 0; Ab_j, 0) : j = 1, \dots, k\} &\leftrightarrow \{(0, b_j; 0, Db_j) : j = 1, \dots, k\}, \\ \{(0, 0; Bb_j, 0) : j = k + 1, \dots, n\} &\leftrightarrow \{(0, A^T \beta_j; 0, \beta_j) : j = k + 1, \dots, n\}, \quad (6) \\ \{(b_j, 0; Ab_j, 0) : j = 1, \dots, n\} &\leftrightarrow \{(0, -b_j; 0, 0) : j = 1, \dots, n\}, \end{aligned}$$

where the  $\{\beta_j\}_{j=k+1}^n$  satisfy

$$\begin{cases} A^T \beta_j \in (\ker B)^\perp = \text{Im } B^T, \\ b_J \cdot B^T \beta_j = \delta_{Jj} \quad \text{for all } J \in \{k+1, \dots, n\}. \end{cases} \tag{7}$$

One advantage of using the particular symplectic basis (6) is that the vectors on the left are all ‘‘horizontal,’’ and the vectors on the right are all ‘‘vertical’’. (The arrows signify the symplectically dual pairs.)

The following proposition implies the existence of  $\{\beta_j\}_{j=k+1}^n$ .

**Proposition 2.** *The set  $\{Ab_1, \dots, Ab_k, Bb_{k+1}, \dots, Bb_n\}$  is a basis for  $\mathbb{R}^n$ .*

*Proof.* Suppose

$$\sum_{j=1}^k \alpha_j Ab_j + \sum_{j=k+1}^n \alpha_j Bb_j = 0.$$

We take the dot product with  $Db_J$ ,  $J \in \{1, \dots, k\}$ , to get  $\alpha_1 = \dots = \alpha_k = 0$ , and the rest are zero by the linear independence of  $\{Bb_{k+1}, \dots, Bb_n\}$ .  $\square$

Thus for  $J \in \{k+1, \dots, n\}$  we can take  $\beta_J$  to be the unique vector orthogonal to the set

$$\{Ab_1, \dots, Ab_k, Bb_{k+1}, \dots, \widehat{Bb_J}, \dots, Bb_n\}$$

(where the wide hat denotes omission) and satisfying

$$\beta_J \cdot Bb_J = 1.$$

We will now describe graph  $\mathcal{H}$  in terms of the above symplectic coordinate system: we write a general linear combination of the  $4n$  vectors and find necessary and sufficient conditions on the coefficients to make the vector in graph  $\mathcal{H}$ . Explicitly, we write the general vector in  $\mathbb{R}^{4n}$  as

$$\begin{aligned} &\sum_{j=1}^k t'_j(0, 0; Ab_j, 0) + \sum_{j=k+1}^n t''_j(0, 0; Bb_j, 0) + \sum_{j=1}^n t''_{n+j}(b_j, 0; Ab_j, 0) \\ &+ \sum_{j=1}^k \theta'_j(0, b_j; 0, Db_j) + \sum_{j=k+1}^n \theta''_j(0, A^T \beta_j; 0, \beta_j) + \sum_{j=1}^n \theta''_{n+j}(0, -b_j; 0, 0) \end{aligned} \tag{8}$$

(the primes are not necessary but are useful for bookkeeping), and we will describe graph  $\mathcal{H}$  as  $(t', \theta')$  as a function of  $(t'', \theta')$ .

We have the following necessary and sufficient conditions for the vector (8) to be in graph  $\mathcal{H}$ :

$$\begin{aligned} \sum_{j=1}^k t'_j A b_j - \sum_{j=k+1}^n \theta''_j A B^T \beta_j + \sum_{j=k+1}^n \theta''_{n+j} B b_j \\ = - \sum_{j=k+1}^n t''_j B b_j - \sum_{j=k+1}^n \theta''_j C B^T \beta_j + \sum_{j=1}^n \theta''_{n+j} D b_j = \sum_{j=1}^n t''_{n+j} C b_j. \end{aligned}$$

In matrix form, this says:

$$\begin{aligned} & \begin{pmatrix} | & & | & & | & & | & & | \\ A b_1 & \cdots & A b_k & (-A B^T \beta_{k+1}) & \cdots & (-A B^T \beta_n) & B b_1 & \cdots & B b_n \\ | & & | & & | & & | & & | \\ & & & & & & & & \\ 0_{n,k} & & (-C B^T \beta_{k+1}) & \cdots & (-C B^T \beta_n) & D b_1 & \cdots & D b_n & \\ | & & | & & | & & | & & | \end{pmatrix} \begin{pmatrix} t' \\ \theta'' \end{pmatrix} \\ & = \begin{pmatrix} | & & | & & & & & & \\ (-B b_{k+1}) & \cdots & (-B b_n) & & 0_{n,n} & & & & \\ | & & | & & & & & & \\ & & & & & & & & \\ & & & & & & & & \\ & & 0_{n,n-k} & & C b_1 & \cdots & C b_n & & \\ & & & & | & & | & & \\ & & & & | & & | & & \end{pmatrix} \begin{pmatrix} t'' \end{pmatrix}. \quad (9) \end{aligned}$$

We would now like to invert the matrix on the left to get  $(t', \theta'')$  as a function of  $(t'', \theta')$ . Once we do that, we are close to our goal of expressing graph  $\mathcal{H}$  in terms of a generating function  $\Phi$ .

Letting  $\Pi$  denote the orthogonal projection onto  $\ker B$ , we find that the inverse of the matrix on the left side of (9) is

$$\begin{pmatrix} \text{---} & D b_1 & \text{---} & & & & & & \\ & \vdots & & & & & & 0_{k,n} & \\ \text{---} & D b_k & \text{---} & & & & & & \\ \text{---} & D(\Pi C^T B - I) b_{k+1} & \text{---} & \text{---} & B b_{k+1} & \text{---} & & & \\ & \vdots & & & \vdots & & & & \\ \text{---} & D(\Pi C^T B - I) b_n & \text{---} & \text{---} & B b_n & \text{---} & & & \\ \text{---} & (D \Pi A^T - I) C b_1 & \text{---} & \text{---} & A b_1 & \text{---} & & & \\ & \vdots & & & \vdots & & & & \\ \text{---} & (D \Pi A^T - I) C b_n & \text{---} & \text{---} & A b_n & \text{---} & & & \end{pmatrix}.$$

Thus, defining the functions

$$f_i''(t'') = \sum_{j=k+1}^n [Bb_i \cdot Db_j]t_j'' + \sum_{j=1}^n [Bb_i \cdot Cb_j]t_{n+j}'' \quad \text{for } i = k+1, \dots, n,$$

$$f_{n+i}''(t'') = \sum_{j=k+1}^n [Cb_i \cdot Bb_j]t_j'' + \sum_{j=1}^n [Ab_i \cdot Cb_j]t_{n+j}'' \quad \text{for } i = 1, \dots, n,$$

we see that (9) is equivalent to the conditions  $t' = 0$ ,  $\theta'' = f''(t'')$ . Noting that

$$\frac{\partial f_i''}{\partial t_j''} = \frac{\partial f_j''}{\partial t_i''} \quad \text{for all } i, j \in k+1, \dots, n,$$

and defining

$$F(t'') = \frac{1}{2} \sum_{i=k+1}^n \sum_{j=k+1}^n t_i'' [Bb_i \cdot Db_j] t_j''$$

$$+ \sum_{i=k+1}^n \sum_{j=1}^n t_i'' [Bb_i \cdot Cb_j] t_{n+j}'' + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n t_{n+i}'' [Ab_i \cdot Cb_j] t_{n+j}'',$$

we conclude that the vector is in graph  $\mathcal{H}$  if and only if

$$t' = 0, \quad \frac{\partial F}{\partial t''}(t'') = \theta''.$$

We now define

$$\varphi(t', t''; \theta', \theta'') = \theta' \cdot t' + F(t'') + (\theta'' - f''(t''))^2.$$

Then in  $(t', t''; \theta', \theta'')$ -coordinates, graph  $\mathcal{H}$  is given as

$$\left\{ \left( t', t''; \frac{\partial \varphi}{\partial t'}, \frac{\partial \varphi}{\partial t''} \right) : \frac{\partial \varphi}{\partial \theta'} = 0, \frac{\partial \varphi}{\partial \theta''} = 0 \right\}.$$

Or, written in terms of the standard basis, graph  $\mathcal{H}$  is the set of values of

$$\sum_{j=1}^k t_j'(0, 0; Ab_j, 0) + \sum_{j=k+1}^n t_j''(0, 0; Bb_j, 0) + \sum_{j=1}^n t_{n+j}''(b_j, 0; Ab_j, 0)$$

$$+ \sum_{j=1}^k \frac{\partial \varphi}{\partial t_j'}(t, \theta)(0, b_j; 0, Db_j) + \sum_{j=k+1}^n \frac{\partial \varphi}{\partial t_j''}(t, \theta)(0, A^T \beta_j; 0, \beta_j)$$

$$+ \sum_{j=1}^n \frac{\partial \varphi}{\partial t_{n+j}''}(t, \theta)(0, -b_j; 0, 0) \quad (10)$$

subject to the condition that  $\frac{\partial \varphi}{\partial \theta}(t, \theta) = 0$ .

We return to complex coordinates, in the standard basis; for that purpose we write the “horizontal” parts of (10) as

$$z := \sum_{j=1}^k it'_j Ab_j + \sum_{j=k+1}^n it''_j Bb_j + \sum_{j=1}^n t''_{n+j} (I + iA)b_j.$$

That is,

$$\begin{aligned} \operatorname{Re} z &= \sum_{j=1}^n t''_{n+j} b_j, \\ \operatorname{Im} z &= \sum_{j=1}^k t'_j Ab_j + \sum_{j=k+1}^n t''_j Bb_j + \sum_{j=1}^n t''_{n+j} Ab_j. \end{aligned}$$

With the same notation as before, the inverse transformation is given by

$$\begin{aligned} t'_j &= -b_j \cdot \operatorname{Re} z + Db_j \cdot \operatorname{Im} z & \text{for } j \in \{1, \dots, k\}, \\ t''_j &= -A^T \beta_j \cdot \operatorname{Re} z + \beta_j \cdot \operatorname{Im} z & \text{for } j \in \{k+1, \dots, n\}, \\ t''_{n+j} &= b_j \cdot \operatorname{Re} z & \text{for } j \in \{1, \dots, n\}. \end{aligned} \quad (11)$$

We write the “vertical” part of (10) as:

$$\begin{aligned} \operatorname{Re} \zeta &= \sum_{j=1}^k \frac{\partial \varphi}{\partial t'_j}(t, \theta) b_j + \sum_{j=k+1}^n \frac{\partial \varphi}{\partial t''_j}(t, \theta) A^T \beta_j - \sum_{j=1}^n \frac{\partial \varphi}{\partial t''_{n+j}}(t, \theta) b_j, \\ \operatorname{Im} \zeta &= \sum_{j=1}^k \frac{\partial \varphi}{\partial t'_j}(t, \theta) Db_j + \sum_{j=k+1}^n \frac{\partial \varphi}{\partial t''_j}(t, \theta) \beta_j. \end{aligned} \quad (12)$$

Using  $t = t(z)$  to denote the transformation (11), we define

$$\Phi(z, \theta) := \varphi(t(z), \theta),$$

so that (12) says

$$\zeta = -2 \frac{\partial \Phi}{\partial z}(z, \theta).$$

In summary, we now have the following expression for  $\operatorname{graph}_{\mathbb{C}} \mathcal{H}$ :

$$\operatorname{graph}_{\mathbb{C}} \mathcal{H} = \left\{ \left( z, -2 \frac{\partial \Phi}{\partial z}(z, \theta) \right) : \frac{\partial \Phi}{\partial \theta}(z, \theta) = 0 \right\}, \quad (13)$$

where the  $\theta \in \mathbb{R}^{2n}$  are considered as auxiliary parameters, as in (5).

As for  $\omega^{\mathbb{C}}|_{\operatorname{graph}_{\mathbb{C}} \mathcal{H}}$ , we use the expression

$$\frac{\partial \Phi}{\partial z}(z, \theta) = \frac{\partial^2 \Phi}{\partial z \partial \theta} \cdot \theta + \frac{\partial^2 \Phi}{\partial z^2} \cdot z + \frac{\partial^2 \Phi}{\partial z \partial \bar{z}} \cdot \bar{z}$$



to compute

$$\begin{aligned} \omega^{\mathbb{C}} & \left( \left( z, -2 \frac{\partial \Phi}{\partial z}(z, \theta) \right), \left( w, -2 \frac{\partial \Phi}{\partial z}(w, \eta) \right) \right) \\ & = 2z \cdot \frac{\partial \Phi}{\partial z}(w, \eta) - 2w \cdot \frac{\partial \Phi}{\partial z}(z, \theta) \\ & = 2 \sum_{j=1}^n \sum_{\ell=1}^{2n} \frac{\partial^2 \Phi}{\partial z_j \partial \theta_\ell}(z_j \eta_\ell - w_j \theta_\ell) + 2 \sum_{j,m=1}^n \frac{\partial^2 \Phi}{\partial z_j \partial \bar{z}_m}(z_j \bar{w}_m - w_j \bar{z}_m), \end{aligned} \tag{14}$$

where the variables are related by the conditions

$$\frac{\partial \Phi}{\partial \theta}(z, \theta) = 0 \quad \text{and} \quad \frac{\partial \Phi}{\partial \theta}(w, \eta) = 0.$$

Of course, from Section 1, we know that (14) is equal to

$$i \begin{pmatrix} x^T & \xi^T \end{pmatrix} \mathfrak{X}(\mathcal{H}) \begin{pmatrix} x' \\ \xi' \end{pmatrix}, \tag{15}$$

where

$$\begin{aligned} z & = x + i(Ax + B\xi), & w & = x' + i(Ax' + B\xi'), \\ -2 \frac{\partial \Phi}{\partial z}(z, \theta) & = \xi + i(Cx + D\xi), & -2 \frac{\partial \Phi}{\partial z}(w, \eta) & = \xi' + i(Cx' + D\xi'). \end{aligned}$$

This completes the proof of the theorem.

We leave it as an illustrative exercise for the reader to compute  $\Phi$  and its derivatives in the special cases when  $B = 0$  and when  $B$  is invertible (to be compared with the generating function (3) in Section 2.1).

### 3. Application: the metaplectic representation

In the previous section, we showed how to associate to a linear symplectomorphism  $\mathcal{H}$  a (real-valued) generating function  $\Phi$ . For the purposes of Fresnel optics and quantum mechanics one then associates to the generating function  $\Phi$  an oscillatory integral operator

$$\mu(\mathcal{H}) : \mathcal{S}(\mathbb{R}^n) \rightarrow \mathcal{S}'(\mathbb{R}^n), \quad u \mapsto a h^{-3n/2} \iint e^{i\Phi(x+iy,\theta)/h} u(x) dx d\theta. \tag{16}$$

The map  $\mu : \mathcal{H} \rightarrow \mu(\mathcal{H})$  is called the *metaplectic representation* of the symplectic group, and  $\mu(\mathcal{H})$  is said to be the “quantization” of the classical object  $\mathcal{H}$ . As defined, the operator  $\mu(\mathcal{H})$  maps Schwartz functions to tempered distributions, but in fact it extends to a bounded operator on  $L^2(\mathbb{R}^n)$ ; we choose  $a$  so that  $\mu(\mathcal{H})$  is unitary on  $L^2(\mathbb{R}^n)$ , and here  $h > 0$  is a small parameter. These are the operators

of “Fresnel optics,” a relatively simple model theory for optics which accounts for interference and diffraction, describing the propagation of light of wavelength  $h$  [Guillemin and Sternberg 1984]. For the analytic details we refer the reader to a text in semiclassical analysis [Dimassi and Sjöstrand 1999]; here we only show that the standard conditions are indeed satisfied.

The above (real-valued) generating function  $\Phi$ , for an arbitrary  $\mathcal{H} \in \text{Sp}(2n, \mathbb{R})$ , has the property that the 1-forms  $d(\partial\Phi/\partial\theta_1), \dots, d(\partial\Phi/\partial\theta_{2n})$  are linearly independent. Equivalently, with the notation from the previous section, the matrix

$$\begin{pmatrix} \frac{\partial^2\Phi}{\partial(\text{Re } z)\partial\theta'} & \frac{\partial^2\Phi}{\partial(\text{Re } z)\partial\theta''} \\ \frac{\partial^2\Phi}{\partial(\text{Im } z)\partial\theta'} & \frac{\partial^2\Phi}{\partial(\text{Im } z)\partial\theta''} \\ \frac{\partial^2\Phi}{\partial\theta'^2} & \frac{\partial^2\Phi}{\partial\theta'\partial\theta''} \\ \frac{\partial^2\Phi}{\partial\theta''\partial\theta'} & \frac{\partial^2\Phi}{\partial\theta''^2} \end{pmatrix} = \begin{pmatrix} \begin{array}{c} | \\ (-b_1) \end{array} & \cdots & \begin{array}{c} | \\ (-b_k) \end{array} & * \\ \begin{array}{c} | \\ | \\ Db_1 \end{array} & \cdots & \begin{array}{c} | \\ | \\ Db_k \end{array} & * \\ \begin{array}{c} | \\ | \\ 0_{k,k} \\ 0_{(2n-k),k} \end{array} & & \begin{array}{c} | \\ | \\ 0_{k,(2n-k)} \\ 2I_{(2n-k),(2n-k)} \end{array} & \end{pmatrix}$$

has linearly independent columns. (The asterisks denote irrelevant components.) This condition says that quadratic form  $\Phi = \Phi(z, \theta)$  is a *nondegenerate phase function* in the sense of semiclassical analysis [Dimassi and Sjöstrand 1999].

Folland writes: “it seems to be a fact of life that there is no simple description of the operator  $\mu(\mathcal{A})$  that is valid for all  $\mathcal{A} \in \text{Sp}$ ” [Folland 1989, p. 193]; however, we believe that (16), combined with our construction of  $\Phi$  in the proof of Theorem 1, is such a description.

### 4. Conclusion

The open problem and results presented in this paper were motivated by the basic question of the relationship between real and complex symplectic linear algebra. Our approach to this question was to consider a real symplectomorphism as a Lagrangian submanifold with regard to the real part of a complex symplectic form. We believe the resulting problem of the nature of the restriction of the imaginary part of the complex symplectic form to this submanifold (formally,  $\mathfrak{X}(\mathcal{H})$  for a symplectomorphism  $\mathcal{H}$ ) is relevant to the structure of the real symplectic group. (We direct the reader to the Appendix for a list of properties of  $\mathfrak{X}$  and reformulations of our open problem which lend credence to this belief.) Accordingly, we view the main result of this paper as primarily a means for further investigation of the open problem of the image of  $\mathfrak{X}$ . In addition to solving our open problem, we believe that, in line with our generating function formulation, it would be interesting to have a “complexified” theory of the calculus of variations. At present we only have trivial extensions of the real theory.

### Appendix

**A. Elementary properties of  $\mathfrak{X}$ .** We first note some standard facts about symplectic matrices that are used throughout the paper; for further information, see, for example, [Cannas da Silva 2001] or [Folland 1989]. We write

$$\mathcal{J} = \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix}$$

for the matrix representing the standard symplectic form.

**Proposition 3 [Folland 1989].** *Let  $\mathcal{H} \in GL(2n, \mathbb{R})$ . The following are equivalent:*

- (1)  $\mathcal{H} \in \text{Sp}(2n, \mathbb{R})$ .
- (2)  $\mathcal{H}^T \mathcal{J} \mathcal{H} = \mathcal{J}$ .
- (3)  $\mathcal{H}^{-1} = \mathcal{J} \mathcal{H}^T \mathcal{J}^{-1} = \begin{pmatrix} D^T & -B^T \\ -C^T & A^T \end{pmatrix}$ .
- (4)  $\mathcal{H}^T \in \text{Sp}(2n, \mathbb{R})$ .
- (5)  $A^T D - C^T B = I$ ,  $A^T C = C^T A$  and  $B^T D = D^T B$ .
- (6)  $AD^T - BC^T = I$ ,  $AB^T = BA^T$  and  $CD^T = DC^T$ .

While  $\mathfrak{X}$  may be extended to all of  $\mathbb{M}^{2n}(\mathbb{R})$ ,

$$\mathfrak{X} : \mathbb{M}^{2n}(\mathbb{R}) \rightarrow \mathfrak{so}(2n, \mathbb{R}), \quad M \mapsto \mathcal{J}M + M^T \mathcal{J}, \tag{1}$$

for purposes of our open problem the resulting linearity of  $\mathfrak{X}$  does not seem to help when  $\mathfrak{X}$  is restricted to  $\text{Sp}(2n, \mathbb{R})$ .

The following proposition presents some of the most interesting elementary linear algebraic properties of  $\mathfrak{X}$ , which follow immediately from the definition.

**Proposition 4.** *Let  $\mathfrak{X} : \mathbb{M}^{2n}(\mathbb{R}) \rightarrow \mathfrak{so}(2n, \mathbb{R})$  be defined as above. Then:*

- (1)  $\ker(\mathfrak{X}) = \mathfrak{sp}(2n, \mathbb{R})$ , the symplectic Lie algebra.
- (2) For any  $\mathcal{H} \in \text{Sp}(2n, \mathbb{R})$ ,  $\mathfrak{X}(\mathcal{H}) = \mathcal{J}(\mathcal{H} + \mathcal{H}^{-1})$ .  
 In particular, for  $\mathcal{U} \in U(n) = \left\{ \begin{pmatrix} A & -B \\ B & A \end{pmatrix} \in \text{Sp}(2n, \mathbb{R}) \right\}$  we have  $\mathcal{U}^{-1} = \mathcal{U}^T$ , so  $\mathfrak{X}(\mathcal{U}) = \mathcal{J}(\mathcal{U} + \mathcal{U}^T)$ .
- (3) For any  $\mathcal{H} \in \text{Sp}(2n, \mathbb{R})$ ,  $\mathfrak{X}(\mathcal{H})$  is invertible (equivalently,  $\text{Im } \omega^{\mathbb{C}}|_{\text{graph } \mathcal{H}}$  is nondegenerate) if and only if  $-1$  is not a member of the spectrum of  $\mathcal{H}^2$ .
- (4) For  $\mathcal{H}, \mathcal{R} \in \text{Sp}(2n, \mathbb{R})$ , we have  $\mathcal{H}^T \mathfrak{X}(\mathcal{R})\mathcal{H} = \mathfrak{X}(\mathcal{H}^{-1}\mathcal{R}\mathcal{H})$ .

We now take some examples.

**Examples of symplectic matrices and their images under  $\mathfrak{X}$ .**

(1) 
$$\begin{pmatrix} A & 0 \\ 0 & (A^T)^{-1} \end{pmatrix} \mapsto \begin{pmatrix} 0 & -A^T - (A^T)^{-1} \\ A + A^{-1} & 0 \end{pmatrix}.$$

In particular,

$$\begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} \mapsto \begin{pmatrix} 0 & -2I \\ 2I & 0 \end{pmatrix} = 2\mathcal{J}.$$

(2) For  $B = B^T$ ,

$$\begin{pmatrix} I & B \\ 0 & I \end{pmatrix} \mapsto \begin{pmatrix} 0 & -2I \\ 2I & 0 \end{pmatrix}.$$

(3) For  $C = C^T$ ,

$$\begin{pmatrix} I & 0 \\ C & I \end{pmatrix} \mapsto \begin{pmatrix} 0 & -2I \\ 2I & 0 \end{pmatrix}.$$

(4) 
$$\mathcal{J} = \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix} \mapsto \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

(5) For  $t \in \mathbb{R}$ ,

$$\begin{pmatrix} (\cos t)I & (-\sin t)I \\ (\sin t)I & (\cos t)I \end{pmatrix} \mapsto \begin{pmatrix} 0 & -2(\cos t)I \\ 2(\cos t)I & 0 \end{pmatrix}.$$

(6) For any  $\mathcal{H} \in \text{Sp}(2n, \mathbb{R})$ , we have  $\mathfrak{X}(\mathcal{H}) = \mathfrak{X}(\mathcal{H}^{-1})$ .

Thus in Examples (2) and (3),  $\text{graph}_{\mathbb{C}} \mathcal{H}$  is an  $RI$ -subspace ( $R$ -Lagrangian and  $I$ -symplectic). And in Example (4),  $\text{graph}_{\mathbb{C}} \mathcal{H}$  is a  $C$ -Lagrangian subspace ( $R$ -Lagrangian and  $I$ -Lagrangian).

The exact nature of the image of  $\mathfrak{X}$  is an open question. The following is a partial result

**Proposition 5.** *For each  $k \in \{0, 1, \dots, n\}$ , there exists  $\mathcal{H}_k \in \text{Sp}(2n, \mathbb{R})$  such that  $\text{rank}(\mathfrak{X}(\mathcal{H}_k)) = 2k$ . Moreover, for any  $\mathcal{H} \in \text{Sp}(2n, \mathbb{R})$ , we have  $\ker \mathfrak{X}(\mathcal{H}) = \ker(\mathcal{H}^2 + I)$ .*

*Proof.* We fix  $k \in \{0, 1, \dots, n\}$  and write

$$(x, \xi) = (x', x'', \xi', \xi''), \quad x', \xi' \in \mathbb{R}^k, \quad x'', \xi'' \in \mathbb{R}^{n-k}.$$

Let

$$\mathcal{H}_k(x', x'', \xi', \xi'') = (x', -\xi'', \xi', x'').$$

The matrix representation of  $\mathcal{H}_k$  is

$$\begin{pmatrix} I_k & 0_k & & \\ & 0_{n-k} & -I_{n-k} & \\ 0_k & I_k & & \\ & I_{n-k} & 0_{n-k} & \end{pmatrix} \in \mathrm{Sp}(2n, \mathbb{R}).$$

Then

$$\mathfrak{X}(\mathcal{H}_k) = \begin{pmatrix} & & -2I_k & \\ & & & 0_{n-k} \\ 2I_k & & & \\ & 0_{n-k} & & \end{pmatrix},$$

so that

$$\mathrm{rank}(\mathfrak{X}(\mathcal{H}_k)) = 2k.$$

The last statement of the proposition follows from (1).  $\square$

**B. Restatement of the problem.** It is sometimes convenient to work with the extension of  $\mathfrak{X}$  to all of  $\mathbb{M}^{2n}(\mathbb{R})$ :

$$\mathfrak{X}(M) = \mathcal{J}M + M^T \mathcal{J}.$$

Then  $\mathfrak{X} : \mathbb{M}(2n, \mathbb{R}) \rightarrow \mathfrak{so}(2n, \mathbb{R})$  is a linear epimorphism with kernel  $\mathfrak{sp}(2n, \mathbb{R})$ , the symplectic Lie algebra (see, for example, [Folland 1989, Proposition 4.2]). Thus the map  $\mathfrak{X}|_{\mathbb{M}(2n, \mathbb{R})}$  is surjective if and only if every element of the quotient space  $\mathbb{M}(2n, \mathbb{R})/\mathfrak{sp}(2n, \mathbb{R})$  contains a symplectic matrix. So our question is:

**Question.** Can every  $M \in \mathbb{M}(2n, \mathbb{R})$  be written as  $M = \mathcal{H} + \mathcal{A}$  for some  $\mathcal{H} \in \mathrm{Sp}(2n, \mathbb{R})$  and some  $\mathcal{A} \in \mathfrak{sp}(2n, \mathbb{R})$ ?

**Proposition 6.** *Every  $M \in \mathbb{M}(2n, \mathbb{R})/\mathfrak{sp}(2n, \mathbb{R})$  has a unique representative of the form*

$$\begin{pmatrix} 0 & \mathcal{S}_2 \\ \mathcal{S}_3 & \mathcal{D} \end{pmatrix},$$

where  $\mathcal{S}_2$  and  $\mathcal{S}_3$  are skew-symmetric.

*Proof.* Existence: let

$$M = \begin{pmatrix} M_1 & M_2 \\ M_3 & M_4 \end{pmatrix} \in \mathbb{M}(2n, \mathbb{R}).$$

Since  $\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \in \mathfrak{sp}(2n, \mathbb{R})$  if and only if  $\delta = -\alpha^T$ ,  $\beta = \beta^T$ ,  $\gamma = \gamma^T$ , we may replace  $M$  by

$$\tilde{M} = M - \begin{pmatrix} M_1 & \frac{1}{2}(M_2 + M_2^T) \\ \frac{1}{2}(M_3 + M_3^T) & -M_1^T \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{2}(M_2 - M_2^T) \\ \frac{1}{2}(M_3 - M_3^T) & M_4 + M_1^T \end{pmatrix}.$$

Uniqueness: suppose

$$\begin{pmatrix} 0 & \mathcal{S}_2 \\ \mathcal{S}_3 & \mathcal{D} \end{pmatrix} = \begin{pmatrix} 0 & \mathcal{S}'_2 \\ \mathcal{S}'_3 & \mathcal{D}' \end{pmatrix} \in \mathbb{M}(2n, \mathbb{R}) / \mathfrak{sp}(2n, \mathbb{R}),$$

with the  $\mathcal{S}_j$  and  $\mathcal{S}'_j$  skew-symmetric. Thus

$$\begin{pmatrix} 0 & \mathcal{S}_2 - \mathcal{S}'_2 \\ \mathcal{S}_3 - \mathcal{S}'_3 & \mathcal{D} - \mathcal{D}' \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ \gamma & -\alpha^T \end{pmatrix} \in \mathfrak{sp}(2n, \mathbb{R}).$$

This shows that  $\mathcal{S}_j - \mathcal{S}'_j$  is symmetric and skew-symmetric, hence zero, and it is clear that  $\mathcal{D} = \mathcal{D}'$ .  $\square$

Thinking geometrically, we are to find the projection of  $\mathrm{Sp}(2n, \mathbb{R})$  onto

$$\left\{ \begin{pmatrix} 0 & \mathcal{S}_2 \\ \mathcal{S}_3 & \mathcal{D} \end{pmatrix} : \mathcal{S}_2, \mathcal{S}_3 \text{ skew-symmetric} \right\}$$

along  $\mathfrak{sp}(2n, \mathbb{R})$ . That is, let  $\mathcal{H} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \in \mathrm{Sp}(2n, \mathbb{R})$ . Then

$$\pi(\mathcal{H}) = \begin{pmatrix} 0 & \frac{1}{2}(B - B^T) \\ \frac{1}{2}(C - C^T) & A^T + D \end{pmatrix}.$$

Is every

$$\begin{pmatrix} 0 & \mathcal{S}_2 \\ \mathcal{S}_3 & \mathcal{D} \end{pmatrix}$$

of this form?

For a possible simplification, the map

$$\mathcal{Y} : \mathrm{Sp}(2n, \mathbb{R}) \rightarrow \mathfrak{so}(2n, \mathbb{R}), \quad \mathcal{H} \mapsto \mathfrak{X}(-\mathcal{J}\mathcal{H}) = \mathcal{H} - \mathcal{H}^T,$$

has the same image as  $\mathfrak{X} : \mathrm{Sp}(2n, \mathbb{R}) \rightarrow \mathfrak{so}(2n, \mathbb{R})$  and may be easier to understand.

### Acknowledgements

Funding for this project was generously provided by Grinnell College, as part of its summer directed research program. VanValkenburgh is currently a faculty member at California State University, Sacramento, but was a visiting faculty member at Grinnell College while the first draft of the paper was written; he thanks both institutions for their support. The authors are also grateful to the anonymous referee for a thoughtful reading of the paper and suggestions for its improvement.

### References

[Abraham and Marsden 1978] R. Abraham and J. E. Marsden, *Foundations of mechanics*, 2nd ed., Benjamin/Cummings, Reading, MA, 1978. [MR 81e:58025](#) [Zbl 0393.70001](#)

- [Arnold 1978] V. I. Arnold, *Mathematical methods of classical mechanics*, Graduate Texts in Mathematics **60**, Springer, New York, 1978. [MR 57 #14033b](#) [Zbl 0386.70001](#)
- [Cannas da Silva 2001] A. Cannas da Silva, *Lectures on symplectic geometry*, Lecture Notes in Mathematics **1764**, Springer, Berlin, 2001. [MR 2002i:53105](#) [Zbl 1016.53001](#)
- [Dimassi and Sjöstrand 1999] M. Dimassi and J. Sjöstrand, *Spectral asymptotics in the semi-classical limit*, London Mathematical Society Lecture Note Series **268**, Cambridge University Press, 1999. [MR 2001b:35237](#) [Zbl 0926.35002](#)
- [Duistermaat 1996] J. J. Duistermaat, *Fourier integral operators*, Progress in Mathematics **130**, Birkhäuser, Boston, MA, 1996. [MR 96m:58245](#) [Zbl 0841.35137](#)
- [Everitt and Markus 2004] W. N. Everitt and L. Markus, *Infinite dimensional complex symplectic spaces*, Mem. Amer. Math. Soc. **810**, Amer. Math. Soc., Providence, RI, 2004. [MR 2005d:46048](#) [Zbl 1054.46019](#)
- [Folland 1989] G. B. Folland, *Harmonic analysis in phase space*, Annals of Mathematics Studies **122**, Princeton University Press, 1989. [MR 92k:22017](#) [Zbl 0682.43001](#)
- [Grigis and Sjöstrand 1994] A. Grigis and J. Sjöstrand, *Microlocal analysis for differential operators: An introduction*, London Mathematical Society Lecture Note Series **196**, Cambridge University Press, 1994. [MR 95d:35009](#) [Zbl 0804.35001](#)
- [Guillemin and Sternberg 1977] V. Guillemin and S. Sternberg, *Geometric asymptotics*, Mathematical Surveys **14**, Amer. Math. Soc., Providence, RI, 1977. [MR 58 #24404](#) [Zbl 0364.53011](#)
- [Guillemin and Sternberg 1984] V. Guillemin and S. Sternberg, *Symplectic techniques in physics*, Cambridge University Press, 1984. [MR 86f:58054](#) [Zbl 0576.58012](#)
- [Hamilton 1828] Sir W. R. Hamilton, “Theory of systems of rays: Part first”, *Transactions of the Royal Irish Academy* **15** (1828), 69–174.
- [Hamilton 1834] Sir W. R. Hamilton, “On a general method in dynamics”, *Philosophical Transactions of the Royal Society* **124** (1834), 247–308.
- [Schapira 1981] P. Schapira, “Conditions de positivité dans une variété symplectique complexe: Application à l’étude des microfonctions”, *Ann. Sci. École Norm. Sup.* (4) **14**:1 (1981), 121–139. [MR 82i:58067](#) [Zbl 0473.58022](#)
- [Sjöstrand 1982] J. Sjöstrand, *Singularités analytiques microlocales*, Astérisque **95**, Société Mathématique de France, Paris, 1982. [MR 84m:58151](#) [Zbl 0524.35007](#)
- [Tao 2012] T. Tao, “The closed graph theorem in various categories”, blog post, 20 November 2012, <http://terrytao.wordpress.com/2012/11/20/the-closed-graph-theorem-in-various-categories/>.
- [Viterbo 1992] C. Viterbo, “Symplectic topology as the geometry of generating functions”, *Math. Ann.* **292**:1 (1992), 685–710. [MR 93b:58058](#) [Zbl 0735.58019](#)
- [Weinstein 1981] A. Weinstein, “Symplectic geometry”, *Bull. Amer. Math. Soc. (N.S.)* **5**:1 (1981), 1–13. [MR 83a:58044](#) [Zbl 0465.58013](#)

Received: 2013-09-20

Revised: 2014-08-24

Accepted: 2014-10-31

[hellmann@grinnell.edu](mailto:hellmann@grinnell.edu)*Department of Mathematics and Statistics, Grinnell College, Grinnell, IA 50112, United States*[langenba@grinnell.edu](mailto:langenba@grinnell.edu)*Department of Mathematics and Statistics, Grinnell College, Grinnell, IA 50112, United States*[mjv@csus.edu](mailto:mjv@csus.edu)*Department of Mathematics and Statistics, California State University, Sacramento, CA 95819, United States*





# Maximization of the size of monic orthogonal polynomials on the unit circle corresponding to the measures in the Steklov class

John Hoffman, McKinley Meyer, Mariya Sardarli and Alex Sherman

(Communicated by Sever S. Dragomir)

We investigate the size of monic, orthogonal polynomials defined on the unit circle corresponding to a finite positive measure. We find an upper bound for the  $L_\infty$  growth of these polynomials. Then we show, by example, that this upper bound can be achieved. Throughout these proofs, we use a method developed by Rahmanov to compute the polynomials in question. Finally, we find an explicit formula for a subsequence of the Verblunsky coefficients of the polynomials.

## 1. Introduction

Let  $V = C(\mathbb{T}; \mathbb{C})$ , where  $\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$ . We define an inner product on  $V$  by

$$\langle f, g \rangle_{d\mu} = \int_{\mathbb{T}} f(z) \overline{g(z)} d\mu,$$

where  $d\mu$  is of the form

$$d\mu = p(\theta) d\theta + \sum_{j=1}^n m_j \delta(\theta - \theta_j),$$

where  $p(\theta)$  is a continuous function,  $\delta$  is the Dirac delta function, and the  $m_j$  are masses placed at the  $\theta_j$  satisfying  $m_j \geq 0$ . We will confine our analysis to measures in the restricted Steklov class of order  $\delta$ , denoted  $S_\delta$ , which consists of measures with the properties

$$p(\theta) > \delta, \quad m_j \geq 0, \quad \langle 1, 1 \rangle_{d\mu} = 2. \quad (1-1)$$

*MSC2010:* 42C05.

*Keywords:* OPUC, classical analysis, approximation theory, orthogonal polynomials on the unit circle.

At the time of writing, Hoffman, Sherman and Meyer were undergraduates at the University of Wisconsin–Madison, and Sardarli was an undergraduate at Princeton University. Hoffman is the corresponding author.

This inner product gives a norm  $\| \cdot \|$  defined as

$$\|f(z)\|_{d\mu} = \sqrt{\langle f, f \rangle_{d\mu}}.$$

Given a measure  $d\mu \in \mathcal{S}_\delta$ , there exists a unique set of monic orthogonal polynomials  $\{\phi_n(z; d\mu)\}$  [Simon 2005]. We will adopt the convention that  $\phi_n(z; d\mu)$  is the polynomial of degree  $n$  in this set. When there is no ambiguity about what measure is being used, we will simply write these polynomials as  $\phi_n(z)$ . Corresponding to the set  $\{\phi_n(z; d\mu)\}$  is the set  $\{\varphi_n(z)\}$  of orthonormal polynomials, defined by

$$\varphi_n(z) = \frac{\phi_n(z)}{\|\phi_n(z)\|}.$$

These polynomials form an orthonormal set. Uniqueness of this set follows from uniqueness of  $\{\phi_n(z)\}$ .<sup>1</sup>

A conjecture of Steklov stated that the sequence

$$M_{n,\delta} = \sup_{d\mu \in \mathcal{S}_\delta} \max_{z \in \mathbb{T}} |\varphi_n(z; d\mu)|$$

is bounded in  $n$ . This was disproven by Rahmanov [1979]. In particular, Rahmanov proved the existence of a probability measure  $d\eta = \phi(\theta) d\theta + \sum_{j=1}^n m_j \delta(\theta - \theta_j)$  such that

$$|\varphi_n(1, d\eta)| \geq C \ln(n) + B$$

for some constants  $B, C$ . The hard part in making such estimates is that, in general, there are few tools available to compute  $\varphi_n(z)$  other than the Gram–Schmidt process. To establish his result, Rahmanov found a formula for computing the  $\phi_n(z; d\eta)$ , where  $d\eta = d\mu + \sum_{k=0}^n m_j \delta(\theta - \theta_j)$  in terms of  $\phi_n(z; d\mu)$ , meaning that  $d\eta$  differs from  $d\mu$  only in its masses. This formula uses the Christoffel–Darboux kernel

$$K_n(z, \xi) = \sum_{j=0}^n \overline{\varphi_j(z)} \varphi_j(\xi). \tag{1-2}$$

The roots of the Christoffel–Darboux kernel are those  $\xi_j$  satisfying

$$K_n(\xi_j, \xi_i) \begin{cases} = 0 & \text{for } i \neq j, \\ \neq 0 & \text{for } i = j. \end{cases} \tag{1-3}$$

Rahmanov’s formula, in light of these definitions, is

$$\phi_n(z; d\eta) = \phi_n(z; d\mu) - \sum_{j=1}^n \frac{m_j \phi_n(\xi_j; d\mu)}{1 + m_j K_{n-1}(\xi_j, \xi_j)} K_{n-1}(z, \xi_j). \tag{1-4}$$

---

<sup>1</sup>Originally, the last condition given for the Steklov class is stated as  $\langle 1, 1 \rangle_{d\mu} = 1$ . This is a minor modification though, because  $\varphi_n(z)/\sqrt{2}$  is in the Steklov class  $\mathcal{S}_{\delta/\sqrt{2}}$ , given the original definition.

We now outline our results. In [Theorem 2.3](#), we use Rahmanov’s method for computing  $\phi_n(\theta)$  for the measure  $d\eta = d\theta/2\pi + \sum_{j=1}^{\lfloor n/4 \rfloor} m_j \delta(\theta - \theta_j)$ , with  $\theta_j = (2\pi j - \pi)/n$ ,  $m_j = 4/n$  and show that the corresponding polynomials  $\phi_n(z; d\mu)$  are uniformly bounded above by  $8/(5\pi^2) \log(\lfloor n/4 \rfloor - 1) + C$ , where  $C$  is a constant. Our next main result is [Theorem 4.1](#), where we construct a family of measures  $d\mu_n$  such that  $\phi_n(1; d\mu_n) > 1/\pi \log n + c$ ,  $c$  a bounded constant. Finally, in [Theorem 5.1](#), we show that, given the measure  $d\mu = d\theta/2\pi + \sum_{j=1}^n m_j \delta(\theta - \theta_j)$ , with  $\theta_j = 2\pi j/n - \theta_0$ , the subsequence  $\{\alpha_{nk-1}\}_{k=1}^\infty$  of the Verblunsky coefficients  $\alpha_j(d\mu)$  satisfies

$$\alpha_{nk-1} = e^{ink\theta_0} \sum_{j=1}^n \frac{m_j}{1 + m_j n k}.$$

The reader may notice that all of our results are stated in terms of  $\phi_n(z)$ , while Steklov’s conjecture is stated in terms of  $\varphi_n(z)$ . We will end this introduction with a lemma, proven by Rahmanov [[1979](#)], that shows why bounds on  $\phi_n(z)$  imply bounds on  $\varphi_n(z)$ , and thus why it is sufficient to evaluate  $\{\phi_n(z)\}$ .

**Lemma 1.1.** *Given a measure  $d\mu \in S_\delta$ ,  $\delta > 0$ , there exists a constant  $C$  such that*

$$\frac{1}{C} \|\phi_n(z, d\mu)\| \leq \|\varphi_n(z, d\mu)\| \leq C \|\phi_n(z, d\mu)\|$$

for all  $n \geq 0$ .

*Proof.* Since

$$|\varphi_n(z)| = \frac{|\phi_n(z)|}{\|\phi_n(z)\|_{d\mu}},$$

it suffices to find constant upper and lower bounds on  $\|\phi_n(z)\|_{d\mu}$ .

To find an upper bound, we first claim that  $\phi_n(z)$  minimizes the integral

$$\int_{\mathbb{T}} |P(z)|^2 d\mu,$$

where  $P(z)$  is any monic polynomial of degree  $n$ . Let  $q(z)$  be an arbitrary polynomial of degree less than  $n$ . Then, since  $\phi_n(z)$  is orthogonal to all polynomials of degree less than  $n$  under the measure  $d\mu$  and the inner product of a polynomial with itself is nonnegative, we have

$$\begin{aligned} &\langle \phi_n(z) + q(z), \phi_n(z) + q(z) \rangle_{d\mu} \\ &= \langle \phi_n(z), \phi_n(z) \rangle_{d\mu} + \langle \phi_n(z), q(z) \rangle_{d\mu} + \langle q(z), \phi_n(z) \rangle_{d\mu} + \langle q(z), q(z) \rangle_{d\mu} \\ &= \langle \phi_n(z), \phi_n(z) \rangle_{d\mu} + \langle q(z), q(z) \rangle_{d\mu} \\ &\geq \langle \phi_n(z), \phi_n(z) \rangle_{d\mu}. \end{aligned}$$

Hence  $\phi_n(z)$  minimizes the integral  $\int_{\mathbb{T}} |P(z)|^2 d\mu$ . In particular this gives us

$$\|\phi_n(z)\|_{d\mu}^2 = \int_0^{2\pi} |\phi_n(z)|^2 d\mu \leq \int_0^{2\pi} |z^n|^2 d\mu = \int_0^{2\pi} 1 d\mu = 2. \tag{1-5}$$

We can derive a lower bound using the fact that  $d\mu \in S_\delta$  and, in particular, that  $d\mu$  satisfies (1-1), which gives

$$\begin{aligned} \|\phi_n(z)\|_{d\mu}^2 &= \frac{1}{2\pi} \int_0^{2\pi} |\phi_n(e^{i\theta})|^2 p(\theta) d\theta + \sum_{j=1}^l m_j |\phi_n(e^{i\theta_j})|^2 \\ &\geq \frac{\delta}{2\pi} \int_0^{2\pi} |\phi_n(e^{i\theta})|^2 d\theta. \end{aligned}$$

Let the coefficient of the  $z^k$  term of  $\phi_n(z)$  be  $a_k$ . In particular,  $a_n = 1$ . Using that the integral of  $e^{ik\theta}$  over the unit circle is 0 for a nonzero integer  $k$ , we get

$$\int_0^{2\pi} |\phi_n(e^{i\theta})|^2 d\theta = \int_0^{2\pi} \sum_{k=0}^n a_k^2 d\theta \geq \int_0^{2\pi} a_n^2 d\theta = 2\pi.$$

Hence,  $\|\phi_n(z)\|_{d\mu}^2 \geq \delta$ .

Combining this with the upper bound on  $\|\phi_n(z)\|_{d\mu}^2$  from (1-5) gives

$$\delta \leq \|\phi_n(z)\|_{d\mu}^2 \leq 2,$$

and as a result

$$\frac{|\phi_n(z)|}{\sqrt{2}} \leq |\varphi_n(z)| \leq \frac{|\phi_n(z)|}{\sqrt{\delta}}. \quad \square$$

## 2. Review of Rahmanov’s result

We begin by reviewing Rahmanov’s argument [1979] to show that the growth of the monic polynomials under Rahmanov’s scheme is bounded below by  $c \log n$ , where  $c$  is a constant. Before doing that, though, we need to prove two lemmas that simplify our future calculations.

We now characterize the roots of the Christoffel–Darboux kernel (which we defined on page 2) for a certain measure:

**Lemma 2.1.** *For  $d\mu = d\theta/2\pi$ , the roots of the Christoffel–Darboux kernel are the  $n$ -th roots of unity times a constant of modulus one.*

*Proof.* Recall the definition of the Christoffel–Darboux kernel and its roots from (1-2) and (1-3). For our  $d\mu$ ,  $\varphi_j = z^j$ , so assuming  $\xi_j$  is of modulus one for all  $j$ ,

$$\begin{aligned} K_{n-1}(\xi_i, \xi_j) &= \sum_{j=0}^{n-1} \xi_i^j / \xi_j^j \\ &= \frac{\xi_i^n / \xi_j^n - 1}{\xi_i / \xi_j - 1} \quad (\text{since this is a geometric series}) \\ &= \frac{\xi_i^n - \xi_j^n}{(\xi_i - \xi_j)\xi_j^{n-1}}. \end{aligned} \tag{2-1}$$

Therefore, by (2-1),  $\xi_j = e^{2i\pi j/n} \xi_0$ , where  $1 \leq j \leq n$  and  $\xi_0$  is an arbitrary point on the unit circle, and so  $\xi_j$  is an  $n$ -th root of unity times a constant.  $\square$

**Lemma 2.2.** *We need only assess  $\phi_n(z; d\mu)$  at  $z = 1$ , since*

$$\sup_{\mu \in S_\delta} \max_{z \in \mathbb{T}} |\phi_n(z; d\mu)| = \sup_{\mu \in S_\delta} |\phi_n(1; d\mu)|.$$

*Proof.* Let

$$d\mu_1 = p(\theta) d\theta + \sum_{j=1}^m m_j \delta(\theta - \theta_j),$$

where  $d\mu_1 \in S_\delta$ . Then  $d\mu_2 \in S_\delta$ , where

$$d\mu_2 = p(\theta - \theta^*) d\theta + \sum_{j=1}^m m_j \delta(\theta - \theta^* - \theta_j), \quad \theta^* \in [0, 2\pi).$$

In particular,  $\phi_n(e^{i\theta}; d\mu_1) = \phi_n(e^{i(\theta+\theta^*)}; d\mu_2)$ .

Hence,

$$\begin{aligned} \max_{z \in \mathbb{T}} |\phi_n(z; d\mu_1)| &= \max_{z \in \mathbb{T}} |\phi_n(z; d\mu_2)|, \\ \sup_{\mu \in S_\delta} \max_{z \in \mathbb{T}} |\phi_n(z; d\mu)| &= \sup_{\mu \in S_\delta} |\phi_n(1; d\mu)|. \end{aligned}$$

Henceforth, we will only look at  $\phi_n(z; d\mu)$  evaluated at  $z = 1$ .  $\square$

**Theorem 2.3.** *Under a finite measure  $d\eta = d\mu + \sum_{j=1}^{\lfloor n/4 \rfloor} m_j \delta(\theta - \theta_j)$ , the monic polynomials are not uniformly bounded from above; specifically, there exists a  $d\eta$  such that the maximums are greater than or equal to  $8/(5\pi^2) \log(\lfloor n/4 \rfloor - 1)$ .*

**Remark 2.4.** This is Rahmanov’s result [1979], whose proof we have included for the reader’s convenience.

*Proof.* First, we will deal generally with some  $d\eta$  without specifying the added masses.

In light of [Lemma 2.1](#), let  $\theta_j = (2\pi j - \pi)/n$  for  $1 \leq j \leq \lfloor n/4 \rfloor$ . Then, using Rahmanov’s formula in [\(1-4\)](#), we have

$$\phi_n(z; d\eta) = \phi_n(z; d\mu) - \sum_{j=1}^{\lfloor n/4 \rfloor} \frac{m_j \phi_n(\xi_j; d\mu)}{1 + m_j K_{n-1}(\xi_j, \xi_j)} K_{n-1}(z, \xi_j), \tag{2-2}$$

which, by noting that  $K_{n-1}(\xi_j, \xi_j) = \sum_{j=1}^{n-1} 1 = n$  and substituting  $z$  and  $\xi_j$  into [\(2-2\)](#), becomes

$$\begin{aligned} \phi_n(z; d\eta) &= z^n - \sum_{j=1}^{\lfloor n/4 \rfloor} \frac{m_j \xi_j^n}{1 + m_j n} \frac{-z^n - 1}{ze^{-i\theta_j} - 1} \\ &= z^n + \sum_{j=1}^{\lfloor n/4 \rfloor} \frac{m_j}{1 + m_j n} \frac{z^n + 1}{1 - ze^{-i\theta_j}}. \end{aligned}$$

Now we want to find a lower bound for  $|\phi_n|$ :

$$\max_{z \in \mathbb{T}} |\phi_n(z; d\eta)| \geq \max_{z \in \mathbb{T}} \left| \operatorname{Im} \left( z^n + \sum_{j=1}^{\lfloor n/4 \rfloor} \frac{m_j}{1 + m_j n} \frac{z^n + 1}{1 - ze^{-i\theta_j}} \right) \right|.$$

We take  $z = 1$ , in line with [Lemma 2.2](#), to get

$$\begin{aligned} \max_{z \in \mathbb{T}} |\phi_n(z; d\eta)| &\geq \left| \operatorname{Im} \left( 1 + 2 \sum_{j=1}^{\lfloor n/4 \rfloor} \frac{m_j}{1 + m_j n} \frac{1}{1 - e^{-i\theta_j}} \right) \right| \\ &= \left| \operatorname{Im} \left( 1 - 2 \sum_{j=1}^{\lfloor n/4 \rfloor} \frac{m_j}{1 + m_j n} \frac{e^{i\theta_j} - 1}{|1 - e^{-i\theta_j}|^2} \right) \right|. \end{aligned}$$

Note that  $0 < \theta_j < \pi/2$ ,  $|1 - e^{-i\theta_j}| \leq \theta_j$ , and

$$\operatorname{Im}(e^{i\theta_j} - 1) = \sin \theta_j \geq \frac{2\theta_j}{\pi} \quad \text{for } \theta \in \left(0, \frac{\pi}{2}\right),$$

which gives

$$\left| \operatorname{Im} \left( 1 - 2 \sum_{j=1}^{\lfloor n/4 \rfloor} \frac{m_j}{1 + m_j n} \frac{e^{i\theta_j} - 1}{|1 - e^{-i\theta_j}|^2} \right) \right| \geq 2 \sum_{j=1}^{\lfloor n/4 \rfloor} \frac{m_j}{1 + m_j n} \frac{2}{\pi \theta_j}. \tag{2-3}$$

Now, we specify the masses of  $d\eta$  to get a precise bound. Let  $m_j = 4/n$  for all  $j$ . This simplifies [\(2-3\)](#) to

$$\max_{z \in \mathbb{T}} |\phi_n(z; d\eta)| \geq \frac{16}{5\pi n} \sum_{j=1}^{\lfloor n/4 \rfloor} \frac{1}{(2j - 1)\frac{\pi}{n}} \geq \frac{8}{5\pi^2} \sum_{j=1}^{\lfloor n/4 \rfloor} \frac{1}{j}.$$

Note that  $\log a = \int_1^a \frac{1}{x} dx \geq \sum_{j=1}^{a-1} \frac{1}{j}$  since  $1/x$  is decreasing.

Therefore,

$$\max_{z \in \mathbb{T}} |\phi_n(z; d\eta)| \geq \frac{8}{5\pi^2} \log\left(\left\lfloor \frac{n}{4} \right\rfloor - 1\right). \tag{2-4}$$

Since  $4/(5\pi^2) \log(\lfloor n/4 \rfloor - 1)$  is strictly increasing in  $n$ ,  $\max_{z \in \mathbb{T}} |\phi_n(z, d\eta)|$  is not uniformly bounded from above. □

### 3. Finding a general upper bound

In this section, we find a general upper bound for the growth of the monic orthogonal polynomials under a  $d\eta$  which differs from  $d\theta/2\pi$  only in the discrete portion. We prove the following theorem by making a sequence of overestimates of  $|\phi_n(1, d\eta)|$ .

**Theorem 3.1.** *Let  $d\eta$  be a measure such that*

$$d\eta = \frac{1}{2\pi} d\theta + \sum_{j=1}^n m_j \delta(\theta - \theta_j),$$

where  $m_j \geq 0$ ,  $\theta_j = 2\pi j/n + \theta_0$  for  $1 \leq j \leq n$ , and  $\theta_0 \in [0, 2\pi)$ .

Then

$$|\phi_n(1, d\eta)| \leq \frac{1}{\pi} \log n + C,$$

where  $C$  is a constant uniformly bounded in  $n$ .

**Remark 3.2.** Note the generalized offset  $\theta_0$  in the theorem. In [Section 2](#), we used the specific offset of  $\theta_0 = -\pi/n$ , but here, we find a general upper bound under any offset.

We prove the theorem using two lemmas. The first, [Lemma 3.3](#), finds an overestimate for  $|\phi_n(1, d\eta)|$  using Rahmanov’s formula [\[1979\]](#). The second, [Lemma 3.4](#), makes another overestimate using Taylor series.

**Lemma 3.3.** *Let  $d\eta$  be a measure such that*

$$d\eta = \frac{1}{2\pi} d\theta + \sum_{j=1}^n m_j \delta(\theta - \theta_j),$$

where  $m_j \geq 0$  and  $\theta_j = 2\pi j/n + \theta_0$ ,  $1 \leq j \leq n$ .

Then

$$|\phi_n(1)| = \begin{cases} \frac{|1 - e^{in\theta_0}|}{2} \left| \sum_{j=1}^n \frac{m_j}{1 + m_j n} \frac{\sin \theta_j}{1 - \cos \theta_j} \right| + c_n & \text{if } \theta_0 \neq 0, \\ c_n & \text{if } \theta_0 = 0, \end{cases}$$

where  $|c_n| < 2$  for all  $n$ .

*Proof.* We first consider the case where  $\theta_0 = 0$ . If  $\theta_0 = 0$ , then  $K_{n-1}(1, e^{i\theta_j}) = 0$  for  $1 \leq j < n$  and  $K_{n-1}(1, e^{i\theta_n}) = n$ . From Rahmanov’s formula (1-4),

$$|\phi_n(1)| = \left| 1 - \frac{m_n}{1 + m_n n} \right| < 1.$$

Therefore,  $\phi_n(1)$  is not increasing in  $n$  for  $\theta_0 = 0$ . Henceforth, we restrict ourselves to working with  $\theta_0 \neq 0$ .

From Rahmanov’s formula in (1-4) and applying Lemma 2.1, we derive

$$\phi_n(1) = 1 - \sum_{j=1}^n \frac{m_j}{1 + m_j n} e^{in\theta_j} \frac{e^{in\theta_j} - 1}{e^{i\theta_j} - 1}.$$

Then, using algebra, we find that

$$\begin{aligned} \phi_n(1) &= 1 + \sum_{j=1}^n \frac{m_j}{1 + m_j n} \frac{1 - e^{in\theta_j}}{1 - e^{i\theta_j}} \\ &= 1 + \frac{1 - e^{in\theta_0}}{2} \sum_{j=1}^n \frac{m_j}{1 + m_j n} \left( 1 - i \frac{\sin \theta_j}{1 - \cos \theta_j} \right), \end{aligned} \tag{3-1}$$

which implies by the triangle inequality that

$$\left| \phi_n(1) - \frac{1 - e^{in\theta_0}}{2} i \sum_{j=1}^n \frac{m_j}{1 + m_j n} \frac{\sin \theta_j}{1 - \cos \theta_j} \right| \leq 1 + \frac{|1 - e^{in\theta_0}|}{2} \left| \sum_{j=1}^n \frac{m_j}{1 + m_j n} \right|.$$

Note that

$$|1 - e^{in\theta_0}| \leq 2 \quad \text{and} \quad 0 \leq \frac{m_j}{1 + m_j n} \leq \frac{1}{n}, \quad \text{so} \quad \frac{|1 - e^{in\theta_0}|}{2} \left| \sum_{j=1}^n \frac{m_j}{1 + m_j n} \right| \leq 1.$$

Hence,

$$\left| \phi_n(1) - \frac{1 - e^{in\theta_0}}{2} i \sum_{j=1}^n \frac{m_j}{1 + m_j n} \frac{\sin \theta_j}{1 - \cos \theta_j} \right| \leq 2. \quad \square$$

Thus, it is sufficient to consider the growth of

$$\frac{|1 - e^{in\theta_0}|}{2} \left| \sum_{j=1}^n \frac{m_j}{1 + m_j n} \frac{\sin \theta_j}{1 - \cos \theta_j} \right|. \tag{3-2}$$

We want to eliminate the magnitude around the sum in (3-2).

Since  $m_j \geq 0$ , and  $\sin \theta_j / (1 - \cos \theta_j)$  is positive on  $(0, \pi)$  and negative on  $(\pi, 2\pi)$ , we have



$$\left| \sum_{j=1}^n \frac{m_j}{1+m_j n} \frac{\sin \theta_j}{1-\cos \theta_j} \right| \leq \max \left\{ \left| \sum_{\theta_j \in (0, \pi)} \frac{m_j}{1+m_j n} \frac{\sin \theta_j}{1-\cos \theta_j} \right|, \left| \sum_{\theta_j \in (\pi, 2\pi)} \frac{m_j}{1+m_j n} \frac{\sin \theta_j}{1-\cos \theta_j} \right| \right\}.$$

Now, if we alter  $d\eta$  so that the masses are instead located at  $\theta_j^* = -2\pi j/n - \theta_0$ , essentially reflecting the discrete portion of the measure over the real axis, we see that (3-2) does not change.

Hence, since we are looking to find an upper bound of (3-1) that is independent of  $m_j$  and  $\theta_0$ , we can assume without loss of generality that we are only looking at  $\theta_j \in (0, \pi)$ , and thus take

$$\frac{|1 - e^{in\theta_0}|}{2} \left| \sum_{j=1}^n \frac{m_j}{1+m_j n} \frac{\sin \theta_j}{1-\cos \theta_j} \right| \leq \frac{|1 - e^{in\theta_0}|}{2} \sum_{\theta_j \in (0, \pi)} \frac{m_j}{1+m_j n} \frac{\sin \theta_j}{1-\cos \theta_j}.$$

Since replacing  $\theta_0$  with  $\theta_0 + 2\pi/n$  and then shifting the index of the  $m_j$  does not affect the value of (3-2), assume  $\theta_0 \in (-2\pi/n, 0)$ . Having made these simplifications, we can now move on to the main lemma, which finds an upper bound as described in the theorem.

**Lemma 3.4.**

$$\frac{|1 - e^{in\theta_0}|}{2} \sum_{\theta_j \in (0, \pi)} \frac{\sin \theta_j}{1-\cos \theta_j} \leq n \left( 1 + \frac{1}{\pi} + \frac{1}{\pi} \log \left\lfloor \frac{n}{2} \right\rfloor \right),$$

where  $\theta_j = 2\pi j/n + \theta_0$ ,  $\theta_0 \in (-2\pi/n, 0)$ .

*Proof.* We separate the first term from the sum, since that term contributes the most to the magnitude. Recall that  $\theta_1 = 2\pi/n + \theta_0$ . Thus,

$$\frac{|1 - e^{in\theta_0}|}{2} \sum_{\theta_j \in (0, \pi)} \frac{\sin \theta_j}{1-\cos \theta_j} \leq \frac{|1 - e^{in\theta_0}|}{2} \frac{\sin \theta_1}{1-\cos \theta_1} + \sum_{\theta_j \in (2\pi/n, \pi)} \frac{\sin \theta_j}{1-\cos \theta_j}$$

since  $|1 - e^{in\theta_0}| \leq 2$ . We now bound these two terms of the sum separately.

We claim that

$$\frac{|1 - e^{in\theta_0}|}{2} \frac{\sin \theta_1}{1-\cos \theta_1} \leq n$$

for  $\theta_0 \in (-2\pi/n, 0)$ . Recall  $\theta_1 = \theta_0 + 2\pi/n$ , so hence  $|1 - e^{in\theta_0}| = |1 - e^{in\theta_1}|$ . Denote  $\theta_1$  by  $t$ , where  $t \in (0, 2\pi/n)$ . We do the calculation

$$\begin{aligned} \frac{|1 - e^{int}|}{2} \frac{\sin t}{1 - \cos t} &= \frac{\sqrt{(1 - \cos nt)^2 + (\sin nt)^2}}{2} \frac{\sin t}{2(\sin \frac{t}{2})^2} = \frac{\sqrt{2 - 2 \cos nt}}{2} \frac{\sin t}{2(\sin \frac{t}{2})^2}. \end{aligned}$$

Because  $\sin(nt/2)$  is nonnegative for  $t \in (0, 2\pi/n)$ , we have

$$\sin \frac{nt}{2} \frac{\sin t}{2(\sin \frac{t}{2})^2} = \frac{\sin \frac{nt}{2} (2 \sin \frac{t}{2} \cos \frac{t}{2})}{2(\sin \frac{t}{2})^2} = \frac{\sin \frac{nt}{2} \cos \frac{t}{2}}{\sin \frac{t}{2}},$$

$\sin(nt/2)/\sin(t/2)$  is nonnegative for  $t \in (0, 2\pi/n)$  and  $\cos t/2$  is bounded above by 1. Hence, the expression is bounded above by  $\sin(nt/2)/\sin(t/2)$ . It remains to show this is bounded above by  $n$ .

This is clearly true for  $n = 1$ . Let  $n > 1$ . Recall that  $nt \in (0, 2\pi)$ . Consider an  $(n + 1)$ -gon inscribed in a unit circle in which  $n$  of the sides of the polygon form a central angle of  $t$ . The last side of the polygon forms a central angle of  $nt$  (this angle may be reflexive.) Recall that the length of a chord of a unit circle which forms a central angle of  $t$  is  $2 \sin(t/2)$ . Similarly the length of the chord which forms a central angle of  $nt$  is  $2 \sin(nt/2)$ . As the polygon is not degenerate, the sum of the lengths of the  $n$  equal side lengths is greater than the length of the remaining side length. Namely  $2n \sin(t/2) \geq 2 \sin(nt/2)$  as desired.

We now handle the second term. To bound

$$\sum_{\theta_j \in (2\pi/n, \pi)} \frac{\sin \theta_j}{1 - \cos \theta_j},$$

note that  $\sin \theta_j / (1 - \cos \theta_j)$  is decreasing on  $(0, 2\pi)$ , so

$$\sum_{\theta_j \in (2\pi/n, \pi)} \frac{\sin \theta_j}{1 - \cos \theta_j} \leq \sum_{j=1}^{\lfloor n/2 \rfloor} \frac{\sin(\frac{2\pi}{n} j)}{1 - \cos(\frac{2\pi}{n} j)}.$$

Recall from the Taylor expansion that we can approximate  $\sin x / (1 - \cos x)$  near 0 by  $2/x$ . In fact, since

$$\lim_{x \rightarrow 0} \frac{\sin x}{1 - \cos x} - \frac{2}{x} = 0$$

and for  $x \in (0, \pi]$ , we have

$$\frac{d}{dx} \left( \frac{\sin x}{1 - \cos x} - \frac{2}{x} \right) < 0,$$

we arrive at the inequality

$$\frac{\sin x}{1 - \cos x} \leq \frac{2}{x} \quad \text{for } x \in (0, \pi].$$

Therefore,

$$\sum_{j=1}^{\lfloor n/2 \rfloor} \frac{\sin(\frac{2\pi}{n} j)}{1 - \cos(\frac{2\pi}{n} j)} \leq \sum_{j=1}^{\lfloor n/2 \rfloor} \frac{2}{\frac{2\pi}{n} j} = \frac{n}{\pi} \sum_{j=1}^{\lfloor n/2 \rfloor} \frac{1}{j}.$$

Recall that  $\log a = \int_1^a 1/x dx \geq \sum_{j=2}^a 1/j$  since  $1/x$  is decreasing, so that

$$\frac{n}{\pi} \sum_{j=1}^{\lfloor n/2 \rfloor} \frac{1}{j} = \frac{n}{\pi} + \frac{n}{\pi} \sum_{j=2}^{\lfloor n/2 \rfloor} \frac{1}{j} \leq \frac{n}{\pi} + \frac{n}{\pi} \log \left\lfloor \frac{n}{2} \right\rfloor,$$

and thus

$$\frac{|1 - e^{in\theta_0}|}{2} \sum_{\theta_j \in (0, \pi)} \frac{\sin \theta_j}{1 - \cos \theta_j} \leq n \left( 1 + \frac{1}{\pi} + \frac{1}{\pi} \log \left\lfloor \frac{n}{2} \right\rfloor \right). \quad \square$$

Returning to the statement of the theorem,

$$\begin{aligned} |\phi_n(1)| &\leq 2 + \frac{|1 - e^{in\theta_0}|}{2} \left| \sum_{j=1}^n \frac{m_j}{1 + m_j n} \frac{\sin \theta_j}{1 - \cos \theta_j} \right| \\ &\leq 2 + \frac{|1 - e^{in\theta_0}|}{2} \left| \sum_{j=1}^n \frac{1}{n} \frac{\sin \theta_j}{1 - \cos \theta_j} \right| \\ &\leq 3 + \frac{1}{\pi} + \frac{1}{\pi} \log \left\lfloor \frac{n}{2} \right\rfloor. \end{aligned}$$

Since  $\log \lfloor n/2 \rfloor$  is equal to  $\log n$  plus some uniformly bounded term, we can conclude that

$$|\phi_n(1)| \leq \frac{1}{\pi} \log n + C,$$

where  $C$  is constant in  $n$ , which completes the proof of the theorem.

**Remark 3.5.** Note that here we used that

$$\frac{m_j}{1 + m_j n} \leq \frac{1}{n}.$$

If we were to use the Rahmanov scheme of distributing masses and set all  $m_j = 1/n$  then

$$\frac{m_j}{1 + m_j n} = \frac{1}{2n},$$

and the monic orthogonal polynomials given by the Rahmanov type of measure would have growth bounded from above by  $1/2\pi \log n + b$ , where  $b$  is a bounded constant.

#### 4. Proving the lower bound

In this section, we construct a measure that achieves the upper bound of  $1/\pi \log n$  plus a bounded term, as described in [Theorem 3.1](#). We accomplish this primarily by applying the technique of Lagrange multipliers to find an optimal measure.

**Theorem 4.1.** *For all  $n \in \mathbb{N}$ , there exists a measure*

$$d\eta = \frac{1}{2\pi} d\theta + \sum_{j=1}^n m_j \delta(\theta - \theta_j),$$

where  $m_j \geq 0$  and  $\sum_{j=1}^n m_j = 1$  such that

$$|\phi_n(1, d\eta)| \geq \frac{1}{\pi} \log n + c,$$

where  $c$  is a bounded constant.

We will prove this theorem as a sequence of lemmas.

The first lemma, [Lemma 4.2](#), finds a lower bound for the expression from [Lemma 3.3](#) which is simpler to manipulate. In the second lemma, [Lemma 4.4](#), we apply the technique of Lagrange multipliers to that lower bound to find a critical “point”, in our case a scheme of  $m_j$ s. Finally, in the third lemma, [Lemma 4.6](#), we insert those derived  $m_j$  into the approximation and find that we achieve the growth stated in the theorem.

Set  $\theta_j = (2\pi j - \pi)/n$ . Inserting those  $\theta_j$  into [\(3-1\)](#), we have that

$$|\phi_n(1)| = \frac{|1 - e^{-\pi i}|}{2} \left| \sum_{j=1}^n \frac{m_j}{1 + m_j n} \frac{\sin \theta_j}{1 - \cos \theta_j} \right| + c_n = \left| \sum_{j=1}^n \frac{m_j}{1 + m_j n} \frac{\sin \theta_j}{1 - \cos \theta_j} \right| + c_n$$

for some constant  $|c_n| < 2$ . We know that  $\sin \theta_j / (1 - \cos \theta_j)$  is positive for  $\theta_j \in (0, \pi)$  and negative for  $\theta_j \in (\pi, 2\pi)$ . Thus, in order to maximize  $|\phi_n(1)|$ , we set  $m_j = 0$  for all  $j$  such that  $\theta_j \in (\pi, 2\pi)$ , which prevents destructive interference from the other side of the circle.

Under this setting, we can say that

$$\left| \sum_{j=1}^n \frac{m_j}{1 + m_j n} \frac{\sin \theta_j}{1 - \cos \theta_j} \right| = \sum_{j=1}^{\lfloor n/2 \rfloor} \frac{m_j}{1 + m_j n} \frac{\sin \theta_j}{1 - \cos \theta_j}.$$

We next bound this equation from below with a simpler expression.

**Lemma 4.2.** *For  $\theta_j = (2\pi j - \pi)/n$  and  $m_j \geq 0$ ,*

$$\sum_{j=1}^{\lfloor n/2 \rfloor} \frac{m_j}{1 + m_j n} \frac{\sin \theta_j}{1 - \cos \theta_j} \geq \frac{1}{\pi} \sum_{j=1}^{\lfloor n/2 \rfloor} \frac{nm_j}{1 + m_j n} \frac{1}{j} + d,$$

where  $d$  is some constant.

**Remark 4.3.** It may appear contradictory that we first find a lower bound when we want the  $n$ -th degree monic polynomial to be as large as possible. However, this lower bound is easier to manipulate, and we show in the subsequent lemmas that it actually achieves the growth stated in the theorem.

*Proof.* We prove this lemma using two approximations. We first approximate  $\sin \theta_j / (1 - \cos \theta_j)$  by  $2/\theta_j$ , from the Taylor series; we then approximate  $2/\theta_j$  by  $1/(\pi j)$ .

First, we show that  $2/\theta_j$  is a good approximation of  $\sin \theta_j / (1 - \cos \theta_j)$ . Let

$$M = \max_{\theta_j \in [0, \pi]} \left| \frac{\sin \theta_j}{1 - \cos \theta_j} - \frac{2}{\theta_j} \right|.$$

This maximum,  $M$ , is achieved because

$$\left| \frac{\sin \theta_j}{1 - \cos \theta_j} - \frac{2}{\theta_j} \right|$$

is continuous in an open neighborhood containing  $[0, \pi]$ . Thus,  $2/\theta_j$  is a good approximation and we can bound the following difference by a constant:

$$\begin{aligned} & \left| \sum_{j=1}^{\lfloor n/2 \rfloor} \frac{m_j}{1 + m_j n} \frac{\sin \theta_j}{1 - \cos \theta_j} - \sum_{j=1}^{\lfloor n/2 \rfloor} \frac{m_j}{1 + m_j n} \frac{2}{\theta_j} \right| \\ & \leq \sum_{j=1}^{\lfloor n/2 \rfloor} \frac{m_j}{1 + m_j n} \left| \frac{\sin \theta_j}{1 - \cos \theta_j} - \frac{2}{\theta_j} \right| \leq \sum_{j=1}^{\lfloor n/2 \rfloor} \frac{1}{n} M \leq M. \end{aligned}$$

Having established this, we can now replace  $2/\theta_j$  with  $2n/((2j - 1)\pi)$  and attain the inequality

$$\sum_{j=1}^{\lfloor n/2 \rfloor} \frac{m_j}{1 + m_j n} \frac{2}{\theta_j} \geq \frac{n}{\pi} \sum_{j=1}^{\lfloor n/2 \rfloor} \frac{m_j}{1 + m_j n} \frac{1}{j}.$$

Combining this and the previous inequality proves the lemma. □

Now that we have the simplified lower bound

$$\frac{1}{\pi} \sum_{j=1}^{\lfloor n/2 \rfloor} \frac{nm_j}{1 + m_j n} \frac{1}{j},$$

we can apply the method of Lagrange multipliers to it in order to construct the  $m_j$  that prove the theorem.

**Lemma 4.4.** *Let  $n \in \mathbb{N}$ . Consider any  $l \in \mathbb{N}$  with  $l \leq n$ . Under the constraints  $m_j \geq 0$  and  $\sum_{j=1}^l m_j = 1$ , we achieve the maximum of*

$$\sum_{j=1}^l \frac{m_j n}{(1 + n m_j) j}$$

by setting

$$m_j = \frac{m_1}{\sqrt{j}} + \frac{1}{n} \left( \frac{1}{\sqrt{j}} - 1 \right)$$

for all  $1 \leq j \leq l$ , where

$$m_1 = \left( 1 + \frac{l}{n} \right) \frac{1}{\sum_{j=1}^{\lfloor n/2 \rfloor} 1/\sqrt{j}} - \frac{1}{n}.$$

*Proof.* Set up  $f$ , the function to be maximized, and the constraint  $g$ , where  $\mathbf{m}$  is the vector listing all  $m_j$ :

$$\begin{aligned} f(\mathbf{m}) &= \sum_{j=1}^l \frac{m_j n}{(1 + n m_j) j}, \\ g(\mathbf{m}) &= \sum_{j=1}^l m_j - 1 = 0. \end{aligned} \tag{4-1}$$

If  $f$  under the constraint  $g$  has a local extremum at  $\mathbf{m}'$  and  $\mathbf{m}'$  is not on the boundary, for example  $m'_j > 0$  for all  $1 \leq j \leq l$ , then there is a  $\lambda \in \mathbb{R}$  such that

$$\nabla f(\mathbf{m}') = \lambda \nabla g(\mathbf{m}'). \tag{4-2}$$

To simplify the following expressions, denote

$$\sum_{j=1}^l \frac{1}{\sqrt{j}} = \alpha(l).$$

Calculations yield that, for all  $j$ ,

$$\frac{n}{(1 + n m'_1)^2} = \frac{n}{j(1 + n m'_j)^2},$$

which, substituting  $m'_1$  and  $m'_j$ , gives  $m'_j$  in terms of  $m'_1$ , that is,

$$m'_j = \frac{m'_1}{\sqrt{j}} + \frac{1}{n} \left( \frac{1}{\sqrt{j}} - 1 \right). \tag{4-3}$$

Inserting that expression for  $m'_j$  into  $g(\mathbf{m}') = \sum_{j=1}^l m'_j - 1 = 0$  yields

$$m'_1 = \left( 1 + \frac{l}{n} \right) \frac{1}{\alpha(l)} - \frac{1}{n}.$$

**Remark 4.5.** For all  $1 \leq j \leq l$ , we have  $m'_j > 0$  since  $\alpha(l) < \int_0^l 1/\sqrt{l} dl = 2\sqrt{l}$ . Thus, we satisfy the condition that  $m'_j \geq 0$ .

To insure that, in the computation for  $\mathbf{m}'$ , the Lagrange multipliers method did in fact give us the  $\mathbf{m}$  that maximized  $f(\mathbf{m})$  under the constraint  $m_j \geq 0$  and  $\sum_{j=1}^l m_j = 1$ , we must check the boundary. We next provide a quick proof that the maximum is not achieved at the boundary.

Consider the Lagrangian  $L(\mathbf{m}) = f(\mathbf{m}) - \lambda g(\mathbf{m})$  defined on  $(-1/n, \infty)^l$ , where  $\lambda$  is the constant in (4-2). Note that  $\mathbf{m}'$  is a critical point of  $L$  since  $\mathbf{m}'$  satisfies  $\nabla L = \nabla f - \lambda \nabla g = 0$ . It suffices to show that  $L$  is concave on  $(-1/n, \infty)^l$ .

We first calculate the entries of the Lagrangian  $L$ :

$$\frac{\partial^2 L}{\partial m_j^2} = -\frac{2n^2}{j} \frac{1}{(1 + nm_j)^3} < 0,$$

$$\frac{\partial^2 L}{\partial m_j \partial m_k} = 0 \quad \text{for } j \neq k.$$

The Hessian of  $L$  is then negative definite and hence  $L$  is concave on  $(-1/n, \infty)^l$ . Therefore,  $\mathbf{m}'$  as computed in (4-3) is a point where  $L$  achieves a global maximum on the open neighborhood  $(-1/n, \infty)^l$ . In particular,  $L(\mathbf{m}')$  is the maximum of  $L$  on the region defined by  $m_j \geq 0$  and  $\sum_{j=1}^l m_j = 1$ , a subset of  $(-1/n, \infty)^l$ . On this region,  $g = 0$ , so  $L = f$ . Hence  $f$ , constrained to the aforementioned region, achieves a global maximum at  $\mathbf{m}'$ . □

We conclude the proof by calculating the value of

$$\sum_{j=1}^l \frac{m_j n}{(1 + nm_j) j}$$

for  $m_j$  as described in Lemma 4.4. Since this function evaluated at  $l = \lfloor n/2 \rfloor$  is a lower bound of  $|\phi_n(1, d\eta)|$ , as proved in Lemma 4.2, this final lemma concludes the proof of the theorem.

**Lemma 4.6.** For the  $m_j$  described in Lemma 4.4 in (4-3),

$$\sum_{j=1}^l \frac{m_j n}{(1 + nm_j) j} = \frac{1}{\pi} \log l + c,$$

where  $c$  is uniformly bounded.

*Proof.* We simply evaluate  $f$  from (4-1) at the  $\mathbf{m}'$  given by (4-3):

$$\begin{aligned} f(\mathbf{m}') &= \sum_{j=1}^l \frac{m'_j}{1 + nm'_j} \frac{n}{j} = \sum_{j=1}^l \frac{\frac{1}{n} \left( \frac{1}{\sqrt{j}} (1 + nm'_1) - 1 \right)}{1 + n \frac{1}{n} \left( \frac{1}{\sqrt{j}} (1 + nm'_1) - 1 \right)} \frac{n}{j} \\ &= \sum_{j=1}^l \frac{\frac{1}{\sqrt{j}} (1 + nm'_1) - 1}{\sqrt{j} (1 + nm'_1)} = \sum_{j=1}^l \frac{1}{j} - \frac{1}{1 + nm'_1} \alpha(l) \\ &= \sum_{j=1}^l \frac{1}{j} - \frac{\alpha(l)^2}{\left(1 + \frac{l}{n}\right)n} \\ &= \sum_{j=1}^l \frac{1}{j} - \frac{\alpha(l)^2}{n + l}. \end{aligned}$$

Now  $\sum_{j=1}^l 1/j$  differs from  $\log l$  by at most 1, and  $\alpha(l)^2/(n + l)$  is bounded in  $n$  and  $l$  since

$$0 \leq \frac{\alpha(l)^2}{n + l} < \frac{(2\sqrt{l})^2}{n + l} = \frac{4l}{n + l} \leq \frac{4n}{n} = 4.$$

Therefore, for the  $\mathbf{m}'$  given by (4-3),  $f(\mathbf{m}') = \log l + d_l$ , where  $d_l$  is a constant bounded uniformly in  $l$ . In light of Lemma 4.2, we have constructed a  $d\eta$  such that  $|\phi_n(1, d\eta)| \geq 1/\pi \log n + c$ , where  $c$  is a bounded constant, completing the proof of Theorem 4.1. □

### 5. Investigating higher degree polynomials

In the previous sections, we described the magnitude of monic polynomials of degree less than or equal to  $n$ , where  $n$  is the number of discrete masses in the measure, using Rahmanov’s formula in (1-4). However, we also want to describe the higher degree monic polynomials, i.e.,  $\phi_{n'}(z; d\eta)$ , where  $n' > n$ . Unfortunately, we are not able to do this for all  $n' > n$ , but we can partially describe  $\phi_{n'}(z; d\eta)$ , where  $n' = kn$ ,  $k \in \mathbb{N}$ .

Recall the definition of Verblunsky coefficients [Simon 2005]:

$$\phi_{n+1}(z) = z\phi_n(z) - \bar{\alpha}_n \phi_n^*(z), \tag{5-1}$$

where

$$\begin{aligned} \phi_n(z) &= \beta_n z^n + \dots + \beta_0, \quad 0 \leq j \leq n, \quad \beta_j \in \mathbb{C}, \\ \phi_n^*(z) &= \bar{\beta}_0 z^n + \dots + \bar{\beta}_n. \end{aligned}$$

In the  $n' = kn$  case, we are able to derive the corresponding Verblunsky coefficients, and do so explicitly for a  $d\rho$  similar to that of Rahmanov’s in Section 2.



**Theorem 5.1.** For a measure  $d\eta = d\theta/2\pi + \sum_{j=1}^n m_j \delta(\theta - \theta_j)$ , with masses located at  $\xi_j = e^{i\theta_j}$  and  $\theta_j = 2\pi j/n + \theta_0$  (cf. Lemma 2.1),

$$\phi_{nk}(z, d\eta) = z^{nk} - \xi_0^{nk} \sum_{j=1}^n \frac{m_j}{1 + m_j nk} K_{nk-1}(z, \xi_j),$$

and

$$\alpha_{nk-1} = \overline{\xi_0^{nk}} \sum_{j=1}^n \frac{m_j}{1 + m_j nk},$$

where  $\alpha_{nk-1}$  is a Verblunsky coefficient. Furthermore, under Rahmanov's scheme, where  $\theta_j = 2\pi j/n$  and

$$d\rho = \frac{d\theta}{2\pi} + \sum_{j=1}^n \frac{\delta(\theta - \theta_j)}{n},$$

the Verblunsky coefficients are

$$\alpha_{nk-1} = \frac{1}{1 + k}.$$

*Proof.* Note that, since  $\phi_n(z; d\eta)$  is a monic polynomial,  $\beta_n$  from the above definition of the Verblunsky coefficients is 1, so

$$\phi_n^*(0; d\eta) = 1,$$

which by (5-1) implies

$$\phi_{n+1}(0; d\eta) = -\bar{\alpha}_n. \tag{5-2}$$

Having set out these preliminaries, we can simply apply Rahmanov's formula [1979] from (1-4) to find a formula for  $\phi_{nk}(z; d\eta)$  under a measure  $d\eta$  as described in the statement of Theorem 5.1:

$$\phi_{nk}(z; d\eta) = z^{nk} - \sum_{j=1}^n \frac{m_j \phi_{nk}(\xi_j; d\mu)}{1 + m_j K_{nk-1}(\xi_j, \xi_j)} K_{nk-1}(z, \xi_j) \tag{5-3}$$

$$= z^{nk} - \xi_0^{nk} \sum_{j=1}^n \frac{m_j}{1 + m_j nk} K_{nk-1}(z, \xi_j). \tag{5-4}$$

**Remark 5.2.** The simplification of the numerator from (5-3) to (5-4) depends upon the  $\xi_j$  being roots of unity times a constant (as in Lemma 2.1). Such a simplification is only possible in the  $\phi_{nk}$  case, which is why the description of other higher-degree monic polynomials is considerably more complicated.

Now consider  $z = 0$  to find the Verblunsky coefficients:

$$\phi_{nk}(0, d\eta) = -\xi_0^{nk} \sum_{j=1}^n \frac{m_j}{1 + m_j nk} K_{nk-1}(0, \xi_j) = -\xi_0^{nk} \sum_{j=1}^n \frac{m_j}{1 + m_j nk},$$

and, applying (5-2), we obtain

$$\begin{aligned} -\bar{\alpha}_{nk-1} &= \phi_{nk}(0, d\eta) = -\xi_0^{nk} \sum_{j=1}^n \frac{m_j}{1 + m_j nk}, \\ \alpha_{nk-1} &= \overline{\xi_0^{nk}} \sum_{j=1}^n \frac{m_j}{1 + m_j nk}. \end{aligned} \tag{5-5}$$

If we now take  $\theta_0 = 0$ , as Rahmanov does, and

$$d\rho = \frac{d\theta}{2\pi} + \sum_{j=1}^n \frac{\delta(\theta - \theta_j)}{n},$$

then (5-5) simplifies to

$$\alpha_{nk-1} = \frac{1}{1 + k}. \tag{5-6}$$

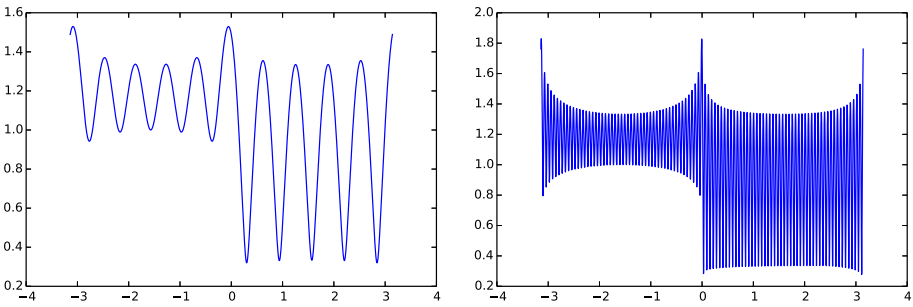
□

**Remark 5.3.** It is noteworthy that, as  $k$  grows, the  $\alpha_{nk-1}$  decay at the rate of  $1/(1 + k)$ . In light of the fact that  $\sum_{j=1}^\infty \alpha_j^2 < \infty$  [Simon 2005], this suggests that the  $\alpha_j$  are small for  $j \in (n(k - 1), nk)$ , where  $k \in \mathbb{N}$ , and increase rapidly near  $j = kn$ . However, as mentioned above, describing  $\phi_j(z; d\eta)$  for  $j \neq kn$  is much more complicated.

### Appendix: Numerical appendix

In order to help visualize the results of this paper, the graphs of the magnitudes of four orthogonal monic polynomials induced by four respective measures have been included at the end of this section. Each measure has a continuous portion of  $d\theta/2\pi$  as well as masses placed at  $\theta_j = \pi/n(2j - 1)$ , where  $1 \leq j \leq n/2$  (cf. Lemma 2.1). For simplicity, throughout this section, we will consider only even  $n$ . For the first two polynomials (displayed in Figure 1), masses of uniform size  $2/n$  are used as suggested by Rahmanov (see Section 2). For the second two (Figure 2), the masses are given their weights according to (4-3).

These graphs have several key features in common, including the presence of two peaks that grow in  $n$ : one at  $\theta = 0$  and another at  $\theta = \pi$ . Also, both have much lower minimums in the range  $0 \leq \theta \leq \pi$  than in  $-\pi \leq \theta \leq 0$ . Upon closer inspection, it can be seen that the two peaks in Figure 1 are equal; in contrast, in

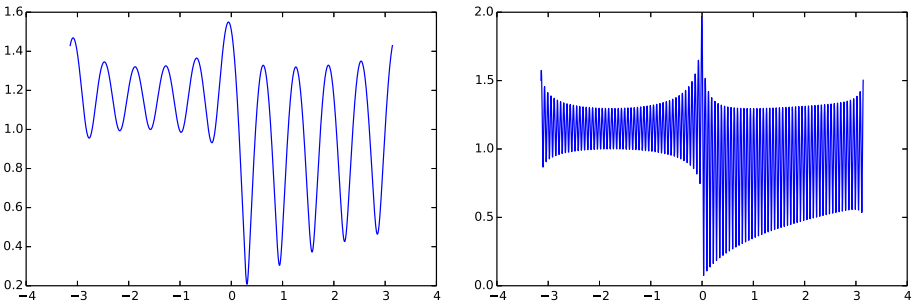


**Figure 1.** Left:  $|\phi_{10}(\theta)|$  for  $\theta_j = \frac{\pi}{10}(2j - 1)$  and  $m_j = \frac{1}{5}$ , where  $1 \leq j \leq 5$ . Right:  $|\phi_{100}(\theta)|$  for  $\theta_j = \frac{\pi}{100}(2j - 1)$  and  $m_j = \frac{1}{50}$ , where  $1 \leq j \leq 50$ .

Figure 2, the peak at  $\theta = 0$  is larger than the peak at  $\theta = \pi$ . Additionally, the peak at  $\theta = 0$  in the latter case is higher than in the former, as predicted by Theorem 4.1.

To explain some of these features, first note that with the above choice of placement of the masses, Rahmanov’s formula (1-4) [1979] reduces to

$$\begin{aligned} \operatorname{Re}(\phi_n(e^{i\theta})) &= (1 + \cos(n\theta)) \left( 1 + \frac{1}{2} \sum_{j=1}^{n/2} \frac{m_j}{1 + nm_j} \right) - \frac{1}{2} \sin(n\theta) \sum_{j=1}^{n/2} \frac{m_j}{1 + nm_j} \frac{\sin(\theta - \theta_j)}{1 - \cos(\theta - \theta_j)}, \\ \operatorname{Im}(\phi_n(e^{i\theta})) &= \sin(n\theta) \left( 1 + \frac{1}{2} \sum_{j=1}^{n/2} \frac{m_j}{1 + nm_j} \right) + \frac{1}{2} (1 + \cos(n\theta)) \sum_{j=1}^{n/2} \frac{m_j}{1 + nm_j} \frac{\sin(\theta - \theta_j)}{1 - \cos(\theta - \theta_j)}. \end{aligned}$$



**Figure 2.** Left:  $|\phi_{10}(\theta)|$  for  $\theta_j = \frac{\pi}{10}(2j - 1)$  and  $m_j$  chosen optimally, where  $1 \leq j \leq 5$ . Right:  $|\phi_{100}(\theta)|$  for  $\theta_j = \frac{\pi}{100}(2j - 1)$  and  $m_j$  chosen optimally, where  $1 \leq j \leq 50$ .

**Analysis of the minima.** Due to its prominent role in each term, let us evaluate both the real and imaginary parts at the extrema of  $1 + \cos(n\theta)$ , that is,  $\theta = \theta_k = (\pi/n)(2k-1)$  and  $\theta = \theta_k^* = 2\pi k/n$ . For  $\theta = \theta_k$ ,  $\sin(n\theta_k)$  and  $1 + \cos(n\theta_k)$  are each zero. However, we must be careful, because for  $1 \leq k \leq n/2$ , one of the terms in the sum will have a denominator of zero. Thus, using L'Hôpital's rule, we take the limits

$$\begin{aligned} \lim_{\theta \rightarrow \theta_k} \frac{\sin(n\theta) \sin(\theta - \theta_k)}{1 - \cos(\theta - \theta_k)} &= -2n, \\ \lim_{\theta \rightarrow \theta_k} \frac{(1 + \cos(n\theta)) \sin(\theta - \theta_k)}{1 - \cos(\theta - \theta_k)} &= 0. \end{aligned}$$

Substituting these values into our formulae, we then have that

$$|\phi_n(e^{i\theta_k})| = \begin{cases} 1 - nm_k/(1 + nm_k) & \text{if } 1 \leq k \leq n/2, \\ 1 & \text{otherwise.} \end{cases}$$

Thus, the minima will be lower in the region where the masses are placed than outside that region. Also, we can now see the reason for the minima increasing as  $\theta$  increases in the cases where the choice of  $m_j$  is optimal, as in [Figure 2](#).

**Analysis of peaks at  $\theta = 0, \pi$ .** Now, let us examine the values of the polynomials at  $\theta = \theta_k^*$ . In this case,  $\sin(n\theta_k^*)$  is still zero, but  $1 + \cos(n\theta_k^*)$  is instead 2, so we need not worry about zero denominators. Immediately, we have that our previous formulae reduce to

$$\begin{aligned} \operatorname{Re}(\phi_n(e^{i\theta_k^*})) &= 1 + \sum_{j=1}^{n/2} \frac{m_j}{1 + nm_j}, \\ \operatorname{Im}(\phi_n(e^{i\theta_k^*})) &= \sum_{j=1}^{n/2} \frac{m_j}{1 + nm_j} \frac{\sin(\theta_k^* - \theta_j)}{1 - \cos(\theta_k^* - \theta_j)}. \end{aligned} \tag{A-1}$$

The real part is constant in  $\theta_k^*$  and can be ignored. For  $k = 0$ , we have precisely the sum that was analyzed in [Section 4](#). For  $k = n/2$ , we obtain the sum

$$\begin{aligned} \operatorname{Im}(\phi_n(e^{i\theta_{n/2}^*})) &= \sum_{j=1}^{n/2} \frac{m_j}{1 + nm_j} \frac{\sin(\pi - \theta_j)}{1 - \cos(\pi - \theta_j)} \\ &= \sum_{j=1}^{n/2} \frac{m_j}{1 + nm_j} \frac{\sin \theta_j}{1 + \cos \theta_j}. \end{aligned}$$

It can easily be seen that, if  $m_j$  is constant, this sum will be identical to the sum for  $k = 0$ , and so the result will be two peaks of equal amplitude as we observed before in [Figure 1](#). If  $m_j$  decreases proportionally to  $1/\sqrt{j}$ , however, this sum

will be very different from the sum for  $k = 0$ , since the largest terms of the sum will now be those  $\theta_j$  close to  $\pi$  rather than zero. The  $m_j$  with corresponding  $\theta_j$  close to  $\pi$  will all be of the order  $1/n$ , and so we would expect that the value of the polynomial here will behave something more similarly to the peaks of the uniform mass case than to those of the optimal  $m$  case.

**Analysis of peaks away from  $\theta = 0, \pi$ .** However, we have not yet explained why the peaks away from  $\theta = 0$  and  $\theta = \pi$  are all smaller, so now we consider the case where  $\theta = \theta_k^*$  for  $0 < k < n/2$ . First, note that

$$\theta_k^* - \theta_j = \frac{\pi}{n}(2(k - j) + 1),$$

and consider the terms in the sum (A-1), where  $j = k$  and  $j = k + 1$ . These terms will be

$$\frac{m_k}{1 + nm_k} \frac{\sin \frac{\pi}{n}}{1 - \cos \frac{\pi}{n}}$$

and

$$-\frac{m_{k+1}}{1 + nm_{k+1}} \frac{\sin \frac{\pi}{n}}{1 - \cos \frac{\pi}{n}}.$$

In the case that all the masses have equal weight, these terms will cancel out completely, and, even in the case of the optimal choice of  $m_j$ , they still mostly cancel out since the difference of  $m_{k+1}$  and  $m_k$  will be small. In general, for the  $j = k - l$  and  $j = k + l + 1$  terms, as long as  $k - l \geq 1$  and  $k + l + 1 \leq n/2$  are satisfied, similar cancellations will occur. Thus, the values at these peaks will be less than those at  $\theta = 0$  and  $\theta = \pi$ .

### Acknowledgements

Our research was done during the 2013 University of Wisconsin-Madison REU, sponsored by NSF grants DMS-1056327 and DMS-1147523. We would like to thank Serguei Denissov for introducing us to this problem and advising us as we wrote this paper.

### References

- [Rahmanov 1979] E. A. Rahmanov, “Steklov’s conjecture in the theory of orthogonal polynomials”, *Mat. Sb. (N.S.)* **108(150)**:4 (1979), 581–608. In Russian; translated in *USSR-Sb* **36**:4 (1980), 549–575. [MR 81j:42042](#) [Zbl 0452.33012](#)
- [Simon 2005] B. Simon, *Orthogonal polynomials on the unit circle, I: Classical theory*, American Mathematical Society Colloquium Publications **54**, Amer. Math. Soc., Providence, RI, 2005. [MR 2006a:42002a](#) [Zbl 1082.42020](#)

[hoffman.locke@gmail.com](mailto:hoffman.locke@gmail.com)

*University of Wisconsin-Madison, Madison, WI 53706,  
United States*

[mtmeyer3@wisc.edu](mailto:mtmeyer3@wisc.edu)

*Department of Applied and Natural Sciences,  
University of Wisconsin-Green Bay, 2420 Nicolet Drive,  
Green Bay, WI 54311, United States*

[sardarli@princeton.edu](mailto:sardarli@princeton.edu)

*Princeton University, Princeton, NJ 08544, United States*

[ajsherman2@wisc.edu](mailto:ajsherman2@wisc.edu)

*University of Wisconsin-Madison, Madison, WI 53706,  
United States*

# A type of multiple integral with log-gamma function

Duokui Yan, Rongchang Liu and Geng-zhe Chang

(Communicated by Kenneth S. Berenhaut)

In this paper, we give a general formula for the multiple integral

$$I = \int_0^1 \int_0^1 \dots \int_0^1 f(x_1 + x_2 + \dots + x_n) dx_1 dx_2 \dots dx_n.$$

As an application, the integral  $I$  with  $f(x) = \log \Gamma(x)$  is evaluated for all  $n \in \mathbb{N}$ . The subsidiary computational challenges are interesting in their own right.

## 1. Introduction

A general idea, when faced with a multiple integral, is to lower its dimension. A well-known example, (see [Chang and Shi 2003], for instance) is the  $n$ -dimensional integral

$$\int_{\substack{x_1+x_2+\dots+x_n \leq 1 \\ x_1, x_2, \dots, x_n \geq 0}} \dots \int f(x_1 + x_2 + \dots + x_n) dx_1 dx_2 \dots dx_n, \quad (1-1)$$

which can be simplified to a one-dimensional integral

$$\frac{1}{(n-1)!} \int_0^1 t^{n-1} f(t) dt.$$

However, to the best of our knowledge, a similar integral,

$$I = \int_0^1 \int_0^1 \dots \int_0^1 f(x_1 + x_2 + \dots + x_n) dx_1 dx_2 \dots dx_n, \quad (1-2)$$

has no such formula.

The aim of this paper is to find a formula for the above integral  $I$  and apply it to the special case when  $f(x) = \log \Gamma(x)$ . The main results are as follows. A general formula of  $I$  is obtained in [Theorem 4.1](#).

*MSC2010:* 05A19, 54C30.

*Keywords:* multiple integral, log-gamma function.

The research of Duokui Yan is supported in part by NSFC (No. 11101221).

**Theorem 4.1.** *The integral  $I$  satisfies*

$$\begin{aligned}
 I &= \int_0^1 \int_0^1 \dots \int_0^1 f(x_1 + x_2 + \dots + x_n) dx_1 dx_2 \dots dx_n \\
 &= \frac{1}{(n-1)!} \sum_{m=1}^n \int_0^1 G_m(t) f(t+m-1) dt,
 \end{aligned}
 \tag{1-3}$$

where

$$G_m(t) = \sum_{i=1}^m (-1)^{i-1} (t+m-i)^{n-1} \binom{n}{i-1}.$$

When  $f(x) = \log \Gamma(x)$ , the value of  $I$  is given in [Theorem 5.1](#). The main challenge of the proof is to find appropriate combinatorial identities to simplify  $I$ .

**Theorem 5.1.**

$$\begin{aligned}
 I &= I(n) = \int_0^1 \int_0^1 \dots \int_0^1 \log \Gamma(x_1 + x_2 + \dots + x_n) dx_1 dx_2 \dots dx_n \\
 &= \frac{1}{2} \log(2\pi) - \frac{n-1}{2} H_n + \sum_{k=2}^{n-1} \frac{(-1)^{n+k+1} k^n}{n!} \binom{n-1}{k} \log k,
 \end{aligned}$$

where the last sum is missing when  $n = 2$  and  $H_n = \sum_{k=1}^n 1/k$ .

The paper is organized as follows. In [Sections 2 and 3](#), we explain the main ideas by using the cases  $n = 2$  and  $3$ . One can see from [Figures 1 and 2](#) how we cut the square and the cube so that the integral  $I$  over each subset becomes a simple one-dimensional integral. In [Section 4](#), a formula of  $I$  is derived in [Theorem 4.1](#), and in [Section 5](#), we evaluate  $I$  when  $f(x) = \log \Gamma(x)$ .

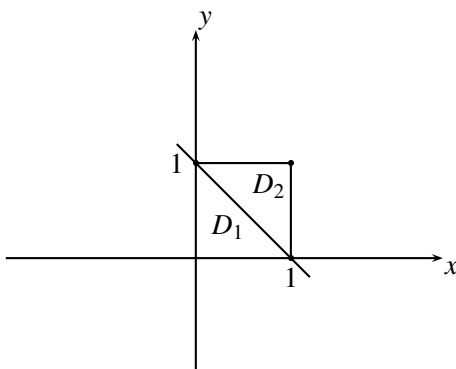
### 2. The case $n = 2$

When  $n = 2$ , the integral  $I$  becomes  $\int_0^1 \int_0^1 f(x+y) dx dy$ , where the integral domain is a unit square. Let  $t = x+y$ . The unit square can then be divided into two domains,  $D_1$  and  $D_2$  as in [Figure 1](#), where

$$\begin{aligned}
 D_1 &= \{(x, y) : 0 \leq x+y \leq 1, 0 \leq x \leq 1, 0 \leq y \leq 1\}, \\
 D_2 &= \{(x, y) : 1 \leq x+y \leq 2, 0 \leq x \leq 1, 0 \leq y \leq 1\}.
 \end{aligned}$$

The following lemma shows that  $t_0^1 \int_0^1 f(x+y) dx dy$  is the sum of two one-dimensional integrals.





**Figure 1.** Domains  $D_1$  and  $D_2$ .

**Lemma 2.1.**

$$\begin{aligned} \int_0^1 \int_0^1 f(x+y) dx dy &= \iint_{D_1} f(x+y) dx dy + \iint_{D_2} f(x+y) dx dy \\ &= \int_0^1 t f(t) dt + \int_0^1 (1-t) f(t+1) dt. \end{aligned} \tag{2-1}$$

*Proof.* It is clear that

$$\int_0^1 \int_0^1 f(x+y) dx dy = \iint_{D_1} f(x+y) dx dy + \iint_{D_2} f(x+y) dx dy.$$

We first consider  $\iint_{D_1} f(x+y) dx dy$ . Note that  $t = x + y$ , and consider the transformation  $(x, y) \mapsto (x, t)$ . It is clear that the Jacobian is 1. Then

$$\iint_{D_1} f(x+y) dx dy = \int_0^1 \int_0^t f(t) dx dt = \int_0^1 t f(t) dt. \tag{2-2}$$

For the integral over domain  $D_2$ , we set  $x_1 = 1 - x$  and  $y_1 = 1 - y$ . Then  $(x_1, y_1) \in D_1$  and

$$\begin{aligned} \iint_{D_2} f(x+y) dx dy &= \iint_{D_1} f(2-x_1-y_1) dx_1 dy_1 \\ &= \int_0^1 t f(2-t) dt. \end{aligned} \tag{2-3}$$

If one sets  $u = 1 - t$ , it follows that  $\int_0^1 t f(2-t) dt = \int_0^1 (1-u) f(u+1) du$ . Then

$$\iint_{D_2} f(x+y) dx dy = \int_0^1 (1-u) f(u+1) du. \tag{2-4}$$

Then, identity (2-1) follows by identities (2-2) and (2-4). □

### 3. The case $n = 3$

When  $n = 3$ , the integral domain of  $I$  is a unit cube. The main idea is to cut the unit cube into several simplexes so that we can apply the integral formula (1-1) over each one.

Let  $E = \{(x, y, z) : 0 \leq x \leq 1, 0 \leq y \leq 1, 0 \leq z \leq 1\}$  be the unit cube. Set

$$E_1 = \{(x, y, z) : 0 \leq x + y + z \leq 1, 0 \leq x \leq 1, 0 \leq y \leq 1, 0 \leq z \leq 1\},$$

$$E_2 = \{(x, y, z) : 1 \leq x + y + z \leq 2, 0 \leq x \leq 1, 0 \leq y \leq 1, 0 \leq z \leq 1\},$$

$$E_3 = \{(x, y, z) : 2 \leq x + y + z \leq 3, 0 \leq x \leq 1, 0 \leq y \leq 1, 0 \leq z \leq 1\}.$$

Then  $E = E_1 \cup E_2 \cup E_3$  and the integral  $I$  satisfies

$$\begin{aligned} I &= \int_0^1 \int_0^1 \int_0^1 f(x + y + z) dx dy dz \\ &= \int_{E_1} f(x + y + z) dx dy dz + \int_{E_2} f(x + y + z) dx dy dz \\ &\quad + \int_{E_3} f(x + y + z) dx dy dz. \end{aligned}$$

Using formula (1-1), it follows that  $\int_{E_1} f(x + y + z) dx dy dz = \frac{1}{2} \int_0^1 t^2 f(t) dt$ . The difficult parts are the integrals over  $E_2$  and  $E_3$ . The following lemma explains how to simplify these two integrals to one-dimensional integrals.

#### Lemma 3.1.

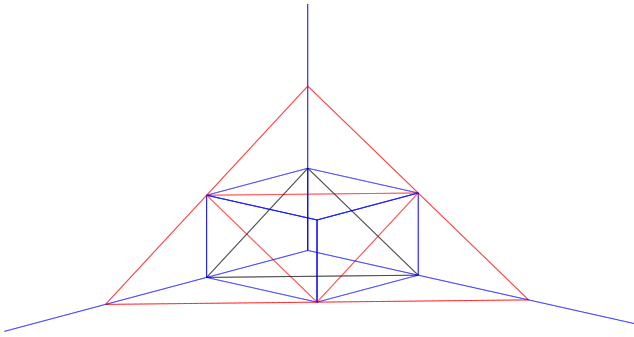
$$\begin{aligned} &\int_0^1 \int_0^1 \int_0^1 f(x + y + z) dx dy dz = \\ &\frac{1}{2} \int_0^1 t^2 f(t) dt + \frac{1}{2} \int_0^1 (-2t^2 + 2t + 1) f(t + 1) dt + \frac{1}{2} \int_0^1 (1 - t)^2 f(t + 2) dt. \quad (3-1) \end{aligned}$$

*Proof.* We introduce the transformation  $(x, y, z) \mapsto (x, y, t)$ . By formula (1-1),

$$\int_{E_1} f(x + y + z) dx dy dz = \frac{1}{2} \int_0^1 t^2 f(t) dt. \quad (3-2)$$

Note that integral (3-2) can be applied to calculate the integral over  $E_3$ . Let  $x_1 = 1 - x$ ,  $y_1 = 1 - y$  and  $z_1 = 1 - z$ . The integral over  $E_3$  becomes

$$\begin{aligned} \int_{E_3} f(x + y + z) dx dy dz &= \int_{E_1} f(3 - x_1 - y_1 - z_1) dx_1 dy_1 dz_1 \\ &= \frac{1}{2} \int_0^1 t^2 f(3 - t) dt. \quad (3-3) \end{aligned}$$



**Figure 2.** Region  $E_{20}$  and its partition:  $E_2, E_{21}, E_{22}, E_{23}$ .

If one sets  $u = 1 - t$ , it implies that  $\frac{1}{2} \int_0^1 t^2 f(3-t) dt = \frac{1}{2} \int_0^1 (1-u)^2 f(2+u) du$ . Hence,

$$\int_{E_3} f(x + y + z) dx dy dz = \frac{1}{2} \int_0^1 (1-t)^2 f(t + 2) dt. \quad (3-4)$$

By equalities (3-2) and (3-4), it is sufficient to show that

$$\int_{E_2} f(x + y + z) dx dy dz = \frac{1}{2} \int_0^1 (-2t^2 + 2t + 1) f(t + 1) dt. \quad (3-5)$$

Consider the domain

$$E_{20} = \{(x, y, z) : 1 \leq x + y + z \leq 2, 0 \leq x \leq 2, 0 \leq y \leq 2, 0 \leq z \leq 2\}.$$

Similar to Figure 1, we can cut  $E_{20}$  into 4 different domains,  $E_2, E_{21}, E_{22}$  and  $E_{23}$ , so that the integral over each domain can be handled easily. A picture of this partition is shown in Figure 2.

$$E_2 = \{(x, y, z) : 1 \leq x + y + z \leq 2, 0 \leq x \leq 1, 0 \leq y \leq 1, 0 \leq z \leq 1\},$$

$$E_{21} = \{(x, y, z) : 1 \leq x + y + z \leq 2, 1 \leq x \leq 2, 0 \leq y \leq 1, 0 \leq z \leq 1\},$$

$$E_{22} = \{(x, y, z) : 1 \leq x + y + z \leq 2, 0 \leq x \leq 1, 1 \leq y \leq 2, 0 \leq z \leq 1\},$$

$$E_{23} = \{(x, y, z) : 1 \leq x + y + z \leq 2, 0 \leq x \leq 1, 0 \leq y \leq 1, 1 \leq z \leq 2\},$$

where  $E_{20} = E_2 \cup E_{21} \cup E_{22} \cup E_{23}$ .

Again by using formula (1-1), the integral over  $E_{20}$  is

$$\int_{E_{20}} f(x + y + z) dx dy dz = \int_1^2 \frac{1}{2} t^2 f(t) dt = \frac{1}{2} \int_0^1 (t + 1)^2 f(t + 1) dt. \quad (3-6)$$

On the other hand, the integral over  $E_{20}$  satisfies

$$\begin{aligned} & \int_{E_{20}} f(x+y+z) dx dy dz \\ &= \int_{E_{21}} f(x+y+z) dx dy dz + \int_{E_{22}} f(x+y+z) dx dy dz \\ & \quad + \int_{E_{23}} f(x+y+z) dx dy dz + \int_{E_2} f(x+y+z) dx dy dz. \end{aligned} \quad (3-7)$$

By the definitions of  $E_{21}$ ,  $E_{22}$  and  $E_{23}$ , it is clear that

$$\int_{E_{21}} f(x+y+z) dx dy dz = \int_{E_{22}} f(x+y+z) dx dy dz = \int_{E_{23}} f(x+y+z) dx dy dz.$$

So we only need to consider  $\int_{E_{21}} f(x+y+z) dx dy dz$ . Let  $\tilde{x} = x - 1$ ; then by equality (3-2),

$$\begin{aligned} \int_{E_{21}} f(x+y+z) dx dy dz &= \int_{E_1} f(\tilde{x} + y + z + 1) d\tilde{x} dy dz \\ &= \frac{1}{2} \int_0^1 t^2 f(t+1) dt. \end{aligned} \quad (3-8)$$

Therefore, (3-6), (3-7) and (3-8) imply that

$$\begin{aligned} & \int_{E_2} f(x+y+z) dx dy dz \\ &= \int_{E_{20}} f(x+y+z) dx dy dz - 3 \int_{E_{21}} f(x+y+z) dx dy dz \\ &= \frac{1}{2} \int_0^1 (t+1)^2 f(t+1) dt - \frac{3}{2} \int_0^1 t^2 f(t+1) dt \\ &= \frac{1}{2} \int_0^1 (-2t^2 + 2t + 1) f(t+1) dt, \end{aligned}$$

which shows equality (3-5). □

#### 4. The general case

In this section, we give a general formula for

$$I = \int_0^1 \int_0^1 \dots \int_0^1 f(x_1 + x_2 + \dots + x_n) dx_1 dx_2 \dots dx_n$$

in [Theorem 4.1](#). In order to prove it, we first find a recursive formula for  $I$  in [Theorem 4.3](#). The proof of [Theorem 4.1](#) then follows by [Theorem 4.4](#) and [Theorem 4.3](#).

**Theorem 4.1.** *The integral  $I$  satisfies*

$$\begin{aligned}
 I &= \int_0^1 \int_0^1 \dots \int_0^1 f(x_1 + x_2 + \dots + x_n) dx_1 dx_2 \dots dx_n \\
 &= \frac{1}{(n-1)!} \sum_{m=1}^n \int_0^1 G_m(t) f(t+m-1) dt,
 \end{aligned}
 \tag{4-1}$$

where

$$G_m(t) = \sum_{i=1}^m (-1)^{i-1} (t+m-i)^{n-1} \binom{n}{i-1}.$$

The idea is to divide the  $n$ -dimensional unit box into  $n$  different polyhedrons and the integral  $I$  over each polyhedron can be simplified to a one-dimensional integral by applying the ideas in the 2D or 3D cases. The  $n$  different polyhedrons are defined as follows:

$$\begin{aligned}
 K_1 &= \{(x_1, x_2, \dots, x_n) : 0 \leq x_1 + x_2 + \dots + x_n \leq 1, \\
 &\hspace{15em} 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1, \dots, 0 \leq x_n \leq 1\}, \\
 K_2 &= \{(x_1, x_2, \dots, x_n) : 1 \leq x_1 + x_2 + \dots + x_n \leq 2, \\
 &\hspace{15em} 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1, \dots, 0 \leq x_n \leq 1\}, \\
 &\vdots \\
 K_n &= \{(x_1, x_2, \dots, x_n) : n-1 \leq x_1 + x_2 + \dots + x_n \leq n, \\
 &\hspace{15em} 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1, \dots, 0 \leq x_n \leq 1\}.
 \end{aligned}$$

By formula (1-1), the integral over  $K_1$  satisfies the following proposition.

**Proposition 4.2.**

$$\int_{K_1} f(x_1 + x_2 + \dots + x_n) dx_1 dx_2 \dots dx_n = \frac{1}{(n-1)!} \int_0^1 t^{n-1} f(t) dt.$$

Let

$$I_m = \int_{K_m} f(x_1 + x_2 + \dots + x_n) dx_1 dx_2 \dots dx_n, \quad m = 1, 2, \dots, n.$$

It is obvious that  $I = \sum_{m=1}^n I_m$ . Then the integral  $I$  reduces to the calculation of each  $I_m$  ( $1 \leq m \leq n$ ). Define

$$J_{s,m} = \int_{K_s} f(x_1 + \dots + x_n + m-s) dx_1 dx_2 \dots dx_n, \tag{4-2}$$

where  $s$  is an integer and  $1 \leq s \leq m$ . Note that  $J_{m,m} = I_m$ . For any  $1 \leq s \leq m-1$ ,  $J_{s,m}$  can be calculated by  $I_s$ . The following theorem shows that  $I_m$  satisfies a recursive formula.

**Theorem 4.3.**

$$I_m = \frac{1}{(n-1)!} \int_0^1 (t+m-1)^{n-1} f(t+m-1) dt - a_1 J_{1,m} - a_2 J_{2,m} - \dots - a_{m-1} J_{m-1,m}, \quad (4-3)$$

where

$$a_i = \binom{m+n-i-1}{n-1}, \quad i = 1, 2, \dots, m-1.$$

*Proof.* We consider the region

$$K_{m0} = \{(x_1, x_2, \dots, x_n) : m-1 \leq x_1 + x_2 + \dots + x_n \leq m, 0 \leq x_1 \leq m, 0 \leq x_2 \leq m, \dots, 0 \leq x_n \leq m\}.$$

By Proposition 4.2,

$$\begin{aligned} \int_{K_{m0}} f(x_1 + x_2 + \dots + x_n) dx_1 dx_2 \dots dx_n &= \frac{1}{(n-1)!} \int_{m-1}^m t^{n-1} f(t) dt \\ &= \frac{1}{(n-1)!} \int_0^1 (t+m-1)^{n-1} f(t+m-1) dt. \end{aligned} \quad (4-4)$$

We define the subset  $K_{i_1 i_2 \dots i_n} \subset K_{m0}$  as follows:

$$K_{i_1 i_2 \dots i_n} = \{(x_1, x_2, \dots, x_n) : m-1 \leq x_1 + x_2 + \dots + x_n \leq m, i_1 - 1 \leq x_1 \leq i_1, i_2 - 1 \leq x_2 \leq i_2, \dots, i_n - 1 \leq x_n \leq i_n\},$$

where  $i_1, i_2, \dots, i_n \in [1, m]$  are positive integers. It is easily seen that the intersection of any two subsets  $K_{i_1 i_2 \dots i_n}$  only happens on their boundaries. We then classify all possible  $K_{i_1 i_2 \dots i_n}$  so that the integral over each one can be evaluated easily. Note that by definition,  $K_{1,1,\dots,1} = K_m$ . To find the integral over  $K_m$ , we need to subtract the integrals over all the other nonempty subsets  $K_{i_1 i_2 \dots i_n}$  ( $i_1, i_2, \dots, i_n \in [1, m]$ ) from  $\int_{K_{m0}} f(x_1 + x_2 + \dots + x_n) dx_1 dx_2 \dots dx_n$ .

The first step is to determine when  $K_{i_1 i_2 \dots i_n}$  ( $i_1, i_2, \dots, i_n \in [1, m]$ ) is nonempty. For any set  $K_{i_1 i_2 \dots i_n}$ , let

$$\tilde{x}_1 = x_1 - (i_1 - 1), \quad \tilde{x}_2 = x_2 - (i_2 - 1), \quad \dots, \quad \tilde{x}_n = x_n - (i_n - 1). \quad (4-5)$$

Then  $K_{i_1 i_2 \dots i_n}$  becomes

$$\begin{aligned} \tilde{K}_{i_1 i_2 \dots i_n} = \{(\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n) : m+n-\alpha-1 \leq \tilde{x}_1 + \tilde{x}_2 + \dots + \tilde{x}_n \leq m+n-\alpha, 0 \leq \tilde{x}_1 \leq 1, 0 \leq \tilde{x}_2 \leq 1, \dots, 0 \leq \tilde{x}_n \leq 1\}. \end{aligned}$$

where  $\alpha = i_1 + i_2 + \dots + i_n$ . Let  $s = m + n - \alpha$ . It is clear that  $K_{i_1 i_2 \dots i_n} \cong \tilde{K}_{i_1 i_2 \dots i_n} = K_s$ . Since  $m + n - s = \sum_{j=1}^n i_j \geq n$ , it follows that  $s \leq m$ . Note that if  $s = m$ , by equality (4-2),  $J_{m,m} = I_m$ . If  $s = 0$ ,  $K_{i_1 i_2 \dots i_n} \cong \tilde{K}_{i_1 i_2 \dots i_n} = \{0\}$ , and if  $s < 0$ ,  $K_{i_1 i_2 \dots i_n} \cong \tilde{K}_{i_1 i_2 \dots i_n} = \emptyset$ . So we only need to consider the case  $1 \leq s \leq m - 1$ . For any given  $s \in [1, m - 1]$ , it follows that

$$\begin{aligned} & \int_{K_{i_1 i_2 \dots i_n}} f(x_1 + x_2 + \dots + x_n) dx_1 dx_2 \dots dx_n \\ &= \int_{\tilde{K}_{i_1 i_2 \dots i_n}} f(\tilde{x}_1 + \dots + \tilde{x}_n + i_1 + \dots + i_n - n) d\tilde{x}_1 \dots d\tilde{x}_n \\ &= \int_{K_s} f(x_1 + \dots + x_n + m - s) dx_1 \dots dx_n \\ &= J_{s,m}. \end{aligned} \tag{4-6}$$

It implies that the subsets  $K_{i_1 i_2 \dots i_n}$  ( $i_1, i_2, \dots, i_n \in [1, m]$ ,  $i_1 + i_2 + \dots + i_n \neq n$ ) with nonzero measure can be classified into  $m - 1$  classes. In each class, every element is identical to some subset  $K_s$  after a shifting transformation in (4-5):  $(x_1, x_2, \dots, x_n) \mapsto (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n)$ .

Next step is to fix  $m$  and  $s$  ( $1 \leq s \leq m - 1$ ), and find out how many subsets are identical to  $K_s$ . Since  $s = m + n - (i_1 + i_2 + \dots + i_n)$ , we have

$$m + n - s = i_1 + i_2 + \dots + i_n, \quad \text{where } i_1, i_2, \dots, i_n \text{ are positive integers.} \tag{4-7}$$

The number of positive integer solutions  $(i_1, i_2, \dots, i_n)$  for (4-7) is  $\binom{m+n-s-1}{n-1}$ . It follows that the total number of subsets identical to  $K_s$  ( $s \in [1, m - 1]$ ) is

$$a_s = \binom{m+n-s-1}{n-1}. \tag{4-8}$$

Therefore, by equalities (4-4), (4-6) and (4-8),  $I_m$  satisfies

$$\begin{aligned} I_m &= \int_{K_m} f(x_1 + x_2 + \dots + x_n) dx_1 dx_2 \dots dx_n \\ &= \frac{1}{(n-1)!} \int_0^1 (t+m-1)^{n-1} f(t+m-1) dt \\ &\quad - a_1 J_{1,m} - a_2 J_{2,m} - \dots - a_{m-1} J_{m-1,m}, \end{aligned} \tag{4-9}$$

where  $a_s$  ( $s = 1, \dots, m - 1$ ) is defined by (4-8). □

By using the cases  $n = 2$  and  $3$ , we can show by induction that

$$I_m = \frac{1}{(n-1)!} \int_0^1 G_m(t) f(t+m-1) dt, \tag{4-10}$$

where  $G_m(t)$  is a polynomial. It follows that

$$\begin{aligned} J_{s,m} &= \int_{K_s} f(x_1 + \dots + x_n + m - s) dx_1 \dots dx_n \\ &= \frac{1}{(n-1)!} \int_0^1 G_s(t) f(t + m - 1) dt, \end{aligned} \tag{4-11}$$

where  $s$  is an integer and  $1 \leq s \leq m$ . The integral  $I$  satisfies

$$I = \int_0^1 \int_0^1 \dots \int_0^1 f(x_1 + x_2 + \dots + x_n) dx_1 dx_2 \dots dx_n = \sum_{m=1}^n I_m. \tag{4-12}$$

In order to find a formula for  $I$ , we only need to compute the polynomial  $G_m(t)$  in equality (4-10) for all  $1 \leq m \leq n$ . For  $m = 1, 2$  and  $3$ , a direct calculation shows that

$$\begin{aligned} G_1(t) &= t^{n-1}, \\ G_2(t) &= (t + 1)^{n-1} - \binom{n}{1} t^{n-1}. \end{aligned} \tag{4-13}$$

By Theorem 4.3 and equality (4-11),

$$\begin{aligned} G_3(t) &= (t + 2)^{n-1} - \binom{n+1}{n-1} G_1(t) - \binom{n}{n-1} G_2(t) \\ &= (t + 2)^{n-1} - \binom{n}{1} (t + 1)^{n-1} + \binom{n}{2} t^{n-1}. \end{aligned}$$

Similarly,

$$G_4(t) = (t + 3)^{n-1} - \binom{n}{1} (t + 2)^{n-1} + \binom{n}{2} (t + 1)^{n-1} - \binom{n}{3} t^{n-1}.$$

It is reasonable to believe that  $G_m(t)$  follows a pattern. The following theorem actually proves this fact.

**Theorem 4.4.** 
$$G_m(t) = \sum_{i=1}^m (-1)^{i-1} (t + m - i)^{n-1} \binom{n}{i-1}. \tag{4-14}$$

*Proof.* The proof is based on the recursive formula (4-3) in Theorem 4.3 and the identity (4-11). By formula (4-3),

$$\begin{aligned} I_m &= \frac{1}{(n-1)!} \int_0^1 (t + m - 1)^{n-1} f(t + m - 1) dt - \sum_{i=1}^{m-1} a_i J_{i,m} \\ &= \frac{1}{(n-1)!} \int_0^1 G_m(t) f(t + m - 1) dt, \end{aligned}$$



where

$$G_m(t) = (t + m - 1)^{n-1} - \sum_{i=1}^{m-1} a_i G_i(t), \quad \text{and} \quad a_i = \binom{m+n-i-1}{n-1}. \quad (4-15)$$

We show this theorem by induction. It is clear that formula (4-14) of  $G_m(t)$  holds for  $m = 1$ . Assume that it holds for any  $1 \leq m \leq k$ . We need to show that formula (4-14) also holds for  $m = k + 1$ .

By (4-15) and the induction assumption, the polynomial  $G_{k+1}(t)$  satisfies

$$G_{k+1}(t) = (t + k)^{n-1} + \sum_{i=1}^k \binom{k+1+n-i-1}{n-1} \sum_{j=1}^i (-1)^j (t + i - j)^{n-1} \binom{n}{j-1}. \quad (4-16)$$

By formula (4-14), we can consider each  $G_m(t)$  ( $1 \leq m \leq k$ ) as a polynomial of  $(t + m - j)^{n-1}$  ( $j = 1, 2, \dots, m$ ) with coefficient  $(-1)^{j-1} \binom{n}{j-1}$ . Then identity (4-16) implies that the coefficient of  $(t + p)^{n-1}$  in  $G_{k+1}(t)$  is

$$L_p(G_{k+1}(t)) = \sum_{i=p+1}^k \binom{k+1+n-i-1}{n-1} (-1)^{i-p} \binom{n}{i-p-1}, \quad (4-17)$$

where  $p \in [0, k - 1]$  is an integer. Similarly,  $G_k(t)$  satisfies

$$G_k(t) = (t + k - 1)^{n-1} + \sum_{i=1}^{k-1} \binom{k+n-i-1}{n-1} \sum_{j=1}^i (-1)^j (t + i - j)^{n-1} \binom{n}{j-1},$$

and the coefficient of  $(t + p)^{n-1}$  ( $p \in [0, k - 2]$ ) in  $G_k(t)$  is

$$\sum_{i=p+1}^{k-1} \binom{k+n-i-1}{n-1} (-1)^{i-p} \binom{n}{i-p-1}. \quad (4-18)$$

Note that  $G_k(t) = \sum_{i=1}^k (-1)^{i-1} (t + k - i)^{n-1} \binom{n}{i-1}$ . It follows that

$$\sum_{i=p+1}^{k-1} \binom{k+n-i-1}{n-1} (-1)^{i-p} \binom{n}{i-p-1} = (-1)^{k-p-1} \binom{n}{k-p-1}. \quad (4-19)$$

If  $p \neq 0$ , let  $q = p - 1$ . By identity (4-19), the coefficient of  $(t + p)^{n-1}$  in (4-17) satisfies

$$\begin{aligned}
 L_p(G_{k+1}(t)) &= \sum_{i=p+1}^k \binom{k+1+n-i-1}{n-1} (-1)^{i-p} \binom{n}{i-p-1} \\
 &= \sum_{i=q+2}^k \binom{k+1+n-i-1}{n-1} (-1)^{i-q-1} \binom{n}{i-q-2} \\
 &= \sum_{i=q+1}^{k-1} \binom{k+n-i-1}{n-1} (-1)^{i-q} \binom{n}{i-q-1} \\
 &= (-1)^{k-q-1} \binom{n}{k-q-1} = (-1)^{k-p} \binom{n}{k-p}. \tag{4-20}
 \end{aligned}$$

Identity (4-20) holds for all integers  $p \in [1, k - 1]$ . It remains to consider the case when  $p = 0$ .

If  $p = 0$ , by (4-17), the coefficient of  $t^{n-1}$  in  $G_{k+1}(t)$  is

$$L_0(G_{k+1}(t)) = \sum_{i=1}^k \binom{k+1+n-i-1}{n-1} (-1)^i \binom{n}{i-1}. \tag{4-21}$$

Next, we show that  $L_0(G_{k+1}(t)) = (-1)^k \binom{n}{k}$ . Note that by the binomial theorem, the coefficient of the term  $x^{k+1}$  in  $(1+x)^{-n}(1+x)^n$  is

$$\begin{aligned}
 &\sum_{i=0}^k (-1)^i \binom{n+i-1}{i} \binom{n}{k-i} \\
 &= \sum_{i=0}^k (-1)^i \binom{n+i-1}{n-1} \binom{n}{k-i} \\
 &= \sum_{j=1}^{k+1} \binom{k+1+n-j-1}{n-1} (-1)^{k+1-j} \binom{n}{j-1} \quad (j = k+1-i) \\
 &= (-1)^{k+1} \left( L_0(G_{k+1}(t)) + (-1)^{k+1} \binom{n}{k} \right). \tag{4-22}
 \end{aligned}$$

On the other hand, for a nonnegative integer  $k$ , the coefficient of the term  $x^{k+1}$  in  $(1+x)^{-n}(1+x)^n = 1$  is always 0. Hence, (4-22) implies that

$$L_0(G_{k+1}(t)) = (-1)^k \binom{n}{k}. \tag{4-23}$$

Therefore, by identities (4-20) and (4-23), it follows that

$$\begin{aligned}
 G_{k+1}(t) &= (t+k)^{n-1} + \sum_{p=0}^{k-1} (-1)^{k-p} \binom{n}{k-p} (t+p)^{n-1} \\
 &= \sum_{i=1}^{k+1} (-1)^{i-1} (t+k+1-i)^{n-1} \binom{n}{i-1}.
 \end{aligned}
 \tag{4-24}$$

This concludes the proof. □

### 5. Application to log-gamma function

In this section, we consider the integral of log-gamma function

$$I = \int_0^1 \int_0^1 \dots \int_0^1 \log \Gamma(x_1 + x_2 + \dots + x_n) dx_1 dx_2 \dots dx_n. \tag{5-1}$$

The integral of log-gamma function has its own importance in many parts of mathematics [Amdeberhan et al. 2011; Choi and Srivastava 2005]. Actually, the case when  $n = 2$  is a problem proposed by Ovidiu Furdui [2010] in the Problems and Solutions section of *The College Mathematics Journal*, and one of its solutions is proposed by Geng-zhe Chang [2011]. When it comes to general dimension  $n$ , it is quite a challenge to evaluate it.

After the preparation of Theorem 4.1 in Section 4, we can evaluate the integral (5-1). A nice formula is given in Theorem 5.1.

**Theorem 5.1.**

$$\begin{aligned}
 I = I(n) &= \int_0^1 \int_0^1 \dots \int_0^1 \log \Gamma(x_1 + x_2 + \dots + x_n) dx_1 dx_2 \dots dx_n \\
 &= \frac{1}{2} \log(2\pi) - \frac{n-1}{2} H_n + \sum_{k=2}^{n-1} \frac{(-1)^{n+k+1} k^n}{n!} \binom{n-1}{k} \log k,
 \end{aligned}
 \tag{5-2}$$

where the last sum is missing when  $n = 2$  and  $H_n = \sum_{k=1}^n 1/k$ .

The proof of this theorem is based on Theorem 4.1 and several combinatorial identities in Jihuai Shi's book [2009].

Note that  $\Gamma(t+1) = t\Gamma(t)$  and  $G_m(t) = \sum_{i=1}^m (-1)^{i-1} (t+m-i)^{n-1} \binom{n}{i-1}$ . By [Theorem 4.1](#), the integral  $I$  becomes

$$\begin{aligned} I &= \frac{1}{(n-1)!} \sum_{m=1}^n \int_0^1 G_m(t) \log \Gamma(t+m-1) dt \\ &= \frac{1}{(n-1)!} \int_0^1 \sum_{m=1}^n G_m(t) \log \Gamma(t) dt \\ &\quad + \frac{1}{(n-1)!} \int_0^1 \sum_{k=2}^n \sum_{m=k}^n G_m(t) \log(t+k-2) dt. \quad (5-3) \end{aligned}$$

Several combinatorial identities are introduced to simplify [\(5-3\)](#).

**Lemma 5.2.**

$$\sum_{m=k}^n G_m(t) = (n-1)! - \sum_{m=1}^{k-1} \binom{n-1}{k-m-1} (-1)^{k-m-1} (t+m-1)^{n-1},$$

and when  $k=1$ ,  $\sum_{m=1}^n G_m(t) = (n-1)!$ .

*Proof.* Note that  $G_m(t) = \sum_{i=1}^m (-1)^{i-1} (t+m-i)^{n-1} \binom{n}{i-1}$ . It follows that

$$\begin{aligned} \sum_{m=1}^k G_m(t) &= \sum_{m=1}^k \sum_{i=1}^m (-1)^{i-1} (t+m-i)^{n-1} \binom{n}{i-1} \\ &= \sum_{m=1}^k \sum_{i=0}^{k-m} (-1)^i \binom{n}{i} (t+m-1)^{n-1}. \end{aligned}$$

By the combinatorial identity  $\sum_{i=0}^m (-1)^i \binom{n}{i} = (-1)^m \binom{n-1}{m}$  ( $m < n$ ), we have

$$\sum_{m=1}^k \sum_{i=0}^{k-m} (-1)^i \binom{n}{i} (t+m-1)^{n-1} = \sum_{m=1}^k \binom{n-1}{k-m} (-1)^{k-m} (t+m-1)^{n-1}.$$

Hence,

$$\sum_{m=1}^k G_m(t) = \sum_{m=1}^k \binom{n-1}{k-m} (-1)^{k-m} (t+m-1)^{n-1}.$$

In the case when  $k = n$ , the combinatorial identity  $\sum_{k=0}^n (-1)^k \binom{n}{k} (x+n-k)^n = n!$  implies

$$\begin{aligned} \sum_{m=1}^n G_m(t) &= \sum_{m=1}^n \binom{n-1}{n-m} (-1)^{n-m} (t+m-1)^{n-1} \\ &= \sum_{k=0}^{n-1} \binom{n-1}{k} (-1)^k (t+n-1-k)^{n-1} \\ &= (n-1)! . \end{aligned}$$

Therefore,

$$\begin{aligned} \sum_{m=k}^n G_m(t) &= \sum_{m=1}^n G_m(t) - \sum_{m=1}^{k-1} G_m(t) \\ &= (n-1)! - \sum_{m=1}^{k-1} \binom{n-1}{k-m-1} (-1)^{k-m-1} (t+m-1)^{n-1} . \quad \square \end{aligned}$$

Let

$$T_k = \sum_{m=1}^k \binom{n-1}{k-m} (-1)^{k-m} (t+m-1)^{n-1} = \sum_{m=0}^{k-1} \binom{n-1}{m} (-1)^m (t+k-m-1)^{n-1} .$$

Then

$$\sum_{m=k}^n G_m(t) = (n-1)! - T_{k-1} .$$

By applying Lemma 5.2, (5-3) becomes

$$\begin{aligned} I &= \int_0^1 \log \Gamma(t) dt + \int_0^1 \sum_{k=0}^{n-2} \log(t+k) dt - \frac{1}{(n-1)!} \int_0^1 \sum_{k=1}^{n-1} T_k \log(t+k-1) dt \\ &= \frac{1}{2} \log(2\pi) + (n-1) \log(n-1) - n + 1 - \frac{1}{(n-1)!} \int_0^1 \sum_{k=1}^{n-1} T_k \log(t+k-1) dt . \end{aligned} \tag{5-4}$$

Then, the calculation of  $I$  reduces to the calculation of

$$\int_0^1 \sum_{k=1}^{n-1} T_k \log(t+k-1) dt .$$

Note that  $T_1 = t^{n-1}$  and

$$\begin{aligned} \int_0^1 T_k \log(t+k-1) dt \\ = \sum_{m=0}^{k-1} \binom{n-1}{m} (-1)^m \int_0^1 (t+k-m-1)^{n-1} \log(t+k-1) dt. \end{aligned}$$

When  $k > 1$ ,

$$\begin{aligned} \int_0^1 (t+k-m-1)^{n-1} \log(t+k-1) dt \\ = \frac{(k-m)^n \log k - (k-m-1)^n \log(k-1)}{n} - \int_0^1 \frac{(t+k-m-1)^n}{n(t+k-1)} dt \\ = \frac{(k-m)^n - (-m)^n}{n} \log k - \frac{(k-m-1)^n - (-m)^n}{n} \\ - \frac{1}{n} \sum_{r=1}^n \frac{k^r - (k-1)^r}{r} \binom{n}{r} (-m)^{n-r}. \end{aligned}$$

Let  $S_1(1) = 0$ ,

$$S_1(k) = \sum_{m=0}^{k-1} \binom{n-1}{m} (-1)^m \left( \frac{(k-m)^n - (-m)^n}{n} \log k - \frac{(k-m-1)^n - (-m)^n}{n} \log(k-1) \right),$$

and

$$S_2(k) = \frac{1}{n} \sum_{m=0}^{k-1} \binom{n-1}{m} (-1)^m \sum_{r=1}^n \frac{k^r - (k-1)^r}{r} \binom{n}{r} (-m)^{n-r}.$$

It follows that

$$\int_0^1 \sum_{k=1}^{n-1} T_k \log(t+k-1) dt = \sum_{k=1}^{n-1} S_1(k) - \sum_{k=1}^{n-1} S_2(k). \quad (5-5)$$

The next lemma calculates  $\sum_{k=1}^{n-1} S_1(k)$ .

**Lemma 5.3.**

$$\sum_{k=1}^{n-1} S_1(k) = \frac{1}{n} \sum_{k=2}^{n-2} \binom{n-1}{k} (-1)^k (-k)^n \log k + \frac{\log(n-1)}{n} (n!(n-1) - (n-1)^n).$$

*Proof.* Note that  $S_1(1) = 0$ .

$$\begin{aligned} & \sum_{k=1}^{n-1} S_1(k) \\ &= \sum_{k=1}^{n-1} \sum_{m=0}^{k-1} \binom{n-1}{m} (-1)^m \left( \frac{(k-m)^n - (-m)^n}{n} \log k - \frac{(k-m-1)^n - (-m)^n}{n} \log(k-1) \right) \\ &= \frac{1}{n} \sum_{k=2}^{n-2} \binom{n-1}{k} (-1)^k (-k)^n \log k \\ & \quad + \frac{1}{n} \sum_{m=0}^{n-2} \binom{n-1}{m} (-1)^m ((n-m-1)^n - (-m)^n) \log(n-1). \end{aligned}$$

Using the combinatorial identity  $\sum_{k=0}^n \binom{n}{k} (-1)^k (x-k)^{n+1} = (x-n/2)(n+1)!$ , we have

$$\sum_{m=0}^{n-2} \binom{n-1}{m} (-1)^m (n-1-m)^n = \sum_{m=0}^{n-1} \binom{n-1}{m} (-1)^m (n-1-m)^n = \frac{n-1}{2} n!,$$

and

$$\begin{aligned} & \sum_{m=0}^{n-2} \binom{n-1}{m} (-1)^m (-m)^n \\ &= \sum_{m=0}^{n-1} \binom{n-1}{m} (-1)^m (-m)^n - (-1)^{n-1} (1-n)^n = (n-1)^n - \frac{n-1}{2} n!. \end{aligned}$$

Hence,

$$\sum_{k=1}^{n-1} S_1(k) = \frac{1}{n} \sum_{k=2}^{n-2} \binom{n-1}{k} (-1)^k (-k)^n \log k + \frac{\log(n-1)}{n} (n!(n-1) - (n-1)^n).$$

□

The following lemma calculates  $\sum_{k=1}^{n-1} S_2(k)$ . Here we only give the result. For reader's convenience, the proof of it is given in the [Appendix](#).

**Lemma 5.4.**

$$\sum_{k=1}^{n-1} S_2(k) = (n-1)!(n-1) - \frac{n-1}{2} H_n(n-1)!,$$

where  $H_n = \sum_{k=1}^n 1/k$ .

Using [Lemma 5.3](#) and [Lemma 5.4](#), we can prove [Theorem 5.1](#) below.

*Proof of Theorem 5.1.* Let  $H_n = \sum_{k=1}^n 1/k$ . By identity (5-5), Lemma 5.3 and Lemma 5.4, we have that

$$\begin{aligned}
 & \int_0^1 \sum_{k=1}^{n-1} T_k \log(t+k-1) dt \\
 &= \sum_{k=1}^{n-1} S_1(k) - \sum_{k=1}^{n-1} S_2(k) \\
 &= \frac{1}{n} \sum_{k=2}^{n-2} \binom{n-1}{k} (-1)^k (-k)^n \log k + \frac{\log(n-1)}{n} (n!(n-1) - (n-1)^n) \\
 & \quad - (n-1)!(n-1) + \frac{n-1}{2} H_n (n-1)!. \quad (5-6)
 \end{aligned}$$

By identities (5-4) and (5-6), it follows that

$$\begin{aligned}
 I &= \frac{1}{2} \log(2\pi) + (n-1) \log(n-1) - n + 1 - \frac{1}{(n-1)!} \int_0^1 \sum_{k=1}^{n-1} T_k \log(t+k-1) dt \\
 &= \frac{1}{2} \log(2\pi) - \frac{n-1}{2} H_n + \frac{1}{n!} \sum_{k=2}^{n-1} \binom{n-1}{k} (-1)^{k+n+1} k^n \log k. \quad \square
 \end{aligned}$$

When  $n = 2, 3$  and  $4$ , the values of the integral  $I$  are

$$\begin{aligned}
 I(2) &= -\frac{3}{4} + \frac{1}{2} \log(2\pi), \\
 I(3) &= \frac{1}{2} \log(2\pi) + \frac{4}{3} \log 2 - \frac{11}{6}, \\
 I(4) &= \frac{1}{2} \log(2\pi) - 2 \log 2 + \frac{27}{8} \log 3 - \frac{25}{8}.
 \end{aligned}$$

### Appendix.

For reader's convenience, the proof of Lemma 5.4 is given here.

#### Lemma 5.4.

$$\sum_{k=1}^{n-1} S_2(k) = (n-1)!(n-1) - \frac{n-1}{2} H_n (n-1)!,$$

where  $H_n = \sum_{k=1}^n 1/k$ .



*Proof.* Note that

$$\begin{aligned} \sum_{k=1}^{n-1} S_2(k) &= \frac{1}{n} \sum_{k=1}^{n-1} \left( \sum_{m=0}^{k-1} \binom{n-1}{m} \right) (-1)^m \sum_{r=1}^n \frac{k^r - (k-1)^r}{r} \binom{n}{r} (-m)^{n-r} \\ &= -\frac{1}{n} \sum_{k=1}^{n-2} \binom{n-1}{k} (-1)^k (-k)^n \sum_{r=1}^n \binom{n}{r} \frac{(-1)^r}{r} \\ &\quad + \frac{1}{n} \sum_{m=0}^{n-2} \binom{n-1}{m} (-1)^m \sum_{r=1}^n \frac{(-m)^{n-r} (n-1)^r}{r} \binom{n}{r}. \end{aligned}$$

Let

$$R_1 = -\frac{1}{n} \sum_{k=1}^{n-2} \binom{n-1}{k} (-1)^k (-k)^n \sum_{r=1}^n \binom{n}{r} \frac{(-1)^r}{r} \tag{A-1}$$

and

$$R_2 = \frac{1}{n} \sum_{m=0}^{n-2} \binom{n-1}{m} (-1)^m \sum_{r=1}^n \frac{(-m)^{n-r} (n-1)^r}{r} \binom{n}{r}. \tag{A-2}$$

Then

$$\sum_{k=1}^{n-1} S_2(k) = R_1 + R_2. \tag{A-3}$$

By applying the combinatorial identities

$$\sum_{k=0}^n \binom{n}{k} (-1)^k (x-k)^{n+1} = \left(x - \frac{n}{2}\right) (n+1)! \quad \text{and} \quad \sum_{k=1}^n \frac{(-1)^{k+1}}{k} \binom{n}{k} = H_n,$$

the sum  $R_1$  can be simplified to

$$\begin{aligned} R_1 &= \frac{1}{n} \sum_{k=1}^{n-2} \binom{n-1}{k} (-1)^k (-k)^n \sum_{r=1}^n \binom{n}{r} \frac{(-1)^{r+1}}{r} \\ &= \frac{H_n}{n} \left( (n-1)^n - \frac{n-1}{2} n! \right). \end{aligned} \tag{A-4}$$

To simplify  $R_2$ , we apply the combinatorial identity

$$\sum_{k=1}^n \frac{(-1)^{k+1}}{k} \binom{n}{k} (1 - (1-x)^k) = \sum_{k=1}^n \frac{x^k}{k},$$

and it follows that

$$\begin{aligned}
 \sum_{r=1}^n \frac{(-m)^{n-r} (n-1)^r}{r} \binom{n}{r} &= -(-m)^n \sum_{r=1}^n \frac{(-1)^{r+1}}{r} \binom{n}{r} \left(1 - \frac{m-n+1}{m}\right)^r \\
 &= (-m)^n \left( \sum_{r=1}^n \frac{1}{r} \left(\frac{m-n+1}{m}\right)^r - \sum_{r=1}^n \frac{(-1)^{r+1}}{r} \binom{n}{r} \right) \\
 &= \sum_{r=1}^n \frac{1}{r} (m-n+1)^r m^{n-r} (-1)^n - (-m)^n H_n.
 \end{aligned}$$

Recalling the formula of  $R_2$  in (A-2), we have

$$nR_2 = \sum_{m=0}^{n-2} \binom{n-1}{m} (-1)^{m+n} \sum_{k=1}^n \frac{1}{k} (m-n+1)^k m^{n-k} - H_n \sum_{m=0}^{n-2} \binom{n-1}{m} (-1)^m (-m)^n. \quad (\text{A-5})$$

By the combinatorial identity  $\sum_{k=0}^n \binom{n}{k} (-1)^k (x-k)^{n+1} = (x-n/2)(n+1)!$ , we see that

$$\sum_{m=0}^{n-2} \binom{n-1}{m} (-1)^m (-m)^n = (n-1)^n - \frac{n-1}{2} n!. \quad (\text{A-6})$$

We then simplify  $\sum_{m=0}^{n-2} \binom{n-1}{m} (-1)^{m+n} \sum_{k=1}^n \frac{1}{k} (m-n+1)^k m^{n-k}$ . Note that

$$\begin{aligned}
 &\sum_{m=0}^{n-2} \binom{n-1}{m} (-1)^{m+n} \sum_{k=1}^n \frac{1}{k} (m-n+1)^k m^{n-k} \\
 &= \sum_{k=1}^n \frac{(-1)^n}{k} \left( \sum_{m=0}^{n-2} \binom{n-1}{m} (-1)^m \sum_{i=0}^k \binom{k}{i} m^{n-k+i} (n-1)^{k-i} (-1)^{k-i} \right). \quad (\text{A-7})
 \end{aligned}$$

Let

$$P(m) = \sum_{i=0}^k \binom{k}{i} m^{n-k+i} (n-1)^{k-i} (-1)^{k-i}.$$

We apply the combinatorial identity  $\sum_{k=0}^n (-1)^k \binom{n}{k} \mathbf{P}(k) = 0$  for any polynomial  $\mathbf{P}(k)$  with  $\deg \mathbf{P}(k) < n$ , and it follows that

$$\begin{aligned} \sum_{m=0}^{n-2} \binom{n-1}{m} (-1)^m \sum_{i=0}^k \binom{k}{i} m^{n-k+i} (n-1)^{k-i} (-1)^{k-i} \\ = \sum_{m=0}^{n-2} \binom{n-1}{m} (-1)^m P(m) \\ = \sum_{m=0}^{n-1} \binom{n-1}{m} (-1)^m P(m) - (-1)^{n-1} P(n-1) \\ = \sum_{m=0}^{n-1} \binom{n-1}{m} (-1)^m (-k(n-1)m^{n-1} + m^n). \end{aligned} \tag{A-8}$$

By the combinatorial identity  $\sum_{k=0}^n (-1)^k \binom{n}{k} (x+n-k)^n = n!$ , we have

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

By the combinatorial identity  $\sum_{k=0}^n \binom{n}{k} (-1)^k (x-k)^{n+1} = (x-n/2)(n+1)!$ , we see that

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

Then equality (A-7) becomes

$$\begin{aligned} \sum_{m=0}^{n-2} \binom{n-1}{m} (-1)^{m+n} \sum_{k=1}^n \frac{1}{k} (m-n+1)^k m^{n-k} \\ = \sum_{k=1}^n \frac{(-1)^n}{k} \left( k(n-1)(-1)^n (n-1)! + (-1)^{n-1} \frac{n-1}{2} n! \right) \\ = n!(n-1) - \frac{n-1}{2} n! H_n, \end{aligned} \tag{A-9}$$

where  $H_n = \sum_{k=1}^n 1/k$ .

Hence, by equalities (A-9) and (A-6),  $nR_2$  in (A-5) can be simplified to

$$nR_2 = n!(n-1) - (n-1)^n H_n. \tag{A-10}$$

That is,

$$R_2 = (n-1)!(n-1) - \frac{H_n}{n} (n-1)^n. \tag{A-11}$$

Therefore, by equalities (A-3), (A-4) and (A-11), it follows that

$$\begin{aligned} \sum_{k=1}^{n-1} S_2(k) &= R_1 + R_2 \\ &= \frac{H_n}{n} \left( (n-1)^n - \frac{n-1}{2} n! \right) + (n-1)!(n-1) - \frac{H_n}{n} (n-1)^n \\ &= (n-1)!(n-1) - \frac{n-1}{2} H_n (n-1)!, \end{aligned}$$

where  $H_n = \sum_{i=1}^n 1/i$ . □

### Acknowledgements

Duokui Yan and Geng-zhe Chang want to express their gratitude to the Department of Mathematics at Brigham Young University for its help and support. We sincerely thank Professor Tiancheng Ouyang for his invitation. All of the authors are greatly indebted to the reviewer for his/her helpful suggestions.

### References

- [Amdeberhan et al. 2011] T. Amdeberhan, M. W. Coffey, O. Espinosa, C. Koutschan, D. V. Manna, and V. H. Moll, “Integrals of powers of loggamma”, *Proc. Amer. Math. Soc.* **139**:2 (2011), 535–545. [MR 2011k:33001](#) [Zbl 1213.33002](#)
- [Chang 2011] G. Chang, “On a double integral with Gamma function as integrant”, *Studies in College Mathematics* **14**:2 (2011), 1–2. In Chinese.
- [Chang and Shi 2003] G. Chang and J. Shi, *A course of mathematical analysis*, Higher Education Press, Beijing, 2003. In Chinese.
- [Choi and Srivastava 2005] J. Choi and H. M. Srivastava, “A family of log-gamma integrals and associated results”, *J. Math. Anal. Appl.* **303**:2 (2005), 436–449. [MR 2005k:11185](#) [Zbl 1064.33003](#)
- [Furdui 2010] O. Furdui, “Problems and solutions: Problem 904”, *The College Mathematics Journal* **41**:3 (2010), 245–246.
- [Shi 2009] J. Shi, *Combinatorial identities*, University of Science and Technology of China Press, Hefei, 2009. In Chinese.

Received: 2014-04-21

Revised: 2014-07-27

Accepted: 2014-07-28

[duokuiyan@buaa.edu.cn](mailto:duokuiyan@buaa.edu.cn)

*School of Mathematics and System Science,  
Beihang University, Beijing 100091, China*

[lewis\\_liou@smss.buaa.edu.cn](mailto:lewis_liou@smss.buaa.edu.cn)

*School of Mathematics and System Science,  
Beihang University, Beijing 100191, China*

[changgz@ustc.edu.cn](mailto:changgz@ustc.edu.cn)

*School of Mathematical Science, University of Science  
and Technology of China, Hefei 230026, China*

# Knight's tours on boards with odd dimensions

Baoyue Bi, Steve Butler, Stephanie DeGraaf and Elizabeth Doebel

(Communicated by Kenneth S. Berenhaut)

A closed knight's tour of a board consists of a sequence of knight moves, where each square is visited exactly once and the sequence begins and ends with the same square. For boards of size  $m \times n$  where  $m$  and  $n$  are odd, a tour is impossible because there are unequal numbers of white and black squares. By deleting a square, we can fix this disparity, and we determine which square to remove to allow for a closed knight's tour.

## 1. Introduction

One popular form of recreational mathematics deals with chess problems [Elkies and Stanley 2003]. While these problems can take many different forms (e.g., placing nonattacking queens or solving endgames), one of the most well-known variations is the knight's tour. In chess, a knight can move in a very restricted way. Namely, it must move one unit in one direction and two units in the perpendicular direction (see Figure 1).

A *knight's tour* is a sequence of legal knight moves where each square on the board is visited once; further, a *closed knight's tour* has the additional condition that it begins and ends with the same square. The problem of determining when a board has a closed knight's tour dates back several hundred years (see for example the work of Euler [1759]), and a full solution using a simple inductive argument was given by Schwenk.

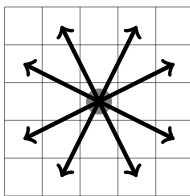
**Theorem 1** [Schwenk 1991]. *For  $m \leq n$ , an  $m \times n$  rectangular board has a closed knight's tour unless one of the three following conditions hold:*

- (1)  $mn$  is odd.
- (2)  $m \in \{1, 2, 4\}$ .
- (3)  $m = 3$  and  $n \in \{4, 6, 8\}$ .

---

MSC2010: primary 05C45; secondary 00A09.

Keywords: knight's tour, expanders, chess boards.



**Figure 1.** Legal knight moves.

Variations of this result have been studied including looking at closed knight's tours on a torus [Watkins and Hoenigman 1997], cylinders [Watkins 2000], spheres [Cairns 2002], and other boards [Lam et al. 1999].

When a knight moves on an  $m \times n$  board, it will alternate between squares which are white and black. When we add in the requirement that we must start and stop at the same square this means that we must take an even number of steps in a closed tour (i.e., to return to our original colored square). However there are  $mn$  steps needed to cover the  $m \times n$  board, and this establishes the first condition of Theorem 1. However, by deleting one square it is possible to leave an equal number of white and black squares on the board opening up the possibility of having a closed knight's tour. This leads to the following pair of questions:

**Question.** *Let  $m, n$  be odd with  $m, n \geq 3$ . Given an  $m \times n$  board, when is it possible to delete one square so that the remaining board has a closed knight's tour? When it is possible to delete a square, which square(s) can we delete?*

An answer to the first question was given by DeMaio and Hippchen [2009] who showed that it is always possible except for the  $3 \times 5$  board. The purpose of this paper is to give an answer to the second question, namely which squares can be deleted when it is possible, which we summarize in the theorem below.

For convention, we will label the squares of the board  $(i, j)$  as we would a matrix, i.e.,  $1 \leq i \leq m$  indicates the row going from top to bottom while  $1 \leq j \leq n$  indicates the column going from left to right. With this labeling we note that a knight move will go from  $(i, j)$  to  $(k, \ell)$ , where  $i + j$  and  $k + \ell$  have different parity. Since there is one more square with  $i + j$  even than there is with  $i + j$  odd, in order for a knight's tour to exist, a necessary condition is that we must delete a square with  $i + j$  even.

**Theorem 2.** *Let  $m, n$  be odd with  $3 \leq m \leq n$ . Then we can delete the square  $(i, j)$  from the  $m \times n$  board and have a closed knight's tour in the remaining board for the following situations:*

- (1) *For the  $3 \times 3$  board,  $(i, j) = (2, 2)$ .*
- (2) *For the  $3 \times 5$  board, there is no single square which can be deleted.*
- (3) *For the  $3 \times 7$  board,  $(i, j) \in \{(2, 2), (2, 6)\}$ .*

- (4) For the  $3 \times 9$  board,  $(i, j) \in \{(1, 1), (1, 5), (1, 9), (3, 1), (3, 5), (3, 9)\}$ .
- (5) For the  $3 \times n$  board with  $n \geq 11$ ,  $i + j$  is even and  $j \notin \{3, 4, n - 3, n - 2\}$ .
- (6) For the  $5 \times 5$  board,  $(i, j) \in \{(1, 1), (1, 5), (5, 1), (5, 5)\}$ .
- (7) For  $m \geq 5$  and  $n \geq 7$ ,  $i + j$  is even.

The problem of which squares can be deleted from a  $3 \times n$  board and having a knight's tour on the remaining board was independently done by Miller and Farnsworth [2013]. We include those results here for completeness and also because the proof of Miller and Farnsworth overlooked the case of removing the  $(2, 8)$  square from the  $3 \times 15$  board.

The rest of this paper is organized as follows. In Section 2 we introduce a method that allows us to expand a closed knight's tour from a smaller board to a larger board. In Sections 3, 4, and 5 we handle the cases of  $3 \times (\text{odd})$ ,  $5 \times (\text{odd})$ , and finally, the remaining cases. Lastly, in Section 6, we give some concluding remarks.

In the remainder of the paper we will make extensive use of symmetry, i.e., if we rotate a board by  $90^\circ$  or take a mirror image, we will still have a closed knight's tour.

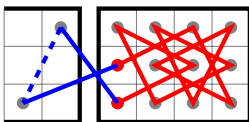
## 2. Gluing on expanders

Our general approach mirrors that which was given in [Schwenk 1991]. Namely, we will form a large collection of base cases and show how to expand these base cases to get the remaining results. Our base cases have been relegated to the appendices, while in this section, we will show how we can expand a board.

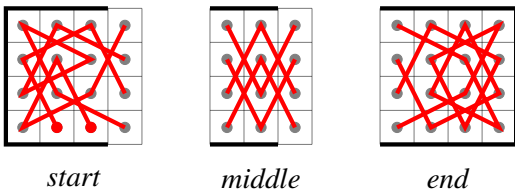
Our tool of choice will be  $m \times p$  expanders which correspond to open knight's tours of the  $m \times p$  board that start at  $(2, 1)$  and end at  $(3, 1)$ . This type of board can be easily connected to corners (since the moves at corners are forced). The following shows how to take a closed knight's tour that uses all or part of a board (i.e., a sub-board) and extend the board in one direction.

**Lemma 3.** *Given a closed knight's tour on a sub-board of the  $m \times n$  board which visits the square  $(1, n)$  and an  $m \times p$  expander, we can find a closed knight's tour on the  $m \times (n + p)$  board which, when restricted to the first  $n$  columns, covers the same sub-board as the original  $m \times n$  board.*

*Proof.* By assumption, our tour visits the  $(1, n)$  square. Therefore, we know that one move on the knight's tour is from  $(1, n)$  to  $(3, n - 1)$ . Deleting this move will result in an open knight's tour that starts at  $(1, n)$  and ends at  $(3, n - 1)$ . Now sequentially place the two boards, first placing the  $m \times n$  board and then the  $m \times p$  expander. Note that the expander is now an open tour that starts at  $(2, n + 1)$  and ends at  $(3, n + 1)$ . Finally, we combine these two open tours to form one single closed tour that visits every square by adding the moves  $(1, n)$  to  $(3, n + 1)$  and  $(3, n - 1)$  to  $(2, n + 1)$ . By construction this will cover the same sub-board as the original  $m \times n$  board.  $\square$



**Figure 2.** An illustration of Lemma 3 for a  $3 \times 4$  expander.



**Figure 3.** The three building blocks to form expanders.

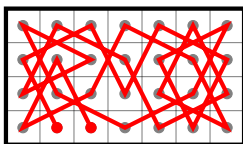
An illustration of Lemma 3 which has a  $3 \times 4$  expander is shown in Figure 2. Note by symmetry that we can also use other corners to glue. Since we will only be deleting one square from the board, we will always have at least one corner on a side available to use. We note that DeMaio and Hippchen [2009] used a similar gluing in their approach.

Following Schwenk, we want to be able to add four rows or columns to boards, which means we want to show that  $n \times 4$  expanders exist. Unfortunately, they do not exist for all  $n$ . However, we will show that they exist when  $n \geq 7$  and is odd. This will be done by appropriately combining the three pieces shown in Figure 3 (where for convenience we have rotated by  $90^\circ$ ).

**Proposition 4.** *A  $n \times 4$  expander exists for odd values  $n \geq 7$ .*

*Proof.* We will use the pieces given above along with induction to show how to do this. First note that these pieces are designed to overlap in a column, so if we take the *start* and *end* together we get the  $7 \times 4$  expander shown in Figure 4.

To finish the proof it suffices to show how we can take an expander and increase its width by 2; i.e., given that we have  $n \times 4$ , we can construct  $(n + 2) \times 4$ . To do this we move the *end* piece over by two spots and in the gap insert a *middle*. For example, for  $n = 9$  and  $n = 11$ , we now get the expanders shown in Figure 5.



**Figure 4.** A  $4 \times 7$  expander.





**Figure 5.**  $4 \times 9$  and  $4 \times 11$  expanders.

Because of the format of the pieces, as we glue these pieces together, we will have degree two at each vertex except for the two special vertices coming from the *start* piece. To show that this is a valid expander, we only need to make sure that we have an open knight's tour (i.e., we visit every square once and we begin and end in different squares). The key to see why this holds is to note that for the *middle* piece we have the relationship shown in [Figure 6](#)

This indicates that the relative ordering of the four “tracks” is the same. In particular, the addition of the *middle* piece will not effect whether or not we have an open knight's tour outside of that piece. But by induction, since we started with an open knight's tour, we still have an open knight's tour, and hence this construction gives a valid expander. □

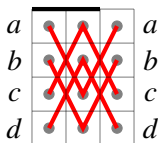
### 3. Closed tours on $3 \times (\text{odd})$ boards

In this section we will work through the cases of  $3 \times n$  for  $n$  odd. We will first look at what happens when  $n \leq 9$  where there are extra constraints on what can be deleted, and then we will establish the general case for  $n \geq 11$ .

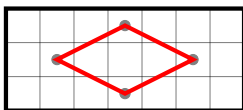
When  $n = 3$ , we note that there is no legal knight move from  $(2, 2)$  to another square. Thus, it cannot be involved in a tour, so it is the only square which can be deleted. Further, there is a closed knight's tour with this square deleted (in [Appendix A](#)), establishing the result.

When  $n = 5$ , each corner would have a move to  $(2, 3)$  and since we only delete one square, we would have to visit the center square multiple times, which is impossible for a closed knight's tour.

When  $n = 7$ , if we keep *both*  $(2, 2)$  and  $(2, 6)$ , then the moves shown in [Figure 7](#) (among others) would be forced to occur. This is impossible to extend to a closed knight's tour of the  $3 \times 7$  board as we already have a cycle just among these four



**Figure 6.** The relationship of the central pieces.



**Figure 7.** Forced moves for  $3 \times 7$ .

vertices. Therefore, we must delete either  $(2, 2)$  or  $(2, 6)$  (which up to symmetry are equivalent). Starting with the  $3 \times 3$  closed knight's tour in [Appendix A](#) and gluing on the  $3 \times 4$  expander as in [Lemma 3](#) to the left (or right) will give a  $3 \times 7$  closed knight's tour with  $(2, 6)$  (or  $(2, 2)$ ) deleted.

Before moving on to analyze the  $3 \times 9$  case, we will establish a general restriction about which square can be deleted.

**Lemma 5.** *It is not possible to construct a closed knight's tour on a  $3 \times n$  board,  $n$  odd, with a deleted square in column  $3, 4, n - 3$ , or  $n - 2$ .*

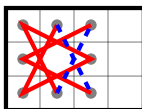
*Proof.* By symmetry it suffices to show that we cannot delete a square in columns 3 or 4. Further note that by parity, we only need to show that  $(1, 3)$ ,  $(3, 3)$  and  $(2, 4)$  cannot be deleted.

Note that to make a complete tour, each square must have an ingoing and outgoing move. This restriction forces the moves of several squares including  $(1, 1)$ ,  $(2, 1)$  and  $(3, 1)$ , as shown in [Figure 8](#) (assuming they have not been deleted).

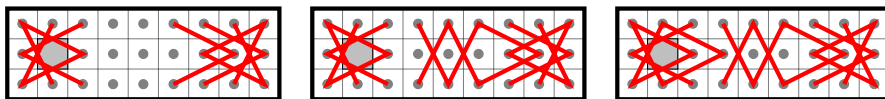
In particular, since  $(2, 1)$  cannot be deleted, both  $(1, 3)$  and  $(3, 3)$  need to be present to be able to connect to  $(2, 1)$ . Thus, we cannot delete a square in column 3.

If we delete  $(2, 4)$ , then the squares  $(1, 2)$  and  $(3, 2)$  *must* connect to  $(3, 3)$  and  $(1, 3)$  respectively. This then forces a small cycle (as shown in [Figure 8](#)) which we cannot then extend to a closed knight's tour. Therefore, we cannot delete  $(2, 4)$ .  $\square$

Applying [Lemma 5](#), we see that for the  $3 \times 9$  board, we cannot delete a square in columns 3, 4, 6, or 7. In [Appendix A](#), we give closed knight's tours for the cases when we delete  $(1, 9)$  and  $(1, 5)$  (which, by symmetry, give tours for when  $(1, 1)$ ,  $(3, 1)$ ,  $(3, 9)$  or  $(3, 5)$  are deleted). It remains to show that we cannot delete  $(2, 2)$ . This is done by examining forced moves. The process is illustrated in [Figure 9](#). First we add in all moves which are forced (near the ends). After this is done, we note that each of the squares  $(1, 5)$  and  $(3, 5)$  only have two possible moves available to them, so their moves are also forced. Finally, this leaves  $(2, 4)$  with



**Figure 8.** The forced moves from the left-hand column of a  $3 \times n$  board.



**Figure 9.** Forced moves for the  $3 \times 9$  board.

only two available moves and so those moves are also forced. But we are now left with a closed cycle that does not cover the entire board and so we cannot extend this to a closed knight's tour.

We are now ready to establish a general result for larger  $3 \times n$  boards.

**Theorem 6.** *Suppose we have a  $3 \times n$  board with  $n \geq 11$  and odd. Then a closed knight's tour is possible on the board after removing the square  $(i, j)$  if and only if  $i + j$  is even and  $j \notin \{3, 4, n - 3, n - 2\}$ .*

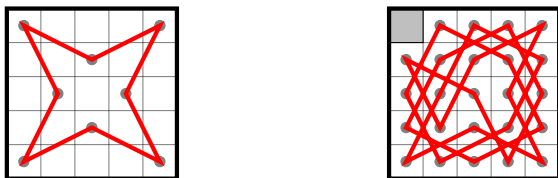
*Proof.* By [Lemma 5](#), we cannot delete a square in column 3, 4,  $n - 3$  or  $n - 2$ .

It remains to show that the deletion of every other square results in a board containing a closed knight's tour. For  $n = 11$ , we show in [Appendix A](#) closed knight's tours with squares  $(1, 1)$  and  $(1, 5)$  deleted (which by symmetry also gives  $(1, 11)$ ,  $(3, 1)$ ,  $(3, 11)$ ,  $(1, 7)$ ,  $(3, 5)$  and  $(3, 7)$ ). In addition, we can take the  $3 \times 3$  board and using [Lemma 3](#), glue on a  $3 \times 4$  expander either twice to the left, twice to the right, or once on each side, giving a closed knight's tours with squares  $(2, 10)$ ,  $(2, 2)$ , or  $(2, 6)$ , respectively, deleted.

For  $n = 13$ , we can use the known solutions for the  $3 \times 9$  board and use [Lemma 3](#) with the  $3 \times 4$  expander to get solutions for the  $3 \times 13$  board with a deleted square. Doing this we get everything except (up to symmetry) boards with squares  $(1, 7)$ ,  $(2, 6)$  or  $(2, 2)$  deleted. These boards are given in [Appendix A](#), establishing this case.

Now assume the result holds true for  $3 \times n$ . Then by taking the collection of closed knight's tours and applying [Lemma 3](#) with a  $3 \times 4$  expander on the left, we will get every closed knight's tour for the  $3 \times (n + 4)$  board which does not have a deleted square in column 1, 2, 3, 4, 7, 8,  $n + 1$ , or  $n + 2$ . Similarly, if we apply [Lemma 3](#) with a  $3 \times 4$  expander on the right, we will get every closed knight's tour for the  $3 \times (n + 4)$  board which does not have a deleted square in column 3, 4,  $n - 3$ ,  $n - 2$ ,  $n + 1$ ,  $n + 2$ ,  $n + 3$ , or  $n + 4$ . The intersection of these sets of columns will contain the mutually common columns 3, 4,  $n + 1$ , and  $n + 2$ . It might also contain additional term(s) if  $\{7, 8\} \cap \{n - 3, n - 2\}$  is nonempty. Because  $n \geq 11$  by assumption, this can only occur when  $n = 11$  and the common column is 8, giving that for  $n \geq 13$ , the intersection is  $\{3, 4, n + 1, n + 2\}$  and  $\{3, 4, 8, 12, 13\}$  if  $n = 11$ .

Therefore, we can get all solutions by building off of the base cases, except for the case when we have a  $3 \times 15$  board and we delete the square  $(2, 8)$ . In [Appendix A](#) we show a closed knight's tour for such a board, and therefore we can construct all such boards.  $\square$



**Figure 10.** Knight’s tour on the  $5 \times 5$  board.

### 4. Closed tours on $5 \times (\text{odd})$ boards

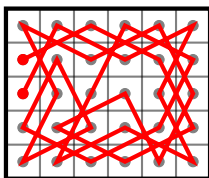
In this section we will work through the cases of  $5 \times n$ . We first handle the exceptional case of  $5 \times 5$  by noting that if we do not delete one of the corner squares and then we draw in the forced moves, we get the board shown on the left in [Figure 10](#). This board has a closed cycle, so we will not be able to form a closed knight’s tour. Therefore, we must delete a corner, and by symmetry, we can delete any corner. On the right in [Figure 10](#) we have given a closed knight’s tour with  $(1, 1)$  deleted.

The remaining cases are handled in the following theorem which makes use of the  $5 \times 6$  expander given in [Figure 11](#).

**Theorem 7.** *Given any  $5 \times n$  board where  $n \geq 7$  is odd, a closed knight’s tour exists after deleting  $(i, j)$  if and only if  $i + j$  is even.*

*Proof.* In [Appendix B](#) we have given a knight’s tour for any appropriate deleted square (up to symmetry) for the  $5 \times 7$ ,  $5 \times 9$  and  $5 \times 11$  boards.

Now suppose we have a  $5 \times n$  board with  $n \geq 13$  and a square  $(i, j)$  with  $i + j$  even. Then we show how to form a closed knight’s tour for this board. First we note that on either the left or the right of the deleted square, there are six full columns in the board. So we repeatedly pull off sets of six columns from one side or the other of the deleted square *until* we have a  $5 \times 7$ ,  $5 \times 9$  or  $5 \times 11$  board with a deleted square (which by construction will be at  $(i', j')$  with  $i' + j'$  even). We now take the closed knight’s tour for this board (which we have already found) and we repeatedly add back in the sets of six columns that we deleted by use of [Lemma 3](#) and the expander shown in [Figure 11](#). The end result will be our desired closed knight’s tour.  $\square$



**Figure 11.** A  $5 \times 6$  expander.

The proof we have just given works by showing how to start with a large board and then showing how to reduce down to a base case which we know is true. An alternative proof approach would be to start with the base cases and then use the expanders in all possible ways to construct a collection of boards and then show that all of the desired boards are in our collection. The latter approach can work but we have opted for the first approach as it gives a simple constructive approach to building the boards. Namely take the board, reduce down to a base case which we know and then reverse the steps to build the desired board. Using the second approach, it is not obvious a priori which board to build off of or how to build up to a larger board; this is especially true for the final result in the next section.

## 5. Closed tours on larger boards

In this section we finish establishing the main result.

**Theorem 8.** *Given an  $m \times n$  board with  $m \leq n$ ,  $m \geq 5$  and  $n \geq 7$  and any square  $(i, j)$  with  $i + j$  even, there is a closed knight's tour of the  $m \times n$  board with  $(i, j)$  deleted.*

*Proof.* We will make use of the  $n \times 4$  expanders from [Proposition 4](#), for odd  $n \geq 7$ , to mimic the proof of the last theorem. By the previous theorem, we know the result holds if  $m = 5$ , so we can assume that  $m \geq 7$ . Further, in [Appendix C](#) we give (up to symmetry) closed knight's tours for the  $7 \times 7$  board. So we know the result also holds for  $m = n = 7$ .

Now, for any  $(i, j)$ , there are either four columns to the left or four columns to the right. We can pull off those four columns and consider the resulting smaller board. By [Lemma 3](#), it follows that if we have a closed knight's tour for this smaller board, we can use the expander to recover a closed knight's tour of our original board. (Note that we might possibly interchange the dimensions by rotating after pulling off these extra columns to maintain that  $m \leq n$ .)

In particular, after finitely many iterations (at most  $(m + n)/4$  since we can only repeat this at most  $m/4$  times for rows and at most  $n/4$  for columns) we will have shrunk the board down to either a  $5 \times n$  or a  $7 \times 7$ , in which case we have a solution. We now take this solution and work backwards to recover the desired original knight's tour.  $\square$

## 6. Conclusion

In this paper we have determined which squares can be deleted in a board with odd dimensions to allow the existence of a closed knight's tour. Reexamining Schwenk's result [\[1991\]](#), we note that there are no closed knight's tours of the  $4 \times n$  board for any  $n$ . DeMaio and Hippchen [\[2009\]](#) were able to show that there are closed tours that exist after deleting two squares (as long as  $n \geq 3$ ). In light of our discussion this raises the following natural question:

**Question.** For the  $4 \times n$  board with  $n \geq 3$ , which pairs of squares can be deleted that result in the existence of a closed knight's tour on the remaining board?

We note that there is the obvious restriction that there must be one square of each parity. There is also a more subtle constraint.

**Proposition 9.** If two squares in the  $4 \times n$  board are deleted and a closed knight's tour exists for the remaining board, then neither square could come from the middle two rows.

*Proof.* In the  $4 \times n$  board, if we have a closed knight's tour, then any move from the first or fourth row must go into the middle two rows. By orienting the tour, we can then create a one-to-one pairing between squares in the first and fourth rows with a subset of the squares in the middle two rows (i.e., by what square follows after in the order given by the tour). Therefore, we can not have deleted both squares from the middle two rows.

Similarly, if we have one square deleted from the middle two rows, then we deleted one square from the first or fourth rows. Therefore, in the closed knight's tour, squares alternate between being in the middle or not. But we also know that squares alternate between different parities, which would imply that the squares in the middle two rows are all the same parity. But this is impossible.  $\square$

This shows that we must delete our two squares from the first and fourth row. Yet, when  $n$  is small, this is not sufficient. However, computational evidence suggests the following.

**Conjecture.** Consider the  $4 \times n$  board with  $n \geq 7$ . For any pair of squares, with one of each parity and neither coming from the middle two rows, there is a closed knight's tour on the board that avoids only these two squares.

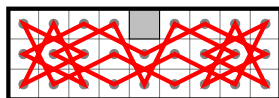
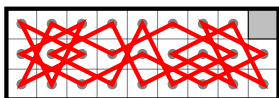
We look forward to seeing the next move in this area.

### Appendix A: Base cases for $3 \times$ (odd)

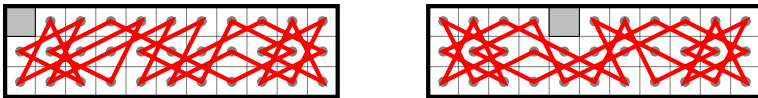
The following is the closed knight's tour of the  $3 \times 3$  board:



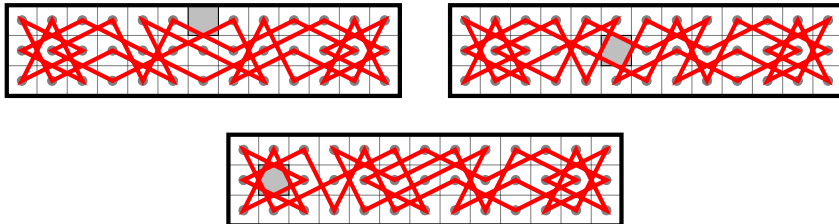
The following are closed knight's tours of the  $3 \times 9$  boards with  $(1, 9)$  and  $(1, 5)$ , respectively, deleted:



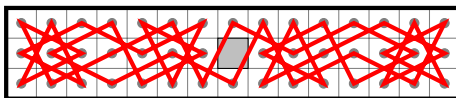
The following are closed knight's tours of the  $3 \times 11$  boards with  $(1, 1)$  and  $(1, 5)$ , respectively, deleted:



The following are closed knight's tours of the  $3 \times 13$  boards with  $(1, 7)$ ,  $(2, 6)$  and  $(2, 2)$ , respectively, deleted:

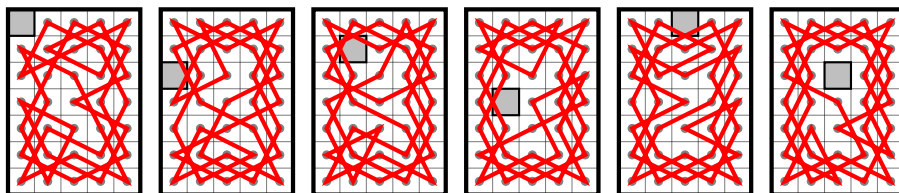


The following is a closed knight's tour of the  $3 \times 15$  board with  $(2, 8)$  deleted:

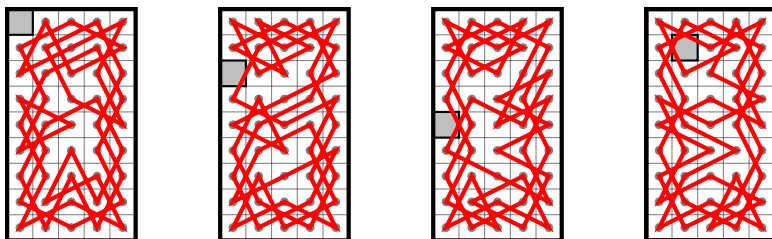


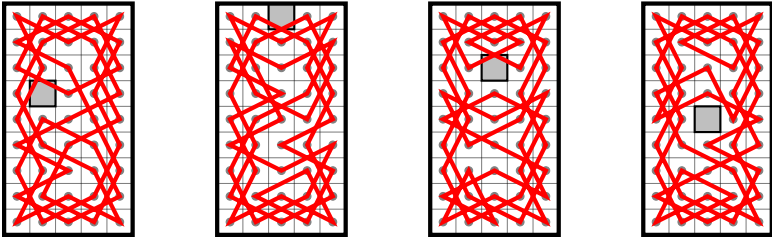
**Appendix B: Base cases for  $5 \times (\text{odd})$**

The following cover the cases (up to symmetry) for the  $5 \times 7$  board:

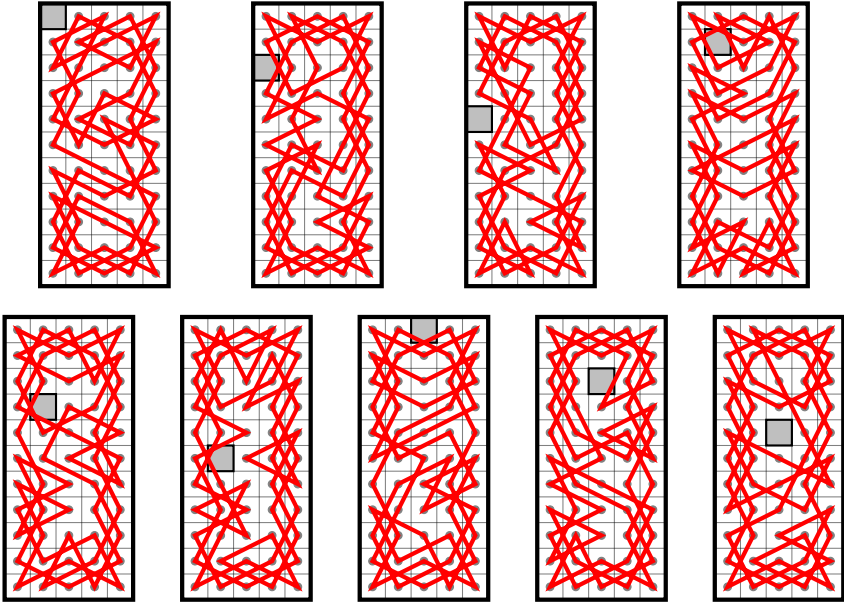


The following cover the cases (up to symmetry) for the  $5 \times 9$  board:



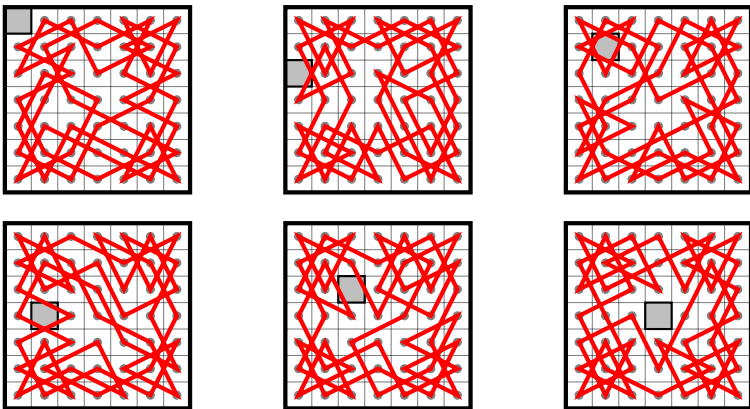


The following cover the cases (up to symmetry) for the  $5 \times 11$  board:



**Appendix C: Cases for  $7 \times 7$**

The following cover the cases (up to symmetry) for the  $7 \times 7$  board:





## References

- [Cairns 2002] G. Cairns, “Pillow chess”, *Math. Mag.* **75**:3 (2002), 173–186. [MR 2005b:91011](#)
- [DeMaio and Hippchen 2009] J. DeMaio and T. Hippchen, “Closed knight’s tours with minimal square removal for all rectangular boards”, *Math. Mag.* **82**:3 (2009), 219–225. [Zbl 1227.97064](#)
- [Elkies and Stanley 2003] N. D. Elkies and R. P. Stanley, “The mathematical knight”, *Math. Intelligencer* **25**:1 (2003), 22–34. [MR 2004c:05113](#)
- [Euler 1759] L. Euler, “Solution d’une question curieuse qui ne paroît soumise à aucune analyse”, *Mém. Acad. Roy. Sci. Belles Lett. (Berlin)* **15** (1759), 310–337. Reprinted in *Commentationes arithmeticae* **1** (1849), 337–355, and in *Commentationes algebraicae ad theoriam combinationum et probabilitatum pertinentes*, edited by L. G. Du Pasquier, Opera Omnia (1), **7** (1923), 26–56.
- [Lam et al. 1999] P. C. B. Lam, W. C. Shiu, and H. L. Cheng, “Knight’s tour on hexagonal nets”, pp. 73–82 in *Proceedings of the Thirtieth Southeastern International Conference on Combinatorics, Graph Theory, and Computing* (Boca Raton, FL, 1999), vol. 141, 1999. [MR 2000k:05176](#) [Zbl 0968.05050](#)
- [Miller and Farnsworth 2013] A. M. Miller and D. L. Farnsworth, “Knight’s tours on  $3 \times n$  chessboards with a single square removed”, *Open J. of Discrete Math.* **3**:1 (2013), 56–59.
- [Schwenk 1991] A. J. Schwenk, “Which rectangular chessboards have a knight’s tour?”, *Math. Mag.* **64**:5 (1991), 325–332. [MR 93c:05081](#) [Zbl 0761.05041](#)
- [Watkins 2000] J. J. Watkins, “Knight’s tours on cylinders and other surfaces”, pp. 117–127 in *Proceedings of the Thirty-first Southeastern International Conference on Combinatorics, Graph Theory and Computing* (Boca Raton, FL, 2000), vol. 143, 2000. [MR 2001k:05136](#) [Zbl 0977.05079](#)
- [Watkins and Hoenigman 1997] J. J. Watkins and R. L. Hoenigman, “Knight’s tours on a torus”, *Math. Mag.* **70**:3 (1997), 175–184. [MR 98i:00003](#) [Zbl 0906.05041](#)

Received: 2014-04-29

Revised: 2014-06-21

Accepted: 2014-08-02

[bby@iastate.edu](mailto:bby@iastate.edu)Department of Mathematics, Iowa State University,  
Ames, IA 50011, United States[butler@iastate.edu](mailto:butler@iastate.edu)Department of Mathematics, Iowa State University,  
Ames, IA 50011, United States[sdegraaf@iastate.edu](mailto:sdegraaf@iastate.edu)Department of Mathematics, Iowa State University,  
Ames, IA 50011, United States[edoebel@iastate.edu](mailto:edoebel@iastate.edu)Department of Mathematics, Iowa State University,  
Ames, IA 50011, United States



# Differentiation with respect to parameters of solutions of nonlocal boundary value problems for difference equations

Johnny Henderson and Xuewei Jiang

(Communicated by Kenneth S. Berenhaut)

For the  $n$ -th order difference equation,  $\Delta^n u = f(t, u, \Delta u, \dots, \Delta^{n-1}u, \lambda)$ , the solution of the boundary value problem satisfying  $\Delta^{i-1}u(t_0) = A_i$ ,  $1 \leq i \leq n-1$ , and  $u(t_1) - \sum_{j=1}^m a_j u(\tau_j) = A_n$ , where  $t_0, \tau_1, \dots, \tau_m, t_1 \in \mathbb{Z}$ ,  $t_0 < \dots < t_0 + n - 1 < \tau_1 < \dots < \tau_m < t_1$ , and  $a_1, \dots, a_m, A_1, \dots, A_n \in \mathbb{R}$ , is differentiated with respect to the parameter  $\lambda$ .

## 1. Introduction

With differences defined by  $\Delta u(t) = u(t+1) - u(t)$  and  $\Delta^i u(t) = \Delta(\Delta^{i-1}u(t))$  for  $i > 1$ , we will be concerned with solutions of the  $n$ -th order difference equation,

$$\Delta^n u = f(t, u, \Delta u, \dots, \Delta^{n-1}u, \lambda), \quad (1-1)$$

satisfying Dirichlet conditions

$$\Delta^{i-1}u(t_0) = A_i, \quad 1 \leq i \leq n-1, \quad (1-2)$$

and nonlocal boundary conditions

$$u(t_1) - \sum_{j=1}^m a_j u(\tau_j) = A_n, \quad (1-3)$$

where  $t_0, \tau_1, \dots, \tau_m \in \mathbb{Z}$ ,  $t_0 + n - 1 < \tau_1 < \dots < \tau_m < t_1$ ,  $A_i \in \mathbb{R}$ ,  $i = 1, \dots, n$ , and  $a_j \in \mathbb{R}$ ,  $j = 1, \dots, m$ .

Let  $\mathbb{Z}$ ,  $\mathbb{R}$ , and  $\mathbb{N}$  denote, respectively, the integers, the real numbers and the natural numbers. Given  $\emptyset \neq S \subseteq \mathbb{R}$ , let  $S_{\mathbb{Z}} := S \cap \mathbb{Z}$ . We assume throughout the paper that for (1-1):

*MSC2010:* primary 39A10, 34B08; secondary 34B10.

*Keywords:* difference equation, boundary value problem, nonlocal, differentiation with respect to parameters.

- (A)  $f(t, s_1, \dots, s_n, \lambda) : \mathbb{Z} \times \mathbb{R}^{n+1} \rightarrow \mathbb{R}$  is continuous.  
 (B)  $(\partial f / \partial s_i)(t, s_1, \dots, s_n, \lambda) : \mathbb{Z} \times \mathbb{R}^{n+1} \rightarrow \mathbb{R}$  is continuous for  $i = 1, \dots, n$ .  
 (C)  $(\partial f / \partial \lambda)(t, s_1, \dots, s_n, \lambda) : \mathbb{Z} \times \mathbb{R}^{n+1} \rightarrow \mathbb{R}$  is continuous.

Given a solution  $u(t)$  of (1-1), two linear equations playing fundamental roles for our results are the *variational equation along  $u(t)$*  given by

$$\Delta^n z = \sum_{i=1}^n \frac{\partial f}{\partial s_i}(t, u(t), \dots, \Delta^{n-1}u(t), \lambda) \Delta^{i-1} z, \quad (1-4)$$

and the corresponding nonhomogeneous equation along  $u(t)$  given by

$$\Delta^n z = \sum_{i=1}^n \frac{\partial f}{\partial s_i}(t, u(t), \dots, \Delta^{n-1}u(t), \lambda) \Delta^{i-1} z + \frac{\partial f}{\partial \lambda}(t, u(t), \dots, \Delta^{n-1}u(t), \lambda). \quad (1-5)$$

Our primary motivation arises from results by Henderson, Horn and Howard [Henderson et al. 1994] dealing with differentiation with respect to parameters for solutions of difference equations satisfying multipoint boundary conditions. Study of the relationship between a solution to a differential or difference equation and the associated variational equation can trace its origin to a result that Hartman [1982] attributed to Peano concerning differentiation of solutions of a differential equation with respect to initial conditions. Since then, these results have been extended and refined in various ways including boundary value problems for differential equations and difference equations [Datta 1998; Ehme and Henderson 1992; Henderson and Lee 1991; Spencer 1975]. Datta and Henderson [1992] did research on differentiation of solutions of difference equations with respect to boundary conditions. Benchohra et al. [2007] extended these results to nonlocal boundary value problems for second order difference equations. Also, interest in multipoint and nonlocal boundary value problems has grown significantly [Ashyralyev et al. 2004; Benchohra et al. 2007; Henderson et al. 2008; Lyons 2011]. Hopkins et al. [2009] proved a theorem about boundary data smoothness for solutions of nonlocal boundary value problems for second order difference equations. Then, Lyons [2014] generalized those results to  $n$ -th order difference equations.

Lyons [2014] has obtained extensive results for solutions of (1-1)–(1-3) when  $f$  is independent of  $\lambda$ . Our main results concern differentiation of solutions of (1-1)–(1-3) with respect to the parameter  $\lambda$ . Section 2 is devoted to results for initial value problems. We state theorems concerning solutions of initial value problems for (1-1) and their continuity and differentiability properties with respect to initial values and parameters. Then, in Section 3, we present two uniqueness assumptions and state theorems concerning continuous dependence with respect to both boundary values and parameters. Finally, in Section 4, we provide our result dealing with solutions of (1-1)–(1-3) and their differentiability properties with respect to the parameter  $\lambda$ .

### 2. Initial value problems

The  $n$ -th order difference equation (1-1) along with the conditions

$$\Delta^{i-1}v(\sigma_0) = c_i, \quad 1 \leq i \leq n, \tag{2-1}$$

where  $\sigma_0 \in \mathbb{Z}$ ,  $c_i \in \mathbb{R}$ ,  $1 \leq i \leq n$ , is called an initial value problem. For notational purposes, we let  $v(t) = v(t, \sigma_0, c_1, \dots, c_n, \lambda)$  denote the solution of the initial value problem (1-1), (2-1) on  $[\sigma_0, +\infty)_{\mathbb{Z}}$ . Results stated in this section concerning continuous dependence and differentiability of  $v$  with respect to initial conditions and parameters can be found in [Datta and Henderson 1992; Henderson and Lee 1991].

**Theorem 2.1** (continuous dependence with respect to initial values). *Assume that condition (A) is satisfied. Let  $\sigma_0 \in \mathbb{Z}$ ,  $c_1, \dots, c_n \in \mathbb{R}$ , and  $\lambda_0 \in \mathbb{R}$  be given. Then, for each  $\varepsilon > 0$  and  $k \in \mathbb{N}$ , there exists a  $\delta(\varepsilon, \sigma_0, k, c_1, \dots, c_n, \lambda_0) > 0$  such that if  $|c_i - d_i| < \delta$ ,  $1 \leq i \leq n$ , and  $|\lambda_0 - p_0| < \delta$ , then*

$$|\Delta^{i-1}v(t, \sigma_0, c_1, \dots, c_n, \lambda_0) - \Delta^{i-1}v(t, \sigma_0, d_1, \dots, d_n, p_0)| < \varepsilon$$

on  $[\sigma_0, k]_{\mathbb{Z}}$  for  $i = 1, \dots, n$ .

**Theorem 2.2** (discrete Peano). *Assume that conditions (A), (B) and (C) are satisfied. Let  $\sigma_0 \in \mathbb{Z}$ ,  $c_1, \dots, c_n \in \mathbb{R}$ , and let  $\lambda \in \mathbb{R}$  be given. Then, for each  $1 \leq j \leq n$ , given  $r_1, \dots, r_n \in \mathbb{R}$  and  $\lambda_0 \in \mathbb{R}$ ,*

$$\alpha_j(t) := \frac{\partial v}{\partial c_j}(t, \sigma_0, r_1, \dots, r_n, \lambda_0), \quad 1 \leq i \leq n,$$

*exists, is the solution of the variational equation (1-4) along  $v(t, \sigma_0, r_1, \dots, r_n, \lambda_0)$  and satisfies the initial conditions*

$$\Delta^{i-1}\alpha_j(\sigma_0) = \delta_{ij}, \quad 1 \leq i \leq n.$$

Moreover,

$$\beta(t) := \frac{\partial v}{\partial \lambda}(t, \sigma_0, r_1, \dots, r_n, \lambda_0)$$

*exists, is the solution of the nonhomogeneous equation (1-5) along  $v(t, \sigma_0, r_1, \dots, r_n, \lambda_0)$ , and satisfies the initial conditions*

$$\Delta^{i-1}\beta(\sigma_0) = 0, \quad 1 \leq i \leq n.$$

### 3. Boundary value problems

In order to establish a relation between the work in the last section and boundary value problems, we need two uniqueness assumptions.

- (D) Given  $\lambda \in \mathbb{R}$ ,  $t_0, \tau_1, \dots, \tau_n, t_1 \in \mathbb{Z}$ ,  $t_0 + n - 1 < \tau_1 < \dots < \tau_n < t_1$ , and  $A_i \in \mathbb{R}$ ,  $1 \leq i \leq n$ , if  $u_1(t)$  and  $u_2(t)$  are solutions of (1-1)–(1-3), then  $u_1(t) \equiv u_2(t)$  on  $[t_0, +\infty)_{\mathbb{Z}}$ .
- (E) For each  $\lambda \in \mathbb{R}$  and  $t_0, \tau_1, \dots, \tau_n, t_1 \in \mathbb{Z}$ , and for each solution  $u(t)$  of (1-1), the only solution  $\rho(t)$  of the boundary value problem for the variational equation (1-4) along  $u(t)$  and satisfying

$$\Delta^{(i-1)}\rho(t_0) = 0, \quad 1 \leq i \leq n - 1,$$

and

$$\rho(t_1) - \sum_{j=1}^m a_j \rho(\tau_j) = 0,$$

where  $t_0 + n - 1 < \tau_1 < \dots < \tau_m < t_1$ , is

$$\rho(t) \equiv 0 \text{ on } [t_0, +\infty)_{\mathbb{Z}}.$$

**Theorem 3.1** (continuous dependence with respect to boundary values and parameters). *Assume conditions (A) and (D) are satisfied. Let  $y(t)$  be a solution of (1-1) for some  $\lambda \in \mathbb{R}$  on  $[a, +\infty)_{\mathbb{Z}}$ . Let  $t_0 < \dots < t_0 + n - 1 < \tau_1 < \dots < \tau_m < t_1$  in  $[a, +\infty)_{\mathbb{Z}}$  be given. Then, there exists  $\varepsilon > 0$  such that if  $|\Delta^{i-1}y(t_0) - A_i| < \varepsilon$ ,  $1 \leq i \leq n - 1$ , and  $|y(t_1) - \sum_{j=1}^m a_j y(\tau_j) - A_n| < \varepsilon$ , and if  $|\lambda - \mu| < \varepsilon$ , then the boundary value problem for (1-1) with respect to the parameter  $\mu$  satisfying*

$$\Delta^{i-1}h(t_0) = A_i, \quad 1 \leq i \leq n - 1,$$

and

$$h(t_1) - \sum_{j=1}^m a_j h(\tau_j) = A_n$$

has a unique solution,  $h(t, t_0, t_1, \tau_1, \dots, \tau_m, A_1, \dots, A_n, \mu)$ , on  $[t_0, +\infty)_{\mathbb{Z}}$ , and moreover,

$$h(t, t_0, t_1, \tau_1, \dots, \tau_m, A_1, \dots, A_n, \mu) \rightarrow y(t),$$

as  $\varepsilon \rightarrow 0$ , on  $[t_0, +\infty)_{\mathbb{Z}}$ .

### 4. Main result

Now, we provide our main result concerning differentiation of solutions of (1-1)–(1-3) with respect to the parameter  $\lambda$ .

**Theorem 4.1.** *Assume conditions (A)–(E) are satisfied. For  $t_0 < \dots < t_0 + n - 1 < \tau_1 < \dots < \tau_m < t_1$  in  $\mathbb{Z}$ , let  $u(t, t_0, t_1, \tau_1, \dots, \tau_m, A_1, \dots, A_n, \lambda)$  denote the solution of (1-1)–(1-3) on  $[t_0, +\infty)_{\mathbb{Z}}$ . Then,  $\partial u / \partial \lambda$  exists on  $[t_0, +\infty)_{\mathbb{Z}}$ , and*

$w(t) := (\partial u / \partial \lambda)(t)$  is the solution of the nonhomogeneous linear equation (1-5) along  $u(t)$  and satisfies

$$\Delta^{i-1}w(t_0) = 0, \quad 1 \leq i \leq n - 1,$$

and

$$w(t_1) - \sum_{j=1}^m a_j w(\tau_j) = 0.$$

*Proof.* Let  $\varepsilon > 0$  be given. For  $0 < |h| < \varepsilon$ , we consider the difference quotient

$$w_h(t) := \frac{1}{h} \left( u(t, t_0, t_1, \tau_1, \dots, \tau_n, A_1, \dots, A_n, \lambda + h) - u(t, t_0, t_1, \tau_1, \dots, \tau_n, A_1, \dots, A_n, \lambda) \right).$$

We show that  $\lim_{h \rightarrow 0} w_h(t)$  exists on  $[t_0, +\infty)_{\mathbb{Z}}$ . For  $h \neq 0$ , we first observe that, for  $1 \leq i \leq n - 1$ ,

$$\begin{aligned} \Delta^{i-1}w_h(t_0) &= \frac{1}{h} \left( \Delta^{i-1}u(t_0, t_0, t_1, \tau_1, \dots, \tau_m, A_1, \dots, A_n, \lambda + h) - \Delta^{i-1}u(t_0, t_0, t_1, \tau_1, \dots, \tau_m, A_1, \dots, A_n, \lambda) \right) \\ &= \frac{1}{h} (A_i - A_i) = 0, \end{aligned}$$

and

$$\begin{aligned} w_h(t_1) - \sum_{j=1}^m \alpha_j w_h(\tau_j) &= \frac{1}{h} \left( u(t_1, t_0, t_1, \tau_1, \dots, \tau_m, A_1, \dots, A_n, \lambda + h) - \sum_{j=1}^m a_j u(\tau_j, t_0, t_1, \tau_1, \dots, \tau_m, A_1, \dots, A_n, \lambda + h) \right. \\ &\quad \left. - u(t_1, t_0, t_1, \tau_1, \dots, \tau_m, A_1, \dots, A_n, \lambda) + \sum_{j=1}^m a_j u(\tau_j, t_0, t_1, \tau_1, \dots, \tau_m, A_1, \dots, A_n, \lambda) \right) \\ &= \frac{1}{h} (A_n - A_n) = 0. \end{aligned}$$

Next, we set

$$D := \Delta^{n-1}u(t_0, t_0, t_1, \tau_1, \dots, \tau_m, A_1, \dots, A_n, \lambda)$$

and

$$\varepsilon_h := \varepsilon_0(h) = \Delta^{n-1}u(t_0, t_0, t_1, \tau_1, \dots, \tau_m, A_1, \dots, A_n, \lambda + h) - D.$$

By Theorem 3.1,  $\varepsilon_h \rightarrow 0$  as  $h \rightarrow 0$ . With  $v(t, t_0, c_1, \dots, c_n, \lambda)$  being our notation for solutions of initial value problems (1-1), (2-1) corresponding to  $\lambda$  in (1-1), we

have, by using a telescoping sum,

$$\begin{aligned} w_h(t) &= \frac{1}{h} (v(t, t_0, A_1, \dots, A_{n-1}, D+\varepsilon, \lambda+h) - v(t, t_0, A_1, \dots, A_{n-1}, D, \lambda)) \\ &= \frac{1}{h} (v(t, t_0, A_1, \dots, A_{n-1}, D+\varepsilon, \lambda+h) - v(t, t_0, A_1, \dots, A_{n-1}, D, \lambda+h) \\ &\quad + v(t, t_0, A_1, \dots, A_{n-1}, D, \lambda+h) - v(t, t_0, A_1, \dots, A_{n-1}, D, \lambda)). \end{aligned}$$

By [Theorem 2.2](#),  $\alpha_n = \partial v / \partial c_n$  and  $\beta = \partial v / \partial \lambda$  both exist. So, by the mean value theorem,

$$\begin{aligned} w_h(t) &= \frac{1}{h} (\alpha_n(t, v(t, t_0, A_1, \dots, A_{n-1}, D + \bar{\varepsilon}, \lambda + h))(D + \varepsilon - D) \\ &\quad + \beta(t, v(t, t_0, A_1, \dots, A_{n-1}, D, \lambda + \bar{h})(\lambda + h - \lambda))) \\ &= \frac{\varepsilon}{h} \alpha_n(t, v(t, t_0, A_1, \dots, A_{n-1}, D + \bar{\varepsilon}, \lambda + h)) \\ &\quad + \beta(t, v(t, t_0, A_1, \dots, A_{n-1}, D, \lambda + \bar{h})), \end{aligned}$$

where

$$\begin{aligned} \alpha_n(t, v(t, t_0, A_1, \dots, A_{n-1}, D + \bar{\varepsilon}, \lambda + h)) &= \frac{\partial v}{\partial c_n}(t, t_0, A_1, \dots, A_{n-1}, D + \bar{\varepsilon}, \lambda + h), \\ \beta(t, v(t, t_0, A_1, \dots, A_{n-1}, D, \lambda + \bar{h})) &= \frac{\partial v}{\partial \lambda}(t, t_0, A_1, A_{n-1}, D, \lambda + \bar{h}), \end{aligned}$$

$\bar{\varepsilon}$  is between 0 and  $\varepsilon$ , and  $\bar{h}$  is between 0 and  $h$ .

To show that  $\lim_{h \rightarrow 0} w_h(t)$  exists, it suffices to show that  $\lim_{h \rightarrow 0} \varepsilon/h$  exists. We have the  $n-1$  conditions,  $\Delta^{i-1} w_h(t_0) = 0, i = 1, \dots, n-1$ , and the condition  $w_h(t_1) - \sum_{j=1}^m a_j w_h(\tau_j) = 0$ . So, from the last condition,

$$\begin{aligned} &\frac{\varepsilon}{h} \frac{\partial v}{\partial c_n}(t_1, t_0, A_1, \dots, A_{n+1}, D + \bar{\varepsilon}, \lambda + h) + \frac{\partial v}{\partial \lambda}(t_1, t_0, A_1, \dots, A_{n-1}, D, \lambda + \bar{h}) \\ &\quad - \frac{\varepsilon}{h} \sum_{j=1}^m a_j \frac{\partial v}{\partial c_n}(t_1, t_0, A_1, \dots, A_{n+1}, D + \bar{\varepsilon}, \lambda + h) \\ &\quad - \sum_{j=1}^m a_j \frac{\partial v}{\partial \lambda}(t_1, t_0, A_1, \dots, A_{n-1}, D, \lambda + \bar{h}) = 0. \end{aligned}$$

Hence, we have

$$\begin{aligned} \frac{\varepsilon}{h} &= \frac{1}{M_{h, \bar{\varepsilon}}} \left( -\beta(t_1, v(t, t_0, A_1, \dots, A_{n-1}, D, \lambda + \bar{h})) \right. \\ &\quad \left. + \sum_{j=1}^m a_j \beta(\tau_j, v(t, t_0, A_1, \dots, A_{n-1}, D, \lambda + \bar{h})) \right), \end{aligned}$$



where

$$M_{h,\bar{\varepsilon}} := \alpha_n(t_1, v(t, t_0, A_1, \dots, A_{n-1}, D + \bar{\varepsilon}, \lambda + h)) - \sum_{j=1}^m a_j \alpha_n(\tau_j, v(t, t_0, A_1, \dots, A_{n-1}, D + \bar{\varepsilon}, \lambda + h)).$$

Now,  $\Delta^{n-1} \alpha_n(t_0, v(t, t_0, A_1, \dots, A_{n-1}, D, \lambda)) = 1$ , so

$$\alpha_n(t, v(t, t_0, A_1, \dots, A_{n-1}, D, \lambda)) \neq 0.$$

By uniqueness assumption (E),

$$\alpha_n(t_1, v(t, t_0, A_1, \dots, A_{n-1}, D, \lambda)) - \sum_{j=1}^m a_j \alpha_n(\tau_j, v(t, t_0, A_1, \dots, A_{n-1}, D, \lambda)) \neq 0.$$

By Theorem 3.1, for  $h$  sufficiently small,  $M_{h,\bar{\varepsilon}} \neq 0$ . So,  $\lim_{h \rightarrow 0} \varepsilon/h$  exists, and

$$\lim_{h \rightarrow 0} \frac{\varepsilon}{h} = \lim_{h \rightarrow 0} \frac{-1}{M_{h,\bar{\varepsilon}}} \left( \beta(t_1, v(t, t_0, A_1, \dots, A_{n-1}, D, \lambda + \bar{h})) - \sum_{j=1}^m a_j \beta(\tau_j, v(t, t_0, A_1, \dots, A_{n-1}, D, \lambda + \bar{h})) \right) := J.$$

Hence,  $\lim_{h \rightarrow 0} w_h(t)$  exists, or in particular,  $(\partial u / \partial \lambda)(t) = \lim_{h \rightarrow 0} w_h(t)$  exists on  $[t_0, +\infty)_{\mathbb{Z}}$ , and

$$\begin{aligned} w(t) &:= \lim_{h \rightarrow 0} w_h(t) \\ &= \frac{\partial u}{\partial \lambda}(t) \\ &= J \cdot \alpha_n(t, v(t, t_0, A_1, \dots, A_{n-1}, D, \lambda)) + \beta(t, v(t, t_0, A_1, \dots, A_{n-1}, D, \lambda)) \\ &= J \cdot \alpha_n(t, u(t, t_0, t_1, \tau_1, \dots, \tau_m, A_1, \dots, A_n, \lambda)) \\ &\quad + \beta(t, u(t, t_0, t_1, \tau_1, \dots, \tau_m, A_1, \dots, A_n, \lambda)), \end{aligned}$$

which is a solution of (1-5) along  $u(t)$ , and from above satisfies the boundary conditions,

$$\Delta^{i-1} w(t_0) = \lim_{h \rightarrow 0} \Delta^{i-1} w_h(t_0) = 0, \quad 1 \leq i \leq n-1,$$

and

$$w(t_1) - \sum_{j=1}^m a_j w(\tau_j) = \lim_{h \rightarrow 0} \left( w_h(t_1) - \sum_{j=1}^m a_j w_h(\tau_j) \right) = 0. \quad \square$$

## References

- [Ashyralyev et al. 2004] A. Ashyralyev, I. Karatay, and P. E. Sobolevskii, “On well-posedness of the nonlocal boundary value problem for parabolic difference equations”, *Discrete Dyn. Nat. Soc.* **2004**:2 (2004), 273–286. MR 2006j:39004 Zbl 1077.39015
- [Benchohra et al. 2007] M. Benchohra, S. Hamani, J. Henderson, S. K. Ntouyas, and A. Ouahab, “Differentiation and differences for solutions of nonlocal boundary value problems for second order difference equations”, *Int. J. Difference Equ.* **2**:1 (2007), 37–47. MR 2008k:39015 Zbl 1177.39003
- [Datta 1998] A. Datta, “Differences with respect to boundary points for right focal boundary conditions”, *J. Differ. Equations Appl.* **4**:6 (1998), 571–578. MR 99k:39007 Zbl 0921.39003
- [Datta and Henderson 1992] A. Datta and J. Henderson, “Differentiation of solutions of difference equations with respect to right focal boundary values”, *Panamer. Math. J.* **2**:1 (1992), 1–16. MR 93a:39002 Zbl 0746.39002
- [Ehme and Henderson 1992] J. Ehme and J. Henderson, “Differentiation of solutions of boundary value problems with respect to boundary conditions”, *Appl. Anal.* **46**:3–4 (1992), 175–194. MR 93g:34028 Zbl 0808.34018
- [Hartman 1982] P. Hartman, *Ordinary differential equations*, 2nd (aka corrected reprint) ed., S. M. Hartman, Baltimore, 1982. Reprinted Birkhäuser, Boston, 1982 and SIAM, Philadelphia, 2002. MR 49 #9294 Zbl 281.34001
- [Henderson and Lee 1991] J. Henderson and L. Lee, “Continuous dependence and differentiation of solutions of finite difference equations”, *Int. J. Math. Math. Sci.* **14**:4 (1991), 747–756. MR 92f:39008 Zbl 0762.39004
- [Henderson et al. 1994] J. Henderson, M. Horn, and L. Howard, “Differentiation of solutions of difference equations with respect to boundary values and parameters”, *Comm. Appl. Nonlinear Anal.* **1**:2 (1994), 47–60. MR 95g:39005 Zbl 0856.39002
- [Henderson et al. 2008] J. Henderson, B. Hopkins, E. Kim, and J. W. Lyons, “Boundary data smoothness for solutions of nonlocal boundary value problems for  $n$ -th order differential equations”, *Involve* **1**:2 (2008), 167–181. MR 2009d:34010 Zbl 1151.34016
- [Hopkins et al. 2009] B. Hopkins, E. Kim, J. W. Lyons, and K. Speer, “Boundary data smoothness for solutions of nonlocal boundary value problems for second order difference equations”, *Comm. Appl. Nonlinear Anal.* **16**:2 (2009), 1–12. MR 2526876 Zbl 1188.39006
- [Lyons 2011] J. W. Lyons, “Differentiation of solutions of nonlocal boundary value problems with respect to boundary data”, *Electron. J. Qual. Theory Differ. Equ.* (2011), Article ID #51. MR 2012i:34024
- [Lyons 2014] J. W. Lyons, “Disconjugacy, differences and differentiation for solutions of non-local boundary value problems for  $n$ th order difference equations”, *J. Differ. Equations Appl.* **20**:2 (2014), 296–311. MR 3173548 Zbl 06259244
- [Spencer 1975] J. D. Spencer, “Relations between boundary value functions for a nonlinear differential equation and its variational equations”, *Canad. Math. Bull.* **18**:2 (1975), 269–276. MR 53 #3402 Zbl 0321.34014

Received: 2014-05-12

Revised: 2014-05-21

Accepted: 2014-05-31

johnny\_henderson@baylor.edu

*Department of Mathematics, Baylor University,  
One Bear Place #97328, Waco, TX 76798, United States*

xuewei\_jiang@baylor.edu

*Department of Mathematics, Baylor University,  
One Bear Place #97328, Waco, TX 76798, United States*

# Outer billiards and tilings of the hyperbolic plane

Filiz Dogru, Emily M. Fischer and Cristian Mihai Munteanu

(Communicated by Kenneth S. Berenhaut)

We present new results regarding the periodicity of outer billiards in the hyperbolic plane around polygonal tables which are tiles in regular two-piece tilings of the hyperbolic plane.

## 1. Introduction

Outer billiards is a simple dynamical system introduced by B. H. Neumann [1959]. J. Moser [1973; 1978] popularized outer billiards as a toy model for planetary motion as a means of finding possible unbounded orbits. Since then, many mathematicians have asked and answered questions about outer billiards systems in various geometries. For example, C. Culter proved in 2004 the existence of periodic orbits for polygonal tables in the Euclidean plane (the proof is presented by S. Tabachnikov [2007]). R. Schwartz [2007; 2009] answered, in the affirmative, Moser's question about the existence of unbounded orbits for certain polygons.

The main motivation for this paper is a result of Vivaldi and Shaidenko [1987] that in the Euclidean case, outer billiards associated to quasirational polygons have all orbits bounded; see also [Kołodziej 1989; Gutkin and Simányi 1992]. As a consequence, all orbits about a lattice polygon in the Euclidean plane are periodic. We continue the work of Dogru and Tabachnikov [2003], who studied the relationship between one-tile regular tilings of the hyperbolic plane and the outer billiards system.

For a detailed account of hyperbolic geometry and the hyperbolic plane, we direct the reader to [Greenberg 1980], and for a survey of outer billiards, see [Tabachnikov and Dogru 2005; Tabachnikov 2005].

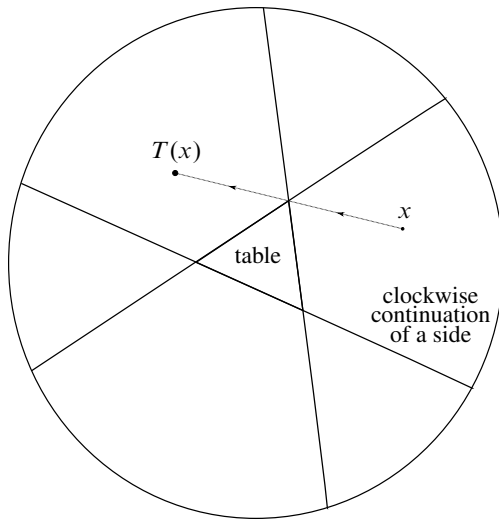
## 2. Definitions

The outer billiard map associated to a convex polygonal table  $P$  in the hyperbolic plane is defined as follows. For a point  $x \in \mathbb{H}^2 \setminus P$ , there are two lines that pass

---

*MSC2010:* 37E15.

*Keywords:* hyperbolic billiards, outer billiards, polygonal billiards, symbolic dynamics, tiling, rotation number, crochet.



**Figure 1.** Outer billiards map in the Klein model.

through  $x$  and are tangent to the table  $P$ . By convention, we consider the tangent line for which  $P$  is on the left, from the point of view of  $x$ . Then we reflect  $x$  about the tangency (support) point to get  $T(x)$  (see Figure 1). The map is well-defined whenever the tangency point is unique and so we are able to define the map  $T$  on the entire hyperbolic plane except for the clockwise continuations of the sides of  $P$  (see Figure 1) and their preimages under  $T$ . An immediate consequence of the definition is that  $T$  is a piecewise isometry.

Likewise, the inverse map  $T^{-1}$  is not defined on the counterclockwise continuations of the sides of  $P$ . We define the *web* associated to  $P$  to be the union of all preimages under  $T$  of the clockwise continuation of the sides and of all preimages under  $T^{-1}$  of the counterclockwise continuation of the sides. For each connected component of the complement of the web, the restriction of the map  $T^n$  to that component is defined by a single isometry of the hyperbolic plane for every  $n \in \mathbb{Z}$ . That means that each connected component of the complement of the web maps as a whole under the iterations of  $T$ .

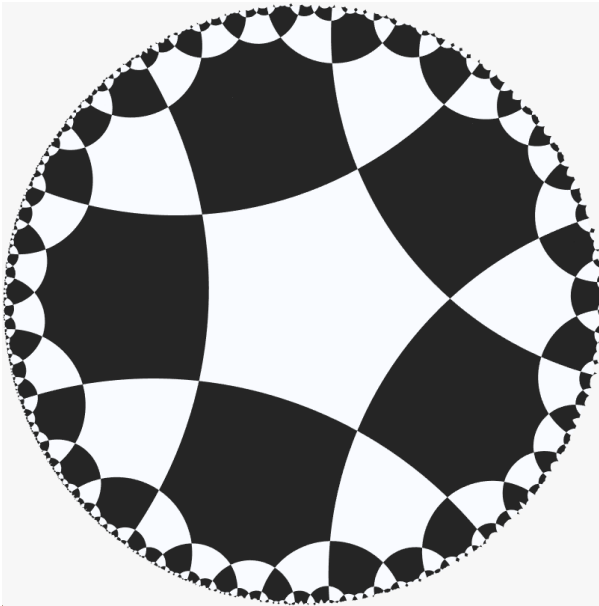
Another feature of the billiards map  $T$  is that it extends continuously to a continuous circle map  $t : S^1 \rightarrow S^1$  at infinity. The map  $t$  is defined using the same reflecting procedure. In this case, the uniqueness of the support point is not needed since the distance between our initial point and the support point is infinite no matter the choice, and hence the map  $t$  is well-defined for every point at infinity. Since  $t$  is a circle map, it has a well-defined Poincaré rotation number  $\rho(t)$ , and we will prove in Section 3 that  $\rho(t)$  encodes information about the combinatorial dynamics of the outer billiards.

### 3. Outer billiards on tilings

We are studying the hyperbolic outer billiards map associated with a polygonal table that is part of a two-piece regular tiling of the hyperbolic plane. These tilings use two polygonal pieces, a regular  $M$ -gon and a regular  $N$ -gon that meet four in each vertex (see [Figure 2](#)). We describe the combinatorial dynamics for outer billiards around one of the  $M$ -gons. We note that the web associated to such a map will fall exactly on the grid lines of the tiling. This is because the reflection around a vertex of the table tile is just a rotation by  $180^\circ$  around vertices in the tiling. It follows that each tile maps as a whole under iterations of  $T$ .

**3.1. Previous results.** Previous results describing outer billiards of tiles in the hyperbolic plane are obtained in [[Dogru and Tabachnikov 2003](#)]. In this paper, the authors have proved that every orbit of the outer billiard map around a right-angled regular  $n$ -gon, for  $n \geq 5$ , is periodic. Any right-angled regular  $n$ -gon generates a tiling of the hyperbolic plane entirely consisting of  $n$ -gons. The theorems proven in the next sections have the same flavor as Theorem 4 in the above mentioned paper.

Define the *rank* of a tile as the minimum number of sides that one has to cross, when starting inside the table, to get to the given tile. This means that tiles that have one common side with the table have rank 1, and tiles that have a common side with a tile of rank 1 have rank 2, and so on.



**Figure 2.** Example of  $(M, N)$ -tiling for  $(M, N) = (6, 7)$ .

**Theorem 1** [Dogru and Tabachnikov 2003]. *For a tiling of regular  $n$ -gons meeting in four,  $n \geq 5$ , the dual billiard map  $T$  preserves the rank of a tile, and every orbit of  $T$  is periodic. The set of rank  $k$  tiles consists of*

$$q_k = n \frac{\lambda_1^k - \lambda_2^k}{\lambda_1 - \lambda_2}$$

elements, where

$$\lambda_{1,2} = \frac{n - 2 \pm \sqrt{n(n - 4)}}{2}$$

are the roots of the equation  $\lambda^2 - (n - 2)\lambda + 1 = 0$ . The action of  $T$  on the set of rank  $k$  tiles is a transitive cyclic permutation  $i \mapsto i + p_k$ , where

$$p_k = \frac{\lambda_1^{k-1} - \lambda_2^{k-1}}{\lambda_1 - \lambda_2} + \frac{\lambda_1^k - \lambda_2^k}{\lambda_1 - \lambda_2}.$$

The rotation number of the dual billiard map at infinity is given by the formula

$$\rho(t) = \lim_{k \rightarrow \infty} \frac{p_k}{q_k} = \frac{n - \sqrt{n(n - 4)}}{2n}.$$

The proof of this theorem uses geometric arguments for the periodicity of orbits and recurrence formulas for computing the number of tiles in each rank and the rotation number of  $t$  (see [Dogru and Tabachnikov 2003] for details). The authors make an important remark that the representation of  $\lambda_1$  (and so the rotation number of the map at infinity) as a continued fraction encodes the dynamics of the tiles under the billiard map  $T$ . We will deduce similar results for two-piece tilings.

**3.2. New results.** Our results extend Theorem 1 to two-piece regular tilings of the hyperbolic plane. We will denote a tiling of regular  $M$ -gons and regular  $N$ -gons as an  $(M, N)$ -tiling, and we will always consider the table to be an  $M$ -gon. Such an  $(M, N)$ -tiling exists if  $\frac{1}{M} + \frac{1}{N} < \frac{1}{2}$ . As mentioned earlier, these tilings have four shapes meeting at each vertex, two  $M$ -gons and two  $N$ -gons.

**3.2.1. Triangles and  $N$ -gons.** Most of the geometric arguments used here are analogous to those used by Dogru and Tabachnikov. Our counting arguments are different, although they are also based on recurrence relations.

Let us introduce a more general notation for rank in order to avoid cumbersome indexing. Observe that the layer of tiles of rank  $k$  includes tiles of the same type (all  $M$ -gons or all  $N$ -gons) and as rank changes by one, that shape changes. So triangles always have even rank and  $N$ -gons always have odd rank. We will say that a rank  $2k - 1$  tile is a rank  $k$   $N$ -gon and a rank  $2k$  tile is a rank  $k$  triangle. The rest of this section is dedicated to describing the dynamics of the billiard map  $T$  in the  $(3, N)$ -tilings through the proof of the following theorem:

**Theorem 2.** *For a  $(3, N)$ -tiling,  $N \geq 7$ , the outer billiard map  $T$  preserves the rank of a tile and every orbit of  $T$  is periodic. The set of rank  $k$   $N$ -gons consists of*

$$q_k = \frac{1}{\sqrt{N-6}}(\Phi_1^{2k-3} + \Phi_2^{2k-3}) + \Phi_1^{2k-2} + \Phi_2^{2k-2}$$

*elements and the set of rank  $k$  triangles consists of*

$$l_k = \frac{N-4}{\sqrt{N-6}}(\Phi_1^{2k-3} + \Phi_2^{2k-3}) + (N-3)(\Phi_1^{2k-2} + \Phi_2^{2k-2})$$

*elements, where*

$$\Phi_{1,2} = \frac{\sqrt{N-6} \pm \sqrt{N-2}}{2}$$

*are the two roots of the equation*

$$\Phi^2 - \sqrt{N-6}\Phi - 1 = 0.$$

*The action of  $T$  on the set of rank  $k$   $N$ -gons is a cyclic permutation  $i \mapsto i + p_k$ , where*

$$p_k = \frac{q_k}{3} + \frac{\Phi_1^{2k-4} - \Phi_2^{2k-4}}{\sqrt{(N-6)(N-2)}} + \frac{\Phi_1^{2k-3} - \Phi_2^{2k-3}}{\sqrt{N-2}},$$

*and the action of  $T$  on the set of rank  $k$  triangles is also a cyclic permutation  $i \mapsto i + j_k$ , where*

$$j_k = \frac{l_k}{3} + (N-4) \frac{\Phi_1^{2k-4} - \Phi_2^{2k-4}}{\sqrt{(N-6)(N-2)}} + (N-3) \frac{\Phi_1^{2k-3} - \Phi_2^{2k-3}}{\sqrt{N-2}}.$$

*The rotation number of the outer billiard map at infinity is given by the formula*

$$\rho(t) = \lim_{k \rightarrow \infty} \frac{p_k}{q_k} = \lim_{k \rightarrow \infty} \frac{j_k}{l_k} = \frac{1}{3} + \frac{1}{3(1 + \Phi_1^2)} = \frac{1}{3} + \frac{1}{3\sqrt{N-2}\Phi_1}.$$

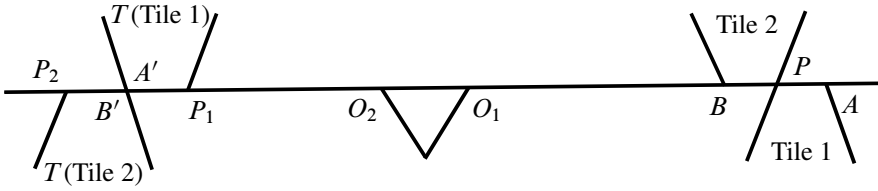
**Theorem 2** contains many independent results and for reasons of clarity we will prove them one by one as claims.

**Claim 3.** *Every orbit of  $T$  is periodic.*

*Proof.* The proof of this result is written in much detail in [Dogru and Tabachnikov 2003]. We will present here a sketch of it and will refer the reader to the above work for detailed explanations. The statement of the claim is a consequence of the following lemma:

**Lemma 4.** *The rank of a tile is preserved under  $T$ .*

*Proof of lemma.* The proof is by induction on the rank, based on geometrical observations. Observe that rank 1 tiles are preserved by  $T$  and notice that every rank  $k$  tile is adjacent to a rank  $k-1$  tile, where these two tiles map together under a single application of  $T$ . These two facts complete the base case and the step of the induction.  $\square$



**Figure 3.** Special case for Lemma 6.

From Lemma 4, since there are finitely many tiles of rank  $k$ , every tile must eventually map back to itself after  $m$  iterations, for some natural number  $m$ . Hence the  $m$ -th iteration of  $T$  maps the entire tile to itself. This implies that  $T^{\circ m}$  (the composition of  $T$  with itself  $m$  times) is a rotation by either  $2\pi j/N$  (for  $N$ -gons) or  $2\pi j/3$  (for triangles) around some point inside the tile. Hence  $T^{\circ Nm}$  restricted to that tile is the identity if the tile is an  $N$ -gon and  $T^{\circ 3m}$  restricted to that tile is the identity if the tile is a triangle. We conclude that every orbit of  $T$  is periodic.  $\square$

**Claim 5.** For every  $k \geq 1$ ,  $T$  permutes the rank  $k$  tiles cyclically.

*Proof.* This claim is an immediate corollary to the following lemma:

**Lemma 6.** Any two consecutive rank  $k$  tiles are mapped to two consecutive rank  $k$  tiles.

*Proof of lemma.* We know by Lemma 4 that the rank of two tiles is preserved under  $T$ . If the two consecutive tiles are not separated by a clockwise continuation of one of the sides of the table then their common point is mapped, together with the two tiles, through the same vertex. Thus the tiles are mapped to two consecutive tiles.

If the two tiles are separated by such a continuation of one side of the table then the argument is more involved. A similar argument is presented in [Dogru and Tabachnikov 2003]. Figure 3 gives a pictorial representation of the situation. The first tile is reflected in  $O_1$ , while the second one is reflected in  $O_2$ . What remains to prove is that  $A'=B'$  so that the images of the two tiles still touch in one point. The following sequence of equalities completes the proof:

$$A'O_2 = A'O_1 - O_1O_2 = BO_1 + AB - O_1O_2 = BO_1 + O_1O_2 = BO_2 = B'O_2. \quad \square$$

In order to compute the formulas for  $q_k, p_k, j_k, l_k$ , we first explain why the tiling we are working with has an intrinsic self-similar geometric structure. We will refer from now on to this self-similar structure as the *crochet pattern*. To describe the crochet pattern, we consider  $N$ -gons to be of two types,  $X$ -type and  $Y$ -type (see Figure 4). Type  $X$   $N$ -gons have two parents in the sense that they touch two  $N$ -gons of the previous rank, while type  $Y$   $N$ -gons touch only one parent. The rank 1  $N$ -gons are of neither of the types, having zero parents, so we call them type 0  $N$ -gons. (This is why our counting argument begins with counting rank 2  $N$ -gons.)



The following claim gives an intuitive explanation of why we call this self-similar structure of the tiling a crochet pattern.

**Claim 7.** *When passing from the  $k$ -th layer of  $N$ -gons to the  $(k+1)$ -th layer of  $N$ -gons, we apply the replacement rules*

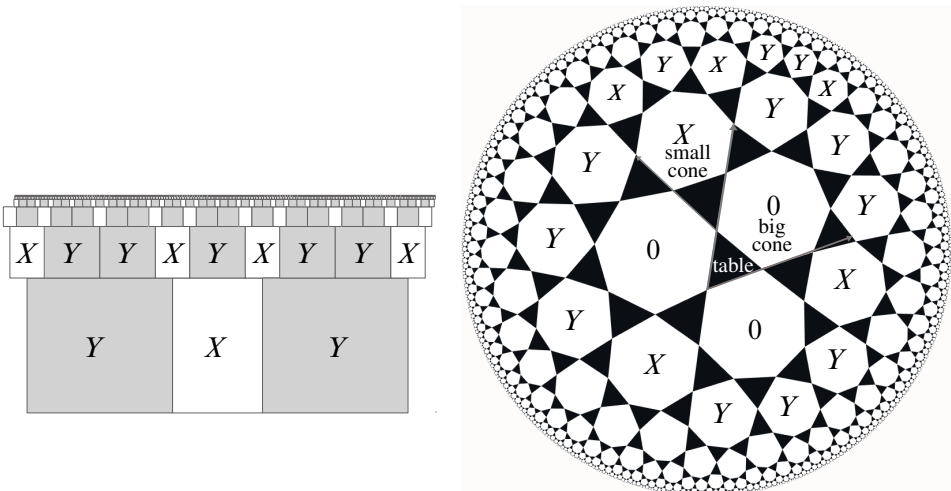
$$\begin{aligned} X &\rightarrow XY^{N-6}, \\ Y &\rightarrow XY^{N-5}, \end{aligned}$$

*i.e., when incrementing rank of the layer by 1, every  $X$  gets replaced by an  $X$  followed by  $N-6$   $Y$ s, and every  $Y$  gets replaced by an  $X$  followed by  $N-5$   $Y$ s.*

*Proof.* The methods used to prove this claim have been developed by Poincaré, and we will not dwell on the details here. The reader can find extensive explanation in *The Symmetry of Things* [Conway et al. 2008].

Instead, we will illustrate the methods used to prove the claim in the case of  $N = 7$  in order to give the geometrical intuition behind the proof. Figure 4 illustrates the local and global behavior of a  $(3, 7)$ -tiling.

In the local picture, the difference between a type  $X$  7-gon and a type  $Y$  7-gon is encoded in the different types of degenerate heptagons we associate to them. We associate to the  $Y$ -type heptagon a rectangle with three additional points on the upper side, while to the  $X$ -type heptagon we associate a rectangle with two additional points on the upper side and one on the lower side since it has two parents. Now by reducing the triangles in the global picture to points, we notice that the heptagons must meet three in each vertex. This results in the crochet pattern shown in Figure 4. This crochet pattern immediately implies the claimed replacement rules.  $\square$



**Figure 4.** The  $(3, 7)$ -tiling.

We can now use this crochet pattern to start our counting argument in order to get the exact numbers in [Theorem 2](#).

**Claim 8.** *The formulas for  $q_k, p_k, j_k, l_k$  hold as stated in [Theorem 2](#).*

*Proof.* Denote the number of  $X$ -type and  $Y$ -type  $N$ -gons of rank  $k$  by  $x_k$  and  $y_k$  and use [Claim 7](#) to obtain the system of linear difference equations

$$\begin{pmatrix} x_k \\ y_k \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ N-6 & N-5 \end{pmatrix} \begin{pmatrix} x_{k-1} \\ y_{k-1} \end{pmatrix}.$$

The initial configuration is  $\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 3(N-4) \end{pmatrix}$  because there must be three rank 2  $N$ -gons with two parents, and the rest of the vertices of the rank 1  $N$ -gons must serve as an anchor for a different  $Y$ -type rank 2  $N$ -gon. Solving this recurrence gives the general term formula

$$\begin{pmatrix} x_k \\ y_k \end{pmatrix} = 3 \begin{pmatrix} \frac{1}{\sqrt{N-6}}(\Phi_1^{2k-3} + \Phi_2^{2k-3}) \\ \Phi_1^{2k-2} + \Phi_2^{2k-2} \end{pmatrix},$$

where

$$\Phi_1 = \frac{\sqrt{N-2} + \sqrt{N-6}}{2} \quad \text{and} \quad \Phi_2 = \frac{-\sqrt{N-2} + \sqrt{N-6}}{2}.$$

From here the formula for  $q_k = x_k + y_k$  follows immediately.

To count the triangles of rank  $k$ , we observe that the triangles of rank  $k$  are the next layer after the  $N$ -gons of rank  $k$ , and each  $X$ -type  $N$ -gon is replaced by  $N-4$  triangles and each  $Y$ -type is replaced by  $N-3$  triangles. Hence the formula for  $l_k = (N-4)x_k + (N-3)y_k$  can be computed.

In order to count how many rank  $k$   $N$ -gons  $T$  jumps, i.e.,  $p_k$ , we need to define  $s_k$  as the number of rank  $k$   $N$ -gons in a small cone, as can be seen in [Figure 4](#). A small cone is opposite one of the triangle’s vertices and doesn’t contain any side of the triangle. In the same way, a big cone (see [Figure 4](#)) is opposite one of the sides of a triangle and contains the table. The number of rank  $k$   $N$ -gons in a big cone is just  $q_k/3 - s_k$  because of the 3-fold symmetry of the tiling.

As above, we need to introduce  $x_k^s$  and  $y_k^s$ , the number of  $X$ -type and  $Y$ -type rank  $k$   $N$ -gons in a small cone. With this,  $s_k = x_k^s + y_k^s$ . The billiard map  $T$  makes any tile jump over two small cones and one big cone so in total it will jump

$$p_k = 2s_k + \left(\frac{q_k}{3} - s_k\right) = \frac{q_k}{3} + s_k.$$

By studying the structure of the small cone, we observe the crochet pattern once again. One notices that the cone that starts at the last  $X$ -type  $N$ -gon of the rank  $k$  ( $k \geq 2$ ) layer looks exactly the same as the initial small cone. That is why  $s_k$  is

equal to the total number of  $N$ -gons obtained by starting with an  $X$ -type  $N$ -gon and using the replacement rules in Claim 7. We express this as a sum,

$$\begin{pmatrix} x_k^s \\ y_k^s \end{pmatrix} = \sum_{i=0}^{k-2} \begin{pmatrix} 1 & 1 \\ N-6 & N-5 \end{pmatrix}^i \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

which, after some computation, becomes

$$\begin{pmatrix} x_k^s \\ y_k^s \end{pmatrix} = \begin{pmatrix} 1 + \frac{\Phi_1^{2k-4} - \Phi_2^{2k-4}}{\sqrt{(N-6)(N-2)}} \\ -1 + \frac{\Phi_1^{2k-3} - \Phi_2^{2k-3}}{\sqrt{N-2}} \end{pmatrix}.$$

The formula for  $p_k = q_k/3 + x_k^s + y_k^s$  follows immediately, and  $j_k$  is computed in the same manner as  $l_k$  was computed. As we have already said, every  $X$  type  $N$ -gon is replaced by  $N-4$  triangles and every  $Y$  type  $N$ -gon is replaced by  $N-3$  triangles on the next level, and this procedure leaves uncounted only one rank  $k$  triangle in the small cone, so  $j_k = (N-4)x_k^s + (N-3)y_k^s + 1$ .  $\square$

**Claim 9.** *The rotation number  $\rho(t)$  equals*

$$\frac{1}{3} + \frac{1}{3(1 + \Phi_1^2)} = \frac{1}{3} + \frac{1}{3\sqrt{N-2}\Phi_1}.$$

*Proof.* The  $k$ -th layer of  $N$ -gons gives a discrete approximation of the circle map at infinity and so  $p_k/q_k$  is an approximation of  $\rho(t)$  as  $k$  goes to  $\infty$ . By taking the limit we obtained the desired formula for the rotation number  $\rho(t)$ .  $\square$

This last claim completes the proof of all the statements in Theorem 2.

**Remark 10.** (1) One might expect the formulas in Theorem 2 to also work for  $N = 6$ , i.e., a  $(3, 6)$ -tiling of the Euclidean plane. That is not the case even though the crochet pattern works exactly the same also in the  $(3, 6)$ -tiling. The difference that appears when computing the formulas in the  $(3, 6)$ -tiling is that the matrix of the difference system is not diagonalizable and so its powers look completely different.

- (2) Note that the determinant of all the matrices given by the crochet pattern is 1. We believe this is true because the crochet pattern replacement can also be reversed, i.e., starting with the rank  $k$  layer, we can construct the rank  $k-1$  layer.
- (3) According to Theorem 2, one can express the eigenvalues  $\Phi_1$  and  $\Phi_2 = 1/\Phi_1$  via the rotation number  $\rho(t)$ . Therefore this rotation number determines the numbers  $q_k, l_k, p_k, j_k$ , and hence the whole dynamics of the map  $T$ .

**3.2.2. General  $(M, N)$ -tilings.** Next we consider the case of a general  $(M, N)$ -tiling. The theorem and subsequent proof are analogous to those in the  $(3, N)$  case in the previous subsection, but we must consider the cases separately due to a difference in the counting method. In the previous section,  $N$ -gons were classified into types  $X$  and  $Y$ , having two parents and one parent, respectively. However, due to the difference in geometry of triangles versus generic  $M$ -gons, the tilings in the  $M \geq 4$  case never produce  $N$ -gons with two parents. In this case,  $N$ -gons either have one parent or no parent, which we denote as types  $Y$  and  $Z$ . This alternate counting method will be explained in detail in the proof, but first we state the theorem:

**Theorem 11.** *For an  $(M, N)$ -tiling with  $M, N \geq 4$  and*

$$\frac{1}{M} + \frac{1}{N} < \frac{1}{2},$$

*the outer billiard map  $T$  preserves the rank of a tile and every orbit of  $T$  is periodic. The set of rank  $k$   $N$ -gons consists of*

$$q_k = \frac{M}{\sqrt{b^2 - 4}} \left( (b + 1)(\alpha_1^{2k-2} - \alpha_2^{2k-2}) - (\alpha_1^{2k-4} - \alpha_2^{2k-4}) \right)$$

*elements, and the set of rank  $k$   $M$ -gons consists of*

$$l_k = \frac{M(N - 2)}{\sqrt{b^2 - 4}} \left( b(\alpha_1^{2k-2} - \alpha_2^{2k-2}) - (\alpha_1^{2k-4} - \alpha_2^{2k-4}) \right)$$

*elements, where  $b = (M - 2)(N - 2) - 2$  and*

$$\alpha_{1,2} = \frac{\sqrt{b - 2} \pm \sqrt{b + 2}}{2}$$

*are the two roots of the equation  $\alpha^2 - \sqrt{b - 2}\alpha - 1 = 0$ . The action of  $T$  on the set of rank  $k$   $N$ -gons is a cyclic permutation  $i \mapsto i + p_k$ , where*

$$p_k = \frac{q_k}{M} + \frac{M - 2}{(b - 2)\sqrt{b + 2}} \left( (b - 1)(\alpha_1^{2k-3} - \alpha_2^{2k-3}) - (\alpha_1^{2k-5} - \alpha_2^{2k-5}) \right),$$

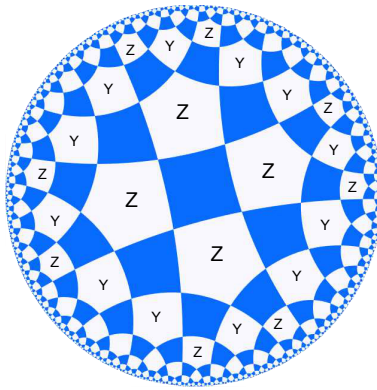
*and the action of  $T$  on the set of rank  $k$   $M$ -gons is also a cyclic permutation  $i \mapsto i + j_k$ , where*

$$j_k = \frac{l_k}{M} + \frac{1}{(b - 2)\sqrt{b + 2}} \left( (b^2 - 2)(\alpha_1^{2k-3} - \alpha_2^{2k-3}) - b(\alpha_1^{2k-5} - \alpha_2^{2k-5}) \right).$$

*The rotation number of the outer billiard map at infinity is given by the formula*

$$\rho(t) = \frac{1}{M} + \frac{M - 2}{M\sqrt{b - 2}\alpha_1} \frac{(b - 1)\alpha_1^2 - 1}{(b + 1)\alpha_1^2 - 1}.$$

**Remark 12.** If  $N = M$ , the statement of [Theorem 11](#) reduces to that of [Theorem 1](#).



**Figure 5.** A  $(4, 5)$ -tiling, with rank 1 and rank 2 pentagons labeled either as type  $Y$  (one parent) or as type  $Z$  (no parents).

The proof of [Theorem 11](#) also consists of several steps.

**Claim 13.** *Every orbit of  $T$  is periodic.*

*Proof.* The proof of this claim is analogous to the proof in the previous section. Because the rank of each tile is preserved under the billiard map, and because there are finitely many tiles of a given rank, every tile must map back to itself after some finite number of iterations  $m$ . When the tile maps back to itself, it has rotated by  $2\pi j/M$  if it is an  $M$ -gon or by  $2\phi j/N$  if it is an  $N$ -gon. Then  $T^{omM}$  is the identity if the tile is an  $M$ -gon and  $T^{omN}$  is the identity if the tile is an  $N$ -gon.  $\square$

**Claim 14.** *For every  $k \geq 1$ ,  $T$  permutes the rank  $k$  tiles cyclically.*

*Proof.* Proof is similar to that for [Claim 5](#).  $\square$

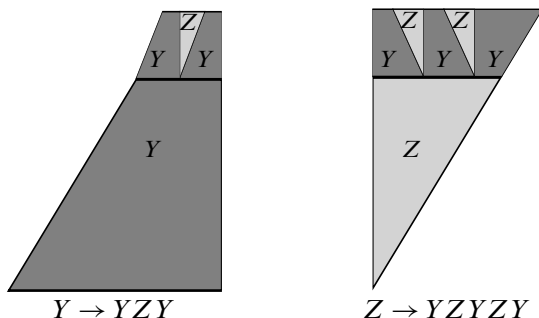
Recall that type  $Y$  tiles have one parent and type  $Z$  tiles have zero parents (see [Figure 5](#)). We now give a crochet pattern for general  $(M, N)$ -tilings,  $M \geq 4$ .

**Claim 15.** *The following replacement rules hold for  $(M, N)$ -tilings:*

$$Y \rightarrow (YZ^{M-3})^{N-4}YZ^{M-4}, \tag{1}$$

$$Z \rightarrow (YZ^{M-3})^{N-3}YZ^{M-4}. \tag{2}$$

*Proof.* In a similar manner to the  $(3, N)$  case, we represent type  $Y$  and  $Z$  tiles as degenerate polygons, with additional vertices. See [Figure 6](#) for illustrations of the  $(4, 5)$  case. Type  $Y$  tiles are represented as quadrilaterals with  $N$  vertices, and type  $Z$  tiles are represented as triangles with  $N$  vertices. Because a  $Y$  tile has  $N-3$  sides available to connect with a tile of higher rank, a rank  $k$   $Y$  tile produces  $N-3$   $Y$  tiles of rank  $k+1$ . Then, since tiles must meet  $M$  to a vertex, there must be  $M-3$   $Z$  tiles between every pair of  $Y$  tiles, and there must be  $M-4$  type  $Z$



**Figure 6.** Tiling of pentagons meeting in fours. Can be extended to a (4, 5)-tiling.

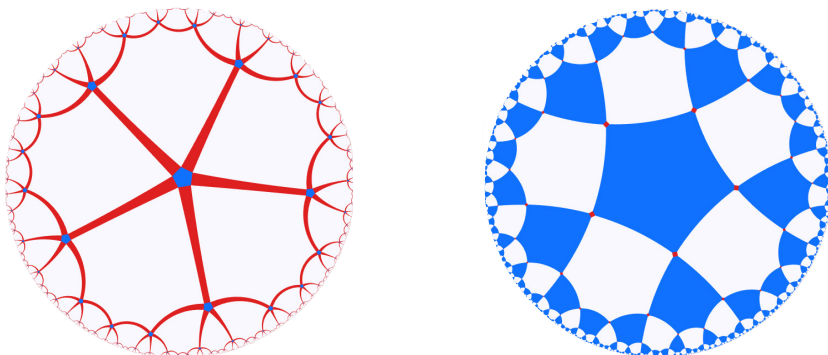
tiles following the last  $Y$ . Similarly, a  $Z$  tile has  $N - 2$  edges free to connect to a tile of higher rank, so a rank  $k$   $Z$  tile produces  $N - 2$   $Y$  tiles of rank  $k + 1$ , again with  $Z$  tiles appropriately interspersed.

This crochet pattern tiles the hyperbolic plane with  $M$   $N$ -gons meeting at every vertex. From this tiling, we obtain the  $(M, N)$ -tiling by considering the points in the tiling becoming  $M$ -gons, as in Figure 7 (compare with [Conway et al. 2008]). The described crochet pattern translates to the replacement rules given above.  $\square$

We can now compute the formulas for the number of  $M$ - and  $N$ -gons of any rank, as well as for the cyclic permutation of  $M$ - and  $N$ -gons of any rank.

**Claim 16.** *The formulas for  $q_k, p_k, j_k, l_k$  hold as stated in Theorem 11.*

*Proof.* Denoting the number of  $Y$ -type and  $Z$ -type  $N$ -gons of rank  $k$  by  $y_k$  and  $z_k$ ,



**Figure 7.** Left: Tiling of the plane by hexagons meeting in fives. Right: by replacing the vertices in the previous picture with pentagons, we achieve a (5, 6)-tiling. Here two hexagons and two pentagons meet at each single vertex.

we obtain the recursion formula

$$\begin{pmatrix} y_k \\ z_k \end{pmatrix} = A \begin{pmatrix} y_{k-1} \\ z_{k-1} \end{pmatrix}, \tag{3}$$

where the matrix  $A$  is obtained from the rules given in (1) and (2), and

$$A = \begin{pmatrix} N-3 & N-2 \\ (M-3)(N-3)-1 & (M-3)(N-2)-1 \end{pmatrix}. \tag{4}$$

As mentioned above, the initial conditions are  $\begin{pmatrix} y_1 \\ z_1 \end{pmatrix} = \begin{pmatrix} 0 \\ M \end{pmatrix}$ .

Solving the recurrence, we find the general formula

$$\begin{pmatrix} y_k \\ z_k \end{pmatrix} = \begin{pmatrix} \frac{M(N-2)(\alpha_1^{2k-2} - \alpha_2^{2k-2})}{\sqrt{b^2-4}} \\ \frac{M((M-3)(N-2)-1)(\alpha_1^{2k-2} - \alpha_2^{2k-2}) + M(\alpha_2^{2k-4} - \alpha_1^{2k-4})}{\sqrt{b^2-4}} \end{pmatrix},$$

where

$$b = (M-2)(N-2)-2, \quad \alpha_1 = \frac{\sqrt{b-2} + \sqrt{b+2}}{2}, \quad \alpha_2 = \frac{\sqrt{b-2} - \sqrt{b+2}}{2}.$$

Then  $q_k = y_k + z_k$ , so

$$q_k = \frac{M}{\sqrt{b^2-4}} ((b+1)(\alpha_1^{2k-2} - \alpha_2^{2k-2}) + \alpha_2^{2k-4} - \alpha_1^{2k-4}).$$

Now that we have counted the  $N$ -gons, we count the  $M$ -gons of rank  $k$  by noticing a pattern in the tiling. We see that a type  $Y$   $N$ -gon of rank  $k$  produces  $N-3$   $M$ -gons of rank  $k$ , and a type  $Z$   $N$ -gon produces  $N-2$   $M$ -gons. Thus the number of  $M$ -gons of rank  $k$  is given by  $l_k = (N-3)y_k + (N-2)z_k$ . The formula for  $l_k$  given in [Theorem 11](#) follows.

Next we determine  $p_k$  by counting how many tiles a rank  $k$   $N$ -gon jumps when  $T$  is applied. As in the previous section, we define  $s_k$  as the number of rank  $k$   $N$ -gons in a small cone. We call  $y_k^s$  and  $z_k^s$  the number of rank  $k$   $Y$ s and  $Z$ s in the small cone. Also, as before, applying  $T$  to any tile causes the tile to jump over two small cones and one big cone. In total, the jump is given by  $p_k = s_k + q_k/M$ .

We observe that

$$\begin{pmatrix} y_k^s \\ z_k^s \end{pmatrix} = \sum_{i=0}^{k-2} A^i \begin{pmatrix} 2 \\ M-4 \end{pmatrix}, \tag{5}$$

where  $A$  is given in (4).

This becomes

$$\begin{pmatrix} y_k^s \\ z_k^s \end{pmatrix} = \begin{pmatrix} \frac{1-\alpha_1^{2k-2}}{1-\alpha_1^2} + \frac{2}{\sqrt{b^2-4}} \frac{1-\alpha_2^{2k-2}}{1-\alpha_2^2} (\alpha_1^2 - N + 3) \\ \frac{1}{\sqrt{b^2-4}} \left( \frac{1-\alpha_1^{2k-2}}{1-\alpha_1^2} (B - \alpha_2^2(M-4)) + \frac{1-\alpha_2^{2k-2}}{1-\alpha_2^2} (-B + \alpha_1^2(M-4)) \right) \end{pmatrix},$$

where  $B = (M - 3)(b - 2) + (M - 4)$ . Then, since  $s_k = y_k^s + z_k^s$ , we have

$$s_k = \frac{M - 2}{(b - 2)\sqrt{b + 2}} ((b - 1)(\alpha_1^{2k-3} + \alpha_2^{2k-3}) + \alpha_2^{2k-5} - \alpha_1^{2k-5}).$$

This allows us to calculate  $p_k$ , and we can compute  $j_k$  by noticing again that every  $Y$ -type  $N$ -gon will be replaced by  $N - 3$   $M$ -gons and every  $Z$ -type  $(N - 2)$ -gon will be replaced by  $N - 3$   $M$ -gons on the next level. This procedure will leave again only one  $M$ -gon out, so  $j_k = (N - 3)y_k^s + (N - 2)z_k^s + 1$ . □

**Claim 17.** *The rotation number is given by*

$$\rho(t) = \frac{1}{M} + \frac{M - 2}{M\sqrt{b - 2}\alpha_1} \frac{(b - 1)\alpha_1^2 - 1}{(b + 1)\alpha_1^2 - 1}.$$

*Proof.* This results from taking the limit of  $p_k/q_k$  as  $k \rightarrow \infty$ . □

### 4. Remarks and acknowledgments

The methods used in this paper both for geometrical and counting arguments can be used also for all other tilings with 2-fold symmetries in the vertices and so we believe that similar theorems and observations can be deduced in a more general setting.

We want to acknowledge the work of two of our colleagues during the program. Stephanie Ger classified the tilings of the hyperbolic plane in terms of symmetry groups, and Ananya Uppal provided us with a Mathematica demonstration that created images that helped to understand the patterns in the tilings. Their contributions were extremely valuable for our work.

We also want to thank our advisors during the Summer@ICERM program, Chaim Goodman-Strauss, Sergei Tabachnikov, Ryan Greene and Tarik Aougab, for all the help and the inspiring discussions. Finally, we thank ICERM for providing the opportunity for us to participate in this great program.

The tiling pictures were created using *KaleidoTile* application created by Jeff Weeks ([geometrygames.org/contact.html](http://geometrygames.org/contact.html)).

### References

[Conway et al. 2008] J. H. Conway, H. Burgiel, and C. Goodman-Strauss, *The symmetries of things*, A K Peters, Wellesley, MA, 2008. [MR 2009c:00002](#) [Zbl 1173.00001](#)



- [Dogru and Tabachnikov 2003] F. Dogru and S. Tabachnikov, “On polygonal dual billiard in the hyperbolic plane”, *Regul. Chaotic Dyn.* **8**:1 (2003), 67–81. [MR 2004b:37064](#) [Zbl 1075.37524](#)
- [Greenberg 1980] M. J. Greenberg, *Euclidean and non-Euclidean geometries: Development and history*, 2nd ed., W. H. Freeman, San Francisco, CA, 1980. [MR 81f:51001](#) [Zbl 0418.51001](#)
- [Gutkin and Simányi 1992] E. Gutkin and N. Simányi, “Dual polygonal billiards and necklace dynamics”, *Comm. Math. Phys.* **143**:3 (1992), 431–449. [MR 92k:58139](#) [Zbl 0749.58032](#)
- [Kołodziej 1989] R. Kołodziej, “The antibilliard outside a polygon”, *Bull. Polish Acad. Sci. Math.* **37**:1-6 (1989), 163–168. [MR 92g:52002](#) [Zbl 0762.58021](#)
- [Moser 1973] J. Moser, *Stable and random motions in dynamical systems: With special emphasis on celestial mechanics*, Annals of Mathematics Studies **77**, Princeton University Press, 1973. [MR 56 #1355](#) [Zbl 0271.70009](#)
- [Moser 1978] J. Moser, “Is the solar system stable?”, *Math. Intelligencer* **1**:2 (1978), 65–71. [MR 58 #14029](#) [Zbl 1288.70008](#)
- [Neumann 1959] B. Neumann, “Sharing ham and eggs”, *Iota* (1959), 14–18.
- [Schwartz 2007] R. E. Schwartz, “Unbounded orbits for outer billiards, I”, *J. Mod. Dyn.* **1**:3 (2007), 371–424. [MR 2008f:37082](#) [Zbl 1130.37021](#)
- [Schwartz 2009] R. E. Schwartz, *Outer billiards on kites*, Annals of Mathematics Studies **171**, Princeton University Press, 2009. [MR 2011a:37081](#) [Zbl 1205.37001](#)
- [Tabachnikov 2005] S. Tabachnikov, *Geometry and billiards*, Student Mathematical Library **30**, Amer. Math. Soc., Providence, RI, 2005. [MR 2006h:51001](#) [Zbl 1119.37001](#)
- [Tabachnikov 2007] S. Tabachnikov, “A proof of Culter’s theorem on the existence of periodic orbits in polygonal outer billiards”, *Geom. Dedicata* **129** (2007), 83–87. [MR 2008m:37062](#) [Zbl 1131.37039](#)
- [Tabachnikov and Dogru 2005] S. Tabachnikov and F. Dogru, “Dual billiards”, *Math. Intelligencer* **27**:4 (2005), 18–25. [MR 2006i:37121](#) [Zbl 1088.37014](#)
- [Vivaldi and Shaidenko 1987] F. Vivaldi and A. V. Shaidenko, “Global stability of a class of discontinuous dual billiards”, *Comm. Math. Phys.* **110**:4 (1987), 625–640. [MR 89c:58067](#) [Zbl 0653.58018](#)

Received: 2014-05-28

Accepted: 2014-06-20

[dogruf@gvsu.edu](mailto:dogruf@gvsu.edu)

*Department of Mathematics, Grand Valley State University,  
Allendale, MI 49401, United States*

[efischer@hmc.edu](mailto:efischer@hmc.edu)

*Harvey Mudd College, Claremont, CA 91711, United States*

[mihaim92@gmail.com](mailto:mihaim92@gmail.com)

*Department of Mathematics, Jacobs University Bremen,  
D-28759 Bremen, Germany*



# Sophie Germain primes and involutions of $\mathbb{Z}_n^\times$

Karena Genzlinger and Keir Lockridge

(Communicated by Kenneth S. Berenhaut)

In the paper “What is special about the divisors of 24?”, Sunil Chebolu proved an interesting result about the multiplication tables of  $\mathbb{Z}_n$  from several different number theoretic points of view: all of the 1s in the multiplication table for  $\mathbb{Z}_n$  are located on the main diagonal if and only if  $n$  is a divisor of 24. Put another way, this theorem characterizes the positive integers  $n$  with the property that the proportion of 1s on the diagonal is precisely  $1/24$ . The present work is concerned with finding the positive integers  $n$  for which there is a given fixed proportion of 1s on the diagonal. For example, when  $p$  is prime, we prove that there exists a positive integer  $n$  such that  $1/p$  of the 1s lie on the diagonal of the multiplication table for  $\mathbb{Z}_n$  if and only if  $p$  is a Sophie Germain prime.

1. Introduction	653
2. The ratio of diagonal units	655
3. Sophie Germain factorizations	656
4. Examples	659
5. The multiplication cube for $\mathbb{Z}_n$	662
References	663

## 1. Introduction

Let  $R$  be a ring and let  $R^\times$  denote its group of units. Call a unit  $u$  in  $R^\times$  a *diagonal unit* if the multiplicative order of  $u$  is at most 2. Such units are more commonly referred to as involutions; our motivation for calling them diagonal units is as follows. The units of  $R$  are in one-to-one correspondence with 1s appearing in its multiplication table, and the diagonal units are in one-to-one correspondence with the 1s appearing on the diagonal. When the order of  $R^\times$  is finite, we will write  $\text{du}(R)$  for the number of diagonal units and

$$\text{pdu}(R) = \frac{\text{du}(R)}{|R^\times|}$$

*MSC2010*: primary 11A41; secondary 16U60.

*Keywords*: Sophie Germain primes, group of units, Gauss–Wantzel theorem.

for the proportion of diagonal units in  $R^\times$ . We will only consider commutative rings, in which case  $R^\times$  is an abelian group. This means that the units of order at most 2 form a subgroup of  $R^\times$ . Hence,  $\text{du}(R)$  divides  $|R^\times|$  by Lagrange’s theorem, so  $\text{pdu}(R)$  is always the reciprocal of an integer. We therefore find it more convenient to work with the *ratio of diagonal units*,

$$\text{rdu}(R) = \frac{|R^\times|}{\text{du}(R)} = \frac{1}{\text{pdu}(R)}.$$

For brevity, we will write  $\text{du}(n)$ ,  $\text{pdu}(n)$  and  $\text{rdu}(n)$  for the quantities  $\text{du}(\mathbb{Z}_n)$ ,  $\text{pdu}(\mathbb{Z}_n)$ , and  $\text{rdu}(\mathbb{Z}_n)$ .

A ring  $R$  is said to *satisfy the diagonal property* if every unit of  $R$  is a diagonal unit; that is,  $R$  satisfies the diagonal property if and only if  $\text{pdu}(R) = \text{rdu}(R) = 1$ . Chebolu [2012] proved that  $\mathbb{Z}_n$  satisfies the diagonal property if and only if  $n$  is a divisor of 24. This leads naturally to a more general study of the equation

$$\text{rdu}(n) = \theta, \tag{1}$$

where  $\theta \geq 1$ . For which values of  $\theta$  does (1) have a solution? If (1) has a solution, can we find the entire solution set? We will answer both of these questions in several cases in Section 4. For example, we will prove the following theorem, which answers both questions when  $\theta$  is prime.

**Theorem 1.1.** *Let  $p$  be a prime. There exists a positive integer  $n$  such that the proportion of diagonal units in  $\mathbb{Z}_n$  is  $1/p$  if and only if  $p$  is a Sophie Germain prime. For a Sophie Germain prime  $p$ , the set of solutions to  $\text{rdu}(n) = p$  is*

$$\begin{aligned} &(2p + 1) \cdot \{\text{divisors of } 24\} \quad \text{if } p > 3, \\ &(2p + 1) \cdot \{\text{divisors of } 24\} \cup p^2 \cdot \{\text{divisors of } 8\} \quad \text{if } p = 3, \\ &(2p + 1) \cdot \{\text{divisors of } 24\} \cup p^4 \cdot \{\text{divisors of } 3\} \quad \text{if } p = 2. \end{aligned}$$

A Sophie Germain prime is a prime  $p$  such that  $2p + 1$  is also prime, in which case  $2p + 1$  is called a safe prime. Such primes arose in Marie-Sophie Germain’s considerable work on Fermat’s last theorem (see [Laubenbacher and Pengelley 1999]).

The remainder of this paper is organized as follows. Section 2 includes background information and a formula for the ratio of diagonal units. We then prove in Section 3 that the equation  $\text{rdu}(n) = \theta$  has a solution if and only if  $\theta$  admits a special type of factorization, and we provide a principle for organizing solutions to this equation given a list of these factorizations. Section 4 is devoted to examples, including proofs of Chebolu’s 24 theorem and Theorem 1.1. We also explore a surprising connection between the proportion of diagonal units and the Gauss–Wantzel theorem on the constructibility of regular polygons (Theorem 4.2). In the last section, we consider a generalization of the current situation and examine 1s on the diagonal of the multiplication cube for  $\mathbb{Z}_n$ .

## 2. The ratio of diagonal units

A common concept in number theory is the notion of a multiplicative function. A function  $f : \mathbb{Z}^+ \rightarrow \mathbb{Z}^+$  is *multiplicative* if  $f(st) = f(s)f(t)$  whenever  $s$  and  $t$  are relatively prime. Euler’s totient function is an example of a multiplicative function par excellence (see [Burton 1989, §7]); it counts the positive integers  $k \leq n$  that are relatively prime to  $n$ . The relevant properties of  $\phi(n)$  are summarized in the next theorem.

**Theorem 2.1** (Euler’s totient function). *Let  $\phi(n)$  denote the number of positive integers less than  $n$  and relatively prime to  $n$ .*

- (A) *The order of  $\mathbb{Z}_n^\times$  is precisely  $\phi(n)$ .*
- (B) *The function  $\phi(n)$  is multiplicative.*
- (C) *For any prime  $p$  and positive integer  $k$ , we have  $\phi(p^k) = p^{k-1}(p - 1)$ .*

We now prove that the functions defined in Section 1 are multiplicative.

**Proposition 2.2.** *The functions  $\text{du}(n)$ ,  $\text{pdu}(n)$ , and  $\text{rdu}(n)$  are multiplicative.*

*Proof.* Certainly,  $\text{rdu}(n)$  is multiplicative if and only if  $\text{pdu}(n)$  is multiplicative. Since  $\text{rdu}(n) = |\mathbb{Z}_n^\times| / \text{du}(n) = \phi(n) / \text{du}(n)$  and  $\phi$  is multiplicative by the previous theorem, it suffices to prove that  $\text{du}(n)$  is multiplicative.

Let  $s$  and  $t$  be relatively prime positive integers. By the Chinese remainder theorem,  $\mathbb{Z}_{st} \cong \mathbb{Z}_s \times \mathbb{Z}_t$ . Since the order of  $(x, y) \in \mathbb{Z}_s \times \mathbb{Z}_t$  is the least common multiple of the orders of  $x$  and  $y$ , the pair  $(x, y)$  is a diagonal unit if and only if  $x$  and  $y$  are diagonal units. Thus,  $\text{du}(st) = \text{du}(s) \text{du}(t)$ . □

Our next goal is to give a formula for  $\text{rdu}(n)$ . To do so, we need one more ingredient.

**Theorem 2.3** (isomorphism class of  $\mathbb{Z}_n^\times$ ). *For any integer  $k \geq 1$  and odd prime  $p$ ,*

$$\mathbb{Z}_{p^k}^\times \cong \mathbb{Z}_{\phi(p^k)} = \mathbb{Z}_{p^{k-1}(p-1)},$$

and

$$\mathbb{Z}_{2^k}^\times \cong \begin{cases} \{1\} & \text{if } k = 1, \\ \mathbb{Z}_2 & \text{if } k = 2, \\ \mathbb{Z}_2 \times \mathbb{Z}_{2^{k-2}} & \text{if } k \geq 3. \end{cases}$$

The odd primary case is a consequence of the primitive root theorem; see [Cohen 2007, 2.1.24] for a short, fairly self-contained proof.

The next proposition provides a formula for the ratio of diagonal units in  $\mathbb{Z}_n$ .

**Proposition 2.4.** *Let  $n$  be a positive integer.*

- (A) *For any odd prime  $p$  and integer  $k \geq 1$ ,*

$$\text{rdu}(p^k) = \phi(p^k) / 2 = p^{k-1}(p - 1) / 2.$$

(B) For any integer  $k \geq 0$ ,

$$\text{rd}u(2^k) = \begin{cases} 1 & \text{if } k = 0, 1, 2, \text{ or } 3, \\ 2^{k-3} & \text{if } k > 3. \end{cases}$$

(C) Let  $n = 2^a 3^b n'$ , where  $a, b \geq 0$  and  $(n', 6) = 1$ . Let  $r$  denote the number of distinct primes dividing  $n'$ . Then,

$$\text{rd}u(n) = \begin{cases} \phi(n')/2^r & \text{if } a \leq 3, b \leq 1, \\ 2^{a-3} \phi(n')/2^r & \text{if } a > 3, b \leq 1, \\ 3^{b-1} \phi(n')/2^r & \text{if } a \leq 3, b > 1, \\ 2^{a-3} 3^{b-1} \phi(n')/2^r & \text{if } a > 3, b > 1. \end{cases}$$

*Proof.* By [Theorem 2.1\(A\)](#),  $\text{rd}u(n) = \phi(n)/\text{d}u(n)$ . Next observe that  $\text{d}u(2) = 1$ ,  $\text{d}u(4) = 2$ ,  $\text{d}u(2^k) = 4$  for  $k \geq 3$ , and  $\text{d}u(p^k) = 2$  for any odd prime  $p$  by [Theorem 2.3](#) (for the last case, note that the group of units is cyclic of even order, so it has a unique subgroup of order 2). Combining these facts with the formula for  $\phi(p^k)$  given in [Theorem 2.1\(C\)](#), one obtains parts (A) and (B). Part (C) follows from the previous two parts and the fact that  $\text{rd}u(n)$  is multiplicative.  $\square$

Though it is likely no surprise that the prime 2 is isolated in [Proposition 2.4\(C\)](#), our reason for isolating the prime 3 may be unclear. For now, we hope the reader is content with the observation that 2 and 3 are the only prime divisors of 24. In slightly more detail, the issue has to do with the fact that if  $p > 3$  is prime, then  $\text{rd}u(p^k) = p$  is impossible, but  $\text{rd}u(3^2) = 3$  and  $\text{rd}u(2^4) = 2$ . Note further that  $\text{rd}u(p^k)$  in [Proposition 2.4\(A\)](#) factors as  $\theta(2\theta + 1)^k$ , where  $2\theta + 1$  is prime. This hints at the relevance of Sophie Germain primes, which appeared in [Theorem 1.1](#), and leads to the study of positive integers that admit the special type of factorization discussed in the following section.

### 3. Sophie Germain factorizations

Given a positive integer  $\theta$ , a *Sophie Germain factorization* of  $\theta$  is a triple

$$F = (s, t, \{(\theta_1, \beta_1), \dots, (\theta_r, \beta_r)\}),$$

where

(A)  $\theta = |F| = 2^s 3^t \prod_{i=1}^r \theta_i (2\theta_i + 1)^{\beta_i}$ ,

(B)  $s \geq 0$  and  $t \geq 0$ ;

(C) for  $i = 1, \dots, r$ ,  $\beta_i \geq 0$  and  $\theta_i > 1$ ; and

(D) the integers  $2\theta_1 + 1, \dots, 2\theta_r + 1$  are distinct primes.

When  $r = 0$ , the set in the third coordinate of  $F$  is empty and the indexed product in (A) is 1. The ordered triple gives the data for the factorization, but the definition

of  $|F|$  gives a far more readable interpretation of what the data represent. We will therefore abuse notation and use the expression defining  $|F|$  in place of  $F$  itself. There is some ambiguity, however, since  $\theta_i$  can be 2 or 3; consequently, we will always include the exponents  $s$  and  $t$  unsimplified, even when  $s = 1$  or  $t = 1$ , unless either is equal to zero, in which case we will omit the corresponding factor entirely. We will not omit zero exponents in the indexed product, and the empty product will appear as 1. For clarification, here are several examples:

$$\begin{aligned} |(0, 0, \emptyset)| &= 1, \\ |(0, 0, \{(3, 0)\})| &= 3 \cdot 7^0, \\ |(0, 1, \emptyset)| &= 3^1 \cdot 1, \\ |(5, 1, \{(3, 0), (5, 2), (9, 4)\})| &= 2^5 \cdot 3^1 \cdot 3 \cdot 7^0 \cdot 5 \cdot 11^2 \cdot 9 \cdot 19^4. \end{aligned}$$

The main difficulty of our current undertaking is to find all possible such factorizations of a given positive integer. However, given a list of the Sophie Germain factorizations of  $\text{rdu}(n)$ , we will see at the end of this section that it is easy to find all solutions to (1).

Let  $\mathcal{S}$  denote the set of all Sophie Germain factorizations of positive integers. We next define two functions,

$$\mathcal{F} : \mathbb{Z}^+ \rightarrow \mathcal{S}$$

and

$$\mathcal{N} : \mathcal{S} \rightarrow \mathbb{Z}^+.$$

The function  $\mathcal{F}(n)$  will select a canonical Sophie Germain factorization of  $n$ , and the function  $\mathcal{N}(F)$  will select a positive integer whose canonical Sophie Germain factorization is  $F$ . Let

$$\mathcal{F}\left(2^a 3^b \prod_{i=1}^r p_i^{\alpha_i}\right) = \begin{cases} \prod_{i=1}^r ((p_i - 1)/2) \cdot p_i^{\alpha_i - 1} & \text{if } a \leq 3, b \leq 1, \quad (1) \\ 2^{a-3} \cdot \prod_{i=1}^r ((p_i - 1)/2) \cdot p_i^{\alpha_i - 1} & \text{if } a > 3, b \leq 1, \quad (2) \\ 3^{b-1} \cdot \prod_{i=1}^r ((p_i - 1)/2) \cdot p_i^{\alpha_i - 1} & \text{if } a \leq 3, b > 1, \quad (3) \\ 2^{a-3} 3^{b-1} \cdot \prod_{i=1}^r ((p_i - 1)/2) \cdot p_i^{\alpha_i - 1} & \text{if } a > 3, b > 1, \quad (4) \end{cases}$$

and let

$$\mathcal{N}\left(2^s 3^t \cdot \prod_{i=1}^r \theta_i (2\theta_i + 1)^{\beta_i}\right) = \begin{cases} \prod_{i=1}^r (2\theta_i + 1)^{\beta_i + 1} & \text{if } s = 0, t = 0, \\ 2^{s+3} \cdot \prod_{i=1}^r (2\theta_i + 1)^{\beta_i + 1} & \text{if } s > 0, t = 0, \\ 3^{t+1} \cdot \prod_{i=1}^r (2\theta_i + 1)^{\beta_i + 1} & \text{if } s = 0, t > 0, \\ 2^{s+3} 3^{t+1} \cdot \prod_{i=1}^r (2\theta_i + 1)^{\beta_i + 1} & \text{if } s > 0, t > 0, \end{cases}$$

The indexed product in the definition of  $\mathcal{F}$  is of course just  $\phi(n')/2^r$ , where  $n' = n/(2^a 3^b)$ . In the definition of  $\mathcal{F}$ , we have labeled the cases 1–4. Every Sophie

Germain factorization falls into precisely one of these four cases, so we will use these numbers to refer to the *type* of a Sophie Germain factorization. If we were to only consider integers relatively prime to 6, the above formulas would each have a single case and these functions would be inverses; interference from the divisors of 24 causes a bit of trouble. We summarize the relevant properties of  $\mathcal{F}$  and  $\mathcal{N}$  in the following theorem.

**Proposition 3.1.** *Let  $\mathcal{F} : \mathbb{Z}^+ \rightarrow \mathcal{S}$  and  $\mathcal{N} : \mathcal{S} \rightarrow \mathbb{Z}^+$  be the functions defined above.*

(A) *For any positive integer  $n$ ,  $\text{rd}(n) = |\mathcal{F}(n)|$ .*

(B) *For any Sophie Germain factorization  $F$ ,*

$$\mathcal{F}(\mathcal{N}(F)) = F.$$

*In particular,  $\mathcal{F}$  is surjective.*

*Proof.* The verification of each statement entails a straightforward computation using the definitions of  $\mathcal{F}$  and  $\mathcal{N}$  combined with [Proposition 2.4\(C\)](#).  $\square$

We now have the following general result.

**Theorem 3.2.** *Fix a positive integer  $\theta$ . The equation*

$$\text{rd}(n) = \theta \tag{2}$$

*has a solution if and only if  $\theta$  admits a Sophie Germain factorization.*

*Proof.* If (2) has a solution, then  $\theta = |\mathcal{F}(n)|$ , so  $\theta$  admits a Sophie Germain factorization. Conversely, if  $|F| = \theta$ , then take  $n = \mathcal{N}(F)$ . Now,

$$\text{rd}(n) = |\mathcal{F}(\mathcal{N}(F))| = |F| = \theta. \quad \square$$

It may feel at this point that we have saddled the reader with a great deal of notation without having accomplished much, given that the true difficulty is finding all possible Sophie Germain factorizations. However, given the set of factorizations, the following proposition provides a nice principle for organizing the solutions to (2). It measures the failure of  $\mathcal{F}$  to be injective, and it is the main reason we have defined  $\mathcal{F}$  and  $\mathcal{N}$ . The proof amounts to a reflection upon the meaning of the conditions used to divide the definition of  $\mathcal{F}$  into four cases.

**Proposition 3.3.** *Let  $F_i$  be a Sophie Germain factorization of type  $i$ . Then,*

$$\mathcal{F}^{-1}(F_1) = \mathcal{N}(F_1) \cdot \{\text{divisors of } 24\},$$

$$\mathcal{F}^{-1}(F_2) = \mathcal{N}(F_2) \cdot \{\text{divisors of } 3\},$$

$$\mathcal{F}^{-1}(F_3) = \mathcal{N}(F_3) \cdot \{\text{divisors of } 8\},$$

$$\mathcal{F}^{-1}(F_4) = \mathcal{N}(F_4) \cdot \{1\}.$$

We will use the above proposition in the next section.



### 4. Examples

Thankfully, it is now time to more concretely investigate the possible proportions of diagonal units using the tools developed above. We begin with Chebolu's theorem [2012].

**4A. Chebolu's 24 theorem.** We include this example for completeness; certainly, the proofs given in [Chebolu 2012] are either more direct or more interesting, or both.

**Theorem 4.1** (Chebolu). *The ring  $\mathbb{Z}_n$  satisfies the diagonal property if and only if  $n$  is a divisor of 24.*

*Proof.* We seek all possible solutions to  $\text{rdu}(n) = 1$ . Since the integer 1 has the unique (type 1) Sophie Germain factorization  $1$ , the solution set is

$$\begin{aligned} \mathcal{N}(1) \cdot \{\text{divisors of } 24\} &= 1 \cdot \{\text{divisors of } 24\} \\ &= \{\text{divisors of } 24\}. \end{aligned}$$

by Proposition 3.3. □

**4B. Proof of Theorem 1.1.** It is straightforward to check that when  $p$  is a Sophie Germain prime, the listed sets provide solutions to  $\text{rdu}(n) = p$ . We therefore turn our attention to the converse.

Let  $p > 3$  be prime and suppose  $\text{rdu}(n) = p$  has a solution, in which case  $p$  admits a Sophie Germain factorization. Any such factorization  $F$  of  $p$  must have  $s = t = 0$  and  $r = 1$  since  $p$  cannot have more than one distinct prime factor. Hence,  $|F| = \theta(2\theta + 1)^\beta$ . Further, since  $\theta(2\theta + 1)^\beta = p$  and  $\theta > 1$ , we must have  $\theta = p$  and  $\beta = 0$ . Thus,

$$p \cdot (2p + 1)^0$$

is the only possible Sophie Germain factorization of  $p$ . This forces  $2p + 1$  to be prime, so  $p$  is a Sophie Germain prime and the set of solutions to  $\text{rdu}(n) = p$  is

$$\mathcal{N}(p \cdot (2p + 1)^0) \cdot \{\text{divisors of } 24\} = (2p + 1) \cdot \{\text{divisors of } 24\}$$

by Proposition 3.3.

For  $p = 2$ , the only Sophie Germain factorizations of 2 are  $2 \cdot 5^0$  and  $2^1 \cdot 1$ . The first factorization has type 1, and the second has type 2. Note that  $\mathcal{N}(2 \cdot 5^0) = 5$  and  $\mathcal{N}(2^1 \cdot 1) = 16$ . Hence, the set of solutions to  $\text{rdu}(n) = 2$  is

$$5 \cdot \{\text{divisors of } 24\} \cup 16 \cdot \{\text{divisors of } 3\}.$$

Finally, for  $p = 3$ , we have the type 1 factorization  $3 \cdot 7^0$  with  $\mathcal{N} = 7$  and the type 3 factorization  $3^1 \cdot 1$  with  $\mathcal{N} = 9$ . Hence, the set of solutions to  $\text{rdu}(n) = 3$  is

$$7 \cdot \{\text{divisors of } 24\} \cup 9 \cdot \{\text{divisors of } 8\}.$$

This completes the proof of Theorem 1.1.

**4C. Prime power ratios.** We now consider the more general case  $\text{rdu}(n) = p^k$  for  $k \geq 1$ . First, assume  $p > 3$  is prime.

Any Sophie Germain factorization of  $p^k$  must have the property that  $s = t = 0$  and each  $\theta_i$  is a positive power of  $p$ . Since  $p$  cannot divide both  $\theta_i$  and  $2\theta_i + 1$ , we must have  $\beta_i = 0$  for all  $i$ . Thus, every Sophie Germain factorization must have the form

$$\prod_{i=1}^r p^{k_i} (2p^{k_i} + 1)^0,$$

where the integers  $2p^{k_1} + 1, \dots, 2p^{k_r} + 1$  are distinct primes and  $\sum k_i = k$  is a partition of  $k$  into distinct odd parts (each  $k_i$  is odd because  $2p^v + 1$  is divisible by 3 whenever  $v$  is even). Each such factorization contributes

$$\prod_{i=1}^r (2p^{k_i} + 1) \cdot \{\text{divisors of } 24\}$$

to the set of solutions to  $\text{rdu}(n) = p^k$ . Here are several examples of what may be gleaned from this discussion:

- (A) There is no solution to  $\text{rdu}(n) = p^k$  when  $p \equiv 1 \pmod{3}$  (since this implies that  $2p^v + 1$  is always divisible by 3).
- (B) There is no solution to  $\text{rdu}(n) = p^2$  since there is no partition of 2 into distinct odd parts.
- (C) There is a solution to  $\text{rdu}(n) = p^4$  if and only if  $2p + 1$  and  $2p^3 + 1$  are both prime.
- (D) There is a solution to  $\text{rdu}(n) = p^7$  if and only if  $2p^7 + 1$  is prime.
- (E) There is a solution to  $\text{rdu}(n) = p^8$  if and only if either  $\{2p + 1, 2p^7 + 1\}$  or  $\{2p^3 + 1, 2p^5 + 1\}$  is a set of primes.

The prime  $p = 5$  illustrates (C). The prime  $p = 677$ , which is not a Sophie Germain prime, illustrates (D). For (E),  $p = 29$  is a prime where both of the indicated sets are sets of primes;  $p = 149$  is a prime where the second set is a set of primes and neither element of the first set is prime;  $p = 179$  is a prime where the first set is a set of primes and neither element of the second set is prime.

The situations for the primes 2 and 3 are similar, so will only discuss the case  $p = 2$ . A Sophie Germain factorization of  $2^k$  must be of type 1 or 2. For type 1 factorizations, one obtains solutions as above:  $k$  must admit a partition into distinct positive integers such that  $2 \cdot 2^{k_i} + 1 = 2^{k_i+1} + 1$  is prime. Such primes are called Fermat primes, and  $k_i + 1$  is forced to be a power of 2, so again each  $k_i$  must be odd. It is unknown whether there are infinitely many Fermat primes, therefore it is unknown whether there are infinitely many powers of 2 such that  $\text{rdu}(n) = 2^k$  admits a type 1 solution. A type 2 factorization must take the form  $2^s \cdot \theta$ , where  $\theta$  is

a type 1 Sophie Germain factorization of  $2^{k-s}$ . Since 3 is a Fermat prime, and the set of all solutions is obtained by multiplying the relevant  $\mathcal{N}$ -values by divisors of 3 or 24, we obtain that  $\text{rdu}(n)$  is a power of 2 if and only if  $n = 2^s p_1 \cdots p_t$ , where  $s \geq 0$  and  $p_1, \dots, p_t$  is a (possibly empty) list of distinct Fermat primes.

This provides an interesting connection between the ratio of diagonal units and a classical result of Gauss and Wantzel (see [Pollack 2009]): it is possible to construct a regular  $n$ -sided polygon in the plane with straightedge and compass if and only if  $n$  takes the form given at the end of the previous paragraph. Gauss proved that the condition on  $n$  is necessary, and Wantzel proved that it is sufficient. Gauss' decision to devote his life to mathematics was in part due to his discovery at age 18 of the constructibility of the regular 17-gon. We summarize our observation in the next theorem.

**Theorem 4.2.** *Let  $n$  be a positive integer. The following statements are equivalent.*

- (A) *The ratio of diagonal units in  $\mathbb{Z}_n$  is a power of 2.*
- (B) *The integer  $n$  has the form  $2^s p_1 \cdots p_t$ , where  $s \geq 0$  and  $p_1, \dots, p_t$  is a (possibly empty) list of distinct Fermat primes.*
- (C) *It is possible to construct a regular  $n$ -gon in the plane with straightedge and compass.*

The authors wish to thank Sunil Chebolu for noticing this connection to the Gauss–Wantzel theorem.

**4D. Pairs of distinct primes.** Call a positive integer  $n$  a *Sophie Germain number* if  $2n+1$  is prime. In all of the cases thus far considered, the integer  $\theta$  is a product of Sophie Germain numbers whenever  $\text{rdu}(n) = \theta$  has a solution. We include this section mainly to give a family of simple examples where this is not necessarily the case.

Let  $3 < p < q$  be distinct primes. The possible Sophie Germain factorizations of  $pq$  are  $p(2p+1)^0 q(2q+1)^0$  (if  $p$  and  $q$  are each Sophie Germain primes),  $(pq)(2pq+1)^0$  (if  $pq$  is a Sophie Germain number), and  $p \cdot q^0$  (if  $p$  is a Sophie Germain prime with safe prime  $q = 2p+1$ ). Each of these factorizations is type 1, so the solution sets (provided they exist) are  $(2p+1) \cdot \{\text{divisors of } 24\}$ ,  $(2p+1)(2q+1) \cdot \{\text{divisors of } 24\}$ , and  $(2p+1)^2 \cdot \{\text{divisors of } 24\}$ , respectively.

The integer  $1081 = 23 \cdot 47$  is not expressible as a product of Sophie Germain numbers since, though  $2 \cdot 23 + 1 = 47$  is prime, neither  $2 \cdot 47 + 1 = 95$  nor  $2 \cdot 23 \cdot 47 + 1 = 2163$  is prime. However,  $\text{rdu}(n) = 1081$  has solution set

$$47^2 \cdot \{\text{divisors of } 24\}.$$

**4E. Further questions.** We conclude this section with a few questions to ponder.

- The set of primes such that  $\text{rdu}(n) = p$  has a solution is precisely the set of Sophie Germain primes. From (B) in Section 4C we see that the set of primes

such that  $\text{rdu}(n) = p^2$  has a solution is the set  $\{2, 3\}$  (since  $\text{rdu}(2^5) = 2^2$  and  $\text{rdu}(3^3) = 3^2$ ). For  $k > 2$ , what can we say about the set of primes such that  $\text{rdu}(n) = p^k$  has a solution? Is it always nonempty? When is it finite?

- If  $p \equiv 2 \pmod{3}$ , must  $\text{rdu}(n) = p^k$  have a solution for some  $k$ ?
- The number of partitions of  $k$  into distinct odd parts is the same as  $s(k)$ , the number of self-conjugate partitions of  $k$ . The maximum number of solutions to  $\text{rdu}(n) = p^k$  (for  $p > 3$  prime) is  $8 \cdot s(k)$ . For each  $k$ , how many primes actually achieve this maximum value?
- Let  $k$  be a positive integer. Call a prime  $p$  a  $k$ -Sophie Germain prime ( $k$ -SGP) if  $k$  admits a partition into distinct odd parts and  $2p^{k_1} + 1, \dots, 2p^{k_r} + 1$  is a list of prime numbers for every partition  $k = k_1 + \dots + k_r$  of  $k$  into distinct odd parts. The value  $k = 1$  corresponds to an ordinary Sophie Germain prime, and there are no 2-SGPs. A prime  $p$  is a 3-SGP if and only if  $2p^3 + 1$  is prime; a prime  $p$  is an 8-SGP if and only if  $2p + 1, 2p^7 + 1, 2p^3 + 1$ , and  $2p^5 + 1$  are prime. Does a  $k$ -SGP exist for each  $k > 2$ ?

### 5. The multiplication cube for $\mathbb{Z}_n$

One could also analyze the multiplication cube for  $\mathbb{Z}_n$ . We know 1s lie exclusively on the diagonal if and only if  $n = 1$  or  $2$  since otherwise  $(-1) \cdot (-1) \cdot 1$  gives a 1 off the diagonal. Since this question seems uninteresting, we might require that every 1 in the multiplication table that is not in a coordinate plane (where one entry in the product is equal to 1) lies on the diagonal. The number of 1s appearing in the multiplication cube for  $\mathbb{Z}_n$  is  $\phi(n)^2$ . (The first and second coordinates may be completely arbitrary units, but then the third coordinate is determined.) The number of 1s off all coordinate planes is  $\phi(n)^2 - 3\phi(n) + 3 - 1$  (by the principle of inclusion/exclusion), and we wish to find values of  $n$  where this quantity is equal to the number of elements of multiplicative order precisely 3 (since the entry for  $1 \cdot 1 \cdot 1$  has been omitted). Put another way, we wish to find values of  $n$  such that  $\phi(n)^2 - 3\phi(n) + 3$  is equal to the number of elements of order dividing 3. In  $\mathbb{Z}_{p^k}^\times$  there is one element of order dividing 3 if  $p \equiv 2 \pmod{3}$ ; three such elements if  $p \equiv 1 \pmod{3}$ ; one such element if  $p = 3$  and  $k = 1$ ; and three such elements if  $p = 3$  and  $k \geq 2$ . Hence, the number of elements of  $\mathbb{Z}_n^\times$  whose order divides 3 is  $3^{r+\epsilon}$ , where  $r$  is the number of prime divisors congruent to 1 modulo 3 and  $\epsilon = 1$  if 9 divides  $n$  and  $\epsilon = 0$  otherwise. We must now consider the equation  $\phi(n)^2 - 3\phi(n) + 3 = 3^{r+\epsilon}$ . If 3 divides the right-hand side, then 3 divides  $\phi(n)^2$ , so in fact 9 divides  $\phi(n)^2 - 3\phi(n)$ . This means 9 cannot divide the right-hand side, so we need only consider  $\phi(n)^2 - 3\phi(n) + 3 = 1$  or  $3$ . This in turn forces  $\phi(n) = 1$  or  $2$  ( $\phi(n)$  cannot equal 3). The only values of  $n$  satisfying either of these equalities are  $n = 1, 2, 3, 4$ , and  $6$ . Conversely, it is easy to check that for  $n = 1, 2, 3, 4$  or  $6$ ,

all 1s in the multiplication cube lie on the diagonal or the coordinate planes. This proves the following theorem.

**Theorem 5.1.** *All 1s in the multiplication cube for  $\mathbb{Z}_n$  lie exclusively on the diagonal or the coordinate planes (where one of the three coordinates is 1) if and only if  $n$  is a divisor of 4 or 6.*

## References

- [Burton 1989] D. M. Burton, *Elementary number theory*, 2nd ed., W. C. Brown, Dubuque, IA, 1989. [MR 90e:11001](#) [Zbl 0696.10002](#)
- [Chebolu 2012] S. K. Chebolu, “What is special about the divisors of 24?”, *Math. Mag.* **85**:5 (2012), 366–372. [Zbl 1274.97016](#)
- [Cohen 2007] H. Cohen, *Number theory, I: Tools and Diophantine equations*, Graduate Texts in Mathematics **239**, Springer, New York, 2007. [MR 2008e:11001](#) [Zbl 1119.11001](#)
- [Laubenbacher and Pengelley 1999] R. Laubenbacher and D. Pengelley, *Mathematical expeditions: Chronicles by the explorers*, Springer, New York, 1999. [MR 99i:01005](#) [Zbl 0919.01001](#)
- [Pollack 2009] P. Pollack, *Not always buried deep: A second course in elementary number theory*, Amer. Math. Soc., Providence, RI, 2009. [MR 2010i:11003](#) [Zbl 1187.11001](#)

Received: 2014-06-09

Revised: 2014-06-09

Accepted: 2014-07-15

[genzka01@alumni.gettysburg.edu](mailto:genzka01@alumni.gettysburg.edu)

*Department of Mathematics, Gettysburg College,  
Gettysburg, PA 17325, United States*

[klockrid@gettysburg.edu](mailto:klockrid@gettysburg.edu)

*Department of Mathematics, Gettysburg College,  
Gettysburg, PA 17325, United States*



# On symplectic capacities of toric domains

Michael Landry, Matthew McMillan and Emmanuel Tsukerman

(Communicated by Michael Dorff)

A toric domain is a subset of  $(\mathbb{C}^n, \omega_{\text{std}})$  which is invariant under the standard rotation action of  $\mathbb{T}^n$  on  $\mathbb{C}^n$ . For a toric domain  $U$  from a certain large class for which this action is not free, we find a corresponding toric domain  $V$  where the standard action is free and for which  $c(U) = c(V)$  for any symplectic capacity  $c$ . Michael Hutchings gives a combinatorial formula for calculating his embedded contact homology symplectic capacities for certain toric four-manifolds on which the  $\mathbb{T}^2$ -action is free. Our theorem allows one to extend this formula to a class of toric domains where the action is not free. We apply our theorem to compute ECH capacities for certain intersections of ellipsoids and find that these capacities give sharp obstructions to symplectically embedding these ellipsoid intersections into balls.

## 1. Introduction

Symplectic capacities, introduced by Gromov and Hofer, are symplectic invariants that assign a nonnegative real number to a subset  $U \subset (\mathbb{C}^n, \omega_{\text{std}})$  and have the following properties:

- (C1) Monotonicity:  $c(U) \leq c(V)$  if  $U \hookrightarrow V$ .
- (C2) Conformality:  $c(\lambda U) = \lambda^2 c(U)$  for  $\lambda \in \mathbb{R}$ .
- (C3) Nontriviality:  $0 < c(B^{2n}(1)) < \infty$ .

Note that combining all three requires a finite capacity for any bounded  $U$ . Sometimes additional nontriviality and normalization axioms are also assumed, but we do not use them here. Many useful symplectic capacities have been defined; some are listed in [Cieliebak et al. 2007].

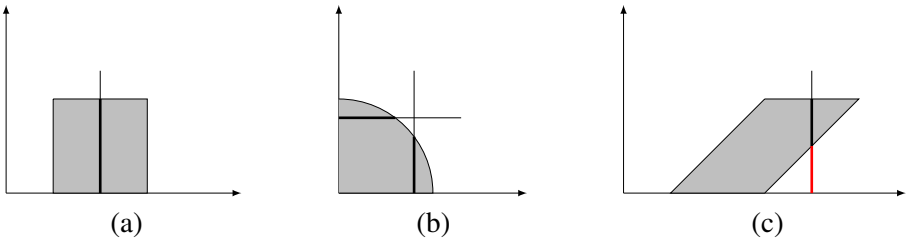
Define the *moment map*  $\mu : \mathbb{C}^n \rightarrow \mathbb{R}^n$  of the symplectic manifold  $(\mathbb{C}^n, \omega_{\text{std}})$  by

$$\mu(z_1, \dots, z_n) = (\pi|z_1|^2, \dots, \pi|z_n|^2),$$

where  $\omega_{\text{std}}$  is the standard symplectic form  $\omega_{\text{std}} = \sum_{i=1}^n dx_i \wedge dy_i$  on  $\mathbb{C}^n$ , and call  $\mu(\mathbb{C}^n)$  the moment space. We call  $U \subset (\mathbb{C}^n, \omega_{\text{std}})$  a *toric domain* when it can

MSC2010: 53D05, 53D20, 53D35.

Keywords: symplectic capacities, toric domain, moment space axes.



**Figure 1.** Appropriate moment regions; (a) and (b) satisfy the conditions of [Criterion 1.1](#), but (c) does not.

be written  $U = \mu^{-1}(A)$  for some *moment region*  $A \subset \mathbb{R}_{\geq 0}^n$  in the moment space, or equivalently when it is invariant under the rotation action of  $\mathbb{T}^n$  on  $\mathbb{C}^n$ . Note that this is a special case of the more general moment map associated with a Hamiltonian action of a Lie group.

Since these toric domains are uniquely represented by their moment regions, we will refer to a symplectic capacity  $c(A)$  of a moment region  $A$ , and by this mean  $c(\mu^{-1}(A))$ . A simple calculation shows that [\(C2\)](#) is equivalent to  $c(\lambda A) = \lambda c(A)$ .

Our main theorem is that for a duly qualified toric domain  $U$  whose moment region satisfies [Criterion 1.1](#) given below, any symplectic capacity of  $U$  is the same as the capacity of a toric domain with a free action, one whose moment region is  $\mu(U)$  translated off the coordinate planes in the moment space.

**Criterion 1.1.** Let  $A \subset \mathbb{R}_{\geq 0}^n$ . If  $A$  intersects a coordinate plane

$$P_i = \{(\rho_1, \dots, \rho_n) \in \mathbb{R}^n \mid \rho_i = 0\},$$

then any line normal to  $P_i$  has connected intersection with  $A \cup P_i$ .

The necessary further qualifications are given in the theorem statement below. [Figure 1](#) illustrates this condition for  $n = 2$ . In this case, [Criterion 1.1](#) ensures that the toric domain is a disk bundle over its projection to the first complex plane of  $\mathbb{C}^2$ ; more generally, for  $A$  satisfying the other conditions below, [Criterion 1.1](#) requires  $\mu^{-1}(A)$  to be a (generalized) disk bundle over its projection to any coordinate plane  $P_i$  which it touches. Disks in the fiber space degenerate to points where  $A$  touches a coordinate plane.

**Theorem 1.2.** Let  $A \subset \mathbb{R}_{\geq 0}^n$  be a moment region which is compact with star-shaped interior and whose boundary intersects transversely the rays from the star-center. Assume that  $A$  satisfies [Criterion 1.1](#). Then  $c(A) = c(A + (1, 1, \dots, 1))$  for any symplectic capacity  $c$ .

The theorem is proved by establishing equal lower and upper bounds on  $c(A)$  in terms of  $c(A + (1, 1, \dots, 1))$ . The lower bound follows readily from properties of toric domains and the axioms [\(C1\)–\(C3\)](#), but for the upper bound we must combine



the axioms with a nontrivial symplectic embedding. Since the proof assumes only the general axioms for capacities, this result holds for all symplectic capacities. Note that the action on a given toric domain  $U = \mu^{-1}(A)$  is free if and only if  $U$  does not intersect the origin in any  $\mathbb{C}$  factor; that is, its moment region does not touch any coordinate plane  $P_i = \{(\rho_1, \dots, \rho_n) \in \mathbb{R}^n \mid \rho_i = 0\}$  in the moment space.

The embedded contact homology (ECH) developed by Michael Hutchings provides a natural way to define certain symplectic capacities called ECH capacities. They are defined for any subset of a symplectic 4-manifold. Hutchings [2011] gives a combinatorial method to compute these capacities for toric domains over convex moment regions that do not touch the axes of the moment space  $\mathbb{R}_{\geq 0}^2$  (that is, the torus action is free). This method is presented in Section 3. In [Hutchings 2014, Remark 4.15] and [Choi et al. 2014, §1.2], it was conjectured that Hutchings’ formula should remain true in most, and probably all, cases where  $\mu(U)$  does touch one or both axes. Theorem 1.2 shows that this is true for the ECH capacities of a large class of toric domains by showing that it is true for all symplectic capacities.

Given  $a, b \in \mathbb{R}^+$ , define the ellipsoid

$$E(a, b) := \left\{ (z_1, z_2) \in \mathbb{C}^2 \mid \frac{\pi |z_1|^2}{a} + \frac{\pi |z_2|^2}{b} \leq 1 \right\}, \tag{1}$$

the ball

$$B(a) := E(a, a),$$

and the polydisk

$$P(a, b) := \{ (z_1, z_2) \in \mathbb{C}^2 \mid \pi |z_1|^2 \leq a, \pi |z_2|^2 \leq b \}, \tag{2}$$

where each inherits the standard symplectic form from  $\mathbb{C}^2$ .

In Section 3, we use Theorem 1.2 to compute ECH capacities of a class of intersections of ellipsoids. We also study symplectic embeddings of domains from this class, proving the following proposition:

**Proposition 1.3.** *Let  $a > b$  and  $c > d$ . Let  $R$  be the radius of the smallest ball containing  $E(a, b) \cap E(c, d)$ , and let  $\rho = \inf\{r \mid E(a, b) \cap E(c, d) \hookrightarrow B(r)\}$ . If  $2a, 2d \geq R$ , then  $\rho = R$ .*

It is known that ECH capacities provide sharp obstructions to symplectically embedding ellipsoids into ellipsoids (proved by McDuff [2011]) and ellipsoids into polydisks [Frenkel and Müller 2012]. Recall that by Gromov’s nonsqueezing theorem [1985], a ball symplectically embeds into a cylinder in  $\mathbb{R}^{2n}$  if and only if the radius of the cylinder exceeds that of the ball. This is an illustration of symplectic rigidity and is easily recovered from the ECH capacities on these domains. The computation of ECH capacities of the ellipsoid intersections above shows that they give sharp obstructions to symplectically embedding those ellipsoid intersections

into balls. Since the balls have much larger volume than the ellipsoid intersections, Proposition 1.3 is another example of symplectic rigidity.

In Proposition 1.3, the ECH capacities give a sharp obstruction. Recent work of Hind and Lisi [2014] shows that neither ECH capacities nor Ekeland–Hofer capacities give sharp obstructions to symplectic embeddings of arbitrary toric domains; in particular the ECH and Ekeland–Hofer obstructions to symplectically embedding a product of polydisks into a ball are not always sharp. The torus action on polydisks and balls is not free, so we might ask whether the situation is any different if we consider only toric domains for which the action is free. However, the case of free torus action is not different in this way, as the following corollary of Theorem 1.2 shows:

**Corollary 1.4.** *Let  $P^*(1, 2) = \mu^{-1}(\mu(P(1, 2)) + (1, 1))$  be a toric domain, let  $a < 3$  and let  $B^*(a) = \mu^{-1}(\mu(B^4(a)) + (1, 1))$ . There is no symplectic embedding  $P^*(1, 2) \hookrightarrow B^*(a)$ .*

This shows that neither ECH nor Ekeland–Hofer capacities are sharp even when we consider only toric domains with a free action because the obstruction given by both of these sequences of capacities is  $a \geq 2$  (see [Hind and Lisi 2014]). This corollary is proved in Section 3B.

## 2. Proof of main theorem

In this section, we prove Theorem 1.2 by constructing symplectomorphisms as the products of area preserving maps. It will be convenient to have the following standard lemma, which shows that translations in the moment space induce symplectomorphisms on toric domains whose moment regions do not touch any coordinate plane.

**Lemma 2.1.** *Suppose  $U \subset (\mathbb{R}^{2n}, \omega_{\text{std}})$  is a toric domain with free torus action such that  $\mu(U) = A$ , and  $B$  is any translate of  $A$  such that the torus action on  $\mu^{-1}$  is also free. Then  $U$  and  $V = \mu^{-1}(B)$  are symplectomorphic. In particular, they have the same symplectic capacity for any capacity.*

*Proof.* We can parametrize  $U$  by  $g : A \times \mathbb{T}^n \rightarrow U$  defined by

$$g(\rho_1, \dots, \rho_n, e^{i\theta_1}, \dots, e^{i\theta_n}) = \left( \sqrt{\frac{\rho_1}{\pi}} e^{i\theta_1}, \dots, \sqrt{\frac{\rho_n}{\pi}} e^{i\theta_n} \right).$$

Then we can pull back the standard symplectic form to  $A \times \mathbb{T}^n$ . A simple calculation shows that for the first term,

$$g^*(dx_1 \wedge dy_1) = \frac{1}{2\pi} d\rho_1 \wedge d\theta_1,$$

and thus

$$g^* \omega_{std} = \frac{1}{2\pi} \sum_{i=1}^n d\rho_i \wedge d\theta_i.$$

Clearly translation in moment space does not affect this last form, so conjugating a translation by this parametrization yields the desired symplectomorphism.  $\square$

Another important fact that can be seen from the proof of [Lemma 2.1](#) is that for a toric domain  $U$  with free torus action and moment region  $A$ , the symplectic volume of  $U$  is equal to the volume of  $A$ :

$$\begin{aligned} \text{vol}(U, \omega_{std}) &= \frac{1}{n!} \int_U \omega_{std}^n = \frac{1}{n!} \int_{A \times \mathbb{T}^n} (g^* \omega_{std})^n \\ &= \frac{1}{(2\pi)^n} \int_{A \times \mathbb{T}^n} d\rho_1 \wedge \cdots \wedge d\rho_n \wedge d\theta_1 \wedge \cdots \wedge d\theta_n \\ &= \int_A d\rho_1 \wedge \cdots \wedge d\rho_n = \text{vol}(A). \end{aligned}$$

So a symplectic embedding of toric domains  $U \hookrightarrow V$  may be possible only if  $\text{vol}(\mu(U)) \leq \text{vol}(\mu(V))$ .

We will also use the following version of the ‘‘Traynor trick’’ (cf. Proposition 5.2 of [[Traynor 1995](#)]):

**Lemma 2.2.** *Given  $\varepsilon > 0$ , there exists an area preserving diffeomorphism*

$$\Psi : B^2(1) \rightarrow SD^2(1 + \varepsilon) = B^2(1 + \varepsilon) - \{x + iy \mid y = 0, x \geq 0\}$$

*from the disk to the slit-disk such that*

$$\delta < |\Psi(z)|^2 < |z|^2 + \varepsilon$$

*for some  $\delta > 0$ .*

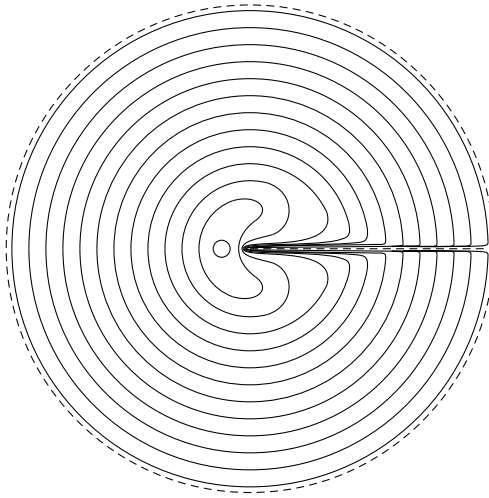
*Proof.* The left inequality follows from continuity (given such a map). For existence and the right inequality, define a family of loops which avoid the slit as in [Figure 2](#), and apply Schlenk [[2005](#), Lemma 3.1].  $\square$

With these tools we can prove [Theorem 1.2](#).

*Proof of Theorem 1.2.*

Our technique is to find upper and lower bounds on  $c(A)$  by producing symplectic embeddings and applying [\(C1\)](#) and [\(C2\)](#). We show that these bounds agree with each other and with  $c(A + (1, 1, \dots, 1))$ .

For what follows, we define the scaling of  $\mathbb{R}^n$  by  $\lambda > 0$  from  $p \in \mathbb{R}^n$  to be the map  $q \mapsto \lambda(q - p) + p$ . Since  $\lambda(q - p) + p = \lambda q + (1 - \lambda)p$ , any scaling by  $\lambda$  from  $p$  is equivalent to a scaling from the origin by  $\lambda$  followed by translation by  $(1 - \lambda)p$ . So



**Figure 2.** A family of loops defining a symplectomorphism  $B^2(1) \rightarrow SD(1 + \varepsilon)$ .

with Lemma 2.1 we may apply conformality of capacities, axiom (C2), on moment regions scaled from points other than the origin. The reason for the requirement that rays from the star-center be transverse to the boundary will become clear in Step 2 with the scaling argument.

*Step 1.* The lower bound may be computed as follows. Let  $p$  be a star-center of  $\text{int } A$ , which means that any other point in  $\text{int } A$  may be connected to  $p$  by a line contained in  $\text{int } A$ . Given any  $\lambda < 1$ , let  $A_\lambda$  be the image of  $A$  under the scaling of the moment space towards  $p$  by  $\lambda$ . Since  $p$  is away from the coordinate planes,  $A_\lambda$  is bounded away from the coordinate planes and contained in  $A$ . By Lemma 2.1 and conformality,  $c(A_\lambda) = \lambda c(A + (1, 1, \dots, 1))$ . Then by monotonicity,  $\lambda c(A + (1, 1, \dots, 1)) \leq c(A)$ , and since  $\lambda < 1$  was arbitrary,

$$c(A + (1, 1, \dots, 1)) \leq c(A).$$

*Step 2.* For the upper bound, we embed  $A$  into an expanded version of  $A$ , and apply an area-preserving map in each dimension in which  $A$  touches a coordinate plane  $P_i$ . We will assume that  $A$  is compact, star-shaped, and that the rays from a star-center  $p$  intersect each  $\partial A_j$  transversely.

Assume without loss of generality that  $A$  touches the first  $k$  coordinate planes and does not touch the others. Let  $p = (\rho_1, \dots, \rho_n)$  be the star-center in  $A$  noted above. The projection  $\tilde{p}_1 = (0, \rho_2, \dots, \rho_n)$  is also a star-center: Choose any other point  $q = (x_1, \dots, x_n) \in A$ . The line from  $\tilde{p}_1$  to  $q$  is entirely below that from  $p$  to  $q$  in the  $\rho_1$  coordinate. By Criterion 1.1, any perpendicular dropped from a point

in  $A$  to  $P_1$  remains in  $A$ . Hence the line from  $\tilde{p}_1$  to  $q$  is also in  $A$ , so  $\tilde{p}_1$  is a star-center. Repeating in the first  $k$  coordinates, we find that  $\tilde{p}_k = (0, \dots, \rho_{k+1}, \dots, \rho_n)$  is a star-center; call this point  $\tilde{p}$ . A simple geometric argument making use of [Criterion 1.1](#) shows that the rays from  $\tilde{p}$  must also be transverse to each  $\partial A_j$ ; we omit that here.

The next step will be to expand  $A$  to  $A_\lambda$  by a finite factor of  $\lambda$ . In order to prevent  $A_\lambda$  from colliding with coordinate planes, first translate  $A$  away from the coordinate planes  $P_{k+1}$  through  $P_n$  by some large amount. Note that this is possible because by assumption  $p_i > 0$  for  $i > k$ , and furthermore translation in the moment spaces induces a symplectomorphism. So we shall instead compute the capacity of this translate, and relabel it  $A$ . Now let  $A_\lambda$  be the scaling of  $A$  from  $\tilde{p}$  by a small  $\lambda > 1$ .

We show that  $A \subset \text{int } A_\lambda$ . Consider any point  $q = (x_1, \dots, x_n) \in A$ . If  $q \in \text{int } A$  then  $q \in \text{int } A_\lambda$ , so suppose  $q \in \partial A$ . Write  $q_{1/\lambda}$  for the point mapped to  $q$  under the scaling;  $q_{1/\lambda}$  will be between  $\tilde{p}$  and  $q$ . Now since the ray from  $\tilde{p}$  to  $q$  is transverse to  $\partial A$ , it follows that  $q_{1/\lambda}$  must be in  $\text{int } A$ , so we can find an open ball  $U$  around  $q_{1/\lambda}$ . That ball maps under the scaling to  $U_\lambda$ , which is an open ball around  $q$  in  $A_\lambda$ . Thus  $q \in \text{int } A_\lambda$ , and  $A \subset \text{int } A_\lambda$ .

Let  $\text{ext } A_\lambda$  denote the exterior of  $A_\lambda$  in  $\mathbb{R}_{\geq 0}^n$ . Both  $A$  and  $A_\lambda$  are compact, so there is some  $d$  so that  $0 < d < d_\lambda = \frac{1}{2} \text{dist}(A, \text{ext } A_\lambda)$ . Now  $A$  is bounded, so let  $a$  be the maximum of the  $\rho_1$  coordinate of  $A$ , and choose  $\varepsilon > 0$  so that  $\varepsilon < d$ . Then by [Lemma 2.2](#), there exists  $\Psi_a : B^2(a) \rightarrow SD^2(a + \varepsilon)$  such that

$$\delta < |\Psi_a(z)|^2 < |z|^2 + \varepsilon \tag{3}$$

for  $\delta > 0$ . Let  $F_\varepsilon = \Psi_a \times \text{id} \times \dots \times \text{id}$ .

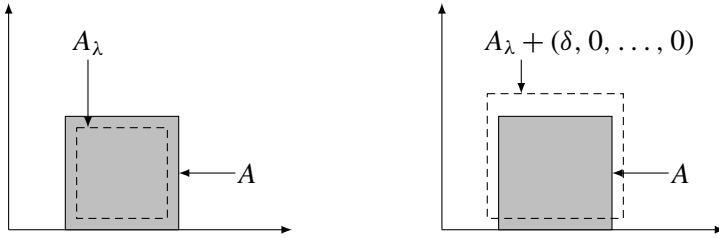
Set  $B = \mu \circ F_\varepsilon(\mu^{-1}(A))$ . Then we claim  $B \subset \text{int } A_\lambda$ . Consider a point  $(z_1, \dots, z_n) \in \mu^{-1}(A)$ , and let

$$(\rho_1, \dots, \rho_n) \equiv \mu(z_1, \dots, z_n) \in A.$$

By the inequality above,  $\mu \circ F_\varepsilon((z_1, \dots, z_n)) = (\tilde{\rho}_1, \dots, \tilde{\rho}_n)$ , where  $\tilde{\rho}_1 < \rho_1 + \varepsilon$  and  $\tilde{\rho}_i = \rho_i$  for  $i > 1$ . Thus every point in  $\mu^{-1}(A)$  is carried by  $F_\varepsilon$  to a point less than  $d$  away from  $A$ , so  $B \subset \text{int } A_\lambda$ ; moreover,  $\text{dist}(B, \text{ext } A_\lambda) > d_\lambda$ . Then let  $\delta = \frac{1}{2} \min\{\delta, d_\lambda\}$  and  $\gamma = \lambda\delta$  (using  $\lambda < 2$ ). Set  $A'_\lambda = A_\lambda + (\gamma, 0, \dots, 0)$ . The lower bound on the left of equation (3), together with the distance from  $B$  to outside  $A_\lambda$ , show that in fact  $B \subset A'_\lambda$ . So by [Lemma 2.1](#),  $c(B) \leq c(A'_\lambda) = \lambda c(A + (\delta, 0, \dots, 0))$ . Now  $\lambda > 1$  was arbitrary, so  $c(B) \leq c(A + (\delta, 0, \dots, 0))$ . Since  $A$  and  $B$  are symplectomorphic,

$$c(A) \leq c(A + (\delta, 0, \dots, 0)).$$

Repeating the same process in the dimensions up to  $k$  and translating up by  $\delta$  in the other coordinates shows that for some  $\delta > 0$ ,  $c(A) \leq c(A + (\delta, \delta, \dots, \delta))$ .



**Figure 3.** Illustration of the conformality argument for the lower bound (left) and the upper bound (right).

Combining with the lower bound, and using [Lemma 2.1](#),

$$c(A) = c(A + (1, 1, \dots, 1)). \quad \square$$

**Remark 2.3.** It is worth noting that we may like to consider regions  $A$  for which  $\partial A$  is not completely smooth. The ellipsoid intersections below are one example. The notion of transversality must then be generalized slightly with the goal that  $A \subset \text{int } A_\lambda$ . If  $\partial A$  is the gluing of multiple hypersurfaces, it is sufficient that the rays from the star-center be transverse to each of the hypersurfaces at the points where they are glued together.

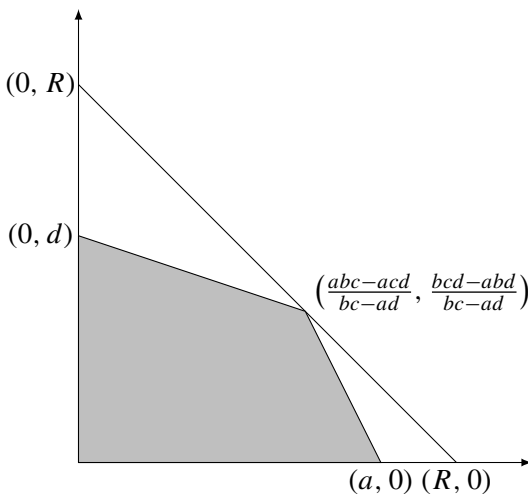
### 3. Applications

**3A. ECH capacities.** The remainder of this paper focuses on 4-dimensional toric domains, with accompanying planar moment regions. Using Michael Hutchings’ theory of embedded contact homology (ECH), one can associate real numbers

$$0 = c_0(M) \leq c_1(M) \leq c_2(M) \leq \dots$$

called *ECH capacities* to any 4-dimensional Liouville domain  $M$ , such that each  $c_i$  is a symplectic capacity for 4-manifolds. For precise definitions of ECH capacities and Liouville domains, see [\[Hutchings 2011\]](#).

We briefly describe the computation of ECH capacities, as given by Theorem 4.14 of [\[Hutchings 2014\]](#). Given a convex body  $A$  in the moment space which does not touch any coordinate plane, we can define a norm  $\ell_A$ , not necessarily symmetric, as follows. Choose an origin in  $A$  from which to draw position vectors to  $\partial A$ . Let  $v_i$  be some vector, and  $q_i$  one of the position vectors on  $\partial A$  such that the outward normal to  $\partial A$  at  $q_i$  is parallel to  $v_i$ . If  $v_i$  has angle between the normals to  $\partial A$  at two incident edges of  $\partial A$ , let  $q_i$  be the corner where the edges meet. Then set  $\ell_A(v_i) = v_i \cdot q_i$ . It is not hard to check that this yields a well-defined norm; see [\[Hutchings 2014\]](#) for details.



**Figure 4.** The image of  $E(a, b) \cap E(c, d)$  under  $\mu$  with suitable  $a, b, c, d$ , and the smallest ball into which it symplectically embeds.

We compute the ECH capacities according to [Hutchings 2011] as follows: for each  $k$ ,  $c_k(A)$  is the shortest perimeter length of an oriented lattice-polygon enclosing  $k + 1$  lattice points, where perimeter length is measured in the norm  $\ell_A$  on the edge vectors of the oriented polygon.

**3A1. Embedding ellipsoid intersections into balls.** We now use Theorem 1.2 to compute the second ECH capacity of a family of ellipsoid intersections. This capacity is in turn used to prove Proposition 1.3. Throughout this section, let  $a, b, c, d > 0$ ,  $a < b$ ,  $c > d$ , and put

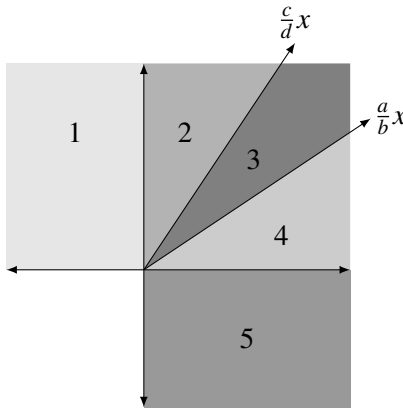
$$R = \frac{abc + bcd - acd - abd}{bc - ad}$$

(see Figure 4). We show that for  $2a, 2d \geq R$ , we have  $c_2(E(a, b) \cap E(c, d)) = R$ . A simple consequence is that  $E(a, b) \cap E(c, d)$  symplectically embeds into a ball if and only if it embeds by inclusion (that is, Proposition 1.3). While in principle that result only requires the easier lower bound of Theorem 1.2, we illustrate the use of Theorem 1.2 to produce the actual ECH capacity, which is sufficient to prove the proposition.

A short computation, or consideration of Figure 4, shows that  $B(R)$  is indeed the smallest ball into which  $E(a, b) \cap E(c, d)$  embeds by inclusion. We first prove the following lemma:

**Lemma 3.1.** *If  $2a, 2d \geq R$ , then  $c_2(E(a, b) \cap E(c, d)) = R$ .*

Assuming Lemma 3.1, observe that Proposition 1.3 is immediate:



**Figure 5.** Calculation of  $\ell_{A'}$ -length by region.

*Proof of Proposition 1.3.* By [Hutchings 2011, Corollary 1.3],  $c_2(B(r)) = r$ , so we have  $\rho \geq R$  by Lemma 3.1. Since  $E(a, b) \cap E(c, d) \subset B(R)$ ,  $\rho \leq R$  and the result follows.  $\square$

*Proof of Lemma 3.1.* Let  $A$  be the moment region of  $E(a, b) \cap E(c, d)$ . Since  $A$  satisfies Criterion 1.1, we know that  $c_2(A) = c_2(A')$  for  $A' = A + (1, 1)$ .

First, we observe that the oriented lattice-polygonal path shown in Figure 6 has  $\ell_{A'}$ -length  $R$  when oriented clockwise, so  $c_2(A) \leq R$ .

Let  $\Gamma$  be an oriented lattice path containing three lattice points with edge vectors  $(\alpha, \beta), (\gamma, \delta), (\epsilon, \zeta)$  (if  $\Gamma$  has only two edge vectors, i.e., is just a line segment, the forthcoming argument applies *mutatis mutandis*). Suppose for a contradiction that  $\ell_{A'}(\Gamma) < R$ .

We first claim that  $\beta, \delta, \zeta \leq 1$  and that at most one is positive. Suppose without loss of generality that  $\beta \geq 2$ . Depending on the region in which  $(\alpha, \beta)$  lies (or its slope  $\beta/\alpha$ , Figure 5), the  $\ell_{A'}$ -length is determined by cases:

$$\ell_{A'}((\alpha, \beta)) = \begin{cases} (\alpha, \beta) \cdot (0, d) & \text{if } \alpha \leq 0 \text{ or } \frac{\beta}{\alpha} \geq \frac{c}{d} \text{ (regions 1, 2),} \\ (\alpha, \beta) \cdot \left( \frac{abc-acd}{bc-ad}, \frac{bcd-abd}{bc-ad} \right) & \text{if } \frac{c}{d} \leq \frac{\beta}{\alpha} \leq \frac{a}{b} \text{ (region 3),} \\ (\alpha, \beta) \cdot (a, 0) & \text{if } 0 < \frac{\beta}{\alpha} \leq \frac{a}{b} \text{ (region 4).} \end{cases}$$

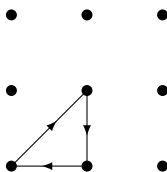
We treat each case separately. In region 1, we have  $(\alpha, \beta) \cdot (0, d) = \beta d \geq 2d \geq R$ , a contradiction. In region 2,

$$\ell_{A'}((\alpha, \beta)) = (\alpha, \beta) \cdot \left( \frac{abc-acd}{bc-ad}, \frac{bcd-abc}{bc-ad} \right)$$

and  $\alpha \geq 1$ . Hence,

$$(\alpha, \beta) \cdot \left( \frac{abc-acd}{bc-ad}, \frac{bcd-abc}{bc-ad} \right) > (1, 1) \cdot \left( \frac{abc-acd}{bc-ad}, \frac{bcd-abc}{bc-ad} \right) = R.$$





**Figure 6.** The minimal path for  $c_2(A)$  in Lemma 3.1.

Lastly, in region 3,  $\ell_{A'}((\alpha, \beta)) = (\alpha, \beta) \cdot (a, 0)$  and  $\alpha > \beta$ , so  $\ell_{A'}((\alpha, \beta)) = \alpha a > 2a \geq R$ . Thus  $\beta, \delta, \zeta \leq 1$ .

To show that at most one of  $\beta, \delta, \gamma$  is positive, assume without loss of generality that  $\beta, \delta \geq 1$ . Another calculation as above shows that both  $\ell_{A'}((\alpha, \beta))$  and  $\ell_{A'}((\gamma, \delta))$  are greater than or equal to  $\min\{a, d\}$ , so  $\ell_{A'}(\Gamma) \geq 2 \min\{a, d\} \geq R$ , a contradiction.

A symmetric argument but with regions 2, 3, 4, and 5 shows that  $\alpha, \gamma, \epsilon \leq 1$  and that at most one is positive. These facts imply that the maximum displacement in either coordinate is 1; that is,  $\Gamma$  lies in  $[0, 1]^2$  up to translation. We check that the shortest lattice path containing three lattice points in  $[0, 1]^2$  has  $\ell_{A'}$ -length  $R$ , so  $\Gamma$  cannot exist. □

**3B. Toric domains with free action.** The proof of Corollary 1.4 simply combines the embeddings involved in the proof of Theorem 1.2 with the result that a symplectic embedding  $P(1, 2) \hookrightarrow B^4(a)$  is possible if and only if  $a \geq 3$  [Hind and Lisi 2014, Theorem 1.1].

*Proof of Corollary 1.4.* Suppose to the contrary that  $a < 3$  is given for which we can find an embedding  $f : P^*(1, 2) \hookrightarrow B^*(a)$ . Let  $\lambda > 1$  be close to 1 such that  $\lambda^2 a < 3$ . Let  $P_\lambda^*(1, 2) = \mu^{-1}(\mu(P(\lambda, 2\lambda)) + (1, 1))$  and  $B_\lambda^*(a) = \mu^{-1}(\mu(B^4(\lambda a)) + (1, 1))$ . After scaling by  $\lambda$ , we can find an embedding  $f_\lambda : P_\lambda^*(1, 2) \hookrightarrow B_\lambda^*(a)$ . This is combined with the embeddings from the proof of Theorem 1.2 as follows:

First, we can find a symplectic embedding  $F : P(1, 2) \hookrightarrow P_\lambda^*(1, 2)$  by the same technique illustrated in that theorem since  $P_\lambda^*(1, 2)$  is just the translated expansion of  $P(1, 2)$ . We also have the inclusion embedding  $\iota : B_\lambda^*(a) \hookrightarrow B(\lambda^2 a)$  because of the translation law (Lemma 2.1) above. Combining these we get

$$\iota \circ f_\lambda \circ F : P(1, 2) \hookrightarrow B(\lambda^2 a).$$

Since  $\lambda^2 a < 3$ , this violates [Hind and Lisi 2014, Theorem 1.1]. Thus no such embedding  $f : P^*(1, 2) \hookrightarrow B^*(a)$  exists. □

By Theorem 1.2, the ECH and Ekeland–Hofer capacities of  $P^*(1, 2)$  and  $B^*(a)$  are the same as those of  $P(1, 2)$  and  $B(a)$ , so neither of these capacities give sharp obstructions to embedding  $P^*(1, 2)$  into  $B^*(a)$ .

## Acknowledgments

We thank our advisor Daniel Cristofaro-Gardiner and the UC Berkeley Geometry, Topology and Operator Algebras RTG Summer Research Program for Undergraduates 2013, supported by NSF grant DMS-0838703. We also thank Michael Hutchings for helpful advice and direction for this work, and the anonymous referee for valuable feedback.

## References

- [Choi et al. 2014] K. Choi, D. Cristofaro-Gardiner, D. Frenkel, M. Hutchings, and V. Ramos, “Symplectic embeddings into four-dimensional concave toric domains”, *J. Topology* (online publication May 2014).
- [Cieliebak et al. 2007] K. Cieliebak, H. Hofer, J. Latschev, and F. Schlenk, “Quantitative symplectic geometry”, pp. 1–44 in *Dynamics, ergodic theory, and geometry*, edited by B. Hasselblatt, Math. Sci. Res. Inst. Publ. **54**, Cambridge Univ. Press, 2007. [MR 2009d:53126](#) [Zbl 1143.53341](#)
- [Frenkel and Müller 2012] D. Frenkel and D. Müller, “Symplectic embeddings of 4-dimensional ellipsoids into cubes”, preprint, 2012. [arXiv 1210.2266](#)
- [Gromov 1985] M. Gromov, “Pseudoholomorphic curves in symplectic manifolds”, *Invent. Math.* **82**:2 (1985), 307–347. [MR 87j:53053](#) [Zbl 0592.53025](#)
- [Hind and Lisi 2014] R. Hind and S. Lisi, “Symplectic embeddings of polydisks”, *Selecta Math.* (online publication January 2014).
- [Hutchings 2011] M. Hutchings, “Quantitative embedded contact homology”, *J. Differential Geom.* **88**:2 (2011), 231–266. [MR 2838266](#) [Zbl 1238.53061](#)
- [Hutchings 2014] M. Hutchings, “Lecture notes on embedded contact homology”, pp. 389–484 in *Contact and symplectic topology*, edited by F. Bourgeois et al., Bolyai Society Mathematica Studies **26**, Springer, New York, 2014.
- [McDuff 2011] D. McDuff, “The Hofer conjecture on embedding symplectic ellipsoids”, *J. Differential Geom.* **88**:3 (2011), 519–532. [MR 2012j:53113](#) [Zbl 1239.53109](#)
- [Schlenk 2005] F. Schlenk, *Embedding problems in symplectic geometry*, de Gruyter Expositions in Mathematics **40**, Walter de Gruyter, Berlin, 2005. [MR 2007c:53125](#) [Zbl 1073.53117](#)
- [Traynor 1995] L. Traynor, “Symplectic packing constructions”, *J. Differential Geom.* **42**:2 (1995), 411–429. [MR 96k:53046](#) [Zbl 0861.52008](#)

Received: 2014-06-20

Revised: 2014-07-30

Accepted: 2014-08-02

[michael.landry@yale.edu](mailto:michael.landry@yale.edu)

Mathematics Department, Yale University,  
10 Hillhouse Avenue, New Haven, CT 06511, United States

[mm2041@cam.ac.uk](mailto:mm2041@cam.ac.uk)

Wheaton College, 501 College Avenue, Wheaton, IL 60187,  
United States

[e.tsukerman@math.berkeley.edu](mailto:e.tsukerman@math.berkeley.edu)

Department of Mathematics, University of California, Berkeley,  
970 Evans Hall, Berkeley, CA 94720, United States

# When the catenary degree agrees with the tame degree in numerical semigroups of embedding dimension three

Pedro A. García-Sánchez and Caterina Viola

(Communicated by Scott T. Chapman)

We characterize numerical semigroups of embedding dimension three having the same catenary and tame degrees.

## 1. Introduction

Let  $S$  be a numerical semigroup minimally generated by  $\{n_1, \dots, n_p\}$ . A factorization of  $s \in S$  is an element  $x = (x_1, \dots, x_p) \in \mathbb{N}^p$  such that  $x_1 n_1 + \dots + x_p n_p = s$  ( $\mathbb{N}$  denotes the set of nonnegative integers). The length of  $x$  is given by  $|x| = x_1 + \dots + x_p$ . Given another factorization  $y = (y_1, \dots, y_p)$ , the distance between  $x$  and  $y$  is  $d(x, y) = \max\{|x - \gcd(x, y)|, |y - \gcd(x, y)|\}$ , where  $\gcd(x, y) = (\min\{x_1, y_1\}, \dots, \min\{x_p, y_p\})$ .

The catenary degree of  $S$  is the minimum nonnegative integer  $N$  such that for every  $s \in S$  and any two factorizations  $x$  and  $y$  of  $s$ , there exists a sequence of factorizations  $x_1, \dots, x_t$  of  $s$  such that

- (1)  $x_1 = x, x_t = y$ ,
- (2) for all  $i \in \{1, \dots, t-1\}$ ,  $d(x_i, x_{i+1}) \leq N$ .

The tame degree of  $S$  is defined also in terms of distances, and it is the minimum  $N$  such that for any  $s \in S$  and any factorization  $x$  of  $s$ , if  $n - n_i \in S$  for some  $i \in \{1, \dots, p\}$ , then there exists another factorization  $x'$  of  $s$  such that  $d(x, x') \leq N$  and the  $i$ -th coordinate of  $x'$  is nonzero ( $n_i$  “occurs” in this factorization).

It is well known that the catenary degree of  $S$  is less than or equal to the tame degree of  $S$  (in greater generality; see [Geroldinger and Halter-Koch 2006]). It

---

*MSC2010:* primary 20M13; secondary 20M14, 13A05.

*Keywords:* numerical semigroup, catenary degree, tame degree.

García-Sánchez is supported by the projects MTM2010-15595, FQM-343, FQM-5849, and FEDER funds. The contents of this article are part of Viola’s master’s thesis. Part of this work was done while she visited the Universidad de Granada under the European Erasmus mobility program. Both authors thank the referee for comments and suggestions.

is also known that in some cases both coincide (for instance for monoids with a generic presentation [Blanco et al. 2011]). In this paper, we want to characterize when this is the case if  $p$  (the embedding dimension of  $S$ ) is three. This description is given in terms of the connectedness of some graphs associated to the elements of  $S$ .

Given  $s \in S$ , we define the graph  $\nabla_s$  as the graph with vertices given by the factorizations of  $s$ , and edges given by the pairs of factorizations  $x$  and  $y$  with  $x \cdot y \neq 0$  (here  $\cdot$  is the dot product; that is,  $x$  and  $y$  have common support). We say that  $s$  is a Betti element of  $S$  if  $\nabla_s$  is not connected. It is well known (see for instance [Rosales and García-Sánchez 2009], where the connected components of  $\nabla_s$  are called  $\mathcal{R}$ -classes of  $s$ ) that the number of Betti elements of  $S = \langle n_1, n_2, n_3 \rangle$  is at most three. We characterize when  $t(S) = c(S)$  in terms of the Betti elements of  $S$ ; this is done in Theorem 25.

## 2. Preliminaries

A numerical semigroup is a submonoid of  $(\mathbb{N}, +)$  with finite complement in  $\mathbb{N}$ . Every submonoid  $M$  of  $(\mathbb{N}, +)$  is isomorphic to the numerical semigroup  $M/\text{gcd}(M)$ . The least positive integer in a numerical semigroup  $S$  is known as its *multiplicity*,  $m(S)$ . Every numerical semigroup  $S$  is minimally generated by  $S^* \setminus (S^* + S^*)$ , and as every two minimal generators are incongruent modulo the multiplicity, this set has finitely many elements. Its cardinality is known as the *embedding dimension* of  $S$ , denoted by  $e(S)$ . Thus, every numerical semigroup admits a unique (and finite) minimal generating system. Its elements are known as *minimal generators* of the semigroup. The largest integer not belonging to  $S$  is the *Frobenius number* of  $S$ ,  $F(S)$ .

For a given nonempty subset  $A$  of  $\mathbb{N}$ , set

$$\langle A \rangle = \{ \lambda_1 a_1 + \dots + \lambda_n a_n \mid n \in \mathbb{N}, a_1, \dots, a_n \in A \},$$

which is the submonoid of  $(\mathbb{N}, +)$  generated by  $A$ .

**2.1. Catenary and tame degrees.** Let  $S$  be minimally generated by  $\{n_1, \dots, n_p\}$ . We recall some key notions from the theory of nonunique factorizations. Consider the monoid epimorphism

$$\varphi : \mathbb{N}^p \rightarrow S, \quad \varphi(a_1, \dots, a_p) = a_1 n_1 + \dots + a_p n_p,$$

known as the *factorization morphism* of  $S$ . The monoid  $S$  is isomorphic to  $\mathbb{N}^p / \sigma$ , where  $\sigma = \{(a, b) \in \mathbb{N}^p \mid \varphi(a) = \varphi(b)\}$  is the kernel congruence of  $\varphi$ . As usual, we write  $a\sigma b$  if  $(a, b) \in \sigma$ . The *set of factorizations* of an element  $n \in S$  is

$$Z(n) = \varphi^{-1}(n) = \{(a_1, \dots, a_p) \in \mathbb{N}^p \mid a_1 n_1 + \dots + a_p n_p = n\}.$$

Let  $a = (a_1, \dots, a_p) \in Z(n)$ . The *length* of the factorization  $a$  is  $|a| = a_1 + \dots + a_p$ .

For  $z = (z_1, \dots, z_p), z' = (z'_1, \dots, z'_p) \in \mathbb{N}^p$ , write

$$\gcd(z, z') = (\min\{z_1, z'_1\}, \dots, \min\{z_p, z'_p\}).$$

Set  $d(z, z') = \max\{|z - \gcd(z, z')|, |z' - \gcd(z, z')|\}$  to be the *distance* between  $z$  and  $z'$ . Given  $x \in \mathbb{N}^p$  and  $Y \subset \mathbb{N}^p$ , we define  $d(x, Y) = \min\{d(x, y) \mid y \in Y\}$  (which exists by Dickson's lemma). The *support* of  $z \in \mathbb{N}^p$  is defined, as usual, by  $\text{Supp}(z) = \{i \in \{1, \dots, p\} \mid z_i \neq 0\}$ . Let  $n \in S$  be such that  $n - n_i \in S$ . Then the set  $Z^i(n) = \{z \in Z(n) \mid i \in \text{Supp}(z)\}$  is not empty.

Given  $n \in S$  and  $z, z' \in Z(n)$ , an  $N$ -chain of factorizations from  $z$  to  $z'$  is a sequence  $z_0, \dots, z_k \in Z(n)$  such that  $z_0 = z, z_k = z'$  and  $d(z_i, z_{i+1}) \leq N$  for all  $i$ . The *catenary degree* of  $n$ ,  $c(n)$ , is the minimal  $N \in \mathbb{N} \cup \{\infty\}$  such that for any two factorizations  $z, z' \in Z(n)$ , there is an  $N$ -chain from  $z$  to  $z'$ . The catenary degree of  $S$ ,  $c(S)$ , is defined by

$$c(S) = \sup\{c(n) \mid n \in S\}.$$

The *tame degree*  $t_S(S', X)$  of  $S' \subseteq S$  and  $X \subset \mathbb{N}^p$  is the minimum of all  $N \in \mathbb{N} \cup \{\infty\}$  such that for all  $s \in S', z \in Z(s)$  and  $x \in X$  with  $s - \varphi(x) \in S$ , there exists  $z' \in Z(s)$  satisfying  $x \leq z'$  and  $d(z, z') \leq N$ . We simply write  $t(S', X)$  when  $S$  is understood. We also simply write  $t(s)$  for  $t(\{s\}, \{n_1, \dots, n_p\})$ , and  $t(S) = t(S, \{n_1, \dots, n_p\})$ , which equals  $\max\{t(s) \mid s \in S\}$ .

A *presentation* for  $S$  is a subset  $\tau$  of  $\sigma$  such that  $\sigma$  is the least congruence (with respect to set inclusion) containing  $\tau$ , or in other words, a system of generators of  $\sigma$ . A *minimal presentation* is a presentation that is minimal with respect to set inclusion (and it can be shown that in this setting it is also minimal with respect to cardinality, see [Rosales and García-Sánchez 2009, Chapter 7]; in monoids these two concepts do not have to be equivalent). We say that  $S$  is *uniquely presented* if for every two minimal presentations  $\tau$  and  $\tau'$  of  $S$  and every  $(a, b) \in \tau$ , either  $(a, b) \in \tau'$  or  $(b, a) \in \tau'$  (see [García-Sánchez and Ojeda 2010]).

Two elements  $z$  and  $z'$  of  $\mathbb{N}^p$  are  $\mathcal{R}$ -related if there exists a chain  $z = z_1, z_2, \dots, z_k = z'$  such that  $\text{Supp}(z_i) \cap \text{Supp}(z_{i+1})$  is not empty for all  $i \in \{1, \dots, k-1\}$ . The number of factorizations of an element in a numerical semigroup is finite, and so is the number of  $\mathcal{R}$ -classes in this set. These classes are crucial, since from them a minimal presentation of  $S$  can be constructed. Moreover, let  $n \in S$  and let  $\mathcal{R}_1^n, \dots, \mathcal{R}_{k_n}^n$  be the different  $\mathcal{R}$ -classes of  $Z(n)$ . Set  $\mu(n) = \max\{r_1^n, \dots, r_{k_n}^n\}$ , where  $r_i^n = \min\{|x| \mid x \in \mathcal{R}_i^n\}$ . Define

$$\mu(S) = \max\{\mu(n) \mid n \in S, k_n \geq 2\}.$$

**Theorem 1** [Chapman et al. 2009, Theorem 1]. *Let  $S$  be numerical semigroup. Then  $c(S) = \mu(S)$ .*

Let  $S$  be a numerical semigroup. An element  $s \in S$  is said to be a Betti element if  $Z(S)$  has more than one  $\mathcal{R}$ -class. Observe that there are finitely many Betti elements in  $S$  if it is finitely presented. The set of Betti elements of  $S$  is denoted by  $\text{Betti}(S)$ . As a consequence of the above theorem, we deduce that

$$c(S) = \max\{c(b) \mid b \in \text{Betti}(S)\}.$$

For the computation of the tame degree of the numerical semigroup  $S$ , a minimal presentation is not, in general, enough as shown in [Chapman et al. 2006]. Let  $\mathcal{I}(S)$  be the set of minimal nonnegative nonzero solutions of the equation

$$n_1x_1 + \cdots + n_px_p - n_1y_1 - \cdots - n_py_p = 0.$$

Let  $(x, y) = (x_1, \dots, x_p, y_1, \dots, y_p) \in \mathbb{N}^{2p}$ . Then  $(x, y)$  is a nonzero solution of the above equation if and only if  $(x_1, \dots, x_p)$  and  $(y_1, \dots, y_p)$  are elements in  $Z(\pi(x_1, \dots, x_p))$ . For  $n \in S$ , we write

$$\mathcal{I}_n(S) = \{(x_1, \dots, x_p, y_1, \dots, y_p) \in \mathcal{I}(S) \mid \pi(x_1, \dots, x_p) = n\}.$$

We have the following.

**Theorem 2** [Chapman et al. 2009, Theorem 2]. *Let  $S$  be a numerical semigroup minimally generated by  $\{n_1, \dots, n_p\}$ . Then*

$$t(S, \{n_i\}) = \max\{d(a, Z^i(\pi(a))) \mid a \in \mathbb{N}^p, \pi(a) - n_i \in S, \mathcal{I}_{\pi(a)}(S) \neq \emptyset\}.$$

And clearly,  $t(S) = \max\{t(S, \{n_i\}) \mid i \in \{1, \dots, p\}\}$ .

Let  $S$  be a numerical semigroup minimally generated by  $\{n_1, \dots, n_p\}$ , with  $p > 1$ . Let  $n \in S$ . Assume that  $n - n_i \in S$  for some  $i \in \{1, \dots, p\}$ . We define  $t_i(n) = \max\{d(z, Z^i(n)) \mid z \in Z(n)\}$ . Hence  $t(n) = \max\{t_i(n) \mid n - n_i \in S, 1 \leq i \leq p\}$ , and we have that  $t(S) = \max\{t(n) \mid n \in S\}$ .

Define

$$\text{Prim}(S) = \{n \in S \mid \text{there are } a, b \in Z(n) \text{ with } (a, b) \in \mathcal{I}(S) \text{ and } a \neq b\},$$

which we call the set of *primitive elements* of  $S$  (note that the condition  $a \neq b$  means  $(a, b) \neq (e_i, e_i)$  for all  $i$ ). As we observed above, the catenary degree of  $S$  is attained in one of its Betti elements. The tame degree, in light of the above theorem, is reached in a primitive element.

Given  $n \in S$ , define  $G_n$  as the graph with vertices given by the minimal generators  $n_i$  such that  $n - n_i \in S$ , and edges given by  $n_in_j$  if  $n - (n_i + n_j) \in S$ . It can be shown that the number of  $\mathcal{R}$ -classes (connected components of  $\nabla_n$ ) equals the number of connected components of  $G_n$  (see for instance [Rosales and García-Sánchez 2009, Chapter 7]). From [Blanco et al. 2011, Lemma 5.4], it can be deduced that if  $n$  is minimal in  $S$  with  $t(S) = t(n)$ , then the graph  $G_n$  is not complete, as

proved by Alfredo Sánchez-R. Navarro in a forthcoming Ph.D. dissertation. Denote by  $\text{NC}(S)$  the set

$$\text{NC}(S) = \{n \in S \mid G_n \text{ is not complete}\}.$$

Then

$$t(S) = \max\{t(s) \mid s \in \text{Prim}(S) \cap \text{NC}(S)\}.$$

**2.2. Symmetric numerical semigroups.** In this subsection we follow the notation used in [Rosales and García-Sánchez 2009, Chapter 3].

A numerical semigroup is *irreducible* if it cannot be expressed as the intersection of two numerical semigroups properly containing it.

A numerical semigroup  $S$  is *symmetric* if it is irreducible and  $F(S)$  is odd.

The following characterization is sometimes used as the definition of a symmetric numerical semigroup.

**Proposition 3.** *Let  $S$  be a numerical semigroup. Then,  $S$  is symmetric if and only if for all  $x \in \mathbb{Z}$ ,  $x \notin S$  implies  $F(S) - x \in S$ .*

**2.3. Gluing of numerical semigroups.** There is an easy way to obtain symmetric numerical semigroups from other symmetric numerical semigroups (this also applies to complete intersections, but for complete intersections this construction fully characterizes them). The proofs of the results in this paragraph can be found in [Rosales and García-Sánchez 2009, Chapters 7 and 8].

**Theorem 4.** *Let  $S$  be a numerical semigroup. Then the cardinality of a minimal presentation for  $S$  is greater than or equal to  $e(S) - 1$ .*

A numerical semigroup is a *complete intersection* if the cardinality of any of its minimal presentations equals its embedding dimension minus one.

Let  $S_1$  and  $S_2$  be two numerical semigroups minimally generated by  $\{n_1, \dots, n_r\}$  and  $\{n_{r+1}, \dots, n_e\}$ , respectively. Let  $\lambda \in S_1 \setminus \{n_1, \dots, n_r\}$  and  $\mu \in S_2 \setminus \{n_{r+1}, \dots, n_e\}$  be such that  $\text{gcd}(\lambda, \mu) = 1$ . We say that

$$S = \langle \mu n_1, \dots, \mu n_r, \lambda n_{r+1}, \dots, \lambda n_e \rangle$$

is a *gluing* of  $S_1$  and  $S_2$ .

The following characterization of complete intersections was first given by Delorme [1976] (though with different notation).

**Theorem 5.** *A numerical semigroup other than  $\mathbb{N}$  is a complete intersection if and only if it is a gluing of two complete intersection numerical semigroups.*

Also the symmetric property is preserved under gluings. As a consequence of this, every complete intersection numerical semigroup is symmetric.

**Proposition 6.** *A gluing of symmetric numerical semigroups is symmetric.*

**Corollary 7.** *Every complete intersection numerical semigroup is symmetric.*

**Corollary 8.** *Every numerical semigroup of embedding dimension two is symmetric.*

If in the process of gluing  $S_1$  and  $S_2$  we always take  $S_2$  to be a copy of  $\mathbb{N}$ , we obtain a special class of complete intersections. A numerical semigroup  $S$  is *free* if it is either  $\mathbb{N}$  or the gluing of a free numerical semigroup with  $\mathbb{N}$ .

**2.4. Numerical semigroups of embedding dimension three.**

**Theorem 9 [Herzog 1970].** *Let  $S$  be a numerical semigroup with embedding dimension three. Then,  $S$  is a complete intersection if and only if it is symmetric.*

Symmetric numerical semigroups with embedding dimension three are free since they are a gluing of a numerical semigroup of embedding dimension two and  $\mathbb{N}$ . This can be used to give an explicit description of the minimal generators of a semigroup of this kind.

**Theorem 10 [Rosales and García-Sánchez 2009, Theorem 10.6].** *Let  $m_1$  and  $m_2$  be two relatively prime integers greater than one. Let  $a, b$  and  $c$  be nonnegative integers with  $a \geq 2, b + c \geq 2$  and  $\gcd(a, bm_1 + cm_2) = 1$ .*

*Then  $S = \langle am_1, am_2, bm_1 + cm_2 \rangle$  is a symmetric numerical semigroup with embedding dimension three. Moreover, every symmetric numerical semigroup of embedding dimension three is of this form.*

Let  $S = \langle n_1 < n_2 < n_3 \rangle$  be a numerical semigroup of embedding dimension three. Define

$$c_i = \min\{k \in \mathbb{N} \setminus \{0\} \mid kn_i \in \langle n_j, n_k \rangle, \{i, j, k\} = \{1, 2, 3\}\}.$$

Then, for all  $\{i, j, k\} = \{1, 2, 3\}$ , there exists some  $r_{ij}, r_{ik} \in \mathbb{N}$  such that

$$c_i n_i = r_{ij} n_j + r_{ik} n_k.$$

From Example 8.23 and Theorem 8.17 in [loc. cit.], we know that

$$\text{Betti}(S) = \{c_1 n_1, c_2 n_2, c_3 n_3\}.$$

Hence  $1 \leq \# \text{Betti}(S) \leq 3$ . Herzog [1970] proved that  $S$  is symmetric if and only if  $r_{ij} = 0$  for some  $i, j \in \{1, 2, 3\}$ , or equivalently,  $\# \text{Betti}(S) \in \{1, 2\}$ . Therefore,  $S$  is nonsymmetric if and only if  $\# \text{Betti}(S) = 3$ .

**3. Catenary and tame degrees in embedding dimension three**

Let  $S$  be a numerical semigroup of embedding dimension three minimally generated by  $\{n_1, n_2, n_3\}$  with  $n_1 < n_2 < n_3$ . Corollary 5.8 in [Blanco et al. 2011] states that  $c(S) = t(S)$  for  $S$  a nonsymmetric numerical semigroup of embedding dimension three. It also gives an explicit formula for  $c(S)$  (and consequently  $t(S)$ ).



For this reason, we focus henceforth on the case when  $S$  is symmetric, and thus  $\# \text{Betti}(S) \in \{1, 2\}$ .

Notice that if  $n \in \text{Betti}(S)$ , then  $G_n$  is not connected, and so it cannot be complete. Hence  $\text{Betti}(S) \subseteq \text{NC}(S)$ . Also the minimality of  $c_i$  forces  $c_i n_i \in \text{Prim}(S)$ . Thus,

$$\text{Betti}(S) \subseteq \text{Prim}(S) \cap \text{NC}(S).$$

(This is true not only for embedding dimension three, but in this case the inclusion is straightforward.)

Numerical experiments were performed using the GAP package `numericalsgps` [GAP; Delgado et al. 2013].

**3.1. When  $S$  has two Betti elements.** We first give several technical lemmas that will be used in the following subcases.

Let  $c_i$  be as above. Denote by  $e_i$  the  $i$ -th row of the  $3 \times 3$  identity matrix.

**Lemma 11.** *Assume that  $c_i n_i = c_j n_j \neq c_k n_k$  for some  $\{i, j, k\} = \{1, 2, 3\}$ . Then*

- (1)  $Z(c_i n_i) = \{c_i e_i, c_j e_j\}$ ,
- (2) *the set  $Z(c_k n_k)$  has two  $\mathcal{R}$ -classes:  $\{c_k e_k\}$  and  $Z(c_k n_k) \setminus \{c_k e_k\}$ ,*
- (3)  *$S$  is uniquely presented if and only if  $Z(c_k n_k) \setminus \{c_k e_k\} = \{r_{ki} e_i + r_{kj} e_j\}$  for some  $r_{ki}, r_{kj} \in \mathbb{N} \setminus \{0\}$ , with  $0 < r_{ki} < c_i$  and  $0 < r_{kj} < c_j$ .*

*Proof.* (1) Assume that there exists  $a_i e_i + a_j e_j + a_k e_k \in Z(c_i n_i) \setminus \{c_i e_i, c_j e_j\}$ . Then  $a_i < c_i$  since otherwise  $(a_i - c_i)n_i + a_j n_j + a_k n_k = 0$ , which leads to  $a_i = c_i$ ,  $a_j = 0$  and  $a_k = 0$ , contradicting that  $a_i e_i + a_j e_j + a_k e_k \neq c_i e_i$ . Hence  $a_j n_j + a_k n_k = (c_i - a_i)n_i$ . The minimality of  $c_i$  forces  $a_i = 0$ . Arguing analogously, we obtain that  $a_j < c_j$ . But then  $(c_j - a_j)n_j = a_k n_k$ , and the minimality of  $c_j$  yields  $a_j = 0$ . Thus  $c_i n_i = c_j n_j = a_k n_k$ . This implies that  $a_k > c_k$  (the equality cannot hold since we are assuming that  $c_i n_i = c_j n_j \neq c_k n_k$ ). Thus,  $c_i n_i = c_j n_j = (a_k - c_k)n_k + r_{kj} n_j + r_{ki} n_i$  for some  $r_{kj}, r_{ki} \in \mathbb{N}$  with  $r_{kj} + r_{ki} \neq 0$ . Assume without loss of generality that  $r_{kj} \neq 0$ . Then the minimality of  $c_j$  forces  $c_j \leq r_{kj}$ , and consequently  $(a_k - c_k)n_k + (r_{kj} - c_j)n_j + r_{ki} n_i = 0$ , which is impossible since  $a_k - c_k \neq 0$ .

(2) We already know that  $c_k n_k \in \text{Betti}(S)$ , and so  $Z(c_k n_k)$  contains at least two  $\mathcal{R}$ -classes. Denote by  $R_1$  the one containing  $c_k e_k$ . If there exists another element in  $R_1$ , then there are some  $a_i, a_j, a_k \in \mathbb{N}$ ,  $a_k \neq 0$ , such that  $c_k n_k = a_i n_i + a_j n_j + a_k n_k$ . From the minimality of  $c_k$  we deduce that  $c_k \leq a_k$ , whence  $a_i n_i + a_j n_j + (a_k - c_k)n_k = 0$ . But this implies that  $a_i = a_j = 0$  and  $a_k = c_k$ , contradicting that  $a_i e_i + a_j e_j + a_k e_k$  was a factorization of  $c_k n_k$  different from  $c_k e_k$ .

Now take any other element in  $Z(c_k n_k) \setminus \{c_k e_k\}$ , say  $a_i e_i + a_j e_j + a_k e_k$ . By the same argument used in the preceding paragraph, we deduce that  $a_k = 0$ . Assume that  $a_i = 0$ . Then  $a_j n_j = c_k n_k$ , and the minimality of  $c_j$  implies that  $a_j > c_j$

(the equality cannot hold since  $c_j n_j \neq c_k n_k$ ). Hence  $(a_j - c_j)n_j + c_i n_i = c_k n_k$ , and  $a_j e_j \mathcal{R} (a_j - c_j)e_j + c_i e_i$ . The same holds if  $a_j = 0$ , and we deduce that all factorizations different from  $c_k e_k$  are  $\mathcal{R}$ -related.

(3) If  $S$  is uniquely presented, then  $Z(c_k n_k)$  has exactly two elements, say  $c_k e_k$  and  $r_{ki} e_i + r_{kj} e_j$ , each in a different  $\mathcal{R}$ -class [García-Sánchez and Ojeda 2010]. Observe that if either  $r_{ki} = 0$  or  $r_{kj} = 0$ , arguing as above, we deduce that  $c_k n_k$  has at least three factorizations, which is impossible. Also  $r_{ki} \geq c_i$  or  $r_{kj} \geq c_k$  yields a new factorization.

For the converse, assume that  $c_k n_k = r_{ki} n_i + r_{kj} n_j$  with  $0 < r_{ki} < c_i$  and  $0 < r_{kj} < c_j$ . If  $(a_k e_k + a_i e_i + a_j e_j) \in Z(c_3 n_3) \setminus \{c_k e_k, r_{ki} e_i + r_{kj} e_j\}$ , as  $Z(c_k n_k)$  has two  $\mathcal{R}$ -classes and one of them is  $\{c_k e_k\}$ , we have that  $a_k = 0$ . Hence  $a_i n_i + a_j n_j = r_{ki} n_i + r_{kj} n_j$ . If  $(a_i, a_j) \geq (r_{ki}, r_{kj})$ , we obtain  $(a_i - r_{ki})n_i + (a_j - r_{kj})n_j = 0$ , which yields  $a_i = r_{ki}$  and  $a_j = r_{kj}$ , which is impossible (here  $\leq$  denotes the usual partial order on  $\mathbb{N}^2$ ; that is,  $(a, b) \leq (c, d)$  if  $(c - a, d - b) \in \mathbb{N}^2$ , and analogously for  $\geq$ ). Also  $(a_i, a_j) \leq (r_{ki}, r_{kj})$  leads to the same contradiction. So, either  $a_i \geq r_{ki}$  and  $a_k \leq r_{kj}$  (and not equality in both), or  $a_i \leq r_{ki}$  and  $a_k \geq r_{kj}$ . By symmetry, and without loss of generality, assume that the first possibility holds. Then  $(a_i - r_{ki})n_i = (r_{kj} - a_j)n_j$ . But this implies that  $r_{kj} - a_j \geq c_j$ , whence  $r_{kj} \geq c_j$ , contradicting the hypothesis.  $\square$

Since we are assuming  $n_1 < n_2 < n_3$ , the following two lemmas are easy to prove.

**Lemma 12.** *The inequality  $c_3 < r_{31} + r_{32}$  holds for any  $r_{31} e_1 + r_{32} e_2 \in Z(c_3 n_3) \setminus \{c_3 e_3\}$ .*

*Proof.* Since  $n_1 < n_2 < n_3$ , we have  $c_3 n_3 = r_{31} n_1 + r_{32} n_2 < r_{31} n_3 + r_{32} n_3$ , and hence  $c_3 < r_{31} + r_{32}$ .  $\square$

**Lemma 13.** *For all  $r_{12} e_2 + r_{13} e_3 \in Z(c_1 n_1) \setminus \{c_1 e_1\}$ , we have  $r_{12} + r_{13} < c_1$ .*

*Proof.* We have  $c_1 n_1 = r_{12} n_2 + r_{13} n_3 > r_{12} n_1 + r_{13} n_1 = (r_{12} + r_{13})n_1$ , and thus  $r_{12} + r_{13} < c_1$ .  $\square$

*The case  $c_1 n_1 = c_2 n_2 \neq c_3 n_3$ .* Recall that we want to compute  $\mu(b)$  for  $b$  a Betti element (Theorem 1). So we must see what factorizations in every  $\mathcal{R}$ -class have minimum length.

In our setting  $c_1 n_1 = c_2 n_2$  implies  $c_2 < c_1$  because  $n_1 < n_2$ .

**Proposition 14.** *Let  $S = \langle n_1, n_2, n_3 \rangle$  with  $n_1 < n_2 < n_3$  and  $c_1 n_1 = c_2 n_2 \neq c_3 n_3$ . Then  $c(S) < t(S)$ .*

*Proof.* By Lemma 11,  $Z(c_3 n_3)$  has two  $\mathcal{R}$ -classes:  $\{c_3 e_3\}$  and  $Z(c_3 n_3) \setminus \{(c_3 e_3)\}$ . Denote  $\{c_3 e_3\}$  by  $R_1$  and its complement in  $Z(c_3 n_3)$  by  $R_2$ . Lemma 12 implies that  $c(c_3 n_3) = \min\{r + s \mid (r, s, 0) \in R_2\}$ , and as  $c(c_1 n_1) = c(c_2 n_2) = c_1$  ( $c_1 > c_2$ ), from Theorem 1, we deduce that

$$c(S) = \max\{c_1, \min\{r + s \mid (r, s, 0) \in R_2\}\}.$$

We distinguish two cases, depending on whether or not  $S$  is uniquely presented. Assume first that  $S$  is not uniquely presented. Let  $(u, v, 0) \in Z(c_3n_3)$  be such that  $u + v = \max\{r + s \mid (r, s, 0) \in R_2\}$ . As  $S$  is not uniquely presented, either  $u \geq c_1$  or  $v \geq c_2$ . If  $v \geq c_2$ , then  $(u + c_1, v - c_2, 0) \in Z(c_3n_3)$ , and  $u + c_1 + v - c_2 = u + v + (c_1 - c_2) > u + v$ , in contradiction with the maximality of  $u + v$ . Hence  $v < c_2$  and  $u \geq c_1$ . If  $u = c_1$ , then  $v \neq 0$  since  $c_1n_1 \neq c_3n_3$ . So  $u + v > c_1$ . Then  $t(S) \geq d((u, v, 0), (0, 0, c_3))$ , which by Lemma 12 equals  $u + v$ . Observe that  $u + v > \min\{r + s \mid (r, s, 0) \in R_2\}$ . Therefore

$$t(S) > \max\{\min\{r + s \mid (r, s, 0) \in R_2\}, c_1\} = c(S).$$

Now assume that  $S$  is uniquely presented. By Lemma 11, there exists one and only one  $(r_{31}, r_{32}) \in \mathbb{N}^2$  such that  $c_3n_3 = r_{31}n_1 + r_{32}n_2$  with  $0 < r_{31} < c_1$  and  $0 < r_{32} < c_2$ , and consequently  $r_{32} < c_1$ .

Take

$$n = (c_2 - r_{32})n_2 + c_3n_3 = r_{31}n_1 + c_2n_2 = (c_1 + r_{31})n_1.$$

Observe that  $n$  has just the three factorizations  $(0, c_2 - r_{32}, c_3)$ ,  $(r_{31}, c_2, 0)$  and  $(c_1 + r_{31}, 0, 0)$ . To see this, assume to the contrary that there exists  $a_1, a_2, a_3 \in \mathbb{N}$  such that  $n = a_1n_1 + a_2n_2 + a_3n_3$  and

$$(a_1, a_2, a_3) \notin \{(0, c_2 - r_{32}, c_3), (r_{31}, c_2, 0), (c_1 + r_{31}, 0, 0)\}.$$

Since  $a_1n_1 + a_2n_2 + a_3n_3 = (c_1 + r_{31})n_1$ , we easily deduce that  $a_1 < c_1 + r_{31}$ . Thus  $a_2n_2 + a_3n_3 = (r_{31} + c_1 - a_1)n_1$ , so  $c_1 + r_{31} - a_1 > c_1$ , and hence  $a_1 < r_{31} < c_1$ .

- If  $c_2 - r_{32} \leq a_2$ , from  $a_1n_1 + a_2n_2 + a_3n_3 = (c_2 - r_{32})n_2 + c_3n_3$ , we obtain  $(c_3 - a_3)n_3 = a_1n_1 + (a_2 - c_2 + r_{32})n_2 > 0$ . Hence  $c_3 - a_3 \geq c_3$ , or equivalently,  $a_3 \leq 0$ , which forces  $a_3 = 0$ . This implies  $c_3n_3 = a_1n_1 + (a_2 - c_2 + r_{32})n_2$ . As  $Z(c_3n_3) = \{c_3e_3, r_{31}e_1 + r_{32}e_2\}$ , we get  $a_2 = c_2$ , which is impossible.
- If, instead,  $a_2 < c_2 - r_{32}$ , from  $a_1n_1 + a_2n_2 + a_3n_3 = r_{31}n_1 + c_2n_2$ , we obtain  $a_3n_3 = (r_{31} - a_1)n_1 + (c_2 - a_2)n_2$  and then  $a_3 \geq c_3$ . Then, from  $a_1n_1 + a_2n_2 + a_3n_3 = (c_2 - r_{32})n_2 + c_3n_3$ , it follows that  $(c_2 - r_{32} - a_2)n_2 = a_1n_1 + (a_3 - c_3)n_3$ , whence  $c_2 - r_{32} - a_2 \geq c_2$ ; that is,  $r_{32} + a_2 \leq 0$ , a contradiction.

Hence we have  $Z(n) = \{(0, c_2 - r_{32}, c_3), (r_{31}, c_2, 0), (c_1 + r_{31}, 0, 0)\}$ . Observe that

$$t(n) \geq d((c_1 + r_{31}, 0, 0), (0, c_2 - r_{32}, c_3)) = \max\{c_2 - r_{32} + c_3, r_{31} + c_1\} = r_{31} + c_1$$

(because  $(r_{31} + c_1)n_1 = (c_2 - r_{32})n_2 + c_3n_3 > (c_2 - r_{32})n_1 + c_3n_1 = (c_2 - r_{32} + c_3)n_1$ , which yields  $r_{31} + c_1 > c_2 - r_{32} + c_3$ ). Then  $t(n) > \max\{c_1, r_{31} + r_{32}\}$ , and hence  $t(S) \geq t(n) > c(S)$ . □

**Example 15.** As an illustration, we offer a numerical semigroup of embedding dimension three  $\langle n_1, n_2, n_3 \rangle$  that is a gluing of  $\langle n_1, n_2 \rangle / \gcd(n_1, n_2)$  and  $\mathbb{N}$ .

We make use of the GAP package `numericalsgps` to perform the calculations. We try it with  $S = \langle 4, 6, 7 \rangle$ . Actually, we first started with  $S_1 = \langle 2, 3 \rangle$  and  $S_2 = \mathbb{N}$ , and glued them together as  $S = \langle 2 \times 2, 2 \times 3, 7 \times 1 \rangle$ ; that is,  $\lambda = 2$  and  $\mu = 7$  with the notations of [Section 2.3](#). The choices of  $\lambda = 2$  and  $\mu = 7$  are restricted by the following facts: they must belong to  $S_2$  and  $S_1$ , respectively, and cannot be minimal generators; we also need  $n_1 < n_2 < n_3$ .

```
gap> s:=NumericalSemigroup(4,6,7);
<Numerical semigroup with 3 generators>
gap> AsGluingOfNumericalSemigroups(s);
[ [ [ 4, 6 ], [ 7 ] ] ]
```

Now we compute a minimal presentation of  $S$  and the Betti elements of  $S$ .

```
gap> MinimalPresentationOfNumericalSemigroup(s);
[ [ [ 2, 1, 0 ], [ 0, 0, 2 ] ], [ [ 3, 0, 0 ], [ 0, 2, 0 ] ] ]
gap> BettiElementsOfNumericalSemigroup(s);
[ 12, 14 ]
```

Finally, we see that  $c(S) < t(S)$ .

```
gap> CatenaryDegreeOfNumericalSemigroup(s);
3
gap> TameDegreeOfNumericalSemigroup(s);
5
```

*The case  $c_1 n_1 \neq c_2 n_2 = c_3 n_3$ .* Observe that  $c_2 n_2 = c_3 n_3$  forces  $c_3 < c_2$ .

**Lemma 16.** *If  $c_1 n_1 \neq c_2 n_2 = c_3 n_3$ , then  $c(S) = \max\{c_1, c_2\}$ .*

*Proof.* By [Theorem 1](#), the catenary degree is reached in one of the two Betti elements:  $\text{Betti}(S) = \{c_1 n_1, c_2 n_2 = c_3 n_3\}$ . From [Lemma 11](#), we have  $c(c_2 n_2) = \max\{c_2, c_3\} = c_2$ , and from [Lemma 13](#),  $c(c_1) = c_1$ . So  $c(S) = \max\{c_1, c_2\}$ .  $\square$

**Proposition 17.** *Let  $S = \langle n_1, n_2, n_3 \rangle$  with  $n_1 < n_2 < n_3$  and  $c_1 n_1 \neq c_2 n_2 = c_3 n_3$ . If  $c_2 n_2 \nmid c_1 n_1$ , then  $t(S) > c(S)$ .*

*Proof.* From [Lemma 16](#), we know that  $c(S) = \max\{c_1, c_2\}$ . As before, we distinguish two cases, depending on whether or not  $S$  is uniquely presented.

Assume first that  $S$  is uniquely presented. In light of [Lemma 11](#), there exists  $r_{12}, r_{13} \in \mathbb{N} \setminus \{0\}$  such that  $Z(c_1 n_1) = \{c_1 e_1, r_{12} e_2 + r_{13} e_3\}$ ,  $r_{12} < c_2$  and  $r_{13} < c_3$  (thus  $r_{13} < c_2$ ). Set

$$n = c_1 n_1 + (c_2 - r_{12}) n_2 = c_2 n_2 + r_{13} n_3 = (c_3 + r_{13}) n_3.$$

As in the proof of [Proposition 14](#), we can see that

$$Z(n) = \{(c_1, c_2 - r_{12}, 0), (0, c_2, r_{13}), (0, 0, c_3 + r_{13})\}.$$

Then  $t(n) \geq d((c_1, c_2 - r_{12}, 0), (0, 0, c_3 + r_{13})) = c_1 + c_2 - r_{12}$  (since  $c_1 > r_{12} + r_{13}$  and  $c_2 > c_3$  imply  $c_1 + c_2 - r_{12} > c_3 + r_{13}$ ). By observing that  $c_1 > r_{12}$ , we get  $c_1 - r_{12} > 0$ , and then  $c_1 + c_2 - r_{12} > c_2$ . Also  $r_{12} < c_2$  implies  $c_1 + c_2 - r_{12} > c_1$ . So  $t(n) \geq c_1 + c_2 - r_{12} > \max\{c_1, c_2\} = c(S)$ , and we conclude that  $t(S) \geq t(n) > c(S)$ .

Now suppose  $S$  is not uniquely presented. From [Lemma 11](#), we deduce that there exists an expression  $c_1n_1 = r_{12}n_2 + r_{13}n_3$ , and we have either  $r_{12} \geq c_2$  or  $r_{13} \geq c_3$ . Without loss of generality suppose that  $r_{13} \geq c_3$ . If  $r_{12} \geq c_2$ , we derive  $c_1n_1 = (r_{12} - c_2)n_2 + (r_{13} + c_3)n_3$ . So we can assume, in addition, that  $r_{12} < c_2$ .

Case 1: If  $r_{12} \neq 0$ , take  $n = (c_3 + r_{13})n_3$ . We prove that the only factorization with nonzero first coordinate of  $n$  is  $(c_1, c_2 - r_{12}, 0)$ . Assume to the contrary that

$$(c_3 + r_{13})n_3 = c_1a_1 + (c_2 - r_{12})n_2 = a_1n_1 + a_2n_2 + a_3n_3,$$

with  $a_1, a_2, a_3 \in \mathbb{N}$ ,  $a_1 \neq 0$  and  $(a_1, a_2, a_3) \neq (c_1, c_2 - r_{12}, 0)$ . Then  $a_3 < c_3 + r_{13}$  since otherwise  $a_1n_1 + a_2n_2 + (a_3 - c_3 - r_{13})n_3 = 0$ , and this forces  $a_1 = 0$ , a contradiction. Hence  $(c_3 + r_{13} - a_3)n_3 = a_1n_1 + a_2n_2$ , and thus  $c_3 + r_{13} - a_3 \geq c_3$ , or equivalently,  $r_{13} \geq a_3$ . Thus  $c_3n_3 + (r_{13} - a_3)n_3 = a_1n_1 + a_2n_2$ , which leads to  $c_2n_2 + (r_{13} - a_3)n_3 = a_1n_1 + a_2n_2$ . Since  $r_{12} \neq 0$ , we derive  $a_1 < c_1$  because otherwise  $(c_2 - r_{12})n_2 = (a_1 - c_1)n_1 + a_2n_2 + a_3n_3$ , and this either leads to  $a_1 = c_1$ ,  $a_2 = c_2 - r_{12}$  and  $a_3 = 0$ , which is impossible, or contradicts the minimality of  $c_2$ . As  $c_2n_2 + (r_{13} - a_3)n_3 = a_1n_1 + a_2n_2$  and  $a_1 < c_1$ , we have  $a_2 \geq c_2$ . Hence  $(r_{13} - a_3)n_3 = a_1n_1 + (a_2 - c_2)n_2$ . This again leads to  $r_{12} - a_3 \geq c_3$ . We can repeat the process and obtain  $(r_{13} - a_3 - kc_3)n_3 = a_1n_1 + (a_2 - (k + 1)c_2)n_2$  for all  $k \in \mathbb{N}$ , which leads also to a contradiction.

Now, we have that

$$t(n) \geq d((0, 0, c_3 + r_{13}), (c_1, c_2 - r_{12}, 0)) = \max\{c_3 + r_{13}, c_1 + c_2 - r_{12}\} = c_1 + c_2 - r_{12}$$

because  $(c_3 + r_{13})n_3 = c_1n_1 + (c_2 - r_{12})n_2 < c_1n_3 + (c_2 - r_{12})n_3 = (c_1 + c_2 - r_{12})n_3$ . Thus this distance is greater than both  $c_1$  and  $c_2$ . In fact,  $c_1 + c_2 - r_{12} > c_1$  follows easily from  $c_2 > r_{12}$ , and  $c_1 + c_2 - r_{12} > c_2$  follows from  $c_1 > r_{12} + r_{13}$  ([Lemma 16](#)).

Case 2: If  $r_{12} = 0$ , then  $c_1n_1 = r_{13}n_3$ , so we get the inequalities  $c_3 < r_{13} < c_1$ . Take  $h = \min\{m \in \mathbb{N} \mid mc_3 > r_{13}\}$  ( $h \geq 2$ ) and let us consider  $n = hc_3n_3$ . Clearly,  $\{(0, 0, hc_3), (c_1, 0, hc_3 - r_{13}), (0, hc_2, 0)\} \subseteq Z(n)$ . We prove that the only factorization of  $n$  with nonzero first coordinate is  $(c_1, 0, hc_3 - r_{13})$ .

To see this, notice that the minimality of  $h$  forces  $hc_3 - r_{13} \leq c_3$  since otherwise  $(h - 1)c_3 > r_{13}$ . Also  $hc_3 - r_{13} = c_3$  implies that  $(h - 1)c_3 = r_{13}$ , and consequently  $c_1n_1 = r_{13}n_3 = (h - 1)c_3n_3 = (h - 1)c_2n_2$ , which means that  $c_2n_2 \mid c_1n_1$ , contradicting

the hypothesis. Hence  $hc_2 - r_{13} < c_3$ . Assume that there is another expression of the form  $n = hc_3n_3 = a_1n_1 + a_2n_2 + a_3n_3$  with  $a_1 \neq 0$ . We can assume that  $a_2 < c_2$  because otherwise  $(a_1, a_2 - c_2, a_3 + c_3)$  is another factorization of  $n$ , and we can repeat this procedure until the second coordinate is less than  $c_2$ . Thus  $(hc_3 - r_{13})n_3 + c_1n_1 = a_1n_1 + a_2n_2 + a_3n_3$ .

- If  $a_3 \geq hc_3 - r_{13}$ , then  $c_1n_1 = a_1n_1 + a_2n_2 + (a_3 + r_{13} - hc_3)n_3$ . The minimality of  $c_1$  forces  $a_1 \geq c_1$ , and consequently  $(a_1 - c_1)n_1 + a_2n_2 + (a_3 + r_{13} - hc_3)n_3 = 0$ . This can only happen if  $(a_1, a_2, a_3) = (c_1, 0, hc_3 - r_{13})$ , a contradiction.
- If  $a_3 < hc_3 - r_{13}$ , then  $(hc_3 - r_{13} - a_3)n_3 + c_1n_1 = a_1n_1 + a_2n_2$ . As  $hc_3 - r_{13} < c_3$ , it follows that  $c_1 > a_1$ , and thus  $(hc_3 - r_{13} - a_3)n_3 + (c_1 - a_1)n_1 = a_2n_2$ . But this forces  $a_2 = 0$  since otherwise  $a_2 \geq c_2$ , contradicting the choice of  $a_2$ . Again we obtain  $(a_1, a_2, a_3) = (c_1, 0, hc_3 - r_{13})$ .

Since  $hc_3 > r_{13}$  and  $hc_2 > c_2$ , we have

$$\begin{aligned} t(S) &\geq d((c_1, 0, hc_3 - r_{13}), (0, hc_2, 0)) \\ &= \max\{c_1 + hc_3 - r_{13}, hc_2\} > \max\{c_1, c_2\} = c(S). \quad \square \end{aligned}$$

**Example 18.** We use the same idea of [Example 15](#). Here we need a gluing of  $\mathbb{N}$  and  $\langle n_2, n_3 \rangle / \gcd(n_2, n_3)$ . We start again with  $\mathbb{N}$  and  $\langle 2, 3 \rangle$ . As we need  $n_1 < n_2 < n_3$ , we choose, for example,  $\lambda = 5$  and  $\mu = 4$ , obtaining  $S = \langle 5, 8, 12 \rangle$ .

```
gap> s:=NumericalSemigroup(5,8,12);;
gap> AsGluingOfNumericalSemigroups(s);
[ [ [ 5 ], [ 8, 12 ] ] ]
```

The minimal presentation and Betti elements of  $S$  are

```
gap> MinimalPresentationOfNumericalSemigroup(s);
[ [ [ 0, 3, 0 ], [ 0, 0, 2 ] ], [ [ 4, 0, 0 ], [ 0, 1, 1 ] ] ]
gap> BettiElementsOfNumericalSemigroup(s);
[ 20, 24 ]
```

Finally, we check that indeed  $c(S) < t(S)$ .

```
gap> CatenaryDegreeOfNumericalSemigroup(s);
4
gap> TameDegreeOfNumericalSemigroup(s);
6
```

**Proposition 19.** *Let  $S = \langle n_1, n_2, n_3 \rangle$  with  $n_1 < n_2 < n_3$  and  $c_1n_1 \neq c_2n_2 = c_3n_3$ . If  $c_2n_2 \mid c_1n_1$ , then  $t(S) = c(S)$ .*

*Proof.* Since  $c_2n_2 \mid c_1n_1$  and  $c_2n_2 \neq c_1n_1$ , we deduce that  $c_1n_1 = kc_2n_2$  for some integer  $k \geq 1$ .

We start by proving that  $\text{Betti}(S) = \text{Prim}(S) \cap \text{NC}(S)$ . Assume that there exists  $n \in (\text{Prim}(S) \cap \text{NC}(S)) \setminus \text{Betti}(S)$ . Then, for some permutation  $(i, j, k)$  of  $(1, 2, 3)$  and some  $a_i, a_j, a_k \in \mathbb{N}$  with  $a_i > 0$  and  $a_j + a_k \geq 2$ , we have  $n = a_i n_i = a_j n_j + a_k n_k$  and  $a_i e_i + a_j e_{j+3} + a_k e_{k+3} \in \mathcal{I}_n(S)$ . We distinguish three cases depending on  $i$ .

Case 1: If  $i = 1$ , then  $n = a_1 n_1 = a_2 n_2 + a_3 n_3$ . Hence  $a_1 \geq c_1$ , and since  $n \notin \text{Betti}(S)$ ,  $a_1 > c_1$ . This implies that

$$n = a_1 n_1 = (a_1 - c_1)n_1 + c_1 n_1 = (a_1 - c_1)n_1 + (k - 1)c_2 n_2 + c_3 n_2,$$

and consequently the graph associated to  $n$  is complete, a contradiction.

Case 2: If  $i = 2$ , then  $n = a_2 n_2 = a_1 n_1 + a_3 n_3$ . As above, we deduce that  $a_2 > c_2$ . Hence  $n = a_2 n_2 = (a_2 - c_2)n_2 + c_3 n_3 = a_1 n_1 + a_3 n_3$ , and in particular the edge  $n_2 n_3$  is in the graph associated to  $n$ .

Assume that  $a_3 \geq c_3$ . Then  $(a_2 - c_2)n_2 = a_1 n_1 + (a_3 - c_3)n_3$ . But this implies that  $(a_1, 0, a_3 - c_3, 0, a_2 - c_2, 0) < (a_1, 0, a_3, 0, a_2, 0)$ , contradicting that  $n \in \text{Prim}(S)$ . Thus,  $a_3 < c_3$ , and then  $(a_2 - c_2)n_2 + (c_3 - a_3)n_3 = a_1 n_1$ . The minimality of  $c_1$  leads to  $a_1 \geq c_1$ . If  $a_1 = c_1$ , then  $a_2 n_2 = kc_2 n_2 + a_3 n_3$ . The fact that  $a_3 < c_3$  forces  $kc_2 \geq a_2$ . But then,  $0 = (kc_2 - a_2)n_2 + a_3 n_3$  which implies that  $a_3 = 0$ , and consequently  $n = c_1 n_1 \in \text{Betti}(S)$ , a contradiction. It follows that  $a_1 > c_1$ . We conclude that

$$\begin{aligned} n = a_2 n_2 = a_1 n_1 + a_3 n_3 &= (a_1 - c_1)n_1 + kc_2 n_2 + a_3 n_3 \\ &= (a_1 - c_1)n_1 + (kc_3 + a_3)n_3, \end{aligned}$$

and thus the graph associated to  $n$  is complete.

Case 3: The case  $i = 3$  is analogous to the previous one.

Hence  $t(S) = \max\{t(c_1 n_1), t(c_2 n_2)\}$ . We already know that  $Z(c_2 n_2) = \{c_2 e_2, c_3 e_3\}$ , and then  $t(c_2 n_2) = c_2$ . Also every factorization of  $c_1 n_1$  is either  $c_1 e_1$  or some  $x e_2 + y e_3$  with  $x + y < c_1$ . It follows that  $t(c_1 n_1) = c_1$ . We conclude the proof by using [Lemma 16](#). □

**Example 20.** We use once more  $S_1 = \mathbb{N}$  and  $S_2 = \langle 2, 3 \rangle$ . We need  $c_2 n_2 \mid c_1 n_1$ . We choose  $\lambda = 12$  and  $\mu = 7$ , obtaining  $S = \langle 12, 14, 21 \rangle$ .

```
gap> s:=NumericalSemigroup(12,14,21);;
gap> AsGluingOfNumericalSemigroups(s);
[ [ [ 12 ] ], [ 14, 21 ] ], [ [ 12, 14 ] ], [ 21 ] ],
                               [ [ 12, 21 ] ], [ 14 ] ] ]
gap> MinimalPresentationOfNumericalSemigroup(s);
[ [ [ 0, 3, 0 ], [ 0, 0, 2 ] ], [ [ 7, 0, 0 ], [ 0, 0, 4 ] ] ]
gap> BettiElementsOfNumericalSemigroup(s);
[ 42, 84 ]
```

Thus  $c_1n_1 = 7 \times 12 = 2^2 \times 3 \times 7$ , which is a multiple of  $c_2n_2 = 3 \times 14 = 2 \times 3 \times 7$ . We check that the tame and catenary degrees agree in this case.

```
gap> CatenaryDegreeOfNumericalSemigroup(s);
7
gap> TameDegreeOfNumericalSemigroup(s);
7
```

The case  $c_1n_1 = c_3n_3 \neq c_2n_2$ .

**Proposition 21.** *Let  $S = \langle n_1, n_2, n_3 \rangle$  with  $n_1 < n_2 < n_3$  and  $c_1n_1 = c_3n_3 \neq c_2n_2$ . Then  $c(S) < t(S)$ .*

*Proof.* The catenary degree is reached in one of the two Betti elements,  $\text{Betti}(S) = \{c_1n_1, c_2n_2\}$ .

We know that  $c(c_1n_1) = c_1$  and that  $Z(c_2n_2)$  has just two  $\mathcal{R}$ -classes, say  $R_1 = \{(0, c_2, 0)\}$  and  $R_2 = Z(c_2n_2) \setminus R_1$  (Lemma 11). Take  $(r_{21}, 0, r_{23}) \in R_2$  such that  $r_{21} + r_{23} = \min\{r + s \mid (r, 0, s) \in R_2\}$ . Hence,  $c(c_2n_2) = \max\{c_2, r_{21} + r_{23}\}$ . So we can conclude that  $c(S) = \max\{c_1, c_2, r_{21} + r_{23}\}$  (Theorem 1).

Since  $c_2n_2 = r_{21}n_1 + r_{23}n_3 > r_{23}n_2$ , we have  $r_{23} < c_2$ . Moreover,  $c_1 > c_3$ , and so if  $r_{21} \geq c_1$ , we have  $r_{21}n_1 + r_{23}n_3 = (r_{21} - c_1)n_1 + (r_{23} + c_3)n_3$ , with  $r_{21} + r_{23} > r_{21} + r_{23} + c_3 - c_1$ , contradicting the minimality of  $r_{21} + r_{23}$ . Therefore,  $r_{21} < c_1$ .

We distinguish two cases.

Case 1: If  $r_{21} \neq 0$ , then take  $n = (c_1 - r_{21})n_1 + c_2n_2 = c_1n_1 + r_{23}n_3 = (c_3 + r_{23})n_3$ . We prove that the only factorization of  $n$  with nonzero second coordinate is  $(c_1 - r_{21}, c_2, 0)$ . Assume that there exists  $(a_1, a_2, a_3) \in Z(n) \setminus \{(c_1 - r_{21}, c_2, 0)\}$  with  $a_2 \neq 0$ . Since  $a_1n_1 + a_2n_2 + a_3n_3 = (c_3 + r_{23})n_3$ , we can easily deduce that  $a_3 < c_3 + r_{23}$ . Thus  $a_1n_1 + a_2n_2 = (c_3 + r_{23} - a_3)n_3$ , so  $c_3 + r_{23} - a_3 > c_3$ , and hence  $a_3 < r_{23}$ .

If  $c_1 - r_{21} \leq a_1$ , from  $a_1n_1 + a_2n_2 + a_3n_3 = (c_1 - r_{21})n_1 + c_2n_2$ , we obtain  $(c_2 - a_2)n_2 = (a_1 - c_1 + r_{21})n_1 + a_3n_3 > 0$ . Hence  $c_2 - a_2 \geq c_2$ , or equivalently  $a_2 \leq 0$ , which forces  $a_2 = 0$ .

If, instead,  $a_1 < c_1 - r_{21}$ , from  $a_1n_1 + a_2n_2 + a_3n_3 = c_1n_1 + r_{23}n_3$ , we obtain  $a_2n_2 = (r_{23} - a_3)n_3 + (c_1 - a_1)n_1$ , and then  $a_2 \geq c_2$ . From  $a_1n_1 + a_2n_2 + a_3n_3 = (c_1 - r_{21})n_1 + c_2n_2$ , it follows that  $(c_1 - r_{21} - a_1)n_1 = (a_2 - c_2)n_2 + a_3n_3$ . Thus,  $c_1 - r_{21} - a_1 \geq c_1$ , that is,  $r_{21} + a_1 \leq 0$ , and then  $a_1 = r_{21} = 0$ , a contradiction.

Hence,  $t(n) \geq d((c_1 - r_{21}, c_2, 0), (0, 0, c_3 + r_{23})) = \max\{c_1 - r_{21} + c_2, c_3 + r_{23}\} = c_1 - r_{21} + c_2$ , since  $(c_3 + r_{23})n_3 = (c_1 - r_{21})n_1 + c_2n_2 < (c_1 - r_{21} + c_2)n_3$ .

Now we have

- $c_1 - r_{21} + c_2 > c_1$  since  $(c_2 - r_{21})n_2 > c_2n_2 - r_{21}n_1 = r_{23}n_3 > 0$  implies  $c_2 - r_{21} > 0$ ;
- $c_1 - r_{21} + c_2 > c_2$  since  $r_{21} > c_1$ ;



- $c_1 - r_{21} + c_2 > r_{21} + r_{23}$  since  $c_1 > r_{21}$  and  $(c_2 - r_{21})n_2 > c_2n_2 - r_{21}n_1 = r_{23}n_3 > r_{23}n_2$  implies  $c_2 - r_{21} > r_{23}$ .

So we finally have that

$$t(S) \geq d((c_1 - r_{21}, c_2, 0), (0, 0, c_3 + r_{23})) > \max\{c_1, c_2, r_{21} + r_{23}\} = c(S).$$

Case 2: If  $r_{21} = 0$ , then  $c_2n_2 = r_{23}n_3$ , so we deduce the inequalities  $c_3 < r_{23} < c_2$ . Take  $h = \min\{m \mid mc_3 > r_{23}\}$  ( $h \geq 2$ ) and let us consider  $n = hc_3n_3$ . It follows that

$$\{(0, 0, hc_3), (0, c_2, hc_3 - r_{23}), (hc_1, 0, 0)\} \subset Z(n).$$

Arguing as in [Proposition 17](#), we can prove that the only possible factorizations with nonzero second coordinate are  $(0, c_2, hc_3 - r_{23})$  and  $(c_1, c_2, 0)$  (this one occurs only if  $hc_3 - r_{23} = c_3$ ).

So we have

- $d((0, c_2, hc_3 - r_{23}), (hc_1, 0, 0)) = \max\{c_2 + hc_3 - r_{23}, hc_1\} > \max\{c_1, c_2, r_{23}\} = \max\{c_1, c_2\} = c(S)$  since  $hc_3 > r_{23}$  and  $hc_1 > c_1$ ;
- if  $hc_3 - r_{23} = c_3$ , then  $c_2n_2 = (h - 1)c_1n_1$ , and consequently  $(h - 1)c_1 > c_2$  and  $h - 1 > 1$  (recall that  $c_2n_2 \neq c_1n_1$ ), whence

$$d((c_1, c_2, 0), (hc_1, 0, 0)) = \max\{(h - 1)c_1, c_2\} > c(S).$$

We conclude that  $t(S) > c(S)$ . □

**Example 22.** As in the preceding example we start with  $S_1 = \mathbb{N}$  and  $S_2 = \langle 2, 3 \rangle$ . We need  $n_1 < n_2 < n_3$ , that is  $2\mu < \lambda < 3\mu$ . For the first case of the proof of [Proposition 21](#) ( $r_{21} \neq 0$ ), we choose  $\lambda = 5$  and  $\mu = 2$ .

```
gap> s:=NumericalSemigroup(4,5,6);;
gap> AsGluingOfNumericalSemigroups(s);
[[ [ 4, 6 ], [ 5 ] ] ]
gap> MinimalPresentationOfNumericalSemigroup(s);
[[ [ 0, 2, 0 ], [ 1, 0, 1 ] ], [ [ 3, 0, 0 ], [ 0, 0, 2 ] ] ]
gap> BettiElementsOfNumericalSemigroup(s);
[ 10, 12 ]
gap> CatenaryDegreeOfNumericalSemigroup(s);
3
gap> TameDegreeOfNumericalSemigroup(s);
4
```

For the second case,  $r_{21} = 0$ , we choose  $\lambda = 18$  and  $\mu = 7$ .

```
gap> s:=NumericalSemigroup(14,18,21);;
gap> AsGluingOfNumericalSemigroups(s);
[[ [ 14 ], [ 18, 21 ] ], [ [ 14, 18 ], [ 21 ] ] ],
```

```

[ [ 14, 21 ], [ 18 ] ] ]
gap> MinimalPresentationOfNumericalSemigroup(s);
[ [ [ 0, 0, 6 ], [ 0, 7, 0 ] ], [ [ 3, 0, 0 ], [ 0, 0, 2 ] ] ]
gap> BettiElementsOfNumericalSemigroup(s);
[ 42, 126 ]
gap> CatenaryDegreeOfNumericalSemigroup(s);
7
gap> TameDegreeOfNumericalSemigroup(s);
9

```

**3.2. When  $S$  has a single Betti element.** Numerical semigroups having a single Betti element are fully characterized in [García Sánchez et al. 2013, Theorem 12]. The following proposition is a particular instance of [loc. cit., Theorem 19]; we include it here for sake of completeness.

**Proposition 23.** *Let  $S = \langle n_1, n_2, n_3 \rangle$  with  $n_1 < n_2 < n_3$  and  $c_1 n_1 = c_2 n_2 = c_3 n_3$ . Then  $c(S) = t(S)$ .*

*Proof.* Take  $h = c_1 n_1 = c_2 n_2 = c_3 n_3$ . The catenary degree of  $S$  is reached in one of the Betti elements; since in our case  $\text{Betti}(S) = \{c_1 n_1 = c_2 n_2 = c_3 n_3 = h\}$ , we get  $c(S) = c(h) = \max\{c_1, c_2, c_3\} = c_1$ .

We know that the tame degree is reached in some  $n \in \text{Prim}(S) \cap \text{NC}(S)$ . Since we have that  $\text{Betti}(S) \subseteq \text{Prim}(S) \cap \text{NC}(S)$  and  $t(h) = \max\{c_1, c_2, c_3\} = c_1$ , in order to prove that  $c(S) = t(S)$ , we show that  $\text{Betti}(S) = \text{Prim}(S) \cap \text{NC}(S)$ . To this end, take  $n \in (\text{Prim}(S) \cap \text{NC}(S)) \setminus \text{Betti}(S)$ . So  $n = a_i n_i = a_j n_j + a_k n_k$  for some  $\{i, j, k\} = \{1, 2, 3\}$ . It follows that  $a_i \geq c_i$  and, since  $n \notin \text{Betti}(S)$ , we have  $a_i \neq c_i$ . So  $a_i > c_i$ . Then we have two cases:

- If  $a_j a_k \neq 0$ , then  $n \notin \text{NC}(S)$  because  $n = (a_i - c_i)n_i + c_j n_j = (a_i - c_i)n_i + c_k n_k$ , and consequently  $G_n$  is a triangle.
- If  $a_j = 0$ , then  $a_k > c_k$ , so we get  $(a_k - c_k)n_k + c_j n_j = a_i n_i = (a_i - c_i)n_i + c_k n_k$ , and then  $G_n$  is a triangle.

In any case we get a contradiction. □

**Example 24.** If we want  $c_1 n_1 = c_2 n_2 = c_3 n_3$ , according to [García Sánchez et al. 2013, Theorem 12], we need three pairwise coprime integers greater than one, and then we need to take all of the products of any two of them. The easiest example is 2, 3, 5, and thus  $n_1 = 2 \times 3$ ,  $n_2 = 2 \times 5$  and  $n_3 = 3 \times 5$ .

```

gap> s:=NumericalSemigroup(6,10,15);
<Numerical semigroup with 3 generators>
gap> AsGluingOfNumericalSemigroups(s);
[ [ [ 6 ], [ 10, 15 ] ], [ [ 6, 10 ], [ 15 ] ],

```

```

[ [ 6, 15 ], [ 10 ] ] ]
gap> BettiElementsOfNumericalSemigroup(s);
[ 30 ]
gap> MinimalPresentationOfNumericalSemigroup(s);
[ [ [ 5, 0, 0 ], [ 0, 0, 2 ] ], [ [ 5, 0, 0 ], [ 0, 3, 0 ] ] ]
gap> CatenaryDegreeOfNumericalSemigroup(s);
5
gap> TameDegreeOfNumericalSemigroup(s);
5

```

#### 4. Main result

Gathering the results from the previous section, we obtain the following theorem.

**Theorem 25.** *Let  $S$  be a numerical semigroup of embedding dimension three minimally generated by  $\{n_1, n_2, n_3\}$ . For every  $\{i, j, k\} = \{1, 2, 3\}$ , define*

$$c_i = \min\{k \in \mathbb{N} \setminus \{0\} \mid kn_i \in \langle n_j, n_k \rangle\}.$$

Then  $c(S) = t(S)$  if and only if

- either  $\# \text{Betti}(S) \neq 2$ ,
- or  $c_1 n_1 \neq c_2 n_2 = c_3 n_3$  and  $c_2 n_2$  divides  $c_1 n_1$ .

#### References

- [Blanco et al. 2011] V. Blanco, P. A. García-Sánchez, and A. Geroldinger, “Semigroup-theoretical characterizations of arithmetical invariants with applications to numerical monoids and Krull monoids”, *Illinois J. Math.* **55**:4 (2011), 1385–1414. MR 3082874 Zbl 1279.20072
- [Chapman et al. 2006] S. T. Chapman, P. A. García-Sánchez, D. Llena, V. Ponomarenko, and J. C. Rosales, “The catenary and tame degree in finitely generated commutative cancellative monoids”, *Manuscripta Math.* **120**:3 (2006), 253–264. MR 2007d:20106 Zbl 1117.20045
- [Chapman et al. 2009] S. T. Chapman, P. A. García-Sánchez, and D. Llena, “The catenary and tame degree of numerical monoids”, *Forum Math.* **21**:1 (2009), 117–129. MR 2010i:20081 Zbl 1177.20070
- [Delgado et al. 2013] M. Delgado, P. A. García-Sánchez, and J. Morais, *NumericalSgps: A package for numerical semigroups*, 2013, <http://www.gap-system.org/Packages/numericalsgps.html>.
- [Delorme 1976] C. Delorme, “Sous-monoïdes d’intersection complète de  $N$ ”, *Ann. Sci. École Norm. Sup.* (4) **9**:1 (1976), 145–154. MR 53 #10821 Zbl 0325.20065
- [GAP] *GAP: Groups, Algorithms, Programming — a system for computational discrete algebra*, The GAP Group, <http://www.gap-system.org>.
- [García-Sánchez and Ojeda 2010] P. A. García-Sánchez and I. Ojeda, “Uniquely presented finitely generated commutative monoids”, *Pacific J. Math.* **248**:1 (2010), 91–105. MR 2011j:20139 Zbl 1208.20052
- [García Sánchez et al. 2013] P. A. García Sánchez, I. Ojeda, and J. C. Rosales, “Affine semigroups having a unique Betti element”, *J. Algebra Appl.* **12**:3 (2013), 1250177. MR 3007913 Zbl 1281.20075

[Geroldinger and Halter-Koch 2006] A. Geroldinger and F. Halter-Koch, *Non-unique factorizations: Algebraic, combinatorial and analytic theory*, Pure and Applied Mathematics **278**, Chapman & Hall/CRC, Boca Raton, FL, 2006. MR 2006k:20001 Zbl 1113.11002

[Herzog 1970] J. Herzog, “Generators and relations of abelian semigroups and semigroup rings”, *Manuscripta Math.* **3** (1970), 175–193. MR 42 #4657 Zbl 0211.33801

[Rosales and García-Sánchez 2009] J. C. Rosales and P. A. García-Sánchez, *Numerical semigroups*, Developments in Mathematics **20**, Springer, New York, 2009. MR 2010j:20091 Zbl 1220.20047

Received: 2014-07-13

Revised: 2014-08-20

Accepted: 2014-09-07

[pedro@ugr.es](mailto:pedro@ugr.es)

*Departamento de Álgebra, Universidad de Granada, Facultad de Ciencias, Av. Fuentenueva, s/n, 18071 Granada, Spain*

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

*Dipartimento di Matematica e Informatica, Università degli Studi di Catania, Viale A. Doria, 6, I-95125 Catania, Italy*

# Cylindrical liquid bridges

Lamont Colter and Ray Treinen

(Communicated by Frank Morgan)

We consider a cylindrical liquid bridge under capillary effects, spanning two horizontal plates and further bounded by a pair of parallel vertical planes. We explicitly formulate the volume-constrained problem and describe a numerical procedure for approximating the solution. Finally, a problem of finding the minimum spanning volume is considered.

## 1. Introduction

We consider a fluid trapped between two horizontal plates  $P_0, P_h$ , and further bounded by two parallel vertical planes  $\Pi_0, \Pi_d$ . Define the distance between  $P_0$  and  $P_h$  to be  $h$ , and that between  $\Pi_0$  and  $\Pi_d$  to be  $d$ . We orient a coordinate system  $(x, y, z)$  so that  $P_0$  is given by  $z \equiv 0$  and  $P_h$  is given by  $z \equiv h$ , while  $\Pi_0$  is given by  $y \equiv 0$  and  $\Pi_d$  is given by  $y \equiv d$ . We assume that the fluid is connected and any wetted portions of the plates are simply connected. The fluid then has a free interface  $\Lambda$  bounding a volume in the  $x$ -direction, and we denote the enclosed volume by  $\mathcal{V}$ . For an example, see [Figure 1](#), where we have not drawn  $\Pi_0$  or  $\Pi_d$ .

We consider dominant energies due to surface tension, wetting energy and gravitational potential energy. This gives the energy functional

$$\mathcal{E}[\Lambda] = \sigma \mathcal{A}[\Lambda] - \sigma \beta \mathcal{W}[\Lambda] + \int_{\mathcal{V}} \rho g z \, dz, \quad (1)$$

where  $\sigma$  is the (constant) surface tension,  $\beta$  is the wetting coefficient, taken to be constant on each plate,  $\rho$  is the uniform fluid density, and  $g$  is the gravitational constant. Further,  $\mathcal{A}$  is the area functional for the free-surface, and  $\mathcal{W}$  is the area functional for the wetted portions of  $P_0, P_h, \Pi_0$  and  $\Pi_d$ .

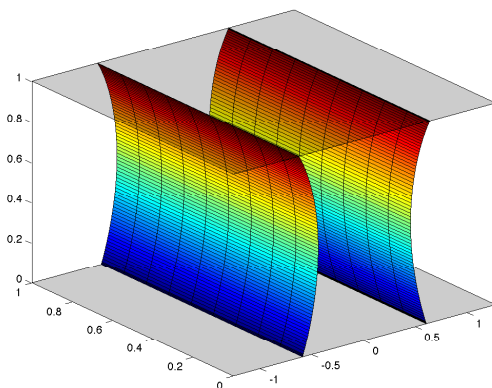
It is well known that the first variation for this functional implies

$$2H = \kappa u - \lambda, \quad (2)$$

---

*MSC2010:* primary 35Q35; secondary 76A02.

*Keywords:* capillarity, liquid bridges, numerical ODE.

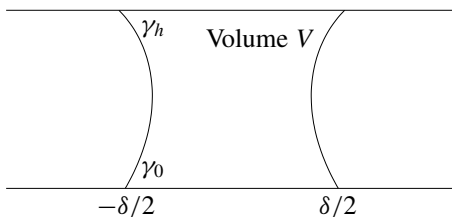


**Figure 1.** A cylindrical bridge.

where  $H$  is the mean curvature of  $\Lambda$ ,  $u$  is the height of the interface, the capillary constant is  $\kappa = \rho g / \sigma$ , and we have included a Lagrange multiplier  $\lambda$ . It may also be derived that  $\beta = \cos \gamma$  for a contact angle  $\gamma$  measured within the fluid. The standard reference is a manuscript by Finn [1986]. In what follows we do not assume that the interface is a graph over a base domain, though we do restrict our attention to the physical case where the interface is embedded. See Theorem 2.1 for details on how we interpret (2).

We make the assumption that  $\beta = 0$  on  $\Pi_0$  and  $\Pi_d$ . This implies a contact angle of  $\pi/2$  along the intersection of  $\Lambda$  with those planes. As we shall see in Section 2, this also implies that the free-surface is generated by curves in the plane  $\Pi_0$  and is extended as a right cylinder. See Figure 2 for an example of the generating curves, where  $\delta/2$  denotes the value of the horizontal displacement of the fluid interface on  $P_0$ . On the plates  $P_0$  and  $P_h$ , we allow the constant  $\beta$  to differ at heights 0 and  $h$  and to be any number in  $[-1, 1]$ . This corresponds to contact angles along the intersection of  $\Lambda$  with those plates, which we will denote by  $\gamma_0$  and  $\gamma_h$  respectively.

In Section 3, we derive a formula for the enclosed volume in terms of the solution to a version of the differential equation (2) when the fluid remains connected. Then we give an algorithm for computing the interface  $\Lambda$  with a volume constraint in



**Figure 2.** A liquid bridge.

**Section 4.** We use this algorithm to give a numerical approximation of  $\Lambda$  for different parameters  $h$ ,  $\gamma_0$ ,  $\gamma_h$ , and volume  $\mathcal{V}$ . Next, in **Section 5**, we present a collection of examples. Finally, we explore the minimum spanning volume for  $(\gamma_0, \gamma_h) \in [0, \pi/2] \times [0, \pi/2]$  in **Section 6**.

As far as we have been able to determine, this is the first exploration of liquid bridges of this type. Three dimensional liquid bridge problems have been studied by Athanassenas [1992]; Concus, Finn and McCuan [Concus et al. 2001]; Finn and Vogel [1992]; and Vogel [1982; 1987; 1989; 2005; 2006; 2013]. In recent work there is a trend to study the lower dimensional versions of certain related fluid mechanics problems. We point to papers by Bhatnagar and Finn [2006], as well as by McCuan and Treinen [2013;  $\geq$  2015], and Wentz [2006] for examples of this approach. In particular, we mention a paper by McCuan [2013] as a model for the present approach.

## 2. Symmetries

There are two types of symmetries in the fluid configurations. The first is the cylindrical symmetry that allows us to restrict our attention to the generating curves in the  $\Pi_0$  plane. The second is a reflective symmetry about the plane  $x = 0$ .

An Alexandrov moving plane argument has been successful in establishing symmetry properties for similar fluid configurations. See Wentz [1980], Treinen [2012], and McCuan [2013]. The following is a direct consequence of first using those methods with a moving plane parallel to  $\Pi_0$ , then a second argument using those methods with a moving plane parallel to  $x = 0$  can be used to show symmetry about  $x = 0$ . The details are left to the interested reader.

**Theorem 2.1.** *The interface  $\Lambda$  is right-cylindrically symmetric with generating curves restricted to the plane  $\Pi_0$ . The generating curves satisfy*

$$\frac{dx}{ds} = \cos \psi, \quad (3)$$

$$\frac{du}{ds} = \sin \psi, \quad (4)$$

$$\frac{d\psi}{ds} = \kappa u - \lambda, \quad (5)$$

and it suffices to compute one generating curve where  $x \geq 0$ .

Note that then the distance  $d$  is not important to our consideration, and hence we can view our problem in this reduced dimensional setting, or as extending infinitely in a horizontal direction. With this perspective, we normalize so that  $d = 1$  so that we are considering volume per unit distance in the  $y$ -direction. The solution may be extended infinitely in both  $y$ -directions and be seen as an infinitely long liquid bridge between two horizontal plates generated by the curves in  $\Pi_0$ . The

solution may also interpreted as a lower dimensional problem, where the interfaces are reduced to curves in the plane  $\Pi_0$ , spanning a *volume* that is more properly seen as an area in  $\Pi_0$ . This last interpretation is the easiest way to visualize the results of our computations, and so is our default for figures, even while we continue to use the terminology of *volume* and *area*, and we use them in the sense of per unit distance  $d$ .

### 3. Computing the fluid volume

Consider solutions to (3)–(5) with the boundary conditions

$$\sin \psi(0) = \cos \gamma_0 \quad \text{at } s = 0, \text{ where } u(0) = 0, \tag{6}$$

$$\sin \psi(\ell) = \cos \gamma_h \quad \text{at } s = \ell, \text{ where } u(\ell) = h. \tag{7}$$

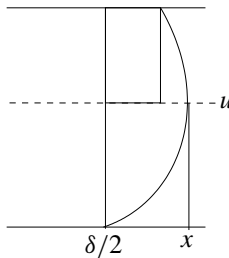
Solutions to this two-point boundary value problem will determine a value of  $x(0)$ , which we denote by  $\delta/2$ . We will later use this as a parameter in the process of constructing approximate solutions, but it is immediately useful in determining a volume formula as follows.

**Theorem 3.1.** *The volume enclosed by the upper plate, lower plate, and the fluid-air interface given by area per unit distance in the  $y$ -direction satisfies*

$$\mathcal{V} = (h - \lambda) \left( x(\ell) - \frac{\delta}{2} \right) + \sin \gamma_0 - \sin \gamma_h, \tag{8}$$

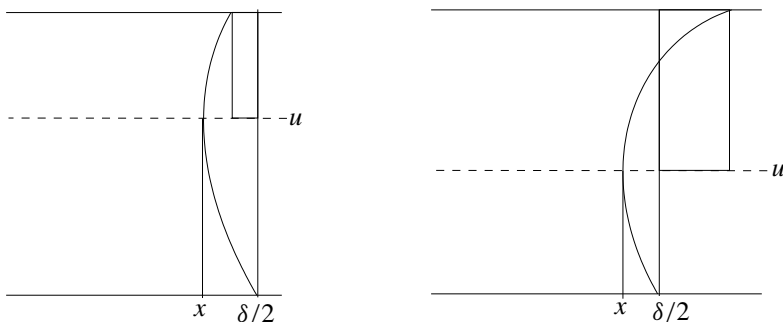
where the solutions  $x, u$ , and  $\psi$  are parametrized by arc length  $s$ , with  $s = 0$  at height  $u = 0$  and  $s = \ell$  at height  $u = h$ .

*Proof.* We find the volume of the enclosed fluid by computing the right half of the volume. The geometric idea is to start with a rectangle with height  $h$  and width  $\delta/2$ , and then add to it the additional volume outside this region. The first configuration is illustrated in Figure 3. This configuration contains a vertical point given by  $(\bar{x}, \bar{u})$ , and this partitions the volume outside of the rectangle into two regions. The lower



**Figure 3.** The configuration used in the volume computation, with only the portion  $x > 0$  shown. Here  $x = \bar{x}$  and  $u = \bar{u}$ , and the plate heights are 0 and  $h$ .





**Figure 4.** Two volume configurations. Here  $x = \bar{x}$  and  $u = \bar{u}$ , and plate heights are 0 and  $h$ .

region is bounded by  $\delta/2$  on the left,  $u = \bar{u}$  above, and the fluid interface on the right. The upper region is bounded by  $x = x(\ell)$  on the left,  $u = \bar{u}$  below, and the fluid interface on the right. Added to this upper region is a second, smaller, rectangle of height  $h - \bar{u}$  and width  $x(\ell) - \delta/2$ . So, we calculate using equations (3)–(5) and integration by parts as follows:

$$\mathcal{V} = \int_{\delta/2}^{\bar{x}} (\bar{u} - u) dx + \int_{x(\ell)}^{\bar{x}} (u - \bar{u}) dx + \left(x(\ell) - \frac{\delta}{2}\right)(h - \bar{u}) \tag{9}$$

$$= \bar{u}\left(\bar{x} - \frac{\delta}{2} + x(\ell) - \bar{x}\right) + \left(x(\ell) - \frac{\delta}{2}\right)(h - \bar{u}) + \int_{x(\ell)}^{\bar{x}} u dx - \int_{\delta/2}^{\bar{x}} u dx \tag{10}$$

$$= h\left(x(\ell) - \frac{\delta}{2}\right) + \bar{u}(0) + \int_{x(\ell)}^{\bar{x}} u dx - \int_{\delta/2}^{\bar{x}} u dx \tag{11}$$

$$= h\left(x(\ell) - \frac{\delta}{2}\right) + \int_{x(\ell)}^{\bar{x}} \left(\frac{d\psi}{ds} + \lambda\right) dx - \int_{\delta/2}^{\bar{x}} \left(\frac{d\psi}{ds} + \lambda\right) dx \tag{12}$$

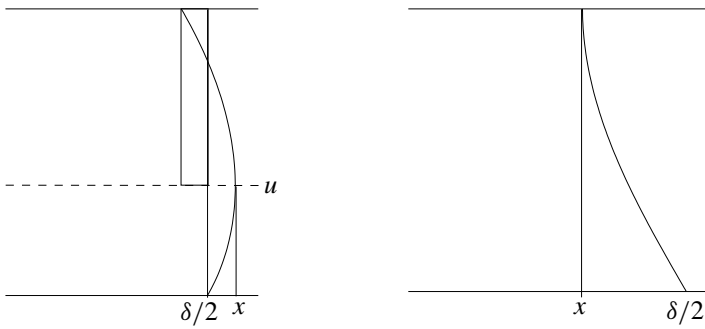
$$= (h - \lambda)\left(x(\ell) - \frac{\delta}{2}\right) + \int_{x(\ell)}^{\bar{x}} \frac{d\psi}{ds} dx - \int_{\delta/2}^{\bar{x}} \frac{d\psi}{ds} dx \tag{13}$$

$$= (h - \lambda)\left(x(\ell) - \frac{\delta}{2}\right) + \int_{\gamma_h - \pi}^{\pi/2} \cos \psi d\psi - \int_{-\gamma_0}^{\pi/2} \cos \psi d\psi \tag{14}$$

$$= (h - \lambda)\left(x(\ell) - \frac{\delta}{2}\right) + \sin \gamma_0 - \sin \gamma_h. \tag{15}$$

There are multiple possible configurations; however, it suffices to adapt the above calculation to these remaining cases:

- $x(s) < \delta/2$  for  $0 < s < \ell$  and  $x(\ell) < \delta/2$ . See Figure 4 (left).
- $x(s) < \delta/2$  for some initial  $s > 0$ , and then  $x(s)$  increases and  $x(\ell) > \delta/2$ . See Figure 4 (right).



**Figure 5.** Two remaining volume configurations. Here  $x = \bar{x}$  and  $u = \bar{u}$ , and plate heights are 0 and  $h$ .

- $x(s) > \delta/2$  for some initial  $s > 0$ , and then  $x(s)$  decreases and  $x(\ell) < \delta/2$ . See [Figure 5](#) (left).
- There is no vertical point on the interface profile curve. There are many such configurations; see [Figure 5](#) (right) for a typical example. The volume computation is straightforward in these cases, only requiring use of (3)–(5).  $\square$

#### 4. Numerical solver

We use a shooting method to solve the two-point boundary value problem of (3)–(5) with boundary conditions (6) and (7). We implement this by nesting two algorithms, namely an inner implementation of an adaptive Runge–Kutta–Fehlberg method and an outer implementation of a multidimensional root finder.

Values for the initial and terminal contact angles  $\gamma_0$ ,  $\gamma_h$ , volume  $\mathcal{V}$ , and height  $h$  are prescribed for the desired solution. The lower conditions for the boundary value problem are

$$r(0) = \frac{\delta}{2}, \quad (16)$$

$$u(0) = 0, \quad (17)$$

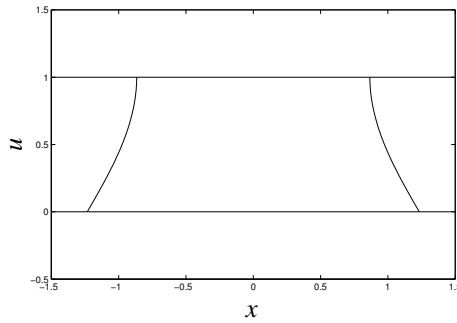
$$\psi(0) = \gamma_0, \quad (18)$$

where the tangent to the curve forms the contact angle  $\gamma_0$  with the lower plate, and the upper boundary conditions are

$$u(\ell) = h, \quad (19)$$

$$\psi(\ell) = -\gamma_h, \quad (20)$$

where the ending arc length  $\ell$  is chosen to terminate at height  $h$  with the tangent to the curve forming the angle  $\gamma_h$  with the upper parallel plate.



**Figure 6.** A liquid bridge with contact angle  $\pi/2$  on the upper plate.

Again, the boundary value problem is solved using a shooting method based on an adaptive ODE solver. The solver uses the adaptive Runge–Kutta–Fehlberg method for 4th and 5th order, implemented by Matlab as ODE45. The absolute and relative tolerances were both set to  $1e-8$ . To begin to solve the problem, reasonable guesses are given for the free parameters: the distance between the generating curves  $\delta$ , the ending arc length  $\ell$ , and the Lagrange multiplier  $\lambda$ . These values are used to generate candidates satisfying the ODE. Then the solutions to (3)–(5) with these values of the free parameters are used to evaluate the equations

$$\mathcal{V} - V(\ell) = 0, \quad (21)$$

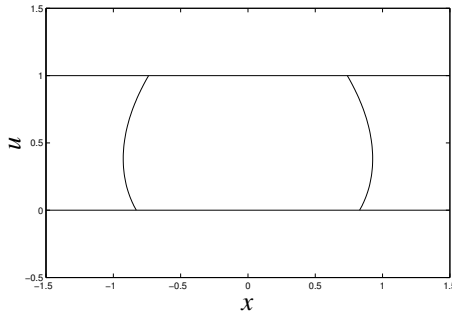
$$h - u(\ell) = 0, \quad (22)$$

$$\gamma_h - \psi(\ell) = 0, \quad (23)$$

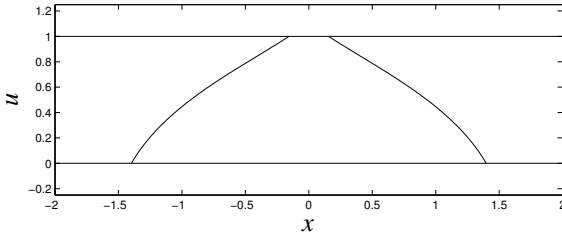
which are not, in general, solved. The parameters  $\delta$ ,  $\ell$ , and  $\lambda$  are adjusted in the multidimensional root finder implemented in Matlab as FSOLVE, which defaults to a trust region method. The tolerances for this portion of the algorithm were set to  $1e-6$ . We recompute the solutions to (3)–(5) with new values of the parameters  $\delta$ ,  $\ell$ , and  $\lambda$  at each step, until (21)–(23) are satisfied to the prescribed tolerance.

## 5. Examples

We present some examples of note generated with the algorithm described in the previous section. In Figure 2 we saw a typical example of a configuration where  $\gamma_0, \gamma_h \in [0, \frac{\pi}{2}]$ . Figure 6 shows a configuration where  $\gamma_h = \pi/2$ , and Figure 7 shows a configuration where both  $\gamma_0, \gamma_h > \pi/2$ . If the volume does not span the gap between  $P_0$  and  $P_h$ , then it will rest on the plate  $P_0$  as a sessile drop. We see in Figure 8 a configuration where  $(\gamma_0, \gamma_h) = (2.57, 1.05)$ , which appears to be close to the maximum height  $h$  before the liquid bridge pinches off of the upper plate  $P_h$  and becomes a sessile drop.



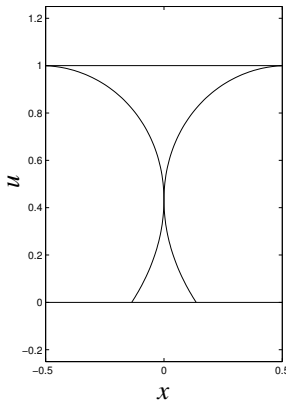
**Figure 7.** A liquid bridge with both  $\gamma_0$  and  $\gamma_h$  larger than  $\pi/2$ .



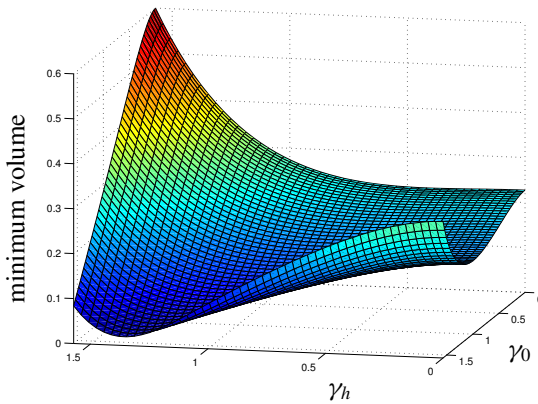
**Figure 8.** A liquid bridge that is visually similar to a sessile drop.

## 6. Minimum spanning volume

Consider configurations where the contact angles  $\gamma_0$  and  $\gamma_h$  are both less than  $\pi/2$ . The phenomenon explored is the minimum volume which admits a solution spanning the two plates  $P_0$  and  $P_h$ . In [Figure 9](#) we see that for angles  $(\gamma_0, \gamma_h) = (0.99, 0)$  and a particular volume, we have a point on the interior of the fluid interface on the right that touches a corresponding point on the interior of the fluid interface on the left.



**Figure 9.** A liquid bridge with interfaces touching on the interior.



**Figure 10.** The minimum spanning volume over a grid of  $50 \times 50$  samples in the  $(\gamma_0, \gamma_h)$ -space.

This is clearly nonphysical and represents an absolute minimum spanning volume. It is apparent that this contact between the left and right interfaces occurs on either  $P_0$  or  $P_h$  if either  $\gamma_0 > \pi/2$  or  $\gamma_h > \pi/2$ . Therefore, we restrict our attention to the region  $0 \leq \gamma_0 \leq \pi/2$  and  $0 \leq \gamma_h \leq \pi/2$ . We seek a minimum volume where  $x(s) = 0$  for some  $s \in [0, \ell]$ .

Observe the crucial fact of the system (3)–(5) that

$$\frac{d\psi}{ds} = \kappa u - \lambda$$

is independent of  $x$ , and so the  $x$  solution may be translated by a constant. We are able to use this to some degree to adjust the volume spanned. If the left and right interfaces are rigidly moved apart in the  $x$ -direction, then the spanned volume increases while still solving the boundary value problem, and conversely, if they are rigidly moved together, they will eventually touch. At this point there exists an arc length  $s$  such that  $x(s) = 0$  for both the left and right portions of the configuration. We are able to use this idea in conjunction with our previous solver to obtain the minimum spanning volume at a fixed height  $h$  for a given pair of contact angles  $(\gamma_0, \gamma_h)$ .

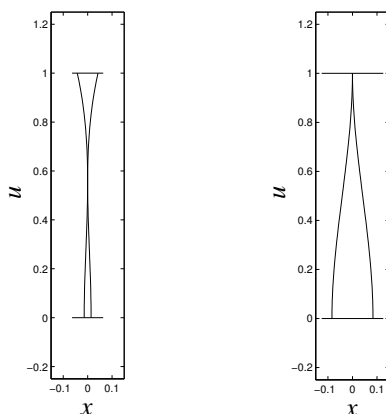
We use the following algorithm to run over a grid of  $50 \times 50$  samples in the  $(\gamma_0, \gamma_h)$ -space. We solve the constrained boundary value problem similar to the method in Section 4, however, we replace the condition

$$\mathcal{V} - V(\ell) = 0$$

with

$$x(s) = 0 \quad \text{for some } s \in [0, \ell]. \quad (24)$$

The results are collected in Figure 10. Here it is worth noting that the examples from Figure 11 are generated from interesting points on the minimum spanning volume



**Figure 11.** Left: The minimum spanning volume. Here  $(\gamma_0, \gamma_h) \approx (\frac{\pi}{2}, 1.35)$ . Right: A very small spanning volume, but not the minimum spanning volume. Here  $(\gamma_0, \gamma_h) = (\frac{\pi}{2}, \frac{\pi}{2})$ .

surface. The minimum spanning volume on the left is actually the minimum volume of all the contact angle pairs, and perhaps surprisingly, it is not the  $(\pi/2, \pi/2)$  case (which is pictured on the right).

## References

- [Athanassenas 1992] M. Athanassenas, “A free boundary problem for capillary surfaces”, *Manuscripta Math.* **76**:1 (1992), 5–19. [MR 93g:35144](#) [Zbl 0768.49025](#)
- [Bhatnagar and Finn 2006] R. Bhatnagar and R. Finn, “Equilibrium configurations of an infinite cylinder in an unbounded fluid”, *Phys. Fluids* **18**:4 (2006), 047103. [MR 2007f:76032](#) [Zbl 1185.76469](#)
- [Concus et al. 2001] P. Concus, R. Finn, and J. McCuan, “Liquid bridges, edge blobs, and Scherk-type capillary surfaces”, *Indiana Univ. Math. J.* **50**:1 (2001), 411–441. [MR 2002g:76023](#) [Zbl 0996.76014](#)
- [Finn 1986] R. Finn, *Equilibrium capillary surfaces*, Grundlehren der Mathematischen Wissenschaften **284**, Springer, New York, 1986. [MR 88f:49001](#) [Zbl 0583.35002](#)
- [Finn and Vogel 1992] R. Finn and T. I. Vogel, “On the volume infimum for liquid bridges”, *Z. Anal. Anwendungen* **11**:1 (1992), 3–23. [MR 95d:76021](#) [Zbl 0760.76015](#)
- [McCuan 2013] J. McCuan, “Extremities of stability for pendant drops”, pp. 157–173 in *Geometric analysis, mathematical relativity, and nonlinear partial differential equations*, edited by M. Ghomi et al., Contemp. Math. **599**, Amer. Math. Soc., Providence, RI, 2013. [MR 3202478](#) [Zbl 1276.00026](#)
- [McCuan and Treinen 2013] J. McCuan and R. Treinen, “Capillarity and Archimedes’ principle of flotation”, *Pacific J. Math.* **265**:1 (2013), 123–150. [MR 3095116](#) [Zbl 06218270](#)
- [McCuan and Treinen  $\geq$  2015] J. McCuan and R. Treinen, “On floating equilibria in a finite container”. To appear.
- [Treinen 2012] R. Treinen, “On the symmetry of solutions to some floating drop problems”, *SIAM J. Math. Anal.* **44**:6 (2012), 3834–3847. [MR 3023432](#) [Zbl 06138483](#)
- [Vogel 1982] T. I. Vogel, “Symmetric unbounded liquid bridges”, *Pacific J. Math.* **103**:1 (1982), 205–241. [MR 84f:53007](#) [Zbl 0504.76025](#)

- [Vogel 1987] T. I. Vogel, “Stability of a liquid drop trapped between two parallel planes”, *SIAM J. Appl. Math.* **47**:3 (1987), 516–525. MR 88e:53010 Zbl 0627.53004
- [Vogel 1989] T. I. Vogel, “Stability of a liquid drop trapped between two parallel planes, II: General contact angles”, *SIAM J. Appl. Math.* **49**:4 (1989), 1009–1028. MR 90k:53013 Zbl 0691.53007
- [Vogel 2005] T. I. Vogel, “Comments on radially symmetric liquid bridges with inflected profiles”, *Discrete Contin. Dyn. Syst.* suppl. (2005), 862–867. MR 2006h:76048 Zbl 1158.53307
- [Vogel 2006] T. I. Vogel, “Convex, rotationally symmetric liquid bridges between spheres”, *Pacific J. Math.* **224**:2 (2006), 367–377. MR 2007f:76033 Zbl 1118.53006
- [Vogel 2013] T. I. Vogel, “Liquid bridges between balls: The small volume instability”, *J. Math. Fluid Mech.* **15**:2 (2013), 397–413. MR 3061769 Zbl 1267.76016
- [Wente 1980] H. C. Wente, “The symmetry of sessile and pendent drops”, *Pacific J. Math.* **88**:2 (1980), 387–397. MR 83j:49042a Zbl 0473.76086
- [Wente 2006] H. C. Wente, “New exotic containers”, *Pacific J. Math.* **224**:2 (2006), 379–398. MR 2007f:76034 Zbl 1118.53007

Received: 2014-07-19

Accepted: 2014-07-28

[icolter08@gmail.com](mailto:icolter08@gmail.com)

*Department of Mathematics, Texas State University,  
601 University Drive San Marcos, TX 78666, United States*

[rt30@txstate.edu](mailto:rt30@txstate.edu)

*Department of Mathematics, Texas State University,  
601 University Drive, San Marcos, TX 78666, United States*





# Some projective distance inequalities for simplices in complex projective space

Mark Fincher, Heather Olney and William Cherry

(Communicated by Michael Dorff)

We prove inequalities relating the absolute value of the determinant of  $n + 1$  linearly independent unit vectors in  $\mathbb{C}^{n+1}$  and the projective distances from the vertices to the hyperplanes containing the opposite faces of the simplices in complex projective  $n$ -space whose vertices or faces are determined by the given vectors.

A basis of unit vectors in  $\mathbb{C}^{n+1}$  determines the vertices (or the faces) of a simplex in  $n$ -dimensional complex projective space. For reasons originally motivated by an inequality in complex function theory proven by Cherry and Eremenko [2011], we investigated the relationship between the determinant of the vectors forming the basis and the projective distances from each vertex of the simplex to the hyperplane containing the face of the opposite side. We show that if  $d_{\min}$  denotes the minimum of these projective distances and if  $D$  denotes the determinant of the basis vectors, then  $d_{\min}^n \leq |D| \leq d_{\min}$ .

Let  $\mathbf{e}_0, \dots, \mathbf{e}_n$  be a basis for  $\mathbb{C}^{n+1}$ . Given two vectors  $\mathbf{a} = a_0\mathbf{e}_0 + \dots + a_n\mathbf{e}_n$  and  $\mathbf{b} = b_0\mathbf{e}_0 + \dots + b_n\mathbf{e}_n$  in  $\mathbb{C}^{n+1}$ , we use  $\mathbf{a} \cdot \mathbf{b}$  to denote the standard dot product,

$$\mathbf{a} \cdot \mathbf{b} = a_0b_0 + \dots + a_nb_n,$$

rather than the Hermitian inner product more typically used with complex vector spaces. Thus, in our notation,

$$|\mathbf{a}|^2 = \mathbf{a} \cdot \bar{\mathbf{a}},$$

where the bar denotes complex conjugation, as usual.

For  $k = 1, \dots, n + 1$ , we let  $\Lambda^k \mathbb{C}^{n+1}$  denote the  $k$ -th exterior power of the vector space  $\mathbb{C}^{n+1}$ , and we recall that

$$\mathbf{e}_0 \wedge \mathbf{e}_1 \wedge \dots \wedge \mathbf{e}_{k-1}, \quad \dots, \quad \mathbf{e}_{i_1} \wedge \mathbf{e}_{i_2} \wedge \dots \wedge \mathbf{e}_{i_k}, \quad \dots, \quad \mathbf{e}_{n+1-k} \wedge \mathbf{e}_{n+2-k} \wedge \dots \wedge \mathbf{e}_n,$$

*MSC2010:* primary 51N15; secondary 32Q45.

*Keywords:* projective height, projective simplex, determinant.

Financial support provided to Fincher and Olney in the form of a UNT SUMS fellowship as part of the UNT mathematics department's NSF funded RTG grant DMS-0943870.

where  $0 \leq i_1 < i_2 < \cdots < i_k \leq n$  form a basis for  $\Lambda^k \mathbb{C}^{n+1}$ . By declaring this basis to be orthonormal in  $\Lambda^k \mathbb{C}^{n+1}$ , the norm and dot product on  $\mathbb{C}^{n+1}$  extend to a norm and inner product on  $\Lambda^k \mathbb{C}^{n+1}$ . For a detailed introduction to exterior algebras and wedge products, see [Bowen and Wang 1976].

**Proposition 1.** *Let  $1 \leq k \leq n+1$  be an integer, and let  $\mathbf{v}_1, \dots, \mathbf{v}_k$  and  $\mathbf{w}_1, \dots, \mathbf{w}_k$  be vectors in  $\mathbb{C}^{n+1}$ . Then,*

$$(\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_k) \cdot (\mathbf{w}_1 \wedge \cdots \wedge \mathbf{w}_k) = \det(\mathbf{v}_i \cdot \mathbf{w}_j)_{1 \leq i, j \leq k}.$$

**Remark.** The matrix of dot products on the right is called a *Gramian* matrix.

*Proof.* This is Exercise 39.3 in [Bowen and Wang 1976]. □

**Corollary 2.** *Let  $\mathbf{v}_1, \dots, \mathbf{v}_k$  be  $k$  vectors in  $\mathbb{C}^{n+1}$ . Then,*

$$|\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_k|^2 = \det(\mathbf{v}_i \cdot \bar{\mathbf{v}}_j)_{1 \leq i, j \leq k}.$$

**Corollary 3.** *Let  $\mathbf{v}_1, \dots, \mathbf{v}_k$  be  $k$  vectors in  $\mathbb{C}^{n+1}$ . Then,*

$$|\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_k| \leq |\mathbf{v}_1| \cdots |\mathbf{v}_k|.$$

*Equality holds if and only if one of the vectors is the zero vector or if  $\mathbf{v}_i \cdot \bar{\mathbf{v}}_j = 0$  for all  $i \neq j$ .*

*Proof.* If any of the vectors  $\mathbf{v}_j$  are the zero vector, then the inequality is obvious. So, assume that none of the  $\mathbf{v}_j$  are zero. Let

$$\mathbf{u}_j = \frac{\mathbf{v}_j}{|\mathbf{v}_j|}$$

be unit vectors in the directions of the  $\mathbf{v}_j$ . Then, clearly,

$$|\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_k| = \left| |\mathbf{v}_1| \mathbf{u}_1 \wedge \cdots \wedge |\mathbf{v}_k| \mathbf{u}_k \right| = |\mathbf{v}_1| \cdots |\mathbf{v}_k| |\mathbf{u}_1 \wedge \cdots \wedge \mathbf{u}_k|.$$

Thus, it suffices to show that  $|\mathbf{u}_1 \wedge \cdots \wedge \mathbf{u}_k| \leq 1$ . To this end, by Corollary 2,

$$|\mathbf{u}_1 \wedge \cdots \wedge \mathbf{u}_k|^2 = \det(\mathbf{u}_i \cdot \bar{\mathbf{u}}_j). \tag{1}$$

The matrix  $(\mathbf{u}_i \cdot \bar{\mathbf{u}}_j)$  is a  $k \times k$  Hermitian matrix with nonnegative eigenvalues  $\lambda_1, \dots, \lambda_k$ . Thus, by the geometric-arithmetic mean inequality,

$$\det(\mathbf{u}_i \cdot \bar{\mathbf{u}}_j) = \lambda_1 \cdots \lambda_k \leq \left( \frac{\lambda_1 + \cdots + \lambda_k}{k} \right)^k = 1,$$

where the equality on the right follows from the fact that

$$\lambda_1 + \cdots + \lambda_k = \text{Trace}(\mathbf{u}_i \cdot \bar{\mathbf{u}}_j) = k,$$

since  $\mathbf{u}_i \cdot \bar{\mathbf{u}}_i = 1$ .

Equality holds in the arithmetic-geometric mean inequality if and only if all the eigenvalues are equal, and hence all equal to one. This is the case if and only if  $(\mathbf{u}_i \cdot \bar{\mathbf{u}}_j)$  is the  $k \times k$  identity matrix, which happens if and only if  $\mathbf{v}_i \cdot \bar{\mathbf{v}}_j = 0$  for all  $i \neq j$ .  $\square$

We will be most interested in the  $n$ -th exterior power of  $\mathbb{C}^{n+1}$ , where

$$\mathbf{e}_1 \wedge \cdots \wedge \mathbf{e}_n, \quad \dots, \quad \mathbf{e}_0 \wedge \cdots \wedge \mathbf{e}_{j-1} \wedge \mathbf{e}_{j+1} \wedge \cdots \wedge \mathbf{e}_n, \quad \dots, \quad \mathbf{e}_0 \wedge \cdots \wedge \mathbf{e}_{n-1}$$

form a basis of  $\Lambda^n \mathbb{C}^{n+1}$ . Let  $L$  denote the isometric isomorphism from  $\Lambda^n \mathbb{C}^{n+1}$  to  $\mathbb{C}^{n+1}$  defined on the basis vectors as follows:

$$\begin{aligned} L(\mathbf{e}_1 \wedge \cdots \wedge \mathbf{e}_n) &= \mathbf{e}_0, \\ &\vdots \\ L(\mathbf{e}_0 \wedge \cdots \wedge \mathbf{e}_{j-1} \wedge \mathbf{e}_{j+1} \wedge \cdots \wedge \mathbf{e}_n) &= (-1)^j \mathbf{e}_j, \\ &\vdots \\ L(\mathbf{e}_0 \wedge \cdots \wedge \mathbf{e}_{n-1}) &= (-1)^n \mathbf{e}_n. \end{aligned}$$

Observe that if  $n = 2$  and  $\mathbf{a}$  and  $\mathbf{b}$  are vectors in  $\mathbb{C}^3$ , then  $L(\mathbf{a} \wedge \mathbf{b}) = \mathbf{a} \times \mathbf{b}$ , where the product on the right is the ordinary cross product in  $\mathbb{C}^3$ .

We will use  $L(\mathbf{b}_1 \wedge \cdots \wedge \mathbf{b}_n)$  as a generalized cross product.

**Proposition 4.** *Let  $\mathbf{a}, \mathbf{b}_1, \dots, \mathbf{b}_n$  be  $n + 1$  vectors in  $\mathbb{C}^{n+1}$ . Then,*

$$\det(\mathbf{a}, \mathbf{b}_1, \dots, \mathbf{b}_n) = \mathbf{a} \cdot L(\mathbf{b}_1 \wedge \cdots \wedge \mathbf{b}_n).$$

*Proof.* If we compute the determinant of the  $(n + 1) \times (n + 1)$  matrix whose rows are  $\mathbf{a}, \mathbf{b}_1, \dots, \mathbf{b}_n$ , then the expression on the right is nothing other than the computation of the determinant by expansion of minors along the first row.  $\square$

**Corollary 5.** *The vector  $L(\mathbf{b}_1 \wedge \cdots \wedge \mathbf{b}_n)$  is orthogonal to each of the  $\mathbf{b}_j$ .*

We define an equivalence relation on  $\mathbb{C}^{n+1} \setminus \{0\}$  by declaring that two nonzero vectors  $\mathbf{v}$  and  $\mathbf{w}$  in  $\mathbb{C}^{n+1}$  are equivalent if there exists a nonzero complex scalar  $c$  such that  $\mathbf{v} = c\mathbf{w}$ . The set of all such equivalence classes is denoted by  $\mathbb{C}\mathbb{P}^n$  and is called the *complex projective space* of dimension  $n$ . A point in  $\mathbb{C}\mathbb{P}^n$  is an equivalence class of vectors in  $\mathbb{C}^{n+1}$  and by the definition of the equivalence relation, we can always represent a point in  $\mathbb{C}\mathbb{P}^n$  by a unit vector in  $\mathbb{C}^{n+1}$ . The set of equivalence classes associated with the vectors in a  $k + 1$  dimensional subspace of  $\mathbb{C}^{n+1}$  is a  $k$ -dimensional subspace of  $\mathbb{C}\mathbb{P}^n$ . When  $k = n - 1$ , such a subspace is called a hyperplane in  $\mathbb{C}\mathbb{P}^n$ . We say that  $n + 1$  points in  $\mathbb{C}\mathbb{P}^n$  are in *general position* if they are not all contained in any one hyperplane. This is equivalent to the vectors representing the points being linearly independent in  $\mathbb{C}^{n+1}$ . Similarly, we say that  $n + 1$  hyperplanes in  $\mathbb{C}\mathbb{P}^n$  are in *general position* if there is no point in

$\mathbb{C}\mathbb{P}^n$  contained in all the hyperplanes. Note that a nonzero vector  $\mathbf{v}$  in  $\mathbb{C}^{n+1}$  can be thought of as representing a hyperplane where the points in the hyperplane are represented by the vectors  $\mathbf{x}$  in  $\mathbb{C}^{n+1}$  such that  $\mathbf{v} \cdot \mathbf{x} = 0$ .

If  $\mathbf{v}$  and  $\mathbf{w}$  are two unit vectors in  $\mathbb{C}^{n+1}$  representing points in  $\mathbb{C}\mathbb{P}^n$ , then the *Fubini–Study distance* between the two points is defined to be  $|\mathbf{v} \wedge \mathbf{w}|$ . Now let  $\mathbf{u}$  and  $\mathbf{v}$  be unit vectors in  $\mathbb{C}^{n+1}$ . We think of  $\mathbf{u}$  as representing a point in  $\mathbb{C}\mathbb{P}^n$  and  $\mathbf{v}$  as representing a hyperplane in  $\mathbb{C}\mathbb{P}^n$ . Then, the Fubini–Study distance from the point represented by  $\mathbf{u}$  to the hyperplane represented by  $\mathbf{v}$  is defined by

$$\begin{aligned} & \text{distance from the point } \mathbf{u} \text{ to the hyperplane } \mathbf{v} \\ &= \min\{\text{distance from } \mathbf{u} \text{ to } \mathbf{x} : \mathbf{v} \cdot \mathbf{x} = 0 \text{ and } |\mathbf{x}| = 1\} \\ &= \min\{|\mathbf{u} \wedge \mathbf{x}| : \mathbf{v} \cdot \mathbf{x} = 0 \text{ and } |\mathbf{x}| = 1\}. \end{aligned}$$

Second perhaps only to hyperbolic geometry, projective geometry, which arose out of the study of perspective in classical painting, is among the most ubiquitous of the non-Euclidean geometries encountered in modern mathematics. See, for instance, [Richter-Gebert 2011] for a recent accessible introduction.

Our first result is a convenient formula for the distance from a vertex of a projective simplex to the hyperplane determined by the opposite face in the simplex.

**Proposition 6.** *Let  $\mathbf{a}, \mathbf{b}_1, \dots, \mathbf{b}_n$  be  $n + 1$  linearly independent unit vectors in  $\mathbb{C}^{n+1}$  representing  $n + 1$  points in general position in  $\mathbb{C}\mathbb{P}^n$ . Then, the Fubini–Study distance  $d$  from the point  $\mathbf{a}$  to the hyperplane in  $\mathbb{C}\mathbb{P}^n$  spanned by  $\mathbf{b}_1, \dots, \mathbf{b}_n$  is given by*

$$d = \frac{|\det(\mathbf{a}, \mathbf{b}_1, \dots, \mathbf{b}_n)|}{|\mathbf{b}_1 \wedge \dots \wedge \mathbf{b}_n|}.$$

*Proof.* Without loss of generality, by making an orthogonal change of coordinates, we may choose our standard basis vectors  $\mathbf{e}_0, \dots, \mathbf{e}_n$  in  $\mathbb{C}^{n+1}$  so that  $\mathbf{e}_0 \cdot \mathbf{b}_j = 0$  for  $j = 1, \dots, n$ . Let  $\mathbf{u}$  be a unit vector in the span of  $\{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ . Then,

$$\mathbf{u} = u_1 \mathbf{e}_1 + \dots + u_n \mathbf{e}_n, \quad \text{with } |u_1|^2 + \dots + |u_n|^2 = 1.$$

Let  $\mathbf{a} = a_0 \mathbf{e}_0 + \dots + a_n \mathbf{e}_n$ . Then, the Fubini–Study distance from the point in  $\mathbb{C}\mathbb{P}^n$  represented by  $\mathbf{a}$  to the point in  $\mathbb{C}\mathbb{P}^n$  represented by  $\mathbf{u}$  is given by  $|\mathbf{a} \wedge \mathbf{u}|$ . Note that

$$\mathbf{a} \wedge \mathbf{u} = a_0 u_1 \mathbf{e}_0 \wedge \mathbf{e}_1 + \dots + a_0 u_n \mathbf{e}_0 \wedge \mathbf{e}_n + \sum_{1 \leq i < j \leq n} (a_i u_j - a_j u_i) \mathbf{e}_i \wedge \mathbf{e}_j. \quad (2)$$

Hence,

$$|\mathbf{a} \wedge \mathbf{u}|^2 \geq |a_0 u_1|^2 + \dots + |a_0 u_n|^2 = |a_0|^2 (|u_1|^2 + \dots + |u_n|^2) = |a_0|^2. \quad (3)$$

Now,

$$\det(\mathbf{a}, \mathbf{b}_1, \dots, \mathbf{b}_n) = \mathbf{a} \cdot L(\mathbf{b}_1 \wedge \dots \wedge \mathbf{b}_n)$$

by [Proposition 4](#). Of course,  $L(\mathbf{b}_1 \wedge \cdots \wedge \mathbf{b}_n)$  is orthogonal to each of the  $\mathbf{b}_j$ . By our choice of basis,  $\mathbf{e}_0$  is also orthogonal to each of the  $\mathbf{b}_j$ . Since the  $\mathbf{b}_j$  form a set of  $n$  linearly independent vectors in an  $(n+1)$ -dimensional vector space, there is only one direction simultaneously orthogonal to all of the  $\mathbf{b}_j$ . Thus,  $L(\mathbf{b}_1 \wedge \cdots \wedge \mathbf{b}_n)$  is in the span of  $\mathbf{e}_0$ , and so

$$|\mathbf{a} \cdot L(\mathbf{b}_1 \wedge \cdots \wedge \mathbf{b}_n)| = |a_0| \cdot |L(\mathbf{b}_1 \wedge \cdots \wedge \mathbf{b}_n)|.$$

Thus, observing that

$$|L(\mathbf{b}_1 \wedge \cdots \wedge \mathbf{b}_n)| = |\mathbf{b}_1 \wedge \cdots \wedge \mathbf{b}_n|,$$

we see from [\(3\)](#) that

$$\begin{aligned} |\mathbf{a} \wedge \mathbf{u}| \geq |a_0| &= \frac{|a_0| \cdot |L(\mathbf{b}_1 \wedge \cdots \wedge \mathbf{b}_n)|}{|\mathbf{b}_1 \wedge \cdots \wedge \mathbf{b}_n|} \\ &= \frac{|\mathbf{a} \cdot L(\mathbf{b}_1 \wedge \cdots \wedge \mathbf{b}_n)|}{|\mathbf{b}_1 \wedge \cdots \wedge \mathbf{b}_n|} \\ &= \frac{|\det(\mathbf{a}, \mathbf{b}_1, \dots, \mathbf{b}_n)|}{|\mathbf{b}_1 \wedge \cdots \wedge \mathbf{b}_n|}. \end{aligned}$$

To complete the proof, we need to show that equality is obtained for some choice of  $\mathbf{u}$ . There are two cases. If  $\mathbf{a}$  is the direction of  $\mathbf{e}_0$ , then equality holds for any choice of  $\mathbf{u}$  since  $a_1 = \cdots = a_n = 0$ . Otherwise, if we choose

$$u_j = \frac{a_j}{\sqrt{|a_1|^2 + \cdots + |a_n|^2}} \quad \text{for } j = 1, \dots, n,$$

we see that the terms in the sum on the far right of [\(2\)](#) are all zero, and so equality holds in [\(3\)](#).  $\square$

**Corollary 7.** *Let  $\mathbf{a}, \mathbf{b}_1, \dots, \mathbf{b}_n$  and  $d$  be as in [Proposition 6](#). Then,*

$$d \geq |\det(\mathbf{a}, \mathbf{b}_1, \dots, \mathbf{b}_n)|.$$

*Equality holds if and only if  $\mathbf{b}_i \cdot \bar{\mathbf{b}}_j = 0$  for all  $i \neq j$ .*

**Example 8.** When  $n = 3$ , let  $0 < s \leq 1$  and consider the projective triangle with vertices represented by the unit vectors

$$\mathbf{a} = \left( \sqrt{\frac{1-s^2}{2}}, \sqrt{\frac{1-s^2}{2}}, s \right), \quad \mathbf{b}_1 = (1, 0, 0), \quad \text{and} \quad \mathbf{b}_2 = (0, 1, 0).$$

Then,  $|\mathbf{b}_1 \wedge \mathbf{b}_2| = 1$ , and so  $d = \det(\mathbf{a}, \mathbf{b}_1, \mathbf{b}_2) = s$ , and equality holds in [Corollary 7](#). We remark that geometrically, these triangles are isosceles with projective side lengths

$$1, \quad \sqrt{\frac{1+s^2}{2}}, \quad \sqrt{\frac{1+s^2}{2}}.$$

*Proof of Corollary 7.* By Corollary 3, we have

$$|\mathbf{b}_1 \wedge \cdots \wedge \mathbf{b}_n| \leq 1.$$

Hence, by the formula for  $d$  in Proposition 6,

$$d = \frac{|\det(\mathbf{a}, \mathbf{b}_1, \dots, \mathbf{b}_n)|}{|\mathbf{b}_1 \wedge \cdots \wedge \mathbf{b}_n|} \geq |\det(\mathbf{a}, \mathbf{b}_1, \dots, \mathbf{b}_n)|.$$

Equality holds if and only if equality holds in Corollary 3.  $\square$

**Proposition 9.** Let  $\mathbf{v}_1, \dots, \mathbf{v}_{n-1}$  be  $n-1$  linearly independent vectors in  $\mathbb{C}^{n+1}$  and let  $\mathbf{w}_1, \dots, \mathbf{w}_n$  be  $n$  linearly independent vectors in  $\mathbb{C}^{n+1}$ . If we let

$$\mathbf{a} = L(\mathbf{w}_1 \wedge \cdots \wedge \mathbf{w}_n) \quad \text{and} \quad \mathbf{b} = L(\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_{n-1} \wedge \mathbf{a}),$$

then

$$\mathbf{b} = (-1)^n \det \begin{pmatrix} \mathbf{w}_1 & \cdots & \mathbf{w}_n \\ \mathbf{v}_1 \cdot \mathbf{w}_1 & \cdots & \mathbf{v}_1 \cdot \mathbf{w}_n \\ \vdots & \vdots & \vdots \\ \mathbf{v}_{n-1} \cdot \mathbf{w}_1 & \cdots & \mathbf{v}_{n-1} \cdot \mathbf{w}_n \end{pmatrix}.$$

**Remark.** Note that the matrix specified in the proposition has vector entries in its first row, and hence its determinant results in a vector. This proposition is a generalization of Lagrange's formula for the vector triple product in  $\mathbb{R}^3$ . The proof of this proposition was inspired by a discussion Cherry had with Charles Conley, and we thank him for his interest. We suspect that Proposition 9 is reasonably well-known, but we were unable to find a reference to it in the literature.

*Proof.* Let

$$\tilde{\mathbf{b}} = \det \begin{pmatrix} \mathbf{w}_1 & \cdots & \mathbf{w}_n \\ \mathbf{v}_1 \cdot \mathbf{w}_1 & \cdots & \mathbf{v}_1 \cdot \mathbf{w}_n \\ \vdots & \vdots & \vdots \\ \mathbf{v}_{n-1} \cdot \mathbf{w}_1 & \cdots & \mathbf{v}_{n-1} \cdot \mathbf{w}_n \end{pmatrix}.$$

We want to show that  $\mathbf{b} = (-1)^n \tilde{\mathbf{b}}$ , and for this, it suffices to show that for all  $\mathbf{z}$  in  $\mathbb{C}^{n+1}$ , we have  $\mathbf{z} \cdot \mathbf{b} = (-1)^n \mathbf{z} \cdot \tilde{\mathbf{b}}$ . Clearly,

$$\mathbf{z} \cdot \tilde{\mathbf{b}} = \det \begin{pmatrix} \mathbf{z} \cdot \mathbf{w}_1 & \cdots & \mathbf{z} \cdot \mathbf{w}_n \\ \mathbf{v}_1 \cdot \mathbf{w}_1 & \cdots & \mathbf{v}_1 \cdot \mathbf{w}_n \\ \vdots & \vdots & \vdots \\ \mathbf{v}_{n-1} \cdot \mathbf{w}_1 & \cdots & \mathbf{v}_{n-1} \cdot \mathbf{w}_n \end{pmatrix}.$$

On the other hand, by [Proposition 4](#),

$$\begin{aligned}
 z \cdot \mathbf{b} &= \det(z, \mathbf{v}_1, \dots, \mathbf{v}_{n-1}, \mathbf{a}) \\
 &= (-1)^n \det(\mathbf{a}, z, \mathbf{v}_1, \dots, \mathbf{v}_{n-1}) \\
 &= (-1)^n \mathbf{a} \cdot L(z \wedge \mathbf{v}_1 \wedge \dots \wedge \mathbf{v}_{n-1}) \\
 &= (-1)^n L(\mathbf{w}_1 \wedge \dots \wedge \mathbf{w}_n) \cdot L(z \wedge \mathbf{v}_1 \wedge \dots \wedge \mathbf{v}_{n-1}) \\
 &= (-1)^n (\mathbf{w}_1 \wedge \dots \wedge \mathbf{w}_n) \cdot (z \wedge \mathbf{v}_1 \wedge \dots \wedge \mathbf{v}_{n-1}) \quad (\text{since } L \text{ is an isometry}) \\
 &= (-1)^n (z \wedge \mathbf{v}_1 \wedge \dots \wedge \mathbf{v}_{n-1}) \cdot (\mathbf{w}_1 \wedge \dots \wedge \mathbf{w}_n) \\
 &= (-1)^n \det \begin{pmatrix} z \cdot \mathbf{w}_1 & \dots & z \cdot \mathbf{w}_n \\ \mathbf{v}_1 \cdot \mathbf{w}_1 & \dots & \mathbf{v}_1 \cdot \mathbf{w}_n \\ \vdots & \vdots & \vdots \\ \mathbf{v}_{n-1} \cdot \mathbf{w}_1 & \dots & \mathbf{v}_{n-1} \cdot \mathbf{w}_n \end{pmatrix} \quad (\text{by Proposition 1}). \quad \square
 \end{aligned}$$

**Proposition 10.** Let  $\mathbf{a}, \mathbf{u}_1, \dots, \mathbf{u}_n$  be  $n+1$  linearly independent vectors in  $\mathbb{C}^{n+1}$ . For  $j = 1, \dots, n$ , let

$$\mathbf{v}_j = L(\mathbf{a} \wedge \mathbf{u}_1 \wedge \dots \wedge \mathbf{u}_{j-1} \wedge \mathbf{u}_{j+1} \wedge \dots \wedge \mathbf{u}_n).$$

Then,  $L(\mathbf{v}_1 \wedge \dots \wedge \mathbf{v}_n) = \pm D^{n-1} \mathbf{a}$ , where  $D = \det(\mathbf{a}, \mathbf{u}_1, \dots, \mathbf{u}_n)$ .

**Remark.** The unspecified sign depends only on  $n$  and can be explicitly determined from the proof. Since the sign will not matter for our purpose, we did not bother to record it here.

*Proof.* By [Proposition 9](#), we get that

$$L(\mathbf{v}_1 \wedge \dots \wedge \mathbf{v}_n) = (-1)^n \det \begin{pmatrix} \mathbf{a} & \mathbf{u}_1 & \dots & \mathbf{u}_{n-1} \\ \mathbf{v}_1 \cdot \mathbf{a} & \mathbf{v}_1 \cdot \mathbf{u}_1 & \dots & \mathbf{v}_1 \cdot \mathbf{u}_{n-1} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{v}_{n-1} \cdot \mathbf{a} & \mathbf{v}_{n-1} \cdot \mathbf{u}_1 & \dots & \mathbf{v}_{n-1} \cdot \mathbf{u}_{n-1} \end{pmatrix}.$$

If  $i \neq j$ , then

$$\mathbf{v}_i \cdot \mathbf{u}_j = L(\mathbf{a} \wedge \dots \wedge \mathbf{u}_{i-1} \wedge \mathbf{u}_{i+1} \wedge \dots \wedge \mathbf{u}_n) \cdot \mathbf{u}_j = 0$$

since  $\mathbf{u}_j$  appears in the wedge product defining  $\mathbf{v}_i$ , and hence  $\mathbf{v}_i$  is orthogonal to  $\mathbf{u}_j$ . Similarly,  $\mathbf{v}_i \cdot \mathbf{a} = 0$ . Moreover,

$$\mathbf{v}_j \cdot \mathbf{u}_j = L(\mathbf{a} \wedge \dots \wedge \mathbf{u}_{j-1} \wedge \mathbf{u}_{j+1} \wedge \dots \wedge \mathbf{u}_n) \cdot \mathbf{u}_j = (-1)^j D$$

by Proposition 4. Hence,

$$L(\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_n) = (-1)^n \det \begin{pmatrix} \mathbf{a} & \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_{n-1} \\ 0 & -D & 0 & \cdots & 0 \\ 0 & 0 & D & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & (-1)^{n-1} D \end{pmatrix} = \pm D^{n-1} \mathbf{a}. \quad \square$$

**Theorem 11.** Let  $\mathbf{u}_0, \dots, \mathbf{u}_n$  be  $n + 1$  linearly independent unit vectors in  $\mathbb{C}^{n+1}$  representing  $n + 1$  points in general position in  $\mathbb{C}\mathbb{P}^n$ , which we think of as the vertices of a projective simplex. For each  $j$  from 0 to  $n$ , let  $d_j$  denote the Fubini–Study distance from the point represented by  $\mathbf{u}_j$  to the hyperplane containing the opposite face of the simplex. Let  $d_{\min}$  denote the minimum of the  $d_j$ . Then,

$$d_{\min}^n \leq |\det(\mathbf{u}_0, \dots, \mathbf{u}_n)|.$$

For equality to hold, at least  $n$  of the  $n + 1$  projective distances  $d_j$  must equal  $d_{\min}$ .

*Proof.* Let  $D = \det(\mathbf{u}_0, \dots, \mathbf{u}_n)$ . Note that  $D \neq 0$  by the linear independence (general position) hypothesis. Without loss of generality, assume that  $d_{\min} = d_n$ . Then,  $d_{\min}^n \leq d_1 d_2 \cdots d_n$ , and equality holds if and only if all of these distances are equal. By Proposition 6,

$$d_j = \frac{|D|}{|\mathbf{u}_0 \wedge \cdots \wedge \mathbf{u}_{j-1} \wedge \mathbf{u}_{j+1} \wedge \cdots \wedge \mathbf{u}_n|}.$$

Thus,

$$d_{\min}^n \leq \frac{|D|^n}{\prod_{j=1}^n |\mathbf{u}_0 \wedge \cdots \wedge \mathbf{u}_{j-1} \wedge \mathbf{u}_{j+1} \wedge \cdots \wedge \mathbf{u}_n|}.$$

For  $j$  from 1 to  $n$ , let

$$\mathbf{v}_j = L(\mathbf{u}_0 \wedge \cdots \wedge \mathbf{u}_{j-1} \wedge \mathbf{u}_{j+1} \wedge \cdots \wedge \mathbf{u}_n),$$

and we now consider  $L(\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_n)$ . By Proposition 10,

$$L(\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_n) = \pm D^{n-1} \mathbf{u}_0.$$

Hence,

$$|L(\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_n)| = |D|^{n-1}$$

since  $|\mathbf{u}_0| = 1$ . We also know that

$$|L(\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_n)| = |\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_n| \leq |\mathbf{v}_1| \cdots |\mathbf{v}_n|$$



by [Corollary 3](#). Moreover, the inequality is strict unless  $\mathbf{v}_i \cdot \bar{\mathbf{v}}_j = 0$  for all  $i \neq j$ . Thus,

$$\begin{aligned} \prod_{j=1}^n |\mathbf{u}_0 \wedge \cdots \wedge \mathbf{u}_{j-1} \wedge \mathbf{u}_{j+1} \wedge \mathbf{u}_n| &= \prod_{j=1}^n |L(\mathbf{u}_0 \wedge \cdots \wedge \mathbf{u}_{j-1} \wedge \mathbf{u}_{j+1} \wedge \mathbf{u}_n)| \\ &= \prod_{j=1}^n |\mathbf{v}_j| \\ &\geq |L(\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_n)| = |D|^{n-1}. \end{aligned}$$

Hence,

$$d_{\min}^n \leq \frac{|D|^n}{\prod_{j=1}^n |\mathbf{u}_0 \wedge \cdots \wedge \mathbf{u}_{j-1} \wedge \mathbf{u}_{j+1} \wedge \cdots \wedge \mathbf{u}_n|} \leq \frac{|D|^n}{|D|^{n-1}} = |D|,$$

as required, with strict inequality unless  $d_1 = \cdots = d_n$  and  $\mathbf{v}_i \cdot \bar{\mathbf{v}}_j = 0$  for all  $i \neq j$ .  $\square$

**Remark.** Equality of the  $n$  distances is not sufficient for equality to hold in [Theorem 11](#), but the proof of [Theorem 11](#) suggests the following conjecture.

**Conjecture 12.** *With notation as in [Theorem 11](#), fix  $0 < D \leq 1$  and consider all configurations of  $\mathbf{u}_0, \dots, \mathbf{u}_n$  such that  $D = |\det(\mathbf{u}_0, \dots, \mathbf{u}_n)|$ . Among all such configurations, the configuration with the largest  $d_{\min}$  will be a regular simplex.*

**Remark.** When  $D < 1$ , equality will not hold in [Theorem 11](#) for the regular simplex with determinant  $D$ .

We now observe that if we like, we could just as easily work with vectors defining the faces of the simplices, rather than the vertices.

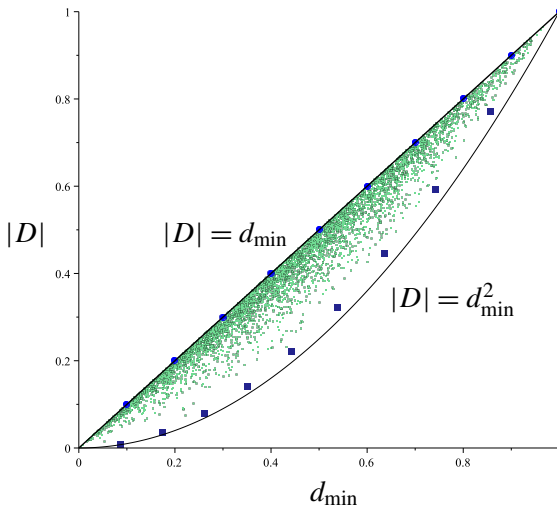
**Proposition 13.** *Let  $\mathbf{a}, \mathbf{b}_1, \dots, \mathbf{b}_n$  be  $n+1$  linearly independent unit vectors in  $\mathbb{C}^{n+1}$ . We think of the vectors as the coefficients of linear forms defining hyperplanes in  $\mathbb{C}\mathbb{P}^n$ . By linear independence, the hyperplanes are in general position and thus determine a simplex. Let  $d$  denote the distance from the hyperplane determined by  $\mathbf{a}$  to the vertex of the simplex where the hyperplanes determined by  $\mathbf{b}_1, \dots, \mathbf{b}_n$  intersect. Then,*

$$d = \frac{|\det(\mathbf{a}, \mathbf{b}_1, \dots, \mathbf{b}_n)|}{|\mathbf{b}_1 \wedge \cdots \wedge \mathbf{b}_n|}.$$

**Remark.** Observe that the distance formula here is identical to that in [Proposition 6](#). Thus, [Theorem 11](#) and [Corollary 7](#) immediately translate to the following corollary.

**Corollary 14.** *Let  $\mathbf{u}_0, \dots, \mathbf{u}_n$  be  $n+1$  linearly independent unit vectors in  $\mathbb{C}^{n+1}$  representing  $n+1$  linear forms defining  $n+1$  hyperplanes in general position in  $\mathbb{C}\mathbb{P}^n$ , which we think of as the faces of a projective simplex. For each  $j$  from 0 to  $n$ , let  $d_j$  denote the Fubini–Study distance from the hyperplane represented by  $\mathbf{u}_j$  to the opposite vertex of the simplex. Let  $d_{\min}$  denote the minimum of the  $d_j$ . Then,*

$$d_{\min}^n \leq |\det(\mathbf{u}_0, \dots, \mathbf{u}_n)| \leq d_{\min}.$$



**Figure 1.**  $|D|$  versus  $d_{\min}$  in the case of dimension  $n = 2$ .

**Remark.** Figure 1 illustrates the inequalities constraining the absolute value of the determinant and the minimum distance in the case when  $n = 2$ , i.e., for the case of projective triangles in the projective plane. The points marked as circles along the line  $|D| = d_{\min}$  illustrate isosceles triangles, as in Example 8. The points marked as squares just above the curve  $|D| = d_{\min}^2$  are from equilateral triangles. The other points are triangles with randomly generated vertices.

*Proof of Proposition 13.* Let

$$\mathbf{u} = \frac{L(\mathbf{b}_1 \wedge \cdots \wedge \mathbf{b}_n)}{|\mathbf{b}_1 \wedge \cdots \wedge \mathbf{b}_n|},$$

which is the unit vector representing the vertex of the simplex where the hyperplanes determined by  $\mathbf{b}_1, \dots, \mathbf{b}_n$  intersect. For  $j = 1, \dots, n$ , let

$$\mathbf{v}_j = L(\mathbf{a} \wedge \mathbf{b}_1 \wedge \cdots \wedge \mathbf{b}_{j-1} \wedge \mathbf{b}_{j+1} \wedge \cdots \wedge \mathbf{b}_n).$$

Then, the vectors  $\mathbf{v}_j$ , which are not necessarily unit vectors, represent the  $n$  other vertices of the simplex. By Proposition 6 and Proposition 4,

$$d = \frac{\left| \det\left(\mathbf{u}, \frac{\mathbf{v}_1}{|\mathbf{v}_1|}, \dots, \frac{\mathbf{v}_n}{|\mathbf{v}_n|}\right) \right|}{\left| \frac{\mathbf{v}_1}{|\mathbf{v}_1|} \wedge \cdots \wedge \frac{\mathbf{v}_n}{|\mathbf{v}_n|} \right|} = \frac{|\mathbf{u} \cdot L(\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_n)|}{|\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_n|}.$$

By [Proposition 10](#),  $L(\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_n) = \pm D^{n-1} \mathbf{a}$ , where  $D = \det(\mathbf{a}, \mathbf{b}_1, \dots, \mathbf{b}_n)$ . Thus,

$$\begin{aligned}
 d &= \frac{|\mathbf{u} \cdot L(\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_n)|}{|\mathbf{v}_1 \wedge \cdots \wedge \mathbf{v}_n|} \\
 &= \frac{|D|^{n-1} |\mathbf{u} \cdot \mathbf{a}|}{|D|^{n-1}} \quad (\text{since } \mathbf{a} \text{ is a unit vector}) \\
 &= \frac{|L(\mathbf{b}_1 \wedge \cdots \wedge \mathbf{b}_n) \cdot \mathbf{a}|}{|\mathbf{b}_1 \wedge \cdots \wedge \mathbf{b}_n|} \quad (\text{by the definition of } \mathbf{u}) \\
 &= \frac{|\det(\mathbf{a}, \mathbf{b}_1, \dots, \mathbf{b}_n)|}{|\mathbf{b}_1 \wedge \cdots \wedge \mathbf{b}_n|} \quad (\text{by } \a href="#">Proposition 4}). \quad \square
 \end{aligned}$$

We conclude by explaining some of the initial motivation coming from complex function theory for this investigation. Let  $\mathbb{D}$  denote the unit disc in the complex plane. J. Dufresnoy [\[1944\]](#) studied complex analytic mappings  $f$  from  $\mathbb{D}$  to  $\mathbb{C}\mathbb{P}^n$  such that the image of  $f$  omits at least  $2n + 1$  hyperplanes in general position in  $\mathbb{C}\mathbb{P}^n$ , where here *general position* means that the linear forms defining any  $n + 1$  of the hyperplanes will be linearly independent. As in [\[Cherry and Eremenko 2011\]](#), we let  $f^\#$  denote the Fubini–Study derivative of  $f$ , which measures how much the mapping  $f$  distorts length, where length in  $\mathbb{D}$  is measured with respect to the standard Euclidean metric and length in  $\mathbb{C}\mathbb{P}^n$  is measured with respect to the Fubini–Study metric. A consequence of Dufresnoy’s work is that  $f^\#(0)$  is bounded above by a constant depending only on the dimension  $n$  and the set of omitted hyperplanes, but Dufresnoy remarked in his 1944 paper that the constant depends on the omitted hyperplanes in a “completely unknown” way. By making a portion (see [\[Eremenko 1999\]](#)) of the potential-theoretic method of Eremenko and Sodin [\[1991\]](#) effective, Cherry and Eremenko [\[2011\]](#) were able to give an explicit and effective estimate on how the constant depends on the omitted hyperplanes. Cherry and Eremenko’s bound was expressed in terms of the singular values of the  $(n + 1) \times (n + 1)$  matrices formed by the coefficients of the normalized linear forms defining  $n + 1$  of the omitted hyperplanes. Let  $P$  be a point in  $\mathbb{C}\mathbb{P}^n$  where  $n$  of the  $2n + 1$  omitted hyperplanes intersect, and let  $Q$  be a point where a different  $n$  of the  $2n + 1$  omitted hyperplanes intersect. Then, the projective line connecting  $P$  with  $Q$  will intersect the  $2n + 1$  omitted hyperplanes in only three points: it will intersect  $n$  of the hyperplanes at  $P$ , another  $n$  at  $Q$  and the last one at some third point  $R$ . Such a line is called a *diagonal* line for the hyperplane configuration. In the event that the hyperplane configuration is such that for some diagonal line, two of the three points  $P$ ,  $Q$ , and  $R$  are very close together, it is not hard to see that one can find a complex analytic map  $f$  from  $\mathbb{D}$  into the diagonal line omitting the three points such that  $f^\#(0)$  is very large. One is then led to ask if this is the only way one can get a very large value of  $f^\#(0)$ . One would thus like to know how this minimum distance among the pairs of points in  $\{P, Q, R\}$

compares to the singular values appearing in Cherry and Eremenko's bound. Rather than look initially at collections of  $2n + 1$  hyperplanes in  $\mathbb{C}\mathbb{P}^n$ , we began with the easier situation of  $n + 1$  hyperplanes in  $\mathbb{C}\mathbb{P}^n$  and did some numerical experiments comparing the singular values of the matrices formed by the coefficients of the defining forms of the hyperplanes and the projective distances from the hyperplanes to the opposite vertices of the simplex whose faces are contained in the given hyperplanes. These opposite vertices would be the points determining the diagonal lines in bigger configurations of hyperplanes. Although Cherry and Eremenko's bound is expressed only in terms of some of the singular values, we realized that we could obtain prettier results for the determinant, whose absolute value is of course the square root of the product of all the singular values. We therefore decided to write this note focusing on the pure projective geometry of the simplices and leave the possible application to complex function theory to another time.

### Acknowledgments

Surya Raghavendran, during a research experiences for undergraduates project supervised by Cherry and funded by a SUMS fellowship as part of the UNT Mathematics Department's NSF funded RTG grant in the summer of 2012, made initial investigations into the relationship between the singular values of the matrix formed by three unit vectors in  $\mathbb{C}^3$  and the projective side lengths of the corresponding projective triangle in  $\mathbb{C}\mathbb{P}^2$ . Our results here build upon his initial work. We also thank Charles Conley for a stimulating discussion that led us to the proof of [Proposition 9](#).

### References

- [Bowen and Wang 1976] R. M. Bowen and C. C. Wang, *Introduction to vectors and tensors, I: Linear and multilinear algebra*, Plenum Press, New York-London, 1976. [MR 57 #2](#) [Zbl 0329.53008](#)
- [Cherry and Eremenko 2011] W. Cherry and A. Eremenko, "Landau's theorem for holomorphic curves in projective space and the Kobayashi metric on hyperplane complements", *Pure Appl. Math. Q.* **7**:1 (2011), 199–221. [MR 2900169](#) [Zbl 1251.32020](#)
- [Dufresnoy 1944] J. Dufresnoy, "Théorie nouvelle des familles complexes normales: Applications à l'étude des fonctions algébroides", *Ann. Sci. École Norm. Sup.* (3) **61** (1944), 1–44. [MR 7,289f](#) [Zbl 0061.15205](#)
- [Eremenko 1999] A. Eremenko, "A Picard type theorem for holomorphic curves", *Period. Math. Hungar.* **38**:1-2 (1999), 39–42. [MR 2000h:32018](#) [Zbl 0940.32010](#)
- [Erëmenko and Sodin 1991] A. È. Erëmenko and M. L. Sodin, "Distribution of values of meromorphic functions and meromorphic curves from the standpoint of potential theory", *Algebra i Analiz* **3**:1 (1991), 131–164. In Russian; translated in *St. Petersburg Math. J.* **3**:1 (1992), 109–136. [MR 93a:32003](#) [Zbl 0748.30026](#)
- [Richter-Gebert 2011] J. Richter-Gebert, *Perspectives on projective geometry: A guided tour through real and complex geometry*, Springer, Heidelberg, 2011. [MR 2012e:51001](#) [Zbl 1214.51001](#)

[mfincher777@gmail.com](mailto:mfincher777@gmail.com)

*Department of Mathematics, University of North Texas,  
1155 Union Circle #311430, Denton, TX 76203, United States*

[heatherolney@my.unt.edu](mailto:heatherolney@my.unt.edu)

*Department of Mathematics, University of North Texas,  
1155 Union Circle #311430, Denton, TX 76203, United States*

[wcherry@unt.edu](mailto:wcherry@unt.edu)

*Department of Mathematics, University of North Texas,  
1155 Union Circle #311430, Denton, TX 76203, United States*



## Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the Submission page at the [Involve website](#).

**Originality.** Submission of a manuscript acknowledges that the manuscript is original and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

**Language.** Articles in *Involve* are usually in English, but articles written in other languages are welcome.

**Required items.** A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

**Format.** Authors are encouraged to use  $\text{\LaTeX}$  but submissions in other varieties of  $\text{\TeX}$ , and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

**References.** Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of  $\text{\BibTeX}$  is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

**Figures.** Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to [graphics@msp.org](mailto:graphics@msp.org) with details about how your graphics were generated.

**White space.** Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

**Proofs.** Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

# involve

2015

vol. 8

no. 4

The $\Delta^2$ conjecture holds for graphs of small order	541
COLE FRANKS	
Linear symplectomorphisms as $R$ -Lagrangian subspaces	551
CHRIS HELLMANN, BRENNAN LANGENBACH AND MICHAEL VANVALKENBURGH	
Maximization of the size of monic orthogonal polynomials on the unit circle corresponding to the measures in the Steklov class	571
JOHN HOFFMAN, MCKINLEY MEYER, MARIYA SARDARLI AND ALEX SHERMAN	
A type of multiple integral with log-gamma function	593
DUOKUI YAN, RONGCHANG LIU AND GENG-ZHE CHANG	
Knight's tours on boards with odd dimensions	615
BAOYUE BI, STEVE BUTLER, STEPHANIE DEGRAAF AND ELIZABETH DOEBEL	
Differentiation with respect to parameters of solutions of nonlocal boundary value problems for difference equations	629
JOHNNY HENDERSON AND XUEWEI JIANG	
Outer billiards and tilings of the hyperbolic plane	637
FILIZ DOGRU, EMILY M. FISCHER AND CRISTIAN MIHAI MUNTEANU	
Sophie Germain primes and involutions of $\mathbb{Z}_n^\times$	653
KARENNA GENZLINGER AND KEIR LOCKRIDGE	
On symplectic capacities of toric domains	665
MICHAEL LANDRY, MATTHEW MCMILLAN AND EMMANUEL TSUKERMAN	
When the catenary degree agrees with the tame degree in numerical semigroups of embedding dimension three	677
PEDRO A. GARCÍA-SÁNCHEZ AND CATERINA VIOLA	
Cylindrical liquid bridges	695
LAMONT COLTER AND RAY TREINEN	
Some projective distance inequalities for simplices in complex projective space	707
MARK FINCHER, HEATHER OLNEY AND WILLIAM CHERRY	