

involve

a journal of mathematics

Editorial Board

Kenneth S. Berenhaut, *Managing Editor*

Colin Adams
John V. Baxley
Arthur T. Benjamin
Martin Bohner
Nigel Boston
Amarjit S. Budhiraja
Pietro Cerone
Scott Chapman
Jem N. Corcoran
Toka Diagana
Michael Dorff
Sever S. Dragomir
Behrouz Emamizadeh
Joel Foisy
Errin W. Fulp
Joseph Gallian
Stephan R. Garcia
Anant Godbole
Ron Gould
Andrew Granville
Jerrold Griggs
Sat Gupta
Jim Haglund
Johnny Henderson
Jim Hoste
Natalia Hritonenko
Glenn H. Hurlbert
Charles R. Johnson
K. B. Kulasekera
Gerry Ladas

David Larson
Suzanne Lenhart
Chi-Kwong Li
Robert B. Lund
Gaven J. Martin
Mary Meyer
Emil Minchev
Frank Morgan
Mohammad Sal Moslehian
Zuhair Nashed
Ken Ono
Timothy E. O'Brien
Joseph O'Rourke
Yuval Peres
Y.-F. S. Pétermann
Robert J. Plemmons
Carl B. Pomerance
Bjorn Poonen
József H. Przytycki
Richard Rebarber
Robert W. Robinson
Filip Saidak
James A. Sellers
Andrew J. Sterge
Ann Trenk
Ravi Vakil
Antonia Vecchio
Ram U. Verma
John C. Wierman
Michael E. Zieve



involve

msp.org/involve

INVOLVE YOUR STUDENTS IN RESEARCH

Involve showcases and encourages high-quality mathematical research involving students from all academic levels. The editorial board consists of mathematical scientists committed to nurturing student participation in research. Bridging the gap between the extremes of purely undergraduate research journals and mainstream research journals, *Involve* provides a venue to mathematicians wishing to encourage the creative involvement of students.

MANAGING EDITOR

Kenneth S. Berenhaut Wake Forest University, USA

BOARD OF EDITORS

Colin Adams	Williams College, USA	Suzanne Lenhart	University of Tennessee, USA
John V. Baxley	Wake Forest University, NC, USA	Chi-Kwong Li	College of William and Mary, USA
Arthur T. Benjamin	Harvey Mudd College, USA	Robert B. Lund	Clemson University, USA
Martin Bohner	Missouri U of Science and Technology, USA	Gaven J. Martin	Massey University, New Zealand
Nigel Boston	University of Wisconsin, USA	Mary Meyer	Colorado State University, USA
Amarjit S. Budhiraja	U of North Carolina, Chapel Hill, USA	Emil Minchev	Ruse, Bulgaria
Pietro Cerone	La Trobe University, Australia	Frank Morgan	Williams College, USA
Scott Chapman	Sam Houston State University, USA	Mohammad Sal Moslehian	Ferdowsi University of Mashhad, Iran
Joshua N. Cooper	University of South Carolina, USA	Zuhair Nashed	University of Central Florida, USA
Jem N. Corcoran	University of Colorado, USA	Ken Ono	Emory University, USA
Toka Diagana	Howard University, USA	Timothy E. O'Brien	Loyola University Chicago, USA
Michael Dorff	Brigham Young University, USA	Joseph O'Rourke	Smith College, USA
Sever S. Dragomir	Victoria University, Australia	Yuval Peres	Microsoft Research, USA
Behrouz Emamizadeh	The Petroleum Institute, UAE	Y.-F. S. Pétermann	Université de Genève, Switzerland
Joel Foisy	SUNY Potsdam, USA	Robert J. Plemmons	Wake Forest University, USA
Errin W. Fulp	Wake Forest University, USA	Carl B. Pomerance	Dartmouth College, USA
Joseph Gallian	University of Minnesota Duluth, USA	Vadim Ponomarenko	San Diego State University, USA
Stephan R. Garcia	Pomona College, USA	Bjorn Poonen	UC Berkeley, USA
Anant Godbole	East Tennessee State University, USA	James Propp	U Mass Lowell, USA
Ron Gould	Emory University, USA	József H. Przytycki	George Washington University, USA
Andrew Granville	Université Montréal, Canada	Richard Rebarber	University of Nebraska, USA
Jerrold Griggs	University of South Carolina, USA	Robert W. Robinson	University of Georgia, USA
Sat Gupta	U of North Carolina, Greensboro, USA	Filip Saidak	U of North Carolina, Greensboro, USA
Jim Haglund	University of Pennsylvania, USA	James A. Sellers	Penn State University, USA
Johnny Henderson	Baylor University, USA	Andrew J. Sterge	Honorary Editor
Jim Hoste	Pitzer College, USA	Ann Trenk	Wellesley College, USA
Natalia Hritonenko	Prairie View A&M University, USA	Ravi Vakil	Stanford University, USA
Glenn H. Hurlbert	Arizona State University, USA	Antonia Vecchio	Consiglio Nazionale delle Ricerche, Italy
Charles R. Johnson	College of William and Mary, USA	Ram U. Verma	University of Toledo, USA
K. B. Kulasekera	Clemson University, USA	John C. Wierman	Johns Hopkins University, USA
Gerry Ladas	University of Rhode Island, USA	Michael E. Zieve	University of Michigan, USA

PRODUCTION

Silvio Levy, Scientific Editor


Cover: Alex Scorpion

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2019 is US\$/year for the electronic version, and \$/year (+\$, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber 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/>

© 2019 Mathematical Sciences Publishers

Optimal transportation with constant constraint

Wyatt Boyer, Bryan Brown, Alyssa Loving and Sarah Tammen

(Communicated by Michael Dorff)

We consider optimal transportation *with constraint*, as did Korman and McCann (2013, 2015), provide simplifications and generalizations of their examples and results, and provide some new examples and results.

1. Introduction

The classical problem of optimal transportation seeks the least-cost way to move material between two locations in \mathbb{R}^n . Monge [1781] sought an optimal mapping. A more general problem, introduced in [Kantorovitch 1942], see also [Villani 2009, Theorem 3.1], seeks a cost-minimizing coupling between two measure spaces. If the coupling is absolutely continuous, it is given by a density H on the product. Recently optimal transportation has been used to better understand Riemannian manifolds and extend concepts such as Ricci curvature to more general spaces; [Cordero-Erausquin et al. 2006; Villani 2009].

Korman and McCann [2015] studied a constraint on the amount of material that can be transported between any two locations, an upper bound $h(x, y)$ on the density H , which goes back at least to [Levin 1984]. It is easy to show (Proposition 2.2) that if h is not prohibitively small, there is an optimal density H which equals 0 or h almost everywhere.

In this paper we specialize to the case of constant h . We assume $h \geq 1$, which is necessary and sufficient for existence (Proposition 2.2). Focusing on the solutions of the form 0 or h almost everywhere, for this paper we define a transportation plan as a map F from X to subsets of Y with measure $1/h$.

Our main section, Section 3, recognizes that many old and new examples of optimal transportation have the stronger “universal” property of minimizing the cost at each point separately. This leads to simplified proofs for many of the results and examples of [Korman and McCann 2013; 2015], as well as explicit examples of optimal transportation plans for all constraints $h \geq 1$. For instance, Example 3.7, due to Korman and McCann [2015, Example 1.1], provides a very short proof

MSC2010: primary 49Q20; secondary 90C08.

Keywords: optimal transportation, isoperimetric, geometry.

that optimal transportation from the unit interval to itself with cost $(x - y)^2$ with constraint $h = 2$ maps each point to whichever half of the unit interval it lies in. Proposition 3.13 proves that the intersection of two optimal transportation plans is optimal under certain conditions. Proposition 3.18 shows that in the torus or any Lie group, every admissible translation-invariant transportation plan is optimal for some continuous cost.

Proposition 4.3 presents a simplified approach to the surprising symmetries for dual cost constraints found by Korman and McCann [2013, Section 4].

Section 5 relates the case of finite spaces to some known combinatorial computations and asymptotic estimates.

2. Existence and uniqueness of optimal transportation plans

Proposition 2.2 provides existence of an optimal transportation plan F for admissible (constant) constraint h .

Definitions 2.1. Let X and Y be smooth manifolds, not necessarily compact, complete, or connected. Let f and g be nonnegative densities on X and Y , yielding probability measures on X and Y . A *transportation plan* F with constant constraint $h \geq 1$ is a measurable map from X to the power set $\mathcal{P}(Y)$ such that $F(x)$ has measure $1/h$ in Y for almost all $x \in X$ and such that $\{x \in X \mid y \in F(x)\}$ has measure $1/h$ in X for almost all $y \in Y$. (By F measurable we mean that the associated density $H(x, y)$, defined as the characteristic function of $F(x)$, is measurable.) For a cost function $c(x, y) \in L^\infty(X \times Y)$, the *total cost of transportation* is defined as

$$c[F] = \int_X \int_{F(x)} c(x, y) dy dx.$$

A transportation plan F is *optimal* if it minimizes cost.

Proposition 2.2. *Let X and Y be smooth (positive-dimensional) manifolds with non-negative densities f and g respectively and total measure 1. There exists an optimal transportation plan $F(x)$ if and only if the (constant) constraint h is at least 1.*

Proof. If $h < 1$, $F(x)$ cannot have measure $1/h$. On the other hand, if $h \geq 1$, the set of transportation densities $\Gamma(X, Y)$ is nonempty, since it includes $H(x, y) = 1$, and an optimal transportation density exists by standard compactness arguments; see [Korman and McCann 2015, Theorem 3.1].

Because $L^\infty(X, Y)$ is the dual of $L^1(X, Y)$, by Alaoglu's theorem, see [Rudin 1991, Section 3.15], the unit ball is compact in the weak-* topology. Thus the set of transportation densities $\Gamma(X, Y)$ is compact as well as convex. By the Krein–Milman theorem, every compact convex set has an extreme point, see [Wikipedia 2014a], and thus $\Gamma(X, Y)$ has an extreme point. The set of optimal transportation densities is a convex face of $\Gamma(X, Y)$ which contains an extreme point H , which is

also an extreme point of $\Gamma(X, Y)$. Such an extreme H must equal 0 or h almost everywhere, i.e., it must be a transportation plan F , see [Korman and McCann 2013, Proposition 3.2], although this can fail for finite sets of points; see Remark 2.3 below. \square

Remark 2.3. Of course there is an optimal transportation plan between finite sets (because there are only finitely many transportation plans), but the same proof does not work because there might be a better transportation density. For example the optimal transportation density from $\{0, 1\}$ to $\{0, \frac{1}{2}, 1\}$ with $h = \frac{3}{2}$ maps 0 to 1 with density $\frac{2}{3}$ and $\frac{1}{2}$ with density $\frac{1}{3}$ and maps 1 to $\frac{1}{2}$ with density $\frac{1}{3}$ and 2 with density $\frac{2}{3}$, and there is no equally good transportation plan $F(x)$; the only transportation plan maps each point to all three points. Actually Proposition 2.2 and its proof work as long as one of the two manifolds is positive-dimensional.

Although we will not need it, we provide the following uniqueness theorem of Korman and McCann.

Proposition 2.4 [Korman and McCann 2013, Theorem 3.3]. *Let X and Y be smooth manifolds with nonnegative densities f and g respectively and total measure 1. If the cost $c(x, y)$ is bounded, twice differentiable, and nondegenerate, i.e., $\det[D_{x^i y^j}^2 c(x, y)] \neq 0$ for almost all $(x, y) \in X \times Y$, then an optimal transportation plan $F(x)$ is unique (up to measure 0).*

Proof. Theorem 3.3 in [Korman and McCann 2013] gives a unique optimal density H . Since at least one optimal transportation density is an extreme point of Γ , H must be an extreme point of Γ and thus a transportation plan F . \square

Additionally, we give necessary and sufficient conditions for a map F from X to subsets of Y to be a transportation plan.

Proposition 2.5. *Let F be a measurable map from X to subsets of Y with constant constraint $h \geq 1$ such that $F(x)$ has measure $1/h$ in Y for almost all $x \in X$. Then F is a transportation plan if and only if for every $A \subset X$ of measure greater than $1/h$, $\bigcap_{x \in A} F(x)$ has measure 0.*

Proof. If F is a transportation plan, the condition holds. Suppose that F is not a transportation plan. Then it is not true that $\{x \in X \mid y \in F(x)\}$ has measure $1/h$ for almost all y . Since by Fubini's theorem the average satisfies

$$\int_Y f(\{x \in X \mid y \in F(x)\}) dy = \int_X g(F(x)) dx = 1/h$$

for some nontrivial subset of Y , we have $\{x \in X \mid y \in F(x)\}$ has measure greater than $1/h$, and the condition fails. \square

3. Universally optimal transportation

Finding optimal transportation plans for a given cost and constraint is hard. For example, the problem of optimal transportation from the unit interval $I = [0, 1]$ to itself with cost $c(x, y) = (x - y)^2$ is still open for $h \neq 2$; see [Korman and McCann 2013, Figure 1; 2015, Example 1.2]. In certain cases, however, it is possible to minimize the cost at each point separately. Further, for every optimal density, the cost function can be adjusted so that the same optimal density is also minimal at each point separately; see Remark 3.2 below.

Definition 3.1. For two smooth manifolds X and Y , a transportation plan F for the cost function c under constant constraint $h \geq 1$ is *universally optimal* if for almost every $x \in X$ it minimizes

$$\int_{F(x)} c(x, y) dy.$$

It follows immediately that F is optimal.

Remark 3.2. Korman, McCann, and Seis [Korman et al. 2015, Theorem 4.2] showed that for continuous densities f, g and $h > 1$, every optimal density is universally optimal for some equivalent cost $c(x, y) + u(x) + v(y)$ and hence for $c(x, y) + v(y)$; by Proposition 2.2 and its proof, this applies to optimal plans in the positive-dimensional case.

Morgan uses this concept of universal optimality to generalize and give shorter proofs of some of the examples of Korman and McCann.

Proposition 3.3 [Korman and McCann 2015, Example 1.3; Morgan 2013, Proposition 1]. *Let X be a Riemannian manifold of unit volume, with a transitive group of measure-preserving isometries, with cost of transportation $c(x, y)$ increasing in distance with constant constraint h . Then unique (universally) optimal transportation is that which maps each $x \in X$ to a geodesic ball about x of volume $1/h$.*

Proof. An optimal transportation plan F with constraint h must map a point $x \in X$ to a set of volume at least $1/h$, and the geodesic ball minimizes cost among such. By the symmetry assumption, all balls of the same radius have the same volume, so the set mapped to a target point $y \in Y$ is the ball about x with volume $1/h$ and the map satisfies the definition of a transportation plan and is clearly uniquely optimal (up to sets of measure 0). \square

Proposition 3.4 [Korman and McCann 2015, Example 1.1; Morgan 2013, Proposition 2]. *Let X and Y be two Riemannian manifolds of unit volume with cost of transportation $c(x, y)$ and constant constraint $h \geq 1$. Suppose that for almost all $x \in X$, $c(x, y)$ is negative for $1/h$ of the y 's in Y and nonnegative for the rest, and for almost all $y \in Y$, $c(x, y)$ is negative for $1/h$ of the x 's in X and nonnegative for*

the rest. Then unique (universally) optimal transportation maps each $x \in X$ to the subset of Y with negative cost.

Proof. By hypothesis, both $F(x)$ and $\{x \in X \mid y \in F(x)\}$ have measure $1/h$ for almost all $x \in X$ and $y \in Y$ respectively, and F is clearly universally and uniquely optimal (up to sets of measure 0). \square

Proposition 3.5. *Every transportation plan for which all images and inverse images have measure $1/h$ is optimal for some cost.*

Proof. Let $c(x, y) = -\chi_{F(x)}(y)$. Then F is optimal by Proposition 3.4. \square

Example 3.6 [Korman and McCann 2015, Example 1.1; Morgan 2013, Example 2.1]. For $h \geq 2$ an integer, let X consist of h equal-volume regions in \mathbb{R}^n such that the maximum diameter of a region is less than the minimum distance between regions. Let $c(x, y)$ be a cost function on $X \times X$ increasing in distance. Then optimal transportation from X to itself with constant constraint h maps the points of each region to itself. (To apply Proposition 3.4, subtract a constant from the cost.)

Example 3.7 [Korman and McCann 2015, Example 1.1; Morgan 2013, Example 2.2]. Let X be a centrally symmetric body in \mathbb{R}^n . For cost $c(x, y) = -2x \cdot y$, which is equivalent to $(x - y)^2$ because its integral differs by a constant, and for constraint $h = 2$, (universally) optimal transportation from X to itself is that which maps x to y with $x \cdot y$ positive (see Figure 1). In \mathbb{R}^1 central symmetry is unnecessary as long as the origin is the median. A similar result holds for any cost having the same sign at each point as $-x \cdot y$. The analysis generalizes to any centrally symmetric probability measure on \mathbb{R}^n for which hyperplanes through the origin have measure 0 and to any probability measure on \mathbb{R}^1 . Optimal transportation from X to itself with cost $-2x \cdot y$ is still open for constraint $h \neq 2$, though numerical estimates from some cases are given in [Korman and McCann 2013, Figure 1; 2015, Example 1.2].

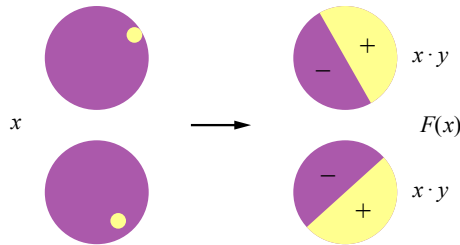


Figure 1. Optimal transportation F from the unit ball in \mathbb{R}^2 to itself with cost $c(x, y) = (x - y)^2$ and constraint $h = 2$ maps each x to the half-ball $\{x \cdot y > 0\}$.

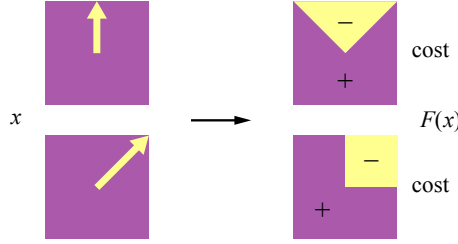


Figure 2. Optimal transportation F maps each x on a ray from the origin to a cone about that ray.

Example 3.8. Unique (universally) optimal transportation from the sphere \mathbb{S}^n to the ball \mathbb{B}^{n+1} with cost $c(x, y) = (x - y)^2$ and constraint $h = 2$ maps a point x to the half-ball $\{x \cdot y > 0\}$.

Proof. As in Example 3.6, the cost is equivalent to $-2x \cdot y$, which is negative precisely on the asserted half-ball, proving the asserted map uniquely universally optimal. \square

Example 3.9 [Morgan 2013, Example 2.3]. Let X be a planar region with h -fold rotational symmetry, such as a square ($h = 4$) as in Figure 2. For cost

$$c(x, y) = \cos(\pi/h)|x||y| - x \cdot y,$$

and constant constraint $h \geq 1$, (universally) optimal transportation maps all points on a ray from the origin to a cone of angle π/h about that ray.

Remark 3.10 [Morgan 2013, Example 2.4]. Such examples of universally optimal transportation plans from X_i to Y_i extend to universally optimal transportation plans from $\prod X_i$ to $\prod Y_i$ with a cost which is negative if and only if the costs of the projections are all negative: optimal transportation with constraint $h = \prod h_i$ maps to points of negative cost. In particular, Example 3.9 generalizes to a product of such actions on \mathbb{R}^{2n} with negative cost if and only if $x_i \cdot y_i \geq (\cos \pi/h_i)|x_i||y_i|$ for all i : optimal transportation with constant constraint $h = \prod h_i$ maps all points with projections on rays from the origin to a product of cones of angle π/h_i about the ray.

Remark 3.11 [Morgan 2013, Example 2.5]. Such examples of universally optimal transportation plans from X to Y with cost $c(x, y)$ extend to universally optimal transportation plans on warped products $A \times X$, $A \times Y$, as long as the cost $c'(a, x, a, y)$ has the same sign as $c(x, y)$. For example, for any $h \geq 1$, Proposition 3.4 on the sphere, with cost $c(x, y) = a|x||y| - x \cdot y$, with a chosen so that optimal transportation maps to points of negative cost, extends to the ball, with points on a ray from the origin mapped to a cone of negative cost about that ray.

Remark 3.12. Although universally optimal transportation plans are by definition optimal transportation plans, the converse is not true in general. Consider transportation from the unit interval to itself with cost of transportation increasing with distance and constant constraint $h = 2$. Minimizing cost for each x does not even give a valid transportation plan because points near 0 and 1 are mapped to by less than half of the interval.

Given two universally optimal transportation plans for two different costs, we seek ways to generate a third cost and a related universally optimal transportation plan.

Proposition 3.13. *Let F_1 and F_2 be optimal transportation plans from X to Y with costs c_1 and c_2 and constant constraints h_1 and h_2 respectively. Suppose that for almost all x , we have $F_i(x) = \{y \in Y \mid c_i(x, y) < 0\}$. If for some $1 \leq h < \infty$, for almost all $x \in X$, $F_1(x) \cap F_2(x)$ has measure $1/h$, and for almost all $y \in Y$, $\{x \in X \mid y \in F_1(x)\} \cap \{x \in X \mid y \in F_2(x)\}$ has measure $1/h$, then $F(x) = F_1(x) \cap F_2(x)$ is a universally optimal transportation plan from X to Y with cost $c(x, y) = \max(c_1, c_2)$ and constraint h .*

Proof. It suffices to show that for almost all $x \in X$, $c(x, y)$ is negative for $1/h$ of the $y \in Y$ and nonnegative for the rest and for almost all $y \in Y$, $c(x, y)$ is negative for $1/h$ of the $x \in X$ and nonnegative for the rest. By hypothesis on F , for almost all $x \in X$, $c(x, y)$ is negative for $1/h$ of the $y \in Y$. It is nonnegative for the rest because $x \notin F(x)$ implies some $c_i(x, y)$ must be nonnegative; thus $c(x, y)$ must also be nonnegative. The reverse condition holds by a similar argument. \square

Corollary 3.14. *Let X be a region with 4-fold rotational symmetry in \mathbb{R}^2 with cost of transportation from X to X given by $c(x, y) = \max((x \cdot y), \det[x \mid y])$, where $\det[x \mid y]$ is the determinant of the matrix with x and y as its column vectors. Mapping each point to the region of negative cost uniquely gives (universally) optimal transportation for $h = 4$ (see Figure 3).*

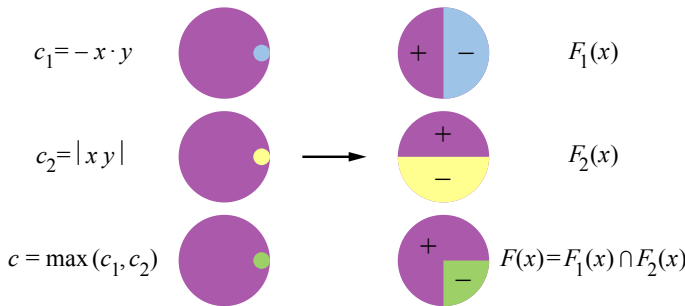


Figure 3. The intersection of optimal transportation plans yields a new optimal transportation plan under certain hypotheses.

Proof. The map $F_1(x) = \{y \in X \mid x \cdot y > 0\}$ is an optimal transportation plan from X to itself with cost $c_1(x, y) = -x \cdot y$ and constraint $h = 2$ (see Example 3.6). Similarly the cost $c_2(x, y) = \det[x \mid y]$ with constraint $h = 2$ satisfies the hypotheses of Proposition 3.4 and thus the map $F_2(x) = \{y \in X \mid \det[x \mid y] < 0\}$ is an optimal transportation plan from X to itself with cost $c_2(x, y)$ and constraint $h = 2$. By Proposition 3.13, if for almost all $x, y \in X$, $F_1(x) \cap F_2(x)$ and $\{x \in X \mid y \in F_1(x)\} \cap \{x \in X \mid y \in F_2(x)\}$ both have constant measure $1/h$ for some $h > 1$, then $F(x) = F_1(x) \cap F_2(x)$ is an optimal transportation plan from X to itself with cost $c(x, y) = \max(c_1, c_2)$ and constraint $1/h$. For almost all $x \in X$, $\partial F_1(x)$ is the line through the origin normal to the line through x and the origin and $\partial F_2(x)$ is the line through x and the origin. Because two normal lines both through the origin partition a region with 4-fold rotational symmetry centered on the origin in \mathbb{R}^2 into four congruent regions, and exactly one of these regions is equivalent to $F_1(x) \cap F_2(x)$, it follows that $F_1(x) \cap F_2(x)$ has constant measure $\frac{1}{4}$. Similarly, the boundary of the set $\{x \in X \mid y \in F_1(x)\}$ is the line through the origin normal to the line through y and the origin and the boundary of the set $\{x \in X \mid y \in F_2(x)\}$ is a line through y and the origin. Thus, by the same argument as above, $\{x \in X \mid y \in F_1(x)\} \cap \{x \in X \mid y \in F_2(x)\}$ has measure $\frac{1}{4}$. By Proposition 3.13, the asserted map is optimal. \square

Remark 3.15. If the hypotheses of Proposition 3.13 hold, then the maps $F(x) = F_1(x) \cup F_2(x)$ and $F(x) = F_1(x) \triangle F_2(x)$ are optimal transportation plans for costs $c = \min(c_1, c_2)$ and $c' = c_1 \cdot c_2$ and some constraints h and h' respectively (the symbol \triangle denotes the symmetric difference of two sets).

Example 3.16. Let X be a region with 4-fold rotational symmetry in \mathbb{R}^2 . Then an optimal transportation plan F for cost

$$c(x, y) = ((\cos 3\pi/4h)|x||y| - x \cdot y)((\cos \pi/4h)|x||y| - x \cdot y)$$

and constraint $h = 2$ maps points on a ray from the origin to two cones (see Figure 4).

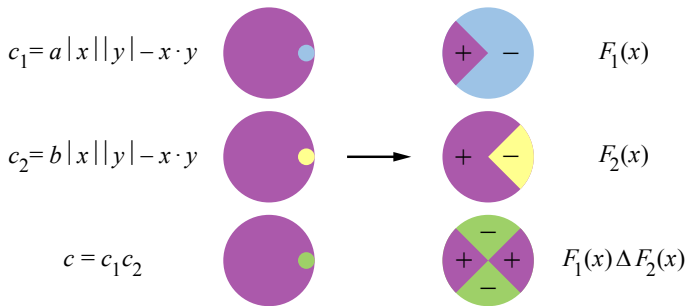


Figure 4. Other set operations yield even more examples of optimal transportation.

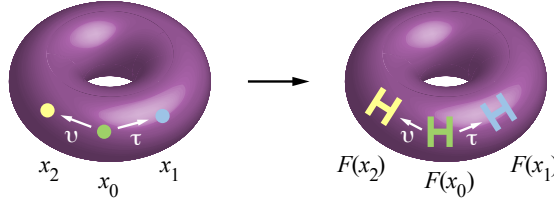


Figure 5. Optimal transportation F maps each x to an H -shaped region.

The condition in Proposition 3.13 that $F_1(x) \cap F_2(x)$ have constant measure $1/h$ for almost all $x \in X$ and some h is independent of the condition that $\{x \in X \mid y \in F_1(x)\} \cap \{x \in X \mid y \in F_2(x)\}$ have constant measure $1/h$ for almost all $y \in Y$.

Example 3.17. Let X be $\{1, 2, 3, 4\}$ or equivalently the unit interval divided into four quarters. Consider transportation F_1, F_2 from X to X such that

$$\begin{aligned} F_1(1) &= \{3, 4\}, & F_1(2) &= \{2, 3\}, & F_1(3) &= \{1, 2\}, & F_1(4) &= \{1, 4\}, \\ F_2(1) &= \{1, 4\}, & F_2(2) &= \{1, 2\}, & F_2(3) &= \{2, 3\}, & F_2(4) &= \{3, 4\}. \end{aligned}$$

Then $F(x) = F_1(x) \cap F_2(x)$ has constant measure $\frac{1}{4}$ but

$$\{x \in X \mid y \in F(x)\} = \{x \in X \mid y \in F_1(x)\} \cap \{x \in X \mid y \in F_2(x)\}$$

has measure $\frac{1}{2}$ for $\{2, 4\}$ and measure 0 for $\{1, 3\}$.

Proposition 3.18. *Let X be a Lie group. Given an open subset A of X with measure $1/h$, there exists a continuous cost function $c(x, y)$ such that the unique (universally) optimal transportation plan F from X to itself with constant constraint h maps the identity to the set A and maps each element $x \in X$ to the set $x \cdot A = xA$.*

Proof. Let the cost $c(x, y)$ equal the distance from y to the boundary of $x \cdot A$, with negative cost on the interior of $x \cdot A$ and nonnegative cost on the complement of $x \cdot A$. By Proposition 3.4, the asserted map is optimal. \square

Example 3.19. Let $X = \mathbb{S}^1 \times \mathbb{S}^1$ with unit area. Given an open subset $A \subset X$ with measure $1/h$, such as the H -shaped region in Figure 5, there exists a continuous cost function $c(x, y)$ such that the unique (universally) optimal transportation plan F from X to itself with constant constraint h maps the origin to the set A and maps almost every x to the set $\tau_x(A)$, where τ_x is the translation that takes the origin to x .

Example 3.20. Optimal transportation from a flat rectangular torus to itself with cost $c(x, y) = \min(d(x_i, y_i))$ and constraint h maps each point to a neighborhood around the coordinate axis centered at that point; see Figure 6.

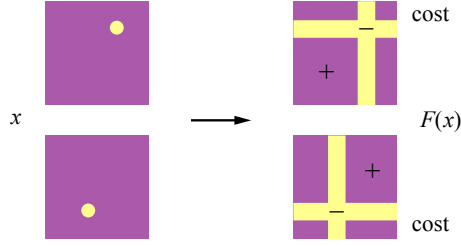


Figure 6. Optimal transportation on the flat rectangular torus maps each x to a small neighborhood around the coordinate axis centered at that point.

4. Transportation and symmetry

Korman and McCann [2013, Section 4] found surprising symmetries between optimal transportation plans with dual constraints. Proposition 4.3 presents a simplified approach.

Definition 4.1. A map f from X' to X is called *measure preserving* if the measure of any $A \subset X$ equals the measure of $f^{-1}(A) \subset X'$.

Proposition 4.2. Let F be an optimal transportation plan from X to Y with cost $c(x, y)$ and constraint h . Let f and g be measure-preserving maps from X' to X and from Y' to Y respectively. Then $G(x') = g^{-1}(F(f(x')))$ provides optimal transportation from X' to Y' with cost $c \circ (f, g)$ and constraint h .

Proof. We need to show that $G(x')$ and $G^{-1}(y')$ both have measure $1/h$ and that the total cost of transportation is minimal. For $x' \in X'$, $G(x') = g^{-1}(F(f(x')))$ must have the same measure as $F(f(x'))$, which is $1/h$ by hypothesis. Similarly, for $y' \in Y'$, $G^{-1}(y') = f^{-1}(F^{-1}(g(y')))$ must have the same measure as $F^{-1}(g(y'))$, which is also $1/h$ by hypothesis. To show that G is optimal, we will show that G and F have the same cost and that any other transportation plan G_2 from X' to Y' has the same cost as an analogous transportation plan F_2 from X to Y and therefore must be of greater total cost than G . The cost of transportation from x' to y' is equal to $c(f(x'), F(f(x')))$; thus G and F have the same total cost of transportation. Let G_2 be another transportation map from X' to Y' . Let $F_2(f(x')) = g(G_2)$. Then $G_2 = g^{-1}(F_2(f(x')))$ and the result follows. \square

Proposition 4.3 [Korman and McCann 2013, Lemma 4.1; Morgan 2013, Proposition 3]. Let M_1, M_2 be subsets of \mathbb{R}^n or Riemannian manifolds with boundary or metric measure spaces of volume V . Let T_i be a measure-preserving map from M_i to itself and let $T = T_1 \times T_2$. Let $c(x, y)$ be a cost satisfying $c \circ T = -c$. If the map F is an optimal transportation plan from M_1 to M_2 with cost $c(x, y)$ with

constraint h , then the map $T_2(F' \circ T_1)$ is an optimal transportation plan from M_1 to M_2 with cost c and constraint h' , where $1/h + 1/h' = 1$ and $F'(x) = F(x)^C$.

Proof. If F is an optimal transportation plan for cost c and constraint h , then $F(x)' = F(x)^C$ is the most expensive transportation plan for cost c with constraint h' , and hence an optimal transportation plan for cost $-c$. Therefore $T_2(F' \circ T_1)$ is an optimal transportation plan for cost $-c \circ T = c$ and constraint h' . \square

Example 4.4 [Morgan 2013, Example after Proposition 3; Korman and McCann 2013, Lemma 4.1]. Let M_1 and M_2 be subsets of \mathbb{R}^n , with M_1 centrally symmetric, and let $c(x, y) = -x \cdot y$ (which is equivalent to $(x - y)^2$). Then central inversion in x carries optimal transportation with constraint h to optimal transportation with constraint h' .

5. Transportation plans on finite sets

Consider the case where X and Y are finite sets, say $X = \{1, 2, \dots, m\}$ and $Y = \{1, 2, \dots, n\}$. In this case we may assume that the constraint h is a common divisor of m and n . A map F from X to Y is equivalent to the $n \times m$ matrix of 0's and 1's with entry $a_{ij} = 1$ if and only if $i \in F(j)$; see [Wikipedia 2014b; Weisstein]. Such a matrix gives a transportation plan if and only if the matrix has m/h 1's in each column and n/h 1's in each row. Thus the number of transportation plans is equal to the number of $n \times m$ binary matrices with constant column sums n/h and constant row sums m/h . Asymptotic estimates exist for large m, n ; see [Canfield and McKay 2005; McKay and Wang 2003].

Acknowledgements

We thank the National Science Foundation and Williams College for supporting this work as part of the Williams College SMALL Undergraduate Research Project. We thank our advisor Frank Morgan for his insight, enthusiasm, and guidance. We thank Robert McCann and Christian Seis for helpful comments. Additionally we thank the Mathematical Association of America, Williams College, and Pomona College for supporting our travel to MathFest 2014.

References

- [Canfield and McKay 2005] E. R. Canfield and B. D. McKay, "Asymptotic enumeration of dense 0-1 matrices with equal row sums and equal column sums", *Electron. J. Combin.* **12** (2005), art. id. R29. MR Zbl
- [Cordero-Erausquin et al. 2006] D. Cordero-Erausquin, R. J. McCann, and M. Schmuckenschläger, "Prékopa–Leindler type inequalities on Riemannian manifolds, Jacobi fields, and optimal transport", *Ann. Fac. Sci. Toulouse Math.* (6) **15**:4 (2006), 613–635. MR Zbl

- [Kantorovitch 1942] L. Kantorovitch, “On the translocation of masses”, *C. R. (Doklady) Acad. Sci. URSS (N.S.)* **37** (1942), 199–201. In Russian; translated in *J. Math. Sci.* **133**:4 (2006), 1381–1382. MR Zbl
- [Korman and McCann 2013] J. Korman and R. McCann, “Insights into capacity-constrained optimal transport”, *Proc. Nat. Acad. Sci. USA* **110**:25 (2013), 10064–10067.
- [Korman and McCann 2015] J. Korman and R. J. McCann, “Optimal transportation with capacity constraints”, *Trans. Amer. Math. Soc.* **367**:3 (2015), 1501–1521. MR Zbl
- [Korman et al. 2015] J. Korman, R. J. McCann, and C. Seis, “Dual potentials for capacity constrained optimal transport”, *Calc. Var. Partial Differential Equations* **54**:1 (2015), 573–584. MR Zbl
- [Levin 1984] V. L. Levin, “The problem of mass transfer in a topological space and probability measures with given marginal measures on the product of two spaces”, *Dokl. Akad. Nauk SSSR* **276**:5 (1984), 1059–1064. In Russian. MR Zbl
- [McKay and Wang 2003] B. D. McKay and X. Wang, “Asymptotic enumeration of 0-1 matrices with equal row sums and equal column sums”, *Linear Algebra Appl.* **373** (2003), 273–287. MR Zbl
- [Monge 1781] G. Monge, “Mémoire sur la théorie des déblais et des remblais”, pp. 666–704 in *Histoire de l’Académie Royale des Sciences de Paris*, De l’Imprimerie Royale, Paris, 1781.
- [Morgan 2013] F. Morgan, “Optimal transportation with constraint”, blog post, 2013, available at <http://sites.williams.edu/Morgan/2013/09/14/optimal-transportation-with-constraint/>.
- [Rudin 1991] W. Rudin, *Functional analysis*, 2nd ed., International Series in Pure and Applied Mathematics, McGraw-Hill, New York, 1991. MR Zbl
- [Villani 2009] C. Villani, *Optimal transport: old and new*, Grundlehren der Mathematischen Wissenschaften **338**, Springer, 2009. MR Zbl
- [Weisstein] E. W. Weisstein, “(0,1)-matrix”, available at <http://tinyurl.com/01matrix>. From MathWorld.
- [Wikipedia 2014a] “Extreme point”, Wikipedia entry, 2014, available at http://en.wikipedia.org/wiki/Extreme_point.
- [Wikipedia 2014b] “Logical matrix”, Wikipedia entry, 2014, available at https://en.wikipedia.org/wiki/Logical_matrix.

Received: 2014-12-22 Revised: 2016-04-19 Accepted: 2016-05-02

wyatt.b.boyer@gmail.com *Department of Mathematics and Statistics, Williams College,
Brighton, MA, United States*

bbrown@math.berkeley.edu *Department of Mathematics, Pomona College,
Claremont, CA, United States*

aloving2@illinois.edu *Department of Mathematics, University of Hawaii at Hilo,
Hilo, HI, United States*

setammen@uga.edu *Department of Mathematics, University of Georgia,
Lawrenceville, GA, United States*

Fair choice sequences

William J. Keith and Sean Grindatti

(Communicated by Kenneth S. Berenhaut)

We consider turn sequences used to allocate of a set of indivisible items between two players who take turns choosing their most desired element of the set, with the goal of minimizing the advantage of the first player. Balanced alternation, while not usually optimal, is fairer than alternation. Strategies for seeking the fairest choice sequence are discussed. We show an unexpected combinatorial connection between partition dominance and fairness, suggesting a new avenue for future investigations in this subject, and conjecture a connection to a previously studied optimality criterion. Several intriguing questions are open at multiple levels of accessibility.

1. Introduction

In the discrete version of the cake-cutting problem [Brams and Taylor 1996], some number of people take turns selecting from among a set of indivisible items (usually $2n$ items for two people). Players' preferences vary (they may not all prefer the same item most, second-most, et cetera). If preferences are not known to other players they are usually assumed to vary with uniform probability. Players' preferences are normally described by a simple ranking, called Borda scoring, which assigns values of 1 through $2n$ to the objects being chosen. An object with value $2n$ is most wanted, and gives twice as much satisfaction as the object valued at n , and so forth. A player assigns a utility to a final distribution of goods at the total of their valuation of all objects they receive. It is easy to see that the sum of all players' utilities is by no means constant as preferences or turn orders vary. This has been a problem of interest for many authors; see [Bouveret and Lang 2011; Hopkins and Jones 2009; Kalinowski and Narodytska 2013; Rubchinsky 2010], and others.

Some investigators (Hopkins [2010], Hopkins and Jones [2009]) consider strategic play when players' preferences are known to each other. If preferences are secret, a player's only strategic move is to take their most-preferred item remaining, and instead the question of interest is whether an administrator who also does not know preferences can vary the policy — the sequence in which turns are allocated —

MSC2010: primary 91A05; secondary 05A17.

Keywords: social choice, fair division, permutations, fairness, egalitarian, partitions, dominance.

to probably maximize some criterion of social interest. Policies have been analyzed for optimality criteria such as min-max (the worst-off agent is likely to do least badly) and social welfare or utilitarian optimality (the expected value of the sum of agents' utilities is as high as possible). For instance, Bouveret and Lang [2011] conjectured that simply taking turns is utilitarian-optimal under uniform distribution of preferences, and Kalinowski and Narodytska [2013] proved this. Data from [Bouveret and Lang 2011] show that alternating turns is not min-max optimal, although it is asymptotically so.

In this paper we define a new optimality criterion, fairness: the expected difference between players' total utilities is minimized. We restrict ourselves to conditions common in the literature (see for instance [Bouveret and Lang 2011; Kalinowski and Narodytska 2013]): Borda scoring, as above, and uniformly distributed preferences. Socially, the criterion is useful when players are intolerant of large inequalities among their outcomes. Such players must also be willing to sacrifice some overall social welfare in order to reduce this, because under the given conditions maximal total utility is known [Kalinowski and Narodytska 2013] to be realized by the simple policy of taking turns: but this obviously advantages the first player. Empirically, fair policies never seem to be too far from utilitarian-optimal policies, but there does not seem to be a strong mathematical connection between the two criteria. However, we were surprised to conjecture from data generated to date that the min-max optimal policy is the fairest among policies that only differ from alternation in which player goes first in a "round". Finally, the fairness criterion also turns out to have a surprising combinatorial property connected with the theory of partitions. This connection is partially proved herein but is still open in general. Thus we think there is significant mathematical interest to be explored here.

We prove a number of results for fairer policies. Our theorems range from the intuitively obvious, to a fascinating combinatorial relationship between the dominance order on partitions and fairness of choice sequences associated to partitions in a natural way. This association is a tool not previously used in the literature, which may yield fruitful lines of analysis for other questions. Stating our main theorems accessibly, deferring technical definitions to the next section, we show the following:

Lemma 1. *Inverting the choice sequence negates advantage.*

Theorem 2. *Moving a player's choices later strictly decreases their advantage.*

Theorem 3. *Altering a four-turn sequence from $LRRL$ to $RLLR$ strictly increases player L 's advantage, if all turn pairs in positions $2k + 1$ and $2k + 2$ are LR or RL .*

In other words, heuristically, *players like earlier choices — but, other things being equal, they would like their earlier choices later.* The first part of this theorem is obvious — the second, we think, quite surprising. This is the dominance connection

we mentioned earlier, and we conjecture but were unable to prove that in fact a stronger connection to partition dominance holds.

In the following section we provide the definitions that a reader will need, including definitions specific to this subject and some background from disparate areas, intending to make the article as self-contained as possible. In Section 3 we describe and prove our theorems on the fairness-of-choice sequences in various relations. In Section 4 we give some numerical, nonrecursive formulas for players' expected values using tools from combinatorics. In the last section we give our collected data to date, and suggest remaining open problems and interesting questions our work raises. Interested investigators will find material for computational projects for students, as well as challenging questions in combinatorial distributions.

For nonmathematical readers only skimming the article to find a “fairest” choice sequence, while not perfect, we recommend the following for players L and R : the “reverse-and-repeat” sequence

$LRRL \ RLLR \ RLLR \ LRRL \ RLLR \ LRRL \ LRRL \ RLLR \ RLLR \ \dots$

Known as balanced alternation [Brams and Taylor 2000], this is well-defined only if the number of turns is a power of 2, but it is easy to then simply take an initial segment of a sufficiently long sequence.

2. Definitions

Two agents, Luis and Rita, each rank a set of $2n$ indivisible items in order of preference; without loss of generality we assume Luis's preference from most-preferred to least is labeled $(2n, 2n-1, \dots, 1)$, and consider Rita's preferences a permutation π of Luis's. We describe Rita's preference order by $\pi^{-1}(2n), \dots, \pi^{-1}(1)$. Reading left to right, we obtain Rita's ranking of the items from most to least preferred. From here on, by “item i ” we mean Luis's label for the item, and specify “Rita's item i ” if required. Neither player knows the other's preference; we assume preferences are uniformly distributed.

Each agent takes an equal number of turns on which they select their most preferred (highest-labeled) item of those remaining to be chosen; e.g., if Luis goes first, he will choose the item labeled $2n$. This ends when each person has exactly half of the items. Following Kalinowski, Narodytska and Walsh [Kalinowski and Narodytska 2013], we refer to an order S in which players are allowed to choose items as a *policy*.

A policy is a word in L and R of length $2n$ containing n L s and n R s. A policy signifies turns at which Luis or Rita chooses from among the remaining set of objects their most preferred item, receiving that item and deleting it from the remaining choices. The set of positions $\{\ell_1, \dots, \ell_n\}$ at which L appears, or R respectively, is the set of *choice positions* for that player.

In a choice of four items in which Rita prefers items at 1432, and the policy is $LRLR$, items in Luis's labeling will be taken in the order 4, 1, 3, 2. We refer to this order as the *path* associated to this policy and preference.

Given a set of preferences and a policy, players will receive some collection of items, which they value as the sum of their valuations for each item received: label these sums $L(S)$ or (S) for a given policy S . Luis's advantage is his value minus Rita's.

The *alternating policy* in which Luis goes first, $LRLR \dots LR$, advantages Luis, since he chooses first in every "round" of a choice for each player. For instance, if Luis and Rita agree on their ranking of the items, Luis will get his 1st, 3rd, 5th, etc. choices, while Rita will get her 2nd, 4th, etc.

For four items, we might argue that $LRRL$ is fairer than $LRLR$; for instance, in the case of agreement, Luis will get items he values at 4 and 1 while Rita gets 3 and 2. We call a policy S_1 *fairer* than policy S_2 if the two players' expected totals (averaging over all possible Rita preference orders) differ by less in absolute value. Equivalently, we may compare their total values over all permutations, $L_{\text{tot}}(S)$ and $R_{\text{tot}}(S)$. Define $L_{\text{adv}}(S) = L_{\text{tot}}(S) - R_{\text{tot}}(S)$. The goal of this article is to seek the fairest policy, minimizing $|L_{\text{adv}}(S)|$.

Remark. Utilitarian-optimality is the usual criterion in the literature. We choose to investigate fairness both for the inherent mathematical interest described above, such as connections to other criteria and mathematical objects further afield, and for the use of cases in which inequality is a phenomenon players accord some negative utility. One could quantize the difference between the social welfare of the fairest and the utilitarian-optimal policies by giving a function to weight the amount by which players disapprove of inequality. Doing so makes all prior criteria instances of a general continuum! If players add to social welfare a "disapproval subtraction" equal to advantage, the resulting optimality criterion is precisely min-max, whereas if they subtract nothing, the criterion is utilitarian-optimality. Our criterion is equivalent to a "disapproval subtraction" of some large multiple of inequality. Study of this generalized criterion could be an interesting route to rigorize the investigation of tradeoffs between various criteria.

Before including a few definitions from other areas of mathematics that will later be useful to us, we state a lemma from [Kalinowski and Narodytska 2013], a very useful recursion that gives the expected value $\bar{u}_i(S)$ of player i for a policy S . Say that a policy S has length p , and denote by \tilde{S} the policy S with the first choice removed — note that this lemma applies to policies of any length, and not necessarily having the same number of places L and R .

Lemma 4 [Kalinowski and Narodytska 2013, Lemma 1]. *Let S be a policy of length p . Denote by $\bar{u}_1(S)$ the expected value of the player choosing first in S , and*

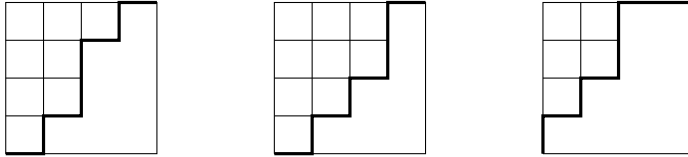
by $\bar{u}_2(S)$ the expected value of the other player. We have

$$\bar{u}_1(S) = p + \bar{u}_1(\tilde{S}), \quad \bar{u}_2(S) = \frac{p+1}{p} \bar{u}_2(\tilde{S}).$$

A *partially ordered set* (poset for short) is a set A with a reflexive, transitive, antisymmetric relation $a \leq b$. If $a \leq b$, $a \neq b$, and no element c satisfies $a \leq c \leq b$, we say that b *covers* a . If for any two elements $c_1, c_k \in A$ all chains $c_1 \leq c_2 \leq \dots \leq c_k$ have equal length when c_{i+1} covers c_i for all i , then the poset is *ranked*; if there is a unique element $c \leq x$ for all $x \in A$ the rank of x may be taken to be the length of any such chain between c and x .

A *partition* of n is a weakly decreasing sequence $\lambda = (\lambda_1, \dots, \lambda_M)$ of nonnegative integers that sum to n . (Typically a partition is defined with positive integers, but it is convenient in this article to speak of a finite number of size 0 parts.) A partition in the $N \times M$ box is one with at most M parts of size at most N ; in this article N will always equal M , and we will refer to a box of size N . The “box” language comes from the *Ferrers diagram* of a partition, which is a collection of squares justified to the axes in the fourth quadrant in which the i -th row has λ_i squares.

Example 5. The Ferrers diagrams of the partitions $(3, 2, 2, 1)$, $(3, 3, 2, 1)$, and $(2, 2, 1, 0)$ in the 4×4 box are illustrated below.



The *profile* of a partition is the set of $E - W$ and $N - S$ segments that form the outer boundary of its Ferrers diagram, possibly including any desired number of segments along the axes. A partition λ *dominates* another partition π if it holds that $\sum_{i=1}^k \lambda_i \geq \sum_{i=1}^k \pi_i$ for all k , assuming an infinite set of trailing zero parts in each. Dominance is a partial order. For instance, $(4, 4, 2, 1, 1)$ dominates $(4, 3, 3, 1, 1)$ but not $(4, 3, 3, 2)$, nor does $(4, 3, 3, 2)$ dominate $(4, 4, 2, 1, 1)$.

A partition $\lambda = (\lambda_1, \dots, \lambda_n)$ *contains* another partition $\sigma = (\sigma_1, \dots, \sigma_n)$ if we have $\lambda_i \geq \sigma_i$ for all i . Containment is a partial order relation which makes partitions into a ranked poset, where the rank of a partition is just the number that it partitions. Thus the set of partitions in the $N \times M$ box ordered by containment is a finite, ranked poset with minimal element $(0, 0, \dots, 0)$. Containment implies dominance, which means that the containment poset on partitions in the $M \times N$ box can be constructed by removing some comparabilities from the dominance poset. In particular, no two partitions of n contain each other, while dominance can be used as a partial order on the set of partitions of n in a given box. Unless we refer to dominance

for a particular theorem, in this paper we mean containment when we speak of the partition poset or the word poset.

We define dominance and covering on words by associating them to partitions. We associate a word of length $2n$ with n each of L and R to a profile in the $n \times n$ box by starting from the upper right corner and drawing an $E - W$ step for an L , and a $N - S$ step for an R . Thus, the partitions in the example above are associated to the sequences $LRLRRLRL$, $LRRLRLRL$, and $LLRRLRLR$ respectively. We say that a word in L and R dominates (resp. covers) another if the associated partition of the first dominates (resp. covers) that of the second. In the figure above, $(3, 3, 2, 1)$ dominates both of the other partitions, and covers $(3, 2, 2, 1)$. These two figures are relevant to an example we give in the next section.

Of particular interest to us is the set of words that lie between the alternating sequences in the word poset which are restricted to consecutive pairs LR or RL for every two $(2k - 1)$ -th and $2k$ -th positions, which we refer to as the Boolean set. These words are associated to the 2^n partitions whose Ferrers diagrams differ only by either containing, or not containing, the squares on the diagonal of the box. The first two partitions in the example are in this set.

3. Effects of changes on the fairness of policies

The policy associated to the minimal element $(0, 0, \dots, 0)$ in the box of size N is $LL \dots RR$, which among all policies with n L s and n R s is obviously the best possible policy for Luis (highest L_{tot} and L_{adv}). We can move through the policy poset by adding one square at a time, exchanging a consecutive pair LR in the sequence for an RL . We should expect that this will always worsen Luis's position, and our first theorem shows this. Slightly less obviously, for any given Rita preference permutation the change happens in a specific way, by the exchange of one pair of items between the two players' outcomes.

Theorem 6. *Let two policies S and S' be such that $S = s_1 \dots s_k s_{k+1} \dots s_{2n}$ where $s_k = L$ and $s_{k+1} = R$, and $S' = s_1 \dots s_{k+1} s_k \dots s_{2n}$. Then $L(S') \leq L(S)$ and $R(S') \geq R(S)$, by exchange of one item between the players' outcomes for any given Rita preference π .*

Proof. The magnitude clause is extremely intuitive and, with Lemma 4, is nearly trivial: Luis's expected value in $LR\sigma$ with σ any following policy of length $p - 2$ is $p + p/(p - 1) \bar{u}_L(\sigma)$, while for $RL\sigma$ it is $(p + 1)/p (p - 1 + \bar{u}_L(\sigma))$. The former is larger than the latter, and Lemma 4 tells us that passage through any prefix to such a word yields an expected value which is some increasing linear function of the input, so the final expected value is larger.

However, by examining the situation in more detail we can say more about the actual change made to the players' outcomes.

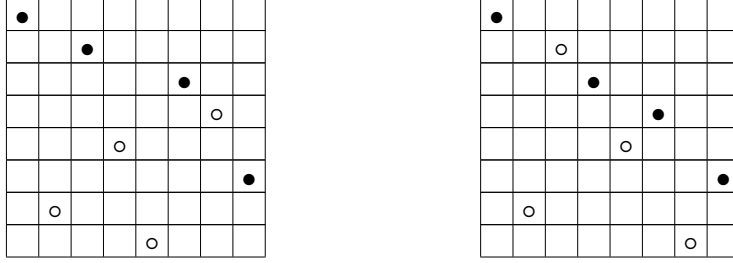


Figure 1. Left: choice procedure for policy S . Right: choice procedure for policy S' . Note that white circles denote items Rita obtains and black circles denote Luis' items.

Let Luis's outcome be $\text{out}_L(S) = \{l_1, l_2, \dots, l_n\}$ and Rita's outcome $\text{out}_R(S) = \{r_1, r_2, \dots, r_n\}$. The first $k - 1$ turns with S' will yield the same results for both players as with S . On turn k , using S , Luis takes some item l_i , and then on turn $k + 1$ Rita takes some item r_j . For S' , turn k is Rita's turn instead of Luis's. She will either take r_j as she did using S or she will take l_i . If she takes r_j , then on the next turn, Luis will take l_i , for he preferred this to all remaining items. Then for the remainder of the items, at any given choice both players will face the same set of remaining items in S' as they did in S , and will make the same choices with S' as they did with S . Thus $\text{out}_L(S) = \text{out}_L(S')$ and $\text{out}_R(S) = \text{out}_R(S')$.

If, instead, Rita takes l_i , then Luis will later take at least one item r_m belonging to Rita's previous outcome $\text{out}_R(S)$ such that both Luis and Rita prefer l_i to r_m . This worsens his outcome and betters hers. We claim that this exchange of l_i for r_m is the only difference in outcomes for S and S' , so $L(S') < L(S)$, and $R(S') > R(S)$.

For example, suppose that $S = LRLRRLRL$, $S' = LRRLRLRL$. Observe that their associated partitions are the first two we illustrated in the example of the previous section, so that the policy (and the associated partition of S' covers and dominates S . We have moved up one step in the word poset, from $(3, 2, 2, 1)$ partitioning 8, to $(3, 3, 2, 1)$ partitioning 9.

Let Rita's preference be given by $\pi = 28741563$. We illustrate the choice procedure for S in Figure 1, left, with white circles denoting items Rita obtains and black circles denoting Luis's items.

Using choice sequence S' , Luis still takes 8 and Rita still takes 2. The third turn is now Rita's instead of Luis's, and since they agree on the best item at that time, Rita takes $l_2 = 7$. On his second turn, Luis takes his $l_3 = 6$. On Rita's third turn, she takes her $r_2 = 4$. On Luis's third turn, he could take his $l_4 = 3$, but he now has access to $r_3 = 5$, so he takes that instead. Now since Rita cannot take her r_3 , she takes $r_4 = 1$ on her fourth turn. On Luis's fourth turn, he also takes his $l_4 = 3$; see Figure 1, right.

In general, if Rita takes l_i , it must follow that she prefers l_i to r_j , her previous choice using S . On each of her subsequent turns g , she will take her r_{g-1} if it is available, or r_g if it is not. Why?

If Luis has not yet taken any item r_g , then on his turns he was taking items he had previously chosen one turn later. In this case, on any of Rita's turns after the change, the items taken so far will consist of items Luis had previously chosen prior to that turn, plus one item Luis had chosen later than her current turn, plus the items Rita had previously taken, except for the item she previously took on the turn before the current turn. Rita preferred that item to all remaining of Luis's items and all remaining of her own, and so she will take it.

In other words, Rita has basically delayed taking items she previously chose because a better item is now available.

On turn $k + 1$, since l_i is no longer available, Luis will either take l_{i+1} or r_j . On each of his subsequent turns h , he will take his l_h if it is available, or either l_{h+1} or r_m , Rita's next choice, if it is not.

On some turn t , Luis must take one of Rita's items r_m because, since Rita took l_i , she can no longer collect all n of the items r_1, r_2, \dots, r_n she collected with choice sequence S . Suppose that turn t is the earliest place this happens. At this point, the items previously chosen constitute the same set as the items that had been chosen at this point in the previous choice sequence; Luis has filled the delay in Rita's choices.

Thus the rest of each player's choices will follow the same with S' as with S , so Luis's and Rita's outcomes using S' will be identical to those for S , with the single exception that Rita now gets l_i and Luis gets r_m . Since Luis chose l_i over r_m using sequence S and Rita chose l_i over r_m using S' , we know l_i is more valuable to both Luis and Rita than r_m . Therefore $L(S') < L(S)$ and $R(S') > R(S)$. \square

The inequalities become strict for L_{tot} and R_{tot} , since there will be at least one strict difference, when Rita's preference permutation is the identity.

Corollary 7. *Let S and S' be choice sequences as before, with S' covering S . Then $L_{\text{tot}}(S') < L_{\text{tot}}(S)$ and $R_{\text{tot}}(S') > R_{\text{tot}}(S)$, and hence $L_{\text{adv}}(S') < L_{\text{adv}}(S)$.*

Proof. Since we know for any given permutation π that $L(S') \geq L(S)$ and $R(S') \leq R(S)$, it is clear that $L_{\text{tot}}(S') \geq L_{\text{tot}}(S)$ and $R_{\text{tot}}(S') \leq R_{\text{tot}}(S)$. For Rita's preference $(2n, \dots, 1)$, the items will be chosen in reverse order. Exchanging s_k and s_{k+1} exchanges items $2n - k + 1$ and $2n - k$. Other choices will be the same for both players. Hence $R(S) + 1 = R(S')$ and $L(S) = L(S') + 1$, so $L(S') < L(S)$ and $R(S') > R(S)$ for this preference. Thus $L_{\text{tot}}(S') < L_{\text{tot}}(S)$ and $R_{\text{tot}}(S') > R_{\text{tot}}(S)$. \square

Clearly $LL \dots RR$ has the highest advantage for Luis and $RR \dots LL$ for Rita. We now know that increasing rank for choice sequences improves Rita's outcome and has the opposite effect on Luis's outcome. While it is not always the case that

an incomparable sequence with higher rank has lower L_{adv} , or even L_{tot} , we can guarantee that this will happen in a limited case: consider any policy S and call its inverse S^{-1} the sequence for which occurrences of L and R are reversed. For example, if $S = LLRRLR$, then $S^{-1} = RRLLRL$. It is true that if S and S^{-1} are sequences such that a series of $LR \rightarrow RL$ moves from S can result in S^{-1} , then any sequence S' constructed from S by some of the same moves will be fairer than S .

Theorem 8. *Consider a policy S such that S^{-1} can be reached by a series of n $LR \rightarrow RL$ moves from S . Construct policy S' by a series of m of the same moves, with $0 < m < n$, so that S' is contained on a path between S and S^{-1} . Then S' is necessarily fairer than S .*

For example,

$LLRRLR$, $LRLRLR$, **$RLRLRL$** , **$RLRLRL$** , **$RLRLRL$** , **$RLRLRL$** , **$RLRLRL$**

is such a sequence from a word to its inverse, where we have bolded the elements moved. Our theorem says that any word within this sequence will be fairer than a word on the ends of the sequence.

Proof. Recall Lemma 1: formally, $L_{\text{adv}}(S) = R_{\text{adv}}(S^{-1})$. Inverting the positions of each player simply swaps their role in procedure, so the truth of this lemma is straightforward.

We now have $L_{\text{adv}}(S^{-1}) = -L_{\text{adv}}(S)$. Since S^{-1} can be reached from S by a series of $LR \rightarrow RL$ moves, we know that $L_{\text{adv}}(S^{-1}) < L_{\text{adv}}(S)$. Since S' can be reached from S and S^{-1} can be reached from S' by a series of such moves, it must hold that $L_{\text{adv}}(S^{-1}) < L_{\text{adv}}(S') < L_{\text{adv}}(S)$. Then since $L_{\text{adv}}(S^{-1}) = -L_{\text{adv}}(S)$, we have $|L_{\text{adv}}(S')| < |L_{\text{adv}}(S)|$, meaning S' is a fairer policy. \square

Because the two alternating sequences are the inversions of each other and can be reached by exchanging every two positions, any sequence on the poset paths between them will be fairer than either. Thus we have the following:

Corollary 9. *Any choice sequence lying in the sequence poset strictly between $LRLR \dots LR$ and $RLRL \dots RL$ is fairer than either of these.*

We now know that a policy which covers another is better for Rita than the latter, and as we would expect, this is generally the case for sequences of higher ranks which are not comparable in the poset. (It does fail occasionally; that is, if λ is of higher rank than σ but does not cover σ , it is possible for R_{tot} to be smaller, though this is not usually the case. For example, this happens with $LLRRRL$ (associated partition $(3, 3, 0)$) and $LRRLRL$ (associated partition $(2, 2, 1)$), and this is the only such pair in the 3×3 box.) What about policies of equal rank?

We have much numerical evidence for an intriguing conjecture we were unable to prove in full generality: that L_{adv} could be associated easily to the dominance order among partitions in a rank.

Conjecture 10. *Among choice sequences σ_i of the same rank, if S_1 dominates S_2 , then S_1 has higher average advantage for Luis than S_2 .*

This has been verified computationally for all ranks of all posets in boxes of sizes up to 10×10 , but not proven generally. We have, however, been able to prove this for a restricted case:

Theorem 11. *If two choice sequences in the Boolean set consist of a prefix α and a suffix β connected by $LRRL$ and $RLLR$ respectively, i.e., $\sigma_1 = \alpha LRRL\beta$ and $\sigma_2 = \alpha RLLR\beta$, then Luis's advantage is strictly greater for σ_2 than for σ_1 .*

Thus, a partition in the Boolean set that dominates another by dominance moves of adjacent, nonoverlapping pairs — in the partition, by moves of one square at a time in the Ferrers board to the next part up in the partition — is better for Luis than the latter. A major difficulty in initially establishing this theorem was that it is *not* the case that Luis's position always worsens going from the former to the latter! Rather, even if Luis's position betters, Rita's does also, and by more.

Proof. The values expected to be obtained by each player as the policy progresses through β do not change; denote these by B_1 and B_2 respectively. Say that β has length r ; it is a straightforward application of Lemma 4 to obtain the expected values of Luis and Rita before and after the change in the connecting word:

$$\begin{aligned}\bar{u}_L(LRRL\beta) &= r + 4 + \frac{r+4}{r+3} \left(\frac{r+3}{r+2} (r+1+B_1) \right) \\ \bar{u}_L(RLLR\beta) &= \frac{r+5}{r+4} \left(2r+5 + \frac{r+2}{r+1} B_1 \right) \\ \bar{u}_R(LRRL\beta) &= \frac{r+5}{r+4} \left(2r+5 + \frac{r+2}{r+1} B_2 \right) \\ \bar{u}_R(RLLR\beta) &= r + 4 + \frac{r+4}{r+3} \left(\frac{r+3}{r+2} (r+1+B_2) \right)\end{aligned}$$

Observe that, given an expected value x that holds as one enters a segment α , Lemma 4 gives that the expected value after exiting α will be some linear function of x . This function will be different for players 1 and 2: say that Luis experiences function $Ax + B$ when passing through α , and Rita experiences $Cx + D$.

Thus we have

$$\begin{aligned}(\bar{u}_L(\alpha LRRL\beta) - \bar{u}_R(\alpha LRRL\beta)) - (\bar{u}_L(\alpha RLLR\beta) - \bar{u}_R(\alpha RLLR\beta)) \\ = (A+C) \frac{r-2}{(r+2)(r+4)} + (AB_1 + CB_2) \frac{-4}{(r+2)(r+4)(r+1)}.\end{aligned}$$

Multiplying through by $(r + 4)(r + 2)$, we find that we wish to show that

$$\frac{AB_1 + CB_2}{A + C} > \frac{1}{4}(r - 2)(r + 1).$$

This statement will be true if the stronger inequality holds:

$$\frac{AB_1 + CB_2}{A + C} \geq \frac{1}{4}r^2.$$

Since we specified that the policy being studied was in the Boolean set, an $LRRL$ can only occur with an even number of places remaining, in which each player has at least one choice in every two adjacent places, and so the minimum possible value of either B_1 or B_2 is

$$1 + 3 + 5 + \cdots + (2(\frac{1}{2}r) - 1) = (\frac{1}{2}r)^2.$$

Thus, the ratio $(AB_1 + CB_2)/(A + C)$ would be reduced by taking the greater of B_1 and B_2 and reducing it to the lesser, giving a ratio above the threshold required. Hence, an adjacent dominance move $LRRL \rightarrow RLLR$ on a choice sequence in the Boolean set improves Luis's advantage. \square

We may observe that this theorem completely characterizes relative relations within ranks in the Boolean set, since any two Boolean set choice sequences with the same rank can be related by adjacent dominance moves. A more general conjecture, which might make a useful intermediate step toward the full conjecture, would be to establish Luis's advantage improvement for dominance by exchange of exactly one position at any distance, regardless of whether a partition was in the Boolean set. This would cover policies $S = \alpha LR\beta RL\gamma$ and $S' = \alpha RL\beta LR\gamma$. The method of proof applied above appears to be insufficient to establish the full conjecture without further insight. We remain interested in the question and invite interested readers to attempt the proof.

Remark. We could certainly have shown Theorem 2 using Lemma 4, but this would have given no information about the exchange made. A similar analysis here shows that for a dominance move, Luis and Rita will either exchange two pairs of items in two nonoverlapping intervals, or a single pair of items. The problem with using this to prove the dominance theorem is that the single exchange may leave Luis worse off. For instance, if Rita's preference is 4312, then in $LRRL$ Luis receives 42, but in $RLLR$ Luis receives 32. So unlike in our previous theorems, there exist some cases in which Luis suffers the opposite of the general effect. Thus, establishing the theorem by these methods would oblige us to estimate the relative frequency of various sizes of exchange — a task we found to be quite difficult.

3.1. Search strategies and heuristics. We can now make a reasonable suggestion for a fair sequence. Due to Theorem 8, one would suspect that a policy near the middle rank would be close to $L_{\text{adv}}(S) = 0$: one suggestion would be the policy known as balanced alternation, or the Thue–Morse sequence, which is an initial segment of

LRRL RLLR RLLR LRRL RLLR LRRL LRRL RLLR RLLR

This will certainly be fairer than either alternating sequence. By rank it lies halfway between the two, and it is near midway between the two extremes of its rank in dominance order, so by our theorems it would reasonably be expected to have L_{adv} close to 0.

Ideally, if our only priority is finding the fairest possible choice sequence, we would like to be able to take as input the length $2n$ of the item set and with a short algorithm place Luis’s choices $\{\ell_1, \dots, \ell_n\}$. We are very far from achieving this, but with the help of the above theorems we can reduce the work considerably from examining all possible choice sequences.

- (1) The poset of partitions in the box of size N has either 1 or 2 middle ranks, depending on whether N is even or odd. Calculate L_{adv} for one such rank; the higher, if N is odd.
- (2) Move up one rank, ignoring choice sequences associated to partitions that cover any that already have negative L_{adv} , and calculate again.
- (3) Repeat until all L_{adv} are nonpositive or an $L_{\text{adv}} = 0$ is found.
- (4) At this point, stop and select the choice sequence with lowest $|L_{\text{adv}}|$. This sequence and its inverse will be the fairest choice sequences for $2n$ items.

If the dominance conjecture were completely true, then a binary search could be run in each rank for the fairest sequence, reducing the work by a factor of $\log_2 n$; if only a sequence in the Boolean set is desired, this can definitely be done.

4. Formulas for expected values

Recall that an *outcome* for Luis is a set of items he receives, and a *path* associated to a given policy and Rita’s preference is the order in which items are taken by both players, as labeled by Luis. Our first theorem in this section gives us a formula, given a particular choice sequence and Luis’s outcome, for the number of paths which yield this outcome, or equivalently, the number of possible sequences of Luis-labeled items that Rita might take under the given conditions. It uses the falling factorial notation

$$(x)_j = x(x-1) \dots (x-(j-1)).$$

Theorem 12. *The number of paths in which Luis has choice positions $\{\ell_1, \dots, \ell_n\}$ and takes values $\{v_1, \dots, v_n\}$ is given by*

$$(\ell_1 - 1)_{\ell_1 - 1} \left(\prod_{j=2}^n (\ell_j + v_{j-1} - 2n - 2)_{\ell_j - \ell_{j-1} - 1} \right) (v_n - 1)_{2n - \ell_n}.$$

Proof. The problem reduces to a question of counting the number of ways to fill each column and row exactly once, in a Ferrers board given by Luis's choices. Consider, for example, the case of length $2n = 10$ in which Luis chooses at positions $\{2, 5, 6, 8, 9\}$ and takes values $\{9, 8, 5, 3, 2\}$:

10	○		■	■			■			■
9		●								
8					●					
7	○		○	○			■			■
6	○		○	○			■			■
5						●				
4	○		○	○			○			■
3								●		
2									●	
1	○		○	○			○			○
	1	2	3	4	5	6	7	8	9	10

Here black circles represent Luis's definite choices, white circles Rita's possibilities, and black squares places forbidden by Luis's choices. Of course, Rita may not choose an item later and higher-valued than a choice of Luis's, else he would have taken this item on an earlier turn in preference to one he selected. It is also easy to observe that Rita must have chosen, say, item 10 at place 1 because Luis chose item 9 at place 2, but our placement of circles is not that keen yet: we merely take all rows and columns not occupied by a Luis choice which are not later and higher than a Luis choice.

Thus, among rows and columns that Luis does not occupy, Rita can take any collection of objects that includes exactly one item in each row and column in the open positions left of and below Luis's choices. Since she may have any preference among the items so chosen, these may come in any order so long as they satisfy the previous conditions.

Such a problem is referred to as counting *full rook placements* within the Ferrers board of shape consisting of the spaces below and left of Luis's choices, in the columns and rows that they do not occupy. This is a standard counting problem, the method for which may be found on page 74 of Stanley's *Enumerative Combinatorics*, Volume 1 [Stanley 1997]. We state here a theorem from that volume for reference:

Theorem 13 [Stanley 1997, Theorem 2.4.1]. *Let $\sum_{k=0}^m r_k x^k$ be the rook polynomial of the Ferrers board B of shape (b_1, \dots, b_m) . Set $s_i = b_i - i + 1$. Then*

$$\sum_{k=0}^m r_k (x)_{m-k} = \prod_{i=1}^m (x + s_i).$$

The constant term of this polynomial, $\prod s_i$, is precisely r_m , the number of ways to place m nonattacking rooks on the board. This is the number we desire.

Our Ferrers board has n parts b_1 through b_n . In order to use the formula of [Stanley 1997], we name parts in ascending order of size, hence the reverse order of their appearance in the choice sequence.

We observe that we have:

- $\ell_1 - 1$ parts of size $2n - n = n$ ($2n$ values, n occupied),
- $\ell_2 - \ell_1 - 1$ parts of size $v_1 - 1 - (n - 1) = v_1 - n$,
- $\ell_3 - \ell_2 - 1$ parts of size $v_2 - 1 - (n - 2) = v_2 - n + 1$,
- \vdots
- $\ell_n - \ell_{n-1} - 1$ parts of size $v_{n-1} - 1 - (1) = v_{n-1} - 2$,
- $2n - \ell_n$ parts of size $v_n - 1$.

Thus among the s_i are $2n - \ell_n$ values $v_n - 1, v_n - 2, v_n - 3, \dots, v_n - (2n - \ell_n - 1)$. The product of these is the falling factorial $(v_n - 1)_{2n - \ell_n}$.

The next s_i start with $s_{2n - \ell_n + 1}$. The associated b_i are all $v_{n-1} - 2$, and there are $\ell_n - \ell_{n-1}$ of them. The s_i thus produced are

$$\begin{aligned} v_{n-1} - 2 - (2n - \ell_n + 1) + 1 &= \ell_n + v_{n-1} - 2n - 2, \\ v_{n-1} - 2 - (2n - \ell_n + 2) + 1 &= \ell_n + v_{n-1} - 2n - 3, \\ &\vdots \end{aligned}$$

$$v_{n-1} - 2 - (2n - \ell_n + \ell_n - \ell_{n-1} - 1) + 1 = \ell_{n-1} + v_{n-1} - 2n.$$

The product of these s_i is the falling factorial $(\ell_n + v_{n-1} - 2n - 2)_{\ell_n - \ell_{n-1} - 1}$.

The other falling factorials in the product arise similarly. \square

Rita may have multiple preferences that give rise to a particular path. For a simple example, in the choice sequence LR , it does not matter whether Rita's preferences are 12 or 21; items will be taken in the sequence 2, 1. The number of Rita preferences that give rise to any path is a constant that depends only on the position of the choices in the sequence, and not on the specific path:

Theorem 14. *Consider a choice sequence S where Luis's choice positions are $\{\ell_1, \dots, \ell_n\}$, and a specific path through S given by $s = s_1, s_2, \dots, s_{2n}$. Then the number of possible preference permutations for Rita resulting in path s through S is $\prod_{j=1}^n (2n - \ell_j + 1)$.*

Proof. Take any S and any path s through S , and construct π by placing into it the items as they occur in s . On each of Rita's turns i , the item r_i she selected must always be placed in the leftmost available position in π since r_i was her most preferred item of those remaining. On each of Luis's turns j , the item l_j he chose may be placed in any of the remaining positions in π since, regardless of its value to Rita, she will not have a chance to take the item after Luis has already taken it. The number of available positions on Luis's turn j is equal to $2n - \ell_j + 1$, where ℓ_j refers to the position of Luis's turn j in the original sequence S . Thus the total number of permutations for a given path s through S is $\prod_{j=1}^n (2n - \ell_j + 1)$. \square

Since the number of preference permutations associated with a given path through a sequence S is dependent only on S , and thus is constant across all paths through S , we can count outcomes by grouping them according to the resulting path.

Suppose Luis's positions are (ℓ_1, \dots, ℓ_n) . Take each possible Luis outcome, in which his values (v_1, \dots, v_n) can range from a maximum of $2n + 1 - i$ for v_i (Rita took no higher-ranked items than his most-preferred) to a minimum of $2n + 1 - \ell_i$ (Rita always took Luis's next-preferred item at her choices). To each outcome we can calculate the number of paths and the number of Rita preferences that give that path. We total Luis's value and sum over all possible outcomes for this choice sequence to get L_{tot} .

Thus, combining Theorems 12 and 14, we have a formula for the total value of Luis's outcomes over the set of all Rita preferences, using a given choice sequence S :

$$L_{\text{tot}} = \left(\prod_{j=1}^n 2n - \ell_j + 1 \right) \sum_{\substack{(v_1, v_2, \dots, v_n) \\ \min(2n+1-i, v_{i-1}-1) \geq v_i \geq 2n+1-\ell_i}} \left(\sum v_n \right) \\ \times (\ell_1 - 1)_{\ell_1-1} \left(\prod_{j=2}^n (\ell_j + v_{j-1} - 2n - 2)_{\ell_j - \ell_{j-1} - 1} \right) (v_n - 1)_{2n - \ell_n}. \quad (1)$$

Dividing by $(2n)!$ gives Luis's expected value.

A second approach to counting outcomes involves making a tree diagram for the possible outcomes with the assumption that Rita's choices are made randomly.

To do this, we let Rita's preferences be arbitrary. Since all possible preferences are considered equally likely, it is also the case that on any of her turns, Rita is equally likely to take any one of the available items, and it is valid to imagine that on each turn she chooses one item at random. Then we can represent the problem using a tree, the nodes of which will contain all possible actions for a turn.

Theorem 15. *The total number of Rita preferences that result in a particular outcome $\{v_1, \dots, v_n\}$ for Luis is $(2n)! P$, where $2n$ is the number of items and P , the probability that the outcome occurs, is equal to the number of paths giving a*

particular outcome for Luis divided by the total number of paths, or

$$\frac{(\ell_1 - 1)_{\ell_1 - 1} \left(\prod_{j=2}^n (\ell_j + v_{j-1} - 2n - 2)_{\ell_j - \ell_{j-1} - 1} \right) (v_n - 1)_{2n - \ell_n}}{\prod_{i=1}^n 2n - r_i + 1},$$

where r_i is the position of Rita's i -th turn in the overall sequence.

Proof. Considering the problem as a tree with Rita's preferences unknown, we know that on each of his turns, Luis will always take the highest-numbered item, and on Rita's turns, she will take any one of the remaining items with equal probability of each. Thus none of Luis's turns will generate additional branches, but on each of Rita's turns, a branch is necessary for each of the remaining items. The number of the remaining items at Rita's turn i is $2n - r_i + 1$. The total number of paths in the tree is then the product of this value over all of her turns, $\prod_{i=1}^n (2n - r_i + 1)$. From Theorem 12, we know that the number of paths giving a particular outcome is $(\ell_1 - 1)_{\ell_1 - 1} \left(\prod_{j=2}^n (\ell_j + v_{j-1} - 2n - 2)_{\ell_j - \ell_{j-1} - 1} \right) (v_n - 1)_{2n - \ell_n}$. Dividing them gives P , the probability the outcome occurs.

Multiplying with P the total number of possible preferences for Rita, $(2n)!$, yields the number of Rita preferences that result in a given outcome. \square

5. Conjectures and open problems

There are a number of open questions that interested researchers from student to faculty might be able to consider for this problem.

Data is often a good start. We begin with the collection of known, guaranteed fairest choice sequences; see Table 1. We list sequences with the Luis-first version; where this gives Luis a negative L_{adv} , the sequence is marked with an asterisk. If the fairest known sequence is not one of those that lies between the alternating sequences, the fairest of those in the Boolean set is given.

length	fairest known	fairest between alternating
2	LR	
4	$LRRL^*$	
6	$LRLRRL$	
8	$LLRRRLRL$	$LRRLRLLR$
10	$LRRLRLRRL$	
12	$LLRRRLLRLLR^*$	$LRLRLRLRRLR^*$
14	$LLRRRLLRRLRL$	$LRRLRRLRLRLRL$
16	$LRLLLLRRRRRRLLL$	$LRRLRLRLRLRLRLR^*$
18	$LRLLRRRRRLRLRLLR$	$LRLLRRLRLRLRLRLRL$

Table 1. The collection of known and guaranteed fairest choice sequences.

Let us pause for a few remarks on Table 1.

The fairest choice sequences seem to be rather generally not within the Boolean set—instead, they seem to be close to $LL \dots RRRR \dots LL$, with half the L 's at the front and back of the sequence. That seems quite surprising and counterintuitive. After all, the simple alternating sequence $LRLR \dots$ maximizes social welfare, and balanced alternation $LRRLRLLR \dots$ has good heuristic arguments for being a relatively fair sequence. Both distribute L and R relatively evenly throughout the sequence. A sequence that “chunks” the players significantly would be very different from these.

Actually using such a choice sequence to divide items would severely strain the assumptions that Luis's and Rita's preferences are independent, and that valuations are linear: Luis would be collecting a quarter of the items before Rita gets a chance to take any of her most preferred items. We assumed no correlation of preferences and items valued in even intervals, but if there is any agreement between Luis and Rita on a small subset of highly valuable items, Luis would be able to seize these immediately. On the other hand, if there is agreement on a small subset of highly undesirable items Luis would also be left with these, so perhaps the distribution would work out. From a strictly mathematical viewpoint, however, it is certainly of intrinsic interest to know if such a sequence is the “typical” fairest sequence.

As mentioned earlier, it is of interest to determine how dominance interacts with other conditions studied in this area, such as the min-max and social welfare conditions described in the introduction. From [Bouveret and Lang 2011] we have a most interesting datum. Bouveret and Lang study policies under min-max (which they refer to as egalitarian), i.e., the policy for which the worse-off player's expected value is highest. In [Bouveret and Lang 2011, Table 1], they list the optimal policies for even lengths up to twelve items. The min-max optimal policy for an even number of items in their listing turns out to be within the Boolean set. It is not always the fairest, but rather, the following appears to be the case:

Conjecture 16. *The fairest policy among those in the Boolean set is the min-max optimal policy.*

It is plausible that a policy where the disadvantaged player does well is a relatively fair policy, and Bouveret and Lang establish that an alternating policy tends toward egalitarian optimality as length grows, so there seems to be multiple pieces of evidence that this conjecture is reasonable. It would be interesting if it turned out to hold.

As a perhaps trivial but astonishing note, we find that for length 14, the fairest choice sequence has L_{adv} exactly 0! Compare this to the alternating sequence, which for length 14 has $L_{\text{adv}} \approx 3.95$, or Luis being advantaged by more than half the number of items each player takes.

An interested investigator might be able to improve (1) by converting the falling factorials into binomial coefficients, and repeatedly applying Abel-type summation identities which sum shifts of binomials. The recurrence of Kalinowski, Narodytska and Walsh is far more useful than this formula, but a closed form might reverse the situation.

Finally, we recall our conjecture on dominance, which in its full generality remains open and which we consider a most intriguing question regarding the fairness condition:

Conjecture 10. *Among choice sequences σ_i of the same rank, if σ_1 dominates σ_2 , then σ_1 has higher average advantage for Luis than σ_2 .*

Partial approaches might include extending the theorem to dominance moves $LR \dots R \dots RL \rightarrow RL \dots R \dots LR$, with bounds on the difference in number of L and R before and after the changing segment.

References

- [Bouveret and Lang 2011] S. Bouveret and J. Lang, “A general elicitation-free protocol for allocating indivisible goods”, pp. 73–78 in *Proc. of 22nd Int. Joint Conference on Artificial Intelligence* (Barcelona, 2011), vol. 1, edited by T. Walsh, AAAI Press, Menlo Park, CA, 2011.
- [Brams and Taylor 1996] S. J. Brams and A. D. Taylor, *Fair division: from cake-cutting to dispute resolution*, Cambridge University Press, 1996. MR Zbl
- [Brams and Taylor 2000] S. J. Brams and A. D. Taylor, *The win-win solution: guaranteeing fair shares to everybody*, W. W. Norton, New York, 2000.
- [Hopkins 2010] B. Hopkins, “Taking turns”, *College Math. J.* **41**:4 (2010), 289–297. MR Zbl
- [Hopkins and Jones 2009] B. Hopkins and M. A. Jones, “Bruhat orders and the sequential selection of indivisible items”, pp. 273–285 in *The mathematics of preference, choice and order*, edited by S. J. Brams et al., Springer, 2009. MR Zbl
- [Kalinowski and Narodytska 2013] T. Kalinowski and N. Narodytska, “A social welfare optimal sequential allocation procedure”, pp. 227–233 in *Proc. of 23rd Int. Joint Conference on Artificial Intelligence* (Beijing, 2013), edited by F. Rossi, AAAI Press, Menlo Park, CA, 2013.
- [Rubchinsky 2010] A. Rubchinsky, “Brams–Taylor model of fair division for divisible and indivisible items”, *Math. Social Sci.* **60**:1 (2010), 1–14. MR Zbl
- [Stanley 1997] R. P. Stanley, *Enumerative combinatorics, I*, Cambridge Studies in Advanced Mathematics **49**, Cambridge University Press, 1997. MR Zbl

Received: 2016-07-08 Revised: 2017-12-10 Accepted: 2017-12-30

wjkeith@mtu.edu

*Department of Mathematical Sciences,
Michigan Tech University, Houghton, MI, United States*

sean.grindatti@gmail.com

*Department of Mathematical Sciences,
Michigan Tech University, Houghton, MI, United States*

Intersecting geodesics and centrality in graphs

Emily Carter, Bryan Ek, Danielle Gonzalez,
Rigoberto Flórez and Darren A. Narayan

(Communicated by Kenneth S. Berenhaut)

In a graph, vertices that are more central are often placed at the intersection of geodesics between other pairs of vertices. This model can be applied to organizational networks, where we assume the flow of information follows shortest paths of communication and there is a required action (i.e., signature or approval) by each person located on these paths. The number of actions a person must perform is linked to both the topology of the network as well as their location within it. The number of expected actions that a person must perform can be quantified by *betweenness centrality*. The betweenness centrality of a vertex v is the ratio of shortest paths between all other pairs of vertices u and w in which v appears to the total number of shortest paths from u to w . We precisely compute the betweenness centrality for vertices in several families of graphs motivated by different organizational networks.

1. Introduction

In a graph, vertices with higher centrality are often placed at the intersection of geodesics between other pairs of vertices. This model can be applied to organizational and social networks, where we assume the flow of information follows shortest paths of communication and there is a required action (i.e., signature or approval) by each person located on these paths.

A simple organizational structure can be designed using a binary tree. An example is shown in Figure 1.

We consider this structure where the CEO oversees a “left wing” and a “right wing”. We first note the employees (E1, E2, E3, and E4) do not have to perform any actions, as they are on the periphery. The CEO will have to perform actions on any correspondence between people in different wings, for a total of 18 possible actions. Vice presidents VP1 and VP2 will have to perform actions on the correspondence between the two employees under them as well as correspondences between their two

MSC2010: 05C12, 05C82.

Keywords: betweenness centrality, shortest paths, distance.

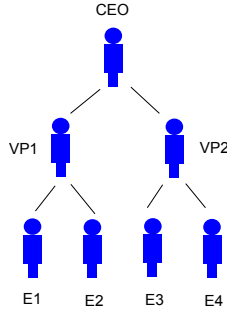


Figure 1. A binary tree organizational structure.

employees and anyone else in the company. This is a total of $2(1 \times 1) + 2(2 \times 4) = 18$ actions. Ironically, in this model, which would appear at first to distribute the work according to rank, the VPs actually have to perform as many actions as the CEO.

We next consider a ternary tree model where each person (except for the employees) oversees three people; see Figure 2. This slight change puts the most amount of work in the hands of the CEO. The CEO has to perform 96 actions, while each of the VPs perform 60 actions, and again the employees are not responsible for any actions.

The determination of the number actions is more complicated if the network contains cycles, since there can be multiple shortest paths, which can be used with equal probability. For our next example we will use the organizational network shown in Figure 3.

Consider the number of actions that B must make. Person B acts on communication between persons A and D and D and A. However A and D could choose to route their correspondence through person C rather than B. So B will only appear on half of the four shortest paths between A and D and D and A. (We will consider these paths to be equally likely to be followed.) Thus the total number of expected

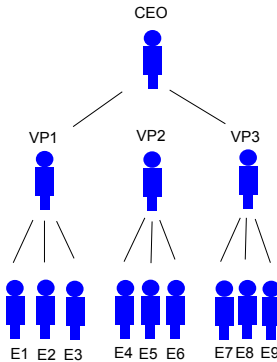


Figure 2. A ternary tree model.

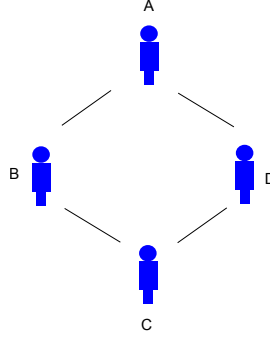


Figure 3. A network with a cycle.

actions by person B is 2, which is the same, by symmetry, for persons A, C, and D. This creates a balance of actions over every employee.

The expected number of actions can be quantified by *betweenness centrality*. This concept was introduced in [Freeman 1977] in the context of social networks. This concept has appeared frequently in both network and neuroscience literature [Brandes et al. 2016; Bullmore and Sporns 2009; Freeman et al. 1991; Guye et al. 2010; Pandit et al. 2013; White and Borgatti 1994]. The betweenness centrality of graphs was computed for various families of graphs including complete bipartite graphs, Cartesian products, wheel graphs, cocktail party graphs, ladder graphs, and cycles [Kumar and Balakrishnan 2016; Kumar et al. 2014].

In this paper, we determine the betweenness centrality for several other families of graphs motivated by organizational networks.

We first give some background with some elementary results.

Definition 1. The *betweenness centrality* of a vertex v , denoted $bc(v)$, measures the frequency at which v appears on a shortest path between two other distinct vertices x and y . Let σ_{xy} be the number of shortest paths between distinct vertices x and y , and let $\sigma_{xy}(v)$ be the number of shortest paths between x and y that contain v . Then

$$bc(v) = \sum_{x,y} \frac{\sigma_{xy}(v)}{\sigma_{xy}}$$

(for all distinct vertices x and y).

In our first lemma, we restate an elementary result on the lower and upper bounds of the betweenness centrality of a vertex. This was found by Gago et al. [2012] and Grassi et al. [2009].

Lemma 2. For a given graph G with n vertices, $0 \leq bc(v) \leq (n-1)(n-2)$ for all vertices v in G . Furthermore these bounds are tight.

It is clear that if a vertex has a betweenness centrality of zero, it means that the vertex is likely to be less vital to the network than a vertex with a higher betweenness centrality. Gago et al. [2012] and Grassi et al. [2009] provided a classification for vertices to have a betweenness centrality of zero. We restate this as our next lemma. We recall that the *closed neighborhood of a vertex* is the subgraph induced by a vertex and its neighbors.

Lemma 3. *Given a vertex v , we have $bc(v) = 0$ if and only if the closed neighborhood of v forms a complete subgraph.*

2. Betweenness centrality

We now investigate the betweenness centralities of vertices in several families of graphs, including star-like graphs, k -ary trees, complete multipartite graphs, and powers of paths and cycles. The following lemma can be implicitly found in [White and Borgatti 1994].

Lemma 4. *Let P_n be a path on vertices v_1, v_2, \dots, v_n . Then $bc(v_i) = 2(i-1)(n-i)$.*

Next we investigate complete multipartite graphs, making a small extension of known results for complete bipartite graphs [Kumar et al. 2014]. The complete multipartite graph K_{n_1, n_2, \dots, n_t} for $t \geq 2$ is the graph where the vertex set is partitioned into t partite sets V_1, V_2, \dots, V_t such that $|V_i| = n_i$ for each $1 \leq i \leq t$ and uv is an edge if and only if u and v belong to different partite sets.

In an application with personnel, people are divided into different groups where there are no direct connections among people in the same group, but there are direct connections between each pair of people in different groups. We give an example of a graph in Figure 4 where there are three vertices in one part, four in a second part, and five in a third part. This graph is denoted by $K_{3,4,5}$. The vertices with the highest betweenness centrality will be in the part of size 3 (since there will be the largest number of shortest paths routed through them) and the vertices with the

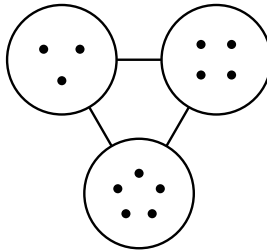


Figure 4. The complete multipartite graph $K_{3,4,5}$. The lines indicate that every vertex in one part is adjacent to every vertex in a different part.

lowest betweenness centrality will be in the part of size 5 (since they will have the smallest number of shortest paths routed through them). We explore this problem for the general class in our next lemma.

Lemma 5. *Let G be the complete multipartite graph K_{n_1, n_2, \dots, n_t} where the vertices in part i are $v_{i,1}, v_{i,2}, \dots, v_{i,n_i}$ for all $1 \leq i \leq t$. Then for all $1 \leq j \leq n_i$.*

$$\text{bc}(v_{i,j}) = \sum_{k=1, k \neq i}^t \frac{\binom{n_k}{2}}{\sum_{r=1, r \neq k}^t n_r}$$

Proof. We will compute $\text{bc}(v_{i,j})$. Consider the shortest paths between vertices $v_{x,y}$ and $v_{x,z}$. We first determine the total number of shortest paths that contain $v_{i,j}$. Let V_1, V_2, \dots, V_t represent the partite sets K_{n_1, n_2, \dots, n_t} . To determine the total number of shortest paths containing $v_{i,j}$ we count the number of pairs of distinct vertices in each part A_k where $k \neq i$, and divide by the number of vertices in $V(G) - A_i$. \square

2.1. Complete and balanced k -ary trees. In a complete and balanced binary trees, there is a root vertex that is adjacent to exactly two other vertices. These vertices then have two “children” vertices. Let $k \geq 2$. A balanced k -ary with t levels will have k^i vertices at the i -th level for all $0 \leq i \leq t-1$. We generalize the class of trees found in the Introduction to include k -ary trees. Here there is a root vertex that has k neighbors and each of these k neighbors have k children. In balanced k -ary trees with t levels there will be k^i vertices at the i -th level for all $0 \leq i \leq t-1$.

We next determine the betweenness centrality of vertices in a k -ary tree.

Theorem 6. *Let G be a complete and balanced k -ary tree with levels $0, 1, \dots, t-1$. Let v_j be a vertex on level j . Then*

$$\text{bc}(v_j) = -\frac{k^{t-j-1} - 1}{(k-1)^2} (k - k^{t+1} - k^{t-j} + k^{t-j-1} + k^{t-j+1} - 1).$$

Proof. Consider a complete and balanced k -ary tree with levels $0, 1, \dots, t-1$. This tree will have $1 + k + k^2 + \dots + k^{t-1} = (k^t - 1)/(k - 1)$ vertices. We note that vertices in the same level will have the same betweenness centrality, so we will use v_i to denote a vertex on level i . Note that vertex v_j has k sets of $(k^{t-j} - 1)/(k - 1)$ vertices beneath it. The paths that pass through v_j will either go between vertices beneath v_j in different subparts, or between any of these vertices and other vertices in the graph besides v_j . Hence

$$\begin{aligned} \text{bc}(v_j) &= \left(\frac{k^{t-(j+1)} - 1}{k-1} \right) (k-1) \left(\frac{k^{t-(j+1)} - 1}{k-1} \right) \\ &\quad + k \left(\frac{k^{t-(j+1)} - 1}{k-1} \right) \left(\frac{k^t - 1}{k-1} - k \left(\frac{k^{t-(j+1)} - 1}{k-1} \right) - 1 \right) \\ &= -\frac{k^{t-j-1} - 1}{(k-1)^2} (k - k^{t+1} - k^{t-j} + k^{t-j-1} + k^{t-j+1} - 1). \end{aligned} \quad \square$$

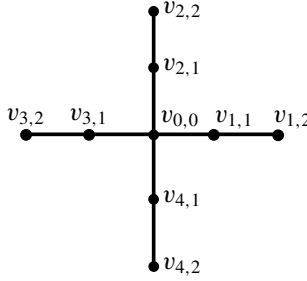


Figure 5. A star-like graph.

2.2. Star-like graphs. As shown earlier, star graphs include vertices with the highest and lowest possible betweenness centrality values. We next expand the investigation to include graphs that are obtained by subdividing the edges of the star graph $K_{1,n-1}$. These graphs will have a “center” and “pendant spokes”. This model appears in a organization where communication moves along different lines that meet at a central processing person. This model is also found frequently in airports where concourses (with multiple gates along them) intersect at a common location. An example of a star-like graph is given in Figure 5.

It is clear that the center $v_{0,0}$ has the highest betweenness centrality, and the betweenness centrality of vertices will be less if they are located farther away from the center. We address the general problem in our next theorem.

Theorem 7. *Let G be a subdivided star graph with the center vertex $v_{0,0}$. Let the m paths pendant to the center have lengths s_1, \dots, s_m and let $v_{l,k}$ be the k -th vertex from $v_{0,0}$ on the l -th pendant path. Then*

$$\text{bc}(v_{0,0}) = 2 \sum_{j=2}^m s_j \sum_{i=1}^{j-1} s_i \quad \text{and} \quad \text{bc}(v_{l,k}) = 2(s_l - k) \left(\sum_{i \neq l} s_i + k \right).$$

Proof. The center vertex will lie on optimal paths between two vertices if and only if the path connects vertices on different spokes. Thus it suffices to sum the number of pairs of vertices between spokes. Vertices on a spoke will lie on an optimal path if and only if the path is between a vertex further along the same spoke ($s_l - k$ vertices) and a vertex closer to the center or on a different spoke yielding $k + \sum_{i \neq l} s_i$ vertices. Finally we double the product to account for paths in either direction. \square

Theorem 8. *Let G be a triangle graph with vertices $v_{0,0}$, $v_{1,0}$, and $v_{2,0}$ and pendant paths of lengths s_0 , s_1 , and s_2 incident to the three vertices, respectively. If $v_{l,k}$ is the k -th vertex on the l -th pendant path then*

$$\text{bc}(v_{l,k}) = 2(s_l - k) \left(k + 2 + \sum_{i \neq l} s_i \right).$$

Proof. The vertex $v_{l,k}$ will be on an optimal path if and only if the path is between a vertex on the l -th pendant path further from the triangle ($s_l - k$ vertices) and a vertex closer to the triangle or on a different pendant path, giving a total of $k + 2 + \sum_{i \neq l} s_i$ vertices. Finally, we double this product to account for both directions. \square

2.3. Powers of cycles and paths. We next consider cycles that have “redundant” connections. Consider a network of 20 people where there are direct links between adjacent people and also links between people that are spaced two apart. We will explore the betweenness centrality of this class of networks.

Recall that the k -th power of a graph G is denoted by G^k , which is defined as follows: $V(G^k) = V(G)$ and $v_i v_j \in E(G^k)$ if and only if the distance between v_i and v_j in G is less than or equal to k . Next, we investigate the betweenness centrality of vertices in powers of cycles.

Theorem 9. *Let $G = C_n^m$ with $n > 2m + 1$ and let $d = \text{diam}(C_n^m) = \lceil (n-1)/(2m) \rceil$. Then*

$$\begin{aligned} \text{bc}(v) &= (d-1)(2\lceil \tfrac{1}{2}(n-1) \rceil - d) - (n - (n-1))(\lceil \tfrac{1}{2}(n-1) \rceil - 1) \\ &= d - \lceil \tfrac{1}{2}n - \tfrac{1}{2} \rceil - 2\lceil \tfrac{1}{2}n - \tfrac{1}{2} \rceil + 2d\lceil \tfrac{1}{2}n - \tfrac{1}{2} \rceil - d^2 + 1. \end{aligned}$$

Proof. Let $G = C_n^m$ with $n > 2m + 1$. Let $r \equiv -\lceil \tfrac{1}{2}(n-1) \rceil \pmod{m}$ such that $m > r \geq 0$. Then $r = dm - \lceil \tfrac{1}{2}(n-1) \rceil$. The maximum number of intermediate vertices on any path is $d-1$. Let P_l be the set of shortest paths of length l where $m+1 \leq l \leq d$. The number of intermediate vertices in each shortest path is $\lceil l/m \rceil$. Let s be the number of internal vertices on a shortest path between two vertices where $1 \leq s \leq d-1$. For each path with length l , where $s = \lceil l/m \rceil - 1$ internal vertices are placed at particular locations, there exist s pair(s) of vertices where the path includes v . In the betweenness centrality this accounts for s terms equal to $1/|P_l|$. Since we can reverse any of these paths, this number is doubled. Then counting all paths of length l , the betweenness centrality for v will be

$$2|P_l| \cdot \frac{s}{|P_l|} = 2s.$$

Summing over all values of l gives

$$\sum_{l=m+1}^{\lceil (n-1)/2 \rceil} 2\left(\left\lceil \frac{l}{m} \right\rceil - 1\right) = (d-1)(2\lceil \tfrac{1}{2}(n-1) \rceil - dm)$$

when n is not divisible by m . When n is divisible by m there will be two paths of the same distance between vertices that are diametrically opposite on the cycle. Hence the final term in the summation must be divided by 2, which yields

$$(d-1)(2\lceil \tfrac{1}{2}(n-1) \rceil - dm) - \left(\left\lceil \frac{\lceil \tfrac{1}{2}(n-1) \rceil}{m} \right\rceil - 1\right).$$

The betweenness centrality values for the two cases can be combined into a single function,

$$c(v) = (d-1)(2\lceil \frac{1}{2}(n-1) \rceil - dm) - \left(\left\lceil \frac{n}{m} \right\rceil - \left\lceil \frac{n-1}{m} \right\rceil \right) \left(\left\lceil \frac{\lceil \frac{1}{2}(n-1) \rceil}{m} \right\rceil - 1 \right). \quad \square$$

2.3.1. Powers of paths. The problem of determining the betweenness centrality of vertices in a power of a path is considerably more difficult than powers of cycles. In the case of cycles, all of the vertices have the same betweenness centrality, but in a path the betweenness centrality of a vertex is dependent upon its location in the path. The m -th power of a path P_n is denoted by P_n^m with vertices v_1, v_2, \dots, v_n and edges $v_i v_j$ whenever $|j-i| \leq m$. For simplicity, an edge that joins two vertices where $j-i = t$ will be referred to as a t -hop.

We first consider P_n^2 . We begin by defining a piecewise function, which will be used in the subsequent lemma:

$$f(i, j, k) = \begin{cases} \frac{j-i+1}{k-i+1} & \text{if } k-i \equiv 1 \pmod{2} \text{ and } j-i \equiv 1 \pmod{2}, \\ \frac{k-j+1}{k-i+1} & \text{if } k-i \equiv 1 \pmod{2} \text{ and } j-i \equiv 0 \pmod{2}, \\ 1 & \text{if } k-i \equiv 0 \pmod{2} \text{ and } j-i \equiv 0 \pmod{2}, \\ 0 & \text{if } k-i \equiv 0 \pmod{2} \text{ and } j-i \equiv 1 \pmod{2}. \end{cases}$$

Lemma 10. *If v_j is a vertex of P_n^2 , then the between centrality of v_1 and v_n is zero and if $1 < j < n$, then the betweenness centrality is*

$$\text{bc}(v_j) = \sum_{\substack{1 < i < j \\ j < k < n}} f(i, j, k). \quad (1)$$

Proof. We prove this proposition for the case where n is even. The case where n is odd is similar. Clearly, $\text{bc}(v_1) = \text{bc}(v_n) = 0$. To calculate the betweenness centrality of any fixed v_j , where $1 < j < n$, we add all values given by the function $f(i, j, k)$ over all i and k , which gives (1).

For the first two cases in our piecewise function we note that since $k-i$ is odd, a shortest path between v_i and v_k must be composed of one 1-hop and $k-i$ 2-hops. In the first case we note that the 1-hop must be before the vertex v_j is reached. Since there are $j-i+1$ possible positions for the 1-hop, $\text{bc}(v_j) = (j-i+1)/(k-i+1)$. In the second case the 1-hop must be after the vertex v_j is reached. Since there are $k-j+1$ possible positions for the 1-hop, $\text{bc}(v_j) = (k-j+1)/(k-i+1)$. For the third and fourth cases since $k-i$ is even, any shortest path between v_i and v_k must be composed of 2-hops. When $k-j$ is even then all of these paths will contain v_j and when $k-j$ is odd then none of these paths contain v_j . \square

We extend this result for P_n^3 in our next lemma. We notice that there are $3^2 = 9$ cases for this step. First we define the following piecewise function. Let

$$g(i, j, k) = \begin{cases} 1 & \text{if } k - i \equiv 0 \pmod{3} \text{ and } j - i \equiv 0 \pmod{3}, \\ \frac{(k - j + 2)(k - j + 5)}{(k - i + 5)(k - i + 2)} & \text{if } k - i \equiv 1 \pmod{3} \text{ and } j - i \equiv 0 \pmod{3}, \\ \frac{(j - i + 2)(j - i + 5)}{(k - i + 5)(k - i + 2)} & \text{if } k - i \equiv 1 \pmod{3} \text{ and } j - i \equiv 1 \pmod{3}, \\ \frac{2(j - i + 1)(k - j + 1)}{(k - i + 5)(k - i + 2)} & \text{if } k - i \equiv 1 \pmod{3} \text{ and } j - i \equiv 2 \pmod{3}, \\ \frac{k - j + 1}{k - i + 1} & \text{if } k - i \equiv 2 \pmod{3} \text{ and } j - i \equiv 0 \pmod{3}, \\ \frac{j - i + 1}{k - i + 1} & \text{if } k - i \equiv 2 \pmod{3} \text{ and } j - i \equiv 2 \pmod{3}, \\ 0 & \text{otherwise.} \end{cases}$$

Lemma 11. *If v_j is a vertex of P_n^3 , then the between centrality of v_1 and v_n is zero and if $1 < j < n$, then*

$$\text{bc}(v_j) = \sum_{\substack{1 < i < j \\ j < k < n}} g(i, j, k).$$

Proof. Clearly, $\text{bc}(v_1) = \text{bc}(v_n) = 0$. To calculate the betweenness centrality of any fixed v_j , where $1 < j < n$, we add all values given by the function $f(i, j, k)$ over all i and k to obtain

$$\text{bc}(v_j) = \sum_{\substack{1 < i < j \\ j < k < n}} f(i, j, k).$$

We consider a series of different cases.

When $k - i \equiv 0 \pmod{3}$, the shortest path between v_i and v_k must consist of 3-hops. Hence these shortest paths will contain v_j if and only if $j - i \equiv 0 \pmod{3}$.

When $k - i \equiv 2 \pmod{3}$ the shortest path between v_i and v_k must consist of a single 2-hop and the rest 3-hops. If $j - i \equiv 1 \pmod{3}$ then v_j will never appear on a shortest path between v_i and v_k . If $j - i \equiv 0 \pmod{3}$ then there will only be 3-hops between v_i and v_j and a single 2-hop and the rest 3-hops between v_j and v_k . There are $\frac{1}{3}(k - j + 1)$ positions in which to place the 2-hop so that v_j lies on a shortest path between v_i and v_k . The total number of shortest paths between v_i and v_k is $\frac{1}{3}(k - i + 1)$. Hence the ratio is $(k - j + 1)/(k - i + 1)$. The case where $j - i \equiv 2 \pmod{3}$ is done similarly.

The case where $k - i \equiv 1 \pmod{3}$ is more complicated as a shortest path between v_i and v_k where $k - i \geq 4$ can have two different forms. The first is a composition

of a single 1-hop and the rest 3-hops. The second is a composition of two 2-hops and the rest 3-hops. Hence from the $\frac{1}{3}(k-i+2)$ positions we must either choose a spot for the single 1-hop or choose two spaces for the two 2-hops. Hence the denominator will be

$$\left(\frac{\frac{1}{3}(k-i+2)}{2}\right) + \frac{1}{3}(k-i+2).$$

Simplifying we obtain that

$$\left(\frac{\frac{1}{3}(k-i+2)}{2}\right) + \frac{1}{3}(k-i+2) = \frac{1}{18}(k-i+2)(k-i+5).$$

When $j-i \equiv 0 \pmod{3}$, we know v_j will be on a shortest path between v_i and v_k if and only if there are only 3-hops between v_i and v_j . Hence the numerator will be

$$\frac{1}{3}(k-j+2) + \left(\frac{\frac{1}{3}(k-j+2)}{2}\right).$$

Simplifying we obtain that

$$\frac{1}{3}(k-j+2) + \left(\frac{\frac{1}{3}(k-j+2)}{2}\right) = \frac{1}{18}(k-j+2)(k-j+5).$$

The case where $j-i \equiv 1 \pmod{3}$ is similar. When $j-i \equiv 2 \pmod{3}$ then there will be a single 2-hop and the rest 3-hops between v_i and v_j , and the same for between v_j and v_k . Hence the numerator will be $\left(\frac{1}{3}(j-i+1)\right)\left(\frac{1}{3}(k-j+1)\right)$. \square

We next investigate higher powers of paths and obtain a complete result for path powers with diameter 2.

We first give an example that shows a connection to the triangular numbers.

Example. Let $G = P_{15}^7$. We note that $\text{bc}(v_j) = \text{bc}(v_{16-j})$.

Clearly $\text{bc}(v_1) = 0$. We next compute $\text{bc}(v_j)$ and consider all shortest paths containing v_j with the form $v_x - v_j - v_y$, where $1 \leq x < j < y \leq 15$. We note that $d(G) = 2$.

$\text{bc}(v_2)$: We first note that any shortest path containing v_2 must start with v_1 and end with v_9 .

Of the paths of length 2 that connect v_1 and v_9 , there are seven possible intermediate vertices v_2, v_3, \dots, v_8 .

Since v_2 is one of these seven possibilities, $\text{bc}(v_2) = \frac{1}{7}$.

$\text{bc}(v_3)$: We first note that any shortest path containing v_3 must have one of the following three forms:

$v_1 - v_9$: Of the shortest paths connecting v_1 and v_9 , there are six possible intermediate vertices v_3, \dots, v_8 .

$v_2 - v_{10}$: Of the shortest paths connecting v_2 and v_{10} , there are seven possible intermediate vertices v_3, \dots, v_9 .

$v_1 - v_{10}$: Of the shortest paths connecting v_1 and v_{10} , there are seven possible intermediate vertices v_3, \dots, v_9 .

$$\text{Hence } bc(v_3) = 2\left(\frac{1}{7}\right) + \frac{1}{6} = \frac{19}{42}.$$

$\underline{bc(v_4)}$: We first note that any shortest path containing v_4 must have one of the following six forms:

$v_1 - v_9$: Of the shortest paths connecting v_1 and v_9 , there are five possible intermediate vertices v_4, \dots, v_8 .

$v_2 - v_{10}$: Of the shortest paths connecting v_2 and v_{10} , there are six possible intermediate vertices v_4, \dots, v_9 .

$v_3 - v_{11}$: Of the shortest paths connecting v_3 and v_{11} , there are seven possible intermediate vertices v_4, \dots, v_{10} .

$v_1 - v_{10}$: Of the shortest paths connecting v_1 and v_{10} , there are six possible intermediate vertices v_4, \dots, v_9 .

$v_2 - v_{11}$: Of the shortest paths connecting v_2 and v_{11} , there are seven possible intermediate vertices v_4, \dots, v_{10} .

$v_1 - v_{11}$: Of the shortest paths connecting v_1 and v_{11} , there are seven possible intermediate vertices v_4, \dots, v_{10} .

$$\text{Hence } bc(v_4) = 3\left(\frac{1}{7}\right) + 2\left(\frac{1}{6}\right) + \frac{1}{5} = \frac{101}{105}.$$

For the sake of brevity we note that this pattern continues with the following observations.

$\underline{bc(v_5)}$: Any shortest path containing v_5 is one of 10 forms where $d(v_x, v_y)$ are 8, 9, 10, or 11.

$$\text{Hence } bc(v_5) = 4\left(\frac{1}{7}\right) + 3\left(\frac{1}{6}\right) + 2\left(\frac{1}{5}\right) + \frac{1}{4} = \frac{241}{140}.$$

$\underline{bc(v_6)}$: Any shortest path containing v_6 is one of 15 forms where $d(v_x, v_y)$ are 8, 9, 10, 11, or 12.

$$\text{Hence } bc(v_6) = 5\left(\frac{1}{7}\right) + 4\left(\frac{1}{6}\right) + 3\left(\frac{1}{5}\right) + 2\left(\frac{1}{4}\right) + \frac{1}{3} = \frac{197}{70}.$$

$\underline{bc(v_7)}$: Any shortest path containing v_7 is one of 21 forms where $d(v_x, v_y)$ are 8, 9, 10, 11, 12, or 13.

$$\text{Hence } bc(v_7) = 6\left(\frac{1}{7}\right) + 5\left(\frac{1}{6}\right) + 4\left(\frac{1}{5}\right) + 3\left(\frac{1}{4}\right) + 2\left(\frac{1}{3}\right) + \frac{1}{2} = \frac{617}{140}.$$

$\underline{bc(v_8)}$: Any shortest path containing v_8 is one of 28 forms where $d(v_x, v_y)$ are 8, 9, 10, 11, 12, 13, or 14.

$$\text{Hence } bc(v_8) = 7\left(\frac{1}{7}\right) + 6\left(\frac{1}{6}\right) + 5\left(\frac{1}{5}\right) + 4\left(\frac{1}{4}\right) + 3\left(\frac{1}{3}\right) + 2\left(\frac{1}{2}\right) + 1 = 7.$$

We note that the number of forms in each of these cases are triangular numbers. This pattern holds in general for $G = P_n^k$, where $n = 2k + 1$. We state this in our next theorem.

P_{15}^7	$\{0, \frac{1}{7}, \frac{19}{42}, \frac{101}{105}, \frac{241}{140}, \frac{197}{70}, \frac{617}{140}, 7, \frac{617}{140}, \frac{197}{70}, \frac{241}{140}, \frac{101}{105}, \frac{19}{42}, \frac{1}{7}, 0\}$
P_{14}^7	$\{0, \frac{1}{7}, \frac{19}{42}, \frac{101}{105}, \frac{241}{140}, \frac{197}{70}, \frac{617}{140}, \frac{617}{140}, \frac{197}{70}, \frac{241}{140}, \frac{101}{105}, \frac{19}{42}, \frac{1}{7}, 0\}$
P_{13}^7	$\{0, \frac{1}{7}, \frac{19}{42}, \frac{101}{105}, \frac{241}{140}, \frac{197}{70}, \frac{197}{70}, \frac{197}{70}, \frac{241}{140}, \frac{101}{105}, \frac{19}{42}, \frac{1}{7}, 0\}$
P_{12}^7	$\{0, \frac{1}{7}, \frac{19}{42}, \frac{101}{105}, \frac{241}{140}, \frac{241}{140}, \frac{241}{140}, \frac{241}{140}, \frac{101}{105}, \frac{19}{42}, \frac{1}{7}, 0\}$
P_{11}^7	$\{0, \frac{1}{7}, \frac{19}{42}, \frac{101}{105}, \frac{101}{105}, \frac{101}{105}, \frac{101}{105}, \frac{19}{42}, \frac{1}{7}, 0\}$
P_{10}^7	$\{0, \frac{1}{7}, \frac{19}{42}, \frac{19}{42}, \frac{19}{42}, \frac{19}{42}, \frac{19}{42}, \frac{19}{42}, \frac{1}{7}, 0\}$
P_9^7	$\{0, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, 0\}$
P_8^7	$\{0, 0, 0, 0, 0, 0, 0\}$

Table 1. Path powers with diameter 2. Note the nested nature of the prefixes ending with $0, \frac{1}{7}, \frac{19}{42}, \frac{101}{105}, \frac{241}{140}, \frac{197}{70}, \frac{617}{140}, 7$.

Theorem 12. Let $G = P_n^k$, where $n = 2k + 1$. Then

$$\text{bc}(v_j) = \sum_{i=1}^{j-1} \frac{j-i}{k+1-i}, \quad (2)$$

where $1 \leq j \leq k$ and $\text{bc}(v_j) = \text{bc}(v_{n+1-j})$.

Proof. When calculating $\text{bc}(v_j)$, we note that v_j is contained in shortest paths between v_x and v_y , where $x < j < y \leq n$ and $y - x = k + i$, where $1 \leq i \leq j - 1$. This will account for $j - i$ pairs where the difference between indices is $k + 1 - i$.

For each pair of vertices v_x and v_y where $y - x = k + i$ there will be $j - i$ sets of paths. Each of these paths will have $k - i + 1$ possible intermediaries. This contributes $(j - i)/(k - i + 1)$ to $\text{bc}(v_j)$. Summing these terms for all $1 \leq i \leq j - 1$ will give the value of $\text{bc}(v_j)$. Hence we have (2). \square

Next we show how the previous theorem can be extended to cover all other path powers of diameter 2. We begin with an example.

Example 13. Let $G = P_{12}^7$. We first note that $\text{bc}(v_j) = \text{bc}(v_{13-j})$.

For v_j where $1 \leq j \leq 5$, the betweenness centrality values are identical to those in P_{15}^7 and can be computed using the exact same method. However the pattern used in the example with P_{15}^7 cannot be extended for $\text{bc}(v_6)$ since $v_y \leq 12$. As a result, the paths used in the computation of $\text{bc}(v_5)$ and $\text{bc}(v_6)$ are identical. Hence, $\text{bc}(v_1) = 0$; $\text{bc}(v_2) = \frac{1}{7}$, $\text{bc}(v_3) = \frac{19}{42}$, $\text{bc}(v_4) = \frac{101}{105}$, and $\text{bc}(v_5) = \text{bc}(v_6) = \frac{241}{140}$.

We observe that the betweenness centrality values in path powers with diameter 2 have a nested pattern (see Table 1).

We formalize this property in our next theorem.

Theorem 14. Let $G = P_n^k$, where $n < 2k + 1$ and $j < k$. Then

$$\text{bc}(v_j) = \sum_{i=1}^{j-1} \frac{j-i}{k-i+1} \quad (3)$$

for all $2 \leq j \leq n - k$ and $\text{bc}(v_j) = \text{bc}(v_{n+1-j})$. For P_n^k , the $\text{bc}(v_j)$ are all equal for all $n - k \leq j \leq \lceil \frac{1}{2}n \rceil$.

Proof. When calculating $\text{bc}(v_j)$, we note that v_j is contained in shortest paths between v_x and v_y where $x < j < y \leq n$ and $y - x = k + i$ where $1 \leq i \leq j - 1$. This will account for $j - i$ pairs where the difference between indices is $k + 1 - i$.

For each pair of vertices v_x and v_y where $y - x = k + i$ (where $k + i \leq n$), there will be $j - i$ sets of paths. Each of these paths will have $k - i + 1$ possible intermediaries. This contributes $(j - i)/(k - i + 1)$ to $\text{bc}(v_j)$. Summing these terms for all $1 \leq i \leq j - 1$ will give the value of $\text{bc}(v_j)$. Hence we have (3). For P_n^k , $\text{bc}(v_j)$ is the same as in P_{2k+1}^k for the first $n - k$ terms. Then since there are the same number of pairs of vertices v_x and v_y where $y - x = k + i$ (where $k + i \leq n$) in P_n^k , the $\text{bc}(v_j)$ are all the same for $n - k \leq j \leq \lceil \frac{1}{2}n \rceil$. \square

3. Conclusion

For path powers with larger diameter the problem becomes more complex. The case of P_n^m involves m^2 different cases, and as n increases the cases become more complicated. Hence the problem for general powers of paths is more difficult. We note that problem is tied to the number of integer partitions with a fixed upper bound on the size of each part [Ratsaby 2008]. The objective is to minimize the number of parts.

We pose the following problem.

Problem 15. Determine the betweenness centrality for all vertices in P_n^m .

Acknowledgements

The authors are grateful to an anonymous referee whose careful reading and comments improved the presentation of this paper. This research was supported by a National Science Foundation Research Experiences for Undergraduates Site Award Grant (#1062128) with cofunding from the Department of Defense. Darren Narayan was also supported by NSF Award #1019532. Rigoberto Flórez was partially supported by The Citadel Foundation.

References

- [Brandes et al. 2016] U. Brandes, S. Borgatti, and L. Freeman, “Maintaining the duality of closeness and betweenness centrality”, *Social Networks* **44** (2016), 153–159.

- [Bullmore and Sporns 2009] E. Bullmore and O. Sporns, “Complex brain networks: graph theoretical analysis of structural and functional systems”, *Nature Rev. Neuroscience* **10**:3 (2009), 186–198.
- [Freeman 1977] L. C. Freeman, “A set of measures of centrality based upon betweenness”, *Sociometry* **40**:1 (1977), 35–41.
- [Freeman et al. 1991] L. C. Freeman, S. P. Borgatti, and D. R. White, “Centrality in valued graphs: a measure of betweenness based on network flow”, *Social Networks* **13**:2 (1991), 141–154. MR
- [Gago et al. 2012] S. Gago, J. Hurajová, and T. Madaras, “Notes on the betweenness centrality of a graph”, *Math. Slovaca* **62**:1 (2012), 1–12. MR Zbl
- [Grassi et al. 2009] R. Grassi, R. Scapellato, S. Stefani, and A. Torriero, “Betweenness centrality: extremal values and structural properties”, pp. 161–175 in *Networks, topology and dynamics: theory and applications to economics and social systems*, edited by A. K. Naimzada et al., Lecture Notes in Economics and Mathematical Systems **613**, Springer, 2009. Zbl
- [Guye et al. 2010] M. Guye, G. Bettus, F. Bartolomei, and P. J. Cozzone, “Graph theoretical analysis of structural and functional connectivity MRI in normal and pathological brain networks”, *Magn. Reson. Mater. Phys.* **23**:5-6 (2010), 409–421.
- [Kumar and Balakrishnan 2016] S. Kumar R. and K. Balakrishnan, “Betweenness centrality of Cartesian product of graphs”, preprint, 2016. arXiv
- [Kumar et al. 2014] S. Kumar Raghavan Unnithan, B. Kannan, and M. Jathavedan, “Betweenness centrality in some classes of graphs”, *Int. J. Comb.* **2014** (2014), art. id. 241723. MR Zbl
- [Pandit et al. 2013] A. S. Pandit, P. Expert, R. Lambiotte, V. Bonnelle, R. Leech, F. E. Turkheimer, and D. J. Sharp, “Traumatic brain injury impairs small-world topology”, *Neurology* **80**:20 (2013), 1826–1833.
- [Ratsaby 2008] J. Ratsaby, “Estimate of the number of restricted integer-partitions”, *Appl. Anal. Discrete Math.* **2**:2 (2008), 222–233. MR Zbl
- [White and Borgatti 1994] D. R. White and S. P. Borgatti, “Betweenness centrality measures for directed graphs”, *Social Networks* **16**:4 (1994), 335–346.

Received: 2017-03-04 Revised: 2017-07-26 Accepted: 2018-01-20

emmacarter2014@gmail.com *School of Mathematical Sciences, Rochester Institute of Technology, Rochester, NY, United States*

bte14@math.rutgers.edu *Department of Mathematics, Rutgers University, Piscataway, NJ, United States*

dng2551@rit.edu *Department of Software Engineering, Rochester Institute of Technology, Rochester, NY, United States*

florezr1@citadel.edu *Department of Mathematics and Computer Science, The Citadel, Charleston, SC, United States*

dansma@rit.edu *School of Mathematical Sciences, Rochester Institute of Technology, Rochester, NY, United States*

The length spectrum of the sub-Riemannian three-sphere

David Klapheck and Michael VanValkenburgh

(Communicated by Kenneth S. Berenhaut)

We determine the lengths of all closed sub-Riemannian geodesics on the three-sphere S^3 . Our methods are elementary and allow us to avoid using explicit formulas for the sub-Riemannian geodesics.

1. Introduction

In the case of a compact Riemannian manifold (M, g) there is a relationship between closed geodesics, representing paths of free classical particles in periodic motion, and eigenfunctions of the Laplacian Δ , representing periodic free quantum “waves” (up to a phase factor). For this reason, the set of lengths of closed geodesics is called the *length spectrum*, in analogy to the spectrum of the Laplacian. There are in fact precise formulas relating lengths to eigenvalues; see for example the announcement [Guillemin and Weinstein 1976] for a readable discussion with references.

So far there is no such formula relating lengths and eigenvalues in the case of a compact sub-Riemannian (sR) manifold. We recall that an sR manifold is a manifold with a specified linear subbundle \mathcal{H} (the “horizontal bundle”) of its tangent bundle, along with a Riemannian metric on \mathcal{H} . Distances between points are then measured using curves that are constrained to have tangent vectors in \mathcal{H} (“horizontal curves”). In fact, when \mathcal{H} is the span of a set of bracket-generating vector fields, then the Chow–Rashevskii theorem says that any two points are connected by a horizontal curve, a result that even experts find surprising [Burago et al. 2001, p. 178]; thus given any two points there is a shortest horizontal curve connecting them; it is called an *sR geodesic*.

Sub-Riemannian geometry is of practical interest; for example, the problem of parallel parking a car, or, even worse, a car with a trailer, is a problem in sR geometry [Burago et al. 2001; Nelson 1967]. And there are further surprises from the purely mathematical point of view, one being Montgomery’s proof of existence of singular sR geodesics, singular in the sense that they do not satisfy the geodesic equations

MSC2010: 53C17.

Keywords: sub-Riemannian geometry.

(Hamilton's equations) [Montgomery 1994; 2002]. This and other relatively recent results in sR geometry inspired renewed interest in the sub-Laplacian: the operator naturally associated with the given (sub-)Riemannian metric on \mathcal{H} .

In this paper, with the goal of understanding a single example, we compute the sR length spectrum of the three-dimensional sphere S^3 with its standard sR structure; this is to be compared with the spectrum of the sub-Laplacian on S^3 , known by Taylor [1986] and generalized to other connected, semisimple Lie groups by Domokos [2015]. We expect that a general theory relating the sR length spectrum to the spectrum of the sub-Laplacian would be amenable to the tools of microlocal analysis, as in the Riemannian setting; [Colin de Verdière et al. 2016] gives hope that this will be accomplished.

We focus on S^3 with its standard sR structure because it is perhaps the simplest compact manifold with an sR structure, and there are no singular sR geodesics on S^3 ; that is, all sR geodesics arise as projections of solutions of Hamilton's equations [Montgomery 2002]. Moreover, we wish to compare the sR setting to the Riemannian setting, in which the spheres S^n are of fundamental importance, as examples of manifolds all of whose geodesics are closed and have the same length T ; in general this is equivalent to most of the spectrum of $\sqrt{-\Delta}$ being concentrated near an arithmetic progression $(2\pi/T)k + \beta$, $k = 1, 2, \dots$, for some constant β [Duistermaat and Guillemin 1975]. As we will see, in the case of S^3 not all sR geodesics are closed, and not all have the same length:

Theorem. *The set of lengths of the closed sR geodesics on S^3 is*

$$\{2\pi\sqrt{n} : n \in \mathbb{N}\}.$$

Others have studied the sR geodesics on S^3 [Calin et al. 2009; Chang et al. 2009; Hurtado and Rosales 2008] (see also the survey article [D'Angelo and Tyson 2010]), but we compute their lengths and differ from the previous work in that we consistently use Hopf coordinates on S^3 and avoid using explicit formulas for the sR geodesics; we believe it clarifies the presentation to *not* use explicit formulas.

We introduce the sR structure and geodesic equations in Section 2 using Hopf coordinates, and in Section 3 we categorize the qualitatively different types of sR geodesics. In Section 4 we determine which sR geodesics are closed, and in Section 5 we compute their lengths, resulting in the theorem above. Finally, in Section 6 we compare the sR length spectrum to the previously known spectrum of the sub-Laplacian.

Remark. During peer review, it was pointed out that the above result is contained in [Chang et al. 2011] (see their Theorem 2). However, our proof is entirely new and has the advantage of being elementary after the introduction of Hamilton's equations (2) in our chosen coordinate system.

2. S^3 in Euclidean and Hopf coordinates

First we consider S^3 as a subset of \mathbb{R}^4 :

$$S^3 = \{(x_1, y_1, x_2, y_2) \in \mathbb{R}^4 : x_1^2 + y_1^2 + x_2^2 + y_2^2 = 1\}.$$

On S^3 we have the orthonormal vector fields

$$\begin{aligned} V &:= -y_1 \frac{\partial}{\partial x_1} + x_1 \frac{\partial}{\partial y_1} - y_2 \frac{\partial}{\partial x_2} + x_2 \frac{\partial}{\partial y_2}, \\ E_1 &:= -x_2 \frac{\partial}{\partial x_1} + y_2 \frac{\partial}{\partial y_1} + x_1 \frac{\partial}{\partial x_2} - y_1 \frac{\partial}{\partial y_2}, \\ E_2 &:= -y_2 \frac{\partial}{\partial x_1} - x_2 \frac{\partial}{\partial y_1} + y_1 \frac{\partial}{\partial x_2} + x_1 \frac{\partial}{\partial y_2}, \end{aligned}$$

which satisfy the Lie bracket relations

$$[V, E_1] = -2E_2, \quad [E_2, V] = -2E_1, \quad [E_1, E_2] = -2V.$$

Thus $\mathcal{H}(S^3) = \text{span}\{E_1, E_2\}$ is a bracket-generating tangent subbundle, and by the Chow–Rashevskii theorem any two points on S^3 are connected by an sR geodesic.

The orbits of the flow generated by V are the circles of the Hopf fibration [Cannas da Silva 2008], so we find it convenient to use Hopf coordinates, see [Wikipedia 2015], on S^3 :

$$\begin{aligned} x_1 &= \cos \theta_1 \sin \theta_0, & y_1 &= \sin \theta_1 \sin \theta_0, \\ x_2 &= \cos \theta_2 \cos \theta_0, & y_2 &= \sin \theta_2 \cos \theta_0 \end{aligned}$$

for $0 < \theta_0 < \frac{\pi}{2}$ and $0 < \theta_j < 2\pi$, $j = 1, 2$. We picture the $(\theta_0, \theta_1, \theta_2)$ -space as “the Hopf cube” $(0, \frac{\pi}{2}) \times (0, 2\pi) \times (0, 2\pi)$. When we have occasion to exit the Hopf cube, we simply return to the definition of Hopf coordinates to make the correct interpretation:

- (i) For the θ_1 - and θ_2 -coordinates the values 0 and 2π are identified.
- (ii) When a point crosses the $\theta_0 = 0$ plane we have that θ_0 changes direction (“bounces”) and (θ_1, θ_2) is identified with $(\theta_1 + \pi, \theta_2)$.
- (iii) When a point crosses the $\theta_0 = \frac{\pi}{2}$ -plane we have that θ_0 changes direction and (θ_1, θ_2) is identified with $(\theta_1, \theta_2 + \pi)$.

The (round) Riemannian metric in Hopf coordinates is

$$ds^2 = d\theta_0^2 + \sin^2 \theta_0 d\theta_1^2 + \cos^2 \theta_0 d\theta_2^2, \tag{1}$$

and the Laplacian is

$$\Delta = \frac{1}{\sin(2\theta_0)} \frac{\partial}{\partial \theta_0} \circ \sin(2\theta_0) \frac{\partial}{\partial \theta_0} + \csc^2 \theta_0 \frac{\partial^2}{\partial \theta_1^2} + \sec^2 \theta_0 \frac{\partial^2}{\partial \theta_2^2}.$$

We now write the sR structure in Hopf coordinates. We can introduce $r > 0$, to give coordinates to \mathbb{R}^4 , allowing us to write the $\partial/\partial x_j$, $\partial/\partial y_j$ in terms of the $\partial/\partial \theta_j$, $\partial/\partial r$. Then restricting to functions on S^3 we get

$$\begin{aligned}\frac{\partial}{\partial x_1} &= \cos \theta_1 \cos \theta_0 \frac{\partial}{\partial \theta_0} - \sin \theta_1 \csc \theta_0 \frac{\partial}{\partial \theta_1}, \\ \frac{\partial}{\partial y_1} &= \sin \theta_1 \cos \theta_0 \frac{\partial}{\partial \theta_0} + \cos \theta_1 \csc \theta_0 \frac{\partial}{\partial \theta_1}, \\ \frac{\partial}{\partial x_2} &= -\cos \theta_2 \sin \theta_0 \frac{\partial}{\partial \theta_0} - \sin \theta_2 \sec \theta_0 \frac{\partial}{\partial \theta_2}, \\ \frac{\partial}{\partial y_2} &= -\sin \theta_2 \sin \theta_0 \frac{\partial}{\partial \theta_0} + \cos \theta_2 \sec \theta_0 \frac{\partial}{\partial \theta_2}.\end{aligned}$$

Our vector fields are then

$$\begin{aligned}V &= \frac{\partial}{\partial \theta_1} + \frac{\partial}{\partial \theta_2}, \\ E_1 &= -\cos(\theta_1 + \theta_2) \frac{\partial}{\partial \theta_0} + \sin(\theta_1 + \theta_2) \cot \theta_0 \frac{\partial}{\partial \theta_1} - \sin(\theta_1 + \theta_2) \tan \theta_0 \frac{\partial}{\partial \theta_2}, \\ E_2 &= -\sin(\theta_1 + \theta_2) \frac{\partial}{\partial \theta_0} - \cos(\theta_1 + \theta_2) \cot \theta_0 \frac{\partial}{\partial \theta_1} + \cos(\theta_1 + \theta_2) \tan \theta_0 \frac{\partial}{\partial \theta_2}.\end{aligned}$$

The commutation relations hold, the same as before, and the vector fields are still orthonormal (of course, with respect to the Riemannian metric in Hopf coordinates).

The sR metric, written in Hopf coordinates, is

$$S = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos^2 \theta_0 \sin^2 \theta_0 & -\cos^2 \theta_0 \sin^2 \theta_0 \\ 0 & -\cos^2 \theta_0 \sin^2 \theta_0 & \cos^2 \theta_0 \sin^2 \theta_0 \end{pmatrix}.$$

Indeed it is easy to check that E_1 and E_2 are orthonormal with respect to S , and V is in the kernel of S . Written as a two-tensor,

$$S = d\theta_0 \otimes d\theta_0 + \cos^2 \theta_0 \sin^2 \theta_0 (d\theta_1 - d\theta_2) \otimes (d\theta_1 - d\theta_2).$$

The sR Laplacian, written in Hopf coordinates, is

$$\Delta_{\text{sR}} = E_1^2 + E_2^2 = \frac{1}{\sin(2\theta_0)} \frac{\partial}{\partial \theta_0} \circ \sin(2\theta_0) \frac{\partial}{\partial \theta_0} + \left(\cot \theta_0 \frac{\partial}{\partial \theta_1} - \tan \theta_0 \frac{\partial}{\partial \theta_2} \right)^2.$$

We can consider the sR metric as being the limit of certain penalty metrics, where the V -direction is penalized by a factor $\lambda > 1$. After simple linear algebra (multiplying the V -direction by λ , multiplying the other directions by 1, and *then*

applying the Riemannian metric), the λ -penalty metric is given by the matrix

$$P_\lambda = \begin{pmatrix} 1 & 0 & 0 \\ 0 & (\lambda^2 - 1) \sin^4 \theta_0 + \sin^2 \theta_0 & (\lambda^2 - 1) \cos^2 \theta_0 \sin^2 \theta_0 \\ 0 & (\lambda^2 - 1) \cos^2 \theta_0 \sin^2 \theta_0 & (\lambda^2 - 1) \cos^4 \theta_0 + \cos^2 \theta_0 \end{pmatrix}.$$

Indeed, one can check that in fact V , E_1 , and E_2 are orthogonal with respect to this metric, that E_1 and E_2 have length 1, and that V has length λ . We can easily compute

$$\det P_\lambda = \lambda^2 \cos^2 \theta_0 \sin^2 \theta_0$$

and

$$P_\lambda^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cot^2 \theta_0 + \lambda^{-2} & \lambda^{-2} - 1 \\ 0 & \lambda^{-2} - 1 & \tan^2 \theta_0 + \lambda^{-2} \end{pmatrix}.$$

From this we find that the λ -penalty Laplacian on S^3 is

$$\Delta_\lambda = \frac{\partial^2}{\partial \theta_0^2} + 2 \cot(2\theta_0) \frac{\partial}{\partial \theta_0} + \left(\cot \theta_0 \frac{\partial}{\partial \theta_1} - \tan \theta_0 \frac{\partial}{\partial \theta_2} \right)^2 + \lambda^{-2} \left(\frac{\partial}{\partial \theta_1} + \frac{\partial}{\partial \theta_2} \right)^2.$$

That is,

$$\Delta_\lambda = E_1^2 + E_2^2 + \lambda^{-2} V^2,$$

as might have been expected.

Montgomery discovered an example in which geodesics with respect to the λ -penalty metric converge (as $\lambda \rightarrow \infty$) to sR geodesics that do *not* solve the sR geodesic equations, in contrast to the Riemannian setting; that is, Montgomery [1994] discovered so-called *singular geodesics*. For the case of S^3 (and more generally, in the contact case), singular geodesics do not exist, so it suffices to study the geodesic equations, or, equivalently, Hamilton's equations [Montgomery 2002].

We denote the dual variable to θ_j by ξ_j . The sR Hamiltonian is then

$$H(\theta, \xi) = \frac{1}{2} \xi_0^2 + \frac{1}{2} (\cot \theta_0 \xi_1 - \tan \theta_0 \xi_2)^2.$$

Hamilton's equations, giving the sR geodesics, are then, for $j = 0, 1, 2$,

$$\dot{\theta}_j = \frac{\partial H}{\partial \xi_j}, \quad \dot{\xi}_j = -\frac{\partial H}{\partial \theta_j}.$$

Explicitly,

$$\begin{aligned} \dot{\theta}_0 &= \xi_0, & \dot{\xi}_0 &= \cot \theta_0 \csc^2 \theta_0 \xi_1^2 - \tan \theta_0 \sec^2 \theta_0 \xi_2^2, \\ \dot{\theta}_1 &= \cot^2 \theta_0 \xi_1 - \xi_2, & \dot{\xi}_1 &= 0, \\ \dot{\theta}_2 &= \tan^2 \theta_0 \xi_2 - \xi_1, & \dot{\xi}_2 &= 0. \end{aligned} \tag{2}$$

One obvious advantage of using Hopf coordinates is that ξ_1 and ξ_2 are constant along the flow; in addition, as always H is constant along the flow, so we already have three conserved quantities. Also, these equations have a clear symmetry; for example,

$$\cot(\tfrac{1}{2}\pi - \theta_0) \csc^2(\tfrac{1}{2}\pi - \theta_0) = \tan \theta_0 \sec^2 \theta_0.$$

The penalty Hamiltonian is

$$\begin{aligned} H_\lambda(\theta, \xi) &= H + \frac{1}{2\lambda^2}(\xi_1 + \xi_2)^2 \\ &= \tfrac{1}{2}\xi_0^2 + \tfrac{1}{2}(\cot \theta_0 \xi_1 - \tan \theta_0 \xi_2)^2 + \frac{1}{2\lambda^2}(\xi_1 + \xi_2)^2. \end{aligned} \quad (3)$$

The corresponding penalty Hamiltonian equations, giving the penalty geodesics, are then

$$\begin{aligned} \dot{\theta}_0 &= \xi_0, & \dot{\xi}_0 &= \cot \theta_0 \csc^2 \theta_0 \xi_1^2 - \tan \theta_0 \sec^2 \theta_0 \xi_2^2, \\ \dot{\theta}_1 &= \cot^2 \theta_0 \xi_1 - \xi_2 + \lambda^{-2}(\xi_1 + \xi_2), & \dot{\xi}_1 &= 0, \\ \dot{\theta}_2 &= \tan^2 \theta_0 \xi_2 - \xi_1 + \lambda^{-2}(\xi_1 + \xi_2), & \dot{\xi}_2 &= 0. \end{aligned} \quad (4)$$

For the case of the Riemannian metric on S^3 , that is, the case $\lambda = 1$, the equations simplify, and we get

$$\dot{\theta}_1 = \csc^2 \theta_0 \xi_1, \quad \dot{\theta}_2 = \sec^2 \theta_0 \xi_2.$$

When $\lambda = 1$, the solutions of Hamilton's equations are great circles on S^3 .

3. Categorizing sR geodesics

Our categorization of sR geodesics is based on a reduced problem. In Hamilton's equations (2), since ξ_1 and ξ_2 are constant along the flow, we can isolate the equations

$$\dot{\theta}_0 = \xi_0, \quad \dot{\xi}_0 = \cot \theta_0 \csc^2 \theta_0 \xi_1^2 - \tan \theta_0 \sec^2 \theta_0 \xi_2^2,$$

which are Hamilton's equations for the sR Hamiltonian H considered as a function of *two* variables

$$H(\theta_0, \xi_0) = \tfrac{1}{2}\xi_0^2 + \tfrac{1}{2}(\cot \theta_0 \xi_1 - \tan \theta_0 \xi_2)^2. \quad (5)$$

Equation (5) can be viewed as a *one-dimensional* energy equation: it is of the form

$$\text{energy} = \text{kinetic energy} + \text{potential energy},$$

with potential function

$$U = \tfrac{1}{2}(\cot \theta_0 \xi_1 - \tan \theta_0 \xi_2)^2.$$

We now list the various disjoint cases:

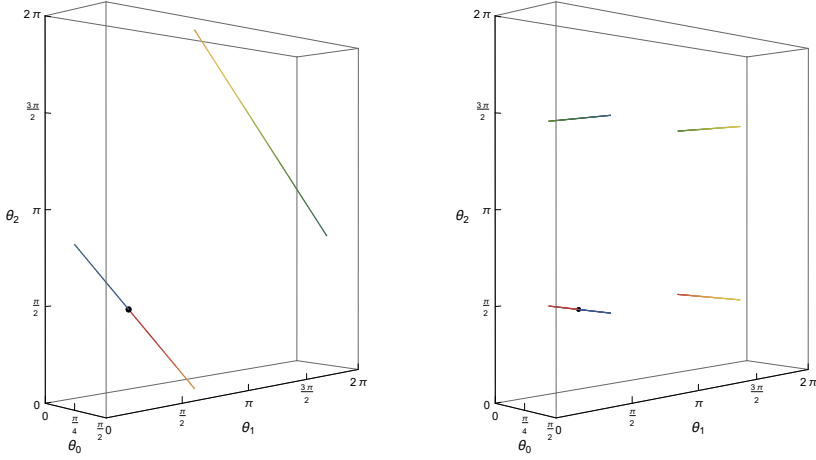


Figure 1. Case 1b (left) and Case 2 (right).

(1) A fixed point in the (θ_0, ξ_0) phase plane. From our original choice of coordinates we may assume that $\theta_0 \equiv \frac{\pi}{4}$, and then $\xi_1^2 = \xi_2^2$.

- (a) $\xi_1 = \xi_2$. (This is precisely the case when $H = 0$.) Then from Hamilton's equations θ_0 , θ_1 , and θ_2 are constant; this gives a degenerate sR geodesic of length 0.
- (b) $\xi_1 = -\xi_2 \neq 0$. Hamilton's equations then say that the speed on the Hopf cube is $\sqrt{2}|\xi_1 - \xi_2|$, and the length of the (simple) closed curve on the Hopf cube is $\sqrt{2}2\pi$, so the period is $2\pi/|\xi_1 - \xi_2|$. On S^3 the speed is $|\xi_1 - \xi_2|$, so the length of this closed sR geodesic is 2π . See Figure 1.

We categorize the remaining cases in terms of the potential function U .

(2) The “free” case $U \equiv 0$. This happens precisely when $\xi_1 = \xi_2 = 0$. (We have already dispensed with the case when θ_0 is constant.) By Hamilton's equations, $\dot{\theta}_1$, $\dot{\theta}_2$, and $\dot{\xi}_0$ are also identically zero, while $\dot{\theta}_0 = \xi_0$. That is, we have a point with speed $|\xi_0|$ moving purely in the θ_0 -direction; the length of this (simple) closed geodesic is 2π . (It is both a geodesic and an sR geodesic.) See Figure 1.

(3) $\xi_1 \neq 0$ and $\xi_2 \neq 0$. Then U is a potential well with a single nondegenerate minimum occurring when $\tan^4 \theta_0 = \xi_1^2 / \xi_2^2$. Typical potential functions are shown in Figure 2 for ξ_1 and ξ_2 with the same and opposite signs.

Since in this case θ_0 is *not* constant, its period is

$$\text{period}(\theta_0) = 2 \int_a^b \frac{d\theta_0}{\sqrt{2(H - U)}} = 2 \int_a^b \frac{d\theta_0}{\sqrt{2H - (\cot \theta_0 \xi_1 - \tan \theta_0 \xi_2)^2}}.$$

Here a and b are the “turning points,” where the kinetic energy is zero.

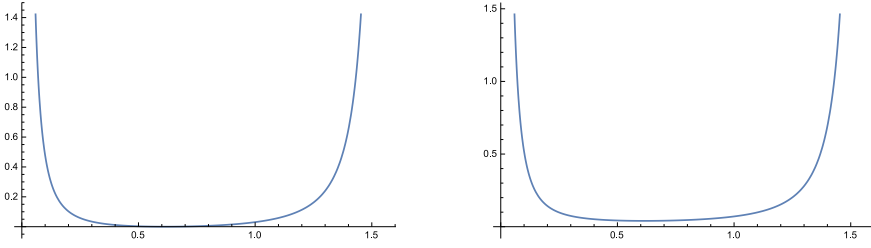


Figure 2. Potential functions with $\xi_1 = 0.1$, $\xi_2 = 0.2$ (left) and $\xi_1 = 0.1$, $\xi_2 = -0.2$ (right).

Fortunately it is possible to evaluate this integral using freshman calculus. Substituting

$$x = \cos^2 \theta_0, \quad 0 < \theta_0 < \frac{\pi}{2},$$

we get

$$\text{period}(\theta_0) = \int_{\cos^2 b}^{\cos^2 a} \frac{dx}{\sqrt{[-2H - (\xi_1 + \xi_2)^2]x^2 + 2(H + \xi_1 \xi_2 + \xi_2^2)x - \xi_2^2}}.$$

The limits of integration are exactly the points where the denominator vanishes (where the velocity is zero), and we recall that the Hamiltonian for Riemannian geodesics is $H_1 = H + \frac{1}{2}(\xi_1 + \xi_2)^2$ (the case $\lambda = 1$), so we have

$$\text{period}(\theta_0) = \frac{1}{\sqrt{2H_1}} \int_{\cos^2 b}^{\cos^2 a} \frac{dx}{\sqrt{(\cos^2 a - x)(x - \cos^2 b)}}.$$

This is an integral known to be solvable by elementary functions. Following [Woods 1934, p. 366],¹ we make the substitution defined by

$$z^2 + 1 = \frac{\cos^2 a - \cos^2 b}{x - \cos^2 b}$$

and finally get the answer

$$\text{period}(\theta_0) = \sqrt{\frac{2}{H_1}} \int_0^\infty \frac{dz}{z^2 + 1} = \frac{\pi}{\sqrt{2H_1}}.$$

For future reference, we note that this is one-half the period of the (Riemannian) geodesic flow; after all, the speed of the geodesic flow is $\sqrt{2H_1}$, and we know the length of each geodesic, a great circle on S^3 , to be 2π . (See Section 4.)

Examples are pictured in Figure 3, where ξ_1 and ξ_2 have the same and opposite signs.

¹This is the book mentioned in [Feynman 1985] as giving him valuable tricks for integration.

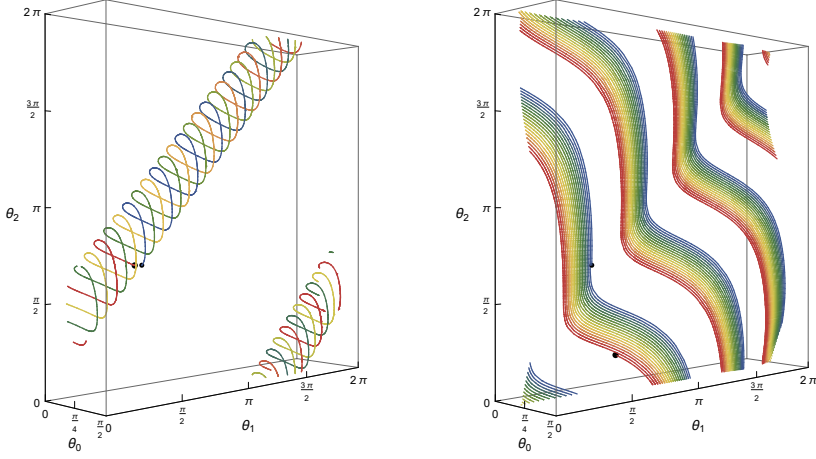


Figure 3. Case 3 when ξ_1 and ξ_2 have the same sign (left) and opposite signs (right).

(4) It remains to check the exceptional cases when $\{\xi_1 = 0 \text{ and } \xi_2 \neq 0\}$ and when $\{\xi_1 \neq 0 \text{ and } \xi_2 = 0\}$. For example, when $\xi_1 = 0$ the potential function is

$$U = \frac{1}{2} \tan^2 \theta_0 \xi_2^2, \quad 0 < \theta_0 < \frac{\pi}{2}.$$

The force induced by this potential causes the point to exit the Hopf cube through the $\theta_0 = 0$ plane; rather we interpret it as bouncing off the plane, returning to the Hopf cube but with θ_1 shifted by π . (See Section 2.) With reasoning as in the previous case, we find that again $\text{period}(\theta_0) = \pi/\sqrt{2H_1}$. The case when $\xi_1 \neq 0$ and $\xi_2 = 0$ follows by renaming the variables $\theta_0 \leftrightarrow \frac{\pi}{2} - \theta_0$ and $\xi_1 \leftrightarrow \xi_2$. In Figure 4 we

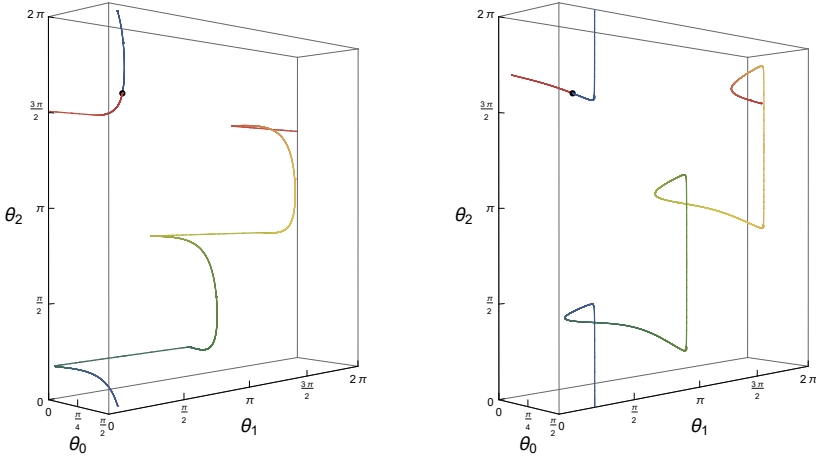


Figure 4. Case 4 with $0 < \xi_1 \ll \xi_2$ (left) and $0 < \xi_2 \ll \xi_1$ (right).

have the cases when $0 < \xi_1 \ll \xi_2$ and $0 < \xi_2 \ll \xi_1$, which illustrate how exiting the Hopf cube and re-entering after a π -shift appears as a limiting case.

Finally, we note that all sR geodesics are simple curves; that is, they do not self-intersect except trivially for closed curves. In Cases 1 and 2 above it is obvious. In Cases 3 and 4 we only need to wait until ξ_0 is zero, corresponding to the θ_0 -particle having zero kinetic energy in the potential well U . When $(\theta_0, \theta_1, \theta_2)$ returns to that value, clearly ξ_0 is zero again, ξ_1, ξ_2 are the same as always, and the $\dot{\theta}_j$ and $\dot{\xi}_j$ return to their values; thus the curve only self-intersects in the case of a closed curve, at the end of a period.

4. Determining which sR geodesics are closed

In this section we identify the closed sR geodesics on S^3 ; we only need to consider Cases 3 and 4, and we may assume that the initial value of ξ_0 is zero. (See the comment at the end of Section 3.) Hurtado and Rosales [2008] found a necessary and sufficient condition in terms of geodesic curvature (see also [D'Angelo and Tyson 2010]):

Theorem [Hurtado and Rosales 2008]. *Let $\gamma : \mathbb{R} \rightarrow S^3$ be a complete sR geodesic of curvature λ . Then γ is a closed curve diffeomorphic to a circle if and only if $\lambda/\sqrt{1+\lambda^2}$ is a rational number. Otherwise γ is diffeomorphic to \mathbb{R} and is dense in some group translate of a Clifford torus.*

Their proof relies on closed-form expressions of the sR geodesics. Here we give a condition which does not rely on closed-form expressions.

From the λ -penalty Hamilton's equations (4), we see that the sR Hamiltonian vector field for the Hamiltonian H is the difference of the Hamiltonian vector fields for the Hamiltonians H_1 and $H_V = \frac{1}{2}(\xi_1 + \xi_2)^2$. Moreover, the vector fields Lie-commute (it is easy to see that the Poisson bracket of H_1 and H_V is zero), so the Hamiltonian flows for H_1 and H_V commute. We can thus consider the H -flow as an H_1 -flow followed by an H_V -flow.

The Hamiltonian for the Riemannian geodesics may be written as

$$H_1(\theta, \xi) = \frac{1}{2}\xi_0^2 + \frac{1}{2}(\csc^2\theta_0 \xi_1^2 + \sec^2\theta_0 \xi_2^2),$$

(the penalty Hamiltonian (3) with $\lambda = 1$), so the first of Hamilton's equations, giving the velocities, are then

$$\dot{\theta}_0 = \xi_0, \quad \dot{\theta}_1 = \csc^2\theta_0 \xi_1, \quad \dot{\theta}_2 = \sec^2\theta_0 \xi_2.$$

We see that the speed, measured using the Riemannian metric (1), is $\sqrt{2H_1}$, which is constant. Moreover, the length of the Riemannian geodesic is 2π , being a great circle, so that the period of the closed orbit is $2\pi/\sqrt{2H_1} = 2 \times \text{period}(\theta_0)$.

On the other hand, the Hamiltonian H_V has Hamiltonian equations

$$\dot{\theta}_1 = \xi_1 + \xi_2, \quad \dot{\theta}_2 = \xi_1 + \xi_2, \quad \dot{\xi}_1 = \dot{\xi}_2 = 0.$$

Thus the speed (with respect to the Euclidean metric on the Hopf cube) is $\sqrt{2}|\xi_1 + \xi_2|$. The length of the orbit (a circle fiber of the Hopf fibration) is $\sqrt{2} \cdot 2\pi$, so the period of the H_V -flow is $2\pi/|\xi_1 + \xi_2|$. (It might seem strange that we find the speed and length with respect to the *Euclidean* metric on the Hopf cube, but the Euclidean metric is sufficient to compute the period of the H_V -flow.)

For a combination of an H_1 -flow and an H_V -flow to result in a closed curve, we need the H_1 -flow to return θ_0 to its original value (since the H_V -flow has no $\partial/\partial\theta_0$ component). Thus the time elapsed must be an integer multiple of $\text{period}(\theta_0) = \pi/\sqrt{2H_1}$. If the integer is odd, the H_1 -flow takes the point to its antipodal point, and we would need a half-period of the H_V -flow to return to the starting point. If the integer is even, the H_1 -flow takes the point back to itself, and we could only allow full periods of the H_V -flow. To summarize, a necessary and sufficient condition for a closed sR geodesic is

$$\text{time elapsed} = p \times \frac{\pi}{|\xi_1 + \xi_2|} = q \times \frac{\pi}{\sqrt{2H_1}},$$

where $p, q \in \{1, 2, 3, \dots\}$ are either both odd or both even. In particular,

$$\frac{p}{q} = \frac{|\xi_1 + \xi_2|}{\sqrt{2H_1}} = \sqrt{1 - \frac{H}{H_1}} \in \mathbb{Q} \cap (0, 1), \quad (6)$$

The quantity p/q is conserved along the flow and is positively homogeneous of degree zero in the ξ -variables. The condition (6) is also sufficient to have a closed sR geodesic. If it holds, then we have

$$H\text{-period} = p \times \frac{\pi}{|\xi_1 + \xi_2|} = q \times \frac{\pi}{\sqrt{2H_1}}$$

for the least such integers $0 < p < q$ that are either both odd or both even.

When plotting sR geodesics in Cases 3 and 4, we can fix any $r \in \mathbb{Q} \cap (0, 1)$ and rewrite the closure condition (6) as

$$\xi_0^2 = \frac{(\xi_1 + \xi_2)^2}{r^2} - \csc^2\theta_0 \xi_1^2 - \sec^2\theta_0 \xi_2^2.$$

We can always find initial conditions satisfying this. Indeed, in Case 3 we can take any nonzero ξ_1 and ξ_2 and then take θ_0 to maximize the right-hand side: $\tan^2\theta_0 = |\xi_1/\xi_2|$. If ξ_1 and ξ_2 have the same sign, the right-hand side is always positive. If ξ_1 and ξ_2 have opposite signs, we need

$$\left| \frac{\xi_1 + \xi_2}{\xi_1 - \xi_2} \right| > r,$$

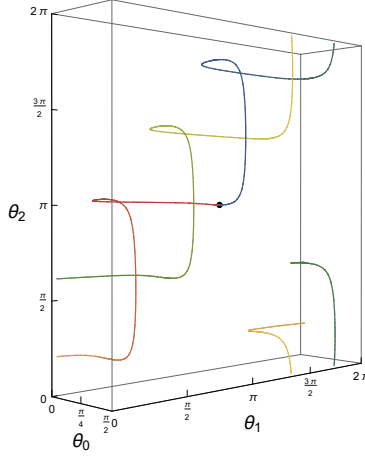


Figure 5. An example with $r = \frac{1}{5}$.

which is only valid for certain ξ_1 and ξ_2 . Case 4 is similar. Then we can solve for ξ_0 , use those numbers as the initial conditions in Hamilton's equations, and then plot the closed sR geodesic. Taking, for example, $r = \frac{1}{5}$, $\xi_1 = 0.6$, and $\xi_2 = 0.7$ we get the sR geodesic in Figure 5.

5. The sR length spectrum

To calculate the lengths of the closed sR geodesics we again only need to consider Cases 3 and 4 (the cases where θ_0 oscillates). We found in the previous section that an sR geodesic is closed when the period of the H_1 -flow and the period of the H_V -flow are commensurable. Then we have

$$\text{period of } H\text{-flow} = p \times \frac{\pi}{|\xi_1 + \xi_2|} = q \times \frac{\pi}{\sqrt{2H_1}} \quad (7)$$

for the least such integers $0 < p < q$ where p, q are either both odd or both even. Since we know the speed of the sR geodesic is a constant $\sqrt{2H}$, we have that the length is

$$\text{length} = \text{period} \times \text{speed} = \frac{\pi q}{\sqrt{2H_1}} \times \sqrt{2H} = \pi q \sqrt{\frac{H}{H_1}} = \pi \sqrt{q^2 - p^2} \quad (8)$$

for the least integers $0 < p < q$ satisfying (7) where p, q are either both odd or both even.

We have another formulation of length that explains the repeating patterns seen in the figures. We know that the distance traveled in one θ_0 -period is

$$\text{period}(\theta_0) \times \text{speed} = \pi \sqrt{\frac{H}{H_1}}.$$

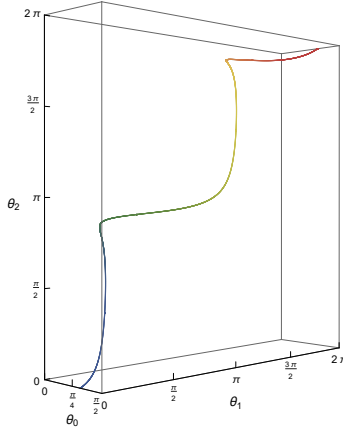


Figure 6. A Riemannian geodesic in Hopf coordinates.

Thus the length of a closed sR geodesic is

$$\text{length} = \pi \times (\text{number of } \theta_0\text{-oscillations}) \times \sqrt{\frac{H}{H_1}}.$$

Comparing with (8), we find that

$$\text{number of } \theta_0\text{-oscillations} = q.$$

Moreover, we see from Hamilton's equations that the curve segments traced out by θ_0 -oscillations are congruent to each other. A similar argument shows that Riemannian geodesics in Hopf coordinates consist of two θ_0 -oscillations, as illustrated in Figure 6.

To summarize, we have found that if an sR geodesic is closed then the initial conditions must satisfy

$$\sqrt{1 - \frac{H}{H_1}} = \frac{|\xi_1 + \xi_2|}{\sqrt{2H_1}} \in \mathbb{Q} \cap (0, 1)$$

and that the length of the closed sR geodesic is

$$\text{length} = \pi \sqrt{q^2 - p^2}$$

for the least integers $0 < p < q$ satisfying (7) where p, q are either both odd or both even.

In fact, every such number is attained as a length; we simply follow the procedure:

- (i) Choose any $p/q \in \mathbb{Q} \cap (0, 1)$, with $\gcd(p, q) = 1$.
- (ii) As seen at the end of Section 4, we can choose initial conditions so that

$$\frac{p}{q} = \sqrt{1 - \frac{H}{H_1}} = \frac{|\xi_1 + \xi_2|}{\sqrt{2H_1}}.$$

Thus

$$p \times \frac{\pi}{|\xi_1 + \xi_2|} = q \times \frac{\pi}{\sqrt{2H_1}}.$$

(iii) If p and q are both odd, then the sR geodesic with those initial conditions has length $\pi\sqrt{q^2 - p^2}$. If one of $\{p, q\}$ is odd and the other is even, the sR geodesic with those initial conditions has length $2\pi\sqrt{q^2 - p^2}$.

Thus the length spectrum consists of 2π and the numbers

$$\pi\sqrt{q^2 - p^2},$$

where $0 < p < q$ are odd integers with $\gcd(p, q) = 1$, and

$$2\pi\sqrt{q^2 - p^2},$$

where $0 < p < q$ are integers, one odd and the other even, with $\gcd(p, q) = 1$.

We now give an alternative characterization of these numbers. It is simpler to work with squares of lengths divided by π^2 . Then we wish to characterize the set S of numbers consisting of 4 and

$$\epsilon(q^2 - p^2),$$

where $0 < p < q$ are integers with $\gcd(p, q) = 1$ and

$$\epsilon = \begin{cases} 1 & \text{if } p \text{ and } q \text{ are both odd,} \\ 4 & \text{if one of } p, q \text{ is odd and the other is even.} \end{cases}$$

In the $\epsilon = 1$ case we take the examples $p = 2k - 1$ and $q = 2k + 1$, $k \in \mathbb{N}$, to get

$$q^2 - p^2 = 4(2k), \quad k \in \mathbb{N}.$$

In the $\epsilon = 4$ case, we take the examples $p = k$ and $q = k + 1$, $k \in \mathbb{N}$, to get

$$4(q^2 - p^2) = 4(2k + 1), \quad k \in \mathbb{N}.$$

This shows that $4\mathbb{N} \subset S$. Now suppose that $n \in S$ and $4 \nmid n$. Then clearly n can only be in the $\epsilon = 1$ case, so there would be odd integers $0 < p < q$ with $\gcd(p, q) = 1$ such that $n = q^2 - p^2$. This is easily seen to be impossible. Thus in fact $4\mathbb{N} = S$.

We note that if $n \in S$ and $8 \mid n$, then n cannot be in the $\epsilon = 4$ case, and that if $n \in S$ and $n = 4(2k + 1)$, $k \in \mathbb{N}$, then n cannot be in the $\epsilon = 1$ case. Both of these statements easily follow from parity arguments.

Converting back to the language of lengths, we find that the set of lengths of the closed sR geodesics is

$$\{2\pi\sqrt{n} : n \in \mathbb{N}\}.$$

By the previous paragraph, odd n correspond to “full periods” of the H_V -flow and geodesic flow (the $\epsilon = 4$ case), and even n correspond to “half periods” of both the H_V -flow and geodesic flow (the $\epsilon = 1$ case).

6. The spectrum of the sub-Laplacian

The sub-Laplacian $-\Delta_{\text{sR}}$ has a compact resolvent, and hence has a pure discrete spectrum $0 < \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n \leq \dots$, with $\lambda_n \rightarrow +\infty$ as $n \rightarrow +\infty$, and a complete orthonormal set of eigenfunctions. (See, for example, the recent paper [Colin de Verdière et al. 2016].) In fact, in the case of S^3 , the eigenfunctions of the sub-Laplacian are the same as the eigenfunctions of the Laplacian. We recall that $\Delta_\lambda = E_1^2 + E_2^2 + \lambda^{-2}V^2$ is the λ -penalty Laplacian, with $\lambda = 1$ giving the Riemannian Laplacian on the sphere Δ_{S^3} , and $\lambda = \infty$ giving the sub-Laplacian on the sphere. In Hopf coordinates we have

$$\Delta_{\text{sR}} = E_1^2 + E_2^2 = \frac{1}{\sin(2\theta_0)} \frac{\partial}{\partial \theta_0} \circ \sin(2\theta_0) \frac{\partial}{\partial \theta_0} + \left(\cot \theta_0 \frac{\partial}{\partial \theta_1} - \tan \theta_0 \frac{\partial}{\partial \theta_2} \right)^2.$$

It is easy to see that $V = \partial/\partial\theta_1 + \partial/\partial\theta_2$ commutes with Δ_{sR} ; hence Δ_{sR} commutes with Δ_{S^3} . Thus Δ_{sR} and Δ_{S^3} have a *common* complete orthonormal set of eigenfunctions [Dirac 1947; von Neumann 1955]; the eigenfunctions of Δ_{sR} are simply the spherical harmonics.

Particularly noteworthy is $(x_1 + iy_1)^k = \sin^k \theta_0 e^{ik\theta_1}$. It is a ‘‘Gaussian beam’’: a family of eigenfunctions of both Δ_{S^3} and Δ_{sR} that concentrates along a great circle. Zelditch [2016, pp. 185–186] singles out this example in the Riemannian setting. It would be interesting to see if it is possible to construct, localized to each sR geodesic, a quasimode or Gaussian beam in the spirit of [Ralston 1976; 1977].

Taylor [1986] used the Peter–Weyl theorem to find the eigenvalues of Δ_{sR} ; Domokos [2015] generalized, using subelliptic Peter–Weyl and Plancherel theorems on compact, connected, semisimple Lie groups. To summarize, the eigenvalues of $-\Delta_{S^3}$ are $m(m+2)$ for $m \in \{0, 1, 2, \dots\}$, and the eigenvalues of $-\Delta_{\text{sR}}$ are (for the same m ; the operators have the same complete orthonormal set of eigenfunctions)

$$4mj - 4j^2 + 2m, \quad j \in \{0, 1, 2, \dots, m\}.$$

For reference, the eigenvalues of the λ -penalty Laplacian

$$-\Delta_\lambda = -\Delta_{\text{sR}} - \lambda^{-2}V^2$$

are

$$(1 - \lambda^{-2})4j(m - j) + m(2 + \lambda^{-2}m),$$

for $m \in \{0, 1, 2, \dots\}$ and $j \in \{0, 1, 2, \dots, m\}$.

At this point we will not conjecture a general formula relating the sR length spectrum of a bracket-generating compact sR manifold (which for S^3 is $\{2\pi\sqrt{n} : n \in \mathbb{N}\}$) to the set of eigenvalues of the sub-Laplacian counted with or without multiplicities (which for S^3 is $\{2m : m = 0, 1, 2, \dots\}$).

Acknowledgement

This work was supported by a SURE (Summer Undergraduate Research Experience) Award at California State University, Sacramento.

References

- [Burago et al. 2001] D. Burago, Y. Burago, and S. Ivanov, *A course in metric geometry*, Graduate Studies in Mathematics **33**, American Mathematical Society, Providence, RI, 2001. MR Zbl
- [Calin et al. 2009] O. Calin, D.-C. Chang, and I. Markina, “SubRiemannian geometry on the sphere \mathbb{S}^3 ”, *Canad. J. Math.* **61**:4 (2009), 721–739. MR Zbl
- [Cannas da Silva 2008] A. Cannas da Silva, *Lectures on symplectic geometry*, 2nd ed., Lecture Notes in Mathematics **1764**, Springer, 2008.
- [Chang et al. 2009] D.-C. Chang, I. Markina, and A. Vasil’ev, “Sub-Riemannian geodesics on the 3-D sphere”, *Complex Anal. Oper. Theory* **3**:2 (2009), 361–377. MR
- [Chang et al. 2011] D.-C. Chang, I. Markina, and A. Vasil’ev, “Hopf fibration: geodesics and distances”, *J. Geom. Phys.* **61**:6 (2011), 986–1000. MR
- [Colin de Verdière et al. 2016] Y. Colin de Verdière, L. Hillairet, and E. Trélat, “Quantum ergodicity and quantum limits for sub-Riemannian Laplacians”, exposé 20, p. 17 in *Séminaire Laurent Schwartz: EDP et applications*, 2014–2015, Éc. Polytech., Palaiseau, 2016. MR Zbl
- [D’Angelo and Tyson 2010] J. P. D’Angelo and J. T. Tyson, “An invitation to Cauchy–Riemann and sub-Riemannian geometries”, *Notices Amer. Math. Soc.* **57**:2 (2010), 208–219. MR Zbl
- [Dirac 1947] P. A. M. Dirac, *The principles of quantum mechanics*, 3rd ed., Oxford University Press, 1947. MR Zbl
- [Domokos 2015] A. Domokos, “Subelliptic Peter–Weyl and Plancherel theorems on compact, connected, semisimple Lie groups”, *Nonlinear Anal.* **126** (2015), 131–142. MR Zbl
- [Duistermaat and Guillemin 1975] J. J. Duistermaat and V. W. Guillemin, “The spectrum of positive elliptic operators and periodic bicharacteristics”, *Invent. Math.* **29**:1 (1975), 39–79. MR Zbl
- [Feynman 1985] R. Feynman, *Surely you’re joking, Mr. Feynman!: adventures of a curious character*, W. W. Norton, New York, 1985.
- [Guillemin and Weinstein 1976] V. Guillemin and A. Weinstein, “Eigenvalues associated with a closed geodesic”, *Bull. Amer. Math. Soc.* **82**:1 (1976), 92–94. Correction in **82**:6 (1976), 966. MR Zbl
- [Hurtado and Rosales 2008] A. Hurtado and C. Rosales, “Area-stationary surfaces inside the sub-Riemannian three-sphere”, *Math. Ann.* **340**:3 (2008), 675–708. MR Zbl
- [Montgomery 1994] R. Montgomery, “Abnormal minimizers”, *SIAM J. Control Optim.* **32**:6 (1994), 1605–1620. MR Zbl
- [Montgomery 2002] R. Montgomery, *A tour of subriemannian geometries, their geodesics and applications*, Mathematical Surveys and Monographs **91**, American Mathematical Society, Providence, RI, 2002. MR Zbl
- [Nelson 1967] E. Nelson, *Tensor analysis*, Princeton University Press, 1967. Zbl
- [von Neumann 1955] J. von Neumann, *Mathematical foundations of quantum mechanics*, Princeton University Press, 1955. MR Zbl
- [Ralston 1976] J. V. Ralston, “On the construction of quasimodes associated with stable periodic orbits”, *Comm. Math. Phys.* **51**:3 (1976), 219–242. MR Zbl

- [Ralston 1977] J. V. Ralston, “Approximate eigenfunctions of the Laplacian”, *J. Differential Geometry* **12**:1 (1977), 87–100. MR Zbl
- [Taylor 1986] M. E. Taylor, *Noncommutative harmonic analysis*, Mathematical Surveys and Monographs **22**, American Mathematical Society, Providence, RI, 1986. MR Zbl
- [Wikipedia 2015] “3-sphere”, Wikipedia entry, 2015, Available at <https://en.wikipedia.org/wiki/3-sphere>.
- [Woods 1934] F. S. Woods, *Advanced calculus*, Ginn, Boston, 1934. Zbl
- [Zelditch 2016] S. Zelditch, “Park City lectures on eigenfunctons”, pp. 111–193 in *Geometric analysis*, edited by H. L. Bray et al., IAS/Park City Math. Ser. **22**, American Mathematical Society, Providence, RI, 2016. MR Zbl

Received: 2017-03-24 Accepted: 2018-03-06

dtk22@csus.edu

Department of Mathematics and Statistics, California State University, Sacramento, Sacramento, CA, United States

mjv@csus.edu

Department of Mathematics and Statistics, California State University, Sacramento, Sacramento, CA, United States

Statistics for fixed points of the self-power map

Matthew Friedrichsen and Joshua Holden

(Communicated by Anant Godbole)

The map $x \mapsto x^x$ modulo p is related to a variation of the ElGamal digital signature scheme in a similar way as the discrete exponentiation map, but it has received much less study. We explore the number of fixed points of this map by a statistical analysis of experimental data. In particular, the number of fixed points can in many cases be modeled by a binomial distribution. We discuss the many cases where this has been successful, and also the cases where a good model may not yet have been found.

1. Introduction and motivation

The security of the ElGamal digital signature scheme against selective forgery relies on the difficulty of solving the congruence $g^{H(m)} \equiv y^r r^s \pmod{p}$ for r and s , given m , g , y , and p but not knowing the discrete logarithm of y modulo p to the base g . (We assume for the moment the security of the hash function $H(m)$.) Similarly, the security of a certain variation of this scheme given in, e.g., [Menezes et al. 1997, Note 11.71] relies on the difficulty of solving

$$g^{H(m)} \equiv y^s r^r \pmod{p}. \quad (1)$$

It is generally expected that the best way to solve either of these congruences is to calculate the discrete logarithm of y , but this is not known to be true. In particular, another possible option would be to choose s arbitrarily and solve the relevant equation for r . In the case of (1), this boils down to solving equations of the form $x^x \equiv c \pmod{p}$. We will refer to these equations as “self-power equations”, and we will call the map $x \mapsto x^x$ modulo p the “self-power map”. This map has been studied in various forms in [Anghel 2013; 2016; Balog et al. 2011; Cilleruelo and Garaev 2016a; 2016b; Crocker 1966; 1969; Somer 1981; Holden 2002a; 2002b; Holden and Moree 2006; Friedrichsen et al. 2010; Holden and Robinson 2012;

MSC2010: primary 11Y99; secondary 11-04, 11T71, 94A60, 11A07, 11D99.

Keywords: self-power map, exponential equation, ElGamal digital signatures, fixed point, random map, number theory.

[Kurlberg et al. 2015]. In this work we will investigate experimentally the number of fixed points of the map, i.e., solutions to

$$x^x \equiv x \pmod{p} \quad (2)$$

between 1 and $p - 1$. In particular, we would like to know whether the distribution across various primes behaves as we would expect if the self-power map were a “random map”. We do this by creating a model in which values of a map are assumed to occur uniformly randomly except as forced by the structure of the self-power map. We can then predict the distribution of the number of fixed points of this random map and compare it statistically to the actual self-power map. If there is “nonrandom” structure in the self-power map, it may be possible to exploit that structure to break the signature scheme mentioned above or others like it.

In this paper, we will give a general heuristic (based on Heuristic 1 below) for the number of fixed points of the self-power map and show that for most cases it appears to accurately predict the behavior of the map. The outlying cases mostly appear to involve elements with order d that are relatively small or large compared to p . We will first show that the number of fixed points for elements with orders 1, 2, $p - 1$, and $(p - 1)/2$ can be predicted exactly. For other small orders which largely don’t follow the general heuristic, we specifically look at the orders 3, 4, and 6 and give a separate model for them. For large orders, we make predictions for the orders $(p - 1)/3$ and $(p - 1)/4$.

Some theoretical work has also been done on bounding the possible number of fixed points of the self-power map. If we denote the number of solutions to (2) which fall between 1 and $p - 1$ by $F(p)$, then we have:

Theorem 1.1 [Cilleruelo and Garaev 2016b, Corollary 2]. *For some absolute constant $c > 0$,*

$$F(p) \leq p^{1/3-c+o(1)}$$

as $p \rightarrow \infty$.

Remark 1. The corollary in [Cilleruelo and Garaev 2016b] is more general and puts a bound on the number of solutions for $x^{f(x)} \equiv 1 \pmod{p}$ for any nonconstant polynomial in $\mathbb{Z}[x]$ without multiple roots in \mathbb{C} .

Remark 2. In the related case of solutions to $x^x \equiv 1 \pmod{p}$, [Cilleruelo and Garaev 2016a] shows that the exponent can be taken to be $\frac{27}{82} + o(1)$ and that is likely also the case here.

As far as a lower bound, every p has at least $x = 1$ as a solution to (2), and at least some primes have only this solution. However, while [Kurlberg et al. 2015; Felix and Kurlberg 2017] give good reason to believe that there are infinitely many such primes, they also prove that these primes are fairly rare:

Theorem 1.2 [Felix and Kurlberg 2017, Corollary 1.2]. *Let $\pi(N)$ be the number of primes less than or equal to N as usual. Let $\mathcal{A}(N)$ denote the set of primes less than or equal to N such that $F(p) = 1$. Then*

$$\#\mathcal{A}(N) \leq \frac{\pi(N)}{(\ln \ln \ln N)^{1-1/e+o(1)}}$$

as $N \rightarrow \infty$.

2. Models and experimental results

Heuristics and normality. Theorem 1.1 gives us a range in which the number of fixed points $F(p)$ can lie, but does not say anything about the distribution of the values within that range. As described above, our goal is to create a random model for the self-power map much like was done for the discrete exponential map in [Holden 2002a; 2002b; Holden and Moree 2006]. Our first attempt assumed that $F(p)$ was normally distributed around the predicted value $\sum_{d|(p-1)} \phi(d)/d$. (The normality assumption had been successfully used for the discrete exponential map in, e.g., [Cloutier and Holden 2010]; see also [Holden and Lindle 2008]. Furthermore, it appeared to be justified by the central limit theorem, given the number of primes we were intending to test.)

In order to calculate the variance of $F(p)$, we use the following heuristic, which is related to those in [Holden and Moree 2006, Section 6], and can also be derived from the assumptions in [Kurlberg et al. 2015, Section 4.1].

Heuristic 1. The map $x \mapsto x^x \bmod p$ is a random map in the sense that for all p , if x, y are chosen uniformly at random from $\{1, \dots, p-1\}$ with $\text{ord}_p x = d$, then

$$\Pr[x^x \equiv y \pmod{p}] \approx \begin{cases} 1/d & \text{if } \text{ord}_p y \mid d, \\ 0 & \text{otherwise.} \end{cases}$$

As some justification, one can use the methods of [Holden and Robinson 2012, Corollary 6.2] to prove the following lemma. This shows that the heuristic holds exactly over the range $1 \leq x \leq (p-1)p$ rather than $1 \leq x \leq p-1$:

Lemma 2.1. *For all p , given fixed $d \mid (p-1)$ and fixed $y \in \{1, \dots, (p-1)p\}$, $p \nmid y$, such that $\text{ord}_p y \mid d$, we have*

$$\#\{x \in \{1, \dots, (p-1)p\} : p \nmid x, x^x \equiv y \pmod{p}, \text{ord}_p x = d\} = (p-1) \frac{\phi(d)}{d}.$$

Similar methods are used in [Holden et al. 2016] to prove the following theorem:

Theorem 2.2 [Holden et al. 2016, Corollary 4]. *Let $G(p)$ be the number of solutions to (2) with $1 \leq x \leq (p-1)p$ and $p \nmid x$. Then*

$$G(p) = (p-1) \sum_{n \mid (p-1)} \frac{\phi(n)}{n}.$$

For more on the self-power map over the range $1 \leq x \leq (p-1)p$, see [Somer 1981, Theorem 1; Holden and Robinson 2012, Sections 6 and 7; Holden et al. 2016].

As far as using Heuristic 1, note that it implies that the “experiment” of testing whether x is a fixed point behaves as a Bernoulli trial. Let $F_d(p)$ be the number of solutions to (2) with $1 \leq x \leq p-1$ and $\text{ord}_p x = d$. Assuming independence of the Bernoulli trials (which is not completely accurate, as we shall see), $F_d(p)$ is distributed as a binomial random variable with $\phi(d)$ trials and success probability $1/d$. (We denote by $\phi(d)$ the Euler ϕ function and it occurs here because it gives the number of elements with order d when $d \mid (p-1)$.)

This distribution has mean $\phi(d)/d$, as expected, and variance $\phi(d)(d-1)/d^2$. Summing over $d \mid (p-1)$ gives the predicted mean and variance of $F(p)$.

We tested the hypothesis that $F(p)$ was normal with this mean and variance by collecting data for 16,405 primes from 100,003 to 299,993 and 10,314 primes from 1,000,003 to 1,142,971. The number of fixed points for each prime was determined using C code originally written by Cloutier [Cloutier and Holden 2010] and modified by Lindle [2008], Hoffman [2009], and Friedrichsen, Larson, and McDowell in [Friedrichsen et al. 2010]. Postprocessing was done using a Python script written by the first author. This data set combined a preliminary set of data from code run on servers maintained by the Rose-Hulman Computer Science & Software Engineering and Mathematics Departments and data from code run on the Tufts High Performance Computing Cluster. The code took a few hours of computational time, with about a day postprocessing work to fully put together the data sets. The postprocessing was the limiting factor in the number of primes we could feasibly work with.

Once the values of $F(p)$ were collected, they were normalized to a z -statistic by subtracting the predicted mean and dividing by the predicted standard deviation (square root of the variance). The z -statistics were grouped separately for the six-digit and seven-digit primes and tested to see if they conformed to the expected standard normal distribution. As you can see in Figures 1 and 2, the distributions appear to be roughly normal to the naked eye, and the standard deviations are close to 1 as expected. The means are a little higher than the expected 0, and there are a few bars which seem significantly off, but these features could be attributed to certain known properties which appear below in Theorem 2.3.

More troubling is the lack of normality revealed by probability plots in Figures 3 and 4. Perfectly normal distributions would lie along the diagonal lines in these figures, and Ryan–Joiner tests confirm that it is very unlikely that $F(p)$ is obeying a normal distribution for these primes. In fact there appear to be more primes in the “tails” than expected, that is, a larger than expected number of primes with significantly more or fewer fixed points than expected. Felix and Kurlberg [2017, Section 1.2] studied the same phenomena with two data sets comprised of seven-digit and ten-digit primes, respectively. They also broke up each data set into

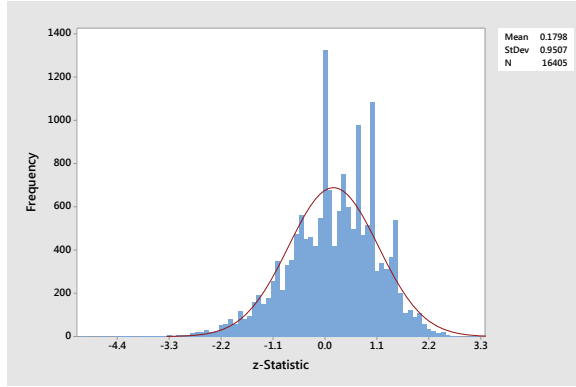


Figure 1. Histogram of z -statistics for six-digit primes.

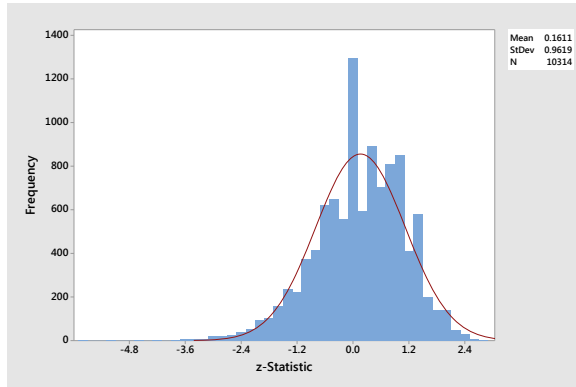


Figure 2. Histogram of z -statistics for seven-digit primes.

different subgroups based on the number of unique prime divisors of $p - 1$. Their analysis matches ours, including a deviation from the binomial model at the tails.

Binomial distribution and goodness of fit. Some modification of the code by the first author allowed us to collect the values of $F_d(p)$ for the same primes as above, in order to see if particular orders were behaving less “randomly” than others. We excluded certain orders where $F_d(p)$ is known to behave predictably:

Theorem 2.3. (1) $F_1(p) = 1$ for all p .

(2) $F_2(p) = 0$ for all p .

(3) $F_{p-1}(p) = 0$ for all p .

(4) $F_{(p-1)/2}(p) = \begin{cases} 0 & \text{if } p \equiv 3 \text{ or } 5 \pmod{8}, \\ & \text{or if } p \equiv 1 \text{ or } 7 \pmod{8} \text{ and } \text{ord}_p 2 \neq (p-1)/2; \\ 1 & \text{if } p \equiv 1 \text{ or } 7 \pmod{8} \text{ and } \text{ord}_p 2 = (p-1)/2. \end{cases}$

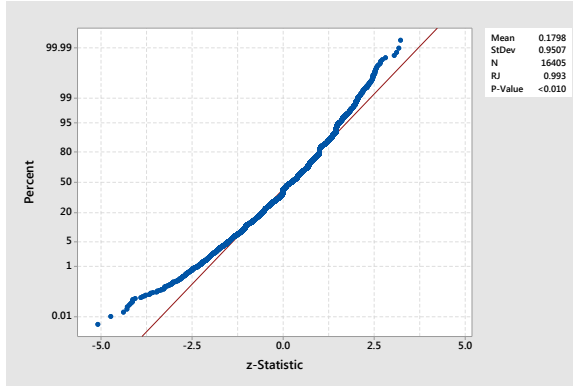


Figure 3. Probability plot of z -statistics for six-digit primes.

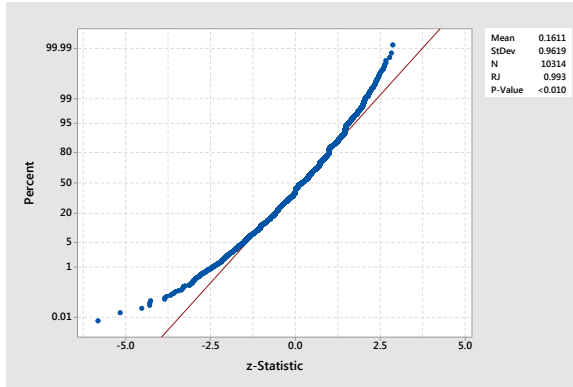


Figure 4. Probability plot of z -statistics for seven-digit primes.

To prove this we use the following lemmas:

Lemma 2.4 [Friedrichsen et al. 2010, Proposition 7]. *Let p be prime. The number x is a solution to (2) if and only if $x \equiv 1 \pmod{\text{ord}_p x}$.*

Corollary 2.5. *Let $d \mid (p - 1)$. The solutions to (2) of order d are exactly the elements of $\mathcal{P} = \{1, d + 1, 2d + 1, \dots, p - d\}$ which have order d .*

Proof of Theorem 2.3. Parts (1) and (2) are clear from the definition. Part (3) is Proposition 6 of [Friedrichsen et al. 2010]. If x is a fixed point such that $\text{ord}_p x = (p - 1)/2$, then Corollary 2.5 implies $x = (p + 1)/2$. Then Proposition 2 of [Friedrichsen et al. 2010] tells us x is a fixed point if and only if 2 is a quadratic residue modulo p , which is if and only if $p \equiv 1$ or $7 \pmod{8}$. Combining this with the fact that $\text{ord}_p(p + 1)/2 = \text{ord}_p 2$ gives part (4). \square

Remark 3. Note that the behavior of fixed points in safe primes, that is, primes where $(p - 1)/2$ is also prime, is completely explained by Theorem 2.3. Safe primes

are important for discrete logarithm-based algorithms because the group $(\mathbb{Z}/p\mathbb{Z})^\times$ will have a subgroup with large prime order. Specifically, it will have a subgroup with order $(p-1)/2$.

We collected values of $F_d(p)$ for each prime and each value of $d \mid (p-1)$ other than $d = 1, 2, p-1$, and $(p-1)/2$. We then attempted to normalize this data, but the resulting z -statistics turned out to be too highly clustered and did not resemble normal data. We therefore decided to do a chi-squared goodness-of-fit test on the data. We used the formula for the mass function of a binomial distribution to predict:

Prediction 1.
$$\Pr[F_d(p) = k] = \binom{\phi(d)}{k} \left(\frac{1}{d}\right)^k \left(\frac{d-1}{d}\right)^{\phi(d)-k}.$$

We chose to use the categories $k = 0, k = 1, k = 2$, and $k > 2$ for our test in order to make sure the categories with large k did not get too small. We summed the predictions over p and d for each of the categories and compared them with the observed numbers of p and d which fell into each category. An initial test using only the primes between 100,003 and 102,677 gave a chi-squared statistic of 4.66 and a statistical p -value of 0.198.¹ Using the common cutoff of 0.05 for statistical significance of p -values, we do not see statistical evidence that our predictions are incorrect. However, using the full set of primes between 100,003 and 299,993 gave a much larger chi-squared statistic of 491.14 and a p -value of less than 10^{-100} .

We hypothesized that not all values of p and d fit the predictions equally well. We tested this by sorting in various ways the values of $F_d(p)$ collected for p between 100,003 and 102,667, and $d \mid (p-1)$ other than $d = 1, 2, p-1$, and $(p-1)/2$. After each sort, we calculated the chi-squared statistics and p -values for a sliding window of 100 values, with predictions and observations calculated as above. (The size of the window was chosen in order to make sure there were enough data points in the window for the chi-squared test to be valid.)

The strongest evidence of a pattern was seen when the data was sorted by value of d . This was confirmed for the full range of primes between 100,003 and 299,993, as can be seen in Figure 5. For data randomly generated according to the relevant binomial distributions, p -values should be evenly distributed between 0 and 1. When p -values are biased towards 0 it indicates statistically significant divergence from the predicted distributions. In other words, dots on the same (approximate) horizontal line should be evenly distributed between the left- and right-hand sides of the graph. (Note that the value of d used to place the dot on the plot is the largest value of d in the window of 100 pairs, so some dots would more accurately “belong” to more than one line.) Horizontal lines where the dots are clustered towards the left-hand side indicate statistically significant divergence.

¹We will use the term “ p -value” in this paper when referring to the statistical concept in order to distinguish it from use of p to indicate a prime.

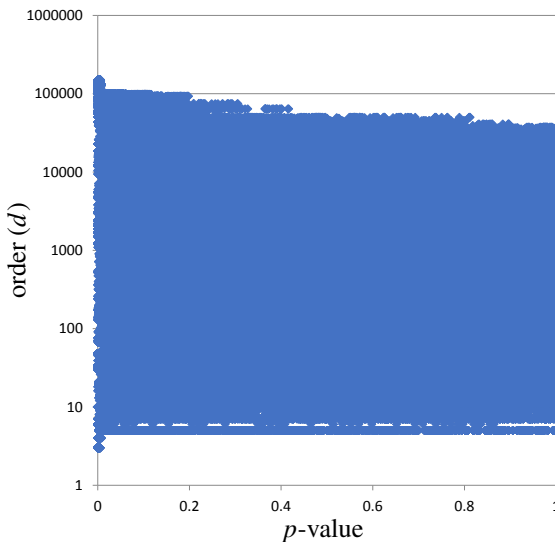


Figure 5. Logarithmic plot showing p -values of the sliding-window goodness-of-fit test, data sorted by order, for six-digit primes.

As you can see, the strongest divergence from the predictions occurs with particularly small and particularly large values of d . (Since the value of d used to place the dot on the plot is the largest value in the window, the effect for small d is even larger than it appears in the plot.) We therefore looked for theoretical explanations of these effects. We observed two significant properties that affected whether or not a given order d followed the formula in Prediction 1. The first is the size of $\phi(d)$ and the second is the size of the set $\mathcal{P} = \{1, d + 1, 2d + 1, \dots, p - d\}$. On the smaller end of the spectrum, the size of $\phi(d)$ is the most influential. On the larger end, the size of the set \mathcal{P} is the most influential. In the next section, we will discuss specific examples of both small and large orders.

3. Small and large orders

Small orders. For $d = 3$ we observed that while $F_3(p) = 2$ should occur roughly one-ninth of the time according to Prediction 1, it never occurred at all in our data. A similar but less striking effect was observed for $d = 4$, while for $d = 6$ it was $F_6(p) = 1$ which was never observed, despite Prediction 1 saying it should happen over one-quarter of the time. It turns out that there is a significant lack of independence in the fixed points for these orders, as we were able to show.

Theorem 3.1. (1) $F_3(p) = 0$ or $F_3(p) = 1$ for all p such that $3 \mid (p - 1)$.

(2) $F_4(p) = 0$ or $F_4(p) = 1$ for all p such that $4 \mid (p - 1)$.

(3) $F_6(p) = 0$ or $F_6(p) = 2$ for all p such that $6 \mid (p - 1)$.

Proof. If $3 \mid (p-1)$, then by Lemma 2.4 the fixed points of order 3 are exactly the elements congruent to 1 modulo 3. In this case there are two elements of order 3, and a direct computation shows that if x is one of them, then $p-1-x$ is the other. Thus the elements of order 3 add up to $p-1 \equiv 0 \pmod{3}$. So at most one of the elements of order 3 can be a fixed point, proving part (1). Part (2) is similar except that the elements of order 4 add up to $p \equiv 1 \pmod{4}$. In part (3) the elements of order 6 add up to $p+1 \equiv 2 \pmod{6}$ so if one is a fixed point then the other must be also. \square

The following lemma says that the elements of a given order f are approximately uniformly distributed across the residue classes modulo any given r .

Lemma 3.2. *Let a, r , and f be positive integers such that $0 \leq a < r \leq p-1$ and $f \mid (p-1)$. Let*

$$\mathcal{Q} = \left\{ a, r+a, 2r+a, \dots, \left\lfloor \frac{p-1-a}{r} \right\rfloor r + a \right\}.$$

Let $\mathcal{Q}' = \{x \in \mathcal{Q} : \text{ord}_p(x) = f\}$. Then

$$\left| \#\mathcal{Q}' - \frac{\phi(f)}{r} \right| \leq 1 + \tau(f)\sqrt{p}(1 + \ln p),$$

where $\tau(f)$ is the number of divisors of f .

Proof. The proof is the same as the proof of equation (7) from [Cobeli and Zaharescu 1999] with the order equal to f instead of $p-1$. \square

In particular, we would expect the elements of order d to be equally likely to be of any residue class modulo d . Since Theorem 3.1 shows that the fixed points of orders $d=3$ and $d=4$ are entirely determined by their residue classes modulo d , this leads us to predict:

- Prediction 2.** (1) $\Pr[F_3(p) = 0] = \frac{1}{3}$ and $\Pr[F_3(p) = 1] = \frac{2}{3}$.
 (2) $\Pr[F_4(p) = 0] = \frac{1}{2}$ and $\Pr[F_4(p) = 1] = \frac{1}{2}$.
 (3) $\Pr[F_6(p) = 0] = \frac{5}{6}$ and $\Pr[F_6(p) = 2] = \frac{1}{6}$.

This is in fact what we observe in the data, as shown in Figure 6. This figure shows the number of primes such that $d \mid (p-1)$ for $d=3, 4$, and 6 , the number of primes for each d with $F_d(p) = 0, 1$, and 2 , and the p -value given by a chi-squared test against the distribution predicted above. We do not see statistical evidence that our predictions are incorrect.

Remark 4. Not all small orders seem to exhibit this lack of independence in a statistically significant way. For example, $d=5$ fits the distribution of the original model with $p = 0.90$ and $d=7$ fits with $p = 0.48$. However, $d=8$, $d=12$, and $d=18$ do not appear to fit the original model. For $d=8$ and $d=12$ the four elements of order d come in pairs which each have a dependence similar to

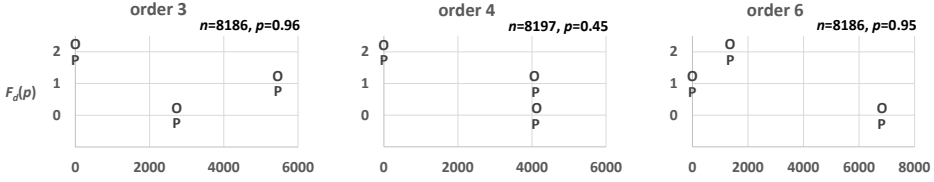


Figure 6. Predicted (P) and observed (O) numbers of primes for fixed points of orders 3, 4, and 6 in six-digit primes.

that for order 4, but we have not worked out the exact model. Other values of d which are multiples of 4 also have dependent pairs but the effect is apparently not large enough to be detected in our data. For $d = 9$ and $d = 18$ the six elements of order d come in two sets of three which each add up to 0 modulo p , producing a dependence pattern related to the ones for orders 3 and 6. We have not worked out the exact model, and it is not clear why the results are statistically significant for $d = 18$ but not $d = 9$. It may be due to chance.

Large orders. We also observed significant deviation from our predictions in the case of large orders. Recall that part (4) of Theorem 2.3 used Proposition 2 of [Friedrichsen et al. 2010] to prove that there was at most one fixed point of order $(p-1)/2$. In fact, that proposition also showed that the fixed point exists if and only if 2 is a quadratic residue modulo p . Similarly, if $3 \mid (p-1)$ then Corollary 2.5 shows that there are at most two fixed points of order $(p-1)/3$, namely $(p+2)/3$ and $(2p+1)/3$. Using methods similar to the above we can show that these residue classes will be fixed points when they are cubic residues modulo p .

Proposition 3.3. *Let p be a prime number equivalent to 1 modulo 3. The residue class $(p+2)/3$ is a fixed point if and only if it is a cubic residue modulo p , and similarly for $(2p+1)/3$.*

Proof. Note that since $1 \leq x \leq p-1$, (2) is equivalent to

$$x^{x-1} \equiv 1 \pmod{p}. \quad (3)$$

Then $(p+2)/3$ is a fixed point if and only

$$\left(\frac{p+2}{3}\right)^{(p-1)/3} \equiv 1 \pmod{p},$$

which by Euler's criterion is equivalent to $(p+2)/3$ being a cubic residue.

Similarly, if $(2p+1)/3$ is a fixed point then

$$\left(\frac{2p+1}{3}\right)^{(2p-2)/3} \equiv 1 \pmod{p}.$$

But then

$$\left(\frac{2p+1}{3}\right)^{(p-1)/3} \equiv \left(\frac{2p+1}{3}\right)^{(4p-4)/3} \equiv 1 \pmod{p}$$

also, where the first equivalence is just Fermat's little theorem. So Euler's criterion is satisfied again. Conversely, if

$$\left(\frac{2p+1}{3}\right)^{(p-1)/3} \equiv 1 \pmod{p}$$

then certainly

$$\left(\frac{2p+1}{3}\right)^{(2p-2)/3} \equiv 1 \pmod{p}$$

so $(2p+1)/3$ is a fixed point. \square

More simplifications show that $(2p+1)/3 \equiv 3^{-1} \pmod{p}$ and $(p+2)/3 \equiv 2(3^{-1}) \pmod{p}$ so $(2p+1)/3$ will be a cubic residue whenever 3 is a cubic residue, and both $(p+2)/3$ and $(2p+1)/3$ will be cubic residues when both 2 and 3 are cubic residues. These same methods can be used to show that all numbers of the form $(m(p-1)/k) + 1$ where $1 \leq m < k$ will be fixed points in the self-power map when the number is a k -th residue.

This is not quite enough to investigate $F_{(p-1)/3}(p)$ since not all cubic residues have order equal to $(p-1)/3$. We thus estimate the probability that a given element of $\{(p+2)/3, (2p+1)/3\}$ has order equal to exactly $(p-1)/3$. Lemma 3.2 suggests that elements of order d occur in \mathcal{P} in approximately the same proportion that they occur in the whole range $1 \leq x \leq p-1$, namely $\phi(d)/(p-1)$. (A more precise statement on the frequency of p such that $kd+1$ has order d would appear to require some variation on Artin's primitive root conjecture.)

We again use a binomial distribution to predict:

Prediction 3. (1) $\Pr[F_{(p-1)/3}(p) = 0] = \left(1 - \frac{\phi((p-1)/3)}{p-1}\right)^2$.

(2) $\Pr[F_{(p-1)/3}(p) = 1] = 2 \left(\frac{\phi((p-1)/3)}{p-1}\right) \left(1 - \frac{\phi((p-1)/3)}{p-1}\right)$.

(3) $\Pr[F_{(p-1)/3}(p) = 2] = \left(\frac{\phi((p-1)/3)}{p-1}\right)^2$.

If $4 \mid (p-1)$, Corollary 2.5 shows that there are at most three fixed points of order $(p-1)/4$, namely $(p+3)/4$, $(p+1)/2$, and $(3p+1)/4$. However, it turns out that they cannot all be fixed points at the same time.

Theorem 3.4. *Let p be a prime number equivalent to 1 modulo 4:*

(1) *If $p \equiv 1 \pmod{8}$ and $p \equiv 1 \pmod{3}$, then $F_{(p-1)/4}(p) \leq 2$.*

(2) If $p \equiv 1 \pmod{8}$ and $p \equiv 2 \pmod{3}$, then $F_{(p-1)/4}(p) \leq 1$.

(3) If $p \equiv 5 \pmod{8}$ and $p \equiv 1 \pmod{3}$, then $F_{(p-1)/4}(p) \leq 1$.

(4) If $p \equiv 5 \pmod{8}$ and $p \equiv 2 \pmod{3}$, then $F_{(p-1)/4}(p) = 0$.

Proof. Suppose $p \equiv 1 \pmod{8}$. Since $(p+1)/2 \equiv 2^{-1} \pmod{p}$ and $(3p+1)/4 \equiv 4^{-1} \pmod{p}$, these two can only be both fixed points of order $(p-1)/4$ if $\text{ord}_p 2 = \text{ord}_p 4 = (p-1)/4$. But we know $8 \mid (p-1)$, so if $\text{ord}_p 2 = (p-1)/4$ then $\text{ord}_p 4 = (p-1)/8$. On the other hand, if $p \equiv 5 \pmod{8}$, then we know $\text{ord}_p 2 \nmid (p-1)/2$ so neither $\text{ord}_p 2$ nor $\text{ord}_p 4$ can be $(p-1)/4$. Now, suppose $p \equiv 2 \pmod{3}$. Then $(p+3)/4$ can only be a fixed point if it is a quartic residue. We know $(p+3)/4 = 3(4^{-1})$ and 4^{-1} is a quadratic residue, but 3 is not a quadratic residue. So, $(p+3)/4$ cannot be quartic since it is not quadratic. \square

To make predictions on the probabilities of each number of fixed points, we again use a binomial distribution. If $p \equiv 1 \pmod{8}$, we keep in mind that the orders of $(p+1)/2$ and $(3p+1)/4$ are dependent so we can treat them together. If $p \equiv 1 \pmod{3}$ also, we know that $(p+3)/4$ might be a fixed point, which is independent of the behavior of $(p+1)/2$ and $(3p+1)/4$:

Prediction 4. Assume $p \equiv 1 \pmod{8}$ and $p \equiv 1 \pmod{3}$; i.e., $p \equiv 1 \pmod{24}$:

$$(1) \Pr[F_{(p-1)/4}(p) = 0] = \left(1 - \frac{2\phi((p-1)/4)}{p-1}\right) \left(1 - \frac{3\phi((p-1)/4)}{(p-1)/2}\right).$$

$$(2) \Pr[F_{(p-1)/4}(p) = 1] = \left(\frac{2\phi((p-1)/4)}{p-1}\right) \left(1 - \frac{3\phi((p-1)/4)}{(p-1)/2}\right) + \left(1 - \frac{2\phi((p-1)/4)}{p-1}\right) \left(\frac{3\phi((p-1)/4)}{(p-1)/2}\right).$$

$$(3) \Pr[F_{(p-1)/4}(p) = 2] = \left(\frac{2\phi((p-1)/4)}{p-1}\right) \left(\frac{3\phi((p-1)/4)}{(p-1)/2}\right).$$

Assume $p \equiv 1 \pmod{8}$ and $p \equiv 2 \pmod{3}$; i.e., $p \equiv 17 \pmod{24}$:

$$(1) \Pr[F_{(p-1)/4}(p) = 0] = \left(1 - \frac{3\phi((p-1)/4)}{(p-1)/2}\right).$$

$$(2) \Pr[F_{(p-1)/4}(p) = 1] = \left(\frac{3\phi((p-1)/4)}{(p-1)/2}\right).$$

If $p \equiv 5 \pmod{8}$, then we simply have:

Prediction 5. Assume $p \equiv 5 \pmod{8}$ and $p \equiv 1 \pmod{3}$; i.e., $p \equiv 13 \pmod{24}$:

$$(1) \Pr[F_{(p-1)/4}(p) = 0] = \left(1 - \frac{2\phi((p-1)/4)}{p-1}\right).$$

$$(2) \Pr[F_{(p-1)/4}(p) = 1] = \left(\frac{2\phi((p-1)/4)}{p-1}\right).$$

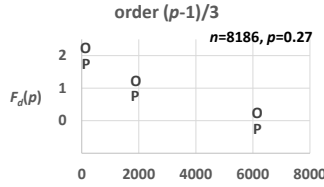


Figure 7. Predicted (P) and observed (O) numbers of primes for fixed points of order $(p - 1)/3$ in six-digit primes.

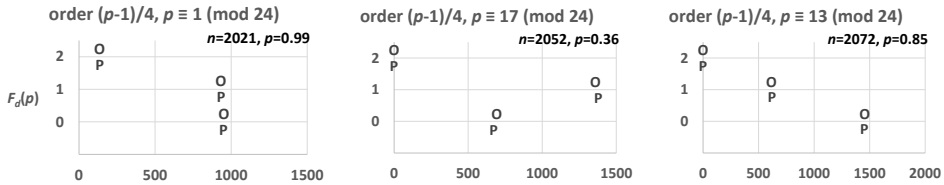


Figure 8. Predicted (P) and observed (O) numbers of primes for fixed points of order $(p - 1)/4$ in six-digit primes.

Assume $p \equiv 5 \pmod{8}$ and $p \equiv 2 \pmod{3}$; i.e., $p \equiv 5 \pmod{24}$:

$$(1) \Pr[F_{(p-1)/4}(p) = 0] = 1.$$

Chi-squared tests on the observed data from six-digit primes against the distributions predicted for orders $(p - 1)/3$ and $(p - 1)/4$ do not show significant deviation, as shown in Figures 7 and 8.

4. Conclusion and future work

In practice, it would certainly be possible for a user of the variant ElGamal digital signature scheme to simply make sure p is a safe prime, or alternatively arrange for r to always be a primitive root. In this way one could avoid the issue of fixed points altogether. However, we feel that it is very likely that a better understanding of the self-power map will help us better understand the security of this and other similar schemes.

We have given some bounds on the number of fixed points of the self-power map and attempted to predict the distribution of the fixed points using a binomial model whose mean is related to these proven bounds. When the order of x is moderate, this binomial model is a good predictor according to the data we collected. When the order of x is small, in particular when it is 3, 4, or 6, the independence assumption of the binomial model is violated in a significant way. However, we were able to find another model which appears to successfully predict the distribution.

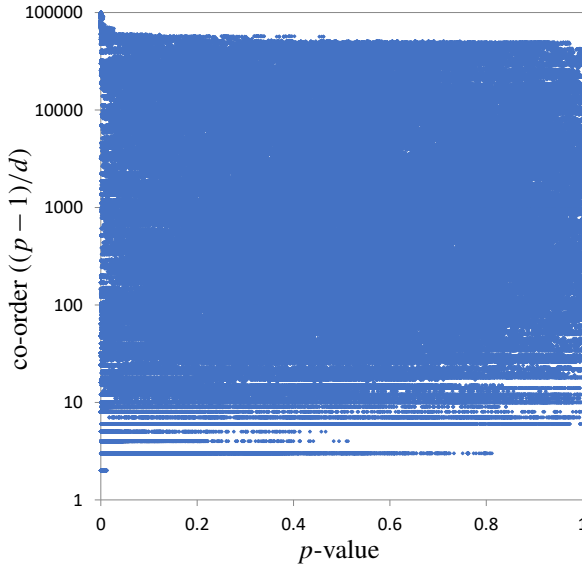


Figure 9. Logarithmic plot showing p -values of the sliding-window goodness-of-fit test, data sorted by co-order, for six-digit primes.

When the order of x is $(p-1)/3$ or $(p-1)/4$, we once again have a significant deviation from our first binomial model. However, a closer look at the set of possible fixed points in each case leads to another binomial model which appears to be successful. Some orders in the range $(p-1)/5$ to $(p-1)/9$ also appear to be showing significant deviation from the original model, as can be seen more clearly in Figure 9. In addition, the sliding-window chi-squared test shows evidence of likely divergence from the predictions in the neighborhood of $(p-1)/16$ and possibly other orders between $(p-1)/20$ and $(p-1)/50$. It is not clear yet whether all of these are true problems with the model, or just “random” consequences of the particular primes that we picked. Further investigation of these orders would appear to be the first item to be considered in future work.

Another very important item of future work would be to consider two-cycles, namely solutions to the equations

$$h^h \equiv a \pmod{p} \quad \text{and} \quad a^a \equiv h \pmod{p}, \quad (4)$$

or more generally k -cycles. Some data has been collected for these larger cycles but the binomial distribution has not yet been calculated or checked. The paper [Friedrichsen et al. 2010] also examined other graph-theoretic statistics of the functional graphs created by the self-power map, especially the number of components. This was also found to obey a nonnormal distribution and one could explore how that distribution is related to the one found here for fixed points.

Acknowledgements

Many thanks to Richard Layton for the design of Figures 6, 7, and 8. We also thank the Rose-Hulman Computer Science & Software Engineering and Mathematics Departments and Tufts University Technology Services for the use of their computers.

References

- [Anghel 2013] C. V. Anghel, *The self-power map and its image modulo a prime*, Ph.D. thesis, University of Toronto, 2013, <https://search.proquest.com/docview/1501462750>. MR
- [Anghel 2016] C. V. Anghel, “The self-power map and collecting all residue classes”, *Math. Comp.* **85**:297 (2016), 379–399. MR Zbl
- [Balog et al. 2011] A. Balog, K. A. Broughan, and I. E. Shparlinski, “On the number of solutions of exponential congruences”, *Acta Arith.* **148**:1 (2011), 93–103. MR Zbl
- [Cilleruelo and Garaev 2016a] J. Cilleruelo and M. Z. Garaev, “The congruence $x^x \equiv \lambda \pmod{p}$ ”, *Proc. Amer. Math. Soc.* **144**:6 (2016), 2411–2418. MR Zbl
- [Cilleruelo and Garaev 2016b] J. Cilleruelo and M. Z. Garaev, “Congruences involving product of intervals and sets with small multiplicative doubling modulo a prime and applications”, *Math. Proc. Cambridge Philos. Soc.* **160**:3 (2016), 477–494. MR Zbl
- [Cloutier and Holden 2010] D. Cloutier and J. Holden, “Mapping the discrete logarithm”, *Involve* **3**:2 (2010), 197–213. MR Zbl
- [Cobeli and Zaharescu 1999] C. Cobeli and A. Zaharescu, “An exponential congruence with solutions in primitive roots”, *Rev. Roumaine Math. Pures Appl.* **44**:1 (1999), 15–22. MR Zbl
- [Crocker 1966] R. Crocker, “On a new problem in number theory”, *Amer. Math. Monthly* **73** (1966), 355–357. MR Zbl
- [Crocker 1969] R. Crocker, “On residues of n^n ”, *Amer. Math. Monthly* **76** (1969), 1028–1029. MR Zbl
- [Felix and Kurlberg 2017] A. T. Felix and P. Kurlberg, “On the fixed points of the map $x \mapsto x^x$ modulo a prime, II”, *Finite Fields Appl.* **48** (2017), 141–159. MR Zbl
- [Friedrichsen et al. 2010] M. Friedrichsen, B. Larson, and E. McDowell, “Structure and statistics of the self-power map”, *Rose-Hulman Undergrad. Math. J.* **11**:2 (2010), art. id. 6.
- [Hoffman 2009] A. Hoffman, “Statistical investigation of structure in the discrete logarithm”, *Rose-Hulman Undergrad. Math. J.* **10**:2 (2009), art. id. 7. Zbl
- [Holden 2002a] J. Holden, “Addenda/corrigenda: fixed points and two-cycles of the discrete logarithm”, preprint, 2002. arXiv
- [Holden 2002b] J. Holden, “Fixed points and two-cycles of the discrete logarithm”, pp. 405–415 in *Algorithmic number theory* (Sydney, 2002), edited by C. Fieker and D. R. Kohel, Lecture Notes in Comput. Sci. **2369**, Springer, 2002. MR Zbl
- [Holden and Lindle 2008] J. Holden and N. Lindle, “A statistical look at maps of the discrete logarithm (abstract only)”, *ACM Commun. Comput. Algebra* **42**:1-2 (2008), 57–59.
- [Holden and Moree 2006] J. Holden and P. Moree, “Some heuristics and results for small cycles of the discrete logarithm”, *Math. Comp.* **75**:253 (2006), 419–449. MR Zbl
- [Holden and Robinson 2012] J. Holden and M. M. Robinson, “Counting fixed points, two-cycles, and collisions of the discrete exponential function using p -adic methods”, *J. Aust. Math. Soc.* **92**:2 (2012), 163–178. MR Zbl

- [Holden et al. 2016] J. Holden, P. A. Richardson, and M. M. Robinson, “Counting fixed points and two-cycles of the singular map $x \mapsto x^{x^n}$ modulo powers of a prime”, preprint, 2016. arXiv
- [Kurlberg et al. 2015] P. Kurlberg, F. Luca, and I. E. Shparlinski, “On the fixed points of the map $x \mapsto x^x$ modulo a prime”, *Math. Res. Lett.* **22**:1 (2015), 141–168. MR Zbl
- [Lindle 2008] N. W. Lindle, *A statistical look at maps of the discrete logarithm*, senior thesis, Rose-Hulman Institute of Technology, 2008, https://scholar.rose-hulman.edu/math_mstr/35.
- [Menezes et al. 1997] A. J. Menezes, P. C. van Oorschot, and S. A. Vanstone, *Handbook of applied cryptography*, CRC Press, Boca Raton, FL, 1997. MR Zbl
- [Somer 1981] L. Somer, “The residues of n^n modulo p ”, *Fibonacci Quart.* **19**:2 (1981), 110–117. MR Zbl

Received: 2017-04-22 Revised: 2018-01-31 Accepted: 2018-02-14

matthew.friedrichsen@tufts.edu *Department of Mathematics, Tufts University, Medford, MA, United States*

holden@rose-hulman.edu *Department of Mathematics, Rose-Hulman Institute of Technology, Terre Haute, IN, United States*

Analytical solution of a one-dimensional thermistor problem with Robin boundary condition

Volodymyr Hrynkyv and Alice Turchaninova

(Communicated by Suzanne Lenhart)

A one-dimensional nonlinear heat conduction equation of steady-state Joule heating in the presence of an electric field in a metal with temperature-dependent conductivities is considered. A technique developed by Young (1986) is adapted and used to derive an analytical solution for the problem with a Robin boundary condition.

1. Introduction

A thermal resistor, or thermistor, is a type of resistor with a highly temperature-dependent electrical conductivity. Thermistors are used as temperature-control elements in a range of equipment, such as spacecraft and air conditioning units, and have applications in the medical field, meteorology, and the chemical industry [Ng 1995; Macklen 1979]. The thermistor problem has been a source of significant mathematical interest and research but, due to the nonlinear nature of the problem, this research has been largely concerned with numerical solutions or existence proofs for a solution [Antontsev and Chipot 1994; Fowler et al. 1992; Howison et al. 1993; Shi et al. 1993; Sidi Ammi and Torres 2008; Xu 2004a; 2004b; Zhou and Westbrook 1997], rather than analytical solutions. In this paper, the thermistor problem is modeled as a nonlinear heat conduction equation of steady-state Joule heating in the presence of an electric field in a metal with temperature-dependent electrical and thermal conductivities. This paper extends the solution found in [Young 1986] to a more general case by introducing a Robin boundary condition on the temperature at an endpoint of the thermistor. This establishes the existing solution in [Young 1986] as a special case.

MSC2010: 35J60.

Keywords: thermistor problem, Robin boundary condition, Joule heating, analytical solution.

2. Formulation of the problem

Assuming that the electrical conductivity $\sigma(T)$ and the thermal conductivity $\kappa(T)$ are smooth functions, the heat conduction in an electrical conductor in the presence of Joule heating due to current can be shown [Young 1986] to satisfy the following two equations, one for the potential Φ and one for the temperature T :

$$\nabla^2 \Phi = -\frac{1}{\sigma} \nabla \sigma \cdot \nabla \Phi \quad \text{in } \Omega, \quad (1)$$

$$\frac{d^2 T}{d\Phi^2} + \left(\frac{1}{\kappa} \frac{d\kappa}{dT} - \frac{1}{\sigma} \frac{d\sigma}{dT} \right) \left(\frac{dT}{d\Phi} \right)^2 = -\frac{\sigma}{\kappa} \quad \text{in } \Omega, \quad (2)$$

with some appropriate boundary conditions. Throughout this section and Section 3 we assume that the given domain Ω lies in \mathbb{R}^n . Equations (1) and (2) respectively describe conservation of charge and the steady diffusion of heat in the presence of Joule heating due to electric current (see [Young 1986] for more details).

3. Derivation of the solution in the general case

For the sake of convenience, we recreate the technique developed in [Young 1986] for obtaining a solution to (1) and (2). Equation (2) can be simplified by introducing a new variable

$$\frac{\sigma(T)}{\kappa(T)} = e^{-\xi(T)}. \quad (3)$$

Differentiating both sides of (3), and after some manipulations, (2) can be written as

$$\frac{d^2 T}{d\Phi^2} + \frac{d\xi}{dT} \left(\frac{dT}{d\Phi} \right)^2 = -e^{-\xi}. \quad (4)$$

Next, setting

$$\theta = \frac{dT}{d\Phi}, \quad (5)$$

(4) becomes

$$\theta \frac{d\theta}{dT} + \frac{d\xi}{dT} \theta^2 = -e^{-\xi}. \quad (6)$$

Observing

$$\frac{1}{2} e^{-2\xi} \frac{d}{dT} (e^{2\xi} \theta^2) = \theta \frac{d\theta}{dT} + \frac{d\xi}{dT} \theta^2$$

allows us to rewrite (6) as

$$\frac{1}{2} e^{-2\xi} \frac{d}{dT} (e^{2\xi} \theta^2) = -e^{-\xi}. \quad (7)$$

Integrating (7), we get

$$e^{2\xi(T)} \theta^2 = C - 2 \int e^{\xi(T)} dT, \quad (8)$$

where C is a constant of integration. Solving for θ^2 in (8) and taking into account (5), we have

$$\left(\frac{dT}{d\Phi}\right)^2 = \frac{C - 2 \int e^{\xi(T)} dT}{e^{2\xi(T)}}.$$

Finally, the following equation for Φ is obtained:

$$\Phi = \int \frac{(\kappa(T)/\sigma(T)) dT}{\sqrt{C - 2 \int (\kappa(s)/\sigma(s)) ds}} + C', \quad (9)$$

where the integration constant of the integral under the square root is absorbed into C . It turns out that for many metals, the ratio of conductivities is proportional to the absolute temperature of the metal. This relationship is known as the Wiedemann–Franz–Lorenz (WFL) law [Berman 1976; Meaden 1965],

$$\frac{\kappa(T)}{\sigma(T)} = \alpha T, \quad (10)$$

where α is the Lorenz number for a given metal and may have slightly different values for different metals. Once the ratio of conductivities is specified using the WFL law, (9) can be integrated to obtain the temperature in terms of the potential, $T(\Phi)$,

$$T(\Phi) = \frac{1}{\sqrt{\alpha}} [C - (\Phi - C')^2]^{1/2}. \quad (11)$$

For (11) to be of any help, we must determine Φ that solves (1). This issue can be dealt with by introducing an auxiliary potential Ψ that satisfies Laplace's equation [Young 1986; Flynn 1969]. Namely, define this auxiliary potential Ψ as

$$\sigma_0 \nabla \Psi \equiv \sigma[T(\Phi)] \nabla \Phi,$$

where σ_0 is the electrical conductivity at some conveniently chosen reference temperature. Then clearly Ψ satisfies Laplace's equation $\nabla^2 \Psi = 0$, and it is easily seen, by isolating $\nabla \Psi$ and integrating, that

$$\Psi = \frac{1}{\sigma_0} \int \sigma[T(\Phi)] d\Phi. \quad (12)$$

Knowing how σ depends on T and using (11) will enable us to perform integration in (12). This will give us Ψ in terms of Φ , and therefore finding the inverse of this function will result in an expression for Φ in terms of a function that satisfies Laplace's equation. In the next section, we will also solve for the constants C and C' in the expression. This ends the derivation of a solution to (1) and (2) when conductivities obey the WFL law.

4. Solution in one dimension with Robin boundary condition

In this section we adapt the technique described in Section 3 for a one-dimensional problem with a Robin boundary condition at one endpoint. Namely, consider a thin rod of length L , where the potential and temperature satisfy (1) and (2), respectively. In addition, the endpoints at $z = 0$ and $z = L$ are held at potentials V_0 and 0, respectively. The boundary condition for T at the right endpoint of the rod $z = L$ is given by a Robin boundary condition, whereas the left endpoint is held at the constant temperature $T = T_0$. Note that it is reasonable to consider a Robin boundary condition for at least one end of the rod, as it models the cooling effect of that end of the thermistor through Newton's law of cooling [Howison 2005]. The boundary conditions are summarized below:

$$T = T_0, \quad \Phi = V_0 \quad \text{at } z = 0, \quad (13)$$

$$\frac{dT}{dz} + \beta(T - T_0) = 0, \quad \Phi = 0 \quad \text{at } z = L. \quad (14)$$

Observe that when β approaches infinity, the boundary condition for T at $z = L$ reduces to $T = T_0$ at $z = L$, which corresponds to that in [Young 1986]. Recall that

$$T(\Phi) = \frac{1}{\sqrt{\alpha}}[C - (\Phi - C')^2]^{1/2}.$$

Now we use the boundary conditions (13) and (14) to determine the constants C and C' . From (13), it is immediate that

$$C - C'^2 = \alpha T_0^2 + V_0^2 - 2V_0 C'. \quad (15)$$

First, we find

$$\begin{aligned} \frac{dT}{dz} + \beta(T - T_0) &= \frac{C' - \Phi}{\sqrt{\alpha}} \frac{1}{[C - (\Phi - C')^2]^{1/2}} \frac{d\Phi}{dz} + \beta \left(\frac{1}{\sqrt{\alpha}}[C - (\Phi - C')^2]^{1/2} - T_0 \right) \end{aligned} \quad (16)$$

and using (14) to evaluate (16) at $z = L$ gives us

$$\frac{C'}{\sqrt{\alpha}} \frac{\Phi_0}{[C - C'^2]^{1/2}} + \beta \left(\frac{1}{\sqrt{\alpha}}[C - C'^2]^{1/2} - T_0 \right) = 0, \quad (17)$$

where we defined $\Phi_0 := \Phi'(L)$. Note that a new parameter Φ_0 has been introduced into the problem. We will address this issue later. Rewriting (17) as

$$\frac{C'}{\beta} \frac{\Phi_0}{[C - C'^2]^{1/2}} + [C - C'^2]^{1/2} = \sqrt{\alpha} T_0 \quad (18)$$

and squaring both sides of (18), we get

$$\left(\frac{C'\Phi_0}{\beta}\right)^2 + \frac{2\Phi_0 C'}{\beta}(C - C'^2) + (C - C'^2)^2 = \alpha T_0^2(C - C'^2).$$

Now using (15) and grouping the result by C'^2 , we obtain the following quadratic equation for C' :

$$C'^2 \left[\left(\frac{\Phi_0}{\beta}\right)^2 - \frac{4V_0\Phi_0}{\beta} + 4V_0^2 \right] + C' \left[\frac{2\Phi_0}{\beta}(\alpha T_0^2 + V_0^2) - 4V_0^3 - 2\alpha V_0 T_0^2 \right] + \alpha T_0^2 V_0^2 + V_0^4 = 0.$$

Defining

$$\begin{aligned} A &= \left(\frac{\Phi_0}{\beta}\right)^2 - \frac{4V_0\Phi_0}{\beta} + 4V_0^2, \\ B &= \frac{2\Phi_0}{\beta}(\alpha T_0^2 + V_0^2) - 4V_0^3 - 2\alpha V_0 T_0^2, \\ D &= \alpha T_0^2 V_0^2 + V_0^4, \end{aligned}$$

the solution for C' is given by

$$C' = \frac{-B \pm \sqrt{B^2 - 4AD}}{2A}, \quad (19)$$

where

$$\begin{aligned} B^2 &= \frac{4\Phi_0(\alpha T_0^2 + V_0^2)^2 - 4\Phi_0\beta(\alpha T_0^2 + V_0^2)(4V_0^3 + 2\alpha V_0 T_0^3) + \beta^2(4V_0^3 + 2\alpha V_0 T_0^2)^2}{\beta^2}, \\ 4AD &= \frac{4\Phi_0^2(\alpha T_0^2 + V_0^2)V_0^2 - 16\Phi_0\beta V_0(\alpha T_0^2 + V_0^2)V_0^2 + \beta^2 16V_0^2(\alpha T_0^2 V_0 + V_0^3)V_0}{\beta^2}, \end{aligned}$$

so that

$$\begin{aligned} B^2 - 4AD &= \frac{4\Phi_0^2\alpha T_0^2(\alpha T_0^2 + V_0^2) - 8\Phi_0\beta\alpha V_0 T_0^2(\alpha T_0^2 + V_0^2) + 4\beta^2\alpha^2 V_0^2 T_0^4}{\beta^2} \\ &= \frac{4\Phi_0^2\alpha T_0^2(\alpha T_0^2 + V_0^2)}{\beta^2} - \frac{8\Phi_0\alpha V_0 T_0^2(\alpha T_0^2 + V_0^2)}{\beta} + 4\alpha^2 V_0^2 T_0^4. \end{aligned}$$

Now it can be verified that this solution to the quadratic equation will match that in [Young 1986] when $\beta \rightarrow \infty$. In this case, the first two terms above disappear and we have

$$\lim_{\beta \rightarrow \infty} A = 4V_0^2, \quad \lim_{\beta \rightarrow \infty} B = -4V_0^3 - 2\alpha V_0 T_0^2, \quad \lim_{\beta \rightarrow \infty} [B^2 - 4AD] = 4\alpha^2 V_0^2 T_0^4.$$

We are thus left with

$$\lim_{\beta \rightarrow \infty} C' = \lim_{\beta \rightarrow \infty} \left[\frac{-B \pm \sqrt{B^2 - 4AD}}{2A} \right] = \frac{4V_0^3 + 2\alpha V_0 T_0^2 \pm \sqrt{4\alpha^2 V_0^2 T_0^4}}{8V_0^2}.$$

Taking the negative square root, we get $C' = \frac{1}{2} V_0$, as in [Young 1986]. The constants C and C' in terms of β are as follows:

$$\begin{aligned} C' &= \frac{-B - 2T_0 \sqrt{\alpha} \sqrt{(\Phi_0 - \beta V_0)^2 (\alpha T_0^2 + V_0^2) - \beta^2 V_0^4}}{2A}, \\ C &= \alpha T_0^2 + V_0^2 - 2V_0 C' + C'^2, \end{aligned} \quad (20)$$

where A and B are defined above. We take the negative root in the quadratic formula for C' because this is the root that reduces to Young's solution when $\beta \rightarrow \infty$. Now we use the auxiliary potential Ψ , given in (12),

$$\Psi = \frac{1}{\sigma_0} \int \sigma[T(\Phi)] d\Phi.$$

First, we note that it is an experimentally verified fact that the thermal conductivity κ varies very little with temperature for many metals; see [Young 1986]. Therefore, it is physically reasonable to assume that $\kappa(T) = \kappa_0$, where κ_0 is a constant. Now, taking $\sigma[T(\Phi)]$ to obey the WFL law (10), we have

$$\sigma[T(\Phi)] = \frac{\kappa[T(\Phi)]}{\alpha T(\Phi)} = \frac{\kappa_0}{\alpha T(\Phi)}. \quad (21)$$

Substituting (21) into (12), we get

$$\Psi = \frac{\kappa_0}{\alpha \sigma_0} \int \frac{d\Phi}{T(\Phi)} = \frac{\kappa_0}{\alpha \sigma_0} \int \frac{d\Phi}{(1/\sqrt{\alpha})[C - (\Phi - C')^2]^{1/2}} = \frac{\kappa_0}{\alpha \sigma_0} \sin^{-1} \left(\frac{\Phi - C'}{\sqrt{C}} \right).$$

Since $\nabla^2 \Psi = 0$, it follows that Ψ is a linear function of z . We set $\Psi(z) = a + bz$ and absorb the constant $\kappa_0/(\alpha \sigma_0)$, as well as the integration constant of $\Psi(z)$, into the coefficients of the linear function. Hence,

$$\sin^{-1} \left(\frac{\Phi - C'}{\sqrt{C}} \right) = a + bz.$$

Using the boundary conditions for Φ to determine a and b ,

$$\begin{aligned} \Phi(z=0) &= \sqrt{C} \sin(a) + C' = V_0, \\ \Phi(z=L) &= \sqrt{C} \sin(a + bL) + C' = 0, \end{aligned}$$

we obtain the expression for the general solution $\Phi(z)$,

$$\Phi(z) = \sqrt{C} \sin(a + bz) + C', \quad (22)$$

fluid 1	transmission surface	fluid 2	β	Φ_0
air	cast iron	air	5.7	-0.03864
air	mild steel	air	7.9	-0.03834
steam	cast iron	air	11.3	-0.03809
steam	mild steel	air	14.2	-0.03796
steam	copper	air	17	-0.03788

Table 1. Φ_0 for realistic values of β .

where

$$a = \sin^{-1}\left(\frac{V_0 - C'}{\sqrt{C}}\right), \quad b = \frac{1}{L}\left[\sin^{-1}\left(\frac{-C'}{\sqrt{C}}\right) - a\right],$$

with C and C' given by (20), and where Φ_0 is determined numerically from the equation

$$\Phi_0 = \Phi'(L) = b\sqrt{C} \cos(a + bL). \quad (23)$$

Equation (23) was obtained by differentiating (22) with respect to z and then evaluating the derivative at $z = L$. Note that since the right-hand side of (23) also contains Φ_0 , we view (23) as an equation where the unknown is Φ_0 . Even though (23) cannot be solved analytically for Φ_0 , as it enters the right-hand side of (23) in a complicated way, we can still solve (23) numerically by choosing physically realistic values for the parameters of the problem. Table 1 gives values of Φ_0 for realistic values of β , provided in [Engineering ToolBox 2003] for transmission surfaces between various combinations of fluids. The units of β are $\text{W}/(\text{m}^2\text{K})$ and the units of Φ_0 are V/m .

To complete the general solution, we substitute (22) back into (11) to obtain $T(z)$:

$$T(z) = \frac{1}{\sqrt{\alpha}} \sqrt{C} \cos(a + bz). \quad (24)$$

Finally, as is expected, $\Phi(z)$ in (22) tends to the one found in [Young 1986] as $\beta \rightarrow \infty$. Indeed, we have

$$\begin{aligned} \lim_{\beta \rightarrow \infty} C' &= \frac{1}{2} V_0, \\ \lim_{\beta \rightarrow \infty} C &= \alpha T_0^2 + \frac{1}{4} V_0^2, \\ \lim_{\beta \rightarrow \infty} a &= \sin^{-1}\left(\frac{\frac{1}{2} V_0}{\sqrt{\alpha T_0^2 + \frac{1}{4} V_0^2}}\right) = \Omega, \\ \lim_{\beta \rightarrow \infty} b &= \frac{1}{L} \left[\sin^{-1}\left(\frac{-\frac{1}{2} V_0}{\sqrt{\alpha T_0^2 + \frac{1}{4} V_0^2}}\right) - \Omega \right] = \frac{1}{L} [-\Omega - \Omega] = -\frac{2\Omega}{L}, \end{aligned}$$

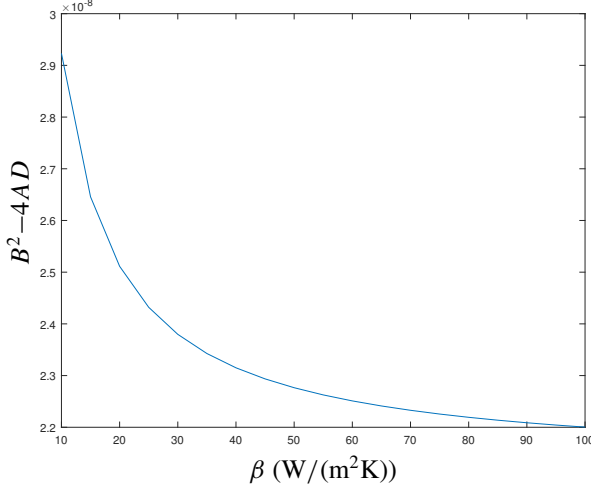


Figure 1. $B^2 - 4AD$ as a function of β for $T_0 = 273$ K, $\alpha = 2.445 \cdot 10^{-8}$ (V/K) 2 , $V_0 = 40$ mV.

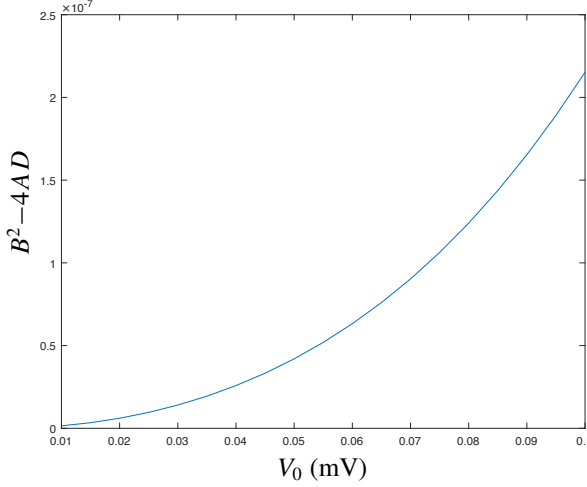


Figure 2. $B^2 - 4AD$ as a function of V_0 for $T_0 = 273$ K, $\alpha = 2.445 \cdot 10^{-8}$ (V/K) 2 , $\beta = 17$ W/(m 2 K).

so that

$$\lim_{\beta \rightarrow \infty} \Phi(z) = \frac{1}{2} V_0 + \sqrt{\alpha T_0^2 + \frac{1}{4} V_0^2} \sin \left[\Omega \left(1 - 2 \frac{z}{L} \right) \right],$$

which coincides with the expression derived in [Young 1986]. Similarly, it can be shown that (24) tends to the expression for $T(z)$ found in [Young 1986] as $\beta \rightarrow \infty$.

Figures 1 and 2 show the graphs of $B^2 - 4AD$ as a function of β and V_0 , respectively.

These figures show that the expression $B^2 - 4AD$ under the square root in (19) is always greater than zero for a physically meaningful range of parameters $\beta > 0$ and $V_0 > 0$. This, in turn, guarantees that there is no “nonexistence” of solution to the given problem.

We can also derive a solution for the case $\beta = 0$ and verify that the general solution (22) reduces to this solution as $\beta \rightarrow 0$. Indeed, when $\beta = 0$, the boundary conditions (13) and (14) are reduced to the boundary conditions

$$\begin{aligned} T &= T_0, \quad \Phi = V_0 \quad \text{at } z = 0, \\ \frac{dT}{dz} &= 0, \quad \Phi = 0 \quad \text{at } z = L. \end{aligned}$$

With the same steps as before, the following expressions for $\Phi(z)$ and $T(z)$ can be derived:

$$\Phi(z) = \sqrt{\alpha T_0^2 + V_0^2} \sin\left[\tilde{\Omega}\left(1 - \frac{z}{L}\right)\right], \quad (25)$$

$$T(z) = \frac{1}{\sqrt{\alpha}} \sqrt{\alpha T_0^2 + V_0^2} \cos\left[\tilde{\Omega}\left(1 - \frac{z}{L}\right)\right], \quad (26)$$

where

$$\tilde{\Omega} := \sin^{-1}\left(\frac{V_0}{\sqrt{\alpha T_0^2 + V_0^2}}\right).$$

It can be easily verified that (22) and (24) approach (25) and (26), respectively, as $\beta \rightarrow 0$.

References

- [Antontsev and Chipot 1994] S. N. Antontsev and M. Chipot, “The thermistor problem: existence, smoothness, uniqueness, blowup”, *SIAM J. Math. Anal.* **25**:4 (1994), 1128–1156. MR Zbl
- [Berman 1976] R. Berman, *Thermal conduction in solids*, Oxford Univ. Press, 1976.
- [Engineering ToolBox 2003] Engineering ToolBox, “Overall heat transfer coefficients for fluids-heat exchanger surface combinations”, 2003, available at <https://tinyurl.com/etoolboxov>.
- [Flynn 1969] D. R. Flynn, “Measurement of thermal conductivity by steady state methods in which the sample is heated directly by passage of an electric current”, pp. 241–300 in *Thermal conductivity*, edited by R. P. Tye, Academic Press, London, 1969.
- [Fowler et al. 1992] A. C. Fowler, I. Frigaard, and S. D. Howison, “Temperature surges in current-limiting circuit devices”, *SIAM J. Appl. Math.* **52**:4 (1992), 998–1011. MR Zbl
- [Howison 2005] S. Howison, *Practical applied mathematics: modelling, analysis, approximation*, Cambridge Univ. Press, 2005. MR Zbl
- [Howison et al. 1993] S. D. Howison, J. F. Rodrigues, and M. Shillor, “Stationary solutions to the thermistor problem”, *J. Math. Anal. Appl.* **174**:2 (1993), 573–588. MR Zbl
- [Macklen 1979] E. D. Macklen, *Thermistors*, Electrochemical Pubs., Ayr, Scotland, 1979.
- [Meaden 1965] G. T. Meaden, *Electrical resistance of metals*, Int. Cryogenics Monograph Series **2**, Springer, 1965.

- [Ng 1995] K. K. Ng, *Complete guide to semiconductor devices*, McGraw-Hill, New York, 1995.
- [Shi et al. 1993] P. Shi, M. Shillor, and X. Xu, “Existence of a solution to the Stefan problem with Joule’s heating”, *J. Differential Equations* **105**:2 (1993), 239–263. MR Zbl
- [Sidi Ammi and Torres 2008] M. R. Sidi Ammi and D. F. M. Torres, “Numerical analysis of a nonlocal parabolic problem resulting from thermistor problem”, *Math. Comput. Simulation* **77**:2-3 (2008), 291–300. MR Zbl
- [Xu 2004a] X. Xu, “Exponential integrability of temperature in the thermistor problem”, *Differential Integral Equations* **17**:5-6 (2004), 571–582. MR Zbl
- [Xu 2004b] X. Xu, “Local regularity theorems for the stationary thermistor problem”, *Proc. Roy. Soc. Edinburgh Sect. A* **134**:4 (2004), 773–782. MR Zbl
- [Young 1986] J. H. Young, “Steady state Joule heating with temperature dependent conductivities”, *Appl. Sci. Res.* **43**:1 (1986), 55–65. Zbl
- [Zhou and Westbrook 1997] S. Zhou and D. R. Westbrook, “Numerical solutions of the thermistor equations”, *J. Comput. Appl. Math.* **79**:1 (1997), 101–118. MR Zbl

Received: 2017-04-26 Revised: 2017-10-05 Accepted: 2018-02-14

hrynkivv@uhd.edu

*Department of Mathematics and Statistics, University
of Houston - Downtown, Houston, TX, United States*

turchaninova2@gator.uhd.edu

*University of Houston - Downtown, Houston, TX,
United States*

On the covering number of S_{14}

Ryan Oppenheim and Eric Swartz

(Communicated by Kenneth S. Berenhaut)

If all elements of a group G are contained in the set-theoretic union of proper subgroups H_1, \dots, H_n , then we define this collection to be a cover of G . When such a cover exists, the cardinality of the smallest possible cover is called the covering number of G , denoted by $\sigma(G)$. Maróti determined $\sigma(S_n)$ for odd $n \neq 9$ and provided an estimate for even n . The second author later determined $\sigma(S_n)$ for $n \equiv 0 \pmod{6}$ when $n \geq 18$, while joint work of the second author with Kappe and Nikolova-Popova also verified that Maróti's rule holds for $n = 9$ and established the covering numbers of S_n for various other small n . Currently, $n = 14$ is the smallest value for which $\sigma(S_n)$ is unknown. In this paper, we prove the covering number of S_{14} is 3096.

1. Introduction

For a group G , a set \mathcal{H} of proper subgroups of G is a *cover* of G if and only if $\bigcup_{A \in \mathcal{H}} A = G$. Further, supposing a cover for G exists, define the *covering number* of G , denoted by $\sigma(G)$, to be the cardinality of the smallest possible cover of G ; that is, $\sigma(G)$ is the size of a minimal cover of G .

Based on the work of Neumann [1954], who showed that a group has a finite cover if and only if it has a finite noncyclic homomorphic image, it suffices to consider covers of finite groups. Covers have enjoyed some degree of attention in recent years, particularly given the property that $\sigma(G)$ serves as an upper bound for $\omega(G)$, defined as the largest integer m such that some subset S of G exists where $|S| = m$ and any two distinct elements of S generate G . This and other related problems have garnered much of the current interest in covering numbers; see [Blackburn 2006; Holmes and Maróti 2010], and, for a general survey of such problems, [Serena 2003].

Tomkinson [1997] determined the covering number for a given solvable group and suggested that it would be of interest to investigate minimal covers of nonsolvable groups. The symmetric and alternating groups have naturally attracted special attention, and there has been significant work to derive formulae for the covering numbers of A_n and S_n . Regarding alternating groups, Maróti [2005] established

MSC2010: 20-04, 20D60.

Keywords: symmetric groups, finite union of proper subgroups, subgroup covering.

that $\sigma(A_n) \geq 2^{n-2}$, where $n \neq 7, 9$ (and $\sigma(A_n) = 2^{n-2}$ if and only if $n \equiv 2 \pmod{4}$). Turning our attention to the symmetric groups, Maróti also showed in the same paper that $\sigma(S_n) = 2^{n-1}$ for odd $n \neq 9$. Later, Kappe, Nikolova-Popova, and the second author [Kappe et al. 2016] showed that this rule holds when $n = 9$ as well, and ascertained the covering numbers of S_8 , S_{10} , and S_{12} . The second author also demonstrated in [Swartz 2016] that $\sigma(S_{18}) = 36773$, and that

$$\sigma(S_n) = \frac{1}{2} \binom{n}{n/2} + \sum_{i=0}^{n/3-1} \binom{n}{i}$$

when $n \equiv 0 \pmod{6}$ and $n \geq 24$; given that $\sigma(S_6)$ and $\sigma(S_{12})$ were already known, this accounts for all multiples of 6. In pursuit of formulae for all yet-unknown $\sigma(S_n)$, this paper is intended to begin the process of finding the general covering number when $n \equiv 2 \pmod{6}$. In determining $\sigma(S_{14})$, or indeed any group whose covering number is unknown, we must establish both the existence of a certain cover of S_{14} and show that no smaller set of proper subgroups could contain among them every element of S_{14} . When considering those groups for which a cover exists (i.e., noncyclic groups), it trivially suffices to consider only maximal subgroups.

The following notation will be used throughout this paper in the discussion of the elements of symmetric groups. We say that $g \in S_n$ has cycle structure (n_1, \dots, n_k) if g can be written as the product of disjoint cycles g_1, \dots, g_k , where the length of each cycle g_i is n_i and $n_1 \leq n_2 \leq \dots \leq n_k$. For example, the element $g = (1\ 2\ 3\ 4\ 5\ 6\ 7)(8\ 9\ 10\ 11\ 12\ 13) \in S_{14}$ has cycle structure $(1, 6, 7)$.

In Section 2, we demonstrate a cover of S_{14} containing 3096 subgroups and prove that $\sigma(S_{14}) = 3096$ by showing that this cover is in fact minimal. The GAP code used in the proof can be found in the online supplement.

2. Covering S_{14}

Let \mathcal{C}_{14} be the set of those maximal subgroups of S_{14} isomorphic to one of A_{14} , $S_7 \text{ wr } S_2$ (here wr denotes the wreath product), S_{13} , $S_3 \times S_{11}$, or $S_4 \times S_{10}$.

Lemma 2.1. *The set \mathcal{C}_{14} is a cover of S_{14} .*

Proof. Any 14-cycle is contained in some subgroup isomorphic to $S_7 \text{ wr } S_2$, and any element of S_{14} that fixes some element of $\{1, \dots, 14\}$ is contained in a subgroup isomorphic to S_{13} . Furthermore, any element without a fixed point that is the product of two cycles is covered by A_{14} , meaning that some element $g \in S_{14}$ could only fail to be covered if it consists of three or more cycles and fixes no points. If the length of one of these cycles is 3 or 4, then g is covered by $S_3 \times S_{11}$ or $S_4 \times S_{10}$, respectively; similarly, if there are two cycles of length 2, then g is covered by $S_4 \times S_{10}$. Furthermore, any element of cycle structure $(2, 6, 6)$ or

isomorphism type	class size
A_{14}	1
$S_7 \text{ wr } S_2$	1716
$S_2 \text{ wr } S_7$	135135
$S_1 \times S_{13}$	14
$S_2 \times S_{12}$	91
$S_3 \times S_{11}$	364
$S_4 \times S_{10}$	1001
$S_5 \times S_9$	2002
$S_6 \times S_8$	3003
$\text{PGL}_2(13)$	39916800

Table 1. Conjugacy classes of maximal subgroups of S_{14} .

$(2, 5, 7)$ stabilizes a decomposition of $\{1, \dots, 14\}$ into two subsets of cardinality 7 and thus is contained in a subgroup isomorphic to $S_7 \text{ wr } S_2$. Since any element of S_{14} which is the product of three or more disjoint cycles must contain a cycle of length 4 or smaller, and we have covered all such elements, we have shown that \mathcal{C}_{14} is indeed a cover. \square

We note that \mathcal{C}_{14} contains 3096 subgroups (see Table 1). We will show that \mathcal{C}_{14} is in fact a minimal cover.

Lemma 2.2. *Any minimal cover of S_{14} contains all subgroups isomorphic to one of A_{14} or S_{13} .*

Proof. We note that $\sigma(A_{14}) = \sigma(S_{13}) = 2^{12} > 3096$, where 3096 is our established upper bound for $\sigma(S_{14})$. Lemma 1 of [Garonzi 2013] states that a maximal subgroup H of a group G with $\sigma(H) > \sigma(G)$ is included in any minimal cover of G containing only maximal subgroups. Thus every minimal cover of the elements of S_{14} must contain every subgroup isomorphic to either A_{14} or S_{13} . \square

Lemma 2.2 shows that we can restrict ourselves to finding a minimal cover of the elements not contained in a subgroup isomorphic to either A_{14} or S_{13} . Let Π denote the set of all $g \in S_{14}$ with cycle structure (14) , $(3, 5, 6)$, or $(4, 5, 5)$. We will divide the elements of Π as follows: Π_0 will be the set of 14-cycles, Π_3 the set of cycles with structure $(3, 5, 6)$, and Π_4 the set of cycles with structure $(4, 5, 5)$. The distribution of these elements among maximal subgroups of S_{14} is shown in Table 2. In Table 2, if the entry in the row indexed by maximal subgroup M_i and column indexed by cycle structure (j) is “ n_m ”, then a subgroup isomorphic to M_i contains n elements with cycle structure (j) , and each element with cycle structure (j) is contained in m maximal subgroups isomorphic to M_i . If the entry in the row indexed by maximal subgroup M_i and the column indexed by cycle

isomorphism type	(14)	(3, 5, 6)	(4, 5, 5)
A_{14}	0	0	0
$S_7 \text{ wr } S_2$	3628800, P	0	0
$S_2 \text{ wr } S_7$	46080, P	0	0
$S_1 \times S_{13}$	0	0	0
$S_2 \times S_{12}$	0	0	0
$S_3 \times S_{11}$	0	2661120, P	0
$S_4 \times S_{10}$	0	0	435456, P
$S_5 \times S_9$	0	483840, P	435456 ₂
$S_6 \times S_8$	0	322560, P	0
$\text{PGL}_2(13)$	468 ₃	0	0

Table 2. Elements of a given cycle structure in S_{14} in each maximal subgroup of a given isomorphism type.

structure (j) is “ n, P ”, then a subgroup isomorphic to M_i contains n elements with cycle structure (j) , and the elements with cycle structure (j) are partitioned among the maximal subgroups isomorphic to M_i .

Let C'_{14} be the set of all subgroups isomorphic to one of $S_7 \text{ wr } S_2$, $S_3 \times S_{11}$, or $S_4 \times S_{10}$. By showing that the set C'_{14} is a minimal cover of the elements of Π , we will show that C_{14} is also a minimal cover of S_{14} .

Lemma 2.3. *Any minimal cover of Π contains all subgroups isomorphic to $S_7 \text{ wr } S_2$.*

Proof. Let \mathcal{B} be a minimal cover of S_{14} . Any cover of S_{14} must contain some mix of subgroups conjugate to $S_7 \text{ wr } S_2$, $S_2 \text{ wr } S_7$, or $\text{PGL}_2(13)$ to cover the elements of Π_0 . Examining Table 2, if M is a maximal subgroup of S_{14} and $M \cap \Pi_0 \neq \emptyset$, then $M \cap \Pi = M \cap \Pi_0$. Hence any minimal cover of the elements of Π must contain a minimal cover of the elements of Π_0 , which is precisely all subgroups isomorphic to $S_7 \text{ wr } S_2$. \square

Lemmas 2.2 and 2.3 show that it suffices to restrict our attention to subgroups isomorphic to one of $S_3 \times S_{11}$, $S_4 \times S_{10}$, $S_5 \times S_9$, or $S_6 \times S_8$ covering elements of $\Pi_3 \cup \Pi_4$ when determining a minimal cover of the permutations in Π . We define $H_1 := \text{Sym}(\{1, 2, 3\}) \times \text{Sym}(\{4, \dots, 14\})$ and will use this notation henceforth.

Lemma 2.4. *If a minimal cover \mathcal{B} of the elements of Π does not contain a subgroup isomorphic to $S_3 \times S_{11}$, then there are at least 11 subgroups isomorphic to $S_3 \times S_{11}$ not contained in \mathcal{B} .*

Proof. Let \mathcal{B} be a minimal cover of the elements of Π . Since we know that C_{14} is a cover of Π , we can compare \mathcal{B} to C_{14} . Define $\mathcal{B}' := \mathcal{B} \setminus C_{14}$ and $\mathcal{C}' := C_{14} \setminus \mathcal{B}$. This

implies

$$\begin{aligned}\mathcal{B} &= (\mathcal{B} \cap \mathcal{C}_{14}) \cup \mathcal{B}', \\ \mathcal{C}_{14} &= (\mathcal{B} \cap \mathcal{C}_{14}) \cup \mathcal{C}'.\end{aligned}$$

Since \mathcal{B} is a minimal cover of the elements of Π , we have $|\mathcal{B}'| \leq |\mathcal{C}'|$. By Lemmas 2.2 and 2.3, \mathcal{B}' consists only of subgroups isomorphic to either $S_5 \times S_9$ or $S_6 \times S_8$, and \mathcal{C}' consists only of subgroups isomorphic to either $S_3 \times S_{11}$ or $S_4 \times S_{10}$. Moreover, we will assume that \mathcal{C}' consists of c_3 subgroups isomorphic to $S_3 \times S_{11}$ and c_4 subgroups isomorphic to $S_4 \times S_{10}$. This means that

$$|\mathcal{B}'| \leq |\mathcal{C}'| = c_3 + c_4,$$

and we want to show that if $c_3 \geq 1$, then $c_3 \geq 11$.

Since we are assuming that \mathcal{B} does not contain a subgroup isomorphic to $S_3 \times S_{11}$, without loss of generality we may assume that $H_1 := \text{Sym}(\{1, 2, 3\}) \times \text{Sym}(\{4, \dots, 14\}) \notin \mathcal{B}$. This means that the subgroups in \mathcal{B}' must cover every element with cycle structure $(3, 5, 6)$ in H_1 . Let $\{4, \dots, 14\} = A \cup A^c$, where $|A| = 5$. If \mathcal{B} is a cover of Π , then, for each such set A , either $\text{Sym}(A) \times \text{Sym}(A^c \cup \{1, 2, 3\})$ or $\text{Sym}(A^c) \times \text{Sym}(A \cup \{1, 2, 3\})$ is contained in \mathcal{B}' . Hence at least $\binom{11}{5} = 462$ subgroups are contained in \mathcal{B}' . Let $\mathcal{B}' = \mathcal{D}_1 \cup \mathcal{D}_2$, where \mathcal{D}_1 consists of the 462 subgroups needed to cover $\Pi_3 \cap H_1$.

We will now bound from above c_4 , the number of groups isomorphic to $S_4 \times S_{10}$ that are in \mathcal{C}_{14} but not in \mathcal{B} . From Table 2, we see that, if M_i is a maximal subgroup isomorphic to $S_i \times S_{14-i}$, then $\Pi_4 \cap M_6 = \emptyset$ and

$$|\Pi_4 \cap M_4| = |\Pi_4 \cap M_5| = 435456.$$

Furthermore, the elements of Π_4 are partitioned among the maximal subgroups isomorphic to $S_4 \times S_{10}$. This means that, if there are n_4 total elements with cycle structure $(4, 5, 5)$ contained in the subgroups of \mathcal{B}' , then \mathcal{B}' can cover the elements from at most $n_4/435456$ subgroups isomorphic to $S_4 \times S_{10}$; in other words,

$$c_4 \leq \frac{n_4}{435456}.$$

To bound n_4 from above, we first observe that \mathcal{D}_2 contains at most $435456 \cdot |\mathcal{D}_2|$ distinct elements with cycle structure $(4, 5, 5)$ (in the case when every subgroup of \mathcal{D}_2 is isomorphic to $S_5 \times S_9$). Consider now \mathcal{D}_1 . The subgroups from \mathcal{D}_1 cover the most elements with cycle structure $(4, 5, 5)$ when each subgroup is isomorphic to $S_5 \times S_9$, so we will assume that each subgroup of \mathcal{D}_1 is isomorphic to $S_5 \times S_9$ to attain an upper bound. Each element with cycle structure $(4, 5, 5)$ is contained in exactly two subgroups isomorphic to $S_5 \times S_9$, and two subgroups $\text{Sym}(A) \times \text{Sym}(\{1, \dots, 14\} \setminus A)$ and $\text{Sym}(B) \times \text{Sym}(\{1, \dots, 14\} \setminus B)$ isomorphic to $S_5 \times S_9$ in \mathcal{D}_1 overlap in these elements precisely when $A \cap B = \emptyset$. Since

both A and B are subsets of $\{4, \dots, 14\}$, and we are assuming that \mathcal{D}_1 contains $\text{Sym}(A) \times \text{Sym}(\{1, \dots, 14\} \setminus A)$ for every subset A of $\{4, \dots, 14\}$ of size 5, each subgroup in \mathcal{D}_1 intersects exactly $\binom{11-5}{5} = 6$ other subgroups of \mathcal{D}_1 in elements of Π_4 . Since each element of Π_4 is contained in exactly two subgroups isomorphic to $S_5 \times S_9$, there are exactly

$$\frac{1}{2} \binom{11}{5} \binom{6}{5} \cdot 3! \cdot 4! \cdot 4! = 4790016$$

elements of Π_4 that are contained in two subgroups of \mathcal{D}_1 . Hence \mathcal{D}_1 contains at most $435456 \cdot |\mathcal{D}_1| - 4790016$ elements with cycle structure $(4, 5, 5)$, which implies

$$c_4 \leq \frac{n_4}{435456} \leq \frac{435456 \cdot |\mathcal{D}_2| + 435456 \cdot |\mathcal{D}_1| - 4790016}{435456} = |\mathcal{D}_2| + |\mathcal{D}_1| - 11 = |\mathcal{B}'| - 11.$$

Therefore,

$$c_3 + c_4 = |\mathcal{C}'| \geq |\mathcal{B}'| \geq 11 + c_4,$$

and so $c_3 \geq 11$, as desired. \square

We now further characterize a hypothetical minimal cover \mathcal{B} of the elements of Π .

Lemma 2.5. *Assume that $H_1 \notin \mathcal{B}$, and let the subgroup $H_2 \cong S_3 \times S_{11}$ of S_{14} stabilize the decomposition $B_2 \cup (\{1, \dots, 14\} \setminus B_2)$, where $|B_2| = 3$. If $H_2 \notin \mathcal{B}$, then $\{1, 2, 3\} \cap B_2 \neq \emptyset$.*

Proof. Let B_2 indeed be such a set without overlap with $\{1, 2, 3\}$ — without loss of generality, say it is $\{4, 5, 6\}$. The output of `PossibleExtensions`($[[1, 2, 3], [4, 5, 6]]$) in GAP (see Function 7 in the online supplement) shows that, up to an automorphism, $\{1, 2, 4\}$ is the only possibility for B_3 , where $H_3 \cong S_3 \times S_{11}$ stabilizes the decomposition of $\{1, \dots, 14\}$ into B_3 and $\{1, \dots, 14\} \setminus B_3$ and $H_3 \notin \mathcal{B}$. The output of `PossibleExtensions`($[[1, 2, 3], [4, 5, 6], [1, 2, 4]]$) reveals that no set of four subgroups not in \mathcal{B} can contain two subgroups whose corresponding 3-sets are disjoint. By Lemma 2.4, there are at least 11 subgroups isomorphic to $S_3 \times S_{11}$ not in \mathcal{B} , and so, without loss of generality, $\{1, 2, 3\} \cap B_2 \neq \emptyset$. \square

We may now use the program `PossibleExtensions_2` (see Function 8 in the online supplement), on the presumption that corresponding fixed 3-sets representing groups isomorphic to $S_3 \times S_{11}$ removed from \mathcal{B} must intersect.

Lemma 2.6. *If a collection H_1, \dots, H_k is not in \mathcal{B} , where H_i stabilizes a decomposition of the set $\{1, \dots, 14\}$ into $B_i \cup \{1, \dots, 14\} \setminus B_i$, $|B_i| = 3$, and $B_1 = \{1, 2, 3\}$, then we may assume $1 \in \bigcap_{i=1}^k B_i$.*

Proof. We observe at the outset that, by Lemma 2.4, $H_1 \notin \mathcal{B}$ implies that $k \geq 11$. Again without loss of generality, we let B_2 be one of $\{1, 2, 4\}$ or $\{1, 4, 5\}$, since $|B_1 \cap B_2| \in \{1, 2\}$. We will first examine the case where $B_2 = \{1, 4, 5\}$. The output of `PossibleExtensions_2`($[[1, 2, 3], [1, 4, 5]]$) shows that,

without loss of generality, the only possibilities for B_3 , when $1 \notin B_3$, are $\{2, 3, 4\}$ and $\{2, 4, 6\}$. The output of $\text{PossibleExtensions_2}([1, 2, 3], [1, 4, 5], [2, 3, 4])$ then shows that if $B_3 = \{2, 3, 4\}$, the only possibility for B_4 is $\{1, 2, 4\}$, and the output of $\text{PossibleExtensions_2}([1, 2, 3], [1, 4, 5], [2, 3, 4], [1, 2, 4])$ shows there is no possibility for B_5 . Meanwhile, if $B_3 = \{2, 4, 6\}$, the output of $\text{PossibleExtensions_2}([1, 2, 3], [1, 4, 5], [2, 4, 6])$ shows that there is no possible B_4 in this case. Therefore, if $|B_1 \cap B_2| = 1$, then we may assume that $1 \in B_i$ for any i , $1 \leq i \leq k$.

Now let $B_2 = \{1, 2, 4\}$; i.e., let $B_1 \cap B_2 = \{1, 2\}$. Then up to symmetry, $1 \in B_3$ is equivalent to $2 \in B_3$; thus, assuming $B_3 \cap \{1, 2\} = \emptyset$, without a loss of generality $\{3, 4\} \subseteq B_3$ and $B_3 = \{3, 4, 5\}$. The output of $\text{PossibleExtensions_2}([1, 2, 3], [1, 2, 4], [3, 4, 5])$ then shows that $B_4 = \{1, 3, 4\}$. Finally, we see that the output of $\text{PossibleExtensions_2}([1, 2, 3], [1, 2, 4], [3, 4, 5], [1, 3, 4])$ shows that there is no possible B_5 . Thus, if $B_1 \cap B_2 = \{1, 2\}$, then $B_i \cap \{1, 2\} \neq \emptyset$ for any i , $1 \leq i \leq k$. Note that this shows that $B_i \cap B_j \cap B_\ell \neq \emptyset$ for any $i, j, \ell \in \{1, \dots, k\}$.

Moreover, if $B_1 \cap B_2 = \{1, 2\}$ and $B_1 \cap B_2 \cap B_3 \cap B_4 = \emptyset$, then without loss of generality we may let $B_3 \cap \{1, 2\} = \{1\}$ and $B_4 \cap \{1, 2\} = \{2\}$. Note that if $B_3 \cap B_1 = \{1\}$, we are done, as in the first case above, as well as if $B_3 \cap B_2 = \{2\}$. Therefore, to continue, we must assume that $B_3 = \{1, 3, 4\}$, and similarly that $B_4 = \{2, 3, 4\}$. However, under these assumptions, $\text{PossibleExtensions_2}([1, 2, 3], [1, 2, 4], [1, 3, 4], [2, 3, 4])$ shows that it is impossible to extend the list to a B_5 . Therefore, all the B_i have nonempty intersection, and without loss of generality, $1 \in \bigcap_{i=1}^k B_i$. \square

Lemma 2.7. \mathcal{B} contains all subgroups isomorphic to $S_3 \times S_{11}$.

Proof. We again observe at the outset that, by Lemma 2.4, $H_1 \notin \mathcal{B}$ implies $k \geq 11$. Lemma 2.6 implies that we may assume each B_i is of the form $\{1, x, y\}$, where $x, y \in \{2, \dots, 14\}$. Hence there are at most $\binom{13}{2} = 78$ subgroups isomorphic to $S_3 \times S_{11}$ omitted from \mathcal{B} , meaning that for any potential list, we have that the output of the GAP function $455\text{Shortage}([\text{list}])$ is at most 78 (see Function 5 in the online supplement). However, we also have $455\text{Shortage}([1, 2, 3], [1, 4, 5]) = \frac{286}{3} > 78$, implying that any two subgroups H_i and H_j not in \mathcal{B} must have $|B_i \cap B_j| = 2$. Without loss of generality we may let $B_1 = \{1, 2, 3\}$ and $B_2 = \{1, 2, 4\}$, and assume that $2 \notin B_3$. Then since $|B_1 \cap B_3| = |B_2 \cap B_3| = 2$, necessarily $B_3 = \{1, 3, 4\}$. However, $455\text{Shortage}([1, 2, 3], [1, 2, 4], [1, 3, 4]) = 106 > 78$, so without loss of generality all B_i contain $\{1, 2\}$, meaning that for all i , there exists some x such that $B_i = \{1, 2, x\}$. Since there are only 12 such x possible and $455\text{Shortage}([1, 2, 3], [1, 2, 4]) = 46 > 12$, we have a contradiction. Thus, all 364 subgroups isomorphic to $S_3 \times S_{11}$ are in any minimal cover \mathcal{B} of S_{14} . \square

Theorem 2.8. \mathcal{C}_{14} is a minimal cover of Π (and therefore of S_{14}), and $\sigma(S_{14}) = 3096$.

Proof. Since subgroups isomorphic to either $S_4 \times S_{10}$ or $S_5 \times S_9$ contain the same number of Π_4 elements (those with $(4, 5, 5)$ cycle structure) — 435456 — the best-case scenario for covering those elements is the number of such elements divided by 435456, namely $\binom{14}{4} \frac{1}{2} \binom{10}{5} \cdot 3! \cdot 4! \cdot 4! / 435456 = 1001$. By Lemmas 2.3 and 2.7, we have already established that every other class of subgroups contained in C'_{14} is shared by \mathcal{B} . Therefore, any minimal cover of $\Pi_3 \cup \Pi_4$ must contain at least $364 + 1001 = 1365$ subgroups, and so any minimal cover of Π (and hence any minimal cover of S_{14}) contains at least $1 + 14 + 1716 + 1365 = 3096$ subgroups. Combined with Lemma 2.1, we have $\sigma(S_{14}) = 3096$. \square

Acknowledgements

The authors would like to thank Luise-Charlotte Kappe for comments on an early version of this paper and the referees for suggestions that greatly improved the final version of this paper.

References

- [Blackburn 2006] S. R. Blackburn, “Sets of permutations that generate the symmetric group pairwise”, *J. Combin. Theory Ser. A* **113**:7 (2006), 1572–1581. MR Zbl
- [Garonzi 2013] M. Garonzi, “Finite groups that are the union of at most 25 proper subgroups”, *J. Algebra Appl.* **12**:4 (2013), art. id. 1350002. MR Zbl
- [Holmes and Maróti 2010] P. E. Holmes and A. Maróti, “Pairwise generating and covering sporadic simple groups”, *J. Algebra* **324**:1 (2010), 25–35. MR Zbl
- [Kappe et al. 2016] L.-C. Kappe, D. Nikolova-Popova, and E. Swartz, “On the covering number of small symmetric groups and some sporadic simple groups”, *Groups Complex. Cryptol.* **8**:2 (2016), 135–154. MR Zbl
- [Maróti 2005] A. Maróti, “Covering the symmetric groups with proper subgroups”, *J. Combin. Theory Ser. A* **110**:1 (2005), 97–111. MR Zbl
- [Neumann 1954] B. H. Neumann, “Groups covered by permutable subsets”, *J. London Math. Soc.* **29** (1954), 236–248. MR Zbl
- [Serena 2003] L. Serena, “On finite covers of groups by subgroups”, pp. 173–190 in *Advances in group theory 2002*, edited by F. de Giovanni and M. L. Newell, Aracne, Rome, 2003. MR Zbl
- [Swartz 2016] E. Swartz, “On the covering number of symmetric groups having degree divisible by six”, *Discrete Math.* **339**:11 (2016), 2593–2604. MR Zbl
- [Tomkinson 1997] M. J. Tomkinson, “Groups as the union of proper subgroups”, *Math. Scand.* **81**:2 (1997), 191–198. MR Zbl

Received: 2017-07-09 Revised: 2017-11-28 Accepted: 2017-12-30

raoppenheim@email.wm.edu Department of Mathematics, College of William and Mary,
Williamsburg, VA, United States

easwartz@wm.edu Department of Mathematics, College of William and Mary,
Williamsburg, VA, United States

Upper and lower bounds on the speed of a one-dimensional excited random walk

Erin Madden, Brian Kidd, Owen Levin,
Jonathon Peterson, Jacob Smith and Kevin M. Stangl

(Communicated by John C. Wierman)

An excited random walk (ERW) is a self-interacting non-Markovian random walk in which the future behavior of the walk is influenced by the number of times the walk has previously visited its current site. We study the speed of the walk, defined as $V = \lim_{n \rightarrow \infty} (X_n/n)$, where X_n is the state of the walk at time n . While results exist that indicate when the speed is nonzero, there exists no explicit formula for the speed. It is difficult to solve for the speed directly due to complex dependencies in the walk since the next step of the walker depends on how many times the walker has reached the current site. We derive the first nontrivial upper and lower bounds for the speed of the walk. In certain cases these upper and lower bounds are remarkably close together.

1. Introduction

A simple random walk on \mathbb{Z} can be thought of as a simple discrete model for random motion where at each time step the “walker” tosses a (possibly biased) coin and steps right if he gets a heads and left if he gets a tails. Mathematically, if we denote the position of the walk after n steps by S_n then we can represent the walk as $S_n = \sum_{i=0}^n \xi_i$, where the sequence of random variables $\xi_1, \xi_2, \xi_3, \dots$ represents the successive steps of the walk. Since the steps are given by the outcomes of repeated tosses of a coin, the random variables $\{\xi_i\}_{i \geq 0}$ are independent and identically distributed (i.i.d.) with $P(\xi_1 = p)$ and $P(\xi_1 = -1) = 1 - p$ (here $p \in (0, 1)$ is the probability that the coin the walker is tossing comes up heads).

Simple random walks are very well known and much is known about them, but in this paper we will focus on a different model for random motion called an excited random walk. In an excited random walk, rather than the steps of the walk being i.i.d. the probability of the walker moving right (+1) or left (−1) from a site on the n -th step is a function of how many times the walker has stepped on that site

MSC2010: primary 60K35; secondary 60G50.

Keywords: excited random walk, Markov chain, stationary distribution.

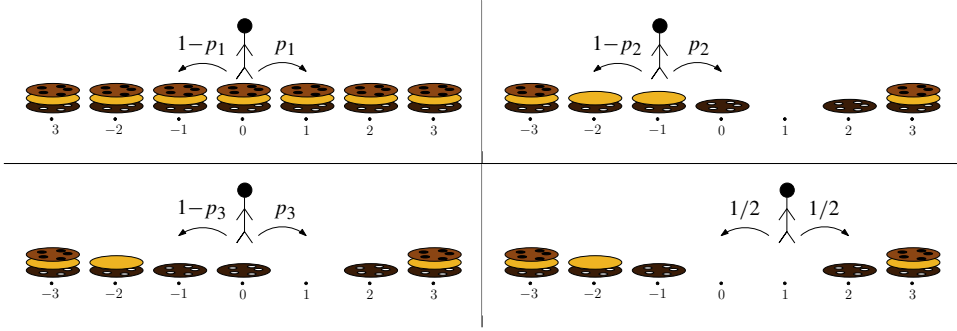


Figure 1. A partial example of an excited random walk with $M = 3$, $\vec{p} = (p_1, p_2, p_3)$, and transition probabilities shown. Top left: the initial state of the walker. Top right: a possible state after 9 steps. Bottom left: 10 steps into the same walk, with the most recent step to the right. Bottom right: 11 steps into the walk, the walker is now in a state with no more cookies left and has equal transition probabilities to the left and right.

by time n . To describe the excited random walk model, we begin by fixing an integer $M \geq 1$ and parameters $p_1, p_2, \dots, p_M \in (0, 1)$. When the walker visits a location i for the j -th time, if $j \leq M$ then the walker tosses a coin with probability of heads p_j , while if $j > M$ the walker tosses a fair coin ($p = \frac{1}{2}$) to determine if the next step is left or right. That is, an excited random walk is a stochastic process $\{X_n\}_{n \geq 0}$ starting at $X_0 = 0$ and such that $X_{n+1} = X_n \pm 1$ and

$$\begin{aligned} \mathbb{P}(X_{n+1} = X_n + 1 \mid X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) \\ = \begin{cases} p_j & \text{if } \#\{k \leq n : x_k = x_n\} = j \leq M, \\ \frac{1}{2} & \text{if } \#\{k \leq n : x_k = x_n\} > M. \end{cases} \end{aligned}$$

Excited random walks are sometimes also called “cookie random walks” due to the following interpretation of the dynamics. We imagine that initially there is an identical stack of M cookies at each site. At every step the random walker takes the top cookie from the stack at the current site (if there is at least one cookie left) and eats it. The cookie induces an “excitement” or drift which causes the walker to step to the right with probability p_j (or left with probability $1 - p_j$). If the walker ever returns to a site where all the cookies have already been eaten then there is nothing to “excite” him and so he steps left/right with equal probability. See Figure 1. Due to this “cookie” interpretation of excited random walks we will often refer to the parameter M as the number of cookies at each site and the parameter p_j as the “strength” of the j -th cookie.

1.1. Background and previous results. Excited random walks were first introduced by Benjamini and Wilson [2003]. In the model they considered, however, there was only one cookie at each site $M = 1$. This model was then generalized by Zerner [2005] to allow for multiple cookies at each site, but with the restriction that all $p_j \geq \frac{1}{2}$; that is, all cookies induced a nonnegative drift for the walker. Kosygina and Zerner [2008] further generalized the model to allow for the possibility of both “positive” ($p_j > \frac{1}{2}$) and “negative” ($p_j < \frac{1}{2}$) cookies in the stack of cookies at each site. In fact, the model of excited random walks is even more general than what we have described here. Certain results have even allowed for placing random cookie stacks at sites (rather than the same cookie stack at each site) and for infinitely many cookies at each site. In this paper, however, we will restrict ourselves to the simpler model described above of M cookies at each site with strengths p_1, p_2, \dots, p_M .

The behavior of simple random walks is quite easy to analyze since, as noted above, the walk $S_n = \sum_{i=1}^n \xi_i$ is the sum of i.i.d. random variables. In particular, the law of large numbers implies $\lim_{n \rightarrow \infty} (S_n/n) = E[\xi_1] = 2p - 1$ with probability 1. That is, the random walk has a deterministic limiting speed of $2p - 1$. Thus, if $p > \frac{1}{2}$ then the walk moves to the right with positive speed, while if $p < \frac{1}{2}$, the walk moves to the left with speed $1 - 2p$ (or equivalently, for any $p \in [0, 1]$ the walker simply moves with *velocity* $2p - 1$). In either of these cases we say that the walk is *transient* since it only visits any site a finite number of times. More generally, if a random walk is transient with nonzero speed, it is *ballistic*. For one-dimensional simple random walks, transience and ballisticity are equivalent, but as we will see in our discussion of excited random walks, this is not always the case. The case $p = \frac{1}{2}$ is more delicate, but it was shown by Pólya [1921] that a one-dimensional simple symmetric random walk is *recurrent*; that is, the walk visits every site infinitely many times.

In contrast to simple random walks, the behavior of excited random walks is much more difficult to determine since the self-interacting nature of the walk creates dependencies among steps of the walk that are very hard to handle. Moreover, the behavior of the walk is at times like a biased random walk (on the first M visits to sites), while at other times it is like a symmetric random walk (after more than M visits to a site). Thus, even the question of determining whether the excited random walk is recurrent or transient is quite difficult. In spite of these difficulties, a number of characteristics of excited random walks have been determined to depend on a single easy to calculate parameter.

$$\delta = \sum_{j=1}^M (2p_j - 1). \quad (1)$$

We will use the notation $\delta_j = 2p_j - 1$ for the drift of the j -th cookie in the cookie stack. Thus, the parameter $\delta = \sum_{j=1}^M \delta_j$ can be thought of as the net total drift contained in all the cookies in the cookie stack at each site.

Theorem 1 [Zerner 2005; Kosygina and Zerner 2008]. *The parameter δ determines the recurrence or transience of the excited random walk:*

(1) *If $\delta > 1$ then the walk is transient to the right, that is,*

$$\mathbb{P}(\lim_{n \rightarrow \infty} X_n = +\infty) = 1.$$

(2) *If $\delta < -1$ then the walk is transient to the left, that is,*

$$\mathbb{P}(\lim_{n \rightarrow \infty} X_n = -\infty) = 1.$$

(3) *If $\delta \in [-1, 1]$ then the walk is recurrent, that is,*

$$\mathbb{P}(\liminf_{n \rightarrow \infty} X_n = -\infty, \limsup_{n \rightarrow \infty} X_n = +\infty) = 1.$$

Zerner [2005] also proved that excited random walks have a limiting speed. That is, given any parameters M and $\vec{p} = (p_1, p_2, \dots, p_M)$ for an excited random walk there is a constant $V_{M, \vec{p}} \in [-1, 1]$ such that

$$\lim_{n \rightarrow \infty} \frac{X_n}{n} = V_{M, \vec{p}}, \quad \text{with probability 1.} \quad (2)$$

Determining the exact value of the speed $V_{M, \vec{p}}$ as a function of M and \vec{p} , however, remains an open problem and is the focus of this paper. While there is still no explicit formula for $V_{M, \vec{p}}$ in general, it is known that the parameter δ determines exactly when the speed is positive, negative or zero.

Theorem 2 [Basdevant and Singh 2008a; Kosygina and Zerner 2008]. *The parameter δ determines the sign of the limiting speed $V_{M, \vec{p}}$ of the excited random walk:*

(1) *If $\delta > 2$ then $V_{M, \vec{p}} > 0$.*

(2) *If $\delta < -2$ then $V_{M, \vec{p}} < 0$.*

(3) *If $\delta \in [-2, 2]$ then $V_{M, \vec{p}} = 0$.*

Remark 3. Note that Theorems 1 and 2 together highlight a very peculiar feature of excited random walks: if $\delta \in (1, 2]$ then the walk is transient to the right, but with zero asymptotic speed. At first this might seem contradictory, but in fact it holds because in this case X_n grows to infinity roughly like $n^{\delta/2}$ if $\delta \in (1, 2)$ or like $n/\log n$ if $\delta = 2$ [Basdevant and Singh 2008b; Kosygina and Zerner 2008].

Example 4. Let $M = 3$ and $\vec{p} = (p, p, p)$. Then $\delta = 6p - 3$.

(1) If $p \in [\frac{1}{3}, \frac{2}{3}]$ then $\delta \in [-1, 1]$, so the walk is recurrent.

(2) If $p \in [\frac{1}{6}, \frac{5}{6}]$ then $\delta \in [-2, 2]$, so the walk is transient with $V_{M, \vec{p}} = 0$.

(3) If $p \in [0, \frac{1}{6})$ then $\delta < -2$, so the walk is ballistic with $V_{M, \vec{p}} < 0$.

(4) If $p \in (\frac{5}{6}, 1]$ then $\delta > 2$, so the walk is ballistic with $V_{M, \vec{p}} > 0$.

Remark 5. It should be noted that if $p_i \in (0, 1)$ for all i , then unless $M \geq 3$, $V_{M,\vec{p}} = 0$. If $M < 3$ then $\delta < 4 \cdot 1 - 2 = 2$. Thus, $V_{M,\vec{p}}$ is nonpositive. A symmetric argument shows that $\delta > -2$ and thus $V_{M,\vec{p}} = 0$ unless $M \geq 3$.

Theorem 2 shows that we can identify the speed of the excited random walk exactly when the speed is zero (when $\delta \in [-2, 2]$). However, as noted above when the speed is nonzero (when $\delta \notin [-2, 2]$), there is no explicit formula for the speed $V_{M,\vec{p}}$. The focus of this paper is to compute explicit upper and lower bounds for the speed in these cases. For simplicity we will restrict ourselves to the case of positive speed ($\delta > 2$) since the negative-speed case can be handled similarly by symmetric arguments. Prior to this paper, when $\delta > 2$ the only known upper and lower bounds on the speed were the trivial ones

$$0 < V_{M,\vec{p}} \leq \max_{j \leq M} (2p_j - 1).$$

The upper bound on the right is the speed of a simple random walk which moves to the right with probability $p^* = \max_{j \leq M} p_j$ on each step. Since this simple random walk is always at least as likely to step right as the excited random walk, it is easy to see that the excited random walk has a speed that is less than or equal to that of this simple random walk. We will develop a method below for obtaining much better bounds than these trivial bounds. In particular, in the case of $M = 3$ cookies per site we will obtain upper and lower bounds which differ by at most 0.0194565.

The rest of the paper will be organized as follows. We begin with a brief introduction to the theory of Markov chains to cover results we will use. Then we describe a particular Markov chain related to excited random walks, known as the backward branching process. We discuss known results about this Markov chain and how they relate to the speed of an excited random walk. Afterward, we derive bounds on the speed using properties of the backward branching process. We end with a discussion of how well these bounds approximate the speed.

2. A related Markov chain

We will introduce a Markov chain that is useful for studying the speed of excited random walks. First, however, we will give a short overview of the notation and terminology of Markov chains and recall a few useful facts about Markov chains.

2.1. Markov chains. Recall that a Markov chain on a countable state space I is a stochastic process $\{Z_n\}_{n \geq 0}$ such that for any choice of $n \geq 1$ and $i_0, i_1, \dots, i_n, i_{n+1} \in I$ we have

$$\begin{aligned} \mathbb{P}(Z_{n+1}=i_{n+1} \mid Z_0=i_0, Z_1=i_1, \dots, Z_{n-1}=i_{n-1}, Z_n=i_n) &= \mathbb{P}(Z_{n+1}=i_{n+1} \mid Z_n=i_n) \\ &= \mathbb{P}(Z_1=i_{n+1} \mid Z_0=i_n). \end{aligned}$$

The transition matrix for the Markov chain is the matrix

$$P = (p(i, j))_{i, j \in I}, \quad \text{where } p(i, j) = \mathbb{P}(X_1 = j \mid X_0 = i).$$

For ease of notation, if the Markov chain starts at $Z_0 = i$ we will write $\mathbb{P}_i(\cdot)$ in place of $\mathbb{P}(\cdot \mid Z_0 = i)$. If the Markov chain starts from a random initial condition given by $\mu = (\mu(i))_{i \in I}$, where $\mu(i)$ is the probability that the Markov chain starts at $Z_0 = i$, then we will denote this with the notation \mathbb{P}_μ ; that is, $\mathbb{P}_\mu(\cdot) = \sum_i \mu(i) \mathbb{P}_i(\cdot)$. Expectations with respect to the probability distributions \mathbb{P}_i and \mathbb{P}_μ for the Markov chain are denoted by \mathbb{E}_i and \mathbb{E}_μ , respectively.

A special choice of an initial distribution is a *stationary distribution*. A probability distribution $\pi = (\pi(i))_{i \in I}$ is a stationary distribution for the Markov chain $Z = \{Z_n\}_{n \geq 0}$ if $\mathbb{P}_\pi(Z_1 = j) = \mathbb{P}_\pi(Z_0 = j) = \pi(j)$ for all $j \in I$, that is, if Z_1 has the same distribution π as Z_0 (and thus, by induction, Z_n has the same distribution as Z_0 for all $n \geq 1$). If π is a stationary distribution then

$$\pi(j) = \mathbb{P}_\pi(Z_1 = j) = \sum_{i \in I} \pi(i) \mathbb{P}_i(X_1 = j) = \sum_{i \in I} \pi(i) p(i, j),$$

so that viewing $\pi = (\pi(i))_{i \in I}$ as a row vector we have $\pi = \pi P$; that is, π is a left eigenvector of the transition matrix P with eigenvalue 1. If the state space I of the Markov chain is finite, then computing the stationary distributions is a simple problem in linear algebra. However, if the state space I is countably infinite then computing stationary distributions is much more difficult and in fact, for some infinite state Markov chains there are no stationary distributions. It is known, however, that if the Markov chain is irreducible (that is, if it is possible starting at any state i to eventually reach any other state j) and there is a stationary distribution then it is unique.

Stationary distributions are important for the analysis of Markov chains because they can be used to determine the long-run asymptotics of the Markov chain. For instance, if the Markov chain is irreducible and a stationary distribution π exists, then it is known that for any initial starting condition

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n Z_k = \mathbb{E}_\pi[Z_0] = \sum_{j \in I} \pi(j) j, \quad \text{with probability 1.}$$

2.2. The backward branching process. Because the transition probabilities of the excited random walk depend on the number of prior visits to the present location and not only on the current location of the walk, an excited random walk is not a Markov chain. However, there is a Markov chain we can study that can give information about the excited random walk. This Markov chain is often referred to in the literature as the “backward branching process” due to some structural similarity with models for population growth known as branching processes. The

backward branching process is related to the excited random walk through an analysis of the number of left (or backward) crossings of edges of the excited random walk before the walk reaches some point to the right for the first time. We refer the reader interested in the details of this connection to [Basdevant and Singh 2008a]. Here we only provide a description of the transition probabilities for this Markov chain and the relevance to the limiting speed of the excited random walk.

To describe the transition probabilities for the backwards branching process, we imagine an infinite sequence of independent coin flips where for the first M flips we use coins which come up heads with probability p_j for $j = 1, 2, \dots, M$ and then for all subsequent flips we use a fair coin. Mathematically we can represent this as the sequence $\{\xi_j\}_{j \geq 1}$ of independent Bernoulli random variables where

$$\mathbb{P}(\xi_j = 1) = \begin{cases} p_j & \text{if } j \leq M, \\ \frac{1}{2} & \text{if } j > M. \end{cases}$$

Next, for any $m \geq 1$ we let

$$F_m = \inf \left\{ k \geq 0 : \sum_{j=1}^{m+k} \xi_j \geq m \right\}.$$

Again viewing the $\{\xi_j\}_{j \geq 1}$ as the outcomes of successive coin tosses, we have that F_m can be interpreted as the number of “tails” before the m -th “heads”. Finally, using this notation we are able to define the backward branching process associated to the excited random walk with parameters M and $\vec{p} = (p_1, p_2, \dots, p_M)$ as the Markov chain $Z = \{Z_n\}_{n \geq 0}$ on $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$ with transition probabilities given by

$$p(i, j) = \mathbb{P}(F_{i+1} = j) \quad \text{for } i, j \geq 0.$$

Example 6. Some transition probabilities which we will use later in Lemma 15 are given below. Also we show the full transition matrix for when $p_1 = p_2 = p_3 = p$. When $M = 3$ cookies per site we have

- $p(0, 0) = p_1$ (no tails before a single heads),
- $p(0, 1) = (1 - p_1)p_2$ (one tail before a single heads),
- $p(0, 2) = (1 - p_1)(1 - p_2)p_3$ (two tails before a single heads),
- $p(0, k) = (1 - p_1)(1 - p_2)(1 - p_3)/2^{k-2}$ for $k \geq 3$ (k tails before a single heads),
- $p(1, 0) = p_1p_2$ (no tails before two heads),
- $p(1, 1) = (1 - p_1)p_2p_3 + p_1(1 - p_2)p_3$ (one tail before two heads),
- $p(k, 0) = p_1p_2p_3/2^{k-2}$ for $k > 3$ (no tails before $k+1$ heads).

In the $M = 3$ case where $p_1 = p_2 = p_3 = p$, (letting $q := 1 - p$), the initial entries of the transition matrix (with $i, j \leq 2$) are

$$\begin{pmatrix} p & pq & pq^2 & \cdots \\ p^2 & 2p^2q & \frac{3}{2}pq^2 & \cdots \\ p^3 & \frac{3}{2}p^2q & \frac{3}{4}(pq^2 + p^2q) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

and the remaining entries (when either $i > 2$ or $j > 2$) are given by

$$p(i, j) = \frac{1}{2^{i+j-2}} \left[\binom{i+j-3}{i-3} p^3 + \binom{i+j-3}{j-3} q^3 + 3 \binom{i+j-3}{i-2} p^2q + 3 \binom{i+j-3}{j-2} pq^2 \right].$$

The Markov chain Z was first introduced in the study of excited random walks by Basdevant and Singh [2008a]. It is easy to see that the Markov chain Z is irreducible since $p(i, j) > 0$ for all $i, j \geq 0$. Moreover, Basdevant and Singh showed that the Markov chain Z has a (unique) stationary distribution π whenever $\delta > 1$ (or equivalently, by Theorem 1, when the excited random walk is transient to the right). Most importantly, Basdevant and Singh proved that the limiting speed $V_{M, \vec{p}}$ for the excited random walk can be expressed in terms of the stationary distribution for the Markov process Z in the following theorem.

Theorem 7 [Basdevant and Singh 2008a]. *Suppose the parameters M and $\vec{p} = (p_1, p_2, \dots, p_M)$ are such that the speed $V_{M, \vec{p}}$ is positive (that is, $\delta > 2$). If π is the stationary distribution for the corresponding backward branching process $Z = \{Z_n\}_{n \geq 0}$, then*

$$V_{M, \vec{p}} = \frac{1}{1 + 2\mathbb{E}_\pi[Z_0]}. \quad (3)$$

A rationalization for and proof sketch of Theorem 7 come from the following. Because $\delta > 2$, the walk X is transient and almost surely $\lim_{n \rightarrow \infty} (X_n/n) = V_{M, \vec{p}} > 0$. In such situations, it holds that almost surely

$$\lim_{n \rightarrow \infty} \frac{X_n}{n} = \frac{1}{\lim_{n \rightarrow \infty} (T_n/n)},$$

where T_n is the hitting time of site n . Essentially, this identity is just noting that distance over time can be expressed in terms of two different quantities for X and each are equivalent to the velocity of the walk.

Now, the hitting-time limit can be expressed in terms of the backward branching process by

$$\lim_{n \rightarrow \infty} \frac{T_n}{n} = \lim_{n \rightarrow \infty} \frac{n + 2 \sum_{k=1}^n Z_k}{n}.$$

To see this, we count the number of steps making up the hitting time to site n . The number of total steps down from positive site k to site $k - 1$ before the walk reaches n is $\sum_{k=1}^n Z_k$. Each of these down steps is canceled by one step back up to site k before reaching n . In addition, we have the final up step from each positive site k up to n , which is n steps. Lastly, T_n contains the total number of steps from 0 to -1 and all the steps contained in the negative half-line. Because X is transient to $+\infty$ when $\delta > 2$, there are a finite (random) number L of these steps and $L/n \rightarrow 0$ almost surely as n goes to ∞ . Then we have the following equalities which imply the conclusion of Theorem 7:

$$\begin{aligned} \frac{1}{V_{M,p}} &= \lim_{n \rightarrow \infty} \frac{T_n}{n} = \lim_{n \rightarrow \infty} \frac{L + n + 2 \sum_{k=1}^n Z_k}{n} \\ &= \lim_{n \rightarrow \infty} \frac{L}{n} + \frac{n}{n} + 2 \frac{1}{n} \sum_{k=1}^n Z_k = 1 + 2\mathbb{E}_\pi[Z_0]. \end{aligned}$$

While Theorem 7 expresses the speed $V_{M,\tilde{p}}$ in terms of the stationary distribution of the backward branching process, unfortunately, this doesn't give an explicit formula for the speed since there is not yet an explicit formula for the stationary distribution π (solving the infinite system of equations $\pi P = \pi$ is too difficult). In the following section, however, we will develop some methods which can be used to obtain rigorous upper and lower bounds on $\mathbb{E}_\pi[Z_0]$ and consequently upper and lower bounds on $V_{M,\tilde{p}}$.

3. Reduction of the formula for the speed

We will show how some recursive formulas for the probability-generating function of the distribution π can be used to get useful approximations (upper and lower bounds) of $\mathbb{E}_\pi[Z_0]$. The starting point of our analysis of the speed of the excited random walk is a recursive formula for the probability-generating function $G(s) := \sum_{k=0}^{\infty} \pi(k)s^k$ of the stationary distribution π for the Markov chain Z . Basdevant and Singh [2008a] showed that $G(s)$ is the unique solution of the functional equation

$$1 - G\left(\frac{1}{2-s}\right) = A(s)[1 - G(s)] + B(s), \quad s \in [0, 1], \quad (4)$$

where

$$A(s) = \frac{1}{(2-s)^{M-1} \mathbb{E}_{M-1}[s^{Z_1}]},$$

and

$$B(s) = 1 - \frac{1}{(2-s)^{M-1} \mathbb{E}_{M-1}[s^{Z_1}]} + \sum_{k=0}^{M-2} \pi(k) \left(\frac{\mathbb{E}_k[s^{Z_1}]}{(2-s)^{M-1} \mathbb{E}_{M-1}[s^{Z_1}]} - \frac{1}{(2-s)^k} \right). \quad (5)$$

While the recursive equation (4) is still too hard to solve explicitly, using the fact that $1/(2-s) \approx s$ when $s \approx 1$, Basdevant and Singh were able to use (4) to obtain asymptotics of the function $G(s)$ near $s = 1$. This is particularly useful because of the property of probability-generating functions that

$$G'(1) = \sum_{k=1}^{\infty} \pi(k)k = \mathbb{E}_{\pi}[Z_0]. \quad (6)$$

By careful analysis of this recursive equation near $s = 1$ and using the formula (3) for the speed, Basdevant and Singh were able to deduce the following implicit formula for the speed of an ERW.

Theorem 8 [Basdevant and Singh 2008a]. *If the speed is nonzero (i.e., if $\delta > 2$), then*

$$\mathbb{E}_{\pi}[Z_0] = G'(1) = \frac{B''(1)}{2(\delta - 2)}$$

and consequently the speed is equal to

$$V_{M,\vec{p}} = \frac{\delta - 2}{\delta - 2 + B''(1)}, \quad (7)$$

where $B(s)$ is defined in (5).

In deriving the representation (7) for the speed, Basdevant and Singh were primarily interested in determining when the speed $V_{M,p}$ was positive. However, an additional consequence of this formula is that it comes much closer to giving an explicit formula for the speed. While computing $\mathbb{E}_{\pi}[Z_0]$ using the standard formula in (6) requires knowing all of the stationary distribution, Theorem 8 shows we can instead compute this using only the $M - 1$ values $\pi(0), \pi(1), \dots, \pi(M - 2)$. This is because all of the probability-generating functions $\mathbb{E}_k[s^{Z_1}]$ can be computed explicitly so that the only unknown terms in $B(s)$ are $\pi(0), \pi(1), \dots, \pi(M - 2)$.

Example 9. In the general case of $M = 3$ cookies, the formula for $B(s)$ involves $\mathbb{E}_k[s^{Z_1}]$ for $k = 0, 1, 2$. These can be explicitly computed using the formulas for the transition probabilities $p(k, j)$ for the backward branching process:

$$\begin{aligned} \mathbb{E}_0[s^{Z_1}] &= p(0, 0) + sp(0, 1) + s^2p(0, 2) + \sum_{k=3}^{\infty} s^k p(0, k) \\ &= p_1 + s[(1 - p_1)p_2] + s^2[(1 - p_1)(1 - p_2)p_3] \\ &\quad + (1 - p_1)(1 - p_2)(1 - p_3) \sum_{k=3}^{\infty} \frac{s^k}{2^{k-2}} \\ &= p_1 + s[(1 - p_1)p_2] + s^2[(1 - p_1)(1 - p_2)p_3] - \frac{(1 - p_1)(1 - p_2)(1 - p_3)s^3}{s - 2}. \end{aligned}$$

Similar explicit calculations show that

$$\mathbb{E}_1[s^{Z_1}] = \frac{s(2p_2(s-1)-s)(2p_3(s-1)-s)}{(s-2)^2} - \frac{p_1(s-1)(p_2(2p_3(3s-4)s-3s^2+4)+2s(s-2p_3(s-1)))}{(s-2)^2},$$

and

$$\mathbb{E}_2[s^{Z_1}] = \frac{(2p_1(s-1)-s)(2p_2(s-1)-s)(2p_3(s-1)-s)}{(s-2)^3}.$$

As noted above, Theorem 8 shows that the speed $V_{M,\vec{p}}$ for an excited random walk can be expressed in terms of only the unknown values $\pi(0), \pi(1), \dots, \pi(M-2)$. The following lemma, however, gives a linear relation among these parameters so that we can actually eliminate one of the unknowns.

Lemma 10. *The unique stationary distribution π of $\{Z_n\}_{n \geq 0}$ satisfies*

$$\delta - 1 = \sum_{k=0}^{M-2} \pi(k)(\mathbb{E}_k[Z_1] - k - 1 + \delta).$$

Remark 11. Note that for any fixed excited-random-walk parameters M and \vec{p} , the expectations $\mathbb{E}_k[Z_1] = \sum_{j=0}^{\infty} jp(k, j)$ appearing in Lemma 10 can be explicitly calculated.

Proof. Due to properties of the stationary distribution we know

$$\mathbb{E}_{\pi}[Z_0] = \mathbb{E}_{\pi}[Z_1],$$

or equivalently

$$\sum_{k=0}^{\infty} k\pi(k) = \sum_{k=0}^{\infty} \pi(k)\mathbb{E}_k[Z_1]. \quad (8)$$

In general, the expectations $\mathbb{E}_k[Z_1]$ have to be calculated individually using the transition probabilities for the Markov chain $\{Z_n\}_{n \geq 0}$. However, Basdevant and Singh [2008a, Lemma 3.3] showed that the following pattern emerges when $k \geq M-1$:

$$\mathbb{E}_k[Z_1] = k + 1 - \delta \quad \text{for all } k \geq M-1. \quad (9)$$

(We provide a proof of (9) in the Appendix.) Using this, and splitting both sums in (8) into $k \leq M-2$ and $k \geq M-1$, we obtain

$$\sum_{k=0}^{M-2} k\pi(k) + \sum_{k=M-1}^{\infty} k\pi(k) = \sum_{k=0}^{M-2} \pi(k)\mathbb{E}_k[Z_1] + \sum_{k=M-1}^{\infty} (k+1-\delta)\pi(k).$$

Noting that $\sum_{k=M-1}^{\infty} k\pi(k)$ appears on both sides, we reduce this to

$$\begin{aligned} \sum_{k=0}^{M-2} k\pi(k) &= \sum_{k=0}^{M-2} \pi(k) \mathbb{E}_k[Z_1] + (1-\delta) \sum_{k=M-1}^{\infty} \pi(k) \\ &= \sum_{k=0}^{M-2} \pi(k) \mathbb{E}_k[Z_1] + (1-\delta) - (1-\delta) \sum_{k=0}^{M-2} \pi(k), \end{aligned}$$

where in the last equality we used that

$$\sum_{k=M-1}^{\infty} \pi(k) = 1 - \sum_{k=0}^{M-2} \pi(k)$$

because π is a probability distribution. The statement of the lemma is then obtained by simplifying. \square

As a special case, when there are $M = 3$ cookies, Lemma 10 gives a simple linear relation between $\pi(0)$ and $\pi(1)$.

Corollary 12. *For $M=3$ cookies with strength $\vec{p} = (p_1, p_2, p_3)$, the linear equation*

$$a\pi(0) + b\pi(1) = c,$$

where (recalling the notation $\delta_j = 2p_j - 1$)

$$a := p_1(\delta_2 + \delta_3) + p_2\delta_3(1 - p_1),$$

$$b := \delta_3 p_1 p_2,$$

$$c := \delta - 1,$$

follows from above.

Proof. When $M = 3$, the equation in Lemma 10 becomes

$$\delta - 1 = [\mathbb{E}_0[Z_1] + \delta - 1] \cdot \pi(0) + [\mathbb{E}_1[Z_1] + \delta - 2] \cdot \pi(1). \quad (10)$$

Next, note that $E_0[Z_1]$ and $E_1[Z_1]$ can be explicitly calculated from the known transition probabilities for Z (compare with Examples 6 and 9 above). For example,

$$\begin{aligned} \mathbb{E}_0[Z_1] &= 0(p_1) + 1(1-p_1)p_2 + 2(1-p_1)(1-p_2)p_3 \\ &\quad + (1-p_1)(1-p_2)(1-p_3) \sum_{k=3}^{\infty} \frac{k}{2^{k-2}} \\ &= (1-p_1)p_2 + 2(1-p_1)(1-p_2)p_3 + 4(1-p_1)(1-p_2)(1-p_3) \\ &= 4 - 4p_1 - 3p_2 - 2p_3 + 3p_1p_2 + 2p_1p_3 + 2p_2p_3 - 2p_1p_2p_3, \end{aligned}$$

and similarly it can be shown that

$$\mathbb{E}_1[Z_1] = 5 - 2(p_1 + p_2 + p_3) - p_1p_2(2p_3 - 1) = 2 - \delta - p_1p_2\delta_3.$$

Substituting these formulas for $\mathbb{E}_0[Z_1]$ and $\mathbb{E}_1[Z_1]$ in (10) and simplifying we obtain the statement of the corollary. \square

4. Bounds on the speed

Theorem 8 and Lemma 10 combined show that the speed $V_{M,\vec{p}}$ of an excited random walk with $\delta > 2$ can be computed in terms of only the unknown values $\pi(0), \pi(1), \dots, \pi(M-3)$. Actually computing this function, however, is rather involved as especially computing $B''(1)$ is a tedious task. Thus, for the remainder of the paper we will restrict ourselves to the case $M = 3$ so that explicit computations can be done. With the aid of Mathematica to compute the derivatives in $B''(1)$, we were able to show the following.

Theorem 13. *For an excited random walk with $M = 3$ cookies of strengths $\vec{p} = (p_1, p_2, p_3)$, if $\delta > 2$, the limiting speed is equal to*

$$V_{3,\vec{p}} = \frac{f_1}{f_2 + f_3 \cdot \pi(0)}, \quad (11)$$

where

$$\begin{aligned} f_1 &= 2p_1 + 2p_2 + 2p_3 - 5, \\ f_2 &= 9 + 8(p_1p_2 + p_1p_3 + p_2p_3) - 10(p_1 + p_2 + p_3), \\ f_3 &= 2(2p_3 - 1)(p_1 + p_2 - 3p_1p_2). \end{aligned}$$

The formula in (11) doesn't quite calculate $V_{3,\vec{p}}$ explicitly since we do not know the value of $\pi(0)$. However, the following lemma shows that we can easily use this formula to compute upper and lower bounds on the speed.

Lemma 14. *Let f_1, f_2 and f_3 be as in Theorem 13. Then, if $\delta = \sum_{j=1}^3 (2p_j - 1) > 2$ the function $x \mapsto f_1/(f_2 + f_3x)$ is strictly positive and increasing for $x \in [0, 1]$.*

Proof. If $g(x) = f_1/(f_2 + f_3x)$, then $g'(x) = -f_1f_3/(f_2 + f_3x)^2$. Thus, to show that $g(x)$ is decreasing we need only to show that $f_1f_3 < 0$ when p_1, p_2, p_3 are such that $\delta > 2$. Note first of all that $\delta > 2$ is equivalent to $p_1 + p_2 + p_3 > \frac{5}{2}$. Therefore,

$$f_1 = 2(p_1 + p_2 + p_3) - 5 > 0,$$

and so it remains to show $f_3 < 0$. To see this, note that since p_1, p_2 and p_3 are each at most 1, the condition $\delta > 2$ implies that they are all strictly larger than $\frac{1}{2}$. Thus,

$$f_3 = 2(2p_3 - 1)(p_1 + p_2 - 3p_1p_2) < 0 \quad \text{if } p_1 + p_2 - 3p_1p_2 < 0.$$

When $\delta > 2$, it follows that $p_1 + p_2 \in (\frac{3}{2}, 2)$. Therefore, if we fix $t \in (\frac{3}{2}, 2)$ and if $p_1 + p_2 = t$ then

$$p_1 + p_2 - 3p_1p_2 = t - 3p_1(t - p_1) = 3p_1^2 + (1 - 3p_1)t$$

and we wish to show that this is negative for all $p_1 \in [t - 1, 1]$. However, since $3p_1^2 + (1 - 3p_1)t$ is convex in p_1 we need only to check the value at the endpoints $p_1 = t - 1$ and $p_1 = 1$, and at both endpoints this evaluates to $3 - 2t < 0$. This completes the proof that $f_3 < 0$ whenever $\delta > 2$ and thus also that $g(x)$ is decreasing for $x \in [0, 1]$.

Since we have already shown that $f_1 > 0$ and $f_3 < 0$ when $\delta > 2$, it will follow that $g(x)$ is nonnegative on $[0, 1]$ if we can show that $f_2 + f_3 > 0$ whenever $\delta > 2$. This will be accomplished by showing that

$$f_2 + f_3 \geq 0 \quad \text{when } \delta = 2, \quad (12)$$

and

$$\frac{\partial}{\partial p_i}(f_2 + f_3) > 0 \quad \text{for } i = 1, 2, 3 \text{ whenever } \delta > 2. \quad (13)$$

To show (12), note that if $\delta = 2$ then $p_1 + p_2 + p_3 = \frac{5}{2}$. Therefore, substituting $p_3 = \frac{5}{2} - p_1 - p_2$ into $f_2 + f_3$ and then factoring we have

$$\begin{aligned} (f_2 + f_3)(p_1, p_2, \tfrac{5}{2} - p_1 - p_2) \\ &= -16 + 28p_1 - 12p_1^2 + 28p_2 - 40p_1p_2 + 12p_1^2p_2 - 12p_2^2 + 12p_1p_2^2 \\ &= 4(1 - p_1)(1 - p_2)(3p_1 + 3p_2 - 4). \end{aligned}$$

However, if $\delta = 2$ then $p_1 + p_2 = \frac{5}{2} - p_3 \geq \frac{3}{2}$ and thus $3p_1 + 3p_2 - 4 \geq \frac{9}{2} - 4 = \frac{1}{2}$. From this, the claim in (12) follows.

To show (13), note that direct computation of derivatives yields

$$\begin{aligned} \frac{\partial(f_2 + f_3)}{\partial p_1} &= -12 + 14p_2 + 12p_3 - 12p_2p_3 = 2p_2 - 12(1 - p_2)(1 - p_3), \\ \frac{\partial(f_2 + f_3)}{\partial p_2} &= -12 + 14p_1 + 12p_3 - 12p_1p_3 = 2p_1 - 12(1 - p_1)(1 - p_3), \\ \frac{\partial(f_2 + f_3)}{\partial p_3} &= -10 + 12p_1 + 12p_2 - 12p_1p_2 = 2 - 12(1 - p_1)(1 - p_2). \end{aligned}$$

For the partial derivative with respect to p_1 , $\delta > 2$ implies $p_3 > \frac{3}{2} - p_2$ so that

$$(1 - p_2)(1 - p_3) < (1 - p_2)(p_2 - 1/2) \leq \frac{1}{16}.$$

Also, since $\delta > 2$ implies $p_2 > \frac{1}{2}$, we have

$$\frac{\partial(f_2 + f_3)}{\partial p_1} > 2\left(\frac{1}{2}\right) - 12\left(\frac{1}{16}\right) = \frac{1}{4} > 0.$$

Similar arguments show that $\partial(f_2 + f_3)/\partial p_2 > \frac{1}{4}$ and $\partial(f_2 + f_3)/\partial p_3 > \frac{5}{4}$ when $\delta > 2$. This completes the proof of (13) and thus also the proof of the lemma. \square

Using Lemma 14, it follows that we can obtain upper and lower bounds on $V_{3,\vec{p}}$ by using the simple bounds $0 \leq \pi(0) \leq 1$; that is,

$$\frac{f_1}{f_2} \leq V_{3,\vec{p}} \leq \frac{f_1}{f_2 + f_3}.$$

However, we can get improved upper bounds on $\pi(0)$ by using the fact that π is not just a probability distribution but also a stationary distribution for the Markov chain $\{Z_n\}_{n \geq 0}$.

Lemma 15. *For an excited random walk with $M = 3$ cookies of strengths $\vec{p} = (p_1, p_2, p_3)$,*

$$\frac{c \cdot p_1 p_2}{b \cdot (1 - p_1) + a \cdot p_1 p_2} \leq \pi(0) \leq c \left(\frac{b \cdot (1 - p_1) p_2}{1 - ((1 - p_1) p_2 p_3 + p_1 (1 - p_2) p_3)} + a \right)^{-1},$$

where a, b , and c are defined in Corollary 12.

Proof. Since π is the stationary distribution of a Markov chain with transition probability matrix $P = (p(i, j))_{i, j \geq 0}$, we know that the (infinite) matrix equation $\pi = \pi P$ holds. That is,

$$\pi(i) = \sum_{k=0}^{\infty} \pi(k) p(k, i) \quad \text{for any } i \geq 0.$$

If we drop all but the first two terms in the sum on the right we then obtain the inequality

$$\pi(i) \geq \pi(0) p(0, i) + \pi(1) p(1, i), \quad (14)$$

where $p(i, j)$ is the transition probability from state i to state j in the backward branching process. For a lower bound on $\pi(0)$ we use $i = 0$ in (14) and then Corollary 12 to get

$$\begin{aligned} \pi(0) &\geq p(0, 0) \pi(0) + p(1, 0) \pi(1) \\ &= p(0, 0) \pi(0) + p(1, 0) \frac{c - a\pi(0)}{b}. \end{aligned}$$

Then, solving for $\pi(0)$ and using the formulas for the transition probabilities yields the lower bound

$$\pi(0) \geq \frac{c \cdot p(1, 0)}{b \cdot (1 - p(0, 0)) + a \cdot p(1, 0)} = \frac{c \cdot p_1 p_2}{b \cdot (1 - p_1) + a \cdot p_1 p_2}. \quad (15)$$

For an upper bound we repeat the same process, this time using $i = 1$ in (14) and applying Corollary 12 to get

$$\frac{c - a\pi(0)}{b} \geq \pi(0) p(0, 1) + \left(\frac{c - a\pi(0)}{b} \right) p(1, 1).$$

Solving this for $\pi(0)$ and then using the formulas for the transition probabilities yields the upper bound

$$\pi(0) \leq c \left(\frac{b \cdot p(0, 1)}{1 - p(1, 1)} + a \right)^{-1} = c \left(\frac{b \cdot (1 - p_1)}{1 - ((1 - p_1)p_2 p_3 + p_1(1 - p_2)p_3)} + a \right)^{-1}. \quad (16)$$

This completes the proof. \square

By applying Lemmas 14 and 15 to Theorem 13, we can obtain explicit upper and lower bounds on the speed of excited random walks with $M = 3$ cookies. The upper/lower bounds are obtained by substituting the respective upper/lower bounds for $\pi(0)$ in Lemma 15 into the formula for the speed in (11). In the special case of $p_1 = p_2 = p_3 > \frac{5}{6}$, this gives the following explicit formulas for upper and lower bounds on the speed:

$$\begin{aligned} \frac{(6p - 5)(p^2 - 2p - 1)}{24p^4 - 42p^3 - 3p^2 + 28p - 9} &\leq V_{3, (p, p, p)}, \\ V_{3, (p, p, p)} &\leq \frac{(6p - 5)(2p^4 - 7p^3 + 5p^2 + p - 3)}{48p^6 - 156p^5 + 180p^4 - 61p^3 - 53p^2 + 51p - 11}. \end{aligned} \quad (17)$$

As is seen in Figure 2, these upper and lower bounds are remarkably close together. In fact, using NMaxValue and NArgMax (Mathematica's numerical optimization functions) one sees that the maximum difference between the upper and lower bounds is at most 0.010326 and is obtained approximately at $p = 0.86649$.

In the general case with $M = 3$ cookies, the upper and lower bounds are again explicit rational functions in (p_1, p_2, p_3) , but these rational functions are extremely long and so we leave it to the interested reader to compute these upper bounds explicitly (with the aid of Mathematica or some other computer algebra software). We note, however, that even in this more general case the upper and lower bounds are remarkably close together. Indeed, again using Mathematica's NMaxValue and NArgMax functions we obtain that the upper and lower bounds differ by at most 0.0194564 and that this maximum is obtained at approximately $\vec{p} = (0.913811, 0.666396, 1)$.

5. Conclusion

Basdevant and Singh showed that the speed of an excited random walk with M cookies per site can be expressed in terms of the expected value of the stationary distribution π of a certain Markov chain on \mathbb{Z}_+ . By using some recursions on the probability-generating function of π that were obtained by Basdevant and Singh, we were able to show that for any fixed values of the parameters p_1, p_2, \dots, p_M , the speed can be expressed as an explicit function of only the $M - 2$ unknown values $\pi(0), \pi(1), \dots, \pi(M - 3)$. In the case of $M = 3$ there is only one unknown parameter, $\pi(0)$, and we can therefore obtain bounds on the speed by obtaining

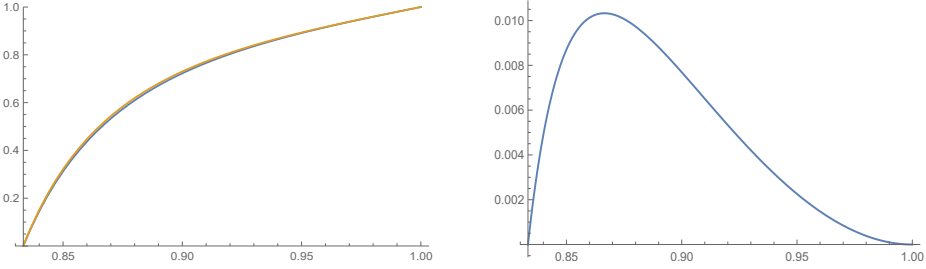


Figure 2. On the left is a plot of the upper and lower bounds for $V_{3,(p,p,p)}$ given in (17). The upper and lower bounds are so close as to be nearly indistinguishable, and so on the right we plot the difference between the upper and lower bounds.

explicit bounds on $\pi(0)$. The bounds we obtain in the case $M = 3$ are very close together, but an exact computation of the speed is at this point still out of reach.

We conclude this paper by stating some remaining open questions related to the results in this paper:

- (1) Can one implement the methods developed in this paper to obtain explicit upper and lower bounds on the speed $V_{M,\vec{p}}$ when $M \geq 4$? The main difficulty here will be that instead of optimizing a function of one variable over an interval, one will need to find the minimum and maximum of a function of $M - 2$ variables over an $(M - 2)$ -dimensional region.
- (2) For any fixed M , is the function $(p_1, p_2, \dots, p_M) \mapsto V_{M,(p_1,p_2,\dots,p_M)}$ differentiable in the region where $\delta = \sum_{j=1}^M (2p_j - 1) > 2$? It was shown in [Basdevant and Singh 2008a] for critical $\vec{p} = (p_1, p_2, \dots, p_M)$ (that is, where $\delta = 2$) that the speed function $\vec{p} \mapsto V_{M,\vec{p}}$ has a positive “right derivative” (that is, the directional derivative is positive in all directions \vec{u} pointing toward the interior of the region where $\delta > 2$). For instance, this implies $p \mapsto V_{3,(p,p,p)}$ has a positive right derivative at $p = \frac{5}{6}$. Since the explicit upper and lower bounds in (17) have the same derivative at $p = 1$, our results show that $p \mapsto V_{3,(p,p,p)}$ is differentiable at $p = 1$ (with derivative equal to 2). It remains open, however, to show that $V_{3,(p,p,p)}$ is differentiable in $(\frac{5}{6}, 1)$.

Appendix: Proof of (9)

We will now give a proof that $\mathbb{E}_k[Z_1] = k + 1 - \delta$ for all $k \geq M - 1$.

Proof. We will compute $\mathbb{E}_k[Z_1]$ by conditioning on $S_M = \sum_{j=1}^M \xi_j$ (the number of successes in the first M Bernoulli trials):

$$\mathbb{E}_k[Z_1] = \sum_{i=0}^M \mathbb{P}(S_M=i) \mathbb{E}[Z_1 \mid Z_0=k \text{ and } S_M=i]. \quad (18)$$

Recall when $Z_0 = k$ that Z_1 is the number of “failures” before the $(k+1)$ -th “success” in the sequence of Bernoulli trials. Given that $S_M = i$ we know that there are i successes and $M - i$ failures in the first M trials, and thus Z_1 is $M - i$ plus the number of failures before the $(k+1-i)$ -th success in a sequence of Bernoulli($\frac{1}{2}$) trials. Since the number of failures before the $(k+1-i)$ -th success is a NegativeBinomial($k+1-i, \frac{1}{2}$) random variable which has mean $k+1-i$, we can therefore conclude that

$$\mathbb{E}[Z_1 \mid Z_0=k \text{ and } S_M=i] = M - i + (k+1-i) = M + k + 1 - 2i.$$

Plugging this into (18) we obtain

$$\begin{aligned} \mathbb{E}_k[Z_1] &= \sum_{i=0}^M \mathbb{P}(S_M=i) \cdot (M + k + 1 - 2i) = M + k + 1 - 2 \sum_{i=0}^M i \cdot \mathbb{P}(S_M=i) \\ &= M + k + 1 - 2\mathbb{E}[S_M] = M + k + 1 - 2 \sum_{j=1}^M \mathbb{E}[\xi_j] \\ &= M + k + 1 - 2 \sum_{j=1}^M p_j = (k+1) - \left(\sum_{j=1}^M 2p_j - 1 \right) = k + 1 - \delta. \quad \square \end{aligned}$$

Acknowledgement

This research was conducted during the 2016 Purdue Research in Mathematics Experience (PRiME) undergraduate math REU. All of the participants are grateful for the support of PRiME provided by NSF grant DMS-1560394 and by the Mathematics Department at Purdue University.

References

- [Basdevant and Singh 2008a] A.-L. Basdevant and A. Singh, “On the speed of a cookie random walk”, *Probab. Theory Related Fields* **141**:3-4 (2008), 625–645. MR Zbl
- [Basdevant and Singh 2008b] A.-L. Basdevant and A. Singh, “Rate of growth of a transient cookie random walk”, *Electron. J. Probab.* **13** (2008), 811–851. MR Zbl
- [Benjamini and Wilson 2003] I. Benjamini and D. B. Wilson, “Excited random walk”, *Electron. Comm. Probab.* **8** (2003), 86–92. MR Zbl
- [Kosygina and Zerner 2008] E. Kosygina and M. P. W. Zerner, “Positively and negatively excited random walks on integers, with branching processes”, *Electron. J. Probab.* **13** (2008), 1952–1979. MR Zbl
- [Pólya 1921] G. Pólya, “Über eine Aufgabe der Wahrscheinlichkeitsrechnung betreffend die Irrfahrt im Straßennetz”, *Math. Ann.* **84**:1-2 (1921), 149–160. MR Zbl
- [Zerner 2005] M. P. W. Zerner, “Multi-excited random walks on integers”, *Probab. Theory Related Fields* **133**:1 (2005), 98–122. MR Zbl

eebossen@gmail.com	<i>Eastern Illinois University, Charleston, IL, United States</i>
Current address:	<i>Department of Mathematics, University of Illinois at Urbana-Champaign, Champaign, IL, United States</i>
bkidd@tamu.edu	<i>Purdue University, West Lafayette, IN, United States</i>
Current address:	<i>Department of Statistics, Texas A&M University, College Station, TX, United States</i>
levin453@umn.edu	olevin2@wisc.edu <i>University of Minnesota, Minneapolis, MN, United States</i>
peterson@purdue.edu	<i>Department of Mathematics, Purdue University, West Lafayette, IN, United States</i>
smit5jb@mail.uc.edu	<i>Franklin College, Franklin, IN, United States</i>
Current address:	<i>Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH, United States</i>
kevin@ttic.edu	<i>University of California, Los Angeles, CA, United States</i>
Current address:	<i>Toyota Technological Institute at Chicago, Chicago, IL, United States</i>

Classifying linear operators over the octonions

Alex Putnam and Tevian Dray

(Communicated by Jim Hoste)

We classify linear operators over the octonions and relate them to linear equations with octonionic coefficients and octonionic variables. Along the way, we also classify linear operators over the quaternions, and show how to relate quaternionic and octonionic operators to real matrices. In each case, we construct an explicit basis of linear operators that maps to the canonical (real) matrix basis; in contrast to the complex case, these maps are surjective. Since higher-order polynomials can be reduced to compositions of linear operators, our construction implies that the ring of polynomials in one variable over the octonions is isomorphic to the product of eight copies of the ring of real polynomials in eight variables.

1. Introduction

The simplest equations are linear and homogeneous; think $y = mx$. However, even linear equations of this form become complicated over number systems other than the reals. What would happen if $mx \neq xm$, or $m(nx) \neq (mn)x$? To address such questions, we analyze here multiplicative operators like mx over the four division algebras, namely the familiar real (\mathbb{R}) and complex (\mathbb{C}) numbers, and the less familiar quaternions (\mathbb{H}), which are not commutative, and octonions (\mathbb{O}), which are neither commutative nor associative.

In the real and complex cases, it is straightforward to rewrite such operators as real matrices. As explained in Section 2, we can generate all such matrices over the reals, but not over the complexes. However, it is initially somewhat surprising to discover that in the remaining cases we can again generate all such matrices, as discussed in Sections 2 and 3. Finally, we discuss some consequences of our work in Section 4, including the immediate generalization to higher-order polynomials.

So far as we are aware, there has not been much previous investigation of octonionic polynomials, linear or otherwise. Serôdio [2007; 2010] considered polynomials with coefficients in \mathbb{O} , but only for real variables. Rodríguez-Ordóñez [2010] classified products of linear equations over \mathbb{O} , and Datta and Nag [1987]

MSC2010: primary 17A35; secondary 15A06.

Keywords: octonions, quaternions, division algebras, linear operators, linear equations.

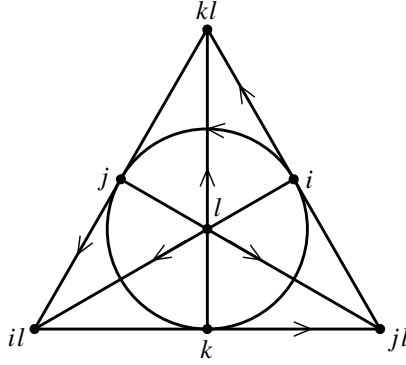


Figure 1. Octonionic multiplication in the Fano plane. Each of the seven oriented lines represents a quaternionic subalgebra; products of two elements on such a line yield \pm the third element, with the sign determined by the arrows.

analyzed the topology of the roots of (some) polynomials over \mathbb{O} . In this work, we provide a classification of all linear equations over \mathbb{O} , and discuss its consequences for polynomials.

Complex numbers can be thought of as a pair of real numbers, the real and imaginary parts; thus, $\mathbb{C} = \mathbb{R} \oplus \mathbb{R}i$, so that $\mathbb{C} \cong \mathbb{R}^2$ as a vector space. In addition, \mathbb{C} admits a product, defined by $i^2 = -1$. Similarly, the quaternions satisfy $\mathbb{H} = \mathbb{C} \oplus \mathbb{C}j$, with multiplication defined by

$$i^2 = j^2 = -1, \quad ji = -ij, \quad (1)$$

from which it follows by associativity that $k = ij$ also satisfies $k^2 = -1$. Multiplication of imaginary quaternions is much like the cross product, and in fact predates it historically. Finally, the octonions (see, e.g., [Dray and Manogue 2015]) satisfy $\mathbb{O} = \mathbb{H} + \mathbb{H}\ell$, where $\ell^2 = -1$; the complete multiplication table can be represented via the oriented Fano plane, as shown in Figure 1. It is easy to check that the octonions are not associative; for instance, $(ij)\ell = k\ell = -i(j\ell)$.

Each of the number systems $\mathbb{K} = \mathbb{R}, \mathbb{C}, \mathbb{H}, \mathbb{O}$ is a *composition algebra*, admitting the operation of conjugation,

$$\bar{x} = 2 \operatorname{Re}(x) - x \quad (2)$$

and an inner product

$$|x|^2 = x\bar{x} \quad (3)$$

satisfying

$$|xy| = |x||y|. \quad (4)$$

Each \mathbb{K} is also a *division algebra*, that is, a vector space on which a compatible multiplication is defined, and in which all nonzero elements are invertible. Explicitly,

the multiplicative inverse of $0 \neq x \in \mathbb{K}$ is given by

$$x^{-1} = \frac{\bar{x}}{|x|^2}. \quad (5)$$

The Hurwitz theorem [1922] asserts that these four algebras are the only (positive-definite) composition algebras over the reals.

2. Real, complex and quaternionic linear operators

We now explore certain linear operators over each division algebra $\mathbb{K} = \mathbb{R}, \mathbb{C}, \mathbb{H}, \mathbb{O}$. Let $\mathcal{L}(\mathbb{K})$ be the set of all multiplicative linear operators from \mathbb{K} to \mathbb{K} , that is, all real-linear operators from \mathbb{K} to \mathbb{K} that can be realized using multiplication (and addition) within \mathbb{K} . More precisely, $\mathcal{L}(\mathbb{K})$ is the group generated by the left and right translations

$$\begin{aligned} m_L : \mathbb{K} &\rightarrow \mathbb{K}, & m_R : \mathbb{K} &\rightarrow \mathbb{K}, \\ x &\mapsto mx, & x &\mapsto xm \end{aligned} \quad (6)$$

for $m \in \mathbb{K}$. These translations are linear over \mathbb{R} by distributivity and the commutativity and associativity of elements of \mathbb{R} in \mathbb{K} . That is,

$$m_L(x + ry) = m_L(x) + rm_L(y) \quad (7)$$

for $x, y \in \mathbb{K}$ and $r \in \mathbb{R}$, and similarly for m_R . Thus, $\mathcal{L}(\mathbb{K})$ must have a matrix representation

$$\pi_{\mathbb{K}} : \mathcal{L}(\mathbb{K}) \rightarrow M_{\dim(\mathbb{K})}(\mathbb{R}), \quad (8)$$

where $M_k(\mathbb{R})$ denotes the set of $k \times k$ real matrices.

Since elements of \mathbb{R} associate and commute, any linear operator over \mathbb{R} can be expressed in the form

$$x \mapsto mx, \quad (9)$$

where $m, x \in \mathbb{R}$. For reasons that will become obvious as we lose commutativity and associativity, we will refer to this linear operator as “ mx ”; that is, we use the image of the operator acting on a “place-holder” variable, x , (also) as the name of the operator. In this sense, $mx \in \mathcal{L}(\mathbb{R})$. Since elements of $M_1(\mathbb{R})$ are matrices of the form $M = (m)$, we have the natural definition

$$\pi_{\mathbb{R}}(mx) = (m). \quad (10)$$

Thus, the set of linear operators on \mathbb{R} is equivalent to the set of real 1×1 matrices, and $\pi_{\mathbb{R}}$ is the trivial map.

Complex numbers also commute and associate, so linear operators over \mathbb{C} can again be expressed in the form (9), where now $m, x \in \mathbb{C}$. Separating each complex

number into real and imaginary parts, e.g., $x = x_1 + x_2 i$, and mapping \mathbb{C} into \mathbb{R}^2 in the natural way,

$$x_1 + x_2 i \mapsto \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad (11)$$

and noting that

$$(m_1 + m_2 i)(x_1 + x_2 i) = m_1 x_1 - m_2 x_2 + (m_1 x_2 + m_2 x_1) i \quad (12)$$

brings the linear operator to the form

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} m_1 x_1 - m_2 x_2 \\ m_1 x_2 + m_2 x_1 \end{pmatrix} \quad (13)$$

so that

$$\pi_{\mathbb{C}}(mx) = \begin{pmatrix} m_1 & -m_2 \\ m_2 & m_1 \end{pmatrix}. \quad (14)$$

Thus, our set of linear operators over \mathbb{C} has only two degrees of freedom, namely the real and imaginary parts of the coefficient m . On the other hand, the set $M_2(\mathbb{R})$ is a vector space with four (real) degrees of freedom. Therefore, there are real 2×2 matrices that cannot be expressed as a (complex-)linear operator over \mathbb{C} . We have therefore shown that $\pi_{\mathbb{C}} : \mathcal{L}(\mathbb{C}) \rightarrow M_2(\mathbb{R})$ cannot be a surjective map. Some simple examples of real 2×2 matrices that are not in the image of $\pi_{\mathbb{C}}$ are projections and complex conjugation.

If we look to the quaternions, we finally start to find more complicated linear operators. Since the quaternions do not commute, all multiplicative linear operators over \mathbb{H} are sums of terms of the form

$$x \mapsto p x q, \quad (15)$$

where $p, q \in \mathbb{H}$. Since we can expand each quaternion p, q , in terms of a basis $\{1, i, j, k\}$ and then distribute over the expanded coefficients, we see that every linear operator over \mathbb{H} can be expressed as a linear combination of terms of the form

$$x \mapsto e_m x e_n \quad (16)$$

for distinct combinations $e_m, e_n \in \{1, i, j, k\}$. Therefore, we only need to consider coefficients that are basis elements of \mathbb{H} . Expanding each quaternion with respect to our basis, e.g., $x = x_1 + x_2 i + x_3 j + x_4 k$, and mapping \mathbb{H} into \mathbb{R}^4 by analogy with (11) leads immediately to, for instance,

$$\pi_{\mathbb{H}}(px) = \begin{pmatrix} p_1 & -p_2 & -p_3 & -p_4 \\ p_2 & p_1 & -p_4 & p_3 \\ p_3 & p_4 & p_1 & -p_2 \\ p_4 & -p_3 & p_2 & p_1 \end{pmatrix}. \quad (17)$$

Although this particular operator has only four (real) degrees of freedom, it is now easy to verify [Putnam 2017] that $\pi_{\mathbb{H}}$ maps the set of all 16 operators $\{e_m x e_n\}$ to a basis of $M_4(\mathbb{R})$.¹ Thus, $\pi_{\mathbb{H}}$ must be a bijection between the linear operators over \mathbb{H} and $M_4(\mathbb{R})$.

An explicit pairing of each elementary matrix in $M_4(\mathbb{R})$ with a corresponding multiplicative linear operator over \mathbb{H} is given in [Putnam 2017]. Intriguing examples are

$$x - ixi - jxj - kxk = 4 \operatorname{Re}(x), \quad (18)$$

$$x + ixi + jxj + kxk = -2\bar{x}, \quad (19)$$

each of which can be verified (or discovered!) by applying $\pi_{\mathbb{H}}$. Conjugation is a linear map over \mathbb{H} !

3. Octonionic linear operators

We are now ready to look at $\mathcal{L}(\mathbb{O})$, the multiplicative linear operators over \mathbb{O} . If we consider operators of the form (15) with $p, q, x \in \mathbb{O}$, then, because \mathbb{O} is not associative, we are really considering two different operators, one of the form $x \mapsto (px)q$, and the other of the form $x \mapsto p(xq)$, unless p, q are in a complex subalgebra of \mathbb{O} (since the octonions are alternative). We can, however, continue to nest more coefficients outside of these two terms. Just as before, because we can distribute over the expanded form of $x \in \mathbb{O}$, we only need to consider linear operators with basis elements as coefficients. Mapping \mathbb{O} into \mathbb{R}^8 again gives a natural definition of, for example,

$$\pi_{\mathbb{O}}(px) = \begin{pmatrix} p_1 & -p_2 & -p_3 & -p_4 & -p_5 & -p_6 & -p_7 & -p_8 \\ p_2 & p_1 & -p_4 & p_3 & -p_6 & p_5 & p_8 & -p_7 \\ p_3 & p_4 & p_1 & -p_2 & p_7 & p_8 & -p_5 & -p_6 \\ p_4 & -p_3 & p_2 & p_1 & p_8 & -p_7 & p_6 & -p_5 \\ p_5 & p_6 & -p_7 & -p_8 & p_1 & -p_2 & p_3 & p_4 \\ p_6 & -p_5 & -p_8 & p_7 & p_2 & p_1 & -p_4 & p_3 \\ p_7 & -p_8 & p_5 & -p_6 & -p_3 & p_4 & p_1 & p_2 \\ p_8 & p_7 & p_6 & p_5 & -p_4 & -p_3 & -p_2 & p_1 \end{pmatrix}. \quad (20)$$

Because we can nest the coefficients of x , we need to count how many nestings we are likely to need to show whether $\pi_{\mathbb{O}}$ is surjective. If we consider operators of

¹Alternatively, one can verify by direct computation that the matrix $\pi_{\mathbb{H}}(\sum a_{m,n} e_m x e_n)$ is

$$\begin{pmatrix} a_{1,1} - a_{2,2} - a_{3,3} - a_{4,4} & -a_{1,2} - a_{2,1} + a_{3,4} - a_{4,3} & -a_{1,3} - a_{2,4} - a_{3,1} + a_{4,2} & -a_{1,4} + a_{2,3} - a_{3,2} - a_{4,1} \\ a_{1,2} + a_{2,1} + a_{3,4} - a_{4,3} & a_{1,1} - a_{2,2} + a_{3,3} + a_{4,4} & a_{1,4} - a_{2,3} - a_{3,2} - a_{4,1} & -a_{1,3} - a_{2,4} + a_{3,1} - a_{4,2} \\ a_{1,3} - a_{2,4} + a_{3,1} + a_{4,2} & -a_{1,4} - a_{2,3} - a_{3,2} + a_{4,1} & a_{1,1} + a_{2,2} - a_{3,3} + a_{4,4} & a_{1,2} - a_{2,1} - a_{3,4} - a_{4,3} \\ a_{1,4} + a_{2,3} - a_{3,2} + a_{4,1} & a_{1,3} - a_{2,4} - a_{3,1} - a_{4,2} & -a_{1,2} + a_{2,1} - a_{3,4} - a_{4,3} & a_{1,1} + a_{2,2} + a_{3,3} - a_{4,4} \end{pmatrix}$$

and then check that the 16 degrees of freedom (the matrix coefficients of $a_{m,n}$) are independent.

the form $x \mapsto pxq$ then at first sight we have $8^2 = 64$ such operators. However, the lack of associativity means that there are $2\binom{7}{2} = 42$ cases where we must count both possible orders of multiplication, resulting in $64 + 42 = 106$ operators with distinct orderings of coefficients. Since $\dim M_8(\mathbb{R}) = 64$, these 106 operators cannot be linearly independent, but it is not obvious whether they span $\mathcal{L}(\mathbb{O})$. Since right multiplication can be expressed in terms of (nested) left multiplication [Conway and Smith 2003], we will instead consider operators with coefficients only on the left. We have the identity operator, $x \mapsto x$, and seven operators of the form $x \mapsto e_n x$ with $e_n \in \{i, j, k, i\ell, j\ell, k\ell, \ell\}$. If we consider one nested coefficient, then we have the form $x \mapsto e_n(e_m x)$, again with $e_n \neq 1 \neq e_m$ and $\binom{7}{2} = 21$ new operators. These singly nested products were shown in [Manogue and Schray 1993] to generate the orthogonal group $\mathrm{SO}(7)$.

Next, we consider two nestings, which yields $\binom{7}{3} = 35$ more operators. Amazingly, this process gives us a total of $1 + 7 + 21 + 35 = 64$ distinct (representations of) operators in $\mathcal{L}(\mathbb{O})$! It was shown in [Putnam 2017] that these 64 linear operators are in fact linearly independent; an explicit pairing with the canonical basis of $M_8(\mathbb{R})$ was also given. Thus, $\pi_{\mathbb{O}}$ is surjective, and doubly nested representations are precisely enough to express all elements $\mathcal{L}(\mathbb{O})$.

In the previous cases, we were only able to construct linear operators for $\dim(\mathbb{K})^2$ different combinations of coefficients of basis elements, because each underlying space was associative. In \mathbb{O} , we can construct the same linear operators with different combinations of coefficients of basis elements. So, operators that appear to be different may have the same image $\pi_{\mathbb{O}}$, and thus in fact correspond to different representations of the same element of $\mathcal{L}(\mathbb{K})$.² It is now straightforward to show that $\mathcal{L}(\mathbb{O})$ forms a group under operator composition, and that the map $\pi_{\mathbb{O}} : \mathcal{L}(\mathbb{O}) \rightarrow M_8(\mathbb{R})$ is a bijection. In particular, it then follows that right multiplication can be expressed in terms of left multiplication, thus verifying the result of [Conway and Smith 2003], and this can be done explicitly by finding a linear combination of the basis given in [Putnam 2017] that yields the same matrix.

Since $\pi_{\mathbb{O}}$ is a surjective map, there must exist elements $f_{n,m} \in \mathcal{L}(\mathbb{O})$ such that $f_{n,m}(x) = x_n e_m$ for $1 \leq n \leq 8$ and $e_m \in \{1, i, j, k, i\ell, j\ell, k\ell, \ell\}$. Some other intriguing elements of $\mathcal{L}(\mathbb{O})$ are given by $x - i(j(kx))$, which projects out the quaternionic part of x , and $x - ixi$, which projects out the complex part of x . It is a useful exercise to work out a representation of the latter operator in terms of nested left multiplication! Again, these assertions can be verified or discovered by applying $\pi_{\mathbb{O}}$.

²An alternative treatment, as in [Putnam 2017], would regard $\mathcal{L}(\mathbb{K})$ as being freely generated by left and right translations, then define an equivalence relation $L \sim M$ on elements $L, M \in \mathcal{L}(\mathbb{K})$ if $\pi_{\mathbb{O}}(L) = \pi_{\mathbb{O}}(M)$. The relation \sim is clearly an equivalence relation, since it is defined by equality of matrices, and what we here call $\mathcal{L}(\mathbb{O})$ would instead be the quotient $\mathcal{L}(\mathbb{O})/\sim$.

4. Conclusion

We have shown that the lack of commutativity of the quaternions, and the lack of associativity of the octonions, conspire to provide just enough degrees of freedom that multiplicative linear operators do indeed generate all real-linear maps in those cases — despite the fact that they do not do so in the complex case. In the quaternionic case, the extra degrees of freedom manifest themselves when considering two-sided operators, whereas in the octonionic case it is the nested nature of iterated multiplication that generates the necessary degrees of freedom. Along the way, we have verified the assertion stated without proof in Section 3 that right multiplication can be expressed in terms of nested left multiplication.

In the octonionic case, we have further shown that it takes precisely three iterated products to generate all 64 independent real-linear maps, noting that $\binom{7}{0} + \binom{7}{1} + \binom{7}{2} + \binom{7}{3} = 64$. This result has an intriguing application to the Clifford algebra $\text{Cl}(6)$, which can be represented precisely as the 64-dimensional matrix algebra $M_8(\mathbb{R})$. As has been noted by Furey [2014], it is therefore possible to represent $\text{Cl}(6)$ entirely in terms of octonionic multiplication, with possible applications to particle physics; see, e.g., [Dray and Manogue 2015].

Having classified multiplicative linear operators over \mathbb{O} , we could consider higher-degree terms, that is, octonionic polynomials. By the distributive law, and because real numbers commute and associate with octonions, we can expand each such term (both coefficients and variables) with respect to a basis. Just as there are $8 = \binom{8}{1}$ (real-)independent components of x , and hence $8 \times 8 = 64$ independent linear operators on \mathbb{O} , there are similarly $\binom{8}{2} + 8 = 36$ quadratic “components” of x^2 , where the last “8” counts coefficients that are squared. Thus, the most general quadratic operator maps x to a linear combination of the $8 \times 36 = 288$ terms $x_m x_n e_p$, where $1 \leq m \leq n \leq 8$ and $1 \leq p \leq 8$. Furthermore, we can realize each such operator (in multiple ways) as a composition of the linear operators $f_{m,n}$, and hence in terms of octonionic multiplication. A similar process can be applied to higher-order terms. It is obvious that any polynomial over \mathbb{O} can be reinterpreted as eight real polynomials in eight variables; our construction shows that the converse is also true, so that $\mathbb{O}[x] \cong (\mathbb{R}[x_1, \dots, x_8])^8$.

Acknowledgments

This work is based on a paper submitted by Putnam in partial fulfillment of the degree requirements for his M.S. in Mathematics at Oregon State University [Putnam 2017].

References

[Conway and Smith 2003] J. H. Conway and D. A. Smith, *On quaternions and octonions: their geometry, arithmetic, and symmetry*, A K Peters, Natick, MA, 2003. MR Zbl

- [Datta and Nag 1987] B. Datta and S. Nag, “Zero-sets of quaternionic and octonionic analytic functions with central coefficients”, *Bull. London Math. Soc.* **19**:4 (1987), 329–336. MR Zbl
- [Dray and Manogue 2015] T. Dray and C. A. Manogue, *The geometry of the octonions*, World Scientific, Hackensack, NJ, 2015. MR Zbl
- [Furey 2014] C. Furey, “Generations: three prints, in colour”, *J. High Energy Phys.* **2014**:10 (2014), art. id. 046.
- [Hurwitz 1922] A. Hurwitz, “Über die Komposition der quadratischen Formen”, *Math. Ann.* **88**:1-2 (1922), 1–25. MR Zbl
- [Manogue and Schray 1993] C. A. Manogue and J. Schray, “Finite Lorentz transformations, automorphisms, and division algebras”, *J. Math. Phys.* **34**:8 (1993), 3746–3767. MR Zbl
- [Putnam 2017] A. Putnam, *Classifying octonionic-linear operators*, master’s thesis, Oregon State University, 2017, available at <http://ir.library.oregonstate.edu/xmlui/handle/1957/61707>.
- [Rodríguez-Ordóñez 2010] H. Rodríguez-Ordóñez, “Homotopy classification of bilinear maps related to octonion polynomial multiplications”, *Linear Algebra Appl.* **432**:12 (2010), 3117–3131. MR Zbl
- [Serôdio 2007] R. Serôdio, “On octonionic polynomials”, *Adv. Appl. Clifford Algebr.* **17**:2 (2007), 245–258. MR Zbl
- [Serôdio 2010] R. Serôdio, “Construction of octonionic polynomials”, *Adv. Appl. Clifford Algebr.* **20**:1 (2010), 155–178. MR Zbl

Received: 2017-07-23 Revised: 2018-01-04 Accepted: 2018-02-14

putnama@math.oregonstate.edu *Department of Mathematics, Oregon State University,
Corvallis, OR, United States*

tevian@math.oregonstate.edu *Department of Mathematics, Oregon State University,
Corvallis, OR, United States*

Spectrum of the Kohn Laplacian on the Rossi sphere

Tawfik Abbas, Madelyne M. Brown,
Allison Ramasami and Yunus E. Zeytuncu

(Communicated by Stephan Garcia)

We study the spectrum of the Kohn Laplacian \square_b^t on the Rossi example $(\mathbb{S}^3, \mathcal{L}_t)$. In particular we show that 0 is in the essential spectrum of \square_b^t , which yields another proof of the global nonembeddability of the Rossi example.

1. Introduction

General setting. Let $\mathbb{S}^3 = \{(z_1, z_2) \in \mathbb{C}^2 : |z_1|^2 + |z_2|^2 = 1\}$ denote the 3-sphere in \mathbb{C}^2 . The space \mathbb{S}^3 is a real three-dimensional manifold and it can be viewed as an abstract CR manifold when one chooses a specific complex vector field that determines the complex tangent vectors. It is a general question whether an abstract CR manifold can be realized as a manifold in \mathbb{C}^N , for some N , where the complex tangent spaces coincide with the ones induced from the ambient space. One way of addressing this question is studying a second-order differential operator, the so-called Kohn Laplacian, that naturally arises on CR manifolds. Many geometric properties of abstract CR manifolds can be studied by analyzing the properties of this differential operator. In this note we address the embeddability question by studying the spectrum of the Kohn Laplacian on a specific abstract CR manifold. In particular we examine the essential spectrum of the Kohn Laplacian. The essential spectrum of a bounded self-adjoint operator is the subset of the spectrum that contains eigenvalues of infinite multiplicity and the limit points. We refer the readers to [Boggess 1991; Chen and Shaw 2001] for the general theory of CR manifolds and the Kohn Laplacian, and to [Davies 1995] for spectral theory.

MSC2010: primary 32V30; secondary 32V05.

Keywords: Kohn Laplacian, spherical harmonics, global embeddability of CR manifolds.

This work is supported by NSF (DMS-1659203). Zeytuncu is also partially supported by a grant from the Simons Foundation (#353525).

Main problem. Rossi [1965] showed that the CR-manifold $(\mathbb{S}^3, \mathcal{L}_t)$ is not CR-embeddable, where

$$\mathcal{L}_t = \bar{z}_1 \frac{\partial}{\partial z_2} - \bar{z}_2 \frac{\partial}{\partial z_1} + \bar{t} \left(z_1 \frac{\partial}{\partial \bar{z}_2} - z_2 \frac{\partial}{\partial \bar{z}_1} \right),$$

and $|t| < 1$. In the case of strictly pseudoconvex CR-manifolds Boutet de Monvel [1975] proved that if the real dimension of the manifold is at least 5, then it can always be globally CR-embedded into \mathbb{C}^N for some N . Later Burns [1979] approached this problem in the $\bar{\partial}$ context and showed that if the tangential operator $\bar{\partial}_{b,t}$ has closed range and the Szegő projection is bounded, then the CR-manifold is CR-embeddable into \mathbb{C}^N . Then Kohn [1985] showed that CR-embeddability is equivalent to showing that the tangential Cauchy–Riemann operator $\bar{\partial}_{b,t}$ has closed range.

In the setting of the Rossi example, as an application of the closed graph theorem, $\bar{\partial}_{b,t}$ has closed range if and only if the Kohn Laplacian

$$\square_b^t = -\mathcal{L}_t \frac{1 + |t|^2}{(1 - |t|^2)^2} \bar{\mathcal{L}}_t$$

has closed range; see [Burns and Epstein 1990, (0.5)]. Furthermore, the closed range property is equivalent to the positivity of the essential spectrum of \square_b^t ; see [Fu 2005] for similar discussion. In this note we tackle the problem of embeddability, from the perspective of spectral analysis. In particular, we show that 0 is in the essential spectrum of \square_b^t , so the Rossi sphere is not globally CR-embeddable into \mathbb{C}^N . This provides a different approach to the results in [Burns 1979; Kohn 1985].

We start our analysis with the spectrum of \square_b^t . We utilize spherical harmonics to construct finite-dimensional subspaces of $L^2(\mathbb{S}^3)$ such that \square_b^t has tridiagonal matrix representations on these subspaces. We then use these matrices to compute eigenvalues of \square_b^t . We also present numerical results obtained by Mathematica that motivate most of our theoretical results. We then present an upper bound for small eigenvalues and we exploit this bound to find a sequence of eigenvalues that converge to 0.

In addition to particular results in this note, our approach can be adopted to study possible other perturbations of the standard CR-structure on the 3-sphere, such as in [Burns and Epstein 1990]. Furthermore, our approach also leads some information on the growth rate of the eigenvalues and possible connections to finite-type (in the sense of commutators) results similar to the ones in [Fu 2008]. We plan to address these issues in future papers.

2. Analysis of \square_b on $\mathcal{H}_{p,q}(\mathbb{S}^3)$

Spherical harmonics. We start with a quick overview of spherical harmonics; we refer to [Axler et al. 2001] for a detailed discussion. We will state the relevant

theorems on \mathbb{C}^2 and $\mathbb{S}^3 \subseteq \mathbb{C}^2$. A polynomial in \mathbb{C}^2 can be written as

$$p(z, \bar{z}) = \sum_{\alpha, \beta} c_{\alpha, \beta} z^\alpha \bar{z}^\beta,$$

where $z \in \mathbb{C}^2$, each $c_{\alpha, \beta}$ is in \mathbb{C} , and $\alpha, \beta \in \mathbb{N}^2$ are multi-indices. That is, $\alpha = (\alpha_1, \alpha_2)$, $z^\alpha = z_1^{\alpha_1} z_2^{\alpha_2}$, and $|\alpha| = \alpha_1 + \alpha_2$.

We denote the space of all homogeneous polynomials on \mathbb{C}^2 of degree m by $\mathcal{P}_m(\mathbb{C}^2)$, and we let $\mathcal{H}_m(\mathbb{C}^2)$ denote the subspace of $\mathcal{P}_m(\mathbb{C}^2)$ that consists of all harmonic homogeneous polynomials on \mathbb{C}^2 of degree m . We use $\mathcal{P}_m(\mathbb{S}^3)$ and $\mathcal{H}_m(\mathbb{S}^3)$ to denote the restriction of $\mathcal{P}_m(\mathbb{C}^2)$ and $\mathcal{H}_m(\mathbb{C}^2)$ onto \mathbb{S}^3 . We denote the space of complex homogeneous polynomials on \mathbb{C}^2 of bidegree p, q by $\mathcal{P}_{p,q}(\mathbb{C}^2)$, and those polynomials that are homogeneous and harmonic by $\mathcal{H}_{p,q}(\mathbb{C}^2)$. As before, we denote by $\mathcal{P}_{p,q}(\mathbb{S}^3)$ and $\mathcal{H}_{p,q}(\mathbb{S}^3)$ the polynomials of the previous spaces, but restricted to \mathbb{S}^3 . We recall that on \mathbb{C}^2 , the Laplacian is defined as

$$\Delta = 4 \left(\frac{\partial^2}{\partial z_1 \partial \bar{z}_1} + \frac{\partial^2}{\partial z_2 \partial \bar{z}_2} \right).$$

As an example, $z_1 \bar{z}_2 - 2z_2 \bar{z}_1 \in \mathcal{P}_{1,1}(\mathbb{C}^2)$, and $z_1 \bar{z}_2^2 \in \mathcal{H}_{1,2}(\mathbb{C}^2)$. We take our first step by stating the following decomposition result.

Proposition 2.1 [Axler et al. 2001, Theorem 5.12]. $L^2(\mathbb{S}^3) = \bigoplus_{m=0}^{\infty} \mathcal{H}_m(\mathbb{S}^3)$.

The spherical harmonics form an orthogonal basis on \mathbb{S}^3 similar to the Fourier series on the unit circle \mathbb{S}^1 . They are also the eigenfunctions of the Laplacian on \mathbb{S}^3 . The summation above is understood as the orthogonal direct sum of Hilbert spaces. This statement is essential to the spectral analysis of \square_b^t on $L^2(\mathbb{S}^3)$ since it decomposes the infinite-dimensional space $L^2(\mathbb{S}^3)$ into finite-dimensional pieces, which is necessary for obtaining the matrix representation of \square_b^t (a special case of the general spectral theory of compact operators). In order to get such a matrix representation, we need a method for obtaining a basis for $\mathcal{H}_k(\mathbb{S}^3)$. Proposition 2.3 presents a method to do so for $\mathcal{H}_m(\mathbb{C}^2)$ and Proposition 2.5 presents a method for $\mathcal{H}_{p,q}(\mathbb{C}^2)$. The dimension of the matrix representation on a particular $\mathcal{H}_m(\mathbb{S}^3)$ is the dimension of the subspace $\mathcal{H}_m(\mathbb{S}^3)$, which is given below and analogously given for $\mathcal{H}_{p,q}(\mathbb{C}^2)$.

Proposition 2.2 [Axler et al. 2001, Proposition 5.8]. For $k, p, q \geq 2$,

$$\begin{aligned} \dim \mathcal{P}_{p,q}(\mathbb{C}^2) &= (p+1)(q+1), \\ \dim \mathcal{H}_{p,q}(\mathbb{C}^2) &= p+q+1 \\ \dim \mathcal{H}_k(\mathbb{C}^2) &= (k+1)^2. \end{aligned}$$

Now we present a method to obtain explicit bases of spaces of spherical harmonics. These bases play an essential role in explicit calculations in the next section. Here,

K denotes the Kelvin transform,

$$K[g](z) = |z|^{-2} g\left(\frac{z}{|z|^2}\right).$$

For multi-indices $\alpha, \beta \in \mathbb{N}^2$, we denote by D^α and \bar{D}^β the differential operators

$$D^\alpha = \frac{\partial^{|\alpha|}}{(\partial^{\alpha_1} z_1)(\partial^{\alpha_2} z_2)} \quad \text{and} \quad \bar{D}^\beta = \frac{\partial^{|\beta|}}{(\partial^{\beta_1} \bar{z}_1)(\partial^{\beta_2} \bar{z}_2)}.$$

Proposition 2.3 [Axler et al. 2001, Theorem 5.25]. *The set*

$$\{K[D^\alpha |z|^{-2}] : |\alpha| = m \text{ and } \alpha_1 \leq 1\}$$

is a vector space basis for $\mathcal{H}_m(\mathbb{C}^2)$, and the set

$$\{D^\alpha |z|^{-2} : |\alpha| = m \text{ and } \alpha_1 \leq 1\}$$

is a vector space basis for $\mathcal{H}_m(\mathbb{S}^3)$.

Homogeneous polynomials of degree k can be written as the sum of polynomials of bidegree p, q such that $p + q = k$.

Proposition 2.4. $\mathcal{P}_k(\mathbb{C}^2) = \bigoplus_{p+q=k} \mathcal{P}_{p,q}(\mathbb{C}^2)$.

Analogous to the version in Proposition 2.3, we use the following method to construct orthogonal bases for $\mathcal{H}_{p,q}(\mathbb{C}^2)$ and $\mathcal{H}_{p,q}(\mathbb{S}^3)$. The proof pretty much follows the proof of [Axler et al. 2001, Theorem 5.25], with changes from single index to double index.

Proposition 2.5. *The set*

$$\{K[\bar{D}^\alpha D^\beta |z|^{-2}] : |\alpha| = p, |\beta| = q, \alpha_1 = 0 \text{ or } \beta_1 = 0\}$$

is a basis for $\mathcal{H}_{p,q}(\mathbb{C}^2)$, and the set

$$\{\bar{D}^\alpha D^\beta |z|^{-2} : |\alpha| = p, |\beta| = q, \alpha_1 = 0 \text{ or } \beta_1 = 0\}$$

is an orthogonal basis for $\mathcal{H}_{p,q}(\mathbb{S}^3)$.

\square_b on $\mathcal{H}_{p,q}(\mathbb{S}^3)$. Before we study the operator \square_b^t , we first need some background on a simpler operator we call \square_b . It arises from the CR-manifold $(\mathbb{S}^3, \mathcal{L})$, and is defined as

$$\square_b = -\mathcal{L}\bar{\mathcal{L}}.$$

Here, $\mathcal{L} = \mathcal{L}_0 = \bar{z}_1(\partial/\partial z_2) - \bar{z}_2(\partial/\partial z_1)$, the standard $(1, 0)$ vector field from the ambient space. We note that this CR-structure is induced from \mathbb{C}^2 and this manifold is naturally embedded. By the machinery above we can compute the eigenvalues of \square_b ; see also [Folland 1972] for a more general discussion.

Theorem 2.6. *Suppose $f \in \mathcal{H}_{p,q}(\mathbb{S}^3)$. Then*

$$\square_b f = (pq + q)f.$$

Proof. Expanding the definition, we get

$$\begin{aligned} \square_b &= -\left(\bar{z}_2 \frac{\partial}{\partial z_1} - \bar{z}_1 \frac{\partial}{\partial z_2}\right) \left(z_2 \frac{\partial}{\partial \bar{z}_1} - z_1 \frac{\partial}{\partial \bar{z}_2}\right) \\ &= -\bar{z}_2 \frac{\partial}{\partial z_1} \left(z_2 \frac{\partial}{\partial \bar{z}_1} - z_1 \frac{\partial}{\partial \bar{z}_2}\right) + \bar{z}_1 \frac{\partial}{\partial z_2} \left(z_2 \frac{\partial}{\partial \bar{z}_1} - z_1 \frac{\partial}{\partial \bar{z}_2}\right) \\ &= -z_2 \bar{z}_2 \frac{\partial^2}{\partial z_1 \partial \bar{z}_1} + \bar{z}_2 \frac{\partial}{\partial \bar{z}_2} + z_1 \bar{z}_2 \frac{\partial^2}{\partial z_1 \partial \bar{z}_2} - z_1 \bar{z}_1 \frac{\partial^2}{\partial z_2 \partial \bar{z}_2} + \bar{z}_1 \frac{\partial}{\partial \bar{z}_1} + z_2 \bar{z}_1 \frac{\partial^2}{\partial z_2 \partial \bar{z}_1}. \end{aligned}$$

Now, let $f \in \mathcal{H}_{p,q}(\mathbb{S}^3)$. Since f is harmonic, we know that

$$\frac{\partial^2}{\partial z_1 \partial \bar{z}_1} = -\frac{\partial^2}{\partial z_2 \partial \bar{z}_2}.$$

Substituting, we get

$$\square_b = z_2 \bar{z}_2 \frac{\partial^2}{\partial z_2 \partial \bar{z}_2} + \bar{z}_2 \frac{\partial}{\partial \bar{z}_2} + z_1 \bar{z}_2 \frac{\partial^2}{\partial z_1 \partial \bar{z}_2} + z_1 \bar{z}_1 \frac{\partial^2}{\partial z_1 \partial \bar{z}_1} + \bar{z}_1 \frac{\partial}{\partial \bar{z}_1} + z_2 \bar{z}_1 \frac{\partial^2}{\partial z_2 \partial \bar{z}_1}.$$

Since f is a polynomial and \square_b is linear, it suffices to show that if $f = z^\alpha \bar{z}^\beta = z_1^{\alpha_1} z_2^{\alpha_2} \bar{z}_1^{\beta_1} \bar{z}_2^{\beta_2}$, where $\alpha_1 + \alpha_2 = p$ and $\beta_1 + \beta_2 = q$, then the claim holds. Using the expansion above, each derivative simply becomes a multiple of f , and we have

$$\begin{aligned} \square_b f &= (\alpha_2 \beta_2 + \beta_2 + \alpha_1 \beta_2 + \alpha_1 \beta_1 + \beta_1 + \alpha_2 \beta_1) f \\ &= ((\alpha_1 + \alpha_2)(\beta_1 + \beta_2) + (\beta_1 + \beta_2)) f \\ &= (pq + q) f. \end{aligned} \quad \square$$

In a similar manner, we can show that $-\bar{\mathcal{L}}\mathcal{L}f = (pq + p)f$. For \square_b , we actually have $\text{spec}(\square_b) = \{pq + q : p, q \in \mathbb{N}\}$; therefore $0 \notin \text{essspec}(\square_b)$ since it is not an accumulation point of the set above.

3. Experimental results in Mathematica

Using the symbolic computation environment provided by Mathematica, we are able to write a program to streamline our calculations¹. We implement the algorithm provided in Proposition 2.5 to construct the vector space basis of $\mathcal{H}_k(\mathbb{S}^3)$ for a

¹Our code for this and the other symbolic computations described below is available in the online supplement.

specified k . As an example, our code produces the following basis of $\mathcal{H}_3(\mathbb{S}^3)$:

$$\{-6\bar{z}_2^3, -6\bar{z}_1\bar{z}_2^2, -6\bar{z}_1^2\bar{z}_2, -6\bar{z}_1^3, 4z_1\bar{z}_1\bar{z}_2 - 2z_2\bar{z}_2^2, 2z_1\bar{z}_1^2 - 4z_2\bar{z}_1\bar{z}_2, -6z_2\bar{z}_1^2, -6z_1\bar{z}_2^2, 4z_1z_2\bar{z}_1 - 2z_2^2\bar{z}_2, -6z_2^2\bar{z}_1, 2z_1^2\bar{z}_1 - 4z_1z_2\bar{z}_2, -6z_1^2\bar{z}_2, -6z_2^3, -6z_1z_2^2, -6z_1^2z_2, -6z_1^3\}.$$

Now, with the basis for $\mathcal{H}_k(\mathbb{S}^3)$, the matrix representation of \square_b^t on $\mathcal{H}_k(\mathbb{S}^3)$ can be computed for each k . In particular, we use this program to construct the matrix representations for $1 \leq k \leq 12$. For a specific k , the code applies \square_b^t to each basis element of $\mathcal{H}_k(\mathbb{S}^3)$ obtained by the results in the previous sections. Then, using the inner product defined by

$$\langle f, g \rangle = \int_{\mathbb{S}^3} f \bar{g} d\sigma,$$

where σ is the standard surface-area measure, the software computes $\langle \square_b^t f_i, f_j \rangle$, where f_i, f_j are basis vectors for $\mathcal{H}_k(\mathbb{S}^3)$. With these results, Mathematica yields the matrix representation for the imputed value of k . For example, for $k = 3$ the program produces the matrix representation

$$h \begin{pmatrix} 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -6\bar{t} & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6\bar{t} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & -6\bar{t} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & -6\bar{t} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{A} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -2\bar{t} & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{A} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2\bar{t} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{A} & 0 & 0 & 0 & 0 & -2\bar{t} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{A} & 0 & 0 & 0 & 0 & 0 & 0 & -2\bar{t} & 0 \\ 0 & 0 & -2t & 0 & 0 & 0 & 0 & 0 & \mathbf{B} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2t & 0 & 0 & 0 & 0 & 0 & \mathbf{B} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2t & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{B} & 0 & 0 & 0 & 0 & 0 \\ -2t & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{B} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -6t & 0 & 0 & 0 & 0 & 0 & 3|t|^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6t & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3|t|^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & -6t & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3|t|^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -6t & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3|t|^2 \end{pmatrix},$$

where $\mathbf{A} = 4 + 3|t|^2$ and $\mathbf{B} = 3 + 4|t|^2$. Since each entry has a common normalization factor,

$$h = \frac{1 + |t|^2}{(1 - |t|^2)^2},$$

this constant has been factored out.

With Mathematica's Eigenvalue function, the eigenvalues are then calculated for these matrix representations. Our numerical results suggest that the smallest nonzero eigenvalue of \square_b^t on $\mathcal{H}_{2k-1}(\mathbb{S}^3)$ decreases as k increases. Conversely, the smallest nonzero eigenvalue of \square_b^t on $\mathcal{H}_{2k}(\mathbb{S}^3)$ increases with k . The smallest

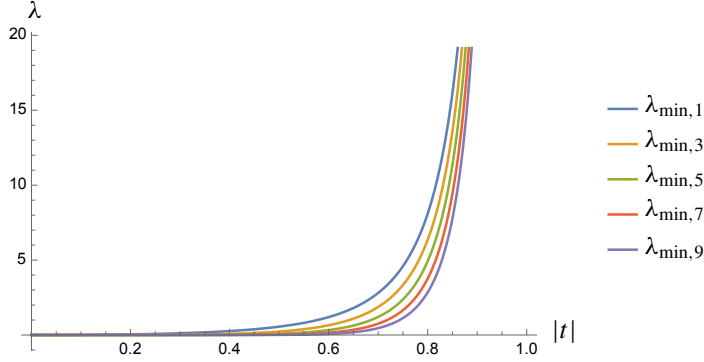


Figure 1. Smallest nonzero eigenvalues for $k = 1, 3, 5, 7, 9$.

eigenvalue of $\mathcal{H}_{2k-1}(\mathbb{S}^3)$ is plotted for $1 \leq k \leq 5$ and $0 < |t| < 1$ in Figure 1. It is apparent that $\lambda_{\min,1} \leq \lambda_{\min,3} \leq \lambda_{\min,5} \leq \lambda_{\min,7} \leq \lambda_{\min,9}$, where $\lambda_{\min,k}$ denotes the smallest nonzero eigenvalue of \square_b^t on $\mathcal{H}_k(\mathbb{S}^3)$. These initial numerical results suggest that $\lim_{k \rightarrow \infty} \lambda_{\min,2k-1} = 0$ for $0 < |t| < 1$, which agrees with our final result.

4. Invariant subspaces of $\mathcal{H}_{2k-1}(\mathbb{S}^3)$ under \square_b^t

In this section we fix $k \geq 1$ and work on $\mathcal{H}_{2k-1}(\mathbb{S}^3)$. As we have seen, \square_b^t can be expanded in the following way:

$$\begin{aligned} \square_b^t &= -(\mathcal{L} + \bar{t}\bar{\mathcal{L}}) \frac{1 + |t|^2}{(1 - |t|^2)^2} (\bar{\mathcal{L}} + t\mathcal{L}) \\ &= -h(\mathcal{L}\bar{\mathcal{L}} + |t|^2\bar{\mathcal{L}}\mathcal{L} + t\mathcal{L}^2 + \bar{t}\bar{\mathcal{L}}^2). \end{aligned} \quad (1)$$

This is because of the linearity of \mathcal{L} and $\bar{\mathcal{L}}$. Now, we need the following property.

Lemma 4.1. *If $\langle f_i, f_j \rangle = 0$ and $f_i, f_j \in \mathcal{H}_{0,2k-1}(\mathbb{S}^3)$, then $\langle \bar{\mathcal{L}}^\sigma f_i, \bar{\mathcal{L}}^\sigma f_j \rangle = 0$ for $0 \leq \sigma \leq 2k - 1$.*

Proof. Choose f_i and f_j in $\mathcal{H}_{0,2k-1}(\mathbb{S}^3)$ and $\langle f_i, f_j \rangle = 0$. We show that $\bar{\mathcal{L}}^\sigma f_i$ and $\bar{\mathcal{L}}^\sigma f_j$ are orthogonal for $0 \leq \sigma \leq 2k - 1$. To do this we use induction on σ . Suppose $\langle \bar{\mathcal{L}}^{\sigma-1} f_i, \bar{\mathcal{L}}^{\sigma-1} f_j \rangle = 0$, and we show that $\langle \bar{\mathcal{L}}^\sigma f_i, \bar{\mathcal{L}}^\sigma f_j \rangle = 0$. Note that, the adjoint of $\bar{\mathcal{L}}$ is $-\mathcal{L}$ and

$$\begin{aligned} \langle \bar{\mathcal{L}}^\sigma f_i, \bar{\mathcal{L}}^\sigma f_j \rangle &= \langle \bar{\mathcal{L}}^{\sigma-1} f_i, -\mathcal{L}\bar{\mathcal{L}}^\sigma f_j \rangle \\ &= \langle \bar{\mathcal{L}}^{\sigma-1} f_i, -(\mathcal{L}\bar{\mathcal{L}})\bar{\mathcal{L}}^{\sigma-1} f_j \rangle \\ &= \langle \bar{\mathcal{L}}^{\sigma-1} f_i, -\square_b \bar{\mathcal{L}}^{\sigma-1} f_j \rangle. \end{aligned}$$

However,² since $\bar{\mathcal{L}}^{\sigma-1} f_j \in \mathcal{H}_{\sigma-1,2k-1-\sigma+1}(\mathbb{S}^3)$, we know that

$$\square_b \bar{\mathcal{L}}^{\sigma-1} f_j = (\sigma)(2k - \sigma - 2)\bar{\mathcal{L}}^{\sigma-1} f_j.$$

²For $f \in \mathcal{H}_{i,j}(\mathbb{S}^3)$, by counting degrees, we notice $\bar{\mathcal{L}}f \in \mathcal{H}_{i-1,j+1}(\mathbb{S}^3)$.

Therefore,

$$\begin{aligned} \langle \bar{\mathcal{L}}^{\sigma-1} f_i, -\square_b \bar{\mathcal{L}}^{\sigma-1} f_j \rangle &= \langle \bar{\mathcal{L}}^{\sigma-1} f_i, -(\sigma)(2k - \sigma - 2) \bar{\mathcal{L}}^{\sigma-1} f_j \rangle \\ &= -(\sigma)(2k - \sigma - 2) \langle \bar{\mathcal{L}}^{\sigma-1} f_i, \bar{\mathcal{L}}^{\sigma-1} f_j \rangle = 0 \end{aligned}$$

by our induction hypothesis as desired. \square

With this, we note that if $\{f_0, \dots, f_{2k-1}\}$ is an orthogonal basis for $\mathcal{H}_{0,2k-1}(\mathbb{S}^3)$, then $\{\bar{\mathcal{L}}^\sigma f_0, \dots, \bar{\mathcal{L}}^\sigma f_{2k-1}\}$ is an orthogonal basis for $\mathcal{H}_{\sigma,2k-1-\sigma}(\mathbb{S}^3)$. Now, we define the following subspaces of $\mathcal{H}_{2k-1}(\mathbb{S}^3)$.

Definition 4.2. Suppose $\{f_0, \dots, f_{2k-1}\}$ is an orthogonal basis for $\mathcal{H}_{0,2k-1}(\mathbb{S}^3)$. Then we define

$$\begin{aligned} V_i &= \text{span}\{f_i, \bar{\mathcal{L}}^2 f_i, \dots, \bar{\mathcal{L}}^{2j-2} f_i, \dots, \bar{\mathcal{L}}^{2k-2} f_i\}, \\ W_i &= \text{span}\{\bar{\mathcal{L}} f_i, \bar{\mathcal{L}}^3 f_i, \dots, \bar{\mathcal{L}}^{2j-1} f_i, \dots, \bar{\mathcal{L}}^{2k-1} f_i\}. \end{aligned}$$

Denote the basis elements for V_i by $v_{i,1}, \dots, v_{i,k}$ and for W_i by $w_{i,1}, \dots, w_{i,k}$. Since each bidegree space $\mathcal{H}_{p,q}(\mathbb{S}^3) \subseteq \mathcal{H}_{2k-1}(\mathbb{S}^3)$ has $2k$ elements, we have $2k$ V_i spaces and $2k$ W_i spaces. We now note the following fact.

Theorem 4.3. $\bigoplus_{i=0}^{2k-1} V_i \oplus W_i = \mathcal{H}_{2k-1}(\mathbb{S}^3)$.

Proof. By Proposition 2.4 and Lemma 4.1, we have

$$\mathcal{H}_{2k-1}(\mathbb{S}^3) = \bigoplus_{i=0}^{2k-1} \mathcal{H}_{i,2k-1-i}(\mathbb{S}^3) = \bigoplus_{i=0}^{2k-1} \bar{\mathcal{L}}^i f_0 \oplus \dots \oplus \bar{\mathcal{L}}^i f_{2k-1}.$$

Manipulating this, we have

$$\begin{aligned} \mathcal{H}_{2k-1}(\mathbb{S}^3) &= \bigoplus_{i=0}^{2k-1} f_i \oplus \bar{\mathcal{L}} f_i \oplus \dots \oplus \bar{\mathcal{L}}^{2k-1} f_i \\ &= \bigoplus_{i=0}^{2k-1} f_i \oplus \bar{\mathcal{L}}^2 f_i \oplus \dots \oplus \bar{\mathcal{L}}^{2k-2} f_i \oplus \bar{\mathcal{L}} f_i \oplus \bar{\mathcal{L}}^3 f_i \oplus \dots \oplus \bar{\mathcal{L}}^{2k-1} f_i \\ &= \bigoplus_{i=0}^{2k-1} V_i \oplus W_i, \end{aligned}$$

which is our goal. \square

The advantage of constructing these spaces in the first place is due to the following fact.

Theorem 4.4. For $0 \leq i \leq 2k-1$, the subspaces V_i and W_i are invariant under \square_b^t .

Proof. By (1), we have

$$\square_b^t = -h(\mathcal{L}\bar{\mathcal{L}} + |t|^2\bar{\mathcal{L}}\mathcal{L} + t\mathcal{L}^2 + \bar{t}\bar{\mathcal{L}}^2).$$

Since the fraction in front is a constant, we can ignore it and only consider the expression in the parentheses. Let $f \in \mathcal{H}_{0,2k-1}(\mathbb{S}^3)$, and define $v_\sigma = \bar{\mathcal{L}}^\sigma f$ to be a basis element of either V_i or W_i , since they have the same form. We first note that $v_\sigma \in \mathcal{H}_{\sigma,2k-1-\sigma}(\mathbb{S}^3)$. Then by our expansion we have

$$\square_b^t v_\sigma = -h(\mathcal{L}\bar{\mathcal{L}}v_\sigma + |t|^2\bar{\mathcal{L}}\mathcal{L}v_\sigma + t\mathcal{L}^2v_\sigma + \bar{t}\bar{\mathcal{L}}^2v_\sigma).$$

We already know $\mathcal{L}\bar{\mathcal{L}}v_\sigma$ and $\bar{\mathcal{L}}\mathcal{L}v_\sigma$ will simply be multiples of v_σ , so we consider \mathcal{L}^2v_σ and $\bar{\mathcal{L}}^2v_\sigma$:

$$\begin{aligned} \mathcal{L}^2v_\sigma &= \mathcal{L}^2\bar{\mathcal{L}}^\sigma f = \mathcal{L}[\mathcal{L}\bar{\mathcal{L}}[\bar{\mathcal{L}}^{\sigma-1}f]] \\ &= -(\sigma)(2k-\sigma)\mathcal{L}\bar{\mathcal{L}}[\bar{\mathcal{L}}^{\sigma-2}f] \\ &= (\sigma)(\sigma-1)(2k+1-\sigma)(2k-\sigma)\bar{\mathcal{L}}^{\sigma-2}f \\ &= (\sigma)(\sigma-1)(2k+1-\sigma)(2k-\sigma)v_{\sigma-2}, \end{aligned} \tag{2a}$$

$$\bar{\mathcal{L}}^2v_\sigma = \bar{\mathcal{L}}^2[\bar{\mathcal{L}}^\sigma f] = \bar{\mathcal{L}}^{\sigma+2}f = v_{\sigma+2}, \tag{2b}$$

so we get multiples of $v_{\sigma-2}$ and $v_{\sigma+2}$. Relating this back to V_i and W_i , we see that if $\sigma = 2j-2$, then $\mathcal{L}^2v_{i,j}$ is a multiple of $v_{i,j-1}$, and $\bar{\mathcal{L}}^2v_{i,j}$ is a multiple of $v_{i,j+1}$. If $\sigma = 2j-1$, we get a similar result for $w_{i,j}$. So we indeed have that both subspaces V_i and W_i are invariant under \square_b^t , and we are done. \square

In light of this fact, we can consider \square_b^t not on the whole space $L^2(\mathbb{S}^3)$ or $\mathcal{H}_{2k-1}(\mathbb{S}^3)$, but rather on these V_i and W_i spaces. In fact, we actually have a representation of \square_b^t on these spaces with respect to the orthogonal bases for V_i and W_i as in Definition 4.2.

Theorem 4.5. *The matrix representation of \square_b^t on V_i and W_i is tridiagonal. That is,*

$$m(\square_b^t) = h \begin{pmatrix} d_1 & u_1 & & & \\ -\bar{t} & d_2 & u_2 & & \\ & -\bar{t} & d_3 & \ddots & \\ & & \ddots & \ddots & u_{k-1} \\ & & & -\bar{t} & d_k \end{pmatrix},$$

where on V_i

$$u_j = -t \cdot (2j)(2j-1)(2k-2j)(2k-1-2j),$$

$$d_j = (2j-1)(2k+1-2j) + |t|^2 \cdot (2j-2)(2k+2-2j),$$

and on W_i

$$u_j = -t \cdot (2j+1)(2j)(2k-2j)(2k-1-2j),$$

$$d_j = (2j)(2k-2j) + |t|^2 \cdot (2j-1)(2k+1-2j).$$

We note that the above definitions don't depend on i ; in other words, each of these matrices are the same on V_i and W_i , regardless of the choice of i .

Proof. Using (2a) and (2b), along with Theorem 2.6, we can entirely describe the action of each piece of \square_b^t on a basis element $v_{i,j}$ or $w_{i,j}$:

$$\begin{aligned} -\mathcal{L}\bar{\mathcal{L}}v_{i,j} &= (2j-1)(2k+1-2j)v_{i,j}, \\ -\mathcal{L}\bar{\mathcal{L}}w_{i,j} &= (2j)(2k-2j)w_{i,j}, \\ -\bar{\mathcal{L}}\mathcal{L}v_{i,j} &= (2j-2)(2k+2-2j)v_{i,j}, \\ -\bar{\mathcal{L}}\mathcal{L}w_{i,j} &= (2j-1)(2k+1-2j)w_{i,j}, \\ -\mathcal{L}^2v_{i,j} &= -(2j-2)(2j-3)(2k+3-2j)(2k+2-2j)v_{i,j-1}, \\ -\mathcal{L}^2w_{i,j} &= -(2j-1)(2j-2)(2k+2-2j)(2k+1-2j)w_{i,j-1}, \\ -\bar{\mathcal{L}}^2v_{i,j} &= -v_{i,j+1}, \\ -\bar{\mathcal{L}}^2w_{i,j} &= -w_{i,j+1}. \end{aligned}$$

By looking at it this way, we notice the tridiagonal structure. So with these observations, we can state that

$$\begin{aligned} \square_b^t v_{i,j} &= h \left(-t \cdot (2j-2)(2j-3)(2k+3-2j)(2k+2-2j)v_{i,j-1} \right. \\ &\quad \left. + ((2j-1)(2k+1-2j) + |t|^2 \cdot (2j-2)(2k+2-2j))v_{i,j} - \bar{t} \cdot v_{i,j+1} \right), \\ \square_b^t w_{i,j} &= h \left(-t \cdot (2j-1)(2j-2)(2k+2-2j)(2k+1-2j)w_{i,j-1} \right. \\ &\quad \left. + ((2j)(2k-2j) + |t|^2 \cdot (2j-1)(2k-1-2j))w_{i,j} - \bar{t} \cdot w_{i,j+1} \right). \end{aligned}$$

Now that we have this formula, we can find $m(\square_b^t)$ on V_i and W_i by computing their effect on the basis vectors $v_{i,j}$ and $w_{i,j}$: When we do this for V_i , we get

$$\begin{aligned} d_j &= (2j-1)(2k+1-2j) + |t|^2 \cdot (2j-2)(2k+2-2j), \\ u_{j-1} &= -t \cdot (2j-2)(2j-3)(2k+3-2j)(2k+2-2j); \end{aligned}$$

hence

$$u_j = -t \cdot (2j)(2j-1)(2k-2j)(2k-1-2j).$$

For W_i , we get

$$\begin{aligned} d_j &= (2j)(2k-2j) + |t|^2 \cdot (2j-1)(2k-1-2j), \\ u_{j-1} &= -t \cdot (2j-1)(2j-2)(2k+2-2j)(2k+1-2j); \end{aligned}$$

hence

$$u_j = -t \cdot (2j+1)(2j)(2k-2j)(2k-1-2j).$$

Finally, by factoring out h and simply substituting in each portion, we obtain the matrix representations above. \square

An immediate consequence of this is that each V_i subspace contributes the same set of eigenvalues to the spectrum of \square_b^t , and similarly for each W_i . Furthermore, we note that the matrices are of rank k (by the tridiagonal structure it is at least of rank $k - 1$ and by Proposition 5.6 the determinant is nonzero, hence rank k). Since the choice of i does not change $m(\square_b^t)$ on these spaces, we will fix an arbitrary i and call the spaces V and W instead.

5. Bottom of the spectrum of \square_b^t

Now that we have a matrix representation for \square_b^t on these V and W spaces inside $\mathcal{H}_{2k-1}(\mathbb{S}^3)$, we can begin to analyze their eigenvalues as k varies. First, we go over some facts about tridiagonal matrices.

Proposition 5.1. *Suppose A is a tridiagonal matrix,*

$$A = \begin{pmatrix} d_1 & u_1 & & & \\ l_1 & d_2 & u_2 & & \\ & l_2 & d_3 & \ddots & \\ & & \ddots & \ddots & u_{k-1} \\ & & & l_{k-1} & d_k \end{pmatrix}$$

and $u_i l_i > 0$ for $1 \leq i < k$. Then A is similar to a symmetric tridiagonal matrix.

Proof. One can verify that if

$$S = \begin{pmatrix} 1 & & & & \\ \sqrt{u_1/l_1} & & & & \\ & \sqrt{u_1 u_2 / (l_1 l_2)} & & & \\ & & \ddots & & \\ & & & \sqrt{u_1 \dots u_{k-1} / (l_1 \dots l_{k-1})} & \end{pmatrix}$$

then $A = S^{-1} B S$, where

$$B = \begin{pmatrix} d_1 & \sqrt{u_1 l_1} & & & \\ \sqrt{u_1 l_1} & d_2 & \sqrt{u_2 l_2} & & \\ & \sqrt{u_2 l_2} & d_3 & \ddots & \\ & & \ddots & \ddots & \sqrt{u_{k-1} l_{k-1}} \\ & & & \sqrt{u_{k-1} l_{k-1}} & d_k \end{pmatrix}.$$

Therefore, A is similar to a symmetric tridiagonal matrix. □

Another special property of tridiagonal matrices is the continuant.

Definition 5.2. Let A be a tridiagonal matrix, like the above. Then we define the *continuant* of A to be a recursive sequence: $f_1 = d_1$, and $f_i = d_i f_{i-1} - u_{i-1} l_{i-1} f_{i-2}$, where $f_0 = 1$.

The reason we define this is because $\det(A) = f_k$. In addition, if we define A_i to mean the square submatrix of A formed by the first i rows and i columns, then $\det(A_i) = f_i$.

With this background, we will now start analyzing \square_b^t on W .

To get bounds on the eigenvalues, we will invoke the Cauchy interlacing theorem; see [Hwang 2004].

Theorem 5.3 (Cauchy interlacing theorem). *Suppose E is an $n \times n$ Hermitian matrix of rank n , and F is an $(n-1) \times (n-1)$ matrix minor of E . If the eigenvalues of E are $\lambda_1 \leq \dots \leq \lambda_n$ and the eigenvalues of F are $\nu_1 \leq \dots \leq \nu_{n-1}$, then the eigenvalues of E and F interlace:*

$$0 < \lambda_1 \leq \nu_1 \leq \lambda_2 \leq \nu_2 \leq \dots \leq \lambda_{n-1} \leq \nu_{n-1} \leq \lambda_n.$$

Now, we can get an intermediate bound on the smallest eigenvalue.

Theorem 5.4. *Suppose A is the Hermitian matrix of rank k , like the above, and $\lambda_1 \leq \dots \leq \lambda_k$ are its eigenvalues. Then*

$$\lambda_1 \leq \frac{\det(A)}{\det(A_{k-1})},$$

where A_{k-1} is A without the last row and column.

Proof. Since A_{k-1} is a $(k-1) \times (k-1)$ matrix minor of A , we can apply the Cauchy interlacing theorem. If the eigenvalues of A_{k-1} are $\nu_1 \leq \dots \leq \nu_{k-1}$, then

$$\lambda_1 \leq \nu_1 \leq \lambda_2 \leq \nu_2 \leq \dots \leq \lambda_{n-1} \leq \nu_{n-1} \leq \lambda_n.$$

Now, we claim that

$$\lambda_1 \det(A_{k-1}) \leq \det(A).$$

To see why this is true, first observe that the determinant of a matrix is simply the product of all its eigenvalues. In particular,

$$\lambda_1 \det(A_{k-1}) = \lambda_1 \nu_1 \dots \nu_{k-1}.$$

But we can simply apply the Cauchy interlacing theorem: since $\nu_1 \leq \lambda_2$, $\nu_2 \leq \lambda_3$, and so on, we get

$$\lambda_1 \nu_1 \dots \nu_{k-1} \leq \lambda_1 \lambda_2 \dots \lambda_k = \det(A).$$

Now, dividing both sides by $\det A_{k-1}$,

$$\lambda_1 \leq \frac{\det(A)}{\det(A_{k-1})},$$

as desired. □

Since $m(\square_b^t)$ on W satisfies the conditions of Proposition 5.1, we find it is similar to the Hermitian tridiagonal matrix

$$A = \begin{pmatrix} a_1 + b_1|t|^2 & c_1|t| & & & \\ c_1|t| & a_2 + b_2|t|^2 & c_2|t| & & \\ & c_2|t| & a_3 + b_3|t|^2 & \ddots & \\ & & \ddots & \ddots & c_{k-1}|t| \\ & & & c_{k-1}|t| & a_k + b_k|t|^2 \end{pmatrix}, \quad (3)$$

where

$$\begin{aligned} a_i &= (2i)(2k - 2i), \\ b_i &= (2i - 1)(2k + 1 - 2i), \\ c_i &= \sqrt{(2i + 1)(2i)(2k - 2i)(2k - 1 - 2i)}. \end{aligned} \quad (4)$$

Note that we are ignoring the constant h for now, which we will add back later. If we can find $\det(A_i)$, then by Theorem 5.4 we can get a closed form for the bound on the smallest eigenvalue. With the following lemma, this is possible:

Lemma 5.5. $a_i b_{i+1} = c_i^2$.

Proof. This is easily verified using the formulas for a_i , b_{i+1} and c_i : $a_i = (2i)(2k - 2i)$, $b_{i+1} = (2i + 1)(2k - 1 - 2i)$, and $c_i^2 = (2i + 1)(2i)(2k - 2i)(2k - 1 - 2i)$. \square

Proposition 5.6. *The determinant of A_i is*

$$\begin{aligned} \det(A_i) &= a_1 a_2 \dots a_{i-1} a_i \\ &\quad + b_1 a_2 \dots a_{i-1} a_i |t|^2 \\ &\quad \vdots \\ &\quad + b_1 b_2 \dots b_{i-1} a_i |t|^{2i-2} \\ &\quad + b_1 b_2 \dots b_{i-1} b_i |t|^{2i}. \end{aligned}$$

In each row, we replace a particular a_j with b_j , and multiply by $|t|^2$. Note that if $i = k$, then $a_k = 0$ and all terms but the last term are 0.

Proof. We will prove this using strong induction on i . We start with the base case $i = 1$, where $\det(A_1) = a_1 + b_1|t|^2$, which does indeed match up with our formula. Next we consider the case $i = 2$, where $\det(A_2) = (a_1 + b_1|t|^2)(a_2 + b_2|t|^2) - c_1^2|t|^2$. By Lemma 5.5 we obtain the desired formula.

Now, assume the formula works for A_{i-1} and A_i . We need to show that the formula works for A_{i+1} . Using the formula for the continuant, we get

$$\det(A_{i+1}) = (a_{i+1} + b_{i+1}|t|^2) \det(A_i) - c_i^2|t|^2 \det(A_{i-1}).$$

By Lemma 5.5,

$$\det(A_{i+1}) = (a_{i+1} + b_{i+1}|t|^2) \det(A_i) - a_i b_{i+1} |t|^2 \det(A_{i-1}).$$

Now, using our induction hypothesis,

$$\begin{aligned} \det(A_{i+1}) &= (a_{i+1} + b_{i+1}|t|^2)(a_1 a_2 \cdots a_i + b_1 a_2 \cdots a_i |t|^2 + \cdots + b_1 b_2 \cdots b_i |t|^{2i}) \\ &\quad - a_i b_{i+1} |t|^2 (a_1 a_2 \cdots a_{i-1} + b_1 a_2 \cdots a_{i-1} |t|^2 + \cdots + b_1 b_2 \cdots b_{i-1} |t|^{2i-2}) \\ &= a_1 a_2 \cdots a_{i+1} + b_1 a_2 \cdots a_{i+1} |t|^2 + \cdots + b_1 b_2 \cdots b_i a_{i+1} |t|^{2i} + a_1 a_2 \cdots a_i b_{i+1} |t|^2 \\ &\quad + b_1 a_2 \cdots a_i b_{i+1} |t|^4 + \cdots + b_1 b_2 \cdots b_{i-1} a_i b_{i+1} |t|^{2i+2} + b_1 b_2 \cdots b_{i+1} |t|^{2i+2} \\ &\quad - a_1 a_2 \cdots a_i b_{i+1} |t|^2 - b_1 a_2 \cdots a_i b_{i+1} |t|^4 - \cdots - b_1 b_2 \cdots b_{i-1} a_i b_{i+1} |t|^{2i+2} \\ &= a_1 a_2 \cdots a_{i+1} + b_1 a_2 \cdots a_{i+1} |t|^2 + \cdots + b_1 b_2 \cdots b_i a_{i+1} |t|^{2i} + b_1 b_2 \cdots b_{i+1} |t|^{2i+2}, \end{aligned}$$

which is the formula for A_{i+1} , and we are done. \square

With this knowledge, we are finally able to prove our main result.

Theorem 5.7. $0 \in \text{essspec}(\square_b^t).$

Proof. By Proposition 5.1, we have that on W in $\mathcal{H}_{2k-1}(\mathbb{S}^3)$ the matrix $m(\square_b^t)$ is similar to the matrix A given in (3)–(4). Now, by Theorem 5.4 we know

$$\lambda_{\min} \leq \frac{\det(A)}{\det(A_{k-1})}.$$

Recall that A_{k-1} denotes the submatrix formed by deleting the last row and column of the $k \times k$ matrix A . To show $0 \in \text{essspec}(\square_b^t)$, we want to show that $\det(A)/\det(A_{k-1}) \rightarrow 0$ as $k \rightarrow \infty$. For this purpose we find an upper bound for $\det(A)/\det(A_{k-1})$ and show that this converges to 0. Notice that Proposition 5.6 implies

$$\begin{aligned} &\frac{\det(A)}{\det(A_{k-1})} \\ &= h \frac{b_1 b_2 \cdots b_{k-1} b_k |t|^{2k}}{a_1 a_2 \cdots a_{k-1} + b_1 a_2 \cdots a_{k-1} |t|^2 + b_1 b_2 \cdots a_{k-1} |t|^4 + \cdots + b_1 b_2 \cdots b_{k-1} |t|^{2k-2}} \\ &\leq h \frac{b_1 b_2 \cdots b_{k-1} b_k |t|^{2k}}{a_1 a_2 \cdots a_{k-1}}, \end{aligned} \tag{5}$$

since, a_j, b_j , and $|t| > 0$. Now using the formulas for a_j and b_j , notice that (5) can be written as

$$h(2k-1)|t|^{2k} \prod_{j=1}^{k-1} \frac{(2j+1)(2k-2j-1)}{(2j)(2k-2j)}.$$

However, we know that for all k and $1 \leq j \leq k-1$,

$$\frac{(2k-2j-1)}{(2k-2j)} < 1,$$

and so,

$$\begin{aligned} h(2k-1)|t|^{2k} \prod_{j=1}^{k-1} \frac{(2j+1)(2k-2j-1)}{(2j)(2k-2j)} &\leq h(2k-1)|t|^{2k} \prod_{j=1}^{k-1} \frac{(2j+1)}{(2j)} \\ &= h(2k-1)|t|^{2k} \prod_{j=1}^{k-1} 1 + \frac{1}{2j}. \end{aligned}$$

Furthermore, we have

$$h(2k-1)|t|^{2k} \prod_{j=1}^{k-1} 1 + \frac{1}{2j} \leq h(2k-1)|t|^{2k} \exp\left(\sum_{j=1}^{k-1} \frac{1}{2j}\right).$$

Note that

$$\sum_{j=1}^{k-1} \frac{1}{2j} \leq \frac{1}{2} \ln k + 1,$$

so our expression becomes

$$\frac{\det(A)}{\det(A_{k-1})} \leq h(2k-1)|t|^{2k} \exp\left(1 + \frac{1}{2} \ln k\right) = eh(2k-1)\sqrt{k}|t|^{2k}$$

and our problem reduces to showing that $\lim_{k \rightarrow \infty} eh(2k-1)\sqrt{k}|t|^{2k} = 0$. We note that h is a constant and $|t| < 1$; therefore, by L'Hospital's rule the last expression indeed goes to 0.

Finally, we have,

$$0 \leq \lim_{k \rightarrow \infty} \lambda_{\min} \leq \lim_{k \rightarrow \infty} \frac{\det(A)}{\det(A_{k-1})} \leq \lim_{k \rightarrow \infty} eh(2k-1)\sqrt{k}|t|^{2k} = 0,$$

and so $\lambda_{\min} \rightarrow 0$. Hence $0 \in \text{essspec}(\square_b^t)$. □

We note that by the discussion in the introduction, this means that the CR-manifold $(\mathcal{L}_t, \mathbb{S}^3)$ is not embeddable into any \mathbb{C}^N .

Acknowledgements

This research was conducted at the NSF REU Site (DMS-1659203) in Mathematical Analysis and Applications at the University of Michigan-Dearborn. We would like to thank the National Science Foundation, the College of Arts, Sciences, and Letters, the Department of Mathematics and Statistics at the University of Michigan-Dearborn, and Al Turfe for their support. We would also like to thank John Clifford, Hyejin Kim, and the other participants of the REU program for fruitful conversations on this topic. We also thank the anonymous referees for constructive comments.

References

- [Axler et al. 2001] S. Axler, P. Bourdon, and W. Ramey, *Harmonic function theory*, 2nd ed., Graduate Texts in Mathematics **137**, Springer, 2001. MR Zbl
- [Bogges 1991] A. Bogges, *CR manifolds and the tangential Cauchy–Riemann complex*, CRC Press, Boca Raton, FL, 1991. MR Zbl
- [Boutet de Monvel 1975] L. Boutet de Monvel, “Intégration des équations de Cauchy–Riemann induites formelles”, exposé 9 in *Séminaire Goulaouic–Lions–Schwartz 1974–1975; équations aux dérivées partielles linéaires et non linéaires*, Centre Math., École Polytech., Paris, 1975. MR Zbl
- [Burns 1979] D. M. Burns, Jr., “Global behavior of some tangential Cauchy–Riemann equations”, pp. 51–56 in *Partial differential equations and geometry* (Park City, Utah, 1977), edited by C. I. Byrnes, Lecture Notes in Pure and Appl. Math. **48**, Dekker, New York, 1979. MR Zbl
- [Burns and Epstein 1990] D. M. Burns and C. L. Epstein, “Embeddability for three-dimensional CR-manifolds”, *J. Amer. Math. Soc.* **3**:4 (1990), 809–841. MR Zbl
- [Chen and Shaw 2001] S.-C. Chen and M.-C. Shaw, *Partial differential equations in several complex variables*, AMS/IP Studies in Advanced Mathematics **19**, American Mathematical Society, Providence, RI, 2001. MR Zbl
- [Davies 1995] E. B. Davies, *Spectral theory and differential operators*, Cambridge Studies in Advanced Mathematics **42**, Cambridge University Press, 1995. MR Zbl
- [Folland 1972] G. B. Folland, “The tangential Cauchy–Riemann complex on spheres”, *Trans. Amer. Math. Soc.* **171** (1972), 83–133. MR Zbl
- [Fu 2005] S. Fu, “Hearing pseudoconvexity with the Kohn Laplacian”, *Math. Ann.* **331**:2 (2005), 475–485. MR Zbl
- [Fu 2008] S. Fu, “Hearing the type of a domain in \mathbb{C}^2 with the $\bar{\partial}$ -Neumann Laplacian”, *Adv. Math.* **219**:2 (2008), 568–603. MR Zbl
- [Hwang 2004] S.-G. Hwang, “Cauchy’s interlace theorem for eigenvalues of Hermitian matrices”, *Amer. Math. Monthly* **111**:2 (2004), 157–159. MR Zbl
- [Kohn 1985] J. J. Kohn, “Estimates for $\bar{\partial}_b$ on pseudoconvex CR manifolds”, pp. 207–217 in *Pseudo-differential operators and applications* (Notre Dame, IN, 1984), edited by F. Trèves, Proc. Sympos. Pure Math. **43**, Amer. Math. Soc., Providence, RI, 1985. MR Zbl
- [Rossi 1965] H. Rossi, “Attaching analytic spaces to an analytic space along a pseudoconcave boundary”, pp. 242–256 in *Proc. Conf. Complex Analysis* (Minneapolis, 1964), Springer, 1965. MR Zbl

Received: 2017-08-22

Revised: 2017-12-02

Accepted: 2017-12-30

abbastaw@msu.edu

Department of Mathematics, Michigan State University,
East Lansing, MI, United States

mmb021@bucknell.edu

Department of Mathematics, Bucknell University,
Lewisburg, PA, United States

rramasam@umich.edu

Department of Mathematics and Statistics,
University of Michigan-Dearborn, Dearborn, MI, United States

zeytuncu@umich.edu

Department of Mathematics and Statistics,
University of Michigan-Dearborn, Dearborn, MI, United States

On the complexity of detecting positive eigenvectors of nonlinear cone maps

Bas Lemmens and Lewis White

(Communicated by Kenneth S. Berenhaut)

In recent work with Lins and Nussbaum, the first author gave an algorithm that can detect the existence of a positive eigenvector for order-preserving homogeneous maps on the standard positive cone. The main goal of this paper is to determine the minimum number of iterations this algorithm requires. It is known that this number is equal to the illumination number of the unit ball B_v of the variation norm, $\|x\|_v := \max_i x_i - \min_i x_i$ on $V_0 := \{x \in \mathbb{R}^n : x_n = 0\}$. In this paper we show that the illumination number of B_v is equal to $\binom{n}{\lfloor n/2 \rfloor}$, and hence provide a sharp lower bound for the running time of the algorithm.

1. Introduction

Classical Perron–Frobenius theory concerns the spectral properties of square non-negative matrices. In recent decades this theory has been extended to a variety of nonlinear maps that preserve a partial ordering induced by a cone (see [Lemmens and Nussbaum 2012] for an up-to-date account).

Of particular interest are order-preserving homogeneous maps $f : \mathbb{R}_{\geq 0}^n \rightarrow \mathbb{R}_{\geq 0}^n$, where

$$\mathbb{R}_{\geq 0}^n := \{x \in \mathbb{R}^n : x_i \geq 0 \text{ for all } i = 1, \dots, n\}$$

is the *standard positive cone*. Recall that $f : \mathbb{R}_{\geq 0}^n \rightarrow \mathbb{R}_{\geq 0}^n$ is *order-preserving* if $f(x) \leq f(y)$ whenever $x \leq y$ and $x, y \in \mathbb{R}_{\geq 0}^n$. Here, $w \leq z$ if $z - w \in \mathbb{R}_{\geq 0}^n$. Furthermore, f is said to be *homogeneous* if $f(\lambda x) = \lambda f(x)$ for all $\lambda \geq 0$ and $x \in \mathbb{R}_{\geq 0}^n$. Such maps arise in mathematical biology [Nussbaum 1989; Schoen 1986] and in optimal control and game theory [Bewley and Kohlberg 1976; Rosenberg and Sorin 2001].

MSC2010: primary 47H07, 47H09; secondary 37C25.

Keywords: nonlinear maps on cones, positive eigenvectors, illumination problem, Hilbert’s metric. White was supported by a London Mathematical Society “Undergraduate Research Bursary” and the School of Mathematics, Statistics and Actuarial Science at the University of Kent.

It is known [Lemmens and Nussbaum 2012, Corollary 5.4.2] that if $f : \mathbb{R}_{\geq 0}^n \rightarrow \mathbb{R}_{\geq 0}^n$ is a continuous, order-preserving, homogeneous map, then there exists $v \in \mathbb{R}_{\geq 0}^n$ such that

$$f(v) = r(f)v,$$

where

$$r(f) := \lim_{k \rightarrow \infty} \|f^k\|_{\mathbb{R}_{\geq 0}^n}^{1/k}$$

is the *cone spectral radius* of f and

$$\|g\|_{\mathbb{R}_{\geq 0}^n} := \sup\{\|g(x)\| : x \in \mathbb{R}_{\geq 0}^n \text{ and } \|x\| \leq 1\}.$$

Thus, as in the case of nonnegative matrices, continuous order-preserving homogeneous maps on $\mathbb{R}_{\geq 0}^n$ have an eigenvector in the cone corresponding to the spectral radius.

In many applications it is important to know if the map has a *positive* eigenvector, i.e., an eigenvector that lies in the interior of $\mathbb{R}_{\geq 0}^n$, that is, $\mathbb{R}_{> 0}^n := \{x \in \mathbb{R}_{\geq 0}^n : x_i > 0 \text{ for } i = 1, \dots, n\}$. This appears to be a much more subtle problem. There exists a variety of sufficient conditions in the literature; see [Cavazos-Cadena 2012; Gaubert and Gunawardena 2004; Lemmens and Nussbaum 2012, Chapter 6; Nussbaum 1988]. Recently, Lemmens, Lins and Nussbaum [Lemmens et al. ≥ 2019 , §5] gave an algorithm that can confirm the existence of a positive eigenvector for continuous, order-preserving, homogeneous maps $f : \mathbb{R}_{\geq 0}^n \rightarrow \mathbb{R}_{\geq 0}^n$. The main goal of this paper is to determine the minimum number of iterations this algorithm needs to perform.

2. Preliminaries

Given a set S in a finite-dimensional vector space V we write S° to denote the interior of S , and we write ∂S to denote the boundary of S with respect to the norm topology on V .

It is known that if $f : \mathbb{R}_{\geq 0}^n \rightarrow \mathbb{R}_{\geq 0}^n$ is an order-preserving homogeneous map and there exists $z \in \mathbb{R}_{> 0}^n$ such that $f(z) \in \partial \mathbb{R}_{\geq 0}^n$, then $f(\mathbb{R}_{> 0}^n) \subset \partial \mathbb{R}_{\geq 0}^n$; see [Lemmens and Nussbaum 2012, Lemma 1.2.2]. Thus to analyse the existence of a positive eigenvector one may as well consider order-preserving homogeneous maps $f : \mathbb{R}_{> 0}^n \rightarrow \mathbb{R}_{> 0}^n$. Moreover, on $\mathbb{R}_{> 0}^n$ we have *Hilbert's metric* d_H , which is given by

$$d_H(x, y) := \log\left(\max_i \frac{x_i}{y_i}\right) - \log\left(\min_i \frac{x_i}{y_i}\right) \quad \text{for } x, y \in \mathbb{R}_{> 0}^n.$$

Note that d_H is not a genuine metric, as $d_H(\lambda x, \mu x) = 0$ for all $x \in \mathbb{R}_{> 0}^n$ and $\lambda, \mu > 0$. In fact, $d_H(x, y) = 0$ if and only if $x = \lambda y$ for some $\lambda > 0$. However, d_H is a metric on the set of rays in $\mathbb{R}_{> 0}^n$.

If $f : \mathbb{R}_{>0}^n \rightarrow \mathbb{R}_{>0}^n$ is order-preserving and homogeneous, then f is nonexpansive under d_H , i.e.,

$$d_H(f(x), f(y)) \leq d_H(x, y) \quad \text{for all } x, y \in \mathbb{R}_{>0}^n;$$

see for example [Lemmens and Nussbaum 2012, Proposition 2.1.1]. In particular, order-preserving homogeneous maps $f : \mathbb{R}_{>0}^n \rightarrow \mathbb{R}_{>0}^n$ are continuous on $\mathbb{R}_{>0}^n$. Moreover, if x and y are eigenvectors of $f : \mathbb{R}_{>0}^n \rightarrow \mathbb{R}_{>0}^n$ with $f(x) = \lambda x$ and $f(y) = \mu y$, then $\lambda = \mu$; see [Lemmens and Nussbaum 2012, Corollary 5.2.2].

In [Lemmens et al. \geq 2019, Theorem 5.1] the following necessary and sufficient conditions were obtained for an order-preserving homogeneous map $f : \mathbb{R}_{>0}^n \rightarrow \mathbb{R}_{>0}^n$ to have a nonempty set of eigenvectors, $E(f) := \{x \in \mathbb{R}_{>0}^n : x \text{ eigenvector of } f\}$, which is bounded under Hilbert's metric.

Theorem 2.1. *If $f : \mathbb{R}_{>0}^n \rightarrow \mathbb{R}_{>0}^n$ is an order-preserving homogeneous map, then $E(f)$ is nonempty and bounded under d_H if and only if for each nonempty proper subset J of $\{1, \dots, n\}$ there exists $x^J \in \mathbb{R}_{>0}^n$ such that*

$$\max_{j \in J} \frac{f(x^J)_j}{x^J_j} < \min_{j \in J^c} \frac{f(x^J)_j}{x^J_j}. \quad (2-1)$$

Note that the assertion is trivial in the case $n = 1$, as each order-preserving homogeneous map $f : \mathbb{R}_{>0} \rightarrow \mathbb{R}_{>0}$ has a nonempty bounded set of eigenvectors. In case $n \geq 2$, Theorem 2.1 yields the following simple algorithm for detecting positive eigenvectors:

Algorithm 2.2. Let $f : \mathbb{R}_{>0}^n \rightarrow \mathbb{R}_{>0}^n$ be an order-preserving homogeneous map. Repeat the following steps until every nonempty proper subset J of $\{1, \dots, n\}$ has been recorded:

- Step 1 Randomly select x , with $x_1 = 1$ and $0 < x_j < 1$ for all $j \in \{2, \dots, n\}$, and compute $f(x)_j/x_j$ for all $j \in \{1, \dots, n\}$.
- Step 2 Record all nonempty proper subsets $J \subset \{1, \dots, n\}$ such that inequality (2-1) holds.

So, if this algorithm halts, then f has an eigenvector in $\mathbb{R}_{>0}^n$ and $E(f)$ is bounded under Hilbert's metric. If $E(f)$ is empty or unbounded under d_H , then the algorithm does not halt. This can happen even if the map is linear. Consider for example the linear map $x \mapsto Ax$ on $\mathbb{R}_{>0}^2$, where

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix},$$

which has no eigenvector in $\mathbb{R}_{>0}^2$. At present no algorithm is known that can decide if an order-preserving homogeneous map on $\mathbb{R}_{>0}^n$ has an empty or an unbounded

set of eigenvectors. It is also unknown if there is an efficient way to generate the vectors x in Step 1.

Note that a randomly chosen x in Step 1 can eliminate multiple subsets J in Step 2. So, it is natural to ask for the least number of vectors required to fulfil the $2^n - 2$ inequalities in (2-1). This number corresponds to the minimum number of times the algorithm has to perform Steps 1 and 2. In this paper we show that one needs at least

$$\binom{n}{\lceil \frac{1}{2}n \rceil}$$

vectors and this lower bound is sharp. Here $\lceil a \rceil$ is the smallest integer $n \geq a$. Likewise we write $\lfloor a \rfloor$ to denote the largest integer $n \leq a$.

3. Connection with the illumination number

Recall that given a compact convex set C with nonempty interior in V , a vector $v \in V$ *illuminates* $z \in \partial C$ if $z + \lambda v \in C^\circ$ for all $\lambda > 0$ sufficiently small. A set S is said to *illuminate* C if for each $z \in \partial C$ there exists $v \in S$ such that v illuminates z . The minimal size of illuminating set for C is called the *illumination number* of C and is denoted by $i(C)$. There is a long-standing open conjecture which asserts that $i(C) \leq 2^n$ for every compact convex body in an n -dimensional vector space; see [Boltyanski et al. 1997, Chapter VI] for further details. It is easy to show, see for example [Lemmens et al. \geq 2019, Lemma 4.1], that if S illuminates every extreme point of C , then S illuminates C .

To proceed we need to discuss the connection between illumination numbers and Theorem 2.1. Firstly, we note that if we let $\Sigma_0 := \{x \in \mathbb{R}_{>0}^n : x_n = 1\}$, then (Σ_0, d_H) is a metric space. Given an order-preserving homogeneous map $f : \mathbb{R}_{>0}^n \rightarrow \mathbb{R}_{>0}^n$ we can consider the *normalised* map $g_f : \Sigma_0 \rightarrow \Sigma_0$ given by

$$g_f(x) := \frac{f(x)}{f(x)_n} \quad \text{for } x \in \Sigma_0.$$

The map g_f is nonexpansive under d_H on Σ_0 . Moreover, $x \in \Sigma_0$ is a fixed point of g_f if and only if x is an eigenvector of f . Thus, if we let $\text{Fix}(g_f) := \{x \in \Sigma_0 : g_f(x) = x\}$, then $\text{Fix}(g_f)$ is nonempty and bounded in (Σ_0, d_H) if and only if $E(f)$ is nonempty and bounded in $(\mathbb{R}_{>0}^n, d_H)$.

It not hard to verify that the map $\text{Log} : \Sigma_0 \rightarrow V_0$ given by

$$\text{Log}(x) := (\log x_1, \dots, \log x_n) \quad \text{for } x = (x_1, \dots, x_n) \in \Sigma_0$$

is an isometry from (Σ_0, d_H) onto $(V_0, \|\cdot\|_v)$, where $V_0 := \{x \in \mathbb{R}^n : x_n = 0\}$ and

$$\|x\|_v := \max_i x_i - \min_i x_i$$

is the *variation norm*.

It follows that the map $h : V_0 \rightarrow V_0$ satisfying $h \circ \text{Log} = \text{Log} \circ g_f$ is nonexpansive under the variation norm, and $\text{Fix}(h)$ is nonempty and bounded in $(V_0, \|\cdot\|_v)$ if and only if $\text{Fix}(g_f)$ is nonempty and bounded in (Σ_0, d_H) .

In [Lemmens et al. ≥ 2019 , Theorem 3.4] the following result concerning fixed point sets of nonexpansive maps on finite-dimensional normed spaces was proved.

Theorem 3.1. *If $h : V \rightarrow V$ is a nonexpansive map on a finite-dimensional normed space V , then $\text{Fix}(h)$ is nonempty and bounded if and only if there exist $w^1, \dots, w^m \in V$ such that $\{f(w^i) - w^i : i = 1, \dots, m\}$ illuminates the unit ball of V .*

For $n \geq 2$, the unit ball B_v of $(V_0, \|\cdot\|_v)$ has $2^n - 2$ extreme points, which are given by

$$\text{ext}(B_v) := \{v_+^I : \emptyset \neq I \subseteq \{1, \dots, n-1\}\} \cup \{v_-^I : \emptyset \neq I \subseteq \{1, \dots, n-1\}\}, \quad (3-1)$$

where $(v_+^I)_i = 1$ if $i \in I$ and 0 otherwise, and $(v_-^I)_i = -1$ if $i \in I$ and 0 otherwise. See [Nussbaum 1994, §2] for details.

In [Lemmens et al. ≥ 2019] the equivalence in Theorem 2.1 was obtained by using Theorem 3.1 and showing that there exists $x^1, \dots, x^m \in \mathbb{R}_{>0}^n$ that fulfil the $2^n - 2$ inequalities in (2-1) if and only if there exist $y^1, \dots, y^m \in V_0$ that illuminate the $2^n - 2$ extreme points of the unit ball B_v . Thus, $i(B_v)$ provides a sharp lower bound for the number of times one needs to repeat Steps 1 and 2 in Algorithm 2.2. In the next section we show the following result concerning $i(B_v)$:

Theorem 3.2. *If B_v is the unit ball of $(V_0, \|\cdot\|_v)$ and $n \geq 2$, then*

$$i(B_v) = \binom{n}{\lceil \frac{1}{2}n \rceil}.$$

4. Proof of Theorem 3.2

Note that the map $(x_1, \dots, x_n) \in V_0 \mapsto (x_1, \dots, x_{n-1}) \in \mathbb{R}^{n-1}$ is an isometry from $(V_0, \|\cdot\|_v)$ onto $(\mathbb{R}^{n-1}, \|\cdot\|_H)$, where

$$\|x\|_H := \left(\max_i x_i\right) \vee 0 - \left(\min_i x_i\right) \wedge 0.$$

Here $a \wedge b := \min(a, b)$ and $a \vee b := \max(a, b)$. Note also that if B_H is the unit ball in $(\mathbb{R}^{n-1}, \|\cdot\|_H)$, then

$$\text{ext}(B_H) = (\{0, 1\}^{n-1} \cup \{0, -1\}^{n-1}) \setminus \{(0, \dots, 0)\}$$

and

$$i(B_H) = i(B_v).$$

For notational simplicity we work with B_H instead of B_v .

The two subsets

$$E_+ := \{0, 1\}^{n-1} \setminus \{(0, \dots, 0)\} \quad \text{and} \quad E_- := \{0, -1\}^{n-1} \setminus \{(0, \dots, 0)\}$$

of $\text{ext}(B_H)$ play a key role in the argument. On $\text{ext}(B_H)$ we have the usual partial ordering $x \leq y$ if $y - x \in \mathbb{R}_{\geq 0}^{n-1}$, which gives rise to two finite partially ordered sets (E_+, \leq) and (E_-, \leq) .

Recall that subset \mathcal{A} of a partially ordered set (P, \leq) is called an *antichain* if $x, y \in \mathcal{A}$ and $x \leq y$ implies $x = y$. A *chain* \mathcal{C} in (P, \leq) is a totally ordered subset if for each $x, y \in \mathcal{C}$ we have that either $x \leq y$ or $y \leq x$. The *length* of a chain \mathcal{C} is the number of distinct elements in \mathcal{C} .

Lemma 4.1. *Let \mathcal{A} be an antichain in (E_+, \leq) or in (E_-, \leq) . If $x \neq y$ in \mathcal{A} are illuminated by v and w , respectively, then $v \neq w$.*

Proof. Suppose that \mathcal{A} is an antichain in (E_+, \leq) and $x \neq y$ are in \mathcal{A} . Then there exist $i \neq j$ such that $0 = x_i < y_i = 1$ and $0 = y_j < x_j = 1$. Now suppose by way of contradiction that z illuminates x and y . So, $\|x + \lambda z\|_H < 1$ and $\|y + \lambda z\|_H < 1$ for all $\lambda > 0$ sufficiently small. Suppose first that $z_i \leq z_j$. Then for $\lambda > 0$ small,

$$1 + \lambda z_j = x_j + \lambda z_j \leq \|x + \lambda z\|_H < 1,$$

and hence $z_j < 0$. So, $z_i \leq z_j < 0$. But then

$$1 + \lambda(z_j - z_i) = x_j + \lambda z_j - \lambda z_i \leq \|x + \lambda z\|_H < 1,$$

which is impossible. On the other hand, if $z_j \leq z_i$, then $1 + \lambda z_i \leq \|y + \lambda z\|_H < 1$, so that $z_j \leq z_i < 0$. But then

$$1 + \lambda(z_i - z_j) = y_i + \lambda z_i - \lambda z_j \leq \|y + \lambda z\|_H < 1,$$

which again is impossible. Thus, z cannot illuminate both x and y .

The argument for the case where \mathcal{A} is an antichain in (E_-, \leq) is similar. \square

Lemma 4.2. *If $x, y \in \text{ext}(B_H)$ are such that $x_i = 1$ and $y_i = -1$ for some i , then one needs two distinct vectors to illuminate x and y .*

Proof. Suppose w illuminates x and y . Then $1 + \lambda w_i = x_i + \lambda w_i \leq \|x + \lambda w\|_H < 1$ for all $\lambda > 0$ sufficiently small, and hence $w_i < 0$. But also $1 - \lambda w_i = -(y_i + \lambda w_i) \leq \|y + \lambda w\|_H < 1$ for all $\lambda > 0$ sufficiently small. This implies that $w_i > 0$, which is impossible. Thus, one needs at least two vectors to illuminate x and y . \square

Corollary 4.3. *If B_H is the unit ball of $(\mathbb{R}^{n-1}, \|\cdot\|_H)$ and $n \geq 2$, then*

$$i(B_H) \geq \binom{n}{\lceil \frac{1}{2}n \rceil}.$$

Proof. For $1 \leq k, m \leq n-1$ define the antichains $\mathcal{A}_+(k) := \{x \in E_+ : \sum_i x_i = k\}$ and $\mathcal{A}_-(m) := \{x \in E_- : \sum_i x_i = -m\}$. If $n > 1$ is odd, then we can take $k := \frac{1}{2}(n-1)$ and $m := \frac{1}{2}(n+1)$ and conclude from Lemmas 4.1 and 4.2 that we need at least

$$\binom{n-1}{\frac{1}{2}(n-1)} + \binom{n-1}{\frac{1}{2}(n+1)} = \binom{n}{\lceil \frac{1}{2}n \rceil}$$

distinct vectors to illuminate the extreme points in $\mathcal{A}_+(k) \cup \mathcal{A}_-(m)$, as for each $x \in \mathcal{A}_+(k)$ and $y \in \mathcal{A}_-(m)$ there exists an i such that $x_i = 1$ and $y_i = -1$.

Likewise if $n > 1$ is even, we can take $k = m = \lceil \frac{1}{2}(n-1) \rceil$, and deduce from Lemmas 4.1 and 4.2 that we need at least

$$\binom{n-1}{\lceil \frac{1}{2}(n-1) \rceil} + \binom{n-1}{\lceil \frac{1}{2}(n-1) \rceil} = \binom{n-1}{\lfloor \frac{1}{2}(n-1) \rfloor} + \binom{n-1}{\lceil \frac{1}{2}(n-1) \rceil} = \binom{n}{\frac{1}{2}n}$$

distinct vectors to illuminate the extreme points in $\mathcal{A}_+(k) \cup \mathcal{A}_-(m)$. \square

Lemma 4.4. *If \mathcal{C} is a chain in (E_+, \leq) or in (E_-, \leq) , then there exists w that illuminates each element of \mathcal{C} .*

Proof. Let \mathcal{C} be a chain in (E_+, \leq) or in (E_-, \leq) . We call a chain $c_1 \leq c_2 \leq \dots \leq c_m$ in (E_+, \leq) or in (E_-, \leq) maximal if it has length $n-1$. The chain \mathcal{C} is contained in a maximal chain. As each coordinate permutation is an isometry of $(\mathbb{R}^{n-1}, \|\cdot\|_H)$ and the map $x \mapsto -x$ is an isometry of $(\mathbb{R}^{n-1}, \|\cdot\|_H)$, we may assume without loss of generality that \mathcal{C} is contained in the maximal chain,

$$\mathcal{C}^* : (1, 0, 0, \dots, 0) \leq (1, 1, 0, \dots, 0) \leq \dots \leq (1, 1, \dots, 1, 0) \leq (1, 1, 1, \dots, 1).$$

Let $w \in \mathbb{R}^{n-1}$ be such that $w_1 < w_2 < \dots < w_{n-1} < 0$. Now if x is the k -th element in the maximal chain and $k < n-1$, then for all $\lambda > 0$ sufficiently small

$$\|x + \lambda w\|_H = (\max_i x_i + \lambda w_i) \vee 0 - (\min_i x_i + \lambda w_i) \wedge 0 = 1 + \lambda w_k - \lambda w_{k+1} < 1.$$

On the other hand, if $x = (1, 1, \dots, 1)$, then clearly $\|x + \lambda w\|_H = 1 + \lambda w_{n-1} < 1$ for all $\lambda > 0$ small. Thus w illuminates each element of \mathcal{C}^* and we are done. \square

To proceed we need to recall a few classical results in the combinatorics of finite partially ordered sets; see [Jukna 2001, §9.1 and 9.2]. Firstly, we recall Dilworth's theorem, which says that if the maximum size of an antichain in a finite partially ordered set (P, \leq) is r , then P can be partitioned into r disjoint chains. In the case where the partially ordered set is $(\{0, 1\}^d, \leq)$, one can combine this result with Sperner's theorem, which says that the maximum size of an antichain in $(\{0, 1\}^d, \leq)$ is $\binom{d}{\lceil d/2 \rceil}$. Thus, $(\{0, 1\}^d, \leq)$ can be partitioned into $\binom{d}{\lceil d/2 \rceil}$ disjoint chains.

To obtain our result we need some more detailed information about the partitions. In particular, we need a result by De Bruijn, Tengbergen, and Kruyswijk [de Bruijn

et al. 1951] concerning symmetric chains; see also [Jukna 2001, Theorem 9.3]. A chain $x^1 \leq \dots \leq x^k$ in $(\{0, 1\}^d, \leq)$ is said to be *symmetric* if

- (a) $(\sum_{j=1}^d x_j^m) + 1 = \sum_{j=1}^d x_j^{m+1}$ for all $1 \leq m < k$, i.e., x^{m+1} is an immediate successor of x^m , and
- (b) $\sum_{j=1}^d x_j^k = d - \sum_{j=1}^d x_j^1$.

Theorem 4.5 [de Bruijn et al. 1951]. *The poset $(\{0, 1\}^d, \leq)$ can be partitioned into $\binom{d}{\lceil d/2 \rceil}$ disjoint symmetric chains.*

Let us now prove the main result of the paper.

Proof of Theorem 3.2. First recall that by Corollary 4.3 it suffices to show that $i(B_H) \leq \binom{n}{\lceil n/2 \rceil}$, as $i(B_V) = i(B_H)$. In other words, we only need to show that $\text{ext}(B_H)$ can be illuminated by $\binom{n}{\lceil n/2 \rceil}$ vectors.

There are two cases to consider: $n \geq 2$ even, and $n \geq 2$ odd.

Let us first consider the case where $n \geq 2$ is even. By Dilworth's theorem and Sperner's theorem we know that the partially ordered set $(\{0, 1\}^{n-1}, \leq)$ can be partitioned into $\binom{n-1}{\lceil (n-1)/2 \rceil}$ disjoint chains. This implies that each of the partially ordered sets (E_+, \leq) and (E_-, \leq) can be partitioned into $\binom{n-1}{\lceil (n-1)/2 \rceil}$ disjoint chains. It now follows from Lemma 4.4 that we need at most

$$\binom{n-1}{\lceil \frac{1}{2}(n-1) \rceil} + \binom{n-1}{\lceil \frac{1}{2}(n-1) \rceil} = \binom{n-1}{\lfloor \frac{1}{2}(n-1) \rfloor} + \binom{n-1}{\lceil \frac{1}{2}(n-1) \rceil} = \binom{n}{\frac{1}{2}n}$$

distinct vectors to illuminate $\text{ext}(B_H)$. This implies that $i(B_V) = i(B_H) \leq \binom{n}{n/2}$.

Now suppose $n \geq 2$ is odd. By Theorem 4.5 we know that $(\{0, 1\}^{n-1}, \leq)$ can be partitioned into $\binom{n-1}{(n-1)/2}$ disjoint symmetric chains.

Let us consider such a symmetric chain decomposition, and let

$$\mathcal{A}_k := \{x \in \{0, 1\}^{n-1} : \sum_i x_i = k\},$$

which is an antichain of size $\binom{n-1}{k}$. Each element of $\mathcal{A}_{(n+1)/2}$ is contained in a distinct symmetric chain, and each of these chains contains an $x \in \{0, 1\}^{n-1}$ with $\sum_i x_i = \frac{1}{2}(n-1)$. Thus, the symmetric chain decomposition of $(\{0, 1\}^{n-1}, \leq)$ consists of

$$\binom{n-1}{\frac{1}{2}(n+1)}$$

chains containing a vector x with $\sum_i x_i = \frac{1}{2}(n+1)$, and

$$\binom{n-1}{\frac{1}{2}(n-1)} - \binom{n-1}{\frac{1}{2}(n+1)}$$

chains consisting of a single vector x with $\sum_i x_i = \frac{1}{2}(n-1)$.

By deleting $(0, 0, \dots, 0)$ from $\{0, 1\}^{n-1}$ we obtain a partition of (E_+, \leq) into disjoint chains. Let \mathcal{S} be the set of vectors $x \in E_+$ which form a singleton chain and $\sum_i x_i = \frac{1}{2}(n-1)$. So,

$$|\mathcal{S}| = \binom{n-1}{\frac{1}{2}(n-1)} - \binom{n-1}{\frac{1}{2}(n+1)}.$$

Now pair each $x \in E_+$ with $x' \in E_-$, where $x'_i = 0$ if $x_i = 1$, and $x'_i = -1$ if $x_i = 0$. In this way we obtain a partition of (E_-, \leq) into disjoint chains with $|\mathcal{S}|$ chains consisting of a single vector. In other words, for each $x \in \mathcal{S}$ we have that $x' \in E_-$ forms a singleton chain in the chain decomposition of (E_-, \leq) .

We know from Lemma 4.4 that we can illuminate the $\binom{n-1}{(n+1)/2}$ chains in (E_+, \leq) containing a vector x with $\sum_i x_i = \frac{1}{2}(n+1)$ using $\binom{n-1}{(n+1)/2}$ vectors. Likewise, we can illuminate the corresponding $\binom{n-1}{(n+1)/2}$ chains in (E_-, \leq) with $\binom{n-1}{(n+1)/2}$ vectors. So, it remains to illuminate the singleton chains in (E_+, \leq) and (E_-, \leq) .

Note that if we can illuminate each pair $\{x, x'\}$, with $x \in \mathcal{S}$ and x' the corresponding vector in E_- , by a single vector, then we need at most

$$2 \binom{n-1}{\frac{1}{2}(n+1)} + \binom{n-1}{\frac{1}{2}(n-1)} - \binom{n-1}{\frac{1}{2}(n+1)} = \binom{n-1}{\frac{1}{2}(n-1)} + \binom{n-1}{\frac{1}{2}(n+1)} = \binom{n}{\lceil \frac{1}{2}n \rceil}$$

vectors to illuminate $\text{ext}(B_H)$, and hence $i(B_V) = i(B_H) \leq \binom{n}{\lceil n/2 \rceil}$ if $n \geq 2$ is odd.

To see how this can be done we consider such a pair $\{x, x'\}$ with $x \in \mathcal{S}$ and let $I := \{i : x_i = 1\}$ and $J := \{i : x_i = 0\}$. So, $I = \{i : x'_i = 0\}$ and $J = \{i : x'_i = -1\}$. Now let $w \in \mathbb{R}^{n-1}$ be such that $w_i < 0$ for all $i \in I$ and $w_i > 0$ for all $i \in J$. Then for all $\lambda > 0$ sufficiently small,

$$\|x + \lambda w\|_H = \max_{i \in I} (1 + \lambda w_i) - 0 < 1$$

and

$$\|x' + \lambda w\|_H = 0 - \min_{i \in J} (-1 + \lambda w_i) < 1.$$

This shows that w illuminates x and x' , which completes the proof. \square

References

- [Bewley and Kohlberg 1976] T. Bewley and E. Kohlberg, “The asymptotic theory of stochastic games”, *Math. Oper. Res.* **1**:3 (1976), 197–208. MR Zbl
- [Boltyanski et al. 1997] V. Boltyanski, H. Martini, and P. S. Soltan, *Excursions into combinatorial geometry*, Springer, 1997. MR Zbl
- [de Bruijn et al. 1951] N. G. de Bruijn, C. van Ebbenhorst Tengbergen, and D. Kruyswijk, “On the set of divisors of a number”, *Nieuw Arch. Wiskunde* (2) **23** (1951), 191–193. MR Zbl
- [Cavazos-Cadena 2012] R. Cavazos-Cadena, “Equivalence of communication and projective boundedness properties for monotone and homogeneous functions”, *Nonlinear Anal.* **75**:2 (2012), 775–785. MR Zbl

- [Gaubert and Gunawardena 2004] S. Gaubert and J. Gunawardena, “The Perron–Frobenius theorem for homogeneous, monotone functions”, *Trans. Amer. Math. Soc.* **356**:12 (2004), 4931–4950. MR Zbl
- [Jukna 2001] S. Jukna, *Extremal combinatorics*, Springer, 2001. MR Zbl
- [Lemmens and Nussbaum 2012] B. Lemmens and R. Nussbaum, *Nonlinear Perron–Frobenius theory*, Cambridge Tracts in Mathematics **189**, Cambridge University Press, 2012. MR Zbl
- [Lemmens et al. \geq 2019] B. Lemmens, B. Lins, and R. Nussbaum, “Detecting fixed points of nonexpansive maps by illuminating the unit ball”, preprint. To appear in *Israel J. Math.* arXiv
- [Nussbaum 1988] R. D. Nussbaum, *Hilbert’s projective metric and iterated nonlinear maps*, Mem. Amer. Math. Soc. **391**, Amer. Math. Soc., Providence, RI, 1988. MR Zbl
- [Nussbaum 1989] R. D. Nussbaum, *Iterated nonlinear maps and Hilbert’s projective metric, II*, Mem. Amer. Math. Soc. **401**, Amer. Math. Soc., Providence, RI, 1989. MR Zbl
- [Nussbaum 1994] R. D. Nussbaum, “Finsler structures for the part metric and Hilbert’s projective metric and applications to ordinary differential equations”, *Differential Integral Equations* **7**:5-6 (1994), 1649–1707. MR Zbl
- [Rosenberg and Sorin 2001] D. Rosenberg and S. Sorin, “An operator approach to zero-sum repeated games”, *Israel J. Math.* **121** (2001), 221–246. MR Zbl
- [Schoen 1986] R. Schoen, “The two-sex multiethnic stable population model”, *Theoret. Population Biol.* **29**:3 (1986), 343–364. MR Zbl

Received: 2017-08-29 Revised: 2017-11-21 Accepted: 2017-12-14

b.lemmens@kent.ac.uk *School of Mathematics, Statistics & Actuarial Science,
University of Kent, Canterbury, United Kingdom*

lcw32@kent.ac.uk *School of Mathematics, Statistics & Actuarial Science,
University of Kent, Canterbury, United Kingdom*

Antiderivatives and linear differential equations using matrices

Yotsanan Meemark and Songpon Sriwongsa

(Communicated by Kenneth S. Berenhaut)

We show how to find the closed-form solutions for antiderivatives of $x^n e^{ax} \sin bx$ and $x^n e^{ax} \cos bx$ for all $n \in \mathbb{N}_0$ and $a, b \in \mathbb{R}$ with $a^2 + b^2 \neq 0$ by using an idea of Rogers, who suggested using the inverse of the matrix for the differential operator. Additionally, we use the matrix to illustrate the method to find the particular solution for a nonhomogeneous linear differential equation with constant coefficients and forcing terms involving $x^n e^{ax} \sin bx$ or $x^n e^{ax} \cos bx$.

1. Matrix inversion

The concepts of basis and matrix for a linear transformation relative to bases are fundamental in linear algebra. Rogers [1997] suggested an application of the inverse of the matrix for the differential operator on $C^\infty(\mathbb{R})$ relative to a given basis \mathcal{B} to obtain antiderivatives of functions in \mathcal{B} . This idea was used with Chebyshev's polynomials and some binomial identities to get a formula for integrating the power of cosines [Meemark and Leela-apiradee 2011]. Also, the integrals of powers of sine and tangent were obtained by Matlak et al. [2014]. This idea provides a useful application of linear algebra to calculus.

Let n be a nonnegative integer and $\mu = a + bi$ a nonzero complex number. In this work, we apply the idea of Rogers with the complex approach to find the antiderivatives of $x^n e^{ax} \sin bx$ and $x^n e^{ax} \cos bx$ for all $n \in \mathbb{N}_0$ and $a, b \in \mathbb{R}$ with $a^2 + b^2 \neq 0$. More precisely, $x^n e^{\mu x} = x^n e^{ax} \cos bx + i x^n e^{ax} \sin bx$. The linearity of the integral operator and comparing the real and imaginary parts yield the desired integrals.

Consider the set of linearly independent functions

$$\mathcal{B}_n = \{e^{\mu x}, x e^{\mu x}, \dots, x^n e^{\mu x}\}.$$

MSC2010: primary 15A09; secondary 34A30.

Keywords: differential operator, inverse of matrix, rectangular form.

Let V be the space with the basis \mathcal{B}_n and $\mathcal{D} : V \rightarrow V$ be the linear operator defined by $\mathcal{D}(f) = f'$ for all $f \in V$. Since V contains no nonzero constant function, $\mathcal{D} : V \rightarrow V$ is invertible. Note that for $j \in \{0, 1, 2, \dots, n\}$, we have

$$\mathcal{D}(x^j e^{\mu x}) = \mu x^j e^{\mu x} + j x^{j-1} e^{\mu x}.$$

This yields the following theorem.

Theorem 1. *The matrix for \mathcal{D} relative to the basis \mathcal{B}_n is*

$$D_n = [\mathcal{D}]_{\mathcal{B}_n} = \begin{bmatrix} \mu & 1 & & & \\ & \mu & 2 & & \\ & & \mu & \ddots & \\ & & & \ddots & n \\ & & & & \mu \end{bmatrix}.$$

According to Rogers' technique [1997], we shall use the inverse of D_n to find the general formula for $\int x^n e^{\mu x} dx$. From the above theorem, D_n is invertible and D_n^{-1} is the upper triangular matrix given by

$$D_n^{-1} = \begin{bmatrix} c_{0,0} & c_{0,1} & \cdots & c_{0,n} \\ & c_{1,1} & \cdots & c_{1,n} \\ & & \ddots & \vdots \\ & & & c_{n,n} \end{bmatrix}.$$

Identifying $\int x^n e^{\mu x} dx$ with the value $D_n^{-1}(x^n e^{\mu x}) \in V$, we get

$$\int x^n e^{\mu x} dx = \sum_{j=0}^n c_{j,n} x^j e^{\mu x},$$

where the $c_{j,n}$, $j \in \{0, 1, \dots, n\}$, satisfy the system of equations

$$\begin{aligned} \mu c_{0,n} + c_{1,n} &= 0, \\ \mu c_{1,n} + 2c_{2,n} &= 0, \\ &\vdots \\ \mu c_{n-1,n} + nc_{n,n} &= 0, \\ \mu c_{n,n} &= 1, \end{aligned}$$

because the product of D_n and D_n^{-1} is the identity matrix. Clearly, $c_{n,n} = 1/\mu$. The back-substitution yields

$$c_{j,n} = c_{n-(n-j),n} = \left(\frac{-n}{\mu}\right) \left(\frac{-(n-1)}{\mu}\right) \cdots \left(\frac{-(j-1)}{\mu}\right) \left(\frac{1}{\mu}\right) = \left(\frac{n!}{j!}\right) \left(\frac{(-1)^{n-j}}{\mu^{n-j+1}}\right)$$

for all $j \in \{0, 1, \dots, n-1\}$. Hence, we have shown:

Theorem 2. For each $j \in \{0, 1, \dots, n\}$, we have

$$c_{j,n} = \binom{n!}{j!} \left(\frac{(-1)^{n-j}}{\mu^{n-j+1}} \right).$$

Note that the integration by parts provides the recursion

$$\int x^n e^{\mu x} dx = \frac{1}{\mu} x^n e^{\mu x} - \frac{n}{\mu} \int x^{n-1} e^{\mu x} dx.$$

It follows that the algorithm presented in Theorem 2, requiring only the last column of D_n^{-1} , is more efficient than integration by parts, which requires the computation of the entire matrix D_n^{-1} .

2. Applications

We use the result from Theorem 2 to find the closed-form of $\int x^n e^{ax} \sin bx dx$ and $\int x^n e^{ax} \cos bx dx$. Moreover, we also use the basis introduced in the above section to find the particular solution for a nonhomogeneous linear differential equation with constant coefficients and forcing terms involving $x^n e^{ax} \sin bx$ or $x^n e^{ax} \cos bx$.

For real μ , the general form of $\int x^n e^{\mu x} dx$ derived in Theorem 2 is the final form. Now, we assume that $\mu = a + ib$ with $b \neq 0$; the rectangular form of $\int x^n e^{\mu x} dx$ still remains to be computed. First, we express $\int x^n e^{\mu x} dx = (p_n(x) - i q_n(x)) e^{\mu x}$ for some polynomials $p_n(x)$ and $q_n(x)$ of degree n in $\mathbb{R}[x]$. Let $\varrho = |\mu|$ and $\varphi = \arg(\mu)$. Then we have

$$\frac{1}{\mu} = \frac{1}{\varrho} e^{-i\varphi} \quad \text{and} \quad \frac{1}{\mu^{n-j+1}} = \frac{1}{\varrho^{n-j+1}} e^{-i\varphi(n-j+1)};$$

hence

$$c_{j,n} = (-1)^{n-j} \binom{n!}{j!} (s_{n-j+1} - i t_{n-j+1}),$$

where

$$s_m = \frac{1}{\varrho^m} \cos m\varphi \quad \text{and} \quad t_m = \frac{1}{\varrho^m} \sin m\varphi \quad \text{for } m \in \mathbb{N}.$$

Since

$$\int x^n e^{\mu x} dx = \sum_{j=0}^n c_{j,n} x^j e^{\mu x} = (p_n(x) - i q_n(x)) e^{\mu x},$$

by comparing the real and imaginary parts, we have

$$p_n(x) = \sum_{j=0}^n (-1)^{n-j} \binom{n!}{j!} s_{n-j+1} x^j \quad \text{and} \quad q_n(x) = \sum_{j=0}^n (-1)^{n-j} \binom{n!}{j!} t_{n-j+1} x^j.$$

Moreover,

$$\begin{aligned} \int x^n e^{\mu x} dx &= (p_n(x) - i q_n(x)) e^{\mu x} = (p_n(x) - i q_n(x)) [e^{ax} (\cos bx + i \sin bx)] \\ &= e^{ax} [p_n(x) \cos bx + q_n(x) \sin bx] - i e^{ax} [q_n(x) \cos bx - p_n(x) \sin bx] \end{aligned}$$

and

$$\int x^n e^{\mu x} dx = \int x^n e^{ax} \cos bx dx + i \int x^n e^{ax} \sin bx dx.$$

In conclusion, we obtain the antiderivatives of $x^n e^{ax} \sin bx$ and $x^n e^{ax} \cos bx$.

Theorem 3. For $n \in \mathbb{N} \cup \{0\}$ and $a, b \in \mathbb{R}$ with $a^2 + b^2 \neq 0$,

$$\begin{aligned} \int x^n e^{ax} \sin bx dx &= -e^{ax} [q_n(x) \cos bx - p_n(x) \sin bx] + C, \\ \int x^n e^{ax} \cos bx dx &= e^{ax} [p_n(x) \cos bx + q_n(x) \sin bx] + C, \end{aligned}$$

where $p_n(x)$ and $q_n(x)$ are polynomials of degree n computed above.

Finally, we remark that to apply the idea of Rogers [1997] and obtain the same results, one may use the basis

$$\mathcal{C}_n = \{e^{ax} \sin bx, e^{ax} \cos bx, x e^{ax} \sin bx, x e^{ax} \cos bx, x^2 e^{ax} \sin bx, x^2 e^{ax} \cos bx, \dots, x^n e^{ax} \sin bx, x^n e^{ax} \cos bx\}$$

instead of \mathcal{B}_n introduced above. But then the matrix for the differential operator relative to \mathcal{C}_n has the block matrix form

$$D = \begin{bmatrix} A & I_2 & & & \\ & A & 2I_2 & & \\ & & A & \ddots & \\ & & & \ddots & nI_2 \\ & & & & A \end{bmatrix},$$

where

$$A = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$$

and I_2 is the 2×2 identity matrix, and the computation for the matrix D^{-1} is tedious. The use of the complex approach and the basis \mathcal{B}_n reduce the complexity of the computation. Moreover, our approach can be used to find the particular solution for a nonhomogeneous linear differential equation with constant coefficients and forcing terms involving $x^n e^{ax} \sin bx$ or $x^n e^{ax} \cos bx$ as follows.

Recall from Theorem 1 that the matrix for the differential operator relative to the basis \mathcal{B}_n is

$$D_n = \begin{bmatrix} \mu & 1 & & & \\ & \mu & 2 & & \\ & & \mu & \ddots & \\ & & & \ddots & n \\ & & & & \mu \end{bmatrix}.$$

It is immediate from the linearity of the differential operator that it suffices to find the particular solution of the equation

$$a_k y^{(k)} + \cdots + a_0 y = x^n e^{\mu x} = (x^n e^{ax} \cos bx) + i(x^n e^{ax} \sin bx),$$

denoted by y_p . Note that $[x^n e^{\mu x}]_{D_n} = (0, \dots, 0, 1)^T$. Let $L = a_k D^k + \cdots + a_0 I$. We shall find a solution of $L[y_p]_{D_n} = (0, \dots, 0, 1)^T$. Then we get that $y_1 = \operatorname{Re} y_p$ and $y_2 = \operatorname{Im} y_p$ are the particular solutions for the equations $a_k y^{(k)} + \cdots + a_0 y = x^n e^{ax} \cos bx$ and $a_k y^{(k)} + \cdots + a_0 y = x^n e^{ax} \sin bx$, respectively.

Example. Consider the equations $y'' - 3y' + 2y = x e^x \sin x$ and $y'' - 3y' + 2y = x e^x \cos x$. As per the set-up above,

$$\mu = 1 + i, \quad L = \begin{bmatrix} \mu^2 - 3\mu + 2 & 2\mu - 3 \\ 0 & \mu^2 - 3\mu + 2 \end{bmatrix},$$

and so the solution $[y_p]_{D_1}$ of $L[y_p]_{D_1} = (0, \dots, 0, 1)^T$ is

$$\left(-\frac{2\mu - 3}{(\mu^2 - 3\mu + 2)^2}, \frac{1}{\mu^2 - 3\mu + 2} \right)^T.$$

Then

$$y_p = -\frac{2\mu - 3}{(\mu^2 - 3\mu + 2)^2} e^{\mu x} + \frac{1}{\mu^2 - 3\mu + 2} x e^{\mu x}.$$

Hence, the particular solution of the first equation is

$$y_1 = \operatorname{Im} y_p = e^x \left(\left(-1 - \frac{1}{2}x \right) \sin x - \left(\frac{1}{2} - \frac{1}{2}x \right) \cos x \right),$$

and the particular solution of the second equation is

$$y_2 = \operatorname{Re} y_p = e^x \left(\left(-1 - \frac{1}{2}x \right) \cos x + \left(\frac{1}{2} - \frac{1}{2}x \right) \sin x \right).$$

3. Acknowledgments

This work grew out of an independent project while Sriwongsa was an undergraduate student at Chulalongkorn University. The project was funded by the Human Resource Development in Science Project (Science Achievement Scholarship of Thailand, SAST).

References

- [Matlak et al. 2014] D. Matlak, J. Matlak, D. Slota, and R. Witula, “Differentiation and integration by using matrix inversion”, *J. Appl. Math. Comput. Mech.* **13**:2 (2014), 63–71.
- [Meemark and Leela-apiradee 2011] Y. Meemark and W. Leela-apiradee, “A change of basis matrix and integrals of power of cosine”, *J. Statist. Plann. Inference* **141**:3 (2011), 1319–1324. MR Zbl
- [Rogers 1997] J. W. Rogers, Jr., “Applications of linear algebra in calculus”, *Amer. Math. Monthly* **104**:1 (1997), 20–26. MR Zbl

Received: 2017-09-03 Revised: 2017-10-26 Accepted: 2017-12-14

yzm101@yahoo.com

*Department of Mathematics and Computer Science, Faculty
of Science, Chulalongkorn University, Bangkok, Thailand*

songpon@uwm.edu

*Department of Mathematical Sciences, University of
Wisconsin-Milwaukee, Milwaukee, WI, United States*

Patterns in colored circular permutations

Daniel Gray, Charles Lanning and Hua Wang

(Communicated by Joshua Cooper)

Pattern containment and avoidance have been extensively studied in permutations. Recently, analogous questions have been examined for colored permutations and circular permutations. In this note, we explore these problems in colored circular permutations. We present some interesting observations, some of which are direct generalizations of previously established results. We also raise some questions and propose directions for future study.

1. Background

Patterns are essentially subpermutations of a bigger permutation. For two permutations π and τ of lengths n and k with $n \geq k$, we say that π *contains* τ as a pattern if there is a subsequence of entries of π , $(\pi_{i_1}, \pi_{i_2}, \pi_{i_3}, \dots, \pi_{i_k})$, which is order isomorphic to τ ; i.e., $\pi_{i_s} \leq \pi_{i_t}$ if and only if $\tau_s \leq \tau_t$. Such a subsequence is called an *occurrence* of τ in π . If no occurrence of τ is present in π , we say that π *avoids* τ .

Most of the earlier work on patterns concerns *pattern avoidance*; see [Bóna 2012] for a nice introduction. A more comprehensive study of pattern containment was first proposed by H. Wilf in 1992 [Liendo 2012]. There are two natural questions one might ask regarding pattern containment. First, what is the shortest permutation that contains every element in some set of permutations? Second, for a given pattern, in what permutation does this pattern occur the most? The former deals with *superpatterns*, whereas the latter concerns *pattern packing*.

Superpatterns. For a set \mathcal{P} of permutations we say that a permutation π is a \mathcal{P} -superpattern if it contains at least one occurrence of every $\tau \in \mathcal{P}$. We also define

$$\text{sp}(\mathcal{P}) = \min\{n : \text{there is a } \mathcal{P}\text{-superpattern of length } n\}$$

and $\text{sp}(k) = \text{sp}(\mathcal{P})$ when \mathcal{P} is the set of all permutations of length k .

For results on the bounds of $\text{sp}(k)$, see [Arratia 1999; Eriksson et al. 2007; Miller 2009]. Bounds of $\text{sp}(\mathcal{P})$ have also been studied for layered permutations [Gray

MSC2010: primary 05A05; secondary 05A15, 05A16.

Keywords: Circular permutations, patterns.

This work was partially supported by a grant from the Simons Foundation (#245307).

2015], 321-avoiding permutations [Bannister et al. 2014], m -colored permutations [Gray and Wang 2016], and words [Burstein et al. 2002/03].

Pattern packing. Letting $f(\pi, \tau)$ be the number of occurrences of τ in π , we define

$$g(n, \tau) = \max\{f(\sigma, \tau) : \sigma \text{ is a permutation of length } n\}.$$

and the *packing density* of τ as

$$\delta(\tau) = \lim_{n \rightarrow \infty} \frac{g(n, \tau)}{\binom{n}{k}}.$$

A permutation π (of length n) with $f(\pi, \tau) = g(n, \tau)$ is called τ -*optimal*.

For packing densities of length-3 and length-4 patterns, see [Albert et al. 2002; Price 1997; Stromquist 1993]. There are three length-4 patterns whose packing densities remain open, as are any longer nonlayered patterns.

Pattern avoidance. Pattern avoidance has been well-studied for permutations; see [Bóna 2012] for details. In the case of colored permutations, [Mansour 2001] provides a formula for the number of permutations avoiding all length-2 permutations whose entries are colorable in r ways. For circular permutations, [Callan 2002] counts the number of circular permutations avoiding 1324, 1342, and 1234. Both topics are relatively new, and there are still many open questions.

Our contribution. There are two natural variations of permutations, colored permutations and circular permutations, where the first one assigns colors to each entry and the second arranges entries around a circle. In colored permutations, superpatterns [Gray and Wang 2016], pattern packing [Just and Wang 2016], and pattern avoidance [Mansour 2001] have been considered. Noncolored pattern containment [Gray et al. 2017] and pattern avoidance [Callan 2002] have been studied for circular patterns. In this paper, we will consider the combination of these two variations, the colored circular permutations. First, we will introduce the necessary terminology and notation in Section 2. We then discuss “supercolored circular permutations” in Section 3, where we point out that many of the results in [Gray and Wang 2016] can be directly generalized to the colored circular permutations. In Section 4, we discuss pattern packing in colored circular permutations, including some generalizations of results in [Just and Wang 2016]. Lastly, in Section 5 we consider pattern avoidance in colored circular permutations. We conclude our work by commenting on the many remaining problems for future work in Section 6.

2. Terminologies in colored and circular permutations

We start with some formal terminologies and notations for colored permutations and patterns.

Definition 2.1. Let k and m be any positive integers. An m -colored permutation of length k is any permutation of length k where each entry is colored one of m given colors; we allow distinct entries of the permutation to be colored differently. We denote the set of all permutations of length k in m colors by $\mathcal{S}_{k,m}$.

In the case that there are only two or three colors, we will color the entries of a permutation “red”, “green”, or “blue”; thus, we may have the colored permutation $2_r 1_b 3_r$, which denotes the permutation 213 whose first and third entries are colored “red” and whose second entry is colored “blue”. If more than three colors are allowed, we will just label the colors with natural numbers; hence, the colored permutation $1_1 3_4 4_1 5_3 2_2$ is the permutation 13452 whose first and third entries are colored 1, fifth entry is colored 2, fourth entry is colored 3, and second entry is colored 4.

Definition 2.2. Let k and m be any positive integers. A *monochromatic* m -colored permutation of length k is any m -colored permutation for which every entry is colored the same color. We denote the set of all monochromatic m -colored permutations of length k by $\mathcal{M}_{k,m}$.

Definition 2.3. Let k and m be any positive integers. A *nonmonochromatic* m -colored permutation of length k is any m -colored permutation for which there exist at least two distinct entries that are colored differently. We denote the set of all nonmonochromatic m -colored permutations of length k by $\mathcal{N}_{k,m}$.

The union of $\mathcal{N}_{k,m}$ and $\mathcal{M}_{k,m}$ is $\mathcal{S}_{k,m}$, the set of all m -colored permutations of length k . For example, $1_r 2_b 3_b$ is nonmonochromatic since the first entry and second entry are colored differently, while $1_r 2_r 3_r$ and $1_b 2_b 3_b$ are both monochromatic. For comparison, we list $\mathcal{S}_{2,2}$, $\mathcal{N}_{2,2}$, and $\mathcal{M}_{2,2}$ below:

$$\begin{aligned}\mathcal{S}_{2,2} &= \{1_r 2_r, 1_r 2_b, 1_b 2_r, 1_b 2_b, 2_r 1_r, 2_r 1_b, 2_b 1_r, 2_b 1_b\}, \\ \mathcal{N}_{2,2} &= \{1_r 2_b, 1_b 2_r, 2_r 1_b, 2_b 1_r\}, \\ \mathcal{M}_{2,2} &= \{1_r 2_r, 1_b 2_b, 2_r 1_r, 2_b 1_b\}.\end{aligned}$$

Definition 2.4. For colored permutations p and q we say that p contains q as a colored pattern if there is some subsequence of p , say P , which satisfies the following two conditions:

- The i -th entry of P is the same color as the i -th entry of q for all i .
- P is order isomorphic to q .

If there is no such P satisfying both conditions, we say that p *avoids* q as a colored pattern.

We will usually drop the phrase “as a colored pattern” and just say “ p contains q ” when it is obvious that we are dealing with colored permutations. For instance, if

$p = 1_r 3_b 2_r$ and $q = 2_b 1_r$, we see that p contains q since the subsequence $(3_b, 2_r)$ of p satisfies both of the conditions above. However, if $q = 2_r 1_b$ then p avoids q since there is no subsequence of p simultaneously satisfying both conditions.

Similar to the noncolored case, for a collection \mathcal{P} of colored permutations we define the \mathcal{P} -superpattern and $\text{sp}(\mathcal{P})$ accordingly. Note that the permutation $p = 1_r 2_b 6_r 5_b 4_r 3_b$ contains every colored permutation in $\mathcal{S}_{2,2}$. Hence, p is an $\mathcal{S}_{2,2}$ -superpattern. Brute force shows that there is no shorter $\mathcal{S}_{2,2}$ -superpattern; therefore $\text{sp}(\mathcal{S}_{2,2}) = 6$.

Next we formalize the concept of pattern containment/avoidance in circular permutations. Note that the following definition also applies to colored permutations.

Definition 2.5. Let $p = p_1 p_2 \cdots p_n$ be a permutation of length n . The *circular shift* of p , denoted $S(p)$, is given by

$$S(p) = p_n p_1 p_2 \cdots p_{n-1}.$$

If we take a permutation, π , and wrap its entries clockwise around a circle, equally spread out within one revolution, then we have created a circular permutation, π_c . We say that $\pi_c = \tau_c$ if τ is just a cyclic shift of π , i.e., $S^i(\tau) = \pi$ for some i .

Definition 2.6. For colored permutations p and q we say that p contains q *circularly* if p contains $S^i(q)$ as a colored pattern for some nonnegative integer i .

Definition 2.7. Let \mathcal{P} be any collection of permutations. A *circular \mathcal{P} -superpattern* is a permutation which contains every $p \in \mathcal{P}$ as a circular pattern. We let $\text{sp}_c(\mathcal{P})$ denote the length of the shortest circular \mathcal{P} -superpattern. When \mathcal{P} is the set of all circular patterns of length k we simply write $\text{sp}_c(k)$.

A useful concept in the study of pattern packing in colored permutations will be “colored blocks”, which we define below.

Definition 2.8. In a colored permutation π , a *colored block* is a maximal monochromatic segment $\pi_i^{(a)}$ in which every entry in this segment has color a and every entry not in this segment is either larger or smaller than each entry in $\pi_i^{(a)}$.

For example, the permutation $\pi = 1_r 2_r 6_b 5_b 3_b 4_r$ has four colored blocks. From left to right, they are $\pi_1^{(r)} = 1_r 2_r$, $\pi_2^{(b)} = 6_b 5_b$, $\pi_3^{(b)} = 3_b$, $\pi_4^{(r)} = 4_r$. Indeed every colored permutation has a unique decomposition into colored blocks: Given a colored permutation, it can first be decomposed into maximal monochromatic subsequences and it is easy to see that there is a unique way to do this. Within each monochromatic subsequence there is a unique way to separate the entries according to their numerical values.

When comparing the numerical values between different blocks, we say that $\pi_i^{(r)} < \pi_j^{(b)}$ when all entries of $\pi_i^{(r)}$ are less than all entries of $\pi_j^{(b)}$. It is easy to see that this concept generalizes naturally to the circular case.

For colored patterns τ and permutations π we define $f_c(\pi, \tau)$ to be the number of occurrences of τ in π wrapped around a circle; i.e.,

$$f_c(\pi, \tau) = f(\pi, \tau) + f(\pi, S(\tau)) + f(\pi, S^2(\tau)) + \cdots + f(\pi, S^{k-1}(\tau)).$$

Then, similar to before,

$$g_c(n, \tau) = \max\{f_c(\sigma, \tau) : \sigma \text{ is a permutation of length } n\}.$$

If π is of length n and $f_c(\pi, \tau) = g_c(n, \tau)$, then we say that π is *circular τ -optimal*.

Definition 2.9. Let τ be a colored permutation of length k . The *circular packing density* of τ , denoted by $\delta_c(\tau)$, is defined by

$$\delta_c(\tau) = \lim_{n \rightarrow \infty} \frac{g_c(n, \tau)}{\binom{n}{k}}.$$

3. Superpatterns

In this section we consider questions related to superpatterns in colored circular permutations. We note that some of the results in this section are direct generalizations from those in [Gray and Wang 2016]. For this reason some details will be omitted.

Theorem 3.1. *For any positive integers k and m , we have that*

$$\text{sp}_c(\mathcal{S}_{k,m}) = m \text{sp}_c(k).$$

Proof. Let p' be a circular $\mathcal{S}_{k,m}$ -superpattern and p'_i be the longest monochromatic subsequence in p' in color i . It follows that p' is a circular k -superpattern and consequently $|p'_i| \geq \text{sp}_c(k)$ for any $1 \leq i \leq m$. Hence

$$|p'| = \sum_{i=1}^m |p'_i| \geq m \text{sp}_c(k).$$

Now, let p be a circular permutation of length $\text{sp}_c(k)$ that contains all noncolored patterns of length k . Consider the m -colored circular permutation p'' , constructed from p by replacing each $1 \leq j \leq \text{sp}_c(k)$ in p with the sequence

$$s_j = [m(j-1) + 1]_1 [m(j-1) + 2]_2 \cdots [m(j-1) + m]_m.$$

It is easy to see that

$$|p''| = m|p| = m \text{sp}_c(k)$$

and that p'' is a $\mathcal{S}_{k,m}$ -superpattern. Thus,

$$\text{sp}_c(\mathcal{S}_{k,m}) \leq |p''| = m \text{sp}_c(k). \quad \square$$

With Theorem 3.1, we can use previously established results [Gray et al. 2017] on $\text{sp}_c(k)$ to bound $\text{sp}_c(\mathcal{S}_{k,m})$.

Corollary 3.2. *For positive integers k and m we have*

$$\text{sp}_c(\mathcal{S}_{k,m}) = m \text{sp}_c(k) \geq mg(k) \frac{k^2}{e^2},$$

where $g(k) \rightarrow 1$ as $k \rightarrow \infty$, and

$$\text{sp}_c(\mathcal{S}_{k,m}) = m \text{sp}_c(k) \leq m(\text{sp}(k-1) + 1) \leq m\left(\frac{1}{2}k(k-1) + 1\right).$$

Consequently,

$$mg(k) \frac{k^2}{e^2} \leq \text{sp}_c(\mathcal{S}_{k,m}) \leq m\left(\frac{1}{2}k(k-1) + 1\right),$$

where $g(k) \rightarrow 1$ as $k \rightarrow \infty$.

As expected, the bounds for $\text{sp}_c(\mathcal{S}_{k,m})$ are simply m times the bounds for $\text{sp}_c(k)$. Next, we restrict our attention to only monochromatic or nonmonochromatic patterns. First we note the following facts on the sizes of $\mathcal{M}_{k,m}$ and $\mathcal{N}_{k,m}$:

- $|\mathcal{S}_{k,m}| = m^k k!$.
- $|\mathcal{M}_{k,m}| = mk!$.
- $|\mathcal{N}_{k,m}| = |\mathcal{S}_{k,m}| - |\mathcal{M}_{k,m}| = (m^{k-1} - 1)|\mathcal{M}_{k,m}|$.

We now establish a lower bound for $\text{sp}_c(\mathcal{N}_{k,m})$.

Theorem 3.3. *For positive integers k and m ,*

$$\text{sp}_c(\mathcal{N}_{k,m}) \geq mg(k, m) \frac{k^2}{e^2},$$

where $g(k, m) \rightarrow 1$ as $k \rightarrow \infty$.

Proof. Let $n = \text{sp}_c(\mathcal{N}_{k,m})$, and our $\mathcal{N}_{k,m}$ -superpattern of length n must contain a circular shift of every permutation in $\mathcal{N}_{k,m}$. Note that at most k such permutations can be circular shifts of each other; hence at least $|\mathcal{N}_{k,m}|/k$ permutations from $\mathcal{N}_{k,m}$ must be contained in the superpattern. Consequently

$$\binom{n}{k} \geq \frac{|\mathcal{N}_{k,m}|}{k} = \frac{(m^k - m)k!}{k} = (m^k - m)(k-1)!.$$

By the fact $n^k/k! \geq \binom{n}{k}$ and Stirling's approximation $k! \geq \sqrt{2\pi k}(k^k/e^k)$, we have

$$\frac{n^k}{k!} \geq (m^k - m)(k-1)!$$

and hence

$$\begin{aligned} n &\geq \left((m^k - m) \frac{(k!)^2}{k} \right)^{1/k} \geq \left((m^k - m) 2\pi \frac{k^{2k}}{e^{2k}} \right)^{1/k} \\ &= m((1 - m^{-k+1})2\pi)^{1/k} \frac{k^2}{e^2} = mg(k, m) \frac{k^2}{e^2} \end{aligned}$$

with $g(k, m) = ((1 - m^{-k+1})2\pi)^{1/k} \rightarrow 1$ as $k \rightarrow \infty$. □

It is interesting to note that this lower bound is similar to that found for $\text{sp}_c(\mathcal{S}_{k,m})$ in Corollary 3.2. To bound $\text{sp}_c(\mathcal{N}_{k,m})$ from above, first note that a circular $\mathcal{M}_{k,m}$ -superpattern must have m copies of a circular k -superpattern, one for each color. Then, we have

$$\text{sp}_c(\mathcal{N}_{k,m}) \leq \text{sp}_c(\mathcal{S}_{k,m}) = m \text{sp}_c(k) = \text{sp}_c(\mathcal{M}_{k,m}).$$

Given the fact that $|\mathcal{N}_{k,m}| = (m^{k-1} - 1)|\mathcal{M}_{k,m}|$, it is rather surprising that the shortest $\mathcal{N}_{k,m}$ -superpattern is not longer than the shortest $\mathcal{M}_{k,m}$ -superpattern. The following further analyzes the relationship between them.

Theorem 3.4. *For any positive integers $k \geq 2$ and m , we have*

$$\text{sp}_c(\mathcal{M}_{k-1,m}) \leq \text{sp}_c(\mathcal{N}_{k,m}) \leq \text{sp}_c(\mathcal{M}_{k,m}).$$

Proof. The second inequality follows from the discussion above.

On the other hand, let q be an m -colored pattern of length k with all but one entry of color i . For some circular shift of q to be contained in a circular $\mathcal{N}_{k,m}$ -superpattern, a circular shift of the length- $(k-1)$ monochromatic pattern in color i must be contained in the superpattern. Hence all length- $(k-1)$ monochromatic patterns (of any color) must occur in a circular $\mathcal{N}_{k,m}$ -superpattern, and $\text{sp}_c(\mathcal{M}_{k-1,m}) \leq \text{sp}_c(\mathcal{N}_{k,m})$. \square

4. Pattern packing

Our results in this section mainly concern the characteristics of the optimal colored circular permutations when the pattern under consideration is described through colored blocks. Again, some of our results here are direct generalizations of those in noncircular case [Just and Wang 2016], for which reason we skip some details.

In the case of having only two colored blocks, we can see that a pattern must be of the form $\pi = \pi_1 \pi_2$ with π_1 in red and π_2 in blue. We will assume, without loss of generality, that $\pi_1 < \pi_2$. In this case, we may simply say that the pattern is of the form rb with $r < b$, and similarly for patterns with more colored blocks.

Theorem 4.1. *For a pattern ρ with two colored blocks of the form rb with $r < b$, there is an optimal circular permutation π of the form RB with $R < B$.*

Proof. Let π be a ρ -optimal permutation of length n with colored blocks $\pi_1 \pi_2 \cdots \pi_k$. We can assume without loss of generality that π_1 is red.

Now, let us take all the red blocks $\pi_{r_1} \pi_{r_2} \cdots \pi_{r_s}$ and blue blocks $\pi_{b_1} \pi_{b_2} \cdots \pi_{b_t}$, and form a new circular permutation $\pi' = \pi_{r_1} \cdots \pi_{r_s} \pi_{b_1} \cdots \pi_{b_t}$. It is easy to see that any occurrence of ρ in π is also in π' .

Next, since ρ is of the form rb with $r < b$, we claim that, in our optimal permutation π' , every red entry must be (numerically) less than every blue entry. Otherwise, one may always “rearrange” the numerical values so that the numerical

ordering stays the same among entries of the same color, and so that all red entries are smaller than the blue ones. The resulting permutation can only contain more occurrences of ρ .

Consequently, all red blocks together simply form a single block in π and so do the blue blocks. Our conclusion, then, follows. \square

Next, we consider patterns with three colored blocks. Note that for circular patterns with three colored blocks and two colors, rb_1b_2 with $b_1 < r < b_2$ is the only case that we needed to investigate: With three colored blocks and two colors there are always one block with one color (say red) and two blocks with the other (say blue). One of the circular shifts of this pattern must be of the form rb_1b_2 . By taking a circular shift of the reversed pattern (i.e., rb_2b_1) if necessary, we may also assume that $b_1 < r < b_2$.

Theorem 4.2. *For a pattern ρ with three colored blocks of the form rb_1b_2 and $b_1 < r < b_2$, there is an optimal circular permutation π of the same form.*

Proof. Let π be a ρ -optimal circular permutation of length n . First, we will show that we can put all blue blocks in increasing order of their numerical values and next to each other. Let $\pi_{r_1}, \pi_{r_2}, \dots, \pi_{r_s}$ be the red blocks of π .

Now, for an occurrence of ρ in π , suppose R_ρ (in π) is the part corresponding to r (in ρ). Let $\pi_{b < R}$ be the collection of all blue blocks (with numerical value) less than R_ρ , and let $\pi_{b > R_\rho}$ be the set of all blue blocks greater than R_ρ . Then, any occurrence of rb_1b_2 with $r \sim R_\rho$ must have b_1 occurring in $\pi_{b < R_\rho}$ and b_2 occurring in $\pi_{b > R_\rho}$. The maximum number of such occurrences (i.e., the maximum possible contribution of R_ρ to $f_c(\pi, \rho)$) is

$$f(\pi_{b < R_\rho}, b_1) f(\pi_{b > R_\rho}, b_2).$$

As far as the ordering of the blue blocks is concerned, arranging the blue blocks in increasing order achieves the above maximum. At this point it is also easy to see that putting blocks of the same color together will not reduce the number of occurrences of ρ . Denote such an optimal permutation by $\pi' = \pi_{r_1} \cdots \pi_{r_s} \pi_{b_1} \cdots \pi_{b_t}$, with $\pi_{b_i} < \pi_{b_{i+1}}$ for any $1 \leq i \leq t-1$.

Next, we show that all red entries form a single block, or equivalently, the numerical value of any red block is between those of $\pi_{b_{j_0-1}}$ and $\pi_{b_{j_0}}$ for some fixed j_0 . Let $\pi_{b \geq j}$ be the collection of blue blocks $\pi_{b_j}, \pi_{b_{j+1}}, \dots, \pi_{b_t}$, and let $\pi_{b < j}$ be the collection of blocks $\pi_{b_1}, \dots, \pi_{b_j}$. Then, there must exist some j_0 that maximizes the occurrences of b_1 in $\pi_{b < j}$ and b_2 in $\pi_{b \geq j}$. In other words,

$$f(\pi_{b < j_0}, b_1) f(\pi_{b \geq j_0}, b_2) \geq f(\pi_{b < j}, b_1) f(\pi_{b \geq j}, b_2)$$

for any $1 < j \leq t$. So,

$$f_c(\pi', \rho) \leq f(\pi_{r_1} \cdots \pi_{r_s}, r) f(\pi_{b < j_0}, b_1) f(\pi_{b \geq j_0}, b_2)$$

with equality when $\pi_{b_{j_0-1}} < \pi_{r_i} < \pi_{b_{j_0}}$ for any $1 \leq i \leq s$. From this it follows that there are exactly one single red block and two blue blocks in $\pi = RB_1B_2$ with $B_1 < R < B_2$. \square

It remains to consider the case when we have three colored blocks in three different colors, i.e., the pattern rbg with $r < b < g$.

Theorem 4.3. *For a pattern ρ of the form rbg with $r < b < g$, there is an optimal circular permutation π of the same form.*

Proof. Let π be a ρ -optimal circular permutation of length n with R' , B' and G' being the collections of all red blocks (in their original order), blue blocks and green blocks respectively. An occurrence of ρ in π must consist of an occurrence of r in R' , an occurrence of b in B' , and an occurrence of g in G' . Hence

$$f(\pi, \rho) \leq f(R', r)f(B', b)f(G', g)$$

with equality if each of R' , B' and G' is a single block, arranged in this order, and $R' < B' < G'$. \square

To summarize the above observations, we have the following.

Corollary 4.4. *For any circular pattern with two or three colored blocks, there is a corresponding optimal circular permutation of the same form.*

Remark 4.5. After seeing the above results on patterns with two or three colored blocks, it is natural to guess that the same holds for patterns with more blocks. Consequently one can ask if there is always an optimal circular permutation of the same form as the pattern. We have not been able to prove either way.

On the other hand, considering $\rho = 1_r 2_b$, it is not hard to check that $\pi = (1_r 3_b 2_r 4_b)_c$ is an optimal length-4 circular permutation for ρ . Thus, there does exist an optimal permutation that is not of the form RB with $R < B$. Evidence seems to suggest that this is the only such case.

5. Pattern avoidance

The numbers of m -colored (noncircular) permutations that avoid one or two 2-letter patterns were presented in [Mansour 2001], together with some discussion of the connection between pattern avoidance in noncolored permutations and colored permutations. In [Callan 2002], pattern avoidance in circular permutations was studied. It was pointed out that, when considered as circular permutations, none avoid any 2-letter patterns and the identity (reverse identity) is the only one avoiding the pattern 132 (123). In this section we extend this study to colored circular permutations, generalizing a little of both [Callan 2002] and [Mansour 2001].

Avoiding a monochromatic length-2 pattern in $\mathcal{S}_{k,m}$. Without loss of generality, we may assume the monochromatic length-2 pattern is $1_1 2_1$. Then, to avoid such a pattern, our permutation π in $\mathcal{S}_{k,m}$ can contain at most one entry of color 1. Consider two cases:

- There is no entry with color 1 in π . Then, there are a total of $(k-1)!$ noncolored circular permutations of length k , and each of the k entries has $m-1$ choices of colors (i.e., the colors $2, 3, \dots, m$); thus the number of such permutations is

$$(k-1)! (m-1)^k.$$

- There is exactly one entry with color 1. Out of the k entries $1, 2, \dots, k$ there are k choices for this particular entry of color 1. There are still $(k-1)!$ ways to wrap the k entries (regardless of their colors) around a circle. Now for each of the remaining $k-1$ entries that are not of color 1, there are $m-1$ choices of colors. Hence the number of such permutations is

$$k! (m-1)^{k-1}.$$

Consequently, we have the following.

Theorem 5.1. *The number of circular permutations in $\mathcal{S}_{k,m}$ that avoid a given monochromatic length-2 pattern is*

$$(k-1)! (m-1)^k + k! (m-1)^{k-1} = (k-1)! (m-1)^{k-1} (k+m).$$

Avoiding nonmonochromatic length-2 pattern in $\mathcal{S}_{k,m}$. Again, without loss of generality, let us assume this pattern to be $1_1 2_2$. For a circular permutation π in $\mathcal{S}_{k,m}$, let E_1 and E_2 be the sets of entries in π that are colored 1 and 2 respectively. It is easy to see that a $1_1 2_2$ pattern will occur if there is any entry in E_1 that is of smaller numerical value than one in E_2 . Thus, all entries in E_1 are larger than those in E_2 . Suppose $|E_1 \cup E_2| = i$ for some $0 \leq i \leq k$. Then, there are $i+1$ ways to partition the entries into E_1 and E_2 (i.e., to find a $j = 0, 1, \dots, i$ such that the smallest j entries are colored 2 and the rest are colored 1).

Thus, still with $(k-1)!$ ways to wrap all entries around a circle, there are $\binom{k}{i}$ ways to choose entries of $E_1 \cup E_2$. After identifying j there are $(m-2)^{k-i}$ ways to color the remaining entries. Consequently the number of such permutations is

$$\begin{aligned} & (k-1)! \sum_{i=0}^k \left((m-2)^{k-i} (i+1) \binom{k}{i} \right) \\ &= (k-1)! \sum_{i=1}^k \left((m-2)^{k-i} i \binom{k}{i} \right) + (k-1)! \sum_{i=0}^k \left((m-2)^{k-i} \binom{k}{i} \right) \\ &= (k-1)! \sum_{i=1}^k \left((m-2)^{(k-1)-(i-1)} k \binom{k-1}{i-1} \right) + (k-1)! \sum_{i=0}^k \left((m-2)^{k-i} \binom{k}{k-i} \right) \end{aligned}$$

$$\begin{aligned}
&= k!((m-2)+1)^{k-1} + (k-1)!((m-2)+1)^k \\
&= (k-1)!(m-1)^k + k!(m-1)^{k-1} = (k-1)!(m-1)^{k-1}(k+m).
\end{aligned}$$

As a result we have the following.

Theorem 5.2. *The number of circular permutations in $S_{k,m}$ that avoid a given nonmonochromatic length-2 pattern is*

$$(k-1)!(m-1)^{k-1}(k+m).$$

Avoiding monochromatic patterns of length 3 in $S_{k,m}$. All circular permutations of length 3 are equivalent under circular shift and reverse. So we will only consider, without loss of generality, the pattern $1_1 3_1 2_1$. It is known that in the noncolored case the only circular permutation that avoids 132 is the identity permutation.

Let π be a permutation that avoids $1_1 3_1 2_1$ in $S_{k,m}$, and let E_1 (of cardinality $i = 0, 1, \dots, k$) be the set of entries of color 1, then:

- If $i \geq 3$, there is only one way to order the entries in E_1 (i.e., in increasing order). Starting with $(k-1)!$ ways to wrap the k entries (regardless of color) around a circle, only one of the $(i-1)!$ orderings of the entries in E_1 can be chosen. Noting that there are $\binom{k}{i}$ ways to pick the numerical values of the entries in E_1 and $(m-1)^{k-i}$ ways to assign colors to the remaining entries, we have the number of such permutations as

$$\sum_{i=3}^k \left(\binom{k}{i} \frac{(k-1)!}{(i-1)!} (m-1)^{k-i} \right).$$

- If $i \leq 2$, then there are $\binom{k}{i}$ ways to pick these i entries, $(k-1)!$ ways to wrap all the entries around the circle and $(m-1)^{k-i}$ ways to color the other entries. The number of such permutations is

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

We may combine the above two formulas and conclude the following.

Theorem 5.3. *The number of circular permutations in $S_{k,m}$ that avoid a given monochromatic length-3 pattern is*

$$(k-1)!(m-1)^k + \sum_{i=1}^k \left(\binom{k}{i} \frac{(k-1)!}{(i-1)!} (m-1)^{k-i} \right).$$

Wilf classes. Theorems 5.1 and 5.2 imply that for any m -colored pattern of length 2, say ϕ , the number of ϕ -avoiding circular permutations in $S_{k,m}$ is $(k-1)!(m-1)^{k-1}(k+m)$. This interesting (and perhaps a little surprising) observation is analogous to the findings in [Mansour 2001] in noncircular case.

This also implies that there is only one Wilf class of colored circular permutations when restricted by one pattern of length 2.

6. Concluding remarks and additional questions

In this short note, we considered questions related to superpatterns, pattern packing, and pattern avoidance in colored circular permutations. We presented some elementary observations, especially those generalized from previously established results on colored (but not circular) permutations, on each of these three questions. Many interesting questions remain to be further explored.

In Section 3, we introduced generalizations of a few facts on colored superpatterns. The arguments of these generalizations follow from direct adjustment of those in [Gray and Wang 2016]. There are, however, also some constructive proofs that cannot be directly generalized to circular cases. It would be interesting to further investigate them.

The several theorems in Section 4 claim that for patterns with two or three colored blocks their corresponding optimal colored circular permutations include at least one with exactly the same format (in terms of the colored blocks). With these facts, one may easily calculate the packing densities of various patterns. It is not clear whether this is true for more colored blocks. It is also mentioned that, for the pattern $\rho = 1_r 2_b$, there exist optimal permutations that have a different format. It seems likely that this is the only such case, though we do not have a proof yet.

The numbers of permutations avoiding various given patterns, as studied in Section 5, lead to an interesting statement on the Wilf classes of colored circular permutations restricted by 2-letter patterns. It appears to be much more complicated to examine the same problem for colored circular permutations restricted by longer patterns or more than one 2-letter patterns.

References

- [Albert et al. 2002] M. H. Albert, M. D. Atkinson, C. C. Handley, D. A. Holton, and W. Stromquist, “On packing densities of permutations”, *Electron. J. Combin.* **9**:1 (2002), art. id. R5. MR Zbl
- [Arratia 1999] R. Arratia, “On the Stanley–Wilf conjecture for the number of permutations avoiding a given pattern”, *Electron. J. Combin.* **6** (1999), art. id. N1. MR Zbl
- [Bannister et al. 2014] M. J. Bannister, W. E. Devanny, and D. Eppstein, “Small superpatterns for dominance drawing”, pp. 92–103 in *2014 Proceedings of the Eleventh Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*, edited by M. Drmota and M. D. Ward, SIAM, Philadelphia, PA, 2014. MR
- [Bóna 2012] M. Bóna, *Combinatorics of permutations*, 2nd ed., CRC Press, Boca Raton, FL, 2012. MR Zbl
- [Burstein et al. 2002/03] A. Burstein, P. Hästö, and T. Mansour, “Packing patterns into words”, *Electron. J. Combin.* **9**:2 (2002/03), art. id. R20. MR Zbl
- [Callan 2002] D. Callan, “Pattern avoidance in circular permutations”, preprint, 2002. arXiv

- [Eriksson et al. 2007] H. Eriksson, K. Eriksson, S. Linusson, and J. Wästlund, “Dense packing of patterns in a permutation”, *Ann. Comb.* **11**:3-4 (2007), 459–470. MR Zbl
- [Gray 2015] D. Gray, “Bounds on superpatterns containing all layered permutations”, *Graphs Combin.* **31**:4 (2015), 941–952. MR Zbl
- [Gray and Wang 2016] D. Gray and H. Wang, “Note on superpatterns”, *Involve* **9**:5 (2016), 797–804. MR Zbl
- [Gray et al. 2017] D. Gray, C. Lanning, and H. Wang, “Pattern containment in circular permutations”, preprint, 2017. To appear in *Integers*.
- [Just and Wang 2016] M. Just and H. Wang, “Note on packing patterns in colored permutations”, *Online J. Anal. Comb.* **11** (2016), art. id. 4. MR Zbl
- [Liendo 2012] M. L. Liendo, *Preferential arrangement containment in strict superpatterns*, master’s thesis, East Tennessee State University, 2012, available at <https://dc.etsu.edu/etd/1428/>.
- [Mansour 2001] T. Mansour, “Pattern avoidance in coloured permutations”, *Sém. Lothar. Combin.* **46** (2001), art. id. B46g. MR Zbl
- [Miller 2009] A. Miller, “Asymptotic bounds for permutations containing many different patterns”, *J. Combin. Theory Ser. A* **116**:1 (2009), 92–108. MR Zbl
- [Price 1997] A. L. Price, *Packing densities of layered patterns*, Ph.D. thesis, University of Pennsylvania, 1997, available at <https://search.proquest.com/docview/304421853>. MR
- [Stromquist 1993] W. Stromquist, “Packing layered posets into posets”, preprint, 1993, available at <http://walterstromquist.com/papers/POSETS.DOC>.

Received: 2017-11-29 Revised: 2018-01-21 Accepted: 2018-02-14

dagray@georgiasouthern.edu *Department of Mathematical Sciences,
Georgia Southern University, Statesboro, GA, United States*

lannin3@clemson.edu *Department of Mathematical Sciences, Clemson University,
Clemson, SC, United States*

hwang@georgiasouthern.edu *Department of Mathematical Sciences,
Georgia Southern University, Statesboro, GA, United States*

Solutions of boundary value problems at resonance with periodic and antiperiodic boundary conditions

Aldo E. Garcia and Jeffrey T. Neugebauer

(Communicated by Johnny Henderson)

We study the existence of solutions of the second-order boundary value problem at resonance $u'' = f(t, u, u')$ satisfying the boundary conditions $u(0) + u(1) = 0$, $u'(0) - u'(1) = 0$, or $u(0) - u(1) = 0$, $u'(0) + u'(1) = 0$. We employ a shift method, making a substitution for the nonlinear term in the differential equation so that these problems are no longer at resonance. Existence of solutions of equivalent boundary value problems is obtained, and these solutions give the existence of solutions of the original boundary value problems.

1. Introduction

Consider the second-order boundary value problem

$$u'' = f(t, u, u'), \quad t \in (0, 1), \quad (1-1)$$

satisfying a combination of antiperiodic and periodic boundary conditions; either

$$u(0) + u(1) = 0, \quad u'(0) - u'(1) = 0. \quad (1-2)$$

or

$$u(0) - u(1) = 0, \quad u'(0) + u'(1) = 0. \quad (1-3)$$

Here we assume $f(t, x, y) : [0, 1] \times \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ is continuous in each of its variables.

Since the boundary value problem $u'' = 0$, (1-2) has the nontrivial solution $u(t) = t - \frac{1}{2}$, the problem (1-1), (1-2) is said to be at resonance. Similarly, since $u'' = 0$, (1-3) has the nontrivial solution $u(t) \equiv 1$, the problem (1-1), (1-3) is also at resonance. Hence, standard methods employing Green's functions cannot be used to show the existence of solutions of these boundary value problems directly. Thus, we consider a shifted boundary value problem so that Green's functions can be employed.

MSC2010: primary 34B15; secondary 34B27.

Keywords: boundary value problem, resonance, shift.

Han [2007] employed a shift argument when studying a three-point boundary value problem

$$\begin{aligned}x''(t) &= f(t, x(t)), \quad t \in (0, 1), \\x'(0) &= 0, \quad x(\eta) = x(1).\end{aligned}$$

Here it was assumed $g(t, x) = f(t, x) + \beta^2 x$ and the equivalent boundary value problem

$$x''(t) + \beta^2 x(t) = g(t, x(t)), \quad x'(0) = 0, \quad x(\eta) = x(1),$$

was studied using the Krasnosel'skii–Guo fixed point theorem [Krasnosel'skii 1964].

Infante, Pietramala, and Tojo [Infante et al. 2016] also employed a shift argument when studying Neumann boundary value problems at resonance

$$\begin{aligned}u''(t) + h(t, u(t)) &= 0, \quad t \in (0, 1), \\u'(0) &= u'(1) = 0.\end{aligned}$$

They assumed $f(t, u) = h(t, u) + \omega^2 u$ and considered the equivalent boundary value problem

$$-u''(t) + \omega^2 u(t) = f(t, u(t)), \quad u'(0) = u'(1) = 0.$$

The Krasnosel'skii–Guo fixed point theorem was also used in their analysis.

More recently, Almansour and Eloë [2015] and Al Mosa and Eloë [2016] studied two-point boundary value problems

$$\begin{aligned}y''(t) &= f(t, y(t)), \quad t \in [0, 1], \\y'(0) &= 0, \quad y'(1) = 0,\end{aligned}$$

and

$$\begin{aligned}y''(t) &= f(t, y(t), y'(t)), \quad t \in [0, 1], \\y(0) &= 0, \quad y'(0) = y'(1),\end{aligned}$$

using shift arguments and the Krasnosel'skii–Guo fixed point theorem, the Schauder fixed point theorem, the Leray–Schauder nonlinear alternative [Zeidler 1990] in the former, and monotone methods coupled with upper and lower solutions in the latter.

When considering the first boundary value problem, they assumed $g(t, y) = f(t, y) + \beta^2 y$ and studied the equivalent boundary value problem

$$y''(t) + \beta^2 y(t) = g(t, y(t)), \quad y'(0) = y'(1) = 0,$$

and when considering second, they assumed $g(t, x, y) = f(t, x, y) + \beta y$ and studied the equivalent boundary value problem

$$y''(t) + \beta y'(t) = g(t, y(t), y'(t)), \quad y(0) = 0, \quad y'(0) = y'(1).$$

Here, we make use of two substitutions, one of which has not been used previously in the literature. In Section 2, we study solutions of (1-1), (1-2) by employing the substitution $g(t, x, y) := f(t, x, y) + \beta y$. The shifted boundary value problem is no longer at resonance, and so a Green's function can be constructed. An appropriate integral operator is defined and fixed point methods are used to show the existence of solutions. In Section 3, we study solutions of (1-1), (1-3). The substitutions mentioned above do not help because in both cases the shifted boundary value problem is still at resonance. Thus, we use the substitution $k(t, x, y) = f(t, x, y) + 2\alpha y + (\alpha^2 + \beta^2)x$. This substitution has not been used in the prior literature. A similar approach to that in Section 2 is then used to show existence of solutions. The construction of the two Green's functions and the shift employed in Section 3 can both lead to more research in this area.

2. Solutions of (1-1), (1-2)

Notice that for $\beta > 0$, $\beta \neq n\pi$, $n \in \mathbb{N}$, the boundary value problem $u'' + \beta^2 u = 0$, (1-2) is at resonance, since $u(t) = \cos \beta t - ((1 + \cos \beta)/\sin \beta) \sin \beta t$ is a nontrivial solution. If $\beta = n\pi$, $n \in \mathbb{N}$, then $u(t) = \sin \beta t$ is a nontrivial solution of the boundary value problem. Thus the substitution $g(t, x, y) = f(t, x, y) + \beta^2 x$ cannot be applied.

Let $\beta > 0$ be a constant and assume $g(t, x, y) := f(t, x, y) + \beta y$. We study the shifted differential equation

$$u'' + \beta u' = g(t, u, u'), \quad t \in (0, 1), \quad (2-1)$$

satisfying boundary conditions (1-2). The boundary value problem (2-1), (1-2) is not at resonance, since the unique solution of $u'' + \beta u' = 0$, (1-2), is $u \equiv 0$. Notice if $u(t)$ is a solution of (2-1), (1-2), then

$$u''(t) = g(t, u(t), u'(t)) - \beta u'(t) = f(t, u(t), u'(t)),$$

implying u is a solution of (1-1), (1-2).

We first construct the Green's function associated with $u'' + \beta u' = 0$, (1-2).

Lemma 2.1. *Let $h(t)$ be a continuous function. Then $u(t)$ is the unique solution of the boundary value problem*

$$u'' + \beta u' = h(t), \quad t \in (0, 1), \quad (2-2)$$

satisfying boundary conditions (1-2) if and only if

$$u(t) = \int_0^1 G(t, s) h(s) ds,$$

where

$$G(t, s) = \frac{1}{2\beta(1-e^{-\beta})} \begin{cases} 2e^{-\beta(1-s)} - 2e^{-\beta} e^{-\beta(t-s)} + e^{-\beta} - 1, & 0 \leq t \leq s \leq 1, \\ 2e^{-\beta(1-s)} - 2e^{-\beta(t-s)} - e^{-\beta} + 1, & 0 \leq s \leq t \leq 1. \end{cases} \quad (2-3)$$

Proof. Using Laplace transforms, one can show the general solution of (2-2) is given by

$$u(t) = c_1 + c_2 e^{-\beta t} + \frac{1}{\beta} \int_0^t (1 - e^{-\beta(t-s)}) h(s) ds.$$

Since $u'(0) - u'(1) = 0$, we have

$$-c_2 \beta + c_2 \beta e^{-\beta} - \int_0^1 e^{-\beta(1-s)} h(s) ds = 0.$$

Solving for c_2 gives

$$c_2 = -\frac{1}{\beta(1 - e^{-\beta})} \int_0^1 e^{-\beta(1-s)} h(s) ds.$$

The boundary condition $u(0) + u(1) = 0$ gives

$$c_1 + c_2 + c_1 + c_2 e^{-\beta} + \frac{1}{\beta} \int_0^1 (1 - e^{-\beta(t-s)}) h(s) ds = 0.$$

By substituting c_2 from above, solving for c_1 , and simplifying, we have

$$c_1 = \frac{1}{2\beta(1 - e^{-\beta})} \int_0^1 (-1 + e^{-\beta} + 2e^{-\beta(1-s)}) h(s) ds.$$

Thus

$$\begin{aligned} u(t) &= \frac{1}{2\beta(1 - e^{-\beta})} \int_0^1 (-1 + e^{-\beta} + 2e^{-\beta(1-s)}) h(s) ds \\ &\quad - \frac{e^{-\beta t}}{\beta(1 - e^{-\beta})} \int_0^1 e^{-\beta(1-s)} h(s) ds + \frac{1}{\beta} \int_0^t (1 - e^{-\beta(t-s)}) h(s) ds \\ &= \int_0^1 G(t, s) h(s) ds, \end{aligned}$$

where

$$G(t, s) = \begin{cases} \frac{-1 + e^{-\beta} + 2e^{-\beta(1-s)}}{2\beta(1 - e^{-\beta})} - \frac{e^{-\beta t} e^{-\beta(1-s)}}{\beta(1 - e^{-\beta})}, & 0 \leq t \leq s \leq 1, \\ \frac{-(1 - e^{-\beta}) + 2e^{-\beta(t-s)}}{2\beta(1 - e^{-\beta})} - \frac{e^{-\beta t} e^{-\beta(1-s)}}{\beta(1 - e^{-\beta})} + \frac{1 - e^{-\beta(t-s)}}{\beta}, & 0 \leq s \leq t \leq 1. \end{cases}$$

Simplifying $G(t, s)$ gives (2-3).

The reverse direction of the proof can be shown by direct computation. \square

Notice that

$$\frac{\partial}{\partial t} G(t, s) = \frac{1}{1 - e^{-\beta}} \begin{cases} e^{-\beta} e^{-\beta(t-s)}, & 0 \leq t \leq s \leq 1, \\ e^{-\beta(t-s)}, & 0 \leq s \leq t \leq 1. \end{cases} \quad (2-4)$$

We point out several properties of the Green's function.

Lemma 2.2. $G(t, s)$ satisfies the following properties:

- (1) $G \in C([0, 1] \times [0, 1])$.
- (2) $G(0, s) = -\frac{1}{2\beta} < 0$ for all $s \in [0, 1]$.
- (3) $G(1, s) = \frac{1}{2\beta} > 0$ for all $s \in [0, 1]$.
- (4) $\frac{\partial}{\partial t} G(t, s) > 0$ for all $(t, s) \in [0, 1] \times [0, 1]$.
- (5) $\max_{t \in [0, 1]} |G(t, s)| = \frac{1}{2\beta}$ for all $s \in [0, 1]$.
- (6) $\max_{t \in [0, 1]} \frac{\partial}{\partial t} G(t, s) \leq \frac{1}{1-e^{-\beta}}$ for all $s \in [0, 1]$.
- (7) $\max_{t \in [0, 1]} \int_0^1 |G(t, s)| ds \leq \frac{(4+\beta)e^\beta + \beta - 4}{2\beta^2(e^\beta - 1)}$.
- (8) $\max_{t \in [0, 1]} \int_0^1 \frac{\partial}{\partial t} G(t, s) ds = \frac{1}{\beta}$.

All of these properties can be shown directly, so a proof is not given. We point out that property (8) is obtained by making all the terms in $G(t, s)$ positive, integrating, and finding an upper bound when $t \in [0, 1]$.

We employ Schauder's fixed point theorem in our analysis. Because of the fact that $G(t, s)$ changes sign, many fixed point theorems using cones cannot be used.

Theorem 2.3 (Schauder fixed point theorem [Hale and Verduyn Lunel 1993]). *If \mathcal{M} is a closed, bounded, convex subset of a Banach space \mathcal{B} and $T : \mathcal{M} \rightarrow \mathcal{M}$ is completely continuous, then T has a fixed point in \mathcal{M} .*

Let $\mathcal{B} = C^{(1)}[0, 1]$ be the Banach space of functions whose first derivatives are continuous endowed with the norm

$$\|u\| = \max\{|u|_0, |u'|_0\},$$

where $|u|_0 = \max_{t \in [0, 1]} |u(t)|$. Let $M > 0$. Define $\mathcal{M} = \{u \in \mathcal{B} : \|u\| \leq M\}$. Notice that \mathcal{M} is a closed, bounded, convex subset of \mathcal{B} .

Define the operator $T : \mathcal{B} \rightarrow \mathcal{B}$ by

$$Tu(t) = \int_0^1 G(t, s) g(s, u(s), u'(s)) ds.$$

Thus if u is a fixed point of T , then u is a solution of (2-1), (1-2). A standard application of the Arzelà–Ascoli theorem gives us that T is completely continuous.

Define

$$\max_{t \in [0, 1]} \int_0^1 |G(t, s)| ds := \bar{G} \quad \text{and} \quad \max_{t \in [0, 1]} \int_0^1 \frac{\partial}{\partial t} G(t, s) ds := \bar{G}'.$$

Theorem 2.4. Assume $f(t, x, y)$ is continuous in $[0, 1] \times \mathbb{R} \times \mathbb{R}$ with

$$|f(t, x, y) + \beta y| \leq \min \left\{ \frac{M}{G}, \frac{M}{G'} \right\}$$

for all $(t, x, y) \in [0, 1] \times [-M, M] \times [-M, M]$. Then (1-1), (1-2) has a solution $u^* \in \mathcal{M}$.

Proof. Since $g(t, x, y) = f(t, x, y) + \beta y$,

$$|g(t, x, y)| \leq \min \left\{ \frac{M}{G}, \frac{M}{G'} \right\}$$

for all $(t, x, y) \in [0, 1] \times [-M, M] \times [-M, M]$.

Now, for $u \in \mathcal{M}$,

$$|Tu(t)| \leq \int_0^1 |G(t, s)| |g(s, u(s), u'(s))| ds \leq \frac{M}{G} \int_0^1 |G(t, s)| ds = M,$$

$$|(Tu)'(t)| \leq \int_0^1 \frac{\partial}{\partial t} G(t, s) |g(s, u(s), u'(s))| ds \leq \beta M \int_0^1 \frac{\partial}{\partial t} G(t, s) ds = M.$$

So $\|Tu\| \leq M$, and $T : \mathcal{M} \rightarrow \mathcal{M}$. Thus T has a fixed point $u^* \in \mathcal{M}$ which is a solution of (2-1), (1-2). Therefore, u^* is a solution of (1-1), (1-2). \square

Example 2.5. Define

$$f(t, x, y) = \frac{5x^2t^2}{y^2 + 2} - 5y.$$

Let $\beta = 5$. Then from Lemma 2.2

$$\min \left\{ \frac{M}{G}, \frac{M}{G'} \right\} \leq \min \left\{ \frac{2\beta^2(e^\beta - 1)}{(4 + \beta)e^\beta + \beta - 4} M, \beta M \right\} = 5M.$$

So

$$|f(t, x, y) + 5y| = \frac{5x^2t^2}{y^2 + 2} \leq 5M^2 \leq 5M$$

if $M \leq 1$. So the boundary value problem

$$\begin{aligned} u'' &= \frac{5u^2t^2}{(u')^2 + 2} - 5u', \quad t \in (0, 1), \\ u(0) + u(1) &= 0, \quad u'(0) - u'(1) = 0, \end{aligned}$$

has a solution u^* with $\|u^*\| \leq 1$.

3. Solutions of (1-1), (1-3)

For $\beta > 0$, the boundary value problem $u'' + \beta^2 u = 0$, (1-3) is at resonance, since

$$u(t) = \cos \beta t - \left(\frac{1 - \cos \beta}{\sin \beta} \right) \sin \beta t$$

gives a nontrivial solution. If $\beta = n\pi$, $n \in \mathbb{N}$, then $u(t) = \cos \beta t$ is a nontrivial solution of the boundary value problem. Thus the substitution $k(t, x, y) = f(t, x, y) + \beta^2 x$ cannot be applied. Also, the boundary value problem $u'' + \beta u' = 0$, (1-3) is at resonance, since $u(t) \equiv 1$ gives a nontrivial solution. This implies the substitution $k(t, x, y) = f(t, x, y) + \beta^2 y$ cannot be used. Thus, neither substitution used in previous literature can be employed.

Let $\alpha > 0$, $\beta \in (0, \frac{\pi}{2})$ and define

$$k(t, x, y) = f(t, x, y) + 2\alpha y + (\alpha^2 + \beta^2)x.$$

Here we consider the equivalent boundary value problem

$$u'' + 2\alpha u' + (\alpha^2 + \beta^2)u = k(t, u, u'), \quad t \in (0, 1), \quad (3-1)$$

satisfying boundary conditions (1-3), which is not at resonance, since the unique solution of $u'' + 2\alpha u' + (\alpha^2 + \beta^2)u = 0$, (1-3) is $u \equiv 0$. If u is a solution of (3-1), (1-3), then u is a solution of (1-1), (1-3).

Again, we construct a corresponding Green's function.

Lemma 3.1. *The unique solution of*

$$u'' + 2\alpha u' + (\alpha^2 + \beta^2)u = h(t), \quad t \in (0, 1), \quad (3-2)$$

satisfying the boundary conditions (1-3) is given by

$$u(t) = \int_0^1 H(t, s) h(s) ds,$$

where

$$H(t, s) = \frac{1}{2\beta(\beta \sinh \alpha - \alpha \sin \beta)} \Psi(t, s), \quad (3-3)$$

with

$$\Psi(t, s) = \begin{cases} e^{-\alpha(t-s)} [-\beta e^{-\alpha} \sin(\beta(s-t)) + 2\alpha \sin(\beta(1-s)) \sin(\beta t) \\ \quad - \beta \sin(\beta t) \cos(\beta(1-s)) + \beta \cos(\beta t) \sin(\beta(1-s))], & 0 \leq t \leq s \leq 1, \\ e^{-\alpha(t-s)} [\beta e^{\alpha} \sin(\beta(t-s)) + 2\alpha \sin(\beta s) \sin(\beta(1-t)) \\ \quad - \beta \sin(\beta s) \cos(\beta(1-t)) + \beta \cos(\beta s) \sin(\beta(1-t))], & 0 \leq s \leq t \leq 1. \end{cases}$$

Proof. If u satisfies (3-2), then, using Laplace transforms,

$$u(t) = e^{-\alpha t} (c_1 \cos(\beta t) + c_2 \sin(\beta t)) + \frac{1}{\beta} \int_0^t (e^{-\alpha(t-s)} \sin(\beta(t-s))) h(s) ds.$$

Solving the system $u(0) - u(1) = 0$, $u'(0) + u'(1) = 0$ gives

$$c_1 = -\frac{1}{2\alpha e^{-\alpha} \sin(\beta) + \beta e^{-2\alpha} - \beta} \times \left[\int_0^1 [e^{-\alpha(1-s)} \sin(\beta(1-s)) - e^{-\alpha(2-s)} \sin(\beta s)] h(s) ds \right],$$

$$c_2 = -\frac{1}{\beta e^{-\alpha} \sin(\beta) (2\alpha e^{-\alpha} \sin(\beta) + \beta e^{-2\alpha} - \beta)} \times \left[\int_0^1 \left[\beta e^{-\alpha(3-s)} [\cos(\beta) \sin(\beta s) + \sin(\beta(1-s))] - e^{-\alpha(2-s)} [\beta \cos(\beta) \sin(\beta(1-s)) + \beta \sin(\beta s) - 2\alpha \sin(\beta) \sin(\beta(1-s))] \right] h(s) ds \right]$$

The Green's function given in (3-3) can then be obtained. \square

Notice

$$\frac{\partial}{\partial t} H(t, s) = \frac{1}{2\beta(\beta \sinh \alpha - \alpha \sin \beta)} \Phi(t, s), \quad (3-4)$$

where

$$\Phi(t, s) = \begin{cases} e^{-\alpha(t-s)} [e^{-\alpha} \beta^2 \cos(\beta(s-t)) + 2\alpha \beta \sin(\beta(1-s)) \cos(\beta t) - \beta^2 \sin(\beta(1-s)) \sin(\beta t) - \beta^2 \cos(\beta(1-s)) \cos(\beta t)] \\ - \alpha e^{-\alpha(t-s)} [2\alpha \sin(\beta(1-s)) \sin(\beta t) - e^{-\alpha} \beta \sin(\beta(s-t)) + \beta \sin(\beta(1-s)) \cos(\beta t) - \beta \cos(\beta(1-s)) \sin(\beta t)], & 0 \leq t \leq s \leq 1, \\ e^{-\alpha(t-s)} [e^{\alpha} \beta^2 \cos(\beta(t-s)) - 2\alpha \beta \sin(\beta s) \cos(\beta(1-t)) - \beta^2 \sin(\beta s) \sin(\beta(1-t)) - \beta^2 \cos(\beta s) \cos(\beta(1-t))] \\ - \alpha e^{-\alpha(t-s)} [2\alpha \sin(\beta s) \sin(\beta(1-t)) + e^{\alpha} \beta \sin(\beta(t-s)) - \beta \sin(\beta s) \cos(\beta(1-t)) + \beta \cos(\beta s) \sin(\beta(1-t))], & 0 \leq s \leq t \leq 1. \end{cases}$$

We point out several properties of the Green's function.

Lemma 3.2. $H(t, s)$ satisfies the following properties:

- (1) $H \in C([0, 1] \times [0, 1])$.
- (2) $H(0, s) = H(1, s) = \frac{e^{\alpha s} (\beta \sin(\beta(1-s)) - e^{-\alpha} \beta \sin(\beta s))}{2\beta(\beta \sinh \alpha - \alpha \sin \beta)}$ for all $s \in [0, 1]$.
- (3) $\max_{t \in [0, 1]} |H(t, s)| \leq \frac{\beta e^{\alpha} + 2\alpha + 2\beta}{2\beta(\beta \sinh \alpha - \alpha \sin \beta)}$ for all $s \in [0, 1]$.
- (4) $\max_{t \in [0, 1]} \left| \frac{\partial}{\partial t} H(t, s) \right| \leq \frac{\alpha \beta e^{\alpha} + 2\alpha^2 + 2\beta^2 + 2\alpha \beta + \beta^2 e^{\alpha}}{2\beta(\beta \sinh \alpha - \alpha \sin \beta)}$ for all $s \in [0, 1]$.
- (5) $\max_{t \in [0, 1]} \int_0^1 |H(t, s)| ds \leq \frac{\beta + \beta \sinh \alpha}{(\alpha^2 + \beta^2)(\beta \sinh \alpha - \alpha \sin \beta)}.$

$$(6) \max_{t \in [0,1]} \int_0^1 \left| \frac{\partial}{\partial t} H(t, s) \right| ds \leq \frac{\alpha^2 e^\alpha + \alpha^2 + \beta^2 + \beta^2 e^\alpha + \alpha e^\alpha + \alpha \beta + 3\beta}{(\alpha^2 + \beta^2)(\beta \sinh \alpha - \alpha \sin \beta)}.$$

Again, a proof is not given, since all these properties can be verified directly. Properties (5) and (6) are obtained by making all the terms in $H(t, s)$ and $(\partial/\partial t)H(t, s)$, respectfully, positive, integrating, and finding an upper bound when $t \in [0, 1]$.

Define the operator $T : \mathcal{B} \rightarrow \mathcal{B}$ by

$$Tu(t) = \int_0^1 H(t, s) k(s, u(s), u'(s)) ds.$$

Thus if u is a fixed point of T , then u is a solution of (3-1), (1-3). A standard application of the Arzelà–Ascoli theorem gives us that T is completely continuous.

Define

$$\max_{t \in [0,1]} \int_0^1 |H(t, s)| ds := \bar{H} \quad \text{and} \quad \max_{t \in [0,1]} \int_0^1 \left| \frac{\partial}{\partial t} H(t, s) \right| ds := \bar{H}'.$$

Theorem 3.3. Assume $f(t, x, y)$ is continuous in $[0, 1] \times \mathbb{R} \times \mathbb{R}$ with

$$|f(t, x, y) + 2\alpha y + (\alpha^2 + \beta^2)x| \leq \min \left\{ \frac{M}{\bar{H}}, \frac{M}{\bar{H}'} \right\}$$

for all $(t, x, y) \in [0, 1] \times [-M, M] \times [-M, M]$. Then (1-1), (1-3) has a solution $u^* \in \mathcal{M}$.

The proof is similar to the proof of Theorem 2.4 and is therefore omitted.

References

- [Al Mosa and Elloe 2016] S. Al Mosa and P. Elloe, “Upper and lower solution method for boundary value problems at resonance”, *Electron. J. Qual. Theory Differ. Equ.* **2016** (2016), art. id. 40. MR Zbl
- [Almansour and Elloe 2015] A. Almansour and P. Elloe, “Fixed points and solutions of boundary value problems at resonance”, *Ann. Polon. Math.* **115**:3 (2015), 263–274. MR Zbl
- [Hale and Verduyn Lunel 1993] J. K. Hale and S. M. Verduyn Lunel, *Introduction to functional-differential equations*, Applied Mathematical Sciences **99**, Springer, 1993. MR Zbl
- [Han 2007] X. Han, “Positive solutions for a three-point boundary value problem at resonance”, *J. Math. Anal. Appl.* **336**:1 (2007), 556–568. MR Zbl
- [Infante et al. 2016] G. Infante, P. Pietramala, and F. A. F. Tojo, “Non-trivial solutions of local and non-local Neumann boundary-value problems”, *Proc. Roy. Soc. Edinburgh Sect. A* **146**:2 (2016), 337–369. MR Zbl
- [Krasnosel’skii 1964] M. A. Krasnosel’skii, *Topological methods in the theory of nonlinear integral equations*, Int. Series of Monographs in Pure Appl. Math. **45**, Macmillan, New York, 1964. MR Zbl
- [Zeidler 1990] E. Zeidler, *Nonlinear functional analysis and its applications, II/A: Linear monotone operators*, Springer, 1990. MR Zbl

Received: 2018-01-24 Revised: 2018-02-13 Accepted: 2018-02-14

aldo_garciaguinto@mymail.eku.edu

*Department of Mathematics and Statistics,
Eastern Kentucky University, Richmond, KY, United States*

jeffrey.neugebauer@eku.edu

*Department of Mathematics and Statistics,
Eastern Kentucky University, Richmond, KY, United States*

Guidelines for Authors

Submissions in all mathematical areas are encouraged. All manuscripts accepted for publication in *Involve* are considered publishable in quality journals in their respective fields, and include a minimum of one-third student authorship. Submissions should include substantial faculty input; faculty co-authorship is strongly encouraged. Authors may submit manuscripts in PDF format on-line at the Submission page at the Involve website.

Originality. Submission of a manuscript acknowledges that the manuscript is original and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles in *Involve* are usually in English, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

Format. Authors are encouraged to use L^AT_EX but submissions in other varieties of T_EX, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibT_EX 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, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@msp.org with details about how your graphics were generated.

White space. Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Proofs. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

involve

2019 vol. 12 no. 1

Optimal transportation with constant constraint	1
WYATT BOYER, BRYAN BROWN, ALYSSA LOVING AND SARAH TAMMEN	
Fair choice sequences	13
WILLIAM J. KEITH AND SEAN GRINDATTI	
Intersecting geodesics and centrality in graphs	31
EMILY CARTER, BRYAN EK, DANIELLE GONZALEZ, RIGOBERTO FLÓREZ AND DARREN A. NARAYAN	
The length spectrum of the sub-Riemannian three-sphere	45
DAVID KLAPHECK AND MICHAEL VANVALKENBURGH	
Statistics for fixed points of the self-power map	63
MATTHEW FRIEDRICHSSEN AND JOSHUA HOLDEN	
Analytical solution of a one-dimensional thermistor problem with Robin boundary condition	79
VOLODYMYR HRYNKIV AND ALICE TURCHANINOVA	
On the covering number of S_{14}	89
RYAN OPPENHEIM AND ERIC SWARTZ	
Upper and lower bounds on the speed of a one-dimensional excited random walk	97
ERIN MADDEN, BRIAN KIDD, OWEN LEVIN, JONATHON PETERSON, JACOB SMITH AND KEVIN M. STANGL	
Classifying linear operators over the octonions	117
ALEX PUTNAM AND TEVIAN DRAY	
Spectrum of the Kohn Laplacian on the Rossi sphere	125
TAWFIK ABBAS, MADELYNE M. BROWN, RAVIKUMAR RAMASAMI AND YUNUS E. ZEYTUNCU	
On the complexity of detecting positive eigenvectors of nonlinear cone maps	141
BAS LEMMENS AND LEWIS WHITE	
Antiderivatives and linear differential equations using matrices	151
YOTSANAN MEEMARK AND SONGPON SRIWONGSA	
Patterns in colored circular permutations	157
DANIEL GRAY, CHARLES LANNING AND HUA WANG	
Solutions of boundary value problems at resonance with periodic and antiperiodic boundary conditions	171
ALDO E. GARCIA AND JEFFREY T. NEUGEBAUER	

