

involve

a journal of mathematics

Editorial Board

Kenneth S. Berenhaut, *Managing Editor*

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



involve

msp.org/involve

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 Scorpan

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2016 is US \$160/year for the electronic version, and \$215/year (+\$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to MSP.

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

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

PUBLISHED BY

 **mathematical sciences publishers**
nonprofit scientific publishing

<http://msp.org/>

© 2016 Mathematical Sciences Publishers

A combinatorial proof of a decomposition property of reduced residue systems

Yotsanan Meemark and Thanakorn Prinyasart

(Communicated by Filip Saidak)

In this paper, we look at three common theorems in number theory: the Chinese remainder theorem, the multiplicative property of the Euler totient function, and a decomposition property of reduced residue systems. We use a grid of squares to give simple transparent visual proofs.

1. Introduction

Let m and n be positive integers. Construct an $m \times n$ grid of squares. We place the sequence of positive integers $1, 2, 3, \dots$ into the grid beginning with the upper left-hand corner cell and moving from the cell numbered i to the cell numbered $i + 1$ by going one box down and one to the right. If this is not possible (at the last row or the rightmost column of our $m \times n$ table), we wrap around to the opposite edge and continue. It is easy to see that the i -th row has numbers that are congruent to i modulo m and the j -th column has numbers that are congruent to j modulo n .

We observe that two positive integers x and y fill the same cell if and only if $x \equiv y \pmod{m}$ and $x \equiv y \pmod{n}$, which is equivalent to $x - y$ is divisible by $[m, n]$, the least common multiple of m and n . From this, it follows that there is a repetition after we get to $[m, n]$ and, of course, that $[m, n]$ is the first integer to arrive at the lower right-hand corner. Thus we have the positive integers from 1 to $[m, n]$ in the table. Notice that we can number all mn boxes in this way if and only if m and n are relatively prime. This follows from $(m, n)[m, n] = mn$. Here (m, n) denotes the greatest common divisor of m and n . When $m = 3$ and $n = 5$, the above explanation can be illustrated by a glued 3×5 table and a discrete torus, which appear in [Terras 1999]; see Figure 1.

In what follows, we point out some applications of this elementary construction. It provides not only a visual verification of two common theorems in number theory, namely, the Chinese remainder theorem and the multiplicative property of the Euler

MSC2010: 11A07.

Keywords: Chinese remainder theorem, reduced residue system.

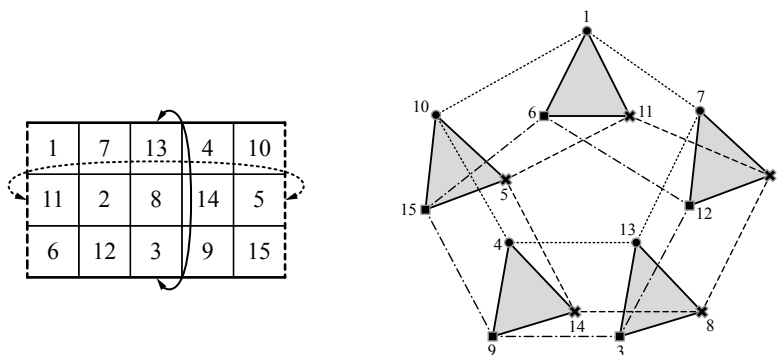


Figure 1. A glued 3×5 table and its corresponding discrete torus.

totient ϕ -function, but also gives a constructive proof for a decomposition property of reduced residue systems, to be defined below. The results are presented in Sections 2 and 3, respectively.

2. The Chinese remainder theorem

Let $d = (m, n)$. We can split the $m \times n$ table into $(m/d) \times (n/d)$ subtables so that each of them is a square $d \times d$ table as shown in Figure 2.

By the above filling method, each subtable has numbers only in its diagonal. For example, the upper left-hand corner subtable will be filled with integers from 1 to d . We move from one subtable to another by going one subtable down and one

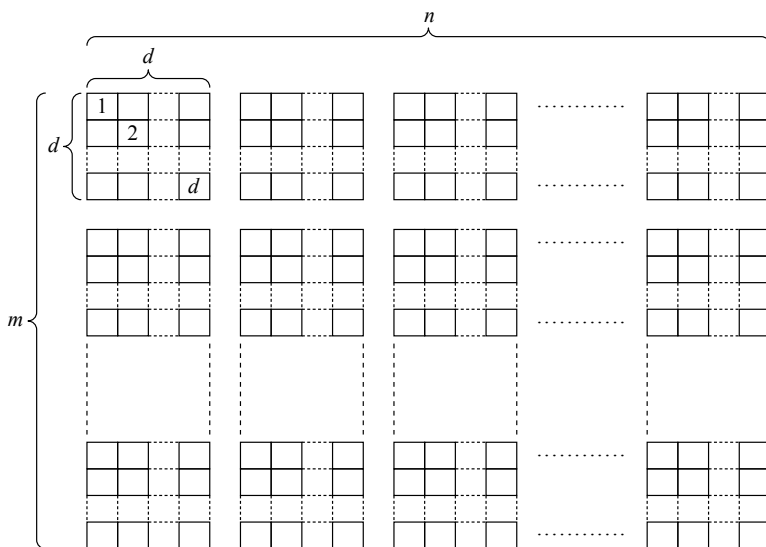


Figure 2. Our division of the $m \times n$ table into $d \times d$ subtables, where $d = (m, n)$.

to the right and wrap around as explained before. Hence a square $d \times d$ subtable can be viewed as a block in an $(m/d) \times (n/d)$ table. Since $(m/d, n/d) = 1$, all $d \times d$ cells have the subsequence

$$(l - 1)d + 1, (l - 1)d + 2, \dots, ld \quad \text{for some } l \in \left\{1, 2, \dots, \frac{mn}{d^2}\right\}$$

in their diagonals. Thus, the $m \times n$ table is transformed into an $(m/d) \times (n/d)$ table with $(m/d, n/d) = 1$ and we can now number all of the mn/d^2 boxes with $1, 2, \dots, mn/d^2$. Now observe that the integers in the original table appear only in the positions $(k + id, k + jd)$, where $k \leq d, i \leq m/d - 1$ and $j \leq n/d - 1$. In other words, the positions of the integers are (a, b) with $a \equiv b \pmod d$, that is, $d \mid (a - b)$. Furthermore, as mentioned earlier, there is a repetition of solutions modulo $[m, n]$. Therefore we have proved the Chinese remainder theorem:

Theorem 1. *Let m and n be positive integers. For integers a and b , the congruences*

$$x \equiv a \pmod m \quad \text{and} \quad x \equiv b \pmod n$$

admit a simultaneous solution if and only if (m, n) divides $a - b$. Moreover, if a solution exists, then it is unique modulo $[m, n]$.

The result when $(m, n) = 1$ was also described by Ledet [2007]. We demonstrate Theorem 1 by the following example.

Example 2. Let $m = 6$ and $n = 8$. Then $(m, n) = 2$ and $[m, n] = 24$. Filling the 6×8 table with the numbers from 1 to 24 as previously described, we obtain

1		19		13		7	
	2		20		14		8
9		3		21		15	
	10		4		22		16
17		11		5		23	
	18		12		6		24

According to this table, one easily sees that $x \equiv 22 \pmod{24}$ is a simultaneous solution for $x \equiv 4 \pmod 6$ and $x \equiv 6 \pmod 8$, and there is no x for which both $x \equiv 5 \pmod 6$ and $x \equiv 4 \pmod 8$. □

If m is a positive integer, the *Euler totient function* $\phi(m)$ is defined to be the number of positive integers not exceeding m which are relatively prime to m . By a *reduced residue system modulo m* , we mean any set of $\phi(m)$ integers, pairwise incongruent modulo m , each of which is relatively prime to m . Notice that if p is a prime, then $\phi(p) = p - 1$ and $\{1, 2, \dots, p - 1\}$ is a reduced residue system modulo p . It is also immediate that $\phi(p^s) = p^s - p^{s-1}$ for all $s \in \mathbb{N}$.

Next, we investigate the decomposition property of the reduced residue systems by our combinatorial technique. Let $a = mn$, where m and n are positive integers.

We arrange the positive integers $1, 2, \dots, [m, n]$ into the $m \times n$ grid of squares by using the above filling method and delete the i -th rows and j -th columns of the table for all i and j with $(m, i) > 1$ and $(n, j) > 1$. For a better understanding of this construction, one may erase all even (second, fourth, ...) rows and all even columns of the table in Example 2. Recall that the i -th row has numbers that are congruent to i modulo m and the j -th column has numbers that are congruent to j modulo n .

Let l be a remaining positive integer in the table. Notice that $l \equiv i \pmod{m}$ with $(m, i) = 1$ and $1 \leq i \leq m$; that is, $l = i + km$ for some nonnegative integer k . Since $(m, i) = 1$, there exist integers x and y such that $mx + iy = 1$. Consequently, we choose $x' = x - ky \in \mathbb{Z}$ and $y' = y \in \mathbb{Z}$. Then $mx' + ly' = 1$, so we have $(l, m) = 1$. Similarly, we can show that $(l, n) = 1$. Since $a = mn$, we also have $(l, a) = 1$. Hence all positive integers left in the table after deletion are relatively prime to a and less than $[m, n]$.

For $(m, n) = 1$, we can place the positive integers from 1 to $[m, n] = mn = a$ in the $m \times n$ grid by the means above. Erase the i -th rows that are not relatively prime to m and cross out the j -th columns that are not relatively prime to n . Then we obtain $\phi(m)\phi(n)$ undeleted cells and eliminate all numbers that are not relatively prime to m and n . Since $(m, n) = 1$, the entries left in the table coincide with positive integers less than and relatively prime to a , so the number of these entries is equal to $\phi(a)$. Hence we can conclude the well-known multiplicative property of the Euler totient ϕ -function, namely, if $(m, n) = 1$, then $\phi(mn) = \phi(a) = \phi(m)\phi(n)$. This combinatorial proof is the one given in the famous book on number theory [Niven et al. 1991]. Since $\phi(p^s) = p^s - p^{s-1} = p^s(1 - p^{-1})$ when p is a prime and $s \geq 1$, the multiplicative property gives a formula for computing

$$\phi(M) = M \prod_{p|M} (1 - p^{-1})$$

for any positive integer M .

3. Decomposition property of reduced residue systems

Let m' be the product of primes in m not in n with the same exponents that they have in m . It is easy to see that m' and n are relatively prime. Place the positive integers from 1 to $m'n$ in the $m' \times n$ grid and erase the rows that are not relatively prime to m' and the columns that are not relatively prime to n . Let l be a positive integer left in the table after deletion. Then $(l, m') = 1 = (l, n)$. Assume that there exists a prime p dividing l and $a = mn$. Thus $p | m$ or $p | n$. But $(l, n) = 1$, so p is not in n and thus p is in m . Therefore $p | m'$, which contradicts the fact

that $(l, m') = 1$. Hence the remaining $\phi(m')\phi(n)$ positive integers in the table are relatively prime to a . Consider them as a $\phi(m') \times \phi(n)$ matrix. The set of all members in each row of this matrix is a reduced residue system modulo n and $x \equiv y \pmod n$ for all integers x and y that are in the same column.

Let A_0 be the above $\phi(m') \times \phi(n)$ matrix and

$$A_i = A_0 + i \begin{bmatrix} m'n & \dots & m'n \\ \vdots & \ddots & \vdots \\ m'n & \dots & m'n \end{bmatrix}_{\phi(m') \times \phi(n)} \quad \text{for } i = 0, 1, \dots, \frac{\phi(mn)}{\phi(m')\phi(n)} - 1.$$

The identity $\phi(M) = M \prod_{p|M} (1 - p^{-1})$ shows that

$$\frac{\phi(mn)}{\phi(m')\phi(n)} = \frac{m}{m'},$$

so the index i ranges from 0 up to $m/m' - 1$, which implies that the entries of A_i do not exceed a . It is also obvious that each entry in A_i is relatively prime to a . We augment A_0 by the matrices

$$A_1, \dots, A_{\frac{\phi(a)}{\phi(m')\phi(n)} - 1},$$

respectively, to form a new $(\phi(a)/\phi(n)) \times \phi(n)$ matrix. Then the entries of this matrix are integers from 1 to a , relatively prime to a , with the condition that the set of the entries in each row is a reduced residue system modulo n and $x \equiv y \pmod n$ for all integers x and y that are in the same column. Hence we have a constructive proof for a theorem on a decomposition property of reduced residue systems modulo a summarized as follows.

Theorem 3. *Let S be a residue system modulo a , and let $n \geq 1$ be a divisor of a . Then we have the following decompositions of S :*

- (1) S is the union of $\phi(a)/\phi(n)$ disjoint sets, each of which is a reduced residue system modulo n .
- (2) S is the union of $\phi(n)$ disjoint sets, each of which consists of $\phi(a)/\phi(n)$ numbers congruent to each other modulo n .

Remark. Another proof of this theorem and its application on character sums can be found in Apostol's book [1976].

Example 4. Consider $a = 48$ with $m = 6$ and $n = 8$. Since $8 = 2^3$ and $6 = 2 \cdot 3$, let $m' = 3$. Filling a 3×8 table with numbers by our technique, we obtain

1	10	19	4	13	22	7	16
17	2	11	20	5	14	23	8
9	18	3	12	21	6	15	24

Delete the rows that contain numbers not relatively prime to 3 and the columns that contain numbers not relatively prime to 8. We have then the 2×4 matrix formed from the remaining numbers given by

$$A = \begin{bmatrix} 1 & 19 & 13 & 7 \\ 17 & 11 & 5 & 23 \end{bmatrix}.$$

Augment this matrix with $\phi(3) = 2$ rows obtained by adding $m'n$ to all entries of A , so we finally reach the decomposition

$$A' = \begin{bmatrix} 1 & 19 & 13 & 7 \\ 17 & 11 & 5 & 23 \\ 25 & 43 & 37 & 31 \\ 41 & 35 & 29 & 47 \end{bmatrix}$$

as desired. □

Acknowledgments

This work grows out of Prinyasart's mini-project at Chulalongkorn University written under the direction of Meemark, to whom he expresses his gratitude. Prinyasart was supported in part by the Development and Promotion of Science and Technology Talents Project.

References

- [Apostol 1976] T. M. Apostol, *Introduction to analytic number theory*, Springer, New York, 1976. MR 0434929 Zbl 0335.10001
- [Ledet 2007] A. Ledet, "Faro shuffles and the Chinese remainder theorem", *Math. Mag.* **80**:4 (2007), 283–289. MR 2356580 Zbl 1219.05002
- [Niven et al. 1991] I. Niven, H. S. Zuckerman, and H. L. Montgomery, *An introduction to the theory of numbers*, 5th ed., Wiley, New York, 1991. MR 1083765 Zbl 0742.11001
- [Terras 1999] A. Terras, *Fourier analysis on finite groups and applications*, London Mathematical Society Student Texts **43**, Cambridge Univ. Press, 1999. MR 1695775 Zbl 0928.43001

Received: 2011-10-16 Revised: 2014-12-21 Accepted: 2015-06-23

yotsanan.m@chula.ac.th *Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Bangkok 10330, Thailand*

thanakorn_dpst@hotmail.com *Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Bangkok 10330, Thailand*

Strong depth and quasigeodesics in finitely generated groups

Brian Gapinski, Matthew Horak and Tyler Weber

(Communicated by Kenneth S. Berenhaut)

A “dead end” in the Cayley graph of a finitely generated group is an element beyond which no geodesic ray issuing from the identity can be extended. We study the so-called “strong dead-end depth” of group elements and its relationship with the set of infinite quasigeodesic rays issuing from the identity. We show that the ratio of strong depth to word length is bounded above by $\frac{1}{2}$ in every finitely generated group and that for any element g in a finitely generated group G , there is an infinite $(3, 0)$ -quasigeodesic ray issuing from the identity and passing through g . Applying the Švarc–Milnor lemma to a finitely generated group acting geometrically on a geodesically connected metric space, we obtain the result that for any two points in such a space, there is an infinite quasigeodesic ray starting at one and passing through the other with quasigeodesic constants independent of the points selected.

1. Introduction

Background and summary of results. Let G be a group and X a finite generating set for G . The Cayley graph for G with respect to X is the graph with vertex set G and an edge from g to gx for every $x \in X \cup X^{-1}$. Throughout, we will use $\Gamma(G, X)$ or simply Γ to denote the Cayley graph of G with respect to X . Assigning all edges in $\Gamma(G, X)$ length 1 determines a metric on $\Gamma(G, X)$, and therefore on G , which we denote by $d(\cdot, \cdot) = d_X(\cdot, \cdot)$. The metric d determines a length function $l_X = l$ on G defined by $l(x) = d(e, x)$, where e is the identity of G .

Many results on discrete groups rely upon an understanding of the structure of geodesics in the Cayley graph. In particular, the question arises of the existence (or nonexistence) of elements g beyond which no geodesic ray from the identity to g can be extended to a longer geodesic. In the current literature, such elements are called *dead ends*. Dead ends have been applied to, for example, the construction in [Lyons

MSC2010: 20F65.

Keywords: Cayley graph, dead end, quasigeodesic.

This research was supported by the NSF REU grant DMS-1062403.

et al. 1996] of a random walk on the lamplighter group that is biased towards the identity but that escapes from the identity faster than a simple random walk. Dead ends also played a role in the proof that infinite commensurable hyperbolic groups are bi-Lipschitz equivalent [Bogopol'skiĭ 1997].

A property of arbitrary metric spaces similar to the nonexistence of dead ends in a group is the geodesic extension property, which states that every finite geodesic segment is contained in an infinite geodesic line. The geodesic extension property appears frequently in the study of nonpositively curved spaces, and especially in the study of CAT(0) spaces and CAT(0) groups. For example, it is shown in [Bridson and Haefliger 1999; Hosaka 2012] that if X is a CAT(0) space with the geodesic extension property, then any geometric action on X of a group of the form $G = G_1 \times G_2$ induces a splitting of X as $X = X_1 \times X_2$ with a geometric action of G_1 on X_1 and of G_2 on X_2 . With further assumptions on G , the action of $G_1 \times G_2$ on $X_1 \times X_2$ is the product action. We refer the reader to Chapter II.6 of the book [Bridson and Haefliger 1999] for a thorough discussion of the role of infinite geodesics in the study of the geometry of nonpositively curved spaces.

One difficulty of extending the above results involving dead ends or geodesic extension to larger classes of groups is that quasi-isometries take geodesics to quasi-geodesics, not geodesics. Therefore, it is possible for a group to have dead ends with respect to one generating set but not another. Even worse, there exist groups with unbounded dead-end depth with respect to one generating set, but no dead ends with respect to another [Riley and Warshall 2006]. This quasi-isometry noninvariance prevents one from using or studying dead ends by way of a geometric action of the group in question on a space, since such an action provides only the quasi-isometric equivalence of the group with the space.

In this paper, we address this problem in two ways. We first analyze the behavior of the *strong depth*, $\sigma(g)$, of an element g , introduced by Lehnert [2009]. Informally, this is the minimum distance back towards the identity that any path in $\Gamma(G, X)$ from g to an element of greater length must travel. Warshall [2011] introduced a similar notion, the *retreat depth*, which is the minimum distance, d , towards the identity that one must travel to enter an unbounded component of the complement of the ball of radius $l(g) - d$. Strong depth and retreat depth are similar and seem to behave roughly the same. Therefore, although we have chosen to phrase all of our theorems in terms of strong depth, they can be restated in terms of retreat depth.

Even though strong depth depends on the generating set, just as for ordinary dead ends, its ratio to length turns out to be well-behaved for all generating sets. In Section 2 we prove the following:

Theorem 2.2. *Let G be an infinite group and X a finite generating set for G . Then for all $g \in G \setminus \{e\}$, we have $\sigma(g)/l(g) \leq \frac{1}{2}$.*

The second way in which we address the strong dependence of dead ends on the quasi-isometry class or the generating set is to relax the question of extending a geodesic path between two elements and instead ask whether there exist universal constants L and A for which one can find an infinite (L, A) -quasigeodesic ray passing through any arbitrary pair of points. If so, then we say that the space in question has *uniform quasigeodesic ray extension*. In Section 3, we show that every infinite, finitely generated group has uniform quasigeodesic ray extension:

Theorem 3.3. *Let G be an infinite group and X a finite generating set for G . Then for all $g \in G$ there exists an infinite $(3, 0)$ -quasigeodesic ray in $\Gamma(G, X)$ starting at the identity of G and passing through g .*

Applying the Švarc–Milnor lemma to a “nice” metric space X admitting a geometric action of a finitely generated group G , we obtain the following corollary:

Corollary 3.8. *Let (X, d_X) be a metric space in which any two points can be joined by a geodesic segment and G a finitely generated group acting by isometries on X . If there exists a ball $B(x_0, R)$ in X whose G -translates cover X with the property that for every $r > 0$ the set $\{g : B(x_0, r) \cap g \cdot B(x_0, r) \neq \emptyset\}$ is finite, then X has uniform quasigeodesic ray extension.*

Definitions. In this section, we review the definitions of the various types of dead ends and dead-end depths that we deal with and summarize some of their basic properties. In what follows, all graphs are assumed to be endowed with the metric d induced by declaring each edge to have length 1.

Definition 1.1. Let Γ be a graph. A *path* in Γ is a function $\rho : I \rightarrow \Gamma$, where I is the intersection of a (possibly infinite) interval of the real line with \mathbb{Z} such that for each $i, j \in I$ with $|i - j| = 1$, we have that $\rho(i)$ and $\rho(j)$ span an edge in Γ .

For convenience and to aid the memory and imagination, we often express a path as

$$\rho = \dots, a_k, a_{k+1}, a_{k+2}, a_{k+3}, \dots,$$

where $a_i = \rho(i)$. We use similar notation for finite paths and paths infinite on only one end.

Definition 1.2. If $\rho = a_0, a_1, a_2, \dots, a_m$ and $\tau = x_0, x_1, x_2, \dots, x_n$ are paths with $a_m = x_0$, then the *concatenation* of ρ and τ is $\rho\tau = a_1, a_2, \dots, a_m, x_1, \dots, x_n$.

Definition 1.3. Let $\rho = a_1, a_2, \dots, a_n$ be a path in a graph Γ . The *path length* between two vertices a_i and a_j in ρ is defined as $p_\rho(a_i, a_j) = |i - j|$.

Definition 1.4. Let $\gamma = a_1, a_2, \dots, a_n$ be a (possibly infinite) path in the graph Γ . We say that γ is a *geodesic* in Γ if for all i, j , we have $p_\gamma(a_i, a_j) = d(a_i, a_j)$.

We now specialize to the case where G is a finitely generated group, X is a fixed generating set for G and $\Gamma(G, X)$ is the Cayley graph of G with respect to X . All of these definitions are dependent on the generating set X , but if there is only one generating set in question, we often omit it from the notation.

Definition 1.5. For $g \in G$ the *word length* of g (with respect to X) is $l(g) = l_X(g) = d(e, g)$ with distance measured in $\Gamma(G, X)$.

Definition 1.6. Let G be a group and X a finite generating set for G . Let $g \in G$ and $n \in \mathbb{N}$. The *sphere of radius n centered at g* (with respect to generating set X) is $S_g(n) := \{h \in G : d_X(g, h) = n\}$.

Definition 1.7. Let G be a group and X a finite generating set for G . Let $g \in G$ and $n \in \mathbb{N}$. The *ball of radius n centered at g* (with respect to generating set X) is $B_g(n) := \{h \in G : d_X(g, h) \leq n\}$.

Definition 1.8. Let G be a group and X a finite generating set for G . The *dead-end depth* (with respect to X) of an element $g \in G$ with $l_X(g) = n$ is the least integer k such that there exists a path of length k in $\Gamma(G, X)$ from g to $S_e(n+1)$. We denote the dead-end depth of g as $\delta_X(g)$ or simply $\delta(g)$ if only one generating set is under investigation. An element $g \in G$ with $\delta(g) > 1$ is called a *dead end*.

Definition 1.9. Let G be a group and X a finite generating set for G . We say that G has *bounded dead-end depth with respect to X* if there exists $N \in \mathbb{N}$ such that, for all $g \in G$, we have $\delta(g) \leq N$. If no such N exists, we say that G has *unbounded dead-end depth with respect to X* .

As previously mentioned, the dead-end elements, dead-end depth, and retreat depth of a group are strongly dependent on the generating set. Riley and Warshall [2006] constructed a group that has bounded dead-end depth with respect to one finite generating set but unbounded dead-end depth with respect to another finite generating set. Lehnert [2009] introduced the following notion of strong depth and showed that Houghton's group H_2 has unbounded strong depth.

Definition 1.10. Let Γ be the Cayley graph of a group G with respect to the finite generating set X and let $g \in G$ with $d(e, g) = n$. The *strong depth* of g (with respect to X) is the minimum k such that there exists a path in $\Gamma(G, X)$ from g to an element of $S_e(n+1)$ that does not enter $B_e(n-k)$. We denote the strong depth of g with respect to X as $\sigma_X(g)$ or simply $\sigma(g)$ if the context is clear.

2. Strong depth

There are two "easy" inequalities satisfied by dead-end depth and strong depth. The first inequality follows from the definitions and states that for every element g of a finitely generated group G , we have that $\sigma(g) \leq \frac{1}{2}\delta(g)$. The second inequality

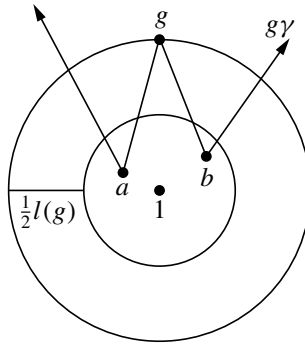


Figure 1. Schematic of a geodesic passing through an element g with $\sigma(g) > \frac{1}{2}l(g)$.

states that for every element g of a finitely generated group G , we have that $\delta(g) \leq 2l(g) + 1$. This is observed by taking a geodesic path from g to the identity and concatenating a geodesic path to an element of greater length than g . To our knowledge, these are the only two inequalities involving dead-end depth known to hold in all finitely generated groups and for all generating sets. In this section, we establish another property of strong depth that holds for every finite generating set of any infinite finitely generated group. Our argument uses the fact that every infinite finitely generated group contains an infinite geodesic line passing through the identity. A sketch of the proof of this fact can be found in [de la Harpe 2000]. We record this as:

Lemma 2.1. *Let G be an infinite group and X a finite generating set for G . Then the Cayley graph $\Gamma(G, X)$ contains a bi-infinite geodesic line passing through the identity.*

Theorem 2.2. *Let G be an infinite group and X a finite generating set for G . Then for all $g \in G \setminus \{e\}$, we have $\sigma(g)/l(g) \leq \frac{1}{2}$.*

Proof. Suppose towards a contradiction that there exists a nonidentity $g \in G$ with $\sigma(g)/l(g) > \frac{1}{2}$. By Lemma 2.1, select an infinite geodesic line

$$\gamma = \dots, w_2, w_1, e, v_1, v_2, \dots$$

in G . Since G acts on Γ by isometries, $g \cdot \gamma$ is an infinite geodesic line that passes through g . Let a be the element in $\{g \cdot w_k : k \in \mathbb{N}\}$ of least length. If two or more such elements exist, select a to be the closest such element to g along $g \cdot \gamma$. Similarly, we let b be the element of $\{g \cdot v_k : k \in \mathbb{N}\}$ of least length, again taking the closest such element to g along $g \cdot \gamma$ if more than one least length element exists. A schematic of this is shown in Figure 1.

Since $\sigma(g) > \frac{1}{2}l(g)$, we have

$$l(a) < \frac{1}{2}l(g), \quad (1)$$

$$l(b) < \frac{1}{2}l(g). \quad (2)$$

Inequalities (1) and (2), together with the facts that $d(a, g) \geq l(g) - l(a)$ and $d(b, g) \geq l(g) - l(b)$, give

$$d(a, g) > \frac{1}{2}l(g), \quad (3)$$

$$d(b, g) > \frac{1}{2}l(g). \quad (4)$$

Now consider the distance along $g \cdot \gamma$ between a and b . Since $g \cdot \gamma$ is a geodesic, inequalities (3) and (4) give

$$p_{g \cdot \gamma}(a, g) = d(a, g) > \frac{1}{2}l(g), \quad (5)$$

$$p_{g \cdot \gamma}(b, g) = d(b, g) > \frac{1}{2}l(g). \quad (6)$$

Since the total path length between a and b is simply the sum of $p_{g \cdot \gamma}(a, g)$ and $p_{g \cdot \gamma}(b, g)$, equations (5) and (6) give

$$\begin{aligned} d(a, b) &= p_{g \cdot \gamma}(a, b) = p_{g \cdot \gamma}(a, g) + p_{g \cdot \gamma}(b, g) \\ &> \frac{1}{2}l(g) + \frac{1}{2}l(g) = l(g). \end{aligned} \quad (7)$$

By the triangle inequality, the distance between a and b is less than or equal to the sum of their lengths. So, by (1) and (2),

$$\begin{aligned} d(a, b) &\leq l(a) + l(b) \\ &< \frac{1}{2}l(g) + \frac{1}{2}l(g) = l(g). \end{aligned} \quad (8)$$

Thus (7) and (8) provide a contradiction, which proves that $\sigma(g)/l(g) \leq \frac{1}{2}$ for all $g \in G \setminus \{e\}$. \square

In practice, groups containing elements with large ratios of strong depth to length seem difficult to find. Indeed, all elements of sufficiently large length of the families of dead ends studied in the papers referenced on page 368 have ratios of strong depth to length that are less than $\frac{1}{6}$. Moreover, we were able to modify the families or generating sets in question to get families of elements with ratios of strong depth to length only as large as $\frac{1}{4}$. This leads one to consider the “limiting” ratio of strong depth to length

$$\Omega(G) = \limsup_{l(g) \rightarrow \infty} \left\{ \frac{\sigma(g)}{l(g)} : g \in G \right\}$$

and ask if there are groups for which $\Omega(G) = \frac{1}{2}$.

Imagining what such a group would look like, one envisions a group G with a sequence of elements (g_n) of increasing length for which it is more and more difficult to reach elements of larger length without retreating closer and closer to halfway back to the identity. If the group had many such elements, one might expect it to be difficult to construct a family of paths, one through each group element, which escape from identity uniformly quickly. In the next section, we examine a property, which we call *uniform quasigeodesic ray extension*, that guarantees the existence of such a family. Unfortunately, however, we cannot establish a connection between $\Omega(G) = \frac{1}{2}$ and a group not having uniform quasigeodesic ray extension. Indeed, the best connection we are able to establish is that uniform quasigeodesic ray extension implies that $\Omega(G) < 1$, which is weaker than the conclusion of Theorem 2.2 for finitely generated groups.

3. Uniform quasigeodesic ray extension

One difficulty in the study of dead ends and in the use of geodesic completeness is that neither existence of dead ends nor the geodesic completeness property is invariant under quasi-isometry. Thus, one cannot apply one of the main strategies of geometric group theory: analyzing a group by understanding its action on a space (or vice versa). In this section we suggest a way of dealing with this difficulty by relaxing the condition of finding geodesic paths to that of finding quasigeodesic paths, all of which have the same multiplicative and additive constants. We begin by reviewing the terminology involved with quasi-isometries and quasigeodesics.

Definition 3.1. Let (X, d_X) and (Y, d_Y) be metric spaces. A function $f : X \rightarrow Y$ is an (L, A) -*quasi-isometric embedding* if there exist constants $L \geq 1$ and $A \geq 0$ such that, for all $x_1, x_2 \in X$,

$$\frac{1}{L}d_X(x_1, x_2) - A \leq d_Y(f(x_1), f(x_2)) \leq Ld_X(x_1, x_2) + A.$$

We refer to L as the *multiplicative constant* and A as the *additive constant* for f . The function f is an (L, A, C) -*quasi-isometry* if there exists an additional constant $C \geq 0$ such that, for all $y \in Y$, there exists $x \in X$ with $d_Y(y, f(x)) \leq C$.

Definition 3.2. Let (X, d_X) be a metric space. A *quasigeodesic* is a (λ, ϵ) -quasi-isometric embedding $g : I \rightarrow X$, where $I = (a, b) \cap \mathbb{Z}$ for some $a, b \in \mathbb{R} \cup \{\infty, -\infty\}$.

We remark that if γ is a geodesic in a graph Γ in the sense of Definition 1.4 then γ is a $(1, 1)$ -quasigeodesic in the sense of Definition 3.2.

Theorem 3.3. *Let G be an infinite group and X a finite generating set for G . Then for all $g \in G$ there exists an infinite $(3, 0)$ -quasigeodesic ray in $\Gamma(G, X)$ starting at the identity of G and passing through g .*

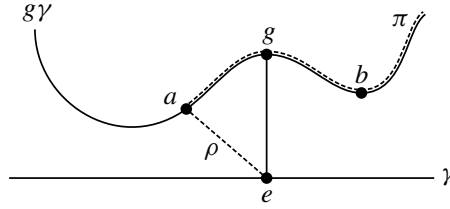


Figure 2. Schematic of the path in Theorem 3.3.

Proof. Let G be an infinite group and X a finite generating set for G and let $\Gamma = \Gamma(G, X)$ be the Cayley graph of G with respect to X . Let γ be an infinite geodesic line in Γ that passes through the identity as given by Lemma 2.1. Let $g \in G$ be an arbitrary element. Since G acts on the Cayley graph by isometries, $g \cdot \gamma$ is an infinite geodesic line passing through g . Let a be an element on $g \cdot \gamma$ such that, for all b on $g \cdot \gamma$, we have $l(a) \leq l(b)$.

Let ρ be any geodesic from the identity to a and let π be the geodesic ray along $g \cdot \gamma$ that starts at a and passes through g . Figure 2 shows a schematic illustration of these geodesic segments. Define $\gamma' : \mathbb{N} \cup \{0\} \rightarrow G$ by

$$\gamma'(t) = \begin{cases} \rho(t) & \text{if } 0 \leq t \leq l(a), \\ \pi(t - l(a)) & \text{if } t > l(a). \end{cases}$$

This infinite segment is indicated by the dotted lines in Figure 2. Observe that γ' is infinite by construction. We break the problem into cases to prove that γ' is a $(3, 0)$ -quasigeodesic by showing that the following inequality holds for all x and y in the domain of γ' :

$$\frac{1}{3}|x - y| \leq d(\gamma'(x), \gamma'(y)) \leq 3|x - y|. \tag{9}$$

Case 1: $x \leq l(a)$ and $y \leq l(a)$. In this case, $\gamma'(x)$ and $\gamma'(y)$ are defined by ρ . Since ρ is a geodesic, we have

$$\frac{1}{3}|x - y| \leq |x - y| = d(\rho(x), \rho(y)) = |x - y| \leq 3|x - y|.$$

Because $\rho(x) = \gamma'(x)$ and $\rho(y) = \gamma'(y)$, we have $d(\rho(x), \rho(y)) = d(\gamma'(x), \gamma'(y))$. Therefore, inequality (9) is satisfied.

Case 2: $x > l(a)$ and $y > l(a)$. This case is similar to the previous one.

Case 3: $x \leq l(a)$ and $y > l(a)$. Note that the right inequality of (9) is trivially true by the definitions of distance and path length because γ' is a path. For the left inequality of (9), first note that $l(\gamma'(x)) \leq l(a) \leq l(\gamma'(y))$. Therefore any geodesic path between $\gamma'(x)$ and $\gamma'(y)$ must intersect $S_e(l(a))$. Since a lies on $S_e(l(a))$, this implies that $d(\gamma'(x), a) \leq d(\gamma'(x), \gamma'(y))$. Because the portion of ρ between

$\gamma'(x)$ and a is a geodesic, we have

$$|x - l(a)| = d(\gamma'(x), a) \leq d(\gamma'(x), \gamma'(y)). \tag{10}$$

By inequality (10) and the triangle inequality,

$$\begin{aligned} |l(a) - y| &= d(a, \gamma'(y)) \\ &\leq d(a, \gamma'(x)) + d(\gamma'(x), \gamma'(y)) \leq 2d(\gamma'(x), \gamma'(y)). \end{aligned} \tag{11}$$

By inequalities (10) and (11),

$$\begin{aligned} \frac{1}{3}|x - y| &= \frac{1}{3}(|x - l(a)| + |l(a) - y|) \\ &\leq \frac{1}{3}(d(\gamma'(x), \gamma'(y)) + 2d(\gamma'(x), \gamma'(y))) = d(\gamma'(x), \gamma'(y)). \end{aligned}$$

Therefore, inequality (9) is satisfied in this case as well.

Therefore, for all x and y in the domain of γ' , inequality (9) holds and γ' is an infinite $(3, 0)$ -quasigeodesic that starts at the identity and passes through g . \square

We have just shown that for any element g in an infinite group G with finite generating set X , there is a $(3, 0)$ -quasigeodesic ray in $\Gamma(G, X)$ starting at the identity and passing through g . We now generalize this property by relaxing the additive and multiplicative constants of the quasigeodesic rays and prove that a large class of metric spaces have this property.

Definition 3.4. A metric space (X, d) has *uniform quasigeodesic ray extension* if there exist real numbers $L \geq 1$ and $A \geq 0$ such that for any $x_1, x_2 \in X$, there is an infinite (L, A) -quasigeodesic ray that starts at x_1 and passes through x_2 . We refer to L as the multiplicative constant and A as the additive constant.

We note that if G is an infinite finitely generated group, then G has uniform quasigeodesic ray extension with respect to any finite generating set. This is because, for any $g_1, g_2 \in G$, Theorem 3.3 provides an infinite $(3, 0)$ -quasigeodesic ray γ starting at e and passing through $g_1^{-1}g_2$. Translating γ by g_1 provides an infinite quasigeodesic ray starting at g_1 and passing through g_2 . We record this as:

Corollary 3.5. *Let G be an infinite group and X a finite generating set for G . Then $\Gamma(G, X)$ has the uniform quasigeodesic ray extension property.*

We now prove that *uniform quasigeodesic ray extension* is invariant under quasi-isometry.

Theorem 3.6. *Let (X, d_X) be a metric space having uniform quasigeodesic ray extension with constants L and A . If a metric space (Y, d_Y) is quasi-isometric to (X, d_X) , then there exist $L' \geq 1$ and $A' \geq 0$ such that (Y, d_Y) has uniform quasigeodesic ray extension with constants L' and A' .*

The proof of this theorem requires the following lemma, whose proof is a standard exercise in quasi-isometries and quasigeodesics.

Lemma 3.7. *Let $\rho : \mathbb{N} \rightarrow X$ be an (L, A) -quasigeodesic ray in a metric space X and $f : X \rightarrow Y$ be an (ϵ, λ) -quasi-isometry from X to a metric space Y . Then $\rho' = f \circ \rho : \mathbb{N} \rightarrow Y$ is an (L', A') -quasigeodesic ray for constants L' and A' depending only on L, A, ϵ , and λ .*

We now prove Theorem 3.6.

Proof. Let (X, d_X) be a metric space with (L, A) -uniform quasigeodesic ray extension, (Y, d_Y) a metric space and $f : X \rightarrow Y$ an (L', A', C) -quasi-isometry. Consider $y_1, y_2 \in Y$. Since f is a quasi-isometry, there exist $x_1, x_2 \in X$ such that $d_Y(y_1, f(x_1)) \leq C$ and $d_Y(y_2, f(x_2)) \leq C$. Since X has uniform quasigeodesic ray extension, there is an infinite (L, A) -quasigeodesic γ that starts at x_1 and passes through x_2 . By Lemma 3.7, $\gamma' = f \circ \gamma$ is an infinite (λ, ϵ) -quasigeodesic that starts at $f(x_1)$ and passes through $f(x_2)$ with λ and ϵ depending only on L, A, L' , and A' . Select $l \in \mathbb{N}$ with $\gamma'(l) = f(x_2)$. We define $\gamma'' : \mathbb{N} \rightarrow Y$ by

$$\gamma''(t) = \begin{cases} y_1 & \text{if } t = 1, \\ \gamma'(t-1) & \text{if } 2 \leq t \leq l+1, \\ y_2 & \text{if } t = l+2, \\ \gamma'(l) & \text{if } t = l+3, \\ \gamma'(t-3) & \text{if } t \geq l+4. \end{cases}$$

Setting the multiplicative constant $\lambda' = \lambda$ and the additive constant $\epsilon' = \epsilon + 2C + 3/\lambda$, one can readily verify that γ'' is a (λ', ϵ') -quasigeodesic. Since the constants of γ'' depend only on λ, ϵ , and C , and not the particular y_1 and y_2 selected, (Y, d_Y) has (λ', ϵ') uniform quasigeodesic ray extension. \square

The Švarc–Milnor lemma is usually phrased in terms of a group G , not known beforehand to be finitely generated, acting properly discontinuously by isometries and with compact quotient on a proper geodesic metric space X , as, for example, in [de la Harpe 2000, Theorem IV.B.23]. In this case, one concludes that G is finitely generated and, when endowed with the word metric with respect to a finite generating set, quasi-isometric with X . However, if one already knows G to be finitely generated, one can drop the requirement that X be proper, replace the condition of a cocompact action with the existence of a ball $B_{x_0}(R)$ of finite radius whose G -translates cover X , and rephrase a properly discontinuous action as one such that, for every $r > 0$, the set $\{g : B_{x_0}(r) \cap g \cdot B_{x_0}(r) \neq \emptyset\}$ is finite. If the action of G satisfies the above conditions, then G is quasi-isometric with X . In this case, by Corollary 3.5 and Theorem 3.6, one may also conclude that X has uniform quasigeodesic ray extension. We formalize this in our final corollary.

Corollary 3.8. *Let (X, d_X) be a metric space in which any two points can be joined by a geodesic segment and G a finitely generated group acting by isometries on X . If there exists a ball $B_{x_0}(R)$ in X whose G -translates cover X with the property that for every $r > 0$ the set $\{g : B_{x_0}(r) \cap g \cdot B_{x_0}(r) \neq \emptyset\}$ is finite then X has uniform quasigeodesic ray extension.*

Acknowledgment

The authors express thanks to Dr. Elizabeth Weaver, the production editor, for her generous help in improving the flow and exposition of the writing.

References

- [Bogopol'skiĭ 1997] O. V. Bogopol'skiĭ, "Infinite commensurable hyperbolic groups are bi-Lipschitz equivalent", *Algebra i Logika* **36**:3 (1997), 259–272, 357. In Russian; translated in *Algebra and Logic* **36**:3 (1997), 155–163. MR 1485595 Zbl 0988.53001
- [Bridson and Haefliger 1999] M. R. Bridson and A. Haefliger, *Metric spaces of non-positive curvature*, Grundlehren der Mathematischen Wissenschaften **319**, Springer, Berlin, 1999. MR 2000k:53038 Zbl 0988.53001
- [de la Harpe 2000] P. de la Harpe, *Topics in geometric group theory*, University of Chicago Press, 2000. MR 1786869 Zbl 0965.20025
- [Hosaka 2012] T. Hosaka, "On splitting theorems for CAT(0) spaces and compact geodesic spaces of non-positive curvature", *Math. Z.* **272**:3–4 (2012), 1037–1050. MR 2995154 Zbl 1293.20040
- [Lehnert 2009] J. Lehnert, "Some remarks on depth of dead ends in groups", *Internat. J. Algebra Comput.* **19**:4 (2009), 585–594. MR 2536193 Zbl 1191.20027
- [Lyons et al. 1996] R. Lyons, R. Pemantle, and Y. Peres, "Random walks on the lamplighter group", *Ann. Probab.* **24**:4 (1996), 1993–2006. MR 1415237 Zbl 0879.60004
- [Riley and Warshall 2006] T. R. Riley and A. D. Warshall, "The unbounded dead-end depth property is not a group invariant", *Internat. J. Algebra Comput.* **16**:5 (2006), 969–983. MR 2274725 Zbl 1111.20034
- [Warshall 2011] A. D. Warshall, "A group with deep pockets for all finite generating sets", *Israel J. Math.* **185** (2011), 317–342. MR 2837139 Zbl 1269.20035

Received: 2014-06-13 Revised: 2015-07-19 Accepted: 2015-07-22

bgapinski@wesleyan.edu	<i>Department of Mathematics and Computer Science, Wesleyan University, Middletown, CT 06459, United States</i>
horakmatt@gmail.com	<i>Department of Mathematics, Statistics and Computer Science, University of Wisconsin-Stout, Menomonie, WI 54751, United States</i>
webert2575@my.uwstout.edu	<i>Department of Mathematics, Statistics and Computer Science, University of Wisconsin-Stout, Menomonie, WI 54751, United States</i>

Generalized factorization in $\mathbb{Z}/m\mathbb{Z}$

Austin Mahlum and Christopher Park Mooney

(Communicated by Vadim Ponomarenko)

Generalized factorization theory for integral domains was initiated by D. D. Anderson and A. Frazier in 2011 and has received considerable attention in recent years. There has been significant progress made in studying the relation τ_n for the integers in previous undergraduate and graduate research projects. In 2013, the second author extended the general theory of factorization to commutative rings with zero-divisors. In this paper, we consider the same relation τ_n over the modular integers, $\mathbb{Z}/m\mathbb{Z}$. We are particularly interested in which choices of $m, n \in \mathbb{N}$ yield a ring which satisfies the various τ_n -atomicity properties. In certain circumstances, we are able to say more about these τ_n -finite factorization properties of $\mathbb{Z}/m\mathbb{Z}$.

1. Introduction and background

D. D. Anderson and A. Frazier [2011] introduced a concept called τ -factorization. This provided a general theory which unified much of the existing literature on factorization theory in integral domains into one general notion of factorization theory. Recently, the second author has used several methods to extend this τ -factorization to commutative rings with zero-divisors; see [Mooney 2015a, 2015b; 2015c; 2016].

There has been a fair amount of research done on a particular τ -relation of interest especially in the integers, \mathbb{Z} . We discuss this in more depth in the following section. In particular, the dissertation of S. M. Hamon [2007] answered the following question, among others: for what $n \in \mathbb{N}$ is \mathbb{Z} τ_n -atomic? A. Florescu [2013] investigated reduced τ_n -factorizations over \mathbb{Z} . These studies helped to give a concrete basis for τ -factorization over the integers.

In this paper, we carry out a similar investigation of $\mathbb{Z}/m\mathbb{Z}$. We again are interested in the τ_n -finite factorization properties, especially the question of τ_n -atomicity. We use the definitions and methods established by D. D. Anderson and S. Valdez-Leon [1996] and generalized by the second author [Mooney 2015a]. In Section 2, we present preliminary definitions and background information in a more rigorous

MSC2010: 13A05, 13E99, 13F15.

Keywords: modular integers, generalized factorization, zero-divisors, commutative rings.

and thorough manner. In Section 3, we present several important properties of $\mathbb{Z}/m\mathbb{Z}$ which play a role in the τ_n -finite factorization properties. In Section 4, we present the main results concerning τ_n -finite factorization properties of $\mathbb{Z}/m\mathbb{Z}$ for various choices of m and n . Finally, in Section 5, we present further thoughts on the remaining questions which were not answered in the present article.

2. Preliminaries

We assume R is a commutative ring with $1 \neq 0$. Let $R^* = R - \{0\}$, $U(R)$ be the set of units of R , and $R^\# = R^* - U(R)$ be the nonzero nonunits of R . As in [Anderson and Valdes-Leon 1996], we let

- $a \sim b$ if $(a) = (b)$,
- $a \approx b$ if there exists $\lambda \in U(R)$ such that $a = \lambda b$,
- $a \cong b$ if (1) $a \sim b$ and (2) $a = b = 0$ or if $a = rb$ for some $r \in R$ then $r \in U(R)$.

We say a and b are *associates* (resp. *strong associates*, *very strong associates*) if $a \sim b$ (resp. $a \approx b$, $a \cong b$). As in [Anderson et al. 2004], a ring R is said to be a *strongly associate* (resp. *very strongly associate*) ring if for any $a, b \in R$, $a \sim b$ implies $a \approx b$ (resp. $a \cong b$).

We leave the routine check that very strong associates are strong associates and strong associates are associates as an exercise for the reader. Both \sim and \approx are equivalence relations, while \cong fails only to be reflexive. It is interesting to see why, in rings with zero-divisors, these associate relations are no longer equivalent. Any nontrivial idempotent $e \in R$ provides an example of an element such that $e \approx e$, but $e \not\cong e$. We have $e = 1 \cdot e$, yet $e \not\cong e$ because e is not a unit in $e = e \cdot e$. This also demonstrates why \cong need not be reflexive. Examples of elements which are associate, but not strongly associate are more difficult to come by. We provide an example first given in [Fletcher 1969] and restated in [Anderson and Valdes-Leon 1996, Example 2.3], where the details are provided. Let $R = F[X, Y, Z]/(X - XYZ)$, where F is a field. Let x, y , and z be the images of X, Y , and Z respectively in R . Then $x = xyz$, so $x \sim xy$, but there is no unit $\lambda \in U(R)$ such that $x = \lambda xy$, so $x \not\approx xy$.

Let τ be a symmetric relation on $R^\#$; that is, $\tau \subseteq R^\# \times R^\#$ and if $(a, b) \in \tau$, then $(b, a) \in \tau$ and we will write $a \tau b$. For nonunits $a, a_i \in R$, and $\lambda \in U(R)$, $a = \lambda a_1 \cdots a_n$ is said to be a τ -factorization if $a_i \tau a_j$ for all $i \neq j$. If $n = 1$, then this is said to be a trivial τ -factorization. Given the above τ -factorization, we would say that a_i is a τ -factor of a or write $a_i |_\tau a$. We note that 0 cannot appear as a τ -factor, except in the trivial factorization $0 = \lambda 0$ for some $\lambda \in U(R)$.

We pause to provide some examples of τ -relations which have been of interest in the literature.

Example 2.1. Let R be a commutative ring with 1.

(1) $\tau_d = R^\# \times R^\#$. This yields the usual factorizations in R and $|\tau_d$ is the same as the usual divides.

(2) $\tau = \emptyset$. For every $a \in R^\#$, there is only the trivial factorization and $a |_\tau b \iff a = \lambda b$ for $\lambda \in U(R) \iff a \approx b$.

(3) Let I be an ideal in R . Set $a \tau_I b$ if and only if $a - b \in I$.

(a) Let $R = \mathbb{Z}$ and $I = (n)$. Then this is τ_n , which was studied extensively in [Florescu 2013; Hamon 2007].

(b) In the present work, we are interested in the case when $R = \mathbb{Z}/m\mathbb{Z}$ and $I = (n)$. We note that $\tau_{(n)}$ is usually written as τ_n and this relation is indeed symmetric since $a - b \in I \iff b - a \in I$.

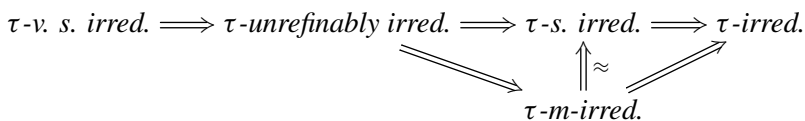
(4) We obtain the comaximal factorizations studied in [McAdam and Swan 2004] by $a \tau b$ if and only if $(a, b) = R$. Furthermore, for any \star -operation, we obtain \star -comaximal factorizations, studied in [Juett 2012], by $a \tau_\star b$ if and only if $(a, b)^\star = R$.

(5) Lastly, for any set S , such as the collection of irreducible or prime elements in a ring R , we can study τ_S -factorizations to obtain the atomic or prime factorizations respectively by saying $a \tau_S b$ if and only if $a \in S$ and $b \in S$.

We now summarize several definitions given in [Mooney 2015a; 2016]. Let $a \in R$ be a nonunit. Then a is said to be τ -irreducible or τ -atomic if for any τ -factorization $a = \lambda a_1 \cdots a_n$, we have $a \sim a_i$ for some i . We say a is τ -strongly irreducible or τ -strongly atomic if for any τ -factorization $a = \lambda a_1 \cdots a_n$, we have $a \approx a_i$ for some a_i . We say that a is τ - m -irreducible or τ - m -atomic if for any τ -factorization $a = \lambda a_1 \cdots a_n$, we have $a \sim a_i$ for all i . Note: the “ m ” is for “maximal” since such an a is maximal among principal ideals generated by elements which occur as τ -factors of a . As in [Mooney 2016], $a \in R$ is said to be a τ -unrefinable atom if a admits only trivial τ -factorizations. We say that a is τ -very strongly irreducible or τ -very strongly atomic if $a \cong a$ and a has no nontrivial τ -factorizations. We refer the reader to [Mooney 2015a; 2016] for a further discussion and more equivalent definitions of these various forms of τ -irreducibility.

We have the following relationship between the various types of τ -irreducibles, which is proved in [Mooney 2015a, Theorem 3.9] as well as [Mooney 2016].

Theorem 2.2. *The following diagram illustrates the relationships between the various types of τ -irreducibility a might satisfy, where \approx represents R being a strongly associate ring:*



Let e be a nontrivial idempotent in R . Let $\tau_\emptyset = \emptyset$. Then there are no non-trivial τ_\emptyset -factorizations. Thus every $a \in R^\#$ is τ_\emptyset -unrefinably atomic. However, $e \cdot e = e$ shows that $e \not\approx e$ and thus e is not τ_\emptyset -very strongly atomic. To see that none of the other reverse implications hold, we may set $\tau = R^\# \times R^\#$ to obtain the usual factorizations. Examples are provided in [Anderson and Valdes-Leon 1996] which show that the other implications are not reversible in rings with zero-divisors.

We are now able to summarize various τ_n -finite factorization properties that a ring may have.

Definition 2.3. Let $\alpha \in \{\text{atomic, strongly atomic, m-atomic, unrefinably atomic, very strongly atomic}\}$. Let $\beta \in \{\text{associate, strongly associate, very strongly associate}\}$.

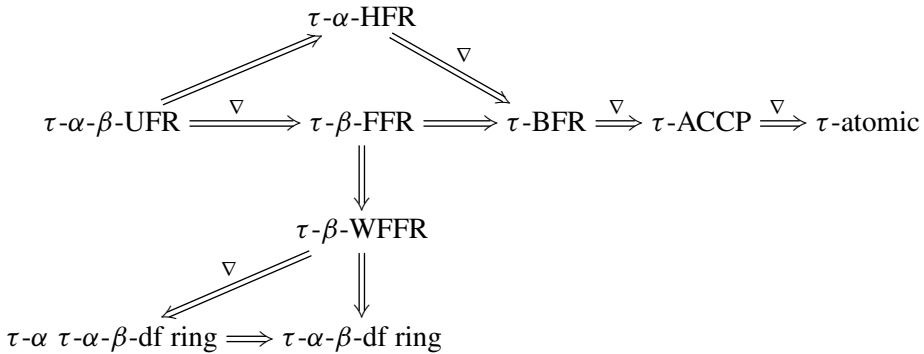
- (1) R is said to be τ - α if every nonunit has a τ -factorization into elements which are τ - α .
- (2) R is said to satisfy τ_n -ACCP if for every nonunit $a_0 \in R$, any ascending chain of principal ideals

$$(a_0) \subseteq (a_1) \subseteq (a_2) \subseteq \cdots \subseteq (a_i) \subseteq (a_{i+1}) \subseteq \cdots$$

such that $a_{i+1} \mid_\tau a_i$ for each i becomes stationary.

- (3) R is said to be a τ_n - α - β -unique factorization ring (UFR) if
 - R is τ_n - α ,
 - every nonunit has a unique τ_n - α factorization up to rearrangement and β .
- (4) R is said to be a τ_n - α -half factorial ring (HFR) if R is τ - α and for each nonunit, the length of every τ_n - α factorization is the same.
- (5) R is said to be a τ_n -bounded factorization ring (BFR) if every nonunit has a finite bound on the length of any τ_n -factorization.
- (6) R is said to be a τ_n - β -finite factorization ring (FFR) if every nonunit has only a finite number of τ_n -factorizations up to rearrangement and β .
- (7) R is said to be a τ_n - β -weak finite factorization ring (WFFR) if every nonunit has only a finite number of τ_n -divisors up to β .
- (8) R is said to be a τ_n - α - β -divisor finite ring (df ring) if every nonunit has only a finite number of τ_n - α -divisors up to β .

We include parts of the diagram from [Mooney 2016] to help the reader visualize the relationship between these τ -finite factorization properties. In the diagram below, ∇ represents τ being refinable and associate-preserving and we direct the reader to [Mooney 2016] for further details:



3. $\mathbb{Z}/m\mathbb{Z}$ is strongly associate

We begin by studying the ring we are interested in, $\mathbb{Z}/m\mathbb{Z}$. As seen in the previous section, the main issue with factorization in rings with zero-divisors is the number of types of irreducibility and atomicity. We find that this ring has several nice properties, which makes our work slightly more manageable. We find that $\mathbb{Z}/m\mathbb{Z}$ is a strongly associate ring and if p is a prime and $e \in \mathbb{N}$, then $\mathbb{Z}/p^e\mathbb{Z}$ is présimplifiable. Equivalently, $\mathbb{Z}/p^e\mathbb{Z}$ is a very strongly associate ring. So if m is a prime power, then for any $a \in R^\#$, all the associate relations and hence types of τ -irreducibility coincide. In general, even if m has multiple prime divisors, we will know that associate and strongly associate coincide; hence τ_n -atomic and τ_n -strongly atomic also coincide.

It was proved, in [Kaplansky 1949], that any Artinian or principal ideal ring is strongly associate. This immediately gives us that our finite (hence Artinian) principal ideal ring, $\mathbb{Z}/m\mathbb{Z}$, is strongly associate. We outline an elementary proof for $\mathbb{Z}/m\mathbb{Z}$ being strongly associate as well as present other useful results about $\mathbb{Z}/m\mathbb{Z}$. We hope this is helpful for the reader, both to become familiar with the ring we are working in and to see the relationships between the various types of associate relations. Many of these results and similar techniques are used later when we analyze the question of τ_n -atomicity of $\mathbb{Z}/m\mathbb{Z}$.

We begin with a remark about the units of a direct product of commutative rings. This is a routine result, which can be found in any modern algebra text, and will be left as an exercise to the reader.

Remark. Let R_1 and R_2 be commutative rings with unity and let $R = R_1 \times R_2$. Then

$$U(R) = \{(\lambda_1, \lambda_2) \mid \lambda_1 \in U(R_1), \lambda_2 \in U(R_2)\} = U(R_1) \times U(R_2) := S.$$

That is, the units in a direct product of rings are the direct product of the collection of units in the individual rings.

Lemma 3.1. $R = R_1 \times R_2$ is strongly associate if and only if R_1 and R_2 are both strongly associate.

Proof. (\Rightarrow) Let $R = R_1 \times R_2$ be a strongly associate ring. Let $(a), (b)$ be ideals in R_1 such that $a \sim b$, i.e., $(a) = (b)$. Consider the ideals $(a) \times R_2 = (a) \times (1)$ and $(b) \times R_2 = (b) \times (1)$. Since $(a) = (b)$, we have

$$((a, 1)) = (a) \times (1) = (b) \times (1) = ((b, 1)).$$

Now $R = R_1 \times R_2$ is strongly associate, so there is a unit $(\lambda_1, \lambda_2) \in U(R)$ such that $(a, 1) = (\lambda_1, \lambda_2)(b, 1)$. Thus $a = \lambda_1 b$. By the above remark, we have shown that $\lambda_1 \in U(R_1)$. Hence $a \approx b$. A symmetric argument demonstrates that R_2 is strongly associate.

(\Leftarrow) Now suppose R_1 and R_2 are strongly associate rings. Let $a, b \in R$ with $a \sim b$. Suppose $a = (a_1, a_2)$ and $b = (b_1, b_2)$. Now $a \sim b$ means $((a_1, a_2)) = ((b_1, b_2))$. We must prove that there exists a $(\lambda_1, \lambda_2) \in U(R)$ with $(a_1, a_2) = (\lambda_1, \lambda_2)(b_1, b_2)$. Now

$$(a_1) \times (a_2) = ((a_1, a_2)) = ((b_1, b_2)) = (b_1) \times (b_2).$$

Thus a_1 is associate with b_1 and a_2 is associate with b_2 . Hence, R_1 and R_2 are strongly associate, so there exists $\lambda_1 \in U(R_1)$ and $\lambda_2 \in U(R_2)$ such that $a_1 = \lambda_1 b_1$ and $a_2 = \lambda_2 b_2$. Therefore $(\lambda_1, \lambda_2) \in U(R)$ with $(a_1, a_2) = (\lambda_1, \lambda_2)(b_1, b_2)$. This demonstrates R is strongly associate as desired. \square

A routine induction argument on n , the number of factors in the product, yields the following result since $R = (R_1 \times R_2 \times \cdots \times R_{n-1}) \times R_n = R_1 \times R_2 \times \cdots \times R_n$.

Lemma 3.2. $R = R_1 \times R_2 \times \cdots \times R_n$ is strongly associate if and only if R_i is strongly associate for each $1 \leq i \leq n$.

Lemma 3.3. Let $a_1, \dots, a_n \in R$. Then $(a_1 a_2 \cdots a_n) = (a_1)(a_2) \cdots (a_n)$.

Proof. Let $x \in (a_1)(a_2) \cdots (a_n)$. Then

$$x = r_{11}a_1r_{12}a_2 \cdots r_{1n}a_n + r_{21}a_1r_{22}a_2 \cdots r_{2n}a_n + \cdots + r_{m1}a_1r_{m2}a_2 \cdots r_{mn}a_n$$

for some $r_{ij} \in R$, with $1 \leq i, j \leq m$, is a typical element of $(a_1)(a_2) \cdots (a_n)$. Notice that we can factor out $a_1 a_2 \cdots a_n$ from each term yielding

$$x = (r_{11}r_{12} \cdots r_{1n} + r_{21}r_{22} \cdots r_{2n} + \cdots + r_{m1}r_{m2} \cdots r_{mn})(a_1 a_2 \cdots a_n). \quad (1)$$

The right-hand side of (1) demonstrates that $x \in (a_1 a_2 \cdots a_n)$. Thus $(a_1 a_2 \cdots a_n) \supseteq (a_1)(a_2) \cdots (a_n)$.

Let $x \in (a_1 a_2 \cdots a_n)$. Then $x = ra_1 a_2 \cdots a_n$ for some $r \in R$. Then we can write $x = ra_1 a_2 \cdots a_n = (ra_1)(1a_2) \cdots (1a_n)$, demonstrating $x \in (a_1)(a_2) \cdots (a_n)$. Thus $(a_1 a_2 \cdots a_n) \subseteq (a_1)(a_2) \cdots (a_n)$. \square

Lemma 3.4. Let $p \in \mathbb{N}$ be a prime number and $e \in \mathbb{N}$. Then $R = \mathbb{Z}/p^e\mathbb{Z}$ is very strongly associate; equivalently, $\mathbb{Z}/p^e\mathbb{Z}$ is présimplifiable. Moreover, this means that $\mathbb{Z}/p^e\mathbb{Z}$ is a strongly associate ring.

Proof. Suppose $a \sim b$. We will show $a \cong b$. Since $a \sim b$, we have $(a) = (b)$ by definition. Thus we must prove that either $a = b = 0$ or if $a = rb$ for some $r \in R$ then $r \in U(R)$.

If $a = 0$ or $b = 0$ we are done, so we may assume that neither a nor b is 0. If a or b are units, then $(a) = (b) = R$ and $r = ab^{-1}$, which is a unit. Thus we may assume a and b are nonzero nonunits. Thus $p \mid a$ and $p \mid b$. Let e_a be the largest integer such that p^{e_a} divides a , but no larger power still divides a . Define e_b similarly. Now $(a) = (b)$, so $a \mid b$ and $p^{e_a} \mid a$ and therefore $p^{e_a} \mid b$. This means $e_a \leq e_b$. Similarly, $b \mid a$ so $e_b \leq e_a$. This means $e_a = e_b$, but by comparing the number of factors of p in both sides of $a = rb$, we see that p cannot divide r . Thus $\gcd(r, p) = 1$ and $r \in U(\mathbb{Z}/p^e\mathbb{Z})$. Hence, R has been shown to be a very strongly associate ring, which is equivalent to présimplifiable in the language of Bouvier [1971; 1972a; 1972b; 1974]. Every présimplifiable ring is certainly a strongly associate ring. □

The following theorem now follows easily from the lemmas and the Chinese remainder theorem.

Theorem 3.5. *Let $m \in \mathbb{N}$ with $m \geq 2$ and $m = p_1^{e_1} \cdots p_n^{e_n}$. Then*

$$\mathbb{Z}/m\mathbb{Z} \cong \mathbb{Z}/p_1^{e_1}\mathbb{Z} \times \mathbb{Z}/p_2^{e_2}\mathbb{Z} \times \cdots \times \mathbb{Z}/p_n^{e_n}\mathbb{Z}$$

is a strongly associate ring.

This means associate and strongly associate are always the same relation and hence τ_n -atomic and τ_n -strongly atomic coincide for our rings $\mathbb{Z}/m\mathbb{Z}$. We also needed R to be a strongly associate ring to conclude that τ_n -m-atomic implies τ_n -strongly atomic in Theorem 2.2. We find that this property of $\mathbb{Z}/m\mathbb{Z}$ greatly streamlines much of the research.

4. τ_n -factorization properties of $\mathbb{Z}/m\mathbb{Z}$

Here we begin our analysis of which choices of $m, n \in \mathbb{N}$ yield a τ_n -atomic (or -strongly atomic, -m-atomic, -unrefinably atomic, -very strongly atomic) ring. Moreover, when possible, we indicate if the ring satisfies other nice τ_n -finite factorization properties.

$\mathbb{Z}/p\mathbb{Z}$. We first consider the simplest case, $R = \mathbb{Z}/p\mathbb{Z}$ when p is prime.

Lemma 4.1. *Let $p \in \mathbb{N}$ be a prime number. Then $R = \mathbb{Z}/p\mathbb{Z}$ is a field.*

Proof. Let $a \in R^*$. Then $\gcd(a, p) = 1$, so by the Euclidean algorithm, there are integers $s, t \in \mathbb{Z}$ such that $as + pt = 1$. When reduced modulo p , we see that $as \equiv 1 \pmod{p}$. Thus $\mathbb{Z}/p\mathbb{Z}$ is a commutative ring with unity such that every nonzero element is a unit. Thus $\mathbb{Z}/p\mathbb{Z}$ is a field. □

Theorem 4.2. *Let $p \in \mathbb{N}$ be prime and set $R = \mathbb{Z}/p\mathbb{Z}$. Let $\alpha \in \{\text{atomic, strongly atomic, } m\text{-atomic, unrefinably atomic, very strongly atomic}\}$. Let $\beta \in \{\text{associate, strongly associate, very strongly associate}\}$. Then for any $n \in \mathbb{N}$, we have:*

- (1) R is τ_n - α .
- (2) R satisfies τ_n -ACCP.
- (3) R is a τ_n -BFR.
- (4) R is a τ_n - α - β -UFR.
- (5) R is a τ_n - α -HFR.
- (6) R is a τ_n - β -FFR.
- (7) R is a τ_n - β -WFFR.
- (8) R is a τ_n - α - β -df ring.

Proof. (1) Let $a \in R$ with a a nonunit. Then by Lemma 4.1, $a = 0$ since all nonzero elements are units in a field. The only τ_n -factorizations are $0 = \lambda 0$ since there are no other nonzero nonunits. Furthermore, R is a field, so (0) is a maximal ideal and therefore 0 is m -irreducible and thus τ_n - m -irreducible. Fields are integral domains, which are présimplifiable, so all of the other forms of τ_n - α coincide. Thus R is τ_n - α .

(2) The only proper ideal is (0) since R is a field, so it certainly satisfies ACCP and therefore τ_n -ACCP.

(3) There are no nonzero nonunits, so there can be no nontrivial τ_n -factorizations. Thus all τ_n -factorizations are trivial and have length 1, making R a τ_n -BFR.

(4)–(6) We know R is τ_n - α by (1). Moreover, 0 has only $0 = \lambda 0$ as a τ_n -factorization. Since R is a field, $0 \cong 0$, so we see this is the only factorization up to rearrangement and β . Hence R is a τ_n - α - β -UFR and a τ_n - α -HFR. Again, this is the only τ_n -factorization, not just the only τ_n - α factorization, so R is certainly a τ_n - β -FFR.

(7)–(8) R is a finite ring with p elements. Hence there are a finite number of τ_n - and τ_n - α -divisors in the whole ring. Thus R is a τ_n - β -WFFR and a τ_n - α - β -df ring. \square

$\mathbb{Z}/p^e\mathbb{Z}$, **where $e > 1$.** For $\mathbb{Z}/m\mathbb{Z}$, with $m = p^e$ (where $e \in \mathbb{N}$ and p is prime), we found that $\mathbb{Z}/p^e\mathbb{Z}$ is présimplifiable, or equivalently very strongly associate. As in [Mooney 2016], we have the following, which we state without proof.

Lemma 4.3. *Let R be a présimplifiable ring. Let $a \in R^\#$ be a nonzero nonunit. Then the following are equivalent:*

- (1) a is τ_n -atomic.
- (2) a is τ_n -strongly atomic.
- (3) a is τ_n - m -atomic.
- (4) a is τ_n -unrefinably atomic.
- (5) a is τ_n -very strongly atomic.

Lemma 4.4. *Let $R = \mathbb{Z}/p^e\mathbb{Z}$, where $p, e, n \in \mathbb{N}$ and p is prime. Then p is τ_n - m -atomic and therefore p is τ_n -atomic (-strongly atomic, - m -atomic, -unrefinably atomic, -very strongly atomic).*

Proof. Let $p \in R = \mathbb{Z}/p^e\mathbb{Z}$. We show that (p) is maximal. The following are equivalent:

- An element $a \in \mathbb{Z}/p^e\mathbb{Z}$ is a unit.
- $\gcd(a, p^e) = 1$.
- $\gcd(a, p) = 1$.
- p does not divide a .
- $a \notin (p)$.

Thus (p) is precisely the set of nonunits. If $J \supsetneq (p)$, then let $x \in J \setminus (p)$. Then p does not divide x , so $x \in J$ is a unit, and so $J = R$. This shows that (p) is a maximal ideal (not just among principal ideals). Thus p is m -atomic and therefore τ_n - m -atomic. Moreover, by Lemma 4.3 this means p is τ_n -atomic (-strongly atomic, - m -atomic, -unrefinably atomic, -very strongly atomic). \square

Proposition 4.5. *Let $p, e, n \in \mathbb{N}$, where p is prime and $e > 1$. The only τ_n -atomic (-strongly atomic, - m -atomic, -unrefinably atomic, -very strongly atomic) elements of $R = \mathbb{Z}/p^e\mathbb{Z}$ are p and unit multiples of p .*

Proof. Let $a \in R$ be a τ_n -irreducible (equivalently, -strongly atomic, - m -atomic, -unrefinably atomic, -very strongly atomic) element. Since a must be a nonunit, we know $\gcd(a, p) = p > 1$. Therefore, $p \mid a$. Let j be the largest number of factors of p that we can factor out of a . That is, let j be the integer such that p^j divides a , but p^{j+1} does not divide a . Write $a = \lambda p^j$. Then $\gcd(\lambda, p) = 1$ or else $p^{j+1} \mid a$. This means $\lambda \in U(R)$. If $j > 1$, then $a = \lambda \cdot p^j = \lambda \cdot p \cdots p$ is a τ_n -factorization of a such that $(a) \neq (p)$. This means a is not τ_n -atomic and therefore a is also not τ_n -strongly atomic (- m -atomic, -unrefinably atomic, -very strongly atomic). Thus, $j = 1$ and $a = \lambda p$, showing any τ_n -atomic (or -strongly atomic, - m -atomic, -unrefinably atomic, -very strongly atomic) element of $R = \mathbb{Z}/p^e\mathbb{Z}$ must be a unit multiple of p . \square

Theorem 4.6. *Let $R = \mathbb{Z}/p^e\mathbb{Z}$, where $p, e, n \in \mathbb{N}$ and p is prime. Then we have the following:*

- (1) R is τ_n -atomic.
- (2) R is τ_n -strongly atomic.
- (3) R is τ_n - m -atomic.
- (4) R is τ_n -unrefinably atomic.
- (5) R is τ_n -very strongly atomic.

Proof. Let $a \in R$ be a nonunit. If a is not a unit, then $\gcd(a, p) > 1$; hence $p \mid a$. We let j represent the integer for which $p^j \mid a$, but p^{j+1} does not divide a . Thus $a = p^j \cdot \lambda$ for some $\lambda \in \mathbb{N}$. Moreover, p does not divide λ , so $\gcd(\lambda, p) = 1$ and λ is a unit. Then $a = \lambda p \cdots p$, where p occurs j times. Certainly $p \tau_n p$ for any $n \in \mathbb{N}$ since $p - p = 0 \in (0) \subseteq I$ for any ideal I . Thus we have found a τ_n -atomic (-strongly atomic, -m-atomic, -unrefinably atomic, -very strongly atomic) factorization of a by Lemma 4.3. \square

Proposition 4.7. *Let $R = \mathbb{Z}/p^e\mathbb{Z}$, where $p, e, n \in \mathbb{N}$ and p is prime. Let $\alpha \in \{\text{atomic, strongly atomic, m-atomic, unrefinably atomic, very strongly atomic}\}$ and let $\beta \in \{\text{associate, strongly associate, very strongly associate}\}$. Then we have the following:*

- (1) R is a τ_n - β -WFFR.
- (2) R is a τ_n - α - β -idf ring.
- (3) R satisfies τ_n -ACCP.

Proof. This is immediate again since R is a finite ring. \square

Remark. We note here that this ring nearly satisfies further τ_n -finite factorization properties; however, we have the following issue. For any $j \geq e$, we have $0 = p \cdots p = p^j$ is a τ_n -atomic (-strongly atomic, -m-atomic, -unrefinably atomic, -very strongly atomic) factorization of 0. This means that R fails to be a τ_n -BFR (or $-\alpha$ -HFR, $-\alpha$ - β -UFR, $-\beta$ -FFR). We do, on the other hand, have some positive results for nonzero elements of $\mathbb{Z}/p^e\mathbb{Z}$.

Theorem 4.8. *Let $p, e, n \in \mathbb{N}$, where p is prime. Let $\alpha \in \{\text{atomic, strongly atomic, m-atomic, unrefinably atomic, very strongly atomic}\}$. Let $\beta \in \{\text{associate, strongly associate, very strongly associate}\}$. Let $a \in \mathbb{Z}/p^e\mathbb{Z}$, a nonzero nonunit. Then we have the following:*

- (1) Any two τ_n - α factorizations of a have the same length.
- (2) The element a not only has a τ_n - α factorization, but it is unique up to rearrangement and β .
- (3) The element a has a finite number of τ_n -factorizations up to rearrangement and β .
- (4) There is a bound on the length of any τ_n -factorization of a .

Proof. (1) Let $a \in R$ be a nonzero nonunit. We know by Theorem 4.6 that there is a τ_n - α factorization of a . As Proposition 4.5 demonstrated, p and unit multiples of p are the only τ_n - α elements in $\mathbb{Z}/p^e\mathbb{Z}$. Recall that from the construction of the τ_n - α factorization in Theorem 4.6, j is the unique integer such that $p^j \mid a$, but p^{j+1} does not divide a . It is clear then that any τ_n - α factorization of a must have precisely j factors, each being some unit multiple of p .

(2) By Proposition 4.5, the only τ_n - α elements are unit multiples of p . Now $\mathbb{Z}/p^e\mathbb{Z}$ is présimplifiable, so all choices of β are equivalent. Thus since all τ_n - α factorizations have the same length and all τ_n - α elements are β , it is clear that this τ_n - α factorization of a is unique.

(3) Since any τ_n -factorization of a is certainly a factorization of a , it suffices to show that there are only finitely many factorizations of a up to β . Again, let j be as in (1). We claim that j is the largest number of nonunit factors that any factorization can have. If each factor is a nonunit, then it must be divisible by p . By the definition of j , we have $p^j \mid a$, but p^{j+1} does not divide a . Thus there can be no more than j factors in any given factorization of a . In this way, all factorizations of a must come as some grouping of the j factors of p or some unit multiple of p . Hence the number of distinct factorizations up to β is certainly bounded by 2^j . A better bound would be $P(j)$, where $P(n)$ is the number of partitions of a set with n elements.

(4) Since there are only a finite number of τ_n -factorizations up to β , we can simply take the maximum length of these factorizations as the bound on the length of τ_n -factorizations of a . Alternatively, it is clear that j , as defined in the unique factorization in (1), is the longest possible τ_n -factorization since any other τ_n -factorization could be refined into this τ_n - α factorization and it would be at least as long. \square

The above theorem shows that 0 is the only element preventing $\mathbb{Z}/p^e\mathbb{Z}$ from being a τ_n - α - β -UFR (or $-\alpha$ -HFR, $-\beta$ -FFR, $-\beta$ -BFR).

$\mathbb{Z}/m\mathbb{Z}$. When m has multiple distinct prime divisors, matters become more complicated. There are now nontrivial idempotent elements. For instance, consider $\mathbb{Z}/6\mathbb{Z}$ and the element 3. We can factor $3 = 3 \cdot 3 = 3 \cdot 3 \cdot 3 = \dots$. Often the solution to dealing with issues that arise from idempotents is using U-factorization, as in [Mooney 2015b]. We are still able to say a few things about certain finite factorization properties in the affirmative, but further research will need to be conducted to completely answer this question.

We begin with a known result which sheds some light on the situation. If $\gcd(n, m) = 1$, then $(n) = R$ and we have the usual factorization since $\tau_n = \tau_d$, where $\tau_d = R^\# \times R^\#$ yields the usual factorizations. This situation was discussed in [Anderson and Valdes-Leon 1996] and we refer the reader here for the traditional case.

Proposition 4.9. *Let $R = \mathbb{Z}/m\mathbb{Z}$, where $m, n \in \mathbb{N}$. Let $\alpha \in \{\text{atomic, strongly atomic, } m\text{-atomic, unrefinably atomic, very strongly atomic}\}$. Let $\beta \in \{\text{associate, strongly associate, very strongly associate}\}$. Then we have the following:*

- (1) R is a τ_n - β -WFFR.
- (2) R is a τ_n - α - β -idf ring.
- (3) R satisfies τ_n -ACCP.

Proof. This is immediate again since R is a finite ring. \square

Theorem 4.10. *Let $\alpha \in \{\text{atomic, strongly atomic, } m\text{-atomic, unrefinably atomic, very strongly atomic}\}$ and $\beta \in \{\text{associate, strongly associate, very strongly associate}\}$. Let $R = \mathbb{Z}/p_1^{e_1} p_2^{e_2} \cdots p_k^{e_k} \mathbb{Z}$, where $p_i, e_i, n, k \in \mathbb{N}$ with p_i primes. Then we have the following:*

- (1) *If $k = 1$, then R is as in the previous subsection.*
- (2) *If $e_i \neq 1$ for at least one i and $k > 1$, then we have the following:*
 - (a) *R fails to be a τ_n -BFR.*
 - (b) *R fails to be a τ_n - β -FFR.*
 - (c) *R fails to be a τ_n - α -HFR.*
 - (d) *R fails to be a τ_n - α - β -UFR.*
- (3) *If $e_i = 1$ for all $1 \leq i \leq k$, then R is a direct product of fields and we have the following:*
 - (a) *R is not τ_n -unrefinably atomic (or -very strongly atomic).*
 - (b) *R fails to be a τ_n -BFR.*
 - (c) *R fails to be a τ_n - β -FFR.*
 - (d) *R fails to be a τ_n - α -HFR.*
 - (e) *R fails to be a τ_n - α - β -UFR.*

Proof. (1) is immediate.

(2) After reordering the primes if necessary, we may assume that $e_1 > 1$. Then consider the element $(0, 1, \dots, 1)$ and the τ_n -factorizations

$$(0, 1, \dots, 1) = (p, 1, \dots, 1) \cdots (p, 1, \dots, 1) = (p, 1, \dots, 1)^j,$$

where $j \geq e_1$. We notice that this is indeed a τ_n -factorization for any choice of ideal (n) since $(p, 1, \dots, 1) - (p, 1, \dots, 1) = (0, 0, \dots, 0) \in (n)$. Furthermore, $(p, 1, \dots, 1)$ is both regular (not a zero-divisor) and generates a principal ideal which is maximal. This means $(p, 1, \dots, 1)$ is τ_n - α and we have demonstrated arbitrarily long τ_n - α factorizations of a nonunit. This proves R is not a τ_n -BFR (or $-\beta$ -FFR, $-\alpha$ -HFR, $-\alpha$ - β -UFR).

(3a) We observe that the element $e := (0, 1, \dots, 1)$ is neither τ_n -unrefinably atomic nor τ_n -very strongly atomic. To see this, consider the τ_n -factorization

$$e = (0, 1, \dots, 1) = (0, 1, \dots, 1)(0, 1, \dots, 1).$$

This demonstrates that e is an idempotent and hence $e \not\cong e$. Thus we have found a nontrivial τ_n -factorization of e . We now consider any factorization of e . We have

$$e = (0, 1, \dots, 1) = (a_{11}, a_{12}, \dots, a_{1k})(a_{21}, a_{22}, \dots, a_{2k}) \cdots (a_{t1}, a_{t2}, \dots, a_{tk}).$$

We have $0 = a_{11}a_{21} \cdots a_{t1}$ in $\mathbb{Z}/p_1^{e_1} \mathbb{Z}$, which is a field, so $a_{f1} = 0$ for some $1 \leq f \leq t$. In the other coordinates, we have factorizations of 1, and thus a_{ij} must be a unit for

each i and $j \geq 2$. This tells us that any factorization of e must have a factor of the form $(0, \lambda_2, \dots, \lambda_k)$, where $\lambda_2, \dots, \lambda_k$ are units. But this means

$$e = (0, 1, \dots, 1) = (1, \lambda_2^{-1}, \dots, \lambda_k^{-1})(0, \lambda_2, \dots, \lambda_k).$$

This factor is a strong associate of e which is neither τ_n -unrefinably atomic nor τ_n -very strongly atomic. Thus there is no possible τ_n -unrefinably atomic or τ_n -very strongly atomic factorization of e . On the other hand, $R/(e) \cong \mathbb{Z}/p_1\mathbb{Z}$, which is a field, and R is a strongly associate ring, so e is τ_n -atomic (-strongly atomic, -m-atomic).

(3b–3e) We again consider $e := (0, 1, \dots, 1)$. We observe that $e = e^2 = e^3 = \dots = e^j = \dots$ yields τ_n -factorizations for any $j > 1$. This demonstrates that R is neither a τ_n -FFR nor a τ_n -BFR. Furthermore, this gives τ_n -atomic (-strongly atomic, -m-atomic) factorizations of e of different lengths, proving R is not a τ_n -atomic-(-strongly atomic-, -m-atomic-) HFR or a τ_n -atomic-(-strongly atomic-, -m-atomic-) β -UFR. Lastly, from (3a), we know R is not even τ_n -unrefinably atomic (or -very strongly atomic), so it is certainly not a τ_n -unrefinably atomic- (or -very strongly atomic-) HFR or a τ_n -unrefinably atomic- (or -very strongly atomic-) β -UFR. \square

5. Further thoughts on $\mathbb{Z}/m\mathbb{Z}$ with multiple prime factors

We have answered many questions regarding τ_n -finite factorization properties in the negative; however, there are certainly some remaining open questions. When there are multiple prime divisors, the question of whether $R = \mathbb{Z}/m\mathbb{Z}$ is τ_n -atomic (or -strongly atomic, -m-atomic) appears much more complicated and sensitive to the choice of the ideal picked. Further research would need to be done. Indeed, this question appears difficult even in the integers; see [Florescu 2013; Hamon 2007]. For fixed $n \in \mathbb{Z}$, τ_n -atomicity and τ_n -finite factorization properties, even for small n , have been and continue to be studied in depth in \mathbb{Z} , especially by Reyes M. Ortiz Albino and many of his students at The University of Puerto Rico at Mayagüez. It seems fertile ground for future research.

The fact that $\mathbb{Z}/m\mathbb{Z}$ is strongly associate simplifies (or at least unifies) some of these questions to make it more tractable. The existence of idempotent elements when m has multiple prime divisors suggests that looking at τ -U-factorization, as in [Mooney 2015b], may be a better path to take. The τ -U-factorizations are particularly effective in dealing with direct products of rings. It was often idempotent elements that were preventing the ring from satisfying further τ_n -finite factorization properties. As initiated by C. R. Fletcher [1969; 1970] and studied extensively by M. Axtell, S. Forman, N. Roersma, and J. Stickles [Axtell 2002; Axtell et al. 2003], the method of U-factorizations is helpful for this. When using U-factorization, rings like $\mathbb{Z}/6\mathbb{Z}$ go from not being even bounded factorization rings ($3 = 3^i$ for all i) to being U-unique factorization rings.

Acknowledgments

The authors would like to thank Viterbo University, in particular, the Viterbo Summer Research Fellowship Program which provided the funding to carry out this research in the summer of 2014. Mooney would also like to acknowledge the work done at The University of Iowa with the VIGRE REU program under the supervision of Professor Daniel D. Anderson, which tackled this problem over \mathbb{Z} and provided the inspiration for this particular study over $\mathbb{Z}/m\mathbb{Z}$ as a possible project suitable for undergraduate research. The authors would also like to thank the referee for diligent work and careful reading of the article. Their suggestions have improved the quality of the article.

References

- [Anderson and Frazier 2011] D. D. Anderson and A. M. Frazier, “On a general theory of factorization in integral domains”, *Rocky Mountain J. Math.* **41**:3 (2011), 663–705. MR 2012g:13003 Zbl 1228.13001
- [Anderson and Valdes-Leon 1996] D. D. Anderson and S. Valdes-Leon, “Factorization in commutative rings with zero divisors”, *Rocky Mountain J. Math.* **26**:2 (1996), 439–480. MR 97h:13001 Zbl 0865.13001
- [Anderson et al. 2004] D. D. Anderson, M. Axtell, S. J. Forman, and J. Stickles, “When are associates unit multiples?”, *Rocky Mountain J. Math.* **34**:3 (2004), 811–828. MR 2005k:13001 Zbl 1092.13002
- [Axtell 2002] M. Axtell, “U-factorizations in commutative rings with zero divisors”, *Comm. Algebra* **30**:3 (2002), 1241–1255. MR 2003d:13001 Zbl 1046.13002
- [Axtell et al. 2003] M. Axtell, S. Forman, N. Roersma, and J. Stickles, “Properties of U-factorizations”, *Int. J. Commut. Rings* **2**:2 (2003), 83–99. MR 2005j:13003 Zbl 1120.13001
- [Bouvier 1971] A. Bouvier, “Sur les anneaux de fractions des anneaux atomiques présimplifiables”, *Bull. Sci. Math.* (2) **95** (1971), 371–377. MR 45 #6810 Zbl 0219.13020
- [Bouvier 1972a] A. Bouvier, “Anneaux présimplifiables”, *C. R. Acad. Sci. Paris Sér. A-B* **274** (1972), A1605–A1607. MR 45 #6797 Zbl 0244.13009
- [Bouvier 1972b] A. Bouvier, “Résultats nouveaux sur les anneaux présimplifiables”, *C. R. Acad. Sci. Paris Sér. A-B* **275** (1972), A955–A957. MR 47 #4982 Zbl 0242.13002
- [Bouvier 1974] A. Bouvier, “Anneaux présimplifiables”, *Rev. Roumaine Math. Pures Appl.* **19** (1974), 713–724. MR 52 #13811 Zbl 0289.13010
- [Fletcher 1969] C. R. Fletcher, “Unique factorization rings”, *Proc. Cambridge Philos. Soc.* **65** (1969), 579–583. MR 39 #189 Zbl 0174.33401
- [Fletcher 1970] C. R. Fletcher, “The structure of unique factorization rings”, *Proc. Cambridge Philos. Soc.* **67** (1970), 535–540. MR 40 #5596 Zbl 0192.38401
- [Florescu 2013] A. A. Florescu, *Reduced $\tau_{(n)}$ factorizations in \mathbb{Z} and $\tau_{(n)}$ -factorizations in \mathbb{N}* , Ph.D. thesis, University of Iowa, 2013, available at <http://search.proquest.com/docview/1444307443>.
- [Hamon 2007] S. M. Hamon, *Some topics in τ -factorizations*, Ph.D. thesis, University of Iowa, 2007, available at <http://search.proquest.com/docview/304860971>.
- [Juett 2012] J. Juett, “Generalized comaximal factorization of ideals”, *J. Algebra* **352** (2012), 141–166. MR 2862178 Zbl 1253.13005

- [Kaplansky 1949] I. Kaplansky, “Elementary divisors and modules”, *Trans. Amer. Math. Soc.* **66** (1949), 464–491. MR 11,155b Zbl 0036.01903
- [McAdam and Swan 2004] S. McAdam and R. G. Swan, “Unique comaximal factorization”, *J. Algebra* **276**:1 (2004), 180–192. MR 2004m:13006 Zbl 1081.13008
- [Mooney 2015a] C. P. Mooney, “Generalized factorization in commutative rings with zero-divisors”, *Houston J. Math.* **41**:1 (2015), 15–32. MR 3347935 Zbl 06522510
- [Mooney 2015b] C. P. Mooney, “Generalized U-factorization in commutative rings with zero-divisors”, *Rocky Mountain J. Math.* **45**:2 (2015), 637–660. MR 3356632 Zbl 06475249
- [Mooney 2015c] C. P. Mooney, “ τ -regular factorization in commutative rings with zero-divisors”, preprint, 2015, available at <http://projecteuclid.org/euclid.rmjm/1411945723>. To appear in *Rocky Mountain J. Math.*
- [Mooney 2016] C. P. Mooney, “ τ -complete factorization in commutative rings with zero-divisors”, *Houston J. Math.* **42**:1 (2016), 23–44.

Received: 2014-09-27 Revised: 2015-04-07 Accepted: 2015-06-06

amahlu04769@viterbo.edu *Department of Mathematics, Viterbo University,
La Crosse, WI 54601, United States*

christopher.mooney@westminster-mo.edu *Department of Mathematics, Westminster College,
Fulton, MO 65251, United States*

Cocircular relative equilibria of four vortices

Jonathan Gomez, Alexander Gutierrez,
John Little, Roberto Pelayo and Jesse Robert

(Communicated by Martin Bohner)

We study the cocircular relative equilibria (planar central configurations) in the four-vortex problem using methods suggested by the study of cocircular central configurations in the Newtonian four-body problem in recent work of Cors and Roberts. Using mutual distance coordinates, we show that the set of four-vortex relative equilibria is a two-dimensional surface with boundary curves representing kite configurations, isosceles trapezoids, and degenerate configurations with one zero vorticity. We also show that there is a constraint on the signs of the vorticities in these configurations; either three or four of the vorticities must have the same sign, in contrast to the noncocircular cases studied by Hampton, Roberts, and Santoprete.

1. Introduction

Understanding central configurations is a problem of fundamental importance in celestial mechanics (for instance, see [Saari 2011]). Recent years have seen heightened interest in the study of central configurations, in part due to the fact that advances in computing power have made it possible to utilize tools from algebraic geometry to study such problems. These tools have led to breakthroughs such as the proof that there are only finitely many central configurations for each collection of positive masses in the four-body problem [Hampton and Moeckel 2006], and the proof of finiteness in generic cases of the five-body problem [Hampton and Jensen 2011; Albouy and Kaloshin 2012].

Similarly useful is the study of relative equilibrium configurations of collections of *Helmholtz vortices* [Hampton and Moeckel 2009; Saari 2011]. Helmholtz vortices, thought of as whirlpools lying in an infinite plane composed of a perfect fluid, were first introduced as a means of modeling the interactions of two-dimensional slices of collections of columnar vortex filaments. The study of relative equilibria of

MSC2010: primary 76B99; secondary 70F10, 13P10.

Keywords: relative equilibria, vortices, central configurations.

The PURE Math 2012 program was made possible through National Science Foundation grants DMS-1045082 and DMS-1045147 and a grant from the National Security Agency.

vortices has applications that range from basic fluid mechanics to the study of how cyclones and hurricanes interact and evolve over time.

Algebraically, the equations defining relative equilibria of vortices are very similar to those defining relative equilibria of masses. Suppose vortices of strengths Γ_i (unlike the masses in the Newtonian problem, these can have positive or negative real values) are initially located at positions $q_i \in \mathbb{R}^2$. Writing $r_{ij} = \|q_i - q_j\|$ for the mutual distance, we have a relative equilibrium if for all i ,

$$\sum_{j \neq i} \Gamma_j \frac{q_i - q_j}{r_{ij}^2} = -\lambda(q_i - c), \quad (1-1)$$

where λ is a constant and c is the center of rotation. The equations (1-1) differ from their Newtonian equivalents because of the r_{ij}^2 in the denominators (where r_{ij}^3 appears in the equations for relative equilibria of masses). The difference is caused by a logarithmic potential in the vortex case that replaces the gravitational potential in the Newtonian case.

In this paper, we study relative equilibria of collections of four point vortices whose locations lie on a circle in the plane (the cocircular configurations in the title). The inspiration for this study can be found in a recent paper in which Cors and Roberts [2012] study the corresponding problem for four cocircular masses under Newtonian gravity. Other articles devoted to the study of cocircular central configurations include [Hampton 2005; Llibre and Valls 2015]. We also use a number of general results on the vortex problem from a second recent article by Hampton, Roberts, and Santoprete [Hampton et al. 2014]. We first present a set of equations in mutual distance coordinates whose solutions correspond to these configurations in Section 2. By analyzing the set of solutions of these equations, in Section 3 we obtain a surface in \mathbb{R}^3 whose points parametrize the family of cocircular relative equilibria. Next, in Section 4, we prove a result concerning the possible signs of the vorticities for a cocircular relative equilibrium. We discuss some constraints on the positions q_i and the vorticities Γ_i in relative equilibria in Section 5. Finally, we follow [Cors and Roberts 2012], *mutatis mutandis*, and analyze two symmetric cases (kites and isosceles trapezoids) in Sections 6 and 7. These cases correspond to boundary points of our surface.

2. Equations for relative equilibria in mutual distance coordinates

By using results from [Hampton et al. 2014] on the general four-vortex problem and adapting results from [Cors and Roberts 2012] on the cocircular case of the four-body problem, in this section we will derive a set of equations characterizing the cocircular relative equilibria in the four-vortex problem.

By equation (10) of [Hampton et al. 2014], the following relation (a consequence of the Dziobek relations in the vortex case) is necessary and sufficient for the existence of a four-vortex relative equilibrium with mutual distances $r_{ij} > 0$,

where $1 \leq i < j \leq 4$:

$$(r_{13}^2 - r_{12}^2)(r_{23}^2 - r_{34}^2)(r_{24}^2 - r_{14}^2) - (r_{12}^2 - r_{14}^2)(r_{24}^2 - r_{34}^2)(r_{13}^2 - r_{23}^2) = 0. \quad (2-1)$$

For future reference, we note that this equation can be rearranged algebraically in many different ways. We will also need the forms

$$(r_{14}^2 - r_{24}^2)(r_{13}^2 - r_{34}^2)(r_{12}^2 - r_{23}^2) - (r_{14}^2 - r_{34}^2)(r_{13}^2 - r_{23}^2)(r_{12}^2 - r_{24}^2) = 0, \quad (2-2)$$

$$(r_{23}^2 - r_{24}^2)(r_{14}^2 - r_{34}^2)(r_{12}^2 - r_{13}^2) - (r_{24}^2 - r_{34}^2)(r_{13}^2 - r_{14}^2)(r_{12}^2 - r_{23}^2) = 0, \quad (2-3)$$

$$(r_{24}^2 - r_{23}^2)(r_{13}^2 - r_{34}^2)(r_{12}^2 - r_{14}^2) - (r_{34}^2 - r_{23}^2)(r_{13}^2 - r_{14}^2)(r_{12}^2 - r_{24}^2) = 0. \quad (2-4)$$

Now we impose the condition that the locations of the four vortices lie on a single circle in the plane. Numbering the positions sequentially around that circle, it follows that $r_{12}, r_{23}, r_{34}, r_{14}$ are the lengths of the exterior edges of a cyclic quadrilateral, and r_{13}, r_{24} are the lengths of the diagonals. Letting

$$a = r_{12}r_{34} + r_{14}r_{23}, \quad b = r_{12}r_{14} + r_{23}r_{34}, \quad c = r_{12}r_{23} + r_{14}r_{34}, \quad (2-5)$$

from the law of cosines and the fact that opposite interior angles in the quadrilateral are supplementary, it follows that

$$r_{13}^2 = \frac{ab}{c}, \quad (2-6)$$

$$r_{24}^2 = \frac{ac}{b}. \quad (2-7)$$

Multiplying the two equations above and taking square roots gives Ptolemy's theorem on cyclic quadrilaterals

$$r_{13}r_{24} = r_{12}r_{34} + r_{14}r_{23}. \quad (2-8)$$

As in [Cors and Roberts 2012], we will always fix the numbering of the vortices so that r_{12} is the largest exterior side length, and we will normalize the unit of distance so $r_{12} = 1$. Then

$$r_{23}, r_{34}, r_{14} \leq 1. \quad (2-9)$$

As noted in [Cors and Roberts 2012], we also have

$$\frac{r_{13}}{r_{24}} = \frac{b}{c},$$

so

$$r_{13} - r_{24} \geq 0 \iff b - c \geq 0 \iff (r_{14} - r_{23})(r_{12} - r_{34}) \geq 0.$$

Since $r_{12} \geq r_{34}$ by our choice of labeling,

$$r_{14} \geq r_{23} \iff r_{13} \geq r_{24}. \quad (2-10)$$

We note some additional useful consequences of the equations above relating the diagonals of the cyclic quadrilateral to the exterior sides. In words, these inequalities

will say that *the diagonals of the cyclic quadrilateral are longer than any exterior side on the opposite side of the diagonal from the longest exterior side*. For instance, from (2-6), notice that

$$r_{13}^2 - r_{14}^2 = r_{34} \left(\frac{r_{34}r_{23} + r_{23}^2r_{14} + r_{14} - r_{14}^3}{r_{23} + r_{14}r_{34}} \right) > 0 \tag{2-11}$$

(since $r_{14} - r_{14}^3 \geq 0$ by (2-9)). By similar computations, we also have

$$r_{13}^2 - r_{34}^2 > 0, \tag{2-12}$$

$$r_{24}^2 - r_{23}^2 > 0, \tag{2-13}$$

$$r_{24}^2 - r_{34}^2 > 0. \tag{2-14}$$

Let $\Gamma_i \in \mathbb{R} \setminus \{0\}$, where $i = 1, \dots, 4$, denote the strengths (vorticities) of the four vortices. The derivation of (2-1) above and a computation analogous to that giving equations (16)–(18) in [Cors and Roberts 2012] leads to the vorticity ratio formulas

$$\frac{\Gamma_2}{\Gamma_1} = \frac{r_{23}r_{24}(r_{13}^2 - r_{14}^2)}{r_{13}r_{14}(r_{24}^2 - r_{23}^2)}, \tag{2-15}$$

$$\frac{\Gamma_3}{\Gamma_1} = \frac{r_{23}r_{34}(1 - r_{14}^2)}{r_{14}(r_{23}^2 - r_{34}^2)}, \tag{2-16}$$

$$\frac{\Gamma_4}{\Gamma_1} = \frac{r_{24}r_{34}(r_{13}^2 - 1)}{r_{13}(r_{24}^2 - r_{34}^2)}. \tag{2-17}$$

We can always normalize (choose units for vorticity) to set $\Gamma_1 = 1$. By (2-3) and (2-11)–(2-14), the numerator in the formula for Γ_2 and the denominators in the formulas for Γ_2 and Γ_4 are always nonzero, so the values of Γ_2 and Γ_4 are always determined by these. Equation (2-16) gives a well-defined value for Γ_3 unless $r_{23}^2 - r_{34}^2 = 0$. Looking at (2-4), (2-12), and (2-13), we see that this implies $1 - r_{14}^2 = 0$, so the quotient is actually indeterminate. If, on the other hand, the factor $1 - r_{14}^2$ vanishes, then (2-4) and (2-11) show that $r_{23}^2 - r_{34}^2 = 0$, or $1 - r_{24}^2 = 0$. When $r_{23}^2 - r_{34}^2 = 0$, an alternate formula for Γ_3 can be derived using (2-1):

$$\Gamma_3 = \frac{(r_{13}^2 - 1)(r_{24}^2 - 1)r_{23}^2}{(r_{24}^2 - r_{23}^2)(r_{13}^2 - r_{23}^2)}. \tag{2-18}$$

There are solutions with $r_{14} = r_{12} = r_{24} = 1$ corresponding to degenerate configurations with vortices 1, 2 and 4 forming an equilateral triangle and $\Gamma_3 = 0$. Similarly, there are degenerate configurations with $r_{13} = r_{12} = r_{23} = 1$ and $\Gamma_4 = 0$. The configurations with $r_{14} = r_{12} = 1$ and $r_{23} = r_{34}$ are the symmetric kites to be studied in Section 6.

Collecting all of the results stated above, we see the following statement.

Theorem 2.1. *A cocircular configuration of four vortices with mutual distances r_{ij} , vorticities Γ_i , and with $r_{12}=1$, $r_{14} < 1$ and $\Gamma_1=1$ is a relative equilibrium if and only if the r_{ij} and Γ_i give a common zero of the following set of six polynomial equations:*

$$\begin{aligned}
 F_1 &= r_{13}^2(r_{23} + r_{34}r_{14}) - (r_{34} + r_{14}r_{23})(r_{14} + r_{23}r_{34}), \\
 F_2 &= r_{24}^2(r_{14} + r_{23}r_{34}) - (r_{34} + r_{14}r_{23})(r_{23} + r_{14}r_{34}), \\
 F_3 &= (r_{13}^2 - 1)(r_{23}^2 - r_{34}^2)(r_{24}^2 - r_{14}^2) - (1 - r_{14}^2)(r_{24}^2 - r_{34}^2)(r_{13}^2 - r_{23}^2), \\
 F_4 &= r_{13}r_{14}(r_{24}^2 - r_{23}^2)\Gamma_2 - r_{23}r_{24}(r_{13}^2 - r_{14}^2), \\
 F_5 &= r_{14}(r_{23}^2 - r_{34}^2)\Gamma_3 - r_{23}r_{34}(1 - r_{14}^2), \\
 F_6 &= r_{13}(r_{24}^2 - r_{34}^2)\Gamma_4 - r_{23}r_{24}(r_{13}^2 - 1).
 \end{aligned}
 \tag{2-19}$$

When $r_{12} = r_{14} = 1$, the equation $F_5 = 0$ is replaced by a similar equation $F'_5 = 0$ derived from (2-18).

3. The surface of cocircular relative equilibria

As suggested by the naive count of variables and equations in the system (2-19), with our normalizations, the set of cocircular relative equilibria is two-dimensional. The equations $F_4 = F_5 = F_6 = 0$ in Theorem 2.1 express the vorticities $\Gamma_2, \Gamma_3, \Gamma_4$ in terms of the r_{ij} . Moreover, we may use the equations $F_1 = 0$ and $F_2 = 0$ to write the squared diagonals r_{13}^2 and r_{24}^2 as functions of the other mutual distances as in (2-6) and (2-7) above. Using these two relations, one can think of F_3 as a function of the three exterior side lengths r_{23}, r_{34}, r_{14} :

$$F_3(r_{23}, r_{34}, r_{14}) = (r_{13}^2 - 1)(r_{23}^2 - r_{34}^2)(r_{24}^2 - r_{14}^2) - (1 - r_{14}^2)(r_{24}^2 - r_{34}^2)(r_{13}^2 - r_{23}^2).$$

Then $F_3 = 0$ defines an algebraic surface in \mathbb{R}^3 with coordinates r_{23}, r_{34}, r_{14} .

By (2-9), we can plot the set of points on which $F_3 = 0$ implicitly in the unit cube. Figure 1 shows the view of the surface looking along the positive r_{14} -axis toward the (r_{23}, r_{34}) -plane. There is a nearly vertical portion of the surface that is obscured from this viewpoint, but visible in the rotated view on the right in Figure 1. However, the entire implicit plot is symmetric across the plane $r_{14} = r_{23}$ (this can be seen by the fact that interchanging r_{14} and r_{23} takes (2-1) to (2-2)).

Therefore we can assume without loss of generality that $r_{14} \geq r_{23}$, and so also $r_{13} \geq r_{24}$ by (2-10). We will only consider that portion of the graph in the following. Because of the shape, we will refer to it as the *bowtie surface*.

We next consider what configurations correspond to points on the boundary curves. Note that if $r_{14} = r_{23}$, then (2-5), (2-6), and (2-7) imply that $r_{13} = r_{24}$ as well, so the only cases where $r_{14} = r_{23}$ are the configurations known as isosceles trapezoids. These will be studied in more detail in Section 7. We next note that since $1 = r_{12} \geq r_{14} \geq r_{23}$, the rest of the boundary is defined by $r_{14} = 1$. Substituting

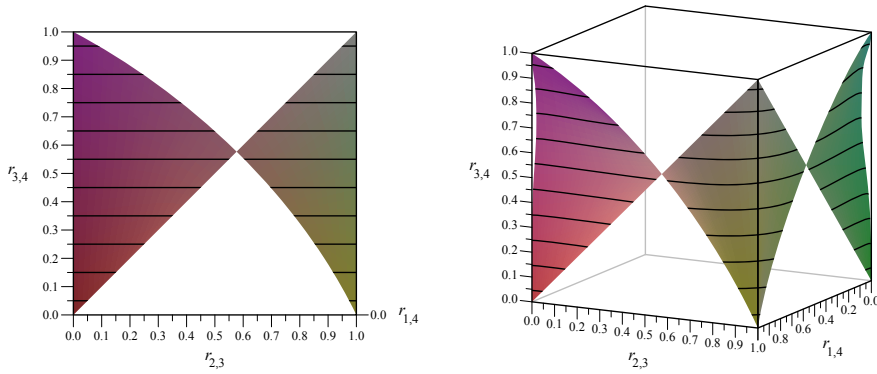


Figure 1. Two views of the surface $F_3(r_{23}, r_{34}, r_{14}) = 0$.

this into F_3 and factoring yields

$$F_3(r_{23}, r_{34}, 1) = (r_{23} - r_{34})(r_{23} + r_{34})^2(r_{23}^2 + r_{23}r_{34} + r_{34}^2 - 1).$$

The first factor vanishes on points corresponding to kite configurations where $r_{23} = r_{34}$. The kite cases will be completely characterized in Section 6.

The second factor is never zero for positive mutual distances. Hence it is left to consider cases where

$$r_{23}^2 + r_{23}r_{34} + r_{34}^2 - 1 = 0.$$

Examining (2-7), we see that when $r_{12} = r_{14} = 1$, this equation is equivalent to $r_{24}^2 = 1$. Therefore, the vortices 1, 2, and 4 are at the corners of an equilateral triangle, and it follows by (2-18) that $\Gamma_3 = 0$. Thus, the points on this curved component of the boundary shown in Figure 2 correspond to degenerate configurations.

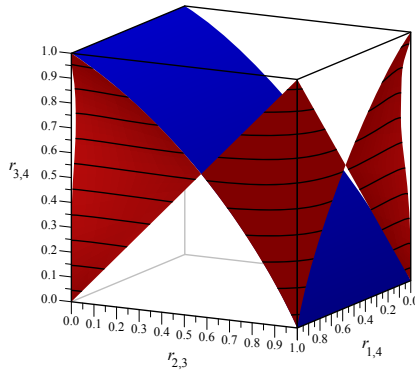


Figure 2. Plot of $r_{23}^2 + r_{23}r_{34} + r_{34}^2 - 1 = 0$ with the graph of $F_3(r_{23}, r_{34}, r_{14}) = 0$.

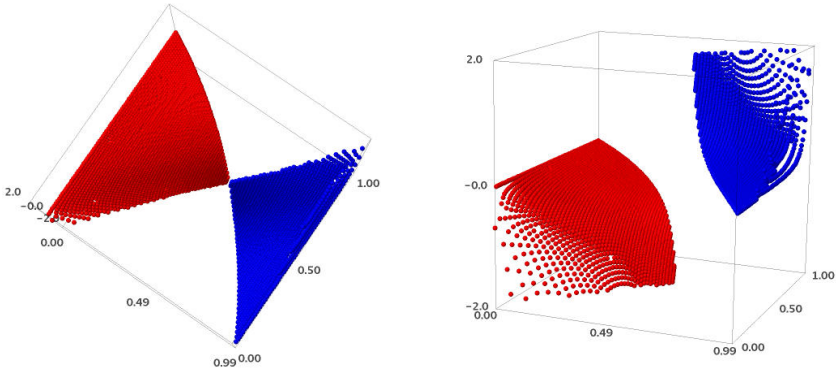


Figure 3. Views of Γ_3 along r_{14} axis (left) and from the side (right).

4. The signs of the vorticities

In this section, we will analyze the possible signs of the Γ_i in solutions of the system of equations from Theorem 2.1. We will see that, in fact, in any such relative equilibrium either all of the Γ_i have the same sign, or else three of the Γ_i have the same sign and the remaining vorticity has the opposite sign.

We were led to conjecture these patterns by plots showing the values for the vorticity Γ_3 obtained from the equation $F_5 = 0$ in (2-19) on the points of the bowtie surface defined by $F_3 = 0$. To generate the plots in Figure 3, we solved the equation $F_3 = 0$ numerically for r_{14} as a function of r_{23} and r_{34} at a collection of points in the projection of the bowtie onto the (r_{23}, r_{34}) -plane, then plotted positive Γ_3 values in blue and negative Γ_3 values in red. Figure 3 (left) shows a top view along the direction of the r_{14} -axis. Figure 3 (right) shows the same plot of Γ_3 -values, but from one side.

In the remainder of this section, we will give an analytic proof that Γ_3 takes opposite signs on the two lobes of the bowtie surface. We will need the following fact; this depends only on the geometry of the cyclic quadrilateral.

Lemma 4.1 [Cors and Roberts 2012, Lemma 4.6]. *Under the assumption $r_{14} \geq r_{23}$, and the consequence noted above in (2-10), it follows that*

$$\frac{r_{13}}{r_{24}} \leq \frac{r_{14}}{r_{23}}.$$

Proof. For the convenience of the reader, we reproduce the proof from [Cors and Roberts 2012]. From (2-6) and (2-7), and using the assumptions $r_{12} = 1$ and $r_{23} \leq r_{14}$, we have

$$\frac{r_{13}}{r_{24}} = \frac{b}{c} = \frac{r_{14} + r_{23}r_{34}}{r_{23} + r_{14}r_{34}} \leq \frac{r_{14}(1 + r_{34})}{r_{23}(1 + r_{34})},$$

which implies the claim. □

Lemma 4.2. *In all cocircular four-vortex relative equilibria as above, $\Gamma_2 > 0$.*

Proof. From the equation $F_4 = 0$, we have

$$\Gamma_2 = \frac{r_{24} r_{23} (r_{13}^2 - r_{14}^2)}{r_{13} r_{14} (r_{24}^2 - r_{23}^2)}. \tag{4-1}$$

The inequality $\Gamma_2 > 0$ follows from (2-11) and (2-13). □

The portion of the bowtie surface with $r_{14} \geq r_{23}$ off the boundary curves is composed of two lobes: one (on the left in Figure 1 (left)) on which $r_{23} < r_{34}$, and a second on which $r_{23} > r_{34}$. We will call these *open subsets* of the bowtie surface lobe I and lobe II, respectively. The closures of the two lobes of the surface intersect only at the point corresponding to a degenerate configuration that is also a kite.

We will deal with the points in the interior of lobe II first, since they follow essentially the same patterns as those found by Cors and Roberts in the cocircular four-body central configurations. We note that in [Cors and Roberts 2012, Section 2.2], the inequality $r_{23} \geq r_{34}$ was deduced from the positivity of the masses m_i . However, this inequality holds by definition on our lobe II.

Theorem 4.3. *On lobe II, we have*

$$\Gamma_2 \geq \Gamma_4 \geq \Gamma_3 > 0.$$

Hence all four of the vorticities have the same sign on lobe II.

Proof. The inequality $\Gamma_2 \geq \Gamma_4$ follows from the equations $F_4 = 0$ and $F_6 = 0$, or from (2-15) and (2-17). These say

$$\Gamma_2 = \frac{r_{23}r_{24}(r_{13}^2 - r_{14}^2)}{r_{13}r_{14}(r_{24}^2 - r_{23}^2)}, \quad \Gamma_4 = \frac{r_{34}r_{24}(r_{13}^2 - 1)}{r_{13}(r_{24}^2 - r_{34}^2)},$$

and the inequalities $r_{23} > r_{34}$, $r_{14} \leq 1$, and $r_{13} \geq r_{14}$ combine to give $\Gamma_2 \geq \Gamma_4$. Finally, $\Gamma_4 \geq \Gamma_3 > 0$ follows using Lemma 4.1 just as in the proof of Theorem 4.4 of [Cors and Roberts 2012]. □

Now we analyze the situation on lobe I:

Theorem 4.4. *On lobe I, we have*

$$\Gamma_4 > \Gamma_2 > 0 > \Gamma_3.$$

Hence three of the vorticities are positive and one is negative on lobe I.

Proof. The inequality $\Gamma_2 > 0$ follows again from Lemma 4.2. On lobe I, $r_{23} < r_{34}$ and the equation $F_5 = 0$ from (2-19) imply that $\Gamma_3 < 0$. Hence to finish the proof, we only need to show that $\Gamma_4 > \Gamma_2$ on this lobe of the bowtie.

We begin with the equations $F_4 = 0$ and $F_6 = 0$ from (2-19). Solving for Γ_2, Γ_4 and multiplying, we have

$$\Gamma_2\Gamma_4 = \frac{r_{23}r_{24}r_{34}}{r_{13}^2r_{14}} \frac{(r_{13}^2 - r_{14}^2)}{(r_{24}^2 - r_{23}^2)} \frac{(r_{13}^2 - 1)}{(r_{24}^2 - r_{34}^2)}.$$

We will show first that $\Gamma_2\Gamma_4 > 0$. From (2-3), we also have

$$\frac{r_{13}^2 - r_{14}^2}{r_{24}^2 - r_{23}^2} = \frac{(r_{14}^2 - r_{34}^2)(r_{13}^2 - 1)}{(1 - r_{23}^2)(r_{24}^2 - r_{34}^2)}. \tag{4-2}$$

Substituting into the previous equation, we have

$$\Gamma_2\Gamma_4 = \frac{r_{23}r_{24}r_{34}}{r_{13}^2r_{14}} \frac{(r_{14}^2 - r_{34}^2)}{(1 - r_{23}^2)} \left(\frac{r_{13}^2 - 1}{r_{24}^2 - r_{34}^2} \right)^2.$$

Hence the sign of $\Gamma_2\Gamma_4$ is determined by the sign of the factor $r_{14}^2 - r_{34}^2$.

By rearranging (2-2) and (2-3) (with $r_{12} = 1$), we obtain the equations

$$\frac{r_{14}^2 - r_{34}^2}{1 - r_{23}^2} = \frac{(r_{13}^2 - r_{34}^2)(r_{24}^2 - r_{14}^2)}{(r_{13}^2 - r_{23}^2)(r_{24}^2 - 1)} = \frac{(r_{13}^2 - r_{14}^2)(r_{24}^2 - r_{34}^2)}{(r_{13}^2 - 1)(r_{24}^2 - r_{23}^2)}. \tag{4-3}$$

In the rightmost expression in (4-3), all of the factors except $r_{13}^2 - 1$ are known to be positive by (2-11), (2-13), and (2-14). Similarly from (2-12) and $r_{23} < r_{34}$, the factors $r_{13}^2 - r_{34}^2$ and $r_{13}^2 - r_{23}^2$ in the middle product are also positive.

We consider the following possible cases. If $r_{24}^2 - r_{14}^2$ and $r_{24}^2 - 1$ have the same sign, then $\Gamma_2\Gamma_4 > 0$ and we are done.

On the other hand, we claim that the case where these factors have opposite signs, that is, $r_{24}^2 - 1 < 0$ but $r_{24}^2 - r_{14}^2 > 0$, is not possible for a four-vortex relative equilibrium (even though these relations are certainly possible for a cyclic quadrilateral). We note that in this remaining potential “bad” case, from (4-3), we have $r_{13}^2 - 1 < 0$, so the edge lengths are ordered as

$$r_{12} = 1 > r_{13} > r_{24} > r_{34} > r_{14} > r_{23}. \tag{4-4}$$

We will show that this is incompatible with the equation $F_3 = 0$, but in the rearranged form given in (2-4).

Denote the factors in that equation as $ABC - abc = 0$. Under the assumptions that the lengths are ordered as in (4-4), we see

$$A = r_{24}^2 - r_{23}^2 > a = r_{34}^2 - r_{23}^2 > 0.$$

We claim that it is also true that $BC > bc > 0$, so the equation $ABC - abc = 0$ cannot hold. First, $BC > 0$ and $bc > 0$ by (4-4). Expand out the products in $BC - bc$, noting one cancellation, to obtain

$$r_{13}^2r_{24}^2 + r_{14}^2 + r_{34}^2r_{14}^2 - r_{13}^2r_{14}^2 - r_{34}^2 - r_{14}^2r_{24}^2.$$

By Ptolemy's theorem from (2-8), we can substitute for the first term and simplify to obtain

$$BC - bc = r_{14}^2(r_{23}^2 + r_{34}^2 + 1 - r_{13}^2 - r_{24}^2) + 2r_{14}r_{23}r_{34}.$$

By the law of cosines as before, we have

$$r_{23}^2 + r_{34}^2 = r_{24}^2 + 2r_{23}r_{34} \cos \theta_3,$$

where θ_3 is the interior angle of the quadrilateral at vortex 3. Hence

$$BC - bc = r_{14}^2(1 - r_{13}^2) + 2r_{14}r_{23}r_{34}(1 + r_{14} \cos \theta_3) > 0.$$

This shows that this case cannot occur. Hence $\Gamma_2\Gamma_4 > 0$ and in addition, $r_{14}^2 - r_{34}^2 > 0$.

It remains to show that $\Gamma_4 > \Gamma_2$. By (2-15) and (2-17),

$$\frac{\Gamma_4}{\Gamma_2} = \frac{r_{34}r_{14}}{r_{23}} \frac{(r_{13}^2 - 1)(r_{24}^2 - r_{23}^2)}{(r_{13}^2 - r_{14}^2)(r_{24}^2 - r_{34}^2)}.$$

As noted above, from (2-3) (with $r_{12} = 1$), we obtain

$$\frac{(r_{13}^2 - 1)(r_{24}^2 - r_{23}^2)}{(r_{13}^2 - r_{14}^2)(r_{24}^2 - r_{34}^2)} = \frac{1 - r_{23}^2}{r_{14}^2 - r_{34}^2}. \quad (4-5)$$

Hence

$$\frac{\Gamma_4}{\Gamma_2} = \frac{r_{34}r_{14}(1 - r_{23}^2)}{r_{23}(r_{14}^2 - r_{34}^2)}.$$

Note that both the numerator and the denominator are positive by the argument showing $\Gamma_2\Gamma_4 > 0$. We subtract the denominator in the last expression from the numerator and factor to obtain

$$(r_{34} - r_{14}r_{23})(r_{14} + r_{23}r_{34}).$$

The first factor is positive since $r_{34} > r_{23}$ on lobe I and $r_{14} < 1$. The second factor is automatically positive since the r_{ij} are distances. Hence $\Gamma_4 > \Gamma_2$. \square

5. Further constraints on the q_i and the Γ_i

We have already seen that, as in the Newtonian case, not every cyclic quadrilateral can appear in a relative equilibrium of four vortices; there are additional geometric constraints imposed by (2-1). The following lemma is inspired by the proof of Conley's perpendicular bisector theorem for Newtonian central configurations from [Moeckel 1990] and gives another type of constraint. To our knowledge, this sort of argument has not been used before for vortices and this sort of approach could be useful in other situations. However, the fact that the Γ_i can be positive or negative makes it somewhat difficult to foresee the circumstances where something of this sort might be used (other than for cases where it is assumed that all the Γ_i are

positive, for instance). We continue to assume that the positions of the vortices are labeled in sequential order around the circumscribed circle, $r_{12} = 1$ is the longest exterior side of the quadrilateral, $r_{23} \leq r_{14}$, and $\Gamma = 1$.

Lemma 5.1. *Let L be the perpendicular bisector of the chord of the circle connecting q_2 and q_3 . Then q_1 and q_4 lie on opposite sides of L . In particular, the arc from q_1 to q_2 along the circle not containing q_3 and q_4 is less than a semicircle.*

Proof. We begin with the observation that, by Theorems 4.3 and 4.4, $\Gamma_1 = 1$ and $\Gamma_4 > 0$ have the same sign in all of our relative equilibria. From (1-1) with $i = 2, 3$, we have the equations

$$\begin{aligned} \Gamma_1 \frac{q_2 - q_1}{r_{12}^2} + \Gamma_3 \frac{q_2 - q_3}{r_{23}^2} + \Gamma_4 \frac{q_2 - q_4}{r_{24}^2} &= -\lambda(q_2 - c), \\ \Gamma_1 \frac{q_3 - q_1}{r_{13}^2} + \Gamma_2 \frac{q_3 - q_2}{r_{23}^2} + \Gamma_4 \frac{q_3 - q_4}{r_{34}^2} &= -\lambda(q_3 - c). \end{aligned}$$

Subtracting these two equations and rearranging, we see that the vector

$$\Gamma_1 \left(\frac{q_2 - q_1}{r_{12}^2} - \frac{q_3 - q_1}{r_{13}^2} \right) + \Gamma_4 \left(\frac{q_2 - q_4}{r_{24}^2} - \frac{q_3 - q_4}{r_{34}^2} \right) \tag{5-1}$$

is a scalar multiple of $q_2 - q_3$. Let v be a unit vector orthogonal to $q_2 - q_3$. The standard inner (dot) product of v and $q_2 - q_3$ is $\langle v, q_2 - q_3 \rangle = 0$. Hence $\langle v, q_2 - q_1 \rangle = \langle v, q_3 - q_1 \rangle$ and $\langle v, q_2 - q_4 \rangle = \langle v, q_3 - q_4 \rangle$. Call the first of these scalars d_1 and the second d_4 . Then taking the inner product of (5-1) and v , we obtain

$$\Gamma_1 d_1 \left(\frac{1}{r_{12}^2} - \frac{1}{r_{13}^2} \right) + \Gamma_4 d_4 \left(\frac{1}{r_{24}^2} - \frac{1}{r_{34}^2} \right) = 0. \tag{5-2}$$

We claim that this relation can only hold when q_1 and q_4 lie on opposite sides of L . Note that $1/r_{12}^2 - 1/r_{13}^2$ (respectively, $1/r_{24}^2 - 1/r_{34}^2$) is zero only if q_1 (respectively, q_4) lies on the perpendicular bisector L . Moreover the sign is positive if q_1 (respectively, q_4) lies in the half-plane bounded by L and containing q_2 and negative on the half-plane containing q_3 . On the other hand, d_1 and d_4 both have the same sign since q_1 and q_4 lie in the same half-plane bounded by the chord through q_2 and q_3 . Hence the only way the left side of (5-2) can cancel to zero is if q_1 and q_4 lie on opposite sides of L . □

Theorem 5.2. *In all of our relative equilibria, $\Gamma_2 \leq 1$.*

Proof. In a cyclic quadrilateral, it is a standard fact that the angle between an exterior side and a diagonal is equal to the angle between the opposite side and the other diagonal. It follows that the four triangles formed by the two diagonals and the exterior sides are similar in pairs. In particular, the angle at q_4 in the triangle formed by q_1, q_2, q_4 and the angle at q_3 in the triangle formed by q_1, q_2, q_3 are

equal. Denote this angle by θ . By Lemma 5.1, $\theta < \pi/2$, so $\cos \theta > 0$. By the law of cosines in these triangles,

$$\begin{aligned} r_{13}^2 + r_{23}^2 &= r_{12}^2 + 2r_{13}r_{23} \cos \theta, \\ r_{14}^2 + r_{24}^2 &= r_{12}^2 + 2r_{14}r_{24} \cos \theta. \end{aligned}$$

By Lemma 4.1, $r_{13}r_{23} \leq r_{14}r_{24}$, and hence since $\cos \theta > 0$, it follows that $r_{13}^2 + r_{23}^2 \leq r_{14}^2 + r_{24}^2$. Thus $r_{13}^2 - r_{14}^2 \leq r_{24}^2 - r_{23}^2$ and the statement to be proved follows since each of the three factors in the product giving Γ_2 in (4-1) is at most 1. \square

It follows from this result that $\Gamma_1 = 1 \geq \Gamma_2 \geq \Gamma_4 \geq \Gamma_3 > 0$ on lobe II from the previous section. On lobe I, we have $\Gamma_4 > \Gamma_2 > 0 > \Gamma_3$, but at present we do not see how to get good bounds on Γ_4 or Γ_3 .

6. The kite configurations

We call a convex quadrilateral a kite if two opposite vertices lie on an axis of symmetry of the configuration (see Figure 4). Thus a cocircular relative equilibrium forms a kite if and only if one pair of opposite vortices lie on the diameter of the circumscribed circle. There are also kites that are not cocircular, but we will not consider them. In the following, we will assume, as in Figure 4, that the axis of symmetry passes through vortices 1 and 3.

The definition of a kite implies that adjacent sides are equal for the two vortices that lie on the diameter of the circle. Thus the conditions $r_{12} = r_{14} = 1$ and $r_{23} = r_{34}$ hold. For any kite inscribed in a circle, each side of the line of symmetry forms a right triangle. This gives us the Pythagorean relation

$$r_{13}^2 = 1 + r_{34}^2. \tag{6-1}$$

To analyze this case, we will use (2-19), but with $F_5 = 0$ replaced by the equivalent form $F'_5 = 0$ from (2-18). We will make use of Gröbner bases for the ideals generated by these polynomials. See [Cox et al. 2007] for general background

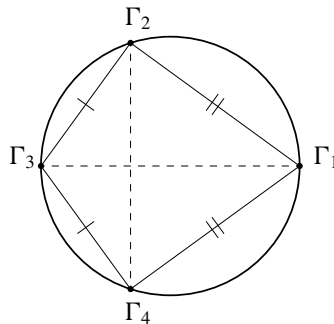


Figure 4. Kite configuration with line of symmetry through vortices 1 and 3.

on this algebraic technique. Equations for the kite configurations are obtained by substituting $r_{14} = 1$ and $r_{23} = r_{34}$. We adjoin an additional equation,

$$1 - tr_{13}r_{24}r_{34}\Gamma_2\Gamma_3\Gamma_4,$$

to force the variables appearing there to be nonzero. Using Sage [Stein et al. 2012], we compute a Gröbner basis for the substituted ideal with respect to the lexicographic order with the variables ordered as

$$t > r_{13} > r_{24} > r_{34} > \Gamma_2 > \Gamma_3 > \Gamma_4.$$

The resulting Gröbner basis contains 24 polynomials, one of which depends only on Γ_3, Γ_4 . After factoring, we see that this polynomial is

$$(4\Gamma_4^2 + \Gamma_4\Gamma_3 + \Gamma_4 - 2\Gamma_3)(-4\Gamma_4^2 + \Gamma_4\Gamma_3 + \Gamma_4 + 2\Gamma_3). \tag{6-2}$$

The next polynomial in the Gröbner basis is

$$\Gamma_2 - \Gamma_4,$$

which shows that $\Gamma_2 = \Gamma_4$ for all kite configurations, as we expect from the symmetry.

The real vanishing locus of each of the two factors in (6-2) is a hyperbola in the (Γ_3, Γ_4) -plane and each of these equations can be solved for Γ_3 in terms of Γ_4 :

$$\Gamma_3 = \frac{\mp 4\Gamma_4^2 - \Gamma_4}{\Gamma_4 \mp 2} \tag{6-3}$$

(the $-$ sign gives the solution of the equation from the left-hand factor in (6-2) and the $+$ gives the solution of the equation from the right-hand factor).

Adjoining each factor in (6-2) to the ideal individually and computing Gröbner bases again, all of the other variables can be expressed in terms of Γ_4 . From the system using the left-hand factor in (6-2), for instance, we obtain

$$r_{34}^2 = \frac{3\Gamma_4}{\Gamma_4 - 2}, \quad r_{24}^2 = \frac{6\Gamma_4}{2\Gamma_4 - 1}, \quad r_{13}^2 = \frac{4\Gamma_4 - 2}{\Gamma_4 - 2}.$$

All of the right sides must be positive since r_{ij} must be nonzero and real. In addition, $r_{34} \leq 1$ forces $-1 \leq \Gamma_4 \leq 0$. However, since $\Gamma_4 > 0$ on the interiors of lobes I and II of the bowtie surface from Theorems 4.4 and 4.3, we see that the left-hand factor from (6-2) is satisfied only for points on the surface $F_3 = 0$ with $r_{23} > r_{14}$.

With the right-hand factor in (6-2), we obtain

$$r_{34}^2 = \frac{3\Gamma_4}{\Gamma_4 + 2}, \quad r_{24}^2 = \frac{6\Gamma_4}{2\Gamma_4 + 1}, \quad r_{13}^2 = \frac{4\Gamma_4 + 2}{\Gamma_4 + 2}. \tag{6-4}$$

(The last equation also follows from (6-1).) Now the equation for r_{34}^2 shows that to get $0 < r_{34} \leq 1$, we must have $0 < \Gamma_4 \leq 1$. Using the $+$ signs in (6-3), it follows

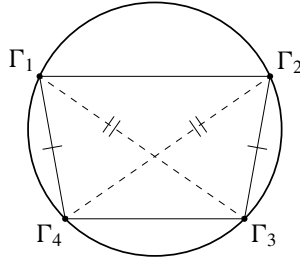


Figure 5. An isosceles trapezoid.

that $\Gamma_3 < 0$ for $0 < \Gamma_4 < \frac{1}{4}$ and $\Gamma_3 > 0$ for $\frac{1}{4} < \Gamma_4 \leq 1$. The points with $\Gamma_3 < 0$ form one of the boundary curves of lobe I of the bowtie surface considered above, and the points with $\Gamma_3 > 0$ give one boundary curve of lobe II. When $\Gamma_4 = \frac{1}{4}$, it follows that $r_{34} = 1/\sqrt{3}$, and the corresponding configuration is the symmetric degenerate configuration mentioned before: an equilateral triangle configuration with $\Gamma_1 = 1$, $\Gamma_2 = \Gamma_4 = \frac{1}{4}$, and an additional vortex with $\Gamma_3 = 0$. When $\Gamma_4 = 1$, we have a geometric square configuration with all exterior sides equal to 1, diagonals equal to $\sqrt{2}$, and all vorticities $\Gamma_i = 1$.

We have proved the following statements.

Theorem 6.1. *There is exactly one kite configuration corresponding to each point on the intersection of the bowtie surface $F_3 = 0$ and the plane given by $r_{23} = r_{34}$. These configurations are parametrized by the value of the vorticity Γ_4 with $0 < \Gamma_4 \leq 1$ as in (6-4). The other vorticities are $\Gamma_2 = \Gamma_4$ and*

$$\Gamma_3 = \frac{4\Gamma_4^2 - \Gamma_4}{\Gamma_4 + 2}.$$

The values $0 < \Gamma_4 \leq \frac{1}{4}$ give the portion of the boundary curve in the closure of lobe I and the values $\frac{1}{4} \leq \Gamma_4 \leq 1$ give the portion of the boundary curve in the closure of lobe II.

7. The isosceles trapezoid configurations

We will call a convex quadrilateral possessing a line of symmetry passing through the midpoints of two opposite edges an isosceles trapezoid. Any such quadrilateral has a circumscribed circle. If we label the vertices as in Figure 5, then the equal pairs of distances are $r_{13} = r_{24}$ and $r_{14} = r_{23}$. The corresponding four-vortex relative equilibria have been described already in Section 7 of [Hampton et al. 2014]. Hence we will only briefly discuss how the results of Hampton, Roberts and Santoprete can be recovered with our setup.

To analyze this case, we will use (2-19). Equations for the isosceles trapezoid configurations are obtained by substituting $r_{23} = r_{14}$ and $r_{24} = r_{13}$. We adjoin an

additional equation,

$$1 - tr_{14}r_{13}r_{34}\Gamma_2\Gamma_3\Gamma_4,$$

to force the variables appearing there to be nonzero. Using Sage [Stein et al. 2012], we compute a Gröbner basis for the substituted ideal with respect to the lexicographic order with the variables ordered as

$$t > r_{14} > r_{13} > r_{34} > \Gamma_3 > \Gamma_4 > \Gamma_2.$$

The resulting Gröbner basis contains 35 polynomials. In factored form, the equations from the polynomials with the three smallest lex leading terms are

$$\begin{aligned} (\Gamma_2 - 1)(r_{34} + 1) &= 0, \\ (\Gamma_4 - 1)(\Gamma_3 - \Gamma_4)(r_{34} + 1) &= 0, \\ (r_{34} - 1)(r_{34} + 1)(\Gamma_3 - \Gamma_4) &= 0. \end{aligned}$$

The first implies that $\Gamma_2 = 1$, since $r_{34} > 0$. Similarly, the second implies either $\Gamma_4 = 1$ or $\Gamma_4 = \Gamma_3$ and the third implies $r_{34} = 1$ or $\Gamma_4 = \Gamma_3$. If $r_{34} = 1$, then the configuration must be a geometric square and $\Gamma_i = 1$ for $i = 1, \dots, 4$. Hence we see the symmetry of the vorticities directly from the form of the Gröbner basis polynomials.

From the subsequent polynomials in the basis, we can solve for the remaining distances in terms of Γ_3 with the triangular form system

$$r_{34}^2 = \frac{2\Gamma_3 + \Gamma_3^2}{2\Gamma_3 + 1}, \quad r_{13}^2 = \frac{\Gamma_3 r_{34}^2 - r_{34}}{\Gamma_3 - r_{34}}, \quad r_{14}^2 = \frac{\Gamma_3 r_{34}^2 + 2r_{13}^2 - \Gamma_3 - 2}{2r_{13}^2 - 2}. \quad (7-1)$$

From the first equation here, we see that $0 < r_{34} \leq 1$ only when $-2 < \Gamma_3 \leq -1$ or $0 < \Gamma_3 \leq 1$. The last equation then shows $r_{14}^2 > 0$ only when $0 < \Gamma_3 \leq 1$.

Theorem 7.1. *There is exactly one isosceles trapezoid configuration corresponding to each point on the intersection of the bowtie surface $F_3 = 0$ and the plane given by $r_{14} = r_{23}$. With the labeling in Figure 5, these configurations are parametrized by the value of the vorticity Γ_3 with $0 < \Gamma_3 \leq 1$ as in (7-1). The point with $\Gamma_3 = 1$ corresponds to the geometric square configuration.*

Acknowledgments

The work reported in this article was begun at the Pacific Undergraduate Research Experience (PURE Math) program at the University of Hawai'i at Hilo in summer 2012 by students Jonathan Gomez, Alexander Gutierrez, and Jesse Robert under the supervision of John Little and Roberto Pelayo. Thanks to everyone at UH Hilo for an enjoyable and productive program. Thanks also to the referees for a careful reading of the manuscript and for identifying a gap in our original proof of the inequality $\Gamma_2 \leq 1$.

References

- [Albouy and Kaloshin 2012] A. Albouy and V. Kaloshin, “Finiteness of central configurations of five bodies in the plane”, *Ann. of Math. (2)* **176**:1 (2012), 535–588. MR 2925390 Zbl 06074021
- [Cors and Roberts 2012] J. M. Cors and G. E. Roberts, “Four-body co-circular central configurations”, *Nonlinearity* **25**:2 (2012), 343–370. MR 2876872 Zbl 1235.70033
- [Cox et al. 2007] D. Cox, J. Little, and D. O’Shea, *Ideals, varieties, and algorithms: An introduction to computational algebraic geometry and commutative algebra*, 3rd ed., Springer, New York, 2007. MR 2007h:13036 Zbl 1118.13001
- [Hampton 2005] M. Hampton, “Co-circular central configurations in the four-body problem”, pp. 993–998 in *EQUADIFF 2003*, edited by F. Dumortier et al., World Sci. Publ., Hackensack, NJ, 2005. MR 2185162 Zbl 1100.70008
- [Hampton and Jensen 2011] M. Hampton and A. Jensen, “Finiteness of spatial central configurations in the five-body problem”, *Celestial Mech. Dynam. Astronom.* **109**:4 (2011), 321–332. MR 2012d:70019 Zbl 1270.70038
- [Hampton and Moeckel 2006] M. Hampton and R. Moeckel, “Finiteness of relative equilibria of the four-body problem”, *Invent. Math.* **163**:2 (2006), 289–312. MR 2008c:70019 Zbl 1083.70012
- [Hampton and Moeckel 2009] M. Hampton and R. Moeckel, “Finiteness of stationary configurations of the four-vortex problem”, *Trans. Amer. Math. Soc.* **361**:3 (2009), 1317–1332. MR 2009m:76023 Zbl 1161.76011
- [Hampton et al. 2014] M. Hampton, G. E. Roberts, and M. Santoprete, “Relative equilibria in the four-vortex problem with two pairs of equal vorticities”, *J. Nonlinear Sci.* **24**:1 (2014), 39–92. MR 3162500 Zbl 1302.76042
- [Llibre and Valls 2015] J. Llibre and C. Valls, “The co-circular central configurations of the 5-body problem”, *J. Dynam. Differential Equations* **27**:1 (2015), 55–67. MR 3317391 Zbl 06425778
- [Moeckel 1990] R. Moeckel, “On central configurations”, *Math. Z.* **205**:4 (1990), 499–517. MR 92b:70012 Zbl 0684.70005
- [Saari 2011] D. G. Saari, “Central configurations—a problem for the twenty-first century”, pp. 283–298 in *Expeditions in mathematics*, edited by T. Shubin et al., Mathematical Association of America, Washington, DC, 2011. MR 2012g:00008 Zbl 1214.00002
- [Stein et al. 2012] W. A. Stein et al., *Sage mathematics software*, Version 5.0, Sage Development Team, 2012, available at <http://www.sagemath.org>.

Received: 2014-11-14 Revised: 2015-03-12 Accepted: 2015-05-09

gomezjon@math.hawaii.edu	<i>Department of Mathematics, University of Hawai’i at Manoa, Honolulu, HI 96822, United States</i>
alexg@umn.edu	<i>Department of Mathematics, University of Minnesota, Minneapolis, MN 55455, United States</i>
jlittle@holycross.edu	<i>Department of Mathematics and Computer Science, College of the Holy Cross, Worcester, MA 01610, United States</i>
robertop@hawaii.edu	<i>Department of Mathematics, University of Hawai’i at Hilo, Hilo, HI 96720, United States</i>
jesse20@hawaii.edu	<i>Department of Mathematics, University of Hawai’i at Hilo, Hilo, HI 96720, United States</i>

On weak lattice point visibility

Neil R. Nicholson and Rebecca Rachan

(Communicated by John C. Wierman)

We say that a point Q in a specific rectangular array of lattice points is weakly visible from a lattice point P not in the array if no point in the array other than Q lies on the line connecting the external point P to Q . A necessary and sufficient condition for determining if a point in the array is weakly viewable by the external point, as well as the number of points that are weakly visible from the external point, is determined.

1. Introduction

Imagine a photographer attempting to capture a picture in which every member of a band in a rectangular array formation is visible, with all persons, including the photographer, standing on lattice points. The photographer must stand at a fixed position, and each band member must have a straight-line view of the photographer, unobstructed by all other band members. Laison and Schick [2007] describe this situation and prove that there are positions for the photographer to stand but these may be quite far away from the marching band. If the photographer decides to stand closer, how can we determine which band members she can see?

These are examples of the questions arising in lattice point visibility that have been investigated for decades. For example, another question that has gotten much attention considers two sets A and B of lattice points. When is every point in A visible from every point in B ? Are there relationships between the sizes of the sets when this is the case, and if so, as the set B grows does its size act predictably? Much work has been done looking at questions such as these [Adhikari and Granville 2009; Adhikari and Balasubramanian 1996; Adhikari and Chen 1999; 2002; Chen and Cheng 2003; Herzog and Stewart 1971].

Here, we fix this second set to be a single point P , playing the role of the photographer trying to see the members of the marching band, those points in set A . In [Nicholson and Sharp 2010], a lower bound was placed on the distance P must be from a rectangular array of lattice points to weakly view every point in the array.

MSC2010: 11H06.

Keywords: weak visibility, lattice point.

Our main result can be used to prove this result in an alternate manner and provides another tool to address one of the primary questions in weak visibility: is there a formula for this minimum distance only dependent upon the dimensions of the lattice points?

2. Definitions

In this paper, all points are assumed to be lattice points in the first quadrant. Let $m, n \in \mathbb{Z}^+$ with $n \leq m$. Define $\Delta_{m,n} = \{1, 2, \dots, m\} \times \{1, 2, \dots, n\}$. We say $Q \in \Delta_{m,n}$ is *weakly visible* from a point $P \notin \Delta_{m,n}$ if no other point in $\Delta_{m,n}$ lies on the segment PQ .

It was proven in [Nicholson and Sharp 2010] that the $m \times m$ square of lattice points with its lower left-hand corner on (m, n) (that is, the square of lattice points with corners (m, n) , $(m, n + m - 1)$, $(2m - 1, n)$, and $(2m - 1, n + m - 1)$), called the *adjacency square to $\Delta_{m,n}$* and denoted $\text{Adj}_{m,n}$, contains no point that weakly views every point in $\Delta_{m,n}$. As a corollary to this, there is a lower bound that can be placed on how close a point viewing every point in $\Delta_{m,n}$ can be to $\Delta_{m,n}$:

Theorem 2.1 [Nicholson and Sharp 2010]. *If a point P weakly views every point in $\Delta_{m,n}$, then P is at least $\sqrt{m^2 + 1}$ units from (m, n) .*

What follows is a complete classification of which points in $\Delta_{m,n}$ are weakly visible from a general point $P \notin \Delta_{m,n}$ as well as two corollaries that follow from that classification.

3. Determining weak visibility

We begin this section with our main result: necessary and sufficient conditions for a point $Q \in \Delta_{m,n}$ to be weakly visible from a point $P \notin \Delta_{m,n}$.

Theorem 3.1. *The point $Q = (x_0, y_0) \in \Delta_{m,n}$ is not weakly visible by the point $P = (a, b)$ if and only if all of the following conditions hold:*

- (1) $\gcd(a - x_0, b - y_0) > 1$.
- (2) $m - x_0 \geq (a - x_0) / \gcd(a - x_0, b - y_0)$.
- (3) $n - y_0 \geq (b - y_0) / \gcd(a - x_0, b - y_0)$.

Proof. Suppose Q is not weakly viewable by P . Then, there exist $t \geq 1$ points on the interior of the segment PQ , and let $R = (x_1, y_1)$ be the first of these points to the right of Q . Thus,

$$\begin{aligned} b - y_0 &= (y_1 - y_0)(t + 1), \\ a - x_0 &= (x_1 - x_0)(t + 1), \end{aligned} \tag{1}$$

implying that

$$\begin{aligned} \gcd(a - x_0, b - y_0) &\geq t + 1 \\ &> 1. \end{aligned} \tag{2}$$

Moreover, in order to have $(x_1, y_1) \in \Delta_{m,n}$, we have

$$\begin{aligned} m - x_0 &\geq x_1 - x_0 \\ &= \frac{a - x_0}{t + 1} \\ &\geq \frac{a - x_0}{\gcd(a - x_0, b - y_0)}, \end{aligned} \tag{3}$$

with the third property following similarly.

Now, assume the three properties hold and $d = \gcd(a - x_0, b - y_0) > 1$, with

$$\begin{aligned} a - x_0 &= dp, \\ b - y_0 &= dq. \end{aligned} \tag{4}$$

We claim that $(x_0 + p, y_0 + q) \in \Delta_{m,n}$ lies on the segment PQ . To see this, note that the slope of PQ is q/p , so that PQ has equation

$$y - y_0 = \frac{q}{p}(x - x_0). \tag{5}$$

The point $(x_0 + p, y_0 + q)$ satisfies this equation, and

$$\begin{aligned} x_0 + p &= x_0 + \frac{a - x_0}{d} \\ &\leq x_0 + m - x_0 \\ &= m. \end{aligned} \tag{6}$$

Similarly, $y_0 + q \leq n$, showing $(x_0 + p, y_0 + q) \in \Delta_{m,n}$ and consequently that Q is not weakly viewable by P . \square

What points then can a particularly chosen external point P weakly view? It is only natural to insist P lies strictly above the line $y = n$ and to the right of the line $x = m$. Thus, the closest such point (with distance measured to the nearest point, (m, n) , in $\Delta_{m,n}$) would be $P = (m + 1, n + 1)$. The aforementioned result from [Nicholson and Sharp 2010] guarantees P cannot weakly view every point in $\Delta_{m,n}$ (for sufficiently large values of m and n). Which points then can P weakly view? Corollary 3.2 follows immediately from Theorem 3.1.

Corollary 3.2. *The point $(x, 1) \in \Delta_{m,n}$ is weakly viewable by the point $P = (m + 1, n + 1)$ if and only if $\gcd((m + 1) - x, n) = 1$.*

This corollary states that the number of points in the first row of $\Delta_{m,n}$ that are weakly visible by P is the number of positive integers less than or equal to m that

are relatively prime to n . This is a variation of the *Euler totient function* (or *Euler phi function*, $\phi(m)$), defined on positive integers m as the number of positive integers less than or equal to m that are relatively prime to m). For $n \leq m$, call this $\phi(n, m)$, precisely the number of points in the first row of $\Delta_{m,n}$ weakly viewable by this particular point P . This allows us to count the total number of points in $\Delta_{m,n}$ that P weakly views:

Corollary 3.3. *The number of points of $\Delta_{m,n}$ weakly viewable by the point $P = (m + 1, n + 1)$ is $\sum_{i=1}^n \phi(i, m)$.*

Proof. The number of points in the j -th row of $\Delta_{m,n}$ that are weakly viewable by P is $\phi(n - j + 1, m)$. \square

We conclude by noting that the main question, amongst numerous other interesting questions, related to the results here remains open. Is there a formula dependent only upon m and n for the point closest to $\Delta_{m,n}$ that weakly views every point of $\Delta_{m,n}$? Such a formula would lend itself not only to deeper development in other lattice point visibility questions and graph theory but would potentially have applications in a multitude of fields [Ghosh and Goswami 2013].

References

- [Adhikari and Balasubramanian 1996] S. D. Adhikari and R. Balasubramanian, “On a question regarding visibility of lattice points”, *Mathematika* **43**:1 (1996), 155–158. MR 97k:11105 Zbl 0855.11009
- [Adhikari and Chen 1999] S. D. Adhikari and Y.-G. Chen, “On a question regarding visibility of lattice points, II”, *Acta Arith.* **89**:3 (1999), 279–282. MR 2000i:11152 Zbl 0936.11039
- [Adhikari and Chen 2002] S. D. Adhikari and Y.-G. Chen, “On a question regarding visibility of lattice points, III”, *Discrete Math.* **259**:1-3 (2002), 251–256. MR 2004a:11107 Zbl 1033.11049
- [Adhikari and Granville 2009] S. D. Adhikari and A. Granville, “Visibility in the plane”, *J. Number Theory* **129**:10 (2009), 2335–2345. MR 2010m:11117 Zbl 1176.11027
- [Chen and Cheng 2003] Y.-G. Chen and L.-F. Cheng, “Visibility of lattice points”, *Acta Arith.* **107**:3 (2003), 203–207. MR 2004g:11053 Zbl 1116.11048
- [Ghosh and Goswami 2013] S. K. Ghosh and P. P. Goswami, “Unsolved problems in visibility graphs of points, segments and polygons”, *ACM Comput. Surv.* **46**:2 (2013), 22:1–22:29. Zbl 1288.05056
- [Herzog and Stewart 1971] F. Herzog and B. M. Stewart, “Patterns of visible and nonvisible lattice points”, *Amer. Math. Monthly* **78** (1971), 487–496. MR 44 #1630 Zbl 0217.03501
- [Laison and Schick 2007] J. D. Laison and M. Schick, “Seeing dots: visibility of lattice points”, *Math. Mag.* **80**:4 (2007), 274–282. MR 2008j:11079 Zbl 1208.11082
- [Nicholson and Sharp 2010] N. Nicholson and R. Sharp, “Weakly viewing lattice points”, *Involve J. of Math.* **3**:1 (2010), 9–16. Zbl 1269.11064

Received: 2014-12-01 Revised: 2015-03-20 Accepted: 2015-04-05

nrnicholson@noctrl.edu

*Department of Mathematics, North Central College,
30 North Brainard Street, Naperville, IL 60540, United States*

rarachan@noctrl.edu

*Department of Mathematics, North Central College,
30 North Brainard Street, Naperville, IL 60540, United States*

Connectivity of the zero-divisor graph for finite rings

Reza Akhtar and Lucas Lee

(Communicated by Scott T. Chapman)

We study the vertex-connectivity and edge-connectivity of the zero-divisor graph Γ_R associated to a finite commutative ring R . We show that the edge-connectivity of Γ_R always coincides with the minimum degree, and that vertex-connectivity also equals the minimum degree when R is nonlocal. When R is local, we provide conditions for the equality of all three parameters to hold, give examples showing that the vertex-connectivity can be much smaller than minimum degree, and prove a general lower bound on the vertex-connectivity.

1. Introduction

Let R be a commutative ring with unit element $1 \neq 0$. The set of zero-divisors in R does not in general possess a convenient algebraic structure; hence, nonalgebraic methods are often needed to study this set. One attempt in this direction involves the so-called *zero-divisor graph* Γ_R , whose definition was first given by Beck [1988] and later adjusted slightly by Anderson and Livingston [1999]. The vertices of Γ_R are precisely the nonzero zero-divisors of R , with two vertices adjacent if and only if the product of the ring elements they represent is zero. The idea is that by studying combinatorial properties of Γ_R , one might hope to draw conclusions about the structure of the set of zero-divisors in R . Since the paper [Anderson and Livingston 1999], considerable work has been done on this topic; for details, see the recent survey articles [Anderson et al. 2011; Coykendall et al. 2012].

One of the first results proved was that for any R , the graph Γ_R is connected, and in fact has diameter at most 3 [Anderson and Livingston 1999, Theorem 2.3]. A more refined combinatorial notion than connectedness is that of *connectivity*. For a graph G , the *vertex-connectivity*, denoted $\kappa(G)$, is the size of the smallest subset of vertices whose removal renders the graph disconnected or leaves a single vertex, while the *edge-connectivity*, denoted $\lambda(G)$, is the size of the smallest subset of edges whose removal renders the graph disconnected. In general, connectivity

MSC2010: 05C25, 13A99.

Keywords: zero-divisor graph, connectivity, finite ring.

of either type is rather difficult to compute; however, when graphs have a lot of symmetry — as is the case with zero-divisor graphs — it is sometimes possible to perform calculations, or at least give meaningful bounds.

The vertex connectivity of $\Gamma(\mathbb{Z}_n)$, with $n \geq 2$, was studied by Aaron Lauve [1999], who later discovered a mistake in his proof of the key formula in Section 4. The present article started as a project to correct this mistake, but later developed into a more comprehensive study of both the vertex- and edge-connectivity of $\Gamma(R)$ for arbitrary finite rings. An obvious starting point is the set of bounds $\kappa(G) \leq \lambda(G) \leq \delta(G)$ (see Proposition 2.2), valid for any graph G ; here $\delta(G)$ is the minimum degree of a vertex in G . In this article, we show that for all finite rings R , we have $\lambda(\Gamma_R) = \delta(\Gamma_R)$, and for nonlocal R , we also have $\kappa(\Gamma_R) = \delta(\Gamma_R)$. When R is local, however, $\kappa(\Gamma_R)$ is not nearly as well-behaved. For example, if R is a principal ideal domain, we always have $\kappa(\Gamma_R) = \delta(\Gamma_R)$; however, one can construct infinite families of rings for which $\kappa(\Gamma_R)$ is of order $\delta(\Gamma_R)^{3/4}$. We give more precise conditions under which $\kappa(\Gamma_R) = \delta(\Gamma_R)$ holds, and show that for any R , the vertex-connectivity $\kappa(\Gamma_R)$ must at least be of order $\delta(\Gamma_R)^{1/3}$.

Problems related to the focus of the present article have been studied in the recent literature. The structure of minimal vertex cuts in Γ_R was studied in [Coté et al. 2011]; however, that article does not investigate the *size* of such cuts, as is the focus of the present article. Our results are of a distinctly different flavor and thus complement rather than duplicate those of [Coté et al. 2011]. The papers [Axtell et al. 2011; Redmond 2012] are more focused in scope, and study graphs whose vertex-connectivity is 1.

2. Preliminaries

Throughout this paper, all rings are finite and commutative with $1 \neq 0$, and all graphs are finite, with no loops or multiple edges.

If R is a ring, we denote by $Z(R)$ the set of zero-divisors in R .

Definition 2.1. Let R be a ring. The *zero-divisor graph* of R , denoted Γ_R , is the graph whose vertex set is the set $Z(R) - \{0\}$, and in which $\{x, y\}$ is an edge if x and y are distinct zero-divisors of R such that $xy = 0$.

By abuse of notation, we blur the distinction between elements of $Z(R) - \{0\}$ and elements of $V(\Gamma_R)$. For $x \in Z(R) - \{0\}$, we denote by $\text{ann } x$ the annihilator of x . Hence, the degree of x (viewed as a vertex of Γ_R) is $|\text{ann } x - \{0, x\}|$.

We also recall various conventions and definitions from graph theory; see [West 1996] or any reference on graph theory for further details. For a graph G , we denote by $V(G)$ its vertex set and by $E(G)$ its edge set. For a vertex v , we denote by $N_G(v)$ (or simply $N(v)$ if the context is clear) the set of neighbors of v in G . We denote by $\delta(G)$ the minimum vertex degree in G .

If $S \subseteq V(G)$, we write $G - S$ to denote the graph with vertex set $\bar{S} = V(G) - S$ and edge set $E(G) - \{\{x, y\} : \{x, y\} \cap S \neq \emptyset\}$. If $T \subseteq E(G)$ is any subset, we denote by $G - T$ the graph with vertex set $V(G)$ and edge set $E(G) - T$. A *vertex cut* is a subset $S \subseteq V(G)$ such that $G - S$ is disconnected, and a *disconnecting set of edges* of G is a subset $T \subseteq E(G)$ such that the graph $G - T$ is disconnected; an *edge cut* is a disconnecting set of edges which is minimal (with respect to inclusion). Writing $[A, B]$ for the set of edges in G with one endpoint in each of the subsets A, B of $V(G)$, it is easily shown (see [West 1996, Remark 4.1.8]) that any edge cut in G must be of the form $[S, \bar{S}]$ for some subset $S \subseteq V(G)$. The *vertex-connectivity* of G , denoted $\kappa(G)$, is the size of the smallest set $S \subseteq V(G)$ such that S is a vertex cut or $G - S$ has only one vertex. Similarly, the *edge-connectivity* of G , denoted $\lambda(G)$, is the size of the smallest edge cut in G . For convenience, we write κ_R (respectively, λ_R, δ_R) instead of $\kappa(\Gamma_R)$ (respectively, $\lambda(\Gamma_R), \delta(\Gamma_R)$). A well-known result relating these parameters is the following statement, due to Whitney.

Proposition 2.2 [West 1996, Theorem 4.1.9]. *For any graph G , we have*

$$\kappa(G) \leq \lambda(G) \leq \delta(G).$$

3. Results

Theorem 3.1. *Let R be a finite nonlocal ring. Then $\kappa_R = \lambda_R = \delta_R$.*

Proof. By the structure theorem for Artin rings, $R \cong R_1 \times \cdots \times R_k$, where $k \geq 2$ and each R_i is a finite local ring. In light of Proposition 2.2, it suffices to show $\kappa_R \geq \delta_R$. To this end, let $S \subseteq V(\Gamma_R)$ be a subset with $|S| < \delta_R$; we will show that $H = \Gamma_R - S$ is connected. For i , with $1 \leq i \leq k$, define

$$C_i = \{(0, \dots, 0, a_i, 0, \dots, 0) \in R_1 \times \cdots \times R_k : a_i \in Z(R_i) - \{0\}\}.$$

We claim that every vertex in H is adjacent to a vertex in $C_i \cap V(H)$ for some $1 \leq i \leq k$. Since vertices of C_i are clearly adjacent to vertices of C_j when $i \neq j$, it will then follow that H is connected. Toward this goal, suppose $b = (b_1, \dots, b_k) \in V(H)$, and fix i , with $1 \leq i \leq k$, such that $b_i \neq 0$. If we define $b' = (1, \dots, 1, b_i, 1, \dots, 1)$, then clearly $N_{\Gamma_R}(b') \subseteq C_i$. In particular, this implies $|C_i| \geq \delta > |S|$, so H must contain some vertex $v \in N_{\Gamma_R}(b')$. Since $N_{\Gamma_R}(b) \supseteq N_{\Gamma_R}(b')$, we see that $v \in N_{\Gamma_R}(b) \cap C_i$, as desired. \square

From this point on, R will denote a finite local ring with maximal ideal \mathfrak{m} . Since R is Artinian, it follows from Nakayama's lemma (see [Atiyah and Macdonald 1969, Proposition 8.6]) that $\mathfrak{m}^n = 0$ for some positive integer n . We will reserve the symbol r for the smallest $n > 0$ satisfying this property. If $r = 1$, then R is a field and Γ_R is the empty graph. If $r = 2$, then Γ_R is a complete graph; so clearly $\kappa_R = \lambda_R = \delta_R = |\mathfrak{m}| - 2$. For the balance of the article, we assume $r \geq 3$, so in

particular, $\mathfrak{m}^2 \neq 0$. Since $\mathfrak{m}^{r-1} \subseteq \text{ann } \mathfrak{m}$, it follows immediately that $A_R = \text{ann } \mathfrak{m} - \{0\}$ is nonempty, and also that Γ_R is not complete. Viewed as a subset of $V(\Gamma_R)$, we have that A_R is a dominating set in Γ_R . Clearly any vertex cut in Γ_R must contain A_R ; thus, writing $\alpha_R = |A_R|$ and using Proposition 2.2, we have the elementary bounds

$$\alpha_R \leq \kappa_R \leq \lambda_R \leq \delta_R. \tag{1}$$

The following condition is important in that its presence forces all the inequalities in (1) to be equalities, but its absence typically has the opposite effect:

$$\text{There exists } x \in \mathfrak{m} \text{ such that } \text{ann } x = \text{ann } \mathfrak{m}. \tag{2}$$

Proposition 3.2. *Suppose condition (2) holds. Then $\alpha_R = \kappa_R = \lambda_R = \delta_R$.*

Proof. If $x^2 = 0$, then $x \in \text{ann } x = \text{ann } \mathfrak{m}$. Thus, $\mathfrak{m} = \text{ann } x = \text{ann } \mathfrak{m}$, and so $\mathfrak{m}^2 = 0$. Hence, we may assume $x^2 \neq 0$. In this case,

$$\delta_R \leq \deg(x) = |\text{ann } x - \{x, 0\}| = |\text{ann } x - \{0\}| = |\text{ann } \mathfrak{m} - \{0\}| = \alpha_R. \quad \square$$

If R is a principal ideal ring, condition (2) is certainly satisfied; thus, we have this:

Corollary 3.3. *Let p be a prime number and $n \geq 3$. Then*

$$\kappa(\mathbb{Z}/p^n\mathbb{Z}) = \lambda(\mathbb{Z}/p^n\mathbb{Z}) = p - 1.$$

It turns out that for local rings, the edge-connectivity is much better behaved than the vertex-connectivity. Recalling that vertices of A_R are dominant in Γ_R , the determination of λ_R is strictly graph-theoretic and follows immediately from the following easily verified fact:

Proposition 3.4. *Let G be a graph with a dominant vertex. Then $\lambda(G) = \delta(G)$.*

Proof. Choose $S \subseteq V(\Gamma_R)$ such that $T = [S, \bar{S}] \subseteq E(\Gamma_R)$ is an edge cut. We may assume without loss of generality that \bar{S} contains a dominant vertex v . Since v is adjacent to all vertices of S , we must have $|T| \geq |S|$. On the other hand, every vertex in S has at least $\delta - |S| + 1$ neighbors in \bar{S} ; so $\delta \geq |T| \geq |S|(\delta - |S| + 1)$. Rearranging the inequality $|S|(\delta - |S| + 1) \leq \delta$ gives $\delta(|S| - 1) \leq |S|(|S| - 1)$. If $|S| > 1$, then cancellation gives $\delta \leq |S|$ and so $|S| = |T| = \delta$. If $|S| = 1$, then all edges incident at the sole vertex in S must be in T , so $|T| = \delta$ in this case also. \square

Corollary 3.5. *Let R be a local ring with $\mathfrak{m}^2 \neq 0$. Then $\lambda_R = \delta_R$.*

We now turn our attention to the vertex-connectivity of Γ_R . It is natural to ask how tight the bounds $\alpha_R \leq \kappa_R \leq \delta_R$ are. In the absence of condition (2), the lower bound is usually not met.

Proposition 3.6. *Let R be a local ring with $r \geq 4$ such that condition (2) fails. Then $\kappa_R > \alpha_R$.*

Proof. First suppose $r \geq 5$. Any vertex cut must contain A_R , so it suffices to show that $H = \Gamma_R - A_R$ is connected. Because $\mathfrak{m}^{r-1} = \mathfrak{m}^{r-2}\mathfrak{m} \neq 0$, there exists some $x \in \mathfrak{m}^{r-2}$ such that $x \notin A_R$. Moreover, x is a finite sum of products of the form uv , where $u \in \mathfrak{m}^{r-3}$ and $v \in \mathfrak{m}$. Since $x \neq 0$ and $A_R \cup \{0\}$ is an ideal (hence closed under addition), at least one of these products must not be in A_R . Thus, we may assume without loss of generality that $x = uv$, where $u \in \mathfrak{m}^{r-3}$ and $v \in \mathfrak{m}$. Clearly u and v are also vertices of H , and because $r \geq 5$, we have $ux \in \mathfrak{m}^{2r-5} \subseteq \mathfrak{m}^r = 0$, so u is adjacent to x in H .

We claim that there is a path in H from every $y \in V(H)$ to x . If $y = u$ or $y = x$, this is clear, so assume otherwise. Since condition (2) fails, y has a neighbor z in H , so $yz = 0$. Now consider the product zu . If $zu = 0$, then y, z, u, x is a path. If $zu \neq 0$ but $zu \in A_R$, then $zx = (zu)v = 0$ and y, z, x is a path. Finally, if $zu \neq 0$ and $zu \notin A_R$, then zu is a vertex of H ; moreover, $y(zu) = 0$ and $x(zu) = (xu)z = 0$, so y, zu, x is a path.

Now suppose $r = 4$. Then $\mathfrak{m}^4 = 0$ but $\mathfrak{m}^3 \neq 0$, so there exists $x \in \mathfrak{m}^2$ such that x is a vertex of $H = \Gamma_R - A_R$. It suffices to show that there is a path from any vertex of H to x . To this end, let y be a vertex of H distinct from x . Since condition (2) fails, y has a neighbor z in H , i.e., $yz = 0$. If $zm \subseteq A_R$, then $zm^2 = 0$ and z is adjacent to x . If $zm \not\subseteq A_R$, then there exists $w \in \mathfrak{m}$ such that zw is a vertex of H . Now zw is a neighbor of y ; however, $zw \in \mathfrak{m}^2$, so it is also a neighbor of x . \square

Remark. The hypothesis $r \geq 4$ in Proposition 3.6 is necessary: when $r = 3$, there exist rings R not satisfying condition (2) for which $\kappa_R = \alpha_R$ and others for which $\kappa_R > \alpha_R$.

As an example of the former, let \mathbb{F}_2 be the field with two elements and consider

$$R = \frac{\mathbb{F}_2[x, y]}{(x^2, y^2)}.$$

By abuse of notation, we will use elements of $\mathbb{F}_2[x, y]$ to describe the cosets they represent in R . Then $\mathfrak{m} = (x, y)$ has eight elements and $\mathfrak{m}^2 = \text{ann } \mathfrak{m} = \{0, xy\}$. Thus, Γ_R has seven vertices, with xy a dominant vertex; moreover, $\Gamma_R - \{xy\}$ is a graph on six vertices with three connected components $\{x, x + xy\}$, $\{y, y + xy\}$ and $\{x + y, x + y + xy\}$, so $\kappa_R = \alpha_R = 1$. Note also that for any $t \in R$, $\text{ann } t$ contains (t) . Since (t) has at least four elements for any $t \neq 0$, there is no way for the equality $\text{ann } t = \text{ann } \mathfrak{m}$ to hold for any $t \in V(\Gamma_R)$. Hence, condition (2) necessarily fails.

As an example of the latter, consider

$$R = \frac{\mathbb{F}_2[x, y, z, w]}{(x^2, y^2, z^2, w^2, xy, yz, zw, wx)}.$$

It is easily seen that R is a local ring satisfying $t^2 = 0$ for all $t \in R$, whose maximal ideal $\mathfrak{m} = (x, y, z, w)$ satisfies $\mathfrak{m}^3 = 0$, $\mathfrak{m}^2 \neq 0$. Moreover, $\text{ann } \mathfrak{m} = (xz, yw)$, so $\alpha_R = 3$. As in the previous example, $t \in \text{ann } t$ for all $t \in R$, so it is easily seen

that condition (2) is not satisfied. Now let $H = \Gamma_R - A_R$; we will show that H is connected, and hence that $\kappa_R > 3$. Observe first that every vertex of H is of the form $c_1x + c_2y + c_3z + c_4w + c_5xz + c_6yw$, where the c_i are elements of \mathbb{F}_2 , and c_1, \dots, c_4 are not all 0. Evidently each such vertex is adjacent to $c_1x + c_2y + c_3z + c_4w$. Since x, y, z, w, x is a cycle in H , it will suffice (to show that H is connected) to construct a path from any vertex of the form $c_1x + c_2y + c_3z + c_4w$ (with not all c_i equal to 0) to one of the vertices of the abovementioned cycle. If v_1, v_2 are distinct elements of $\{x, y, z, w\}$ which are adjacent in H , then $v_1 + v_2$ is adjacent to v_1 . If v_1, v_2 are not adjacent, then choose v_3 from this set, distinct from v_1 and v_2 ; then v_3 will be adjacent to $v_1 + v_2$. If v_1, v_2, v_3 are distinct elements of $\{x, y, z, w\}$, then we may assume without loss of generality that v_2 is adjacent to both v_1 and v_3 . It follows that $v_1 + v_2 + v_3$ is adjacent to v_2 . Finally, $x + y + z + w$ is adjacent to $x + z$. Thus H is connected, and so $\kappa_R > 3 = \alpha_R$.

The next family of examples shows that both bounds $\alpha_R \leq \kappa_R \leq \delta_R$ can be quite loose.

Proposition 3.7. *Let F be a field of order $f = 2^s$ and*

$$R = \frac{F[x, y, z]}{(x^2, y^2, z^2)}.$$

Then $\alpha_R = f - 1$, $\kappa_R = f^3 - 1$, and $\delta_R = f^4 - 2$.

Proof. Observe that R is a local ring with maximal ideal $\mathfrak{m} = (x, y, z)$ such that $t^2 = 0$ for all $t \in R$. Moreover, $\mathfrak{m}^2 = (xy, xz, yz)$, $\mathfrak{m}^3 = (xyz)$, and $\mathfrak{m}^4 = 0$.

Clearly R is generated (as an F -vector space) by $\{1, x, y, z, xy, xz, yz, xyz\}$; from this description, it is easily seen that $|R| = f^8$, $|\mathfrak{m}| = f^7$, $|\mathfrak{m}^2| = f^4$, and $|\mathfrak{m}^3| = f$. Also, $\text{ann } \mathfrak{m} = \mathfrak{m}^3$, so $\alpha_R = f - 1$. Now since $t^2 = 0$ for all $t \in R$, it follows that $\text{ann } t \supseteq (t)$; because $|\text{ann } t| \cdot |(t)| = |R|$, we have $|\text{ann } t| \geq |R|^{1/2} = f^4$ for all $t \in R$. Direct computation shows that $\text{ann } x = (x)$, so x is a vertex in Γ_R of minimum degree $\delta_R = f^4 - 2$.

Let $S = (\text{ann } x \cap \mathfrak{m}^2) - \{0\}$. Also, any element in $(x) - S - \{0\}$ is associate to x and hence has the same neighborhood in Γ_R ; in fact, $(x) - S - \{0\}$ is a clique and a connected component of $\Gamma_R - S$. Thus there is no path in $\Gamma_R - S$ from x to y , and so $\kappa_R \leq |S| = f^3 - 1$.

Now suppose $T \subseteq V(\Gamma_R)$ is a set of vertices such that $|T| < f^3 - 1$. Given $t \in \mathfrak{m}$, consider the multiplication-by- t map $\mathfrak{m}^2 \rightarrow t\mathfrak{m}^2$. This is an R -module homomorphism whose kernel is $\text{ann } t \cap \mathfrak{m}^2$; hence

$$|\mathfrak{m}^3| \geq |t\mathfrak{m}^2| = \frac{|\mathfrak{m}^2|}{|\text{ann } t \cap \mathfrak{m}^2|},$$

and so $|\text{ann } t \cap \mathfrak{m}^2| \geq |\mathfrak{m}^2|/|\mathfrak{m}^3| = f^3$. Taking into account that 0 and possibly t itself are elements of $\text{ann } t$, this implies that every vertex of $H = \Gamma_R - T$ has a

neighbor (in H) lying in \mathfrak{m}^2 . To show that H is connected, let a and b be vertices of H . Then a has a neighbor $c \in \mathfrak{m}^2$ in H and b has a neighbor $d \in \mathfrak{m}^2$ in H . Now $cd \in \mathfrak{m}^4 = 0$, so c and d are adjacent in H , proving that there exists a path from a to b .

This shows that $\kappa_R = f^3 - 1$. □

In the example of Proposition 3.7, κ_R is roughly $(1/|F|)\delta_R$, so by taking F to be arbitrarily large, we see that there is no hope for a general upper bound on κ_R which is linear in δ_R ; in fact, in this family, κ_R is roughly $\delta_R^{3/4}$. It is natural, then, to ask for the maximum value of a , with $0 < a \leq 3/4$, such that κ_R can be bounded below (for all finite rings R) by a function of order δ_R^a . As a first step in this direction, we offer this:

Proposition 3.8. *Let R be a finite ring. Then $\kappa_R \geq (\frac{1}{2}\delta_R)^{1/3} - (\sqrt{3})^{-1}$.*

The proof relies crucially on the following observation:

Lemma 3.9. *Let R be a ring and S a vertex cut of Γ_R such that $V(G)$ is the disjoint union of two nonempty sets A and B with no edges between A and B . Suppose $|S| < \delta_R$. If $a \in A$ and $b \in B$, then $ab \in S$, $|\text{ann } a| \geq |B|/|S|$ and $|\text{ann } b| \geq |A|/|S|$.*

Proof. The hypothesis $|S| < \delta_R$ implies that a has some neighbor $x \in A$ and that b has some neighbor $y \in B$. Then $ab \neq 0$, but ab is a neighbor of both $x \in A$ and $y \in B$; thus, $ab \in S$. Now let $B = \{b_1, \dots, b_n\}$. Since each of the products ab_1, \dots, ab_n is an element of S , some element $s \in S$ appears at least $|B|/|S|$ times in this list; without loss of generality, we may assume that $ab_1 = \dots = ab_k = s$, where $k \geq |B|/|S|$. Thus, $0, b_2 - b_1, \dots, b_k - b_1$ are distinct elements of $\text{ann } a$ and hence $|\text{ann } a| \geq k \geq |B|/|S|$. The proof of the remaining assertion is similar. □

Proof of Proposition 3.8. If $\kappa_R = \delta_R$, there is nothing to prove, so assume $\kappa_R < \delta_R$ and let $S \subseteq V(\Gamma_R) = \mathfrak{m} - \{0\}$ be a minimal vertex cut. Partition the vertices of $H = \Gamma_R - S$ into two disjoint nonempty sets A and B such that there are no edges between A and B ; we may assume without loss of generality that B is the larger of these two sets, i.e.,

$$|A| \leq \frac{|\mathfrak{m}| - 1 - |S|}{2} \leq |B|.$$

Now if $x \in A$ and $y \in B$, Lemma 3.9 implies that H contains no vertices from $\text{ann } x \cap \text{ann } y$. Since the zero element is not a vertex of Γ_R , we have, again using Lemma 3.9, that

$$|S| \geq |\text{ann } x \cap \text{ann } y| - 1 = \frac{|\text{ann } x| |\text{ann } y|}{|\text{ann } x + \text{ann } y|} - 1 \geq \frac{|B|/|S| \cdot |A|/|S|}{|\mathfrak{m}|} - 1.$$

Thus,

$$\begin{aligned} |S|^3 &\geq \frac{|A||B|}{|\mathfrak{m}|} - |S|^2 \geq |A| \frac{|\mathfrak{m}| - 1 - |S|}{2|\mathfrak{m}|} - |S|^2 \\ &= \frac{|A|}{2} - \frac{|S|}{2} \frac{|S| + 1}{|S|} \frac{|A|}{|\mathfrak{m}|} - |S|^2 \geq \frac{|A|}{2} - \frac{|S|}{2} - |S|^2. \end{aligned}$$

However, the neighbors of $x \in A$ in Γ_R are all members of $A \cup S$. Thus, $|A| + |S| \geq \delta_R + 1$ and so, continuing the calculation from above, we have

$$|S|^3 + |S|^2 + \frac{|S|}{2} \geq \frac{|A|}{2} \geq \frac{\delta_R - |S| + 1}{2},$$

which, upon rearrangement, gives

$$2(|S|^3 + |S|^2 + |S| + \frac{1}{2}) \geq \delta_R.$$

Hence, $2(|S| + 1/\sqrt{3})^3 \geq \delta_R$. Rearranging the inequality gives the desired result. \square

Acknowledgements

The authors thank Miami University for supporting this research during the summer of 2003. We would also like to express our gratitude to Aaron Lauve for introducing this problem to us and for providing us with his work on the topic.

References

- [Anderson and Livingston 1999] D. F. Anderson and P. S. Livingston, “The zero-divisor graph of a commutative ring”, *J. Algebra* **217**:2 (1999), 434–447. MR 2000e:13007 Zbl 0941.05062
- [Anderson et al. 2011] D. F. Anderson, M. C. Axtell, and J. A. Stickles, Jr., “Zero-divisor graphs in commutative rings”, pp. 23–45 in *Commutative algebra—Noetherian and non-Noetherian perspectives*, edited by M. Fontana et al., Springer, New York, 2011. MR 2012b:13023 Zbl 1225.13002
- [Atiyah and Macdonald 1969] M. F. Atiyah and I. G. Macdonald, *Introduction to commutative algebra*, Addison-Wesley, Reading, MA, 1969. MR 39 #4129 Zbl 0175.03601
- [Axtell et al. 2011] M. Axtell, N. Baeth, and J. Stickles, “Cut vertices in zero-divisor graphs of finite commutative rings”, *Comm. Algebra* **39**:6 (2011), 2179–2188. MR 2012i:13043 Zbl 1226.13007
- [Beck 1988] I. Beck, “Coloring of commutative rings”, *J. Algebra* **116**:1 (1988), 208–226. MR 89i:13006 Zbl 0654.13001
- [Coté et al. 2011] B. Coté, C. Ewing, M. Huhn, C. M. Plaut, and D. Weber, “Cut-sets in zero-divisor graphs of finite commutative rings”, *Comm. Algebra* **39**:8 (2011), 2849–2861. MR 2012i:13014 Zbl 1228.13011
- [Coykendall et al. 2012] J. Coykendall, S. Sather-Wagstaff, L. Sheppardson, and S. Spiroff, “On zero-divisor graphs”, pp. 241–299 in *Progress in commutative algebra 2: Closures, finiteness and factorization*, edited by C. Francisco et al., Walter de Gruyter, Berlin, 2012. MR 2932598 Zbl 1243.13017
- [Lauve 1999] A. Lauve, *Zero-divisor graphs of finite commutative rings*, Senior thesis, University of Oklahoma, 1999.
- [Redmond 2012] S. P. Redmond, “Cut vertices and degree-one vertices of zero-divisor graphs”, *Comm. Algebra* **40**:8 (2012), 2749–2756. MR 2968909 Zbl 1263.13004
- [West 1996] D. B. West, *Introduction to graph theory*, Prentice Hall, Upper Saddle River, NJ, 1996. MR 96i:05001 Zbl 0845.05001

Received: 2015-01-30 Revised: 2015-02-10 Accepted: 2015-03-04

akhtarr@miamioh.edu

*Department of Mathematics, Miami University,
Oxford, OH 45056, United States*

lucas@boldlee.com

*Department of Mathematics, Miami University,
Oxford, OH 45056, United States*

Enumeration of m -endomorphisms

Louis Rubin and Brian Rushton

(Communicated by Vadim Ponomarenko)

An m -endomorphism on a free semigroup is an endomorphism that sends every generator to a word of length $\leq m$. Two m -endomorphisms are combinatorially equivalent if they are conjugate under an automorphism of the semigroup. In this paper, we specialize an argument of N. G. de Bruijn to produce a formula for the number of combinatorial equivalence classes of m -endomorphisms on a rank- n semigroup. From this formula, we derive several little-known integer sequences.

1. Introduction

Let D be a nonempty set of symbols, and let D^+ be the set of all finite strings of one or more elements of D . That is, $D^+ = \{d_1 \cdots d_k : k \in \mathbb{N}, d_i \in D\}$. Paired with the operation of string concatenation, D^+ forms the *free semigroup* on D . If $d_1, \dots, d_k \in D$, then we refer to the natural number k as the *length* of the string $d_1 \cdots d_k$. Denote the length of $W \in D^+$ by $|W|$.

By a *semigroup endomorphism* (or, simply, an *endomorphism*) on D^+ , we mean a mapping $\phi : D^+ \rightarrow D^+$ satisfying $\phi(W_1 W_2) = \phi(W_1)\phi(W_2)$ for all $W_1, W_2 \in D^+$. Note that if ϕ is an endomorphism on D^+ and $d_1, \dots, d_k \in D$, then $\phi(d_1 \cdots d_k) = \phi(d_1) \cdots \phi(d_k)$; this shows that an endomorphism on D^+ is determined by its action on the elements of D . On the other hand, any mapping $f : D \rightarrow D^+$ extends uniquely to the endomorphism $\phi_f : D^+ \rightarrow D^+$ defined by $\phi_f(d_1 \cdots d_k) = f(d_1) \cdots f(d_k)$, and it is straightforward to verify that ϕ_f is an automorphism (that is, a bijective endomorphism) precisely when f is a bijection on D .

Example 1. Let $D = \{a, b\}$, and let $f : D \rightarrow D^+$ be defined by $f(a) = ab$ and $f(b) = a$. Then, for example,

$$\phi_f(ababa) = f(a)f(b)f(a)f(b)f(a) = abaabaab.$$

Let $\text{End}(D^+)$ be the collection of all endomorphisms on D^+ , and let $m \in \mathbb{N}$. Then $\phi \in \text{End}(D^+)$ is called an *m -endomorphism* if and only if $|\phi(d)| \leq m$ for

MSC2010: primary 05A99; secondary 20M15.

Keywords: enumeration, free semigroup endomorphisms, semigroup.

all $d \in D$. Note that the mapping ϕ_f from Example 1 is an m -endomorphism for all $m \geq 2$. Now let Γ be the set of all m -endomorphisms on D^+ . That is,

$$\Gamma = \{\phi \in \text{End}(D^+) : \phi(D) \subseteq R\},$$

where $R = \{W \in D^+ : |W| \leq m\}$. Consider the set Ω consisting of all mappings $f : D \rightarrow R$. Then we may write

$$\Gamma = \{\phi_f : f \in \Omega\}.$$

We can put the set Γ into one-to-one correspondence with Ω by sending each m -endomorphism to its restriction to D . Moreover, if $|D| = n \in \mathbb{N}$, then the size of these sets is easily evaluated in view of the fact that $|R| = \sum_{i=1}^m n^i$. In particular, if $n > 1$, then $|R| = (n^{m+1} - n)/(n - 1)$, and

$$|\Gamma| = |\Omega| = \left(\frac{n^{m+1} - n}{n - 1}\right)^n.$$

However, in this paper we are interested in counting the number of *classes* of m -endomorphisms under a particular equivalence relation. To motivate our definition of equivalence on Γ , we define a relation \sim on Ω as follows:

$$f_1 \sim f_2 \iff \text{there exists a bijection } g : D \rightarrow D \text{ such that } f_2 \circ g = \phi_g \circ f_1.$$

As an exercise, the reader may wish to verify that \sim satisfies the reflexive, symmetric, and transitive properties required of any equivalence relation. In Section 1.1, however, it will be shown that \sim is a specific instance of a well-known equivalence relation induced by a group acting on a nonempty set.

Example 2. Let f be as in Example 1 (with $D = \{a, b\}$). Consider the bijection $g : D \rightarrow D$ defined by $g(a) = b$ and $g(b) = a$. Now let $f_1 : D \rightarrow D^+$ be given by $f_1(a) = b$ and $f_1(b) = ba$. Then

$$(f_1 \circ g)(a) = f_1(g(a)) = f_1(b) = ba = g(a)g(b) = \phi_g(ab) = \phi_g(f(a)) = (\phi_g \circ f)(a),$$

$$(f_1 \circ g)(b) = f_1(g(b)) = f_1(a) = b = g(a) = \phi_g(a) = \phi_g(f(b)) = (\phi_g \circ f)(b),$$

which shows that $f \sim f_1$.

Remark 3. Perhaps a more intuitive illustration of \sim is as follows. If we let f and f_1 be as in Example 2, then the respective graphs of f and f_1 are $\{(a, ab), (b, a)\}$ and $\{(a, b), (b, ba)\}$. But the graph of f_1 can be obtained by applying the bijection g to each element of D that appears in the graph of f . In other words,

$$\{(g(a), g(a)g(b)), (g(b), g(a))\} = \{(a, b), (b, ba)\}.$$

Since the graphs of f and f_1 are “the same” up to a permutation of a and b , we wish to consider these mappings equivalent, and \sim provides the desired equivalence relation.

Extending \sim to an equivalence relation on Γ leads to the following definition. If $f, h \in \Omega$, then ϕ_f is *combinatorially equivalent* to ϕ_h if and only if there exists a bijection $g : D \rightarrow D$ such that $\phi_h \circ \phi_g = \phi_g \circ \phi_f$. To state precisely the aim of this paper: given a set of symbols D with $|D| = n$, we wish to produce a formula for the number of equivalence classes in Γ under the relation of combinatorial equivalence. To this end, we shall specialize an argument of N. G. de Bruijn [1972] (namely, that used for his Theorem 1) to produce a formula for the number of classes in Ω under the relation \sim . But it is easy to check that for all $f, h \in \Omega$, we have $f \sim h$ if and only if ϕ_f is combinatorially equivalent to ϕ_h . Hence, there is a well-defined correspondence given by

$$[f] \leftrightarrow [\phi_f]$$

between the equivalence classes in Ω and those in Γ , and it follows that our formula will also provide the number of m -endomorphisms on D^+ up to combinatorial equivalence. Moreover, once this formula is obtained, we can fix one of the variables n, m and let the other run through the natural numbers in order to derive integer sequences, many of which appear to be little-known.

1.1. Group actions. For the reader’s convenience, we review group actions. The following material (through Proposition 4) is paraphrased from [Malik et al. 1997]. Let G be a group and S a nonempty set. A *left action* of G on S is a function

$$\cdot : G \times S \rightarrow S, \quad (g, s) \mapsto g \cdot s,$$

such that, for all $g_1, g_2 \in G$ and for all $s \in S$,

- (1) $(g_1 g_2) \cdot s = g_1 \cdot (g_2 \cdot s)$, where $g_1 g_2$ denotes the product of g_1, g_2 in G , and
- (2) $e \cdot s = s$, where e is the identity element of G .

A left action induces the well-known equivalence relation E on the set S given by

$$(a, b) \in E \iff g \cdot a = b \quad \text{for some } g \in G$$

for all $a, b \in S$. We refer to the equivalence classes under this relation as the *orbits* of G on S . The following result (known as Burnside’s lemma) gives an expression for the number of these, provided that G and S are finite.

Proposition 4 [Malik et al. 1997]. *Let S be a finite, nonempty set, and suppose there is a left action of a finite group G on S . Then the number of orbits of G on S is*

$$\frac{1}{|G|} \sum_{g \in G} |\{s \in S : g \cdot s = s\}|.$$

Thus, the number of orbits of G on S equals the average number of elements of S that are “fixed” by an element of G . We now show that the relation \sim from Section 1 is a specific instance of the relation E described above. To see this, let D

be a finite nonempty set, and let $\text{Sym}(D)$ denote the symmetric group on D (i.e., the group of all bijections on D). Then $\text{Sym}(D)$ acts on the set Ω according to the rule

$$g \cdot f = \phi_g \circ f \circ g^{-1}$$

for all $g \in \text{Sym}(D)$, $f \in \Omega$. (One can easily verify that \cdot defined in this way is indeed a left action.) Now, for any $f_1, f_2 \in \Omega$, we have

$$\begin{aligned} f_1 \sim f_2 &\iff f_2 \circ g = \phi_g \circ f_1 \text{ for some } g \in \text{Sym}(D) \\ &\iff f_2 = \phi_g \circ f_1 \circ g^{-1} \text{ for some } g \in \text{Sym}(D) \\ &\iff g \cdot f_1 = f_2 \text{ for some } g \in \text{Sym}(D) \\ &\iff (f_1, f_2) \in E. \end{aligned}$$

It follows that the equivalence classes in Ω under the relation \sim are just the orbits of $\text{Sym}(D)$ on Ω . Enumerating the elements of $\text{Sym}(D)$ by $g_1, \dots, g_{n!}$, we find the number of orbits to be

$$\frac{1}{n!} \sum_{r=1}^{n!} |\{f \in \Omega : f \circ g_r = \phi_{g_r} \circ f\}|. \tag{1}$$

For any permutation g of a finite set, and for each natural number j , let $c(g, j)$ denote the number of cycles of length¹ j occurring in the cycle decomposition of g . (This notation comes from [de Bruijn 1972].) The quantities $c(g, j)$ will play a role in the evaluation of $|\{f \in \Omega : f \circ g_r = \phi_{g_r} \circ f\}|$, which occurs in the next section. Our evaluation is a modification of de Bruijn’s counting argument [1964, § 5.12].

2. Main results

We now produce a formula for the number of equivalence classes in Ω under the relation \sim . Let D be a finite set, and suppose that $g \in \text{Sym}(D)$ is the product of disjoint cycles of lengths k_1, k_2, \dots, k_ℓ , where $k_1 \leq k_2 \leq \dots \leq k_\ell$. Then the sequence k_1, k_2, \dots, k_ℓ is called the *cycle type* of g . For example, if $D = \{a, b, c, d, e\}$, then the permutation $g = (a)(b, c)(d, e)$ has cycle type 1, 2, 2. The following lemma will be useful.

Lemma 5. *Let D be a finite set, and let $g \in \text{Sym}(D)$ have cycle type k_1, k_2, \dots, k_ℓ . For each $1 \leq i \leq \ell$, select a single $d_i \in D$ from the cycle corresponding to k_i . (Thus, k_i is the smallest natural number such that $g^{k_i}(d_i) = d_i$.) Now suppose that $f \in \Omega$. Then $f \circ g = \phi_g \circ f$ if and only if for each $1 \leq i \leq \ell$,*

- (1) $(f \circ g^j)(d_i) = (\phi_g^j \circ f)(d_i)$ for all $j \in \mathbb{N}$,
- (2) $f(d_i)$ is of the form $d'_1 \cdots d'_{k \leq m}$, where $d'_1, \dots, d'_k \in D$ each belong to a cycle in g whose length divides k_i .

¹There should be no confusion between the notions of “string length” and “cycle length”.

Proof. First assume that $f \circ g = \phi_g \circ f$. Then condition (1) follows from an inductive argument. But $f(d_i) = f(g^{k_i}(d_i)) = \phi_g^{k_i}(f(d_i))$. Write $f(d_i) = d'_1 \cdots d'_k$, where $d'_1, \dots, d'_k \in D$ and $k \leq m$. Then

$$d'_1 \cdots d'_k = \phi_g^{k_i}(d'_1 \cdots d'_k) = g^{k_i}(d'_1) \cdots g^{k_i}(d'_k).$$

In particular, for each $1 \leq t \leq k$, we have $d'_t = g^{k_i}(d'_t)$. This implies that

$$(d'_t, g(d'_t), g^2(d'_t), \dots, g^{k_i-1}(d'_t))$$

is a cycle whose length divides k_i . The conclusion follows.

Conversely, suppose that condition (1) holds. (Condition (2) is superfluous here.) Let $d \in D$. Then there exist $i, j \in \mathbb{N}$ such that $d = g^j(d_i)$. Now,

$$\begin{aligned} f(g(d)) &= f(g(g^j(d_i))) = f(g^{1+j}(d_i)) \\ &= \phi_g^{1+j}(f(d_i)) = \phi_g(\phi_g^j(f(d_i))) = \phi_g(f(g^j(d_i))) = \phi_g(f(d)). \end{aligned}$$

Therefore, $f \circ g = \phi_g \circ f$, so the proof is complete. □

Once again, suppose that $|D| = n$, and label the elements of $\text{Sym}(D)$ by $g_1, \dots, g_{n!}$. For each $1 \leq r \leq n!$, we can find the number of $f \in \Omega$ satisfying

$$f \circ g_r = \phi_{g_r} \circ f. \tag{2}$$

Suppose that g_r has cycle type $k_{r1}, k_{r2}, \dots, k_{r\ell_r}$. For each $1 \leq i \leq \ell_r$, select a single element $d_{ri} \in D$ from the cycle corresponding to k_{ri} . Then Lemma 5 implies that any $f \in \Omega$ satisfying (2) is determined by its values on each d_{ri} . Hence, to find the number of f satisfying (2), we need only count the number of possible images of d_{ri} under such an f , and then take the product over all i . But the m or fewer elements of D comprising the string $f(d_{ri})$ must each belong to a cycle in the decomposition of g_r whose length divides k_{ri} . For each $1 \leq k \leq m$, there are

$$\left(\sum_{j | k_{ri}} j c(g_r, j) \right)^k$$

choices of $f(d_{ri})$ such that $|f(d_{ri})| = k$. Hence, there are

$$\sum_{k=1}^m \left(\sum_{j | k_{ri}} j c(g_r, j) \right)^k$$

total choices of $f(d_{ri})$. Taking the product over all i , it follows that the number of f satisfying (2) is

$$\prod_{i=1}^{\ell_r} \left(\sum_{k=1}^m \left(\sum_{j | k_{ri}} j c(g_r, j) \right)^k \right). \tag{3}$$

Thus, we've evaluated $|\{f \in \Omega : f \circ g_r = \phi_{g_r} \circ f\}|$, and putting together (1) and (3) gives an expression for the number of equivalence classes in Ω under the relation \sim . Recalling that these classes are in one-to-one correspondence with the classes in Γ under the relation of combinatorial equivalence, we obtain our main result:

Theorem 6. *If $|D| = n$, then the number of m -endomorphisms on D^+ , up to combinatorial equivalence, is the value of*

$$\frac{1}{n!} \sum_{r=1}^{n!} \left(\prod_{i=1}^{\ell_r} \left(\sum_{k=1}^m \left(\sum_{j|k_i} jc(g_r, j) \right)^k \right) \right), \tag{4}$$

where $g_1, \dots, g_{n!}$ are the elements of $\text{Sym}(D)$, and $k_{r1}, \dots, k_{r\ell_r}$ is the cycle type of g_r .

Example 7. Let $D = \{a, b\}$. We find the number of classes of 1-endomorphisms on D^+ . The elements of $\text{Sym}(D)$ (in cycle notation) are $g_1 = (a)(b)$ and $g_2 = (a, b)$. Evidently, $c(g_1, 1) = 2$, $c(g_2, 1) = 0$, and $c(g_2, 2) = 1$. Using Theorem 6, there are

$$\frac{1}{2}(c(g_1, 1)^2 + 2c(g_2, 2)) = \frac{1}{2}(2^2 + 2) = 3$$

classes of 1-endomorphisms on D^+ . These are given by

$$\left\{ \begin{matrix} a \rightarrow a \\ b \rightarrow b \end{matrix} \right\}, \quad \left\{ \begin{matrix} a \rightarrow b \\ b \rightarrow a \end{matrix} \right\} \quad \text{and} \quad \left\{ \begin{matrix} a \rightarrow a & \equiv & a \rightarrow b \\ b \rightarrow a & \equiv & b \rightarrow b \end{matrix} \right\}.$$

We can extend the result of Example 7 by fixing $n = 2$ and letting m be arbitrary. From (4), we find that the number of classes of m -endomorphisms on D^+ , where $|D| = 2$, is

$$\frac{1}{2}((2^{m+1} - 2)^2 + (2^{m+1} - 2)).$$

Running m through the natural numbers, we obtain values 3, 21, 105, 465, 1953, \dots . This is the sequence A134057 in the On-line Encyclopedia of Integers [OEIS 1996]. However, for $n = 3$, the number of classes of m -endomorphisms becomes

$$\frac{1}{6} \left(\left(\frac{3^{m+1} - 3}{2} \right)^3 + 3m \frac{3^{m+1} - 3}{2} + 2 \frac{3^{m+1} - 3}{2} \right).$$

Letting $m = 1, 2, 3, 4, \dots$ gives values 7, 304, 9958, 288280, \dots . This sequence appears to be little-known, and has been submitted by the authors to the OEIS.

2.1. An alternative formulation of Theorem 6. We now present a slight rewording of Theorem 6. In order to compute the number of equivalence classes of m -endomorphisms (where $|D| = n$), we need not, in practice, consider each element of $\text{Sym}(D)$ individually. Rather, we need only consider the cycle types of these permutations. The following well-known result gives the number of permutations in $\text{Sym}(D)$ of a given cycle type.

Proposition 8 [Dummit and Foote 2004]. *Let $|D| = n$, and let $g \in \text{Sym}(D)$. Suppose that m_1, m_2, \dots, m_s are the distinct integers appearing in the cycle type of g . For each $j \in \{1, 2, \dots, s\}$, abbreviate $c_j = c(g, m_j)$. Let C_g be the set of all permutations in $\text{Sym}(D)$ whose cycle type is that of g . Then*

$$|C_g| = \frac{n!}{\prod_{j=1}^s c_j! m_j^{c_j}}. \tag{5}$$

For convenience, we shall say that $g \in \text{Sym}(D)$ fixes the mapping $f \in \Omega$ if and only if $f \circ g = \phi_g \circ f$. Now, two bijections in $\text{Sym}(D)$ with the same cycle type must fix the same number of $f \in \Omega$. Therefore, in order to derive an expression for the number of classes of m -endomorphisms on D^+ , we can select a single representative in $\text{Sym}(D)$ of each possible cycle type, then determine the number of $f \in \Omega$ fixed by each representative using expression (3), multiply this number by the corresponding value of (5), and then sum up over all of our representatives and divide by $n!$. But the cycle types in $\text{Sym}(D)$ are precisely the *integer partitions* of n , namely, the nondecreasing sequences of natural numbers whose sum is n . If $p(n)$ denotes the number of integer partitions of n , then we may restate Theorem 6 as follows.

Corollary 9. *Let $|D| = n$, and suppose that $g_1, \dots, g_{p(n)} \in \text{Sym}(D)$ have distinct cycle types. Then the number of m -endomorphisms on D^+ , up to combinatorial equivalence, is the value of*

$$\frac{1}{n!} \sum_{r=1}^{p(n)} \left(|C_{g_r}| \prod_{i=1}^{\ell_r} \left(\sum_{k=1}^m \left(\sum_{j|k_{r_i}} j c(g_r, j) \right)^k \right) \right), \tag{6}$$

where $k_{r1}, \dots, k_{r\ell_r}$ is the cycle type of g_r , and C_{g_r} is as in Proposition 8.

Example 10. To illustrate Corollary 9, we compute the number of classes of m -endomorphisms when $|D| = 4$. Let $D = \{a, b, c, d\}$. As previously mentioned, the cycle types in $\text{Sym}(D)$ are the integer partitions of 4:

$$1 + 1 + 1 + 1, \quad 1 + 1 + 2, \quad 2 + 2, \quad 1 + 3, \quad 4.$$

Hence, the bijections

$$\begin{aligned} g_1 &= (a)(b)(c)(d), & g_2 &= (a)(b)(c, d), & g_3 &= (a, b)(c, d), \\ g_4 &= (a)(b, c, d), & g_5 &= (a, b, c, d) \end{aligned}$$

encompass all possible cycle types in $\text{Sym}(D)$. Direct calculation using (5) yields

$$|C_{g_1}| = 1, \quad |C_{g_2}| = 6, \quad |C_{g_3}| = 3, \quad |C_{g_4}| = 8, \quad |C_{g_5}| = 6.$$

Thus, by Corollary 9, the number of classes of m -endomorphisms when $n = 4$ is

$$\frac{1}{24} (\Lambda_4^4 + 6\Lambda_2^2 \Lambda_4 + 3\Lambda_4^2 + 8m \Lambda_4 + 6\Lambda_4),$$

where $\Lambda_k = (k^{m+1} - k)/(k - 1)$.

	$n = 1$	$n = 2$	$n = 3$	$n = 4$
$m = 1$	1	3	7	19
$m = 2$	2	21	304	6,915
$m = 3$	3	105	9,958	2,079,567
$m = 4$	4	465	288,280	556,898,155
$m = 5$	5	1,953	7,973,053	144,228,436,231
$m = 6$	6	8,001	217,032,088	37,030,504,349,475

	$n = 5$	$n = 6$
$m = 1$	47	130
$m = 2$	207,258	7,773,622
$m = 3$	746,331,322	409,893,967,167
$m = 4$	2,406,091,382,736	19,560,646,482,079,624
$m = 5$	7,567,019,254,708,782	916,131,223,607,107,471,135
$m = 6$	23,677,181,825,841,420,408	42,770,482,829,102,570,213,645,988

Table 1. Values of (6) for $n, m \leq 6$.

Proceeding along the lines of Example 10, we find that there are

$$\frac{1}{120}(\Lambda_5^5 + 10\Lambda_3^3 \Lambda_5 + 15m\Lambda_5^2 + 20\Lambda_2^2 \Lambda_5 + 20\Lambda_2 \Lambda_3 + 30m \Lambda_5 + 24 \Lambda_5)$$

classes of m -endomorphisms when $n = 5$, and

$$\begin{aligned} \frac{1}{720}(\Lambda_6^6 + 15\Lambda_4^4 \Lambda_6 + 45\Lambda_2^2 \Lambda_6^2 + 15\Lambda_6^3 + 40\Lambda_3^3 \Lambda_6 \\ + 120m\Lambda_3 \Lambda_4 + 40\Lambda_6^2 + 90\Lambda_2^2 \Lambda_6 + 90\Lambda_2 \Lambda_6 + 144m\Lambda_6 + 120\Lambda_6) \end{aligned}$$

classes of m -endomorphisms when $n = 6$. Letting m run through \mathbb{N} in these cases, we again obtain sequences that are not well-known. Table 1 displays the values of (6) for $n, m \leq 6$.

Remark 11. The sequence 1, 3, 7, 19, 47, 130, . . . is sequence A001372 in [OEIS 1996].

3. Two natural variations

In this section, we highlight two natural variations of Corollary 9. First, we restrict our attention to endomorphisms on D^+ that send each element of D to a string of length exactly m . We then consider m -endomorphisms of the so-called free monoid, which contains the empty string. Expressions analogous to those in Section 2 are derived in each case.

3.1. m -uniform endomorphisms. Fix $n, m \in \mathbb{N}$, and suppose that $|D| = n$. Then $\phi \in \text{End}(D^+)$ is called an m -uniform endomorphism if and only if $|\phi(d)| = m$ for

each $d \in D$. In this section, we produce a formula for the number of m -uniform endomorphisms on D^+ up to combinatorial equivalence. To begin, let $g_1, \dots, g_{p(n)} \in \text{Sym}(D)$ have distinct cycle types. We now put $R = \{W \in D^+ : |W| = m\}$ and take Ω to be the set of all mappings of D into R . For each $1 \leq r \leq p(n)$, we ask for the number of $f \in \Omega$ satisfying

$$f \circ g_r = \phi_{g_r} \circ f.$$

Once again, if g_r has cycle type $k_{r1}, \dots, k_{r\ell_r}$, then for each $1 \leq i \leq \ell_r$ we select an element d_{ri} from the cycle corresponding to k_{ri} , and count the number of possible values of $f(d_{ri})$. In this case, we must have $|f(d_{ri})| = m$, where the elements of D comprising the string $f(d_{ri})$ each belong to a cycle whose length divides k_{ri} . Hence, there are

$$\left(\sum_{j|k_{ri}} jc(g_r, j) \right)^m$$

choices of $f(d_{ri})$, and multiplying over all i yields

$$\prod_{i=1}^{\ell_r} \left(\sum_{j|k_{ri}} jc(g_r, j) \right)^m$$

as the value of $|\{f \in \Omega : f \circ g_r = \phi_{g_r} \circ f\}|$. Noting that permutations in $\text{Sym}(D)$ of the same cycle type fix the same number of $f \in \Omega$, we multiply by $|C_{g_r}|$, sum with respect to r , and divide by $n!$ to obtain the following.

Corollary 12. *If $|D| = n$ and $g_1, \dots, g_{p(n)} \in \text{Sym}(D)$ have distinct cycle types, then the number of m -uniform endomorphisms on D^+ , up to combinatorial equivalence, is the value of*

$$\frac{1}{n!} \sum_{r=1}^{p(n)} \left(|C_{g_r}| \prod_{i=1}^{\ell_r} \left(\sum_{j|k_{ri}} jc(g_r, j) \right)^m \right), \tag{7}$$

where $k_{r1}, \dots, k_{r\ell_r}$ is the cycle type of g_r , and C_{g_r} is as in Proposition 8.

When $n = 2$, the number of m -uniform endomorphisms on D^+ , up to combinatorial equivalence, is

$$\frac{1}{2}(2^{2m} + 2^m).$$

Letting $m = 1, 2, 3, 4, \dots$ gives values 3, 10, 36, 136, \dots . This is the sequence A007582 from [OEIS 1996]. Moreover, when $n = 3$ there are

$$\frac{1}{6}(3^{3m} + 3 \cdot 3^m + 2 \cdot 3^m)$$

classes of m -uniform endomorphisms, and letting m run through \mathbb{N} gives the sequence 7, 129, 3303, 88641, \dots , which is not well known. Continuing, the

	$n = 1$	$n = 2$	$n = 3$	$n = 4$
$m = 1$	1	3	7	19
$m = 2$	1	10	129	2,836
$m = 3$	1	36	3,303	700,624
$m = 4$	1	136	88,641	178,981,696
$m = 5$	1	528	7,973,053	45,813,378,304
$m = 6$	1	2,080	64,570,689	11,728,130,323,456

	$n = 5$	$n = 6$
$m = 1$	47	130
$m = 2$	83,061	3,076,386
$m = 3$	254,521,561	141,131,630,530
$m = 4$	794,756,352,216	6,581,201,266,858,896
$m = 5$	2,483,530,604,092,546	307,047,288,863,992,988,160
$m = 6$	7,761,021,959,623,948,401	14,325,590,271,500,876,382,987,456

Table 2. Values of (7) for $n, m \leq 6$.

expressions when $n = 4, 5, 6$ are

$$\begin{aligned} & \frac{1}{24}(4^{4m} + 6 \cdot 2^{2m} \cdot 4^m + 3 \cdot 4^{2m} + 8 \cdot 4^m + 6 \cdot 4^m), \\ & \frac{1}{120}(5^{5m} + 10 \cdot 3^{3m} \cdot 5^m + 15 \cdot 5^{2m} + 20 \cdot 2^{2m} \cdot 5^m + 20 \cdot 2^m \cdot 3^m + 30 \cdot 5^m + 24 \cdot 5^m), \\ & \frac{1}{720}(6^{6m} + 15 \cdot 4^{4m} \cdot 6^m + 45 \cdot 2^{2m} \cdot 6^{2m} + 15 \cdot 6^{3m} + 40 \cdot 3^{3m} \cdot 6^m \\ & \quad + 120 \cdot 3^m \cdot 4^m + 40 \cdot 6^{2m} + 90 \cdot 2^{2m} \cdot 6^m + 90 \cdot 2^m \cdot 6^m + 144 \cdot 6^m + 120 \cdot 6^m), \end{aligned}$$

respectively. Table 2 displays the values of (7) for $n, m \leq 6$.

3.2. The free monoid. If we adjoin the unique string of length 0 (denoted by ϵ) to the set D^+ , then we form the set D^* . Paired with the operation of string concatenation, D^* forms the *free monoid* on D . We refer to ϵ as the *empty string*, and it serves as the identity element in D^* . That is, for any $W \in D^*$,

$$W\epsilon = W = \epsilon W.$$

We define an endomorphism on D^* to be a mapping $\phi : D^* \rightarrow D^*$ such that $\phi(W_1W_2) = \phi(W_1)\phi(W_2)$ for all $W_1, W_2 \in D^*$.

Remark 13. Note that if ϕ is an endomorphism on D^* , then $\phi(\epsilon) = \epsilon$. This follows since for any $W \in D^*$, we have

$$\phi(W) = \phi(\epsilon W) = \phi(\epsilon)\phi(W),$$

which implies that $\phi(\epsilon)$ has length 0.

Now, an m -endomorphism on D^* is an endomorphism such that $|\phi(d)| \leq m$ for all $d \in D$. Thus, an m -endomorphism on D^* can map elements of D to ϵ . To determine the number of m -endomorphisms on D^* up to combinatorial equivalence, we put $R = \{W \in D^* : |W| \leq m\}$, and for each $g \in \text{Sym}(D)$, we ask for the number of $f : D \rightarrow R$ that are fixed by g . Again, it suffices to count the number of possible images under such an f of a single $d \in D$ from each cycle in the decomposition of g , and then multiply over all the cycles. But there is now one additional possible value of $f(d)$: the empty string. Hence, if d belongs to a cycle of length k_i , then we have

$$1 + \sum_{k=1}^m \left(\sum_{j|k_i} jc(g_r, j) \right)^k = \sum_{k=0}^m \left(\sum_{j|k_i} jc(g_r, j) \right)^k$$

choices of $f(d)$. From this observation, we deduce the following.

Corollary 14. *Let $|D| = n$, and suppose that $g_1, \dots, g_{p(n)} \in \text{Sym}(D)$ have distinct cycle types. Then the number of m -endomorphisms on D^* , up to combinatorial equivalence, is the value of*

$$\frac{1}{n!} \sum_{r=1}^{p(n)} \left(|C_{g_r}| \prod_{i=1}^{\ell_r} \left(\sum_{k=0}^m \left(\sum_{j|k_{r_i}} jc(g_r, j) \right)^k \right) \right), \tag{8}$$

where $k_{r1}, \dots, k_{r\ell_r}$ is the cycle type of g_r , and C_{g_r} is as in Proposition 8.

When $n = 2$, the number of m -endomorphisms on D^* , up to combinatorial equivalence, is

$$\frac{1}{2}((2^{m+1} - 1)^2 + (2^{m+1} - 1)).$$

This is sequence A006516 from [OEIS 1996]. The corresponding expressions for $n = 3, 4, 5, 6$ are

$$\begin{aligned} & \frac{1}{6}(\Delta_3^3 + 3(m+1)\Delta_3 + 2\Delta_3), \\ & \frac{1}{24}(\Delta_4^4 + 6\Delta_2^2\Delta_4 + 3\Delta_4^2 + 8(m+1)\Delta_4 + 6\Delta_4), \\ & \frac{1}{120}(\Delta_5^5 + 10\Delta_3^3\Delta_5 + 15(m+1)\Delta_5^2 + 20\Delta_2^2\Delta_5 + 20\Delta_2\Delta_3 + 30(m+1)\Delta_5 + 24\Delta_5), \\ & \frac{1}{720}(\Delta_6^6 + 15\Delta_4^4\Delta_6 + 45\Delta_2^2\Delta_6^2 + 15\Delta_6^3 + 40\Delta_3^3\Delta_6 + 120(m+1)\Delta_3\Delta_4 \\ & \quad + 40\Delta_6^2 + 90\Delta_2^2\Delta_6 + 90\Delta_2\Delta_6 + 144(m+1)\Delta_6 + 120\Delta_6), \end{aligned}$$

where $\Delta_k = (k^{m+1} - 1)/(k - 1)$. Once again, the sequences given by these expressions appear to be little-known. Table 3 gives the values of (8) for $n, m \leq 6$.

4. (χ, ζ) -patterns

In closing, we briefly place the relation \sim from Section 1 into a more general context. Let G be a finite group, and let N and M be finite nonempty sets. Suppose

	$n = 1$	$n = 2$	$n = 3$	$n = 4$
$m = 1$	2	6	16	45
$m = 2$	3	28	390	8,442
$m = 3$	4	120	10,760	2,180,845
$m = 4$	5	496	295,603	563,483,404
$m = 5$	6	2,016	8,039,304	144,651,898,755
$m = 6$	7	8,128	217,629,416	37,057,640,711,850

	$n = 5$	$n = 6$
$m = 1$	121	338
$m = 2$	244,910	8,967,034
$m = 3$	770,763,470	419,527,164,799
$m = 4$	2,421,556,983,901	19,636,295,549,860,505
$m = 5$	2,370,422,688,990,078	916,720,535,022,517,503,173
$m = 6$	23,683,244,198,577,149,289	42,775,066,732,111,188,868,070,978

Table 3. Values of (8) for $n, m \leq 6$.

that $\chi : G \rightarrow \text{Sym}(N)$ and $\zeta : G \rightarrow \text{Sym}(M)$ are group homomorphisms. Denote the set of all functions from N into M by M^N . This notation comes from de Bruijn [1972], who also introduced the equivalence relation $E_{\chi, \zeta}$ on M^N defined by

$$(f_1, f_2) \in E_{\chi, \zeta} \iff f_2 \circ \chi(\gamma) = \zeta(\gamma) \circ f_1 \text{ for some } \gamma \in G.$$

Example 15 [de Bruijn 1972]. Suppose that N is a set of size $n \in \mathbb{N}$, and define an equivalence relation S on the set of all mappings of N into itself by

$$(f_1, f_2) \in S \iff f_2 \circ \gamma = \gamma \circ f_1 \text{ for some } \gamma \in \text{Sym}(N).$$

Letting $G = \text{Sym}(N)$, $M = N$, and $\chi = \zeta$ be the identity homomorphism on $\text{Sym}(N)$ shows that S is a special case of the relation $E_{\chi, \zeta}$. Moreover, the sequence in Remark 11 gives the number of equivalence classes under S for $n = 1, 2, 3 \dots$ (See [de Bruijn 1972, § 3].)

The relation $E_{\chi, \zeta}$ stems from the left action of G on M^N given by

$$\gamma \cdot f = \zeta(\gamma) \circ f \circ \chi(\gamma^{-1})$$

for all $\gamma \in G, f \in M^N$. De Bruijn [1972] referred to the orbits of G on M^N as (χ, ζ) -patterns, and provided a formula for the number of these by applying Burnside’s lemma, and then evaluating $|\{f \in M^N : \gamma \cdot f = f\}|$ for each $\gamma \in G$. But the relation \sim on the set $\Omega = \{\text{mappings of } D \text{ into } R\}$, where $0 < |D| < \infty$ and $R = \{W \in D^+ : |W| \leq m\}$, is a special instance of the relation $E_{\chi, \zeta}$. To see this,

take $N = D$, $M = R$, and $G = \text{Sym}(D)$. Let χ be the identity homomorphism on $\text{Sym}(D)$, and define $\zeta : G \rightarrow \text{Sym}(R)$ by

$$\zeta(g) = \phi_g|_R$$

for all $g \in \text{Sym}(D)$. Then for any $g, g' \in \text{Sym}(D)$,

$$\zeta(g \circ g') = \phi_{g \circ g'}|_R = (\phi_g \circ \phi_{g'})|_R = \phi_g|_R \circ \phi_{g'}|_R = \zeta(g) \circ \zeta(g'),$$

so ζ is a group homomorphism. Now, for any $f_1, f_2 \in \Omega$, we have

$$\begin{aligned} f_1 \sim f_2 &\iff f_2 \circ g = \phi_g \circ f_1 = \phi_g|_R \circ f_1 \text{ for some } g \in \text{Sym}(D) \\ &\iff f_2 \circ \chi(g) = \zeta(g) \circ f_1 \text{ for some } g \in \text{Sym}(D) \\ &\iff (f_1, f_2) \in E_{\chi, \zeta}. \end{aligned}$$

It follows that the equivalence classes in Ω under the relation \sim are (χ, ζ) -patterns for χ, ζ chosen as above. In particular, our Theorem 6 is a special case of de Bruijn's formula.

Acknowledgments

We thank the anonymous referee, whose numerous observations and suggestions led to substantial revision. This research was supported by Temple University's Undergraduate Research Program.

References

- [de Bruijn 1964] N. G. de Bruijn, "Pólya's theory of counting", pp. 144–184 in *Applied combinatorial mathematics*, edited by E. F. Bechenbach, Wiley, New York, 1964. Zbl 0144.00601
- [de Bruijn 1972] N. G. de Bruijn, "Enumeration of mapping patterns", *J. Combinatorial Theory Ser. A* **12**:1 (1972), 14–20. MR 0284357 Zbl 0239.05007
- [Dummit and Foote 2004] D. S. Dummit and R. M. Foote, *Abstract algebra*, 3rd ed., Wiley, Hoboken, NJ, 2004. MR 2286236 Zbl 1037.00003
- [Malik et al. 1997] D. S. Malik, J. N. Mordeson, and M. Sen, *Fundamentals of abstract algebra*, McGraw-Hill, New York, 1997.
- [OEIS 1996] OEIS, "The on-line encyclopedia of integer sequences", 1996, available at <http://oeis.org>.

Received: 2015-02-06 Revised: 2015-07-14 Accepted: 2015-07-20

rubinlj@slu.edu

*Department of Mathematics and Computer Science,
St. Louis University, 220 North Grand Boulevard,
St. Louis, MO 63103, United States*

brirush@mathematics.byu.edu

*Department of Mathematics, Brigham Young University,
268 TMCB, Provo, UT 84602, United States*

Quantum Schubert polynomials for the G_2 flag manifold

Rachel E. Elliott, Mark E. Lewers and Leonardo C. Mihalcea

(Communicated by Jim Haglund)

We study some combinatorial objects related to the flag manifold X of Lie type G_2 . Using the moment graph of X , we calculate all the curve neighborhoods for Schubert classes. We use this calculation to investigate the ordinary and quantum cohomology rings of X . As an application, we obtain positive Schubert polynomials for the cohomology ring of X and we find quantum Schubert polynomials which represent Schubert classes in the quantum cohomology ring of X .

1. Introduction

One of the major theorems in algebra is the classification of complex semisimple Lie algebras. There are four classical infinite series (of types A_n , B_n , C_n , D_n) and five exceptional finite series (of types E_6 , E_7 , E_8 , F_4 , G_2). To each algebra, one can associate a group and to each group a certain geometric object called a flag manifold. In type A_n , the points of this flag manifold are sequences $V_1 \subset V_2 \subset \cdots \subset \mathbb{C}^n$ of vector spaces V_i of dimension i . The algebra of type G_2 is considered the simplest among the exceptional series, and we denote by X the flag manifold for type G_2 . The study of flag manifolds has a long and rich history starting in the 1950s, and it lies at the intersection of algebraic geometry, combinatorics, topology and representation theory.

One can associate a ring to the flag manifold X called the cohomology ring $H^*(X)$. This ring has a distinguished basis given by Schubert classes σ_w , indexed by the elements w in the Weyl group W of type G_2 ; see Section 4 below. We recall that W is actually isomorphic to the dihedral group with 12 elements, although we will use a different realization of it which is more suitable for our purposes. This ring is generated by Schubert classes σ_{s_1} , σ_{s_2} for the simple reflections s_1 , s_2 in W . Therefore, at least in principle, the full multiplication table in the ring is determined by a formula to multiply one Schubert class by another for either s_1 or s_2 . This is called a *Chevalley formula*. There has been a substantial amount of work to find

MSC2010: primary 14N15; secondary 14M15, 14N35, 05E15.

Keywords: quantum cohomology, Schubert polynomial, G_2 flag manifold.

Mihalcea was supported in part by NSA Young Investigator Award H98230-13-1-0208 .

Chevalley formulas for this ring, starting with Chevalley [1994] in the 1950s. This formula can be expressed combinatorially in terms of the root system and the Weyl group for type G_2 . Alternatively, the cohomology ring has a “Borel” presentation $H^*(X) = \mathbb{Q}[x_1, x_2]/I$, where I is the ideal generated by $x_1^2 - x_1x_2 + x_2^2$ and x_1^6 . A natural question is to find out what is the relation between this “algebraic” presentation and the “geometric” one which involves the Schubert basis. In other words, one needs to find a polynomial in $\mathbb{Q}[x_1, x_2]$ which represents a Schubert class σ_w under the isomorphism $H^*(X) = \mathbb{Q}[x_1, x_2]$. This is called a *Schubert polynomial*. Such polynomials are not unique, as their class in $\mathbb{Q}[x_1, x_2]/I$ is unchanged if one changes a polynomial by elements in I . In Section 5, we use the Chevalley rule to find Schubert polynomials for σ_w . Some of our polynomials coincide with similar Schubert polynomials found by D. Anderson [2011], via different methods. The polynomials we found are homogeneous and have *positive* coefficients. Given that the positivity of Schubert polynomial coefficients has geometric interpretations in type A_n (see the paper of A. Knutson and E. Miller [2005]), this is a desirable property.

The current paper also focuses on a deformation of the ring above called the *quantum cohomology ring* $\text{QH}^*(X)$. It is a deformation of $H^*(X)$ with the addition of quantum parameters $q^{\mathbf{d}} = q_1^{d_1}q_2^{d_2}$ for degrees $\mathbf{d} = (d_1, d_2)$. If $\mathbf{d} = (0, 0)$, or equivalently $q_1 = q_2 = 0$, the product reduces to the corresponding calculation in $H^*(X)$. More detail will be given in Section 4. See [Fulton and Pandharipande 1997] for more information about the background/history of this ring. Similar to the ring $H^*(X)$, the quantum cohomology ring has a $\mathbb{Z}[q]$ -basis consisting of Schubert classes σ_w (where $q = (q_1, q_2)$ are the quantum parameters), and it is generated as a ring by the classes σ_{s_1} and σ_{s_2} for the simple reflections s_1 and s_2 .

The *quantum Chevalley formula* is a formula for the quantum multiplication $\sigma_w \star \sigma_{s_i}$ ($i = 1, 2$). An explicit form of this formula, which uses combinatorics of the root system of Lie type G_2 , was obtained by Fulton and Woodward [2004]. In this paper, we use the “curve neighborhoods” method to write down the explicit Chevalley formula. This alternative method, obtained by Buch and Mihalcea [2015], involves an interesting graph associated to the flag manifold, called the *moment graph*. Its definition and properties are found in Section 3. It also has the advantage that it leads to a conjectural Chevalley formula in a further deformation of the quantum cohomology ring, called *quantum K-theory*. This will be addressed in a follow-up paper.

Our main application is to obtain a quantum version of the Schubert polynomials. More precisely, it is known [Fulton and Pandharipande 1997, Proposition 11] that $\text{QH}^*(X) = \mathbb{Q}[x_1, x_2, q_1, q_2]/\tilde{I}$, where \tilde{I} a certain ideal which deforms I . Then, as in the classical case, we would like to find the polynomials in $\mathbb{Q}[x_1, x_2, q_1, q_2]$ which represent each Schubert class σ_w via the isomorphism $\text{QH}^*(X) = \mathbb{Q}[x_1, x_2, q_1, q_2]/\tilde{I}$. These are called *quantum Schubert polynomials*. As before, these polynomials are not unique, but we can impose some natural

conditions that they satisfy, such as the fact that they deform the ordinary Schubert polynomials, and that they are homogeneous with respect to a certain grading. To our knowledge, such polynomials have not been explicitly calculated in the literature. As a byproduct, we also use the quantum Chevalley formula to recover the ideal \tilde{I} of quantum relations. This ideal has been, in principle, calculated by Kim [1999] using different techniques, but the explicit polynomials generating this ideal do not seem to appear in the literature. Our results are stated in Theorem 5.2 below.

2. Preliminaries: the root system and the Weyl group of type G_2

2A. The G_2 root system. Denote by R the root system of type G_2 . It consists of 12 roots, which are nonzero vectors in the hyperplane in \mathbb{R}^3 given by the equation $\xi_1 + \xi_2 + \xi_3 = 0$; our main reference is [Bourbaki 2002]. The roots are displayed in Table 1, in terms of the natural coordinates in \mathbb{R}^3 . Each root α can be written uniquely as $\alpha = c_1\alpha_1 + c_2\alpha_2$, where α_1, α_2 are *simple roots* and $c_1c_2 \geq 0$. A root is *positive (negative)* if both c_1, c_2 are nonnegative (resp. nonpositive). The set of simple roots is denoted by $\Delta = \{\alpha_1, \alpha_2\}$, where $\alpha_1 = \epsilon_1 - \epsilon_2$ and $\alpha_2 = -2\epsilon_1 + \epsilon_2 + \epsilon_3$. For later purposes, we need to expand each root in terms of the simple roots. The full results are shown in Table 1. The root vectors in the Δ -basis can be seen in Figure 1.

We also need the *dual* root system consisting of *coroots* α^\vee . The coroot α^\vee of a root α is defined as $\alpha^\vee = 2\alpha/(\alpha, \alpha)$, where (α, α) is the standard inner product in \mathbb{R}^3 . Note that the coroots satisfy the properties $(\alpha^\vee)^\vee = \alpha$ and $(-\alpha)^\vee = -\alpha^\vee$. We denote the full set of coroots by R^\vee and define the set Δ^\vee , which holds the *simple coroots* α_1^\vee and α_2^\vee for R^\vee . Table 1 shows the values for each of the coroots.

2B. The Weyl group of G_2 . The *Weyl group* of G_2 , denoted W , is the group generated by reflections s_α , where $\alpha \in R$. Let $s_i := s_{\alpha_i}$. Geometrically, s_α is the reflection across the line perpendicular to the root α . For example, the reflection s_1

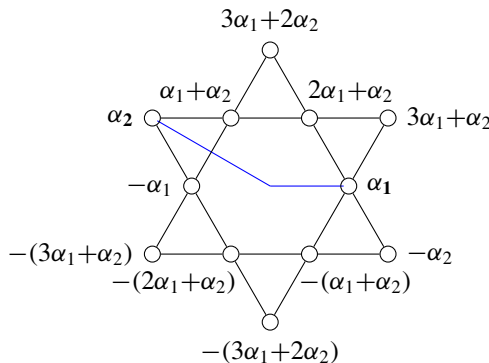


Figure 1. The root system for G_2 . Each node is a root. The blue lines represent the coordinate system using the Δ -basis.

natural coordinates E -basis $(\epsilon_1, \epsilon_2, \epsilon_3)$	\pm	simple roots basis (α_1, α_2)	coroot α^\vee $\alpha^\vee = \lambda\alpha_1 + \mu\alpha_2$
$\epsilon_1 - \epsilon_2$	+	α_1	$\alpha_1^\vee = \alpha_1$
$\epsilon_3 - \epsilon_1$	+	$\alpha_1 + \alpha_2$	$(\alpha_1 + \alpha_2)^\vee = \alpha_1 + \alpha_2$
$\epsilon_3 - \epsilon_2$	+	$2\alpha_1 + \alpha_2$	$(2\alpha_1 + \alpha_2)^\vee = 2\alpha_1 + \alpha_2$
$\epsilon_2 + \epsilon_3 - 2\epsilon_1$	+	α_2	$\alpha_2^\vee = \frac{1}{3}\alpha_2$
$\epsilon_1 + \epsilon_3 - 2\epsilon_2$	+	$3\alpha_1 + \alpha_2$	$(3\alpha_1 + \alpha_2)^\vee = \alpha_1 + \frac{1}{3}\alpha_2$
$-\epsilon_1 - \epsilon_2 + 2\epsilon_3$	+	$3\alpha_1 + 2\alpha_2$	$(3\alpha_1 + 2\alpha_2)^\vee = \alpha_1 + \frac{2}{3}\alpha_2$
$-(\epsilon_1 - \epsilon_2)$	-	$-\alpha_1$	$(-\alpha_1)^\vee = -\alpha_1$
$-(\epsilon_3 - \epsilon_1)$	-	$-(\alpha_1 + \alpha_2)$	$(-\alpha_1 - \alpha_2)^\vee = -\alpha_1 - \alpha_2$
$-(\epsilon_3 - \epsilon_2)$	-	$-(2\alpha_1 + \alpha_2)$	$(-2\alpha_1 - \alpha_2)^\vee = -2\alpha_1 - \alpha_2$
$-(\epsilon_2 + \epsilon_3 - 2\epsilon_1)$	-	$-\alpha_2$	$(-\alpha_2)^\vee = -\frac{1}{3}\alpha_2$
$-(\epsilon_1 + \epsilon_3 - 2\epsilon_2)$	-	$-(3\alpha_1 + \alpha_2)$	$(-3\alpha_1 - \alpha_2)^\vee = -\alpha_1 - \frac{1}{3}\alpha_2$
$-(-\epsilon_1 - \epsilon_2 + 2\epsilon_3)$	-	$-(3\alpha_1 + 2\alpha_2)$	$(-3\alpha_1 - 2\alpha_2)^\vee = -\alpha_1 - \frac{2}{3}\alpha_2$

Table 1. The root system of type G_2 . For each root, we give its sign, the root in terms of Δ -basis, and the corresponding coroot.

(corresponding to s_{α_1}) is the reflection across the line perpendicular to the α_1 -axis (see Figure 2). As Figure 2 shows, for any root α , we have $s_\alpha = s_{-\alpha}$. Therefore only six unique reflections exist for the G_2 root system.

It is known (see, e.g., [Humphreys 1972]) that W has the presentation

$$W = \langle s_1, s_2 : s_1^2 = s_2^2 = 1, (s_1s_2)^6 = 1 \rangle.$$

From this it follows easily that W is isomorphic to the dihedral group with 12 elements. In order to determine the reflections in W , we need the following definitions.

Definition 2.1. Consider $w \in W$. A *reduced expression* for w is one involving products of s_1 and s_2 in as short a way as possible (via the relations in the presentation). If $w \in W$, where w is a reduced expression, the *length* of w , denoted by $\ell(w)$, is the number of simple reflections (s_1 and s_2) that show up in the reduced expression.

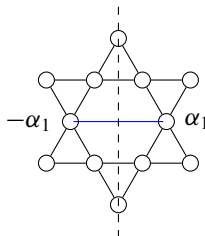


Figure 2. The reflection s_{α_1} (dashed line) which is perpendicular to the α_1 -axis (blue line).

root (in Δ -basis)	reflection ($w \in W$)
$\pm\alpha_1$	s_1
$\pm\alpha_2$	s_2
$\pm(3\alpha_1 + \alpha_2)$	$s_1s_2s_1$
$\pm(\alpha_1 + \alpha_2)$	$s_2s_1s_2$
$\pm(2\alpha_1 + \alpha_2)$	$s_1s_2s_1s_2s_1$
$\pm(3\alpha_1 + 2\alpha_2)$	$s_2s_1s_2s_1s_2$

Table 2. The root reflection corresponding to each root in G_2 .

Example 2.2. Consider $w = s_1s_1s_1s_2s_1s_2$. From the presentation of W , we know that $s_1^2 = s_1s_1 = 1$ and so this expression is not reduced. However, $(s_1s_1)s_1s_2s_1s_2 = (1)s_1s_2s_1s_2 = s_1s_2s_1s_2$. The latter is a reduced expression and $\ell(w) = 4$.

The 12 reduced expressions of the elements in W are

$$W = \{1, s_1, s_2, s_1s_2, s_2s_1, s_1s_2s_1, s_2s_1s_2, s_1s_2s_1s_2, s_1s_2s_1s_2s_1, s_2s_1s_2s_1s_2, s_1s_2s_1s_2s_1s_2, s_2s_1s_2s_1s_2s_1s_2\}.$$

We denote by w_0 the longest element $s_1s_2s_1s_2s_1s_2$. Notice that among the twelve elements, only six of them are the *root reflections* from the root system of G_2 . Because any reflection has order 2, it is easy to check that the root reflections correspond to the reduced expressions of odd length.

Since the reflections s_1 and s_2 generate W , every reflection s_α in the G_2 root system can be expressed as a reduced expression product of s_1s and s_2s . Consider the action of W on the root system R given by the natural action of reflections on vectors in \mathbb{R}^3 . Explicitly, this action is given by $s_\alpha \cdot \beta = s_\alpha(\beta) = \beta - (\beta, \alpha^\vee)\alpha$ (see [Humphreys 1972, p. 43]). The following lemma is proved in [loc. cit.].

Lemma 2.3. *Let $w \in W$ and $\alpha \in R$. Then $ws_\alpha w^{-1} = s_{w \cdot \alpha}$.*

Example 2.4. Consider $w = s_1s_2s_1$. We want to find a reflection s_α that corresponds to w . By Lemma 2.3, $s_1s_2s_1 = s_{s_1(\alpha_2)}$, where s_1 is its own inverse and the action is

$$s_1(\alpha_2) = \alpha_2 - (\alpha_2, \alpha_1^\vee)\alpha_1 = \alpha_2 - \left(\alpha_2, \frac{2\alpha_1}{(\alpha_1, \alpha_1)}\right)\alpha_1.$$

We know $(\alpha_1, \alpha_1) = 2$, (see Table 1) so

$$\begin{aligned} \alpha_2 - \left(\alpha_2, \frac{2\alpha_1}{(\alpha_1, \alpha_1)}\right)\alpha_1 &= \alpha_2 - \left(\alpha_2, \frac{2\alpha_1}{2}\right)\alpha_1 = \alpha_2 - (\alpha_2, \alpha_1)\alpha_1 \\ &= \alpha_2 - (-3)\alpha_1. \end{aligned}$$

Thus $s_1(\alpha_2) = 3\alpha_1 + \alpha_2$. The reflection $s_1s_2s_1$ is the reflection $s_{3\alpha_1 + \alpha_2}$.

Table 2 shows the reflection across the line perpendicular to each root. Notice that roots α and $-\alpha$ have the same reflection and all reflections listed have *odd* length.

coroot $\alpha^\vee = d_1\alpha_1^\vee + d_2\alpha_2^\vee$	degree \mathbf{d} (d_1, d_2)
α_1^\vee	(1, 0)
α_2^\vee	(0, 1)
$(3\alpha_1 + \alpha_2)^\vee$	(1, 1)
$(\alpha_1 + \alpha_2)^\vee$	(1, 3)
$(2\alpha_1 + \alpha_2)^\vee$	(2, 3)
$(3\alpha_1 + 2\alpha_2)^\vee$	(1, 2)

Table 3. The degree for each coroot in the moment graph.

3. The moment graph and curve neighborhoods

3A. Finding the moment graph. Using the properties of the elements in the Weyl group for G_2 , it is possible to define the following graph.

Definition 3.1. The *moment graph* is an oriented graph that consists of a pair (V, E) , where V is the set of vertices and E is the set of edges. To each Weyl group element $v \in W$ there corresponds a vertex $v \in V$ in this graph. For $x, y \in V$, an edge exists from x to y , denoted by

$$x \xrightarrow{\alpha^\vee} y,$$

if there exists a reflection s_α such that $y = xs_\alpha$ and $\ell(y) > \ell(x)$.

Definition 3.2. A *degree \mathbf{d}* is a nonnegative combination $d_1\alpha_1^\vee + d_2\alpha_2^\vee$ of simple coroots. We will denote it as $\mathbf{d} = (d_1, d_2)$.

Since any coroot α^\vee is a linear combination in terms of α_1^\vee and α_2^\vee , it determines a degree. These degrees are given in Table 3.

Example 3.3. An edge exists from s_1 to s_2s_1 . This is so because

$$\ell(s_2s_1) > \ell(s_1) \quad \text{and} \quad s_2s_1 = s_1s_\alpha, \quad \text{where } s_\alpha = s_1s_2s_1.$$

Example 2.4 shows $s_1s_2s_1 = s_{3\alpha_1 + \alpha_2}$. The edge corresponding to these two edges has degree $(3\alpha_1 + \alpha_2)^\vee$; i.e.,

$$s_1 \xrightarrow{(3\alpha_1 + \alpha_2)^\vee} s_2s_1.$$

Notice that $(3\alpha_1 + \alpha_2)^\vee = 1\alpha_1^\vee + 1\alpha_2^\vee$, so $\mathbf{d} = (1, 1)$. The edge from s_1 to s_2s_1 can be represented by the degree $(1, 1)$.

We depict the moment graph as oriented upward, as in Figure 3. To help read the moment graph, a color code has been set up to represent the different edges. We review some of the relevant properties of the moment graph:

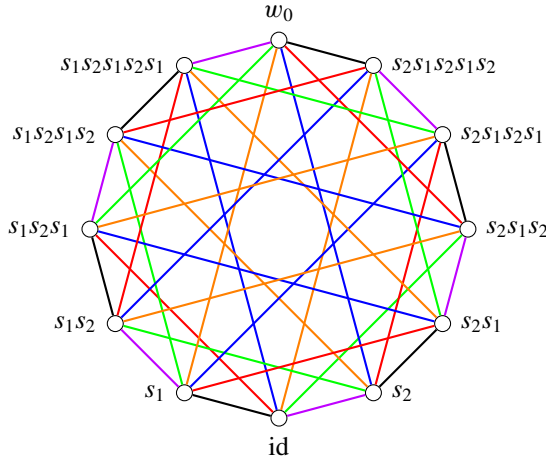


Figure 3. The moment graph for G_2 . The color code for the degrees is black = $(1, 0)$, violet = $(0, 1)$, red = $(1, 1)$, green = $(1, 3)$, blue = $(2, 3)$, orange = $(1, 2)$.

- The vertices correspond to the 12 Weyl group elements.
- The edges represent the root reflections associated to the G_2 root system. There are six different types of edges (different degree values) because there are exactly six reflections in the G_2 root system. Note that edges exist between Weyl group elements if the difference between lengths is odd.
- The bottom vertex is the element with the smallest length (id, where $\ell(\text{id}) = 0$). The vertices in the next “row” have length 1 (s_1 and s_2). The length of these elements increases by one as you travel up the graph. The top vertex is the element with the largest length, w_0 , where $\ell(w_0) = 6$.
- For any vertex, there are six edges connected to it, corresponding to the six different coroots in R^\vee .
- For any $w_1, w_2 \in W$, where $\ell(w_1) = \ell(w_2)$, both w_1 and w_2 will have edges connecting to the same six vertices.

3B. Curve neighborhoods. In Section 3A, we defined the degree d to help simplify the moment graph for use in future calculations. The importance of the moment graph can be realized with the following concept defined by A. Buch and L. Mihalea [2015]:

Definition 3.4. Fix a degree $d = (d_1, d_2)$ and an element u of the Weyl group W . The *curve neighborhood*, $\Gamma_d(u)$, is a subset of W which consists of the maximal elements in the moment graph which can be reached from u with a path of total degree at most d .

Example 3.5. Consider $w = \text{id}$ and $\mathbf{d} = (1, 1)$. We want to determine the “highest” path (starting at the identity) where the total degree traveled is at most $(1, 1)$. By inspecting the moment graph, we see that there are three initial paths starting from id :

- *path* $\mathbf{d} = (1, 0)$, which goes from id to s_1 . Upon reaching s_1 , one is not allowed to travel more than $\mathbf{d}' = (0, 1)$ upwards. Further inspection of the moment graph shows that a path exists with degree $(0, 1)$ from s_1 to s_1s_2 . We now have traveled a total degree of $(1,1)$. Thus we are done and s_1s_2 is the largest element on this path.
- *path* $\mathbf{d} = (0, 1)$ which goes from id to s_2 . Upon reaching s_2 , one is not allowed to travel more than $\mathbf{d}' = (1, 0)$ upwards. Further inspection gives a path with degree $(1, 0)$ from s_2 to s_2s_1 . We now have traveled a total degree of $(1, 1)$. Thus we are done and s_2s_1 is the largest element on this path.
- *path* $\mathbf{d} = (1, 1)$ which goes from id to $s_1s_2s_1$. Since we traveled a total degree of $(1, 1)$, we are done, and $s_1s_2s_1$ is the largest element on this path.

We now take the maximal element that can be reached from id with degree $(1, 1)$. The largest of the three elements above is $s_1s_2s_1$; thus $\Gamma_{(1,1)}(\text{id}) = \{s_1s_2s_1\}$.

It is clear that for any $w \in W$, there exists some degree (a, b) where $\Gamma_{(a,b)}(w) = w_0$. Then for any larger degree (a', b') , where $a' \geq a$ and $b' \geq b$, we have $\Gamma_{(a',b')}(w) = w_0$. Table 6 in the Appendix shows the curve neighborhoods for every element of the Weyl group. For all the examples given, the curve neighborhood for some degree \mathbf{d} at $u \in W$ is always unique, a fact which was initially proved in [Buch and Mihalcea 2015] for all Lie types.

4. Quantum cohomology ring for flag manifold X

Recall that X denotes the flag manifold of type G_2 . The cohomology ring, denoted by $H^*(X)$, consists of elements that can each be written uniquely as finite sums $\sum_{w \in W} a_w \sigma_w$, where $a_w \in \mathbb{Z}$ and σ_w is a (geometrically defined) *Schubert class*. Addition in this ring is given by

$$\sum_{w \in W} a_w \sigma_w + \sum_{w \in W} b_w \sigma_w = \sum_{w \in W} (a_w + b_w) \sigma_w.$$

The quantum cohomology ring $\text{QH}^*(X)$ is a deformation of $H^*(X)$ by adding quantum parameters, $q^{\mathbf{d}} = q_1^{d_1} q_2^{d_2}$ for degrees $\mathbf{d} = (d_1, d_2)$. If $\mathbf{d} = (0, 0)$ for any calculation in $\text{QH}^*(X)$, we reduce down to the corresponding calculation in $H^*(X)$. Similarly to $H^*(X)$, the elements of $\text{QH}^*(X)$ can each be written uniquely as finite sums $\sum_{w \in W} a_w(\mathbf{d}) q^{\mathbf{d}} \sigma_w$, where $a_w(\mathbf{d}) \in \mathbb{Z}$. The addition in this ring is also straightforward:

$$\sum_{w \in W} a_w(\mathbf{d}) q^{\mathbf{d}} \sigma_w + \sum_{w \in W} b_w(\mathbf{d}) q^{\mathbf{d}} \sigma_w = \sum_{w \in W} (a_w(\mathbf{d}) + b_w(\mathbf{d})) q^{\mathbf{d}} \sigma_w.$$

The multiplication in this ring is given by certain integers $c_{u,v}^{w,d}$ called the *Gromov–Witten invariants*:

$$\sigma_u \star \sigma_v = \sum_{w,d} c_{u,v}^{w,d} q^d \sigma_w,$$

where the sum is over $w \in W$ and degrees d which have nonnegative components. The (quantum) cohomology ring has two generators, namely σ_{s_1} and σ_{s_2} , corresponding to the simple reflections $s_1, s_2 \in W$. As a result, every element is a sum of monomials in the σ_{s_i} , and the quantum multiplication $\sigma_u \star \sigma_{s_i}$ by generators σ_{s_i} determines the entire ring multiplication. The formula for $\sigma_w \star \sigma_{s_i}$, the (*quantum*) *Chevalley rule*, is illustrated in Section 4A. We list below a few properties that will help to understand this ring and we refer, e.g., to [Fulton and Pandharipande 1997] for full details.

- (1) The multiplication of quantum parameters is given by $q_i^{d_s} q_i^{d'_s} = q_i^{d_s+d'_s}$.
- (2) The quantum multiplication \star is associative, commutative and has unit $1 = \sigma_{\text{id}}$.
- (3) The quantum multiplication is graded by imposing $\deg(\sigma_w) = \ell(w)$ and for $d = (d_1, d_2)$, we have $\deg q^d = 2(d_1 + d_2)$. This implies that $\deg(\sigma_u \star \sigma_v) = \deg(\sigma_u) + \deg(\sigma_v)$ and that $c_{u,v}^{w,d} = 0$ unless $\ell(u) + \ell(v) = \ell(w) + \deg q^d$.
- (4) If we impose the substitution $q_1 = q_2 = 0$ in $\sigma_u \star \sigma_v$ then we obtain the multiplication $\sigma_u \cdot \sigma_v$ in the ordinary cohomology ring $H^*(X)$.

4A. Quantum Chevalley rule via curve neighborhoods. Recall that each coroot α^\vee can be written as a linear combination $\alpha^\vee = d_1 \alpha_1^\vee + d_2 \alpha_2^\vee$, where $\alpha_1^\vee, \alpha_2^\vee$ are the simple coroots and $d_1, d_2 \in \mathbb{Z}$. It follows that each α^\vee can be identified with the unique degree $d = (d_1, d_2)$. Let $d[i]$ denote the i -th component of the degree d in the decomposition $d = d[1] \alpha_1^\vee + d[2] \alpha_2^\vee$. In other words, $d[i] = d_i$. Note that $\alpha^\vee[i]$ means the same thing as $d[i]$.

The classical *Chevalley rule* [1994] (see also [Fulton and Woodward 2004]) is a formula for the products $\sigma_u \cdot \sigma_{s_i} \in H^*(X)$:

$$\sigma_u \cdot \sigma_{s_i} = \sum_{\alpha} (\alpha^\vee[i]) \sigma_{us_\alpha}, \tag{1}$$

where the sum is over positive roots α such that $\ell(us_\alpha) = \ell(u) + 1$.

The *quantum Chevalley formula* for $\sigma_u \star \sigma_{s_i} = \sum_{w,d} c_{u,s_i}^{w,d} q^d \sigma_w$ was first proved by Fulton and Woodward [2004]. See Theorem 4.3 below. We follow here an approach based on curve neighborhoods, recently proved by Buch and Mihalea [2015]. If $d = (0, 0)$ then the coefficients $c_{u,s_i}^{w,d}$ are those from identity (1) above. If $d \neq (0, 0)$ then the quantum coefficient $c_{u,s_i}^{w,d}$ can be calculated as follows. First, let $w[d] \in W$ be the curve neighborhood $\Gamma_d(w)$. Then

$$c_{u,s_i}^{w,d} = d[i] \cdot \delta_{u,w[d]}, \tag{2}$$

where δ_{v_1,v_2} is the Kronecker symbol and w satisfies $\ell(w) + \deg q^d = \ell(u) + 1$.

Remark 4.1. Although it is not clear from the definition, it turns out that if $\mathbf{d}[i] \neq 0$ then $u = w[\mathbf{d}]$ only if $\mathbf{d} = \alpha^\vee$ for some α such that $\ell(s_\alpha) = \deg q^{\alpha^\vee} - 1$. This recovers the original quantum Chevalley rule from [Fulton and Woodward 2004].

Example 4.2. Consider $\sigma_{s_1} \star \sigma_{s_1}$.

- Assume $\mathbf{d} = (0, 0)$. We need to determine roots α such that $\ell(s_1 s_\alpha) = \ell(s_1) + 1 = 2$. The only possible Weyl group elements to represent s_α are s_2 and $s_1 s_2 s_1$. If $s_\alpha = s_2$ then $\alpha = \alpha_2$. This implies $\alpha^\vee = (0, 1)$, so $\alpha^\vee[1] = 0$. If $s_\alpha = s_1 s_2 s_1$ then $\alpha = 3\alpha_1 + \alpha_2$. This implies $\alpha^\vee = (1, 1)$, so $\alpha^\vee[1] = 1$. Thus

$$\sum_{\alpha} (\alpha^\vee[i]) \sigma_{us_\alpha} = 0 \cdot \sigma_{s_1 s_2} + 1 \cdot \sigma_{s_1 s_1 s_2 s_1} = \sigma_{s_2 s_1}.$$

- Assume $\mathbf{d} \neq (0, 0)$. We need to determine $w \in W$ such that $w[\mathbf{d}] = \Gamma_{\mathbf{d}}(w) = s_1$. According to the curve neighborhood results table in the Appendix, the only possible $w \in W$ are id and s_1 . For both elements, the possible nondegrees are $(N, 0)$, where $N \in \mathbb{N}$. Note that we also need to choose w and \mathbf{d} such that $\ell(w) + \deg q^{\mathbf{d}} = \ell(s_1) + 1 = 2$. Since $\deg q^{\mathbf{d}}$ is never odd, $\ell(w)$ must be even. This eliminates s_1 . As for id , $\ell(\text{id}) = 0$ so then $\deg q^{\mathbf{d}} = 2$, where $\mathbf{d} = (N, 0)$. This implies $N = 1$. Therefore $c_{s_1, s_1}^{\text{id}, (1, 0)} = \mathbf{d}[1] \cdot \delta_{s_1, s_1} = 1 \cdot 1 = 1$ and this represents the only nonzero quantum term. Thus for $\mathbf{d} \neq (0, 0)$,

$$\sum_{w \in W, \mathbf{d}} c_{u, s_i}^{w, \mathbf{d}} q^{\mathbf{d}} \sigma_w = 1 \cdot q^{(1, 0)} \cdot \sigma_{\text{id}} = 1 \cdot q_1 \cdot 1 = q_1.$$

Combining the classical (i.e., from $H^*(X)$) and pure quantum terms gives us $\sigma_{s_1} \star \sigma_{s_1} = \sigma_{s_2 s_1} + q_1$.

Table 4 shows the results of our quantum Chevalley computations.

Theorem 4.3 (the quantum Chevalley rule [Fulton and Woodward 2004; Buch and Mihalcea 2015]). *The following holds in $\text{QH}^*(X)$:*

$$\sigma_u \star \sigma_{s_i} = \sum_{\alpha} (\alpha^\vee[i]) \sigma_{us_\alpha} + \sum_{\beta} (\beta^\vee[i]) q^{\beta^\vee} \sigma_{us_\beta}. \tag{3}$$

The first sum is over positive roots α such that $\ell(us_\alpha) = \ell(u) + 1$ and the second sum is over positive roots β such that $\ell(us_\beta) = \ell(u) + 1 - \deg(q^{\beta^\vee})$.

5. Quantum Schubert polynomials

We know that $\text{QH}^*(X)$ is generated as a $\mathbb{Q}[q] = \mathbb{Q}[q_1, q_2]$ -algebra by the classes σ_{s_1} and σ_{s_2} . (This means that every element in $\text{QH}^*(X)$ can be written as a sum of monomials in the σ_{s_i} with coefficients in $\mathbb{Q}[q]$.) Then there exists a *surjective* homomorphism of $\mathbb{Q}[q]$ -algebras $\Psi : \mathbb{Q}[x_1, x_2; q_1, q_2] \rightarrow \text{QH}^*(X)$ sending

$$\Psi(q_i) = q_i, \quad \Psi(x_1) = \sigma_{s_1}, \quad \Psi(x_1 + x_2) = \sigma_{s_2}.$$

w	$\sigma_w \star \sigma_{s_1}$	$\sigma_w \star \sigma_{s_2}$
s_1	$\sigma_{s_2 s_1} + q_1$	$\sigma_{s_1 s_2} + \sigma_{s_2 s_1}$
s_2	$\sigma_{s_1 s_2} + \sigma_{s_2 s_1}$	$3\sigma_{s_1 s_2} + q_2$
$s_1 s_2$	$\sigma_{s_1 s_2 s_1} + \sigma_{s_2 s_1 s_2}$	$2\sigma_{s_2 s_1 s_2} + q_2 \sigma_{s_1}$
$s_2 s_1$	$2\sigma_{s_1 s_2 s_1} + q_1 \sigma_{s_2}$	$3\sigma_{s_1 s_2 s_1} + \sigma_{s_2 s_1 s_2}$
$s_1 s_2 s_1$	$\sigma_{s_2 s_1 s_2 s_1} + q_1 \sigma_{s_1 s_2} + q_1 q_2$	$\sigma_{s_1 s_2 s_1 s_2} + 2\sigma_{s_2 s_1 s_2 s_1} + q_1 q_2$
$s_2 s_1 s_2$	$\sigma_{s_2 s_1 s_2 s_1} + 2\sigma_{s_1 s_2 s_1 s_2}$	$3\sigma_{s_1 s_2 s_1 s_2} + q_2 \sigma_{s_2 s_1}$
$s_1 s_2 s_1 s_2$	$\sigma_{s_1 s_2 s_1 s_2 s_1} + \sigma_{s_2 s_1 s_2 s_1 s_2}$	$\sigma_{s_2 s_1 s_2 s_1 s_2} + q_2 \sigma_{s_1 s_2 s_1}$
$s_2 s_1 s_2 s_1$	$\sigma_{s_1 s_2 s_1 s_2 s_1} + q_1 \sigma_{s_2 s_1 s_2} + q_1 q_2 \sigma_{s_2}$	$\sigma_{s_2 s_1 s_2 s_1 s_2} + 3\sigma_{s_1 s_2 s_1 s_2 s_1} + q_1 q_2 \sigma_{s_2}$
$s_1 s_2 s_1 s_2 s_1$	$q_1 \sigma_{s_1 s_2 s_1 s_2} + q_1 q_2 \sigma_{s_1 s_2}$	$\sigma_{w_0} + q_1 q_2 \sigma_{s_1 s_2}$
$s_2 s_1 s_2 s_1 s_2$	$\sigma_{w_0} + q_1 q_2^2$	$q_2 \sigma_{s_2 s_1 s_2 s_1} + 2q_1 q_2^2$
$w_0 = (s_1 s_2)^3$	$q_1 \sigma_{s_2 s_1 s_2 s_1 s_2} + q_1 q_2 \sigma_{s_2 s_1 s_2} + q_1 q_2^2 \sigma_{s_1}$	$q_2 \sigma_{s_1 s_2 s_1 s_2 s_1} + q_1 q_2 \sigma_{s_2 s_1 s_2} + 2q_1 q_2^2 \sigma_{s_1}$

Table 4. The quantum Chevalley table.

Note that for any $P, P' \in \mathbb{Q}[x_1, x_2, q_1, q_2]$, we have $\Psi(P \cdot P') = \Psi(P) \star \Psi(P')$. We call Ψ the *quantization map*. Let \tilde{I} be the kernel of this homomorphism. By the first isomorphism theorem, we have an isomorphism

$$\bar{\Psi} : \mathbb{Q}[x_1, x_2, q_1, q_2] / \tilde{I} \rightarrow \text{QH}^*(X),$$

and this gives the presentation of the quantum cohomology ring. A *quantum Schubert polynomial* for the Schubert class σ_w is any polynomial $P_w \in \mathbb{Q}[x_1, x_2, q_1, q_2]$ such that the image of P_w under Ψ gives the class σ_w . Equivalently $\bar{\Psi}(P_w + \tilde{I}) = \sigma_w$.

To find a quantum Schubert polynomial P_w , we proceed by induction on $\ell(w)$, using the quantum Chevalley formula from Table 4, and starting from the “initial conditions” $P_{s_1} = x_1$ and $P_{s_2} = x_1 + x_2$. To obtain the corresponding classical Schubert polynomials for cohomology, set $q_1 = q_2 = 0$.

Example 5.1. In order to calculate $P_{s_2 s_1}$, we use the identity $\sigma_{s_1} \star \sigma_{s_1} = \sigma_{s_2 s_1} + q_1$ (taken from Table 4). Using that Ψ is an algebra homomorphism, we know that

$$\Psi(x_1^2) = \Psi(x_1) \star \Psi(x_1) = \sigma_{s_1} \star \sigma_{s_1} \quad \text{and} \quad \Psi(q_1) = q_1.$$

Since $\Psi(x_1^2 - q_1) = \Psi(x_1^2) - \Psi(q_1)$, it follows that $\Psi(x_1^2 - q_1) = \sigma_{s_2 s_1}$. This shows that $x_1^2 - q_1$ is a quantum Schubert polynomial for $\sigma_{s_2 s_1}$. The corresponding ordinary Schubert polynomial is x_1^2 , obtained by making $q_1 = 0$.

Computations of ordinary Schubert polynomials were done for the ordinary cohomology ring $H^*(X)$ of the G_2 flag manifold in a paper by Anderson [2011]. A classical result of Borel [1953] shows that $H^*(X) = \mathbb{Q}[x_1, x_2] / I$, where $I = \langle x_1^2 - x_1 x_2 + x_2^2, x_1^6 \rangle$. (This can also be deduced from the classical Chevalley formula.) Anderson used this presentation and a different method to obtain different Schubert

σ_{s_α}	our calculation	Anderson's calculation [2011]
w_0	$\frac{1}{2}(x_1^6 + x_1^5x_2)$	$\frac{1}{2}x_1^5x_2$
$s_1s_2s_1s_2s_1$	$\frac{1}{2}x_1^5$	$\frac{1}{2}x_1^5$
$s_2s_1s_2s_1s_2$	$\frac{1}{6}(x_1 + x_2)^3x_1x_2$	$\frac{1}{2}(x_1^3 + x_2x_1^2 + x_2^2x_1 + x_2^3)x_1x_2$
$s_2s_1s_2s_1$	$\frac{1}{2}x_1^4$	$\frac{1}{2}(4x_1^2 - 3x_1x_2 + 3x_2^2)x_1^2$
$s_1s_2s_1s_2$	$\frac{1}{6}(x_1 + x_2)^2x_1x_2$	$\frac{1}{2}(x_1^4 + x_1^3x_2 + x_1^2x_2^2 + x_1x_2^3 + x_2^4)$
$s_1s_2s_1$	$\frac{1}{2}x_1^3$	$\frac{1}{2}(4x_1^2 - 3x_1x_1 + 3x_2^2)x_1$
$s_2s_1s_2$	$\frac{1}{2}(x_1 + x_2)x_1x_2$	$2x_1^3 + \frac{1}{2}x_1^2x_2 + \frac{1}{2}x_1x_2^2 + 2x_2^3$
s_2s_1	x_1^2	$3x_1^2 - 2x_1x_2 + 2x_2^2$
s_1s_2	x_1x_2	$2x_1^2 - x_1x_2 + 2x_2^2$
s_2	$x_1 + x_2$	$x_1 + x_2$
s_1	x_1	x_1
id	1	1

Table 5. Classical Schubert polynomials.

polynomials, but our answers and his must be equal modulo the ideal I . The classical Schubert polynomials we found are shown alongside Anderson's in Table 5. In order to check if our results are equal, we verified that the difference between our resulting classical polynomials was a multiple of one of the elements of the ideal.

We used our quantum Schubert polynomial results, found in Theorem 5.2 below, to compute the ideal \tilde{I} of the quantum cohomology ring $\text{QH}^*(X)$. This ideal is a deformation of the ideal I of $H^*(X)$. As an example, we will derive the degree-2 relation in \tilde{I} . From the quantum Chevalley table on page 447, we know the identities

- $\sigma_{s_1} \star \sigma_{s_1} = \sigma_{s_2s_1} + q_1,$
- $\sigma_{s_1} \star \sigma_{s_2} = \sigma_{s_1s_2} + \sigma_{s_2s_1},$ and
- $\sigma_{s_2} \star \sigma_{s_2} = 3\sigma_{s_1s_2} + q_2.$

These three equalities can be combined to obtain

$$3(\sigma_{s_1} \star \sigma_{s_1}) + (\sigma_{s_2} \star \sigma_{s_2}) = 3(\sigma_{s_1} \star \sigma_{s_2}) + 3q_1 + q_2.$$

Now apply the transformation under $\bar{\Psi}$ to get

$$3(x_1 \cdot x_1) + ((x_1 + x_2) \cdot (x_2 + x_2)) \equiv (3(x_1(x_1 + x_2)) + 3q_1 + q_2) + \tilde{I},$$

which is

$$3x_1^2 + x_1^2 + 2x_1x_2 + x_2^2 \equiv (3x_1^2 + 3x_1x_2 + 3q_1 + q_2) + \tilde{I}.$$

Their difference belongs to $\tilde{I} = \ker \Psi$, so (after simplification) we get

$$x_1^2 - x_1x_2 + x_2^2 - (3q_1 + q_2) \in \tilde{I}.$$

This is the degree-2 relation in \tilde{I} . Notice that this is clearly a deformation of the ideal term $x_1^2 - x_1x_2 + x_2^2$ in I . To get the degree-6 relation, one does a similar manipulation but using the higher-degree terms in the quantum Chevalley table on page 447. The following is the main result of this paper.

Theorem 5.2. *The quantum cohomology ring of the flag manifold of type G_2 is*

$$\text{QH}^*(X) = \mathbb{Q}[q_1, q_2, x_1, x_2] / \langle R_2, R_6 \rangle,$$

where $R_2 := x_1^2 - x_1x_2 + x_2^2 - (3q_1 + q_2)$ and

$$R_6 := x_1^6 + q_1(-2x_1^4 - \frac{13}{3}x_2^3x_2 - \frac{5}{3}x_1^2x_2^2 - \frac{1}{3}x_1x_2^3) + q_1^2(-\frac{10}{3}x_1^2 - \frac{5}{3}x_1x_2 - \frac{1}{3}x_2^2) + q_1q_2(-2x_1^2 - \frac{11}{3}x_1x_2) - \frac{8}{3}q_1^2q_2.$$

Under this presentation, the corresponding quantum Schubert polynomials are

$$\begin{aligned} w_0 = (s_1s_2)^3 &: \frac{1}{2}(x_1^6 + x_1^5x_2) + \frac{1}{2}(-2x_1^4 - 6x_1^3x_2 - 5x_1^2x_2^2 - x_1x_2^3)q_1, \\ &\quad + \frac{1}{2}(-3x_1^2 - 7x_1x_2 - 2x_2^2)q_1q_2 + \frac{1}{2}(-3x_1^2 - 4x_1x_2 - x_2^2)q_1^2 - q_1^2q_2, \\ s_1s_1s_2s_1s_2 &: \frac{1}{6}((x_1+x_2)^3x_1x_2) + \frac{1}{6}((x_1+x_2)^3q_1, \\ &\quad + (-6x_1^3 - 4x_1^2x_2 - x_1x_2^2)q_2 + (8x_1 + 5x_2)q_1q_2), \\ s_1s_2s_1s_2s_1 &: \frac{1}{2}x_1^5 + \frac{1}{2}((-2x_1^3 - 4x_1^2x_2 - x_1x_2^2)q_1 + (-3x_1 - 2x_2)q_1q_2 + (-3x_1 - x_2)q_1^2), \\ s_1s_2s_1s_2 &: \frac{1}{6}((x_1+x_2)^2x_1x_2) + \frac{1}{6}((x_1+x_2)^2q_1 + (-3x_1^2 - x_1x_2)q_2 + 2q_1q_2), \\ s_2s_1s_2s_1 &: \frac{1}{2}x_1^4 + \frac{1}{2}((-2x_1^2 - 3x_1x_2)q_1 - 2q_1q_2 - 2q_1^2), \\ s_2s_1s_2 &: \frac{1}{2}((x_1+x_2)x_1x_2) + \frac{1}{2}((x_1+x_2)q_1 - x_1q_2), \\ s_1s_2s_1 &: \frac{1}{2}x_1^3 + \frac{1}{2}((-2x_1 - x_2)q_1), \\ s_1s_2 &: x_1x_2 + q_1, \\ s_2s_1 &: x_1^2 - q_1, \\ s_2 &: x_1 + x_2, \\ s_1 &: x_1, \\ \text{id} &: 1. \end{aligned}$$

Appendix: Table of curve neighborhood calculations

This appendix contains the curve neighborhoods for all the Weyl group elements. In order to list them as concisely as possible, we need to define

- $\ell, m = 0, 1, 2, 3, \dots$,
- $N, M = 1, 2, 3, \dots$,
- $N', M' = 2, 3, 4, \dots$

If $w \in W$ then $\Gamma_{(0,0)}(w) = w$, so we won't include that condition in the table.

id	s_1	s_2
$\Gamma_{(N,0)}(\text{id}) = s_1$	$\Gamma_{(N,0)}(s_1) = s_1$	$\Gamma_{(N,0)}(s_2) = s_2 s_1$
$\Gamma_{(0,N)}(\text{id}) = s_2$	$\Gamma_{(0,N)}(s_1) = s_1 s_2$	$\Gamma_{(0,N)}(s_2) = s_2$
$\Gamma_{(N,1)}(\text{id}) = s_1 s_2 s_1$	$\Gamma_{(N,1)}(s_1) = s_1 s_2 s_1$	$\Gamma_{(N,1)}(s_2) = s_2 s_1 s_2 s_1$
$\Gamma_{(1,N')}(\text{id}) = s_2 s_1 s_2 s_1 s_2$	$\Gamma_{(N,N')}(s_1) = w_0$	$\Gamma_{(1,N')}(s_2) = s_2 s_1 s_2 s_1 s_2$
$\Gamma_{(N',M')}(\text{id}) = w_0$		$\Gamma_{(N',M')}(s_2) = w_0$
$s_1 s_2$	$s_2 s_1$	$s_1 s_2 s_1$
$\Gamma_{(N,0)}(s_1 s_2) = s_1 s_2 s_1$	$\Gamma_{(N,0)}(s_2 s_1) = s_2 s_1$	$\Gamma_{(N,0)}(s_1 s_2 s_1) = s_1 s_2 s_1$
$\Gamma_{(0,N)}(s_1 s_2) = s_1 s_2$	$\Gamma_{(0,N)}(s_2 s_1) = s_2 s_1 s_2$	$\Gamma_{(0,N)}(s_1 s_2 s_1) = s_1 s_2 s_1 s_2$
$\Gamma_{(N,1)}(s_1 s_2) = s_1 s_2 s_1 s_2 s_1$	$\Gamma_{(N,1)}(s_2 s_1) = s_2 s_1 s_2 s_1$	$\Gamma_{(N,1)}(s_1 s_2 s_1) = s_1 s_2 s_1 s_2 s_1$
$\Gamma_{(N,N')}(s_1 s_2) = w_0$	$\Gamma_{(N,N')}(s_2 s_1) = w_0$	$\Gamma_{(N,N')}(s_1 s_2 s_1) = w_0$
$s_2 s_1 s_2$	$s_1 s_2 s_1 s_2$	$s_2 s_1 s_2 s_1$
$\Gamma_{(N,0)}(s_2 s_1 s_2) = s_2 s_1 s_2 s_1$	$\Gamma_{(N,0)}(s_1 s_2 s_1 s_2) = s_1 s_2 s_1 s_2 s_1$	$\Gamma_{(N,0)}(s_2 s_1 s_2 s_1) = s_2 s_1 s_2 s_1$
$\Gamma_{(0,N)}(s_2 s_1 s_2) = s_2 s_1 s_2$	$\Gamma_{(0,N)}(s_1 s_2 s_1 s_2) = s_1 s_2 s_1 s_2$	$\Gamma_{(0,N)}(s_2 s_1 s_2 s_1) = s_2 s_1 s_2 s_1 s_2$
$\Gamma_{(N,M)}(s_2 s_1 s_2) = w_0$	$\Gamma_{(N,M)}(s_1 s_2 s_1 s_2) = w_0$	$\Gamma_{(N,M)}(s_2 s_1 s_2 s_1) = w_0$
$s_1 s_2 s_1 s_2 s_1$	$s_2 s_1 s_2 s_1 s_2$	w_0
$\Gamma_{(N,0)}(s_1 s_2 s_1 s_2 s_1) = s_1 s_2 s_1 s_2 s_1$	$\Gamma_{(0,N)}(s_2 s_1 s_2 s_1 s_2) = s_2 s_1 s_2 s_1 s_2$	$\Gamma_{(\ell,m)}(w_0) = w_0$
$\Gamma_{(\ell,N)}(s_1 s_2 s_1 s_2 s_1) = w_0$	$\Gamma_{(N,\ell)}(s_2 s_1 s_2 s_1 s_2) = w_0$	

Table 6. The curve neighborhoods for every degree at every $w \in W$.

Acknowledgement

This work is part of an undergraduate research project conducted under the guidance of Prof. Leonardo C. Mihalcea.

References

[Anderson 2011] D. Anderson, “Chern class formulas for G_2 Schubert loci”, *Trans. Amer. Math. Soc.* **363**:12 (2011), 6615–6646. MR 2833570 Zbl 1234.14037

[Borel 1953] A. Borel, “Sur la cohomologie des espaces fibrés principaux et des espaces homogènes de groupes de Lie compacts”, *Ann. of Math. (2)* **57** (1953), 115–207. MR 0051508 Zbl 0052.40001

[Bourbaki 2002] N. Bourbaki, *Lie groups and Lie algebras: Chapters 4–6*, Springer, Berlin, 2002. MR 1890629 Zbl 0983.17001

[Buch and Mihalcea 2015] A. S. Buch and L. C. Mihalcea, “Curve neighborhoods of Schubert varieties”, *J. Differential Geom.* **99**:2 (2015), 255–283. MR 3302040 Zbl 06423472

[Chevalley 1994] C. Chevalley, “Sur les décompositions cellulaires des espaces G/B ”, pp. 1–23 in *Algebraic groups and their generalizations: classical methods* (University Park, PA, 1991), edited by W. J. Haboush and B. J. Parshall, Proc. Sympos. Pure Math. **56**, Amer. Math. Soc., Providence, RI, 1994. MR 1278698 Zbl 0824.14042

[Fulton and Pandharipande 1997] W. Fulton and R. Pandharipande, “Notes on stable maps and quantum cohomology”, pp. 45–96 in *Algebraic geometry* (Santa Cruz, CA 1995), edited by J.

Kollár et al., Proc. Sympos. Pure Math. **62**, Amer. Math. Soc., Providence, RI, 1997. MR 1492534 Zbl 0898.14018

[Fulton and Woodward 2004] W. Fulton and C. Woodward, “On the quantum product of Schubert classes”, *J. Algebraic Geom.* **13**:4 (2004), 641–661. MR 2072765 Zbl 1081.14076

[Humphreys 1972] J. E. Humphreys, *Introduction to Lie algebras and representation theory*, Graduate Texts in Mathematics **9**, Springer, New York, 1972. MR 0323842 Zbl 0254.17004

[Kim 1999] B. Kim, “Quantum cohomology of flag manifolds G/B and quantum Toda lattices”, *Ann. of Math. (2)* **149**:1 (1999), 129–148. MR 1680543 Zbl 1054.14533

[Knutson and Miller 2005] A. Knutson and E. Miller, “Gröbner geometry of Schubert polynomials”, *Ann. of Math. (2)* **161**:3 (2005), 1245–1318. MR 2180402 Zbl 1089.14007

Received: 2015-02-18 Accepted: 2015-05-29

rae1313@vt.edu

*Department of Physics, Virginia Tech University,
306 Robeson Hall, Blacksburg, VA 24061, United States*

mel3jp@virginia.edu

*Department of Mathematics, University of Virginia, 141 Cabell
Drive, Kerchof Hall, Charlottesville, VA 22904, United States*

lmihalce@math.vt.edu

*Department of Mathematics, Virginia Tech University,
460 McBryde Hall, Blacksburg, VA 24060, United States*

The irreducibility of polynomials related to a question of Schur

Lenny Jones and Alicia Lamarche

(Communicated by Kenneth S. Berenhaut)

In 1908, Schur raised the question of the irreducibility over \mathbb{Q} of polynomials of the form $f(x) = (x + a_1)(x + a_2) \cdots (x + a_m) + c$, where the a_i are distinct integers and $c \in \{-1, 1\}$. Since then, many authors have addressed variations and generalizations of this question. In this article, we investigate the irreducibility of $f(x)$ and $f(x^2)$, where the integers a_i are consecutive terms of an arithmetic progression and c is a nonzero integer.

1. Introduction

Throughout this paper, unless indicated otherwise, “reducible polynomial” and “irreducible polynomial” pertain to reducibility and irreducibility over \mathbb{Q} . Schur [1908] raised the question of the irreducibility of polynomials of the form

$$g_{\pm}(x) = (x + a_1)(x + a_2) \cdots (x + a_m) \pm 1,$$

where the a_i are distinct integers. Westlund [1909] showed that $g_{-}(x)$ is always irreducible, and that if $g_{+}(x)$ is reducible, then $g_{+}(x)$ must be the square of a polynomial. Flügel [1909] showed that $g_{+}(x)$ is reducible if and only if there exists an integer z such that

$$g_{+}(x + z) = (x - 1)^2 \quad \text{or} \quad g_{+}(x + z) = (x^2 - 3x + 1)^2.$$

Since that time, numerous authors [Seres 1956; Győry et al. 2011] have addressed variations and generalizations of these questions. For some more recent generalizations, and a complete history and bibliography chronicling these results, see [Győry et al. 2011].

Here we investigate the irreducibility of polynomials $f(x)$ and $f(x^2)$, where

$$f(x) = (x + a_i)(x + a_{i+1}) \cdots (x + a_{i+m-1}) + c, \quad (1-1)$$

MSC2010: 12E05, 11C08.

Keywords: irreducible polynomial.

with the a_j being consecutive terms of an arithmetic progression

$$\mathcal{A} = \{k, k + d, k + 2d, \dots\},$$

where $d > 0$ is the common difference. Since

$$f(x) = (x + k + jd)(x + k + (j + 1)d) \cdots (x + k + (j + m - 1)d) + c$$

is irreducible if and only if

$$f(x) = x(x + d)(x + 2d) \cdots (x + (m - 1)d) + c \tag{1-2}$$

is irreducible, our focus here is on (1-2). While we are placing restrictions on the values of a_i in (1-1), the fact that we are not initially placing any restrictions on c , other than $c \neq 0$, and that we are also concerned with the irreducibility of $f(x^2)$, make this investigation somewhat of a departure from previous ones. In particular, defining $F(x) := f(x^2)$, where $f(x)$ is as in (1-2), we are interested in determining values of d, m and c , with $d > 0$ and $m \geq 2$, for which

- (I) $f(x)$ is reducible,
- (II) $f(x)$ is irreducible, but $F(x)$ is reducible,
- (III) both $f(x)$ and $F(x)$ are irreducible.

Note that if $F(x)$ is irreducible, then $f(x)$ is irreducible. However, the converse is false in general, as the example $f(x) = x - 1$ illustrates, so that situation (II) is not, in general, vacuous. Clearly, a complete answer to (I) and (II), or to (I) and (III), provides an answer to (III), or to (II), respectively. Although a complete answer to (I) seems intractable, a reasonable approach seems to be to place restrictions on one or more of d, m and c . For example, one could place a bound on m and determine the appropriate values of d and c such that $f(x)$ satisfies (I), (II) or (III). This is the strategy we employ in Section 3. However, in this scenario, even small values of m prove to be challenging. In Section 4, by imposing different restrictions on d, m and c , we can establish the following theorem for larger degree polynomials:

Theorem 1.1. *Let $p \geq 3$ be prime, and let $c, d \in \mathbb{Z}$, with $c \neq 0, d > 0$ and $d \not\equiv 0 \pmod{p}$. Let*

$$\begin{aligned} f(x) &= x(x + d)(x + 2d) \cdots (x + (p - 1)d) + c \\ &= x^p + a_{p-1}x^{p-1} + \cdots + a_1x + c. \end{aligned}$$

(1) *If $c \not\equiv 0 \pmod{p}$, then $f(x)$ is irreducible. If, in addition, $c \neq -z^2$ for any $z \in \mathbb{Z}$, then $F(x)$ is irreducible.*

(2) *Let k be a fixed positive integer, and suppose that $|c| = kp^w$, where*

$$p^w > k^{p-1} + a_{p-1}k^{p-2} + \cdots + a_2k + a_1.$$

Then both $f(x)$ and $F(x)$ are irreducible if one of the sets of conditions below holds:

- (a) $c > 0$.
- (b) $c < 0$, $w \equiv 1 \pmod{2}$ and $k \not\equiv 0 \pmod{p}$.
- (c) $c < 0$ and $p \equiv 3 \pmod{4}$.

Computations in this article were performed using either Maple or Magma.

2. Preliminaries

We now present, without proof, some facts that are used to establish the results in this article. The first two theorems for general fields k first appeared in [Schinzel 1982]. For fields $k \subset \mathbb{C}$, they are originally due to Capelli [Schinzel 2000].

Theorem 2.1. *Let k be a field, and let $f(x)$ and $g(x)$ be polynomials in $k[x]$ with $f(x)$ irreducible over k . Suppose that $f(\alpha) = 0$. Then $f(g(x))$ is reducible over k if and only if $g(x) - \alpha$ is reducible over $k(\alpha)$. Furthermore, if*

$$g(x) - \alpha = c_1 u_1(x)^{e_1} \cdots u_r(x)^{e_r},$$

where $c_1 \in k(\alpha)$ and the $u_j(x)$ are distinct monic irreducible polynomials in $k(\alpha)[x]$, then

$$f(g(x)) = c_2 \mathcal{N}(u_1(x))^{e_1} \cdots \mathcal{N}(u_r(x))^{e_r},$$

where $c_2 \in k$, and the norms $\mathcal{N}(u_j(x))$ are distinct monic irreducible polynomials in $k[x]$.

Theorem 2.2. *Let k be a field, and let $r \in \mathbb{Z}$ with $r \geq 2$. Let $\alpha \in k$. Then $x^r - \alpha$ is reducible over k if and only if either $\alpha = \beta^p$ for some prime divisor p of r and $\beta \in k$, or $4 \mid r$ and $\alpha = -4\beta^4$ for some $\beta \in k$.*

The next result follows from direct applications of Theorem 2.1 with $g(x) = x^2$ and Theorem 2.2 with $r = 2$, and equating constant terms.

Theorem 2.3. *Let*

$$f(x) = x^n + \sum_{j=1}^{n-1} a_j x^j + c \in \mathbb{Z}[x],$$

with $f(x)$ irreducible. Then:

- (1) If $n \equiv 0 \pmod{2}$ and $c \neq z^2$ for any $z \in \mathbb{Z}$, then $F(x)$ is irreducible.
- (2) If $n \equiv 1 \pmod{2}$ and $c \neq -z^2$ for any $z \in \mathbb{Z}$, then $F(x)$ is irreducible.

The following result is well-known [Serret 1992].

Theorem 2.4. *Let p be a prime, and let $f(x) = x^p - x + c \in \mathbb{F}_p[x]$. If $c \not\equiv 0 \pmod{p}$, then $f(x)$ is irreducible over \mathbb{F}_p .*

Since the irreducibility of a polynomial over \mathbb{F}_p implies its irreducibility over \mathbb{Q} , we immediately have the following corollary.

Corollary 2.5. *Let $f(x) \in \mathbb{Z}[x]$, and let p be a prime. If $f(x) \equiv x^p - x + c \pmod{p}$ and $c \not\equiv 0 \pmod{p}$, then $f(x)$ is irreducible.*

The next theorem and its corollary are special cases of results of Weisner [1934].

Theorem 2.6. *Let*

$$A(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0 \in \mathbb{Z}[x]$$

be such that $n \geq 2$, $a_n \neq 0$ and

$$|a_0| = kp^w, \quad \text{with } k, w \geq 1,$$

where p is a prime that does not divide a_1 if $w > 1$. Suppose further that there exists L such that $|r| \geq L \geq 1$ for all zeros r of $A(x)$. If $k < L$, then $A(x)$ is irreducible.

Corollary 2.7. *Let $k, w \geq 1$ and $n \geq 2$ be integers, and let*

$$A_{\pm}(x) = x^n + a_{n-1} x^{n-1} + \dots + a_1 x \pm kp^w \in \mathbb{Z}[x],$$

where p is a prime that does not divide a_1 . If

$$p^w > k^{n-1} + |a_{n-1}|k^{n-2} + \dots + |a_2|k + |a_1|,$$

then each of $A_{\pm}(x)$ is irreducible.

3. A first approach

In this section, we investigate an approach to determine the values of c such that each of the conditions (I), (II) and (III) holds. The idea is to analyze the degree-type factorization of $f(x)$. The following proposition, whose proof is immediate from the definition of $f(x)$ in (1-2), represents a modest step in this direction.

Proposition 3.1. *The polynomial $f(x)$ has a zero $n \in \mathbb{Z}$ if and only if*

$$c = -n(n+d)(n+2d) \cdots (n+(m-1)d) \quad \text{for some } n \in \mathbb{Z}.$$

One difficulty in establishing a more general result similar to Proposition 3.1 is that the number of possible degree-type factorizations of $f(x)$ into irreducibles increases as m increases. To avoid this complication, we bound the value of m . However, even for small values of m , such a method proves to be challenging. To illustrate the difficulties that arise, we address the cases (I) and (II) for each value of $m \in \{2, 3, 4\}$. For case (I), we use the straightforward method of equating coefficients. Our investigation of case (II) also uses the method of equating coefficients, but we additionally utilize Theorem 2.1 and Theorem 2.2 with $g(x) = x^2$. Although the techniques are similar, each value of m presents distinct obstacles.

The case of $m = 2$.

Theorem 3.2. *Let $c, d \in \mathbb{Z}$, with $d > 0$, and let*

$$f(x) = x(x + d) + c.$$

Then

(1) *$f(x)$ is reducible if and only if*

$$c \in \{-n(n + d) \mid n \in \mathbb{Z}\},$$

(2) *$f(x)$ is irreducible and $F(x)$ is reducible if and only if*

$$c \in \left\{ \frac{1}{4}(s^2 + d)^2 \mid s \in \mathbb{Z}, \text{ with } s > 0 \text{ and } s^2 \equiv d \pmod{2} \right\}.$$

Proof. Observe that (1) follows immediately from Proposition 3.1. To prove (2), suppose first that $f(x)$ is irreducible and $f(\alpha) = 0$. Suppose that $\alpha = \beta^2$ for some $\beta \in \mathbb{Q}(\alpha)$. Then, by Theorem 2.1,

$$F(x) = x^4 + dx^2 + c \tag{3-1}$$

$$\begin{aligned} &= \mathcal{N}(x + \beta)\mathcal{N}(x - \beta) \\ &= (x^2 + (\beta + \bar{\beta})x + \beta\bar{\beta})(x^2 - (\beta + \bar{\beta})x + \beta\bar{\beta}) \\ &= (x^2 + sx + t)(x^2 - sx + t) \\ &= x^4 + (2t - s^2)x^2 + t^2 \end{aligned} \tag{3-2}$$

for some $s, t \in \mathbb{Z}$. Equating coefficients in (3-1) and (3-2) and solving the resulting system of equations gives $c = \frac{1}{4}(s^2 + d)^2$.

There are two items of concern here. The first item to address is whether there are any restrictions that must be placed on s to guarantee that $c \in \mathbb{Z}$. The second item is whether there are any values of s such that $f(x)$ is reducible. Clearly, we can assume that $s \geq 0$ in any case, and imposing the restriction that $c \in \mathbb{Z}$ tells us that $s^2 \equiv d \pmod{2}$. We must now check if there are any such values of c such that $f(x)$ is reducible. That is, are there values of $s, n \in \mathbb{Z}$ such that

$$\frac{1}{4}(s^2 + d)^2 = -n(n + d)? \tag{3-3}$$

Solving (3-3), we get the single integer solution $s = 0$ and $n = -\frac{1}{2}d$, where $d \equiv 0 \pmod{2}$, which corresponds to $c = \frac{1}{4}d^2$. Hence, we must have $s > 0$ to ensure that $f(x)$ is irreducible. Under these restrictions on c , we have conversely that $f(x)$ is irreducible and that

$$\begin{aligned} F(x) &= x^4 + dx^2 + \frac{1}{4}(s^2 + d)^2 \\ &= (x^2 + sx + \frac{1}{2}(s^2 + d))(x^2 - sx + \frac{1}{2}(s^2 + d)). \end{aligned} \quad \square$$

The case of $m = 3$.

Theorem 3.3. *Let $c, d \in \mathbb{Z}$, with $d > 0$, and let*

$$f(x) = x(x + d)(x + 2d) + c.$$

Then

(1) *$f(x)$ is reducible if and only if*

$$c \in R = \{-n(n + d)(n + 2d) \mid n \in \mathbb{Z}\},$$

(2) *$f(x)$ is irreducible and $F(x)$ is reducible if and only if $c \in S \setminus R$, where*

$$S = \left\{ -\left(\frac{s^4 + 6ds^2 + d^2}{8s} \right)^2 \mid \text{all of the conditions in } A \text{ hold} \right\},$$

and A is the following list:

$$\begin{aligned} & d \not\equiv 2, 3 \pmod{4}, \quad s \in \mathbb{Z}^+, \quad \frac{d^2}{s} \in \mathbb{Z}^+, \\ & s \equiv 0 \pmod{2} \text{ and } \frac{d^2}{8s} \in \mathbb{Z}^+ \text{ if } d \equiv 0 \pmod{4}, \\ & s \equiv 1 \pmod{2} \text{ if } d \equiv 1 \pmod{4}. \end{aligned}$$

Moreover, S contains at most finitely many elements for a fixed value of d .

Proof. As in the case of $m = 2$, observe that (1) follows immediately from Proposition 3.1. To establish (2), we proceed as in Theorem 3.2. We assume that $f(x)$ is irreducible and $f(\alpha) = 0$. Suppose also that $\alpha = \beta^2$ for some $\beta \in \mathbb{Q}(\alpha)$. Then, by Theorem 2.1, we have

$$F(x) = x^6 + 3dx^4 + 2d^2x^2 + c \tag{3-4}$$

$$\begin{aligned} &= (x^3 + sx^2 + tx + u)(x^3 - sx^2 + tx - u) \\ &= x^6 + (2t - s^2)x^4 + (t^2 - 2su)x^2 - u^2 \end{aligned} \tag{3-5}$$

for some $s, t, u \in \mathbb{Z}$. Equating coefficients in (3-4) and (3-5) and solving the resulting system of equations, with $d > 0$, gives

$$c = -\left(\frac{s^4 + 6ds^2 + d^2}{8s} \right)^2,$$

where we can assume that $s > 0$. Since $c \in \mathbb{Z}$, it is necessary that $d^2 \equiv 0 \pmod{s}$. This restriction alone implies that there are at most finitely many such values of c for a fixed d , and therefore all such values of c in S can be effectively computed. Further analysis reveals that $d \not\equiv 2, 3 \pmod{4}$, since $s^4 + 6ds^2 + d^2 \equiv 0 \pmod{8}$. Additionally, we see that $s \equiv 0 \pmod{2}$ when $d \equiv 0 \pmod{4}$, and in this case we get the more restrictive condition that $d^2 \equiv 0 \pmod{8s}$. Finally, $s \equiv 1 \pmod{2}$ when $d \equiv 1 \pmod{4}$.

Conversely, if $c \in S \setminus R$, then $f(x)$ is irreducible and

$$F(x) = x^6 + 3dx^4 + 2d^2x^2 - \left(\frac{s^4 + 6ds^2 + d^2}{8s}\right)^2 = F_1(x)F_2(x),$$

where

$$F_1(x) = x^3 + sx^2 + \frac{1}{2}(s^2 + 3d)x + \frac{s^4 + 6ds^2 + d^2}{8s} \in \mathbb{Z}[x]$$

and

$$F_2(x) = x^3 - sx^2 + \frac{1}{2}(s^2 + 3d)x - \frac{s^4 + 6ds^2 + d^2}{8s} \in \mathbb{Z}[x]. \quad \square$$

As in the proof of Theorem 3.2, a somewhat more explicit description of the values of c such that (II) holds would be desirable. To determine whether any values of $c \in S$ from (2) are such that $f(x)$ is reducible when $d \equiv 0, 1 \pmod{4}$, we must solve the Diophantine equation

$$\left(\frac{s^4 + 6ds^2 + d^2}{8s}\right)^2 = n(n + d)(n + 2d). \tag{3-6}$$

Again, because of the restriction on s for a given value of d , the solutions to (3-6) can be effectively computed. We conjecture that there are no solutions to (3-6) for any value of d , so that $S \cap R = \emptyset$.

Remark 3.4. Any solutions to (3-6) are integral solutions of the so-called ‘‘congruent-number’’ elliptic curve $y^2 = x(x^2 - d^2)$, which has been studied extensively [Bremner et al. 2000; Koblitz 1993; Silverman 2009].

The case of $m = 4$.

Theorem 3.5. *Let $c, d \in \mathbb{Z}$, with $d > 0$, and let*

$$f(x) = x(x + d)(x + 2d)(x + 3d) + c.$$

Then

- (1) $f(x)$ is reducible if and only if $c \in R = R_1 \cup R_2$, where

$$R_1 = \{v(2d^2 - v) \mid v \in \mathbb{Z}\},$$

$$R_2 = \{\frac{1}{4}(u - d)(u - 2d)(u - 4d)(u - 5d) \in \mathbb{Z} \mid u \in \mathbb{Z}\},$$

- (2) $f(x)$ is irreducible and $F(x)$ is reducible if and only if $c \in S \setminus R$, where

$$S = \left\{ \left(\frac{u^2 + 6d^3}{2t}\right)^2 \in \mathbb{Z} \mid \text{all of the conditions in } B \text{ hold} \right\},$$

and B is the following list:

$$u, t \in \mathbb{Z}^+, \quad t = \frac{1}{2}(s^2 + 6d) \text{ for some } s \in \mathbb{Z},$$

$$8u^2 + (-8s^3 - 48sd)u + s^6 + 18s^4 + 64s^2d^2 = 0.$$

Moreover, S contains at most finitely many elements.

Proof. Logically, since

$$f(x) = x^4 + 6dx^3 + 11d^2x^2 + 6d^3x + c \quad (3-7)$$

is a fourth-degree polynomial, there are five possibilities that could occur when factoring $f(x)$ into irreducibles:

- (1) $f(x)$ is irreducible.
- (2) $f(x)$ is the product of a linear factor and an irreducible cubic.
- (3) $f(x)$ is the product of two linear factors and an irreducible quadratic.
- (4) $f(x)$ is the product of two irreducible quadratics.
- (5) $f(x)$ is the product of four linear factors.

Proposition 3.1 gives us conditions under which $f(x)$ has a linear factor, but it is not delicate enough alone to distinguish among possibilities (2), (3) and (5). In fact, it turns out that (2) is vacuous. To see this, first note that if $f(r) = 0$ for some $r \in \mathbb{Z}$, then

$$\begin{aligned} f(-r - 3d) &= (-r - 3d)(-r - 2d)(-r - d)(-r) + c \\ &= r(r + d)(r + 2d)(r + 3d) + c \\ &= f(r) = 0. \end{aligned}$$

If $r \neq -r - 3d$, then $f(x)$ has at least two distinct linear factors. If $r = -r - 3d$, then $4r^3 + 18dr^2 + 22d^2r + 6d^3 = f'(r) = f'(-r - 3d) = -4r^3 - 18dr^2 - 22d^2r - 6d^3$, so that $f'(r) = 0$. Hence, $(x - r)^2$ divides $f(x)$, and therefore (2) does not occur. Thus, to determine exactly the values of c for which $f(x)$ is reducible, we proceed as follows. Assuming $f(x)$ is reducible, we write

$$\begin{aligned} f(x) &= (x^2 + sx + t)(x^2 + ux + v) \\ &= x^4 + (s + u)x^3 + (t + su + v)x^2 + (tu + sv)x + tv. \end{aligned} \quad (3-8)$$

Solving the system of equations that results by equating coefficients in (3-7) and (3-8), we arrive at the two solutions for c ,

$$c = v(2d^2 - v) \quad \text{and} \quad c = \frac{1}{4}(u - d)(u - 2d)(u - 4d)(u - 5d),$$

where if $c = v(2d^2 - v)$, then

$$f(x) = (x^2 + 3dx + (2d^2 - v))(x^2 + 3dx + v),$$

and if $c = \frac{1}{4}(u - d)(u - 2d)(u - 4d)(u - 5d) \in \mathbb{Z}$, then

$$f(x) = (x^2 + (6d - u)x + \frac{1}{2}(u - 5d)(u - 4d))(x^2 + ux + \frac{1}{2}(u - 2d)(u - d)).$$

We note that the infinite sets R_1 and R_2 are not disjoint, and further analysis is required to determine the particular degree-types given in (3), (4) and (5).

We turn now to an examination of when

$$F(x) = x^8 + 6dx^6 + 11d^2x^4 + 6d^3x^2 + c \tag{3-9}$$

is reducible, assuming that $f(x)$ is irreducible. As before, we have from Theorem 2.1 and Theorem 2.2 that

$$\begin{aligned} F(x) &= (x^4 + sx^3 + tx^2 + ux + v)(x^4 - sx^3 + tx^2 - ux + v) \\ &= x^8 + (2t - s^2)x^6 + (t^2 - 2us + 2v)x^4 + (2vt - u^2)x^2 + v^2. \end{aligned} \tag{3-10}$$

Equating coefficients in (3-9) and (3-10), and solving the resulting system of equations yields

$$\begin{aligned} v &= \frac{u^2 + 6d^3}{2t}, \quad t = \frac{1}{2}(s^2 + 6d), \\ 8u^2 - (8s^3 + 48sd)u + s^6 + 18s^4d + 64s^2d^2 &= 0. \end{aligned} \tag{3-11}$$

Note that if $s = 0$ in (3-11), then $u = 0$ and $c = d^4$, so that $f(x) = (x^2 + 3dx + d^2)^2$. Viewing the third equation in (3-11) as a quadratic equation in the variable u , and solving gives

$$u = \frac{1}{4}(2s^3 + 12ds \pm s\sqrt{2s^4 + 12ds^2 + 16d^2}). \tag{3-12}$$

From (3-12), we see that a necessary condition for u to be an integer is that $2s^4 + 12ds^2 + 16d^2$ be a square. To determine when this occurs, we think of s as a variable and we seek nontrivial ($s \neq 0$) integral solutions to the elliptic curve

$$y^2 = 2s^4 + 12ds^2 + 16d^2 = 2(s^2 + 2d)(s^2 + 4d). \tag{3-13}$$

For a given value of d , it is well known that there are at most finitely many nontrivial solutions to (3-13), and these solutions can be found using the command

$$\text{IntegralQuarticPoints}([2, 0, 12d, 0, 16d^2])$$

in Magma. Hence, there are at most finitely many polynomials $f(x)$ that satisfy (II), and for a given value of d , these polynomials can effectively be found.

Conversely, if $c \notin R$, then $f(x)$ is irreducible, and it is straightforward to derive (3-9) by the substitution of conditions (3-11) into (3-10). □

Remark 3.6. For bounds on the number of solutions to (3-13), the interested reader should see [Bennett 1998; Bugeaud et al. 2011].

4. A second approach

In this section, we prove Theorem 1.1, which can be deduced easily using the following theorem and some results presented in Section 2.

Theorem 4.1. *Let p be a prime and let*

$$f(x) = x^n + \sum_{j=1}^{n-1} a_j x^j + c \in \mathbb{Z}[x],$$

where $n \geq 2$ and $c \equiv 0 \pmod{p}$. Suppose that $f(x)$ is irreducible. Then

- (1) *If $n \equiv 0 \pmod{2}$ and $a_1 \not\equiv -z^2 \pmod{p}$ for any $z \in \mathbb{F}_p$, then $F(x)$ is irreducible.*
- (2) *If $n \equiv 1 \pmod{2}$ and $a_1 \not\equiv z^2 \pmod{p}$ for any $z \in \mathbb{F}_p$, then $F(x)$ is irreducible.*

Proof. Since $f(x)$ is irreducible, we can apply Theorem 2.1 and Theorem 2.2 to deduce that if $F(x)$ is reducible, then

$$\begin{aligned} F(x) &= x^{2n} + \sum_{j=1}^{n-1} a_j x^{2j} + c \\ &= \left(x^n + \sum_{j=0}^{n-1} b_j x^j \right) \left(x^n + \sum_{j=0}^{n-1} (-1)^{n-j} b_j x^j \right) \\ &= \begin{cases} x^{2n} + \dots + (2b_0 b_2 - b_1^2) x^2 + b_0^2 & \text{if } n \equiv 0 \pmod{2}, \\ x^{2n} + \dots + (b_1^2 - 2b_0 b_2) x^2 - b_0^2 & \text{if } n \equiv 1 \pmod{2}. \end{cases} \end{aligned}$$

Since $c \equiv 0 \pmod{p}$, equating coefficients gives that $b_0 \equiv 0 \pmod{p}$ and

$$a_1 \equiv \begin{cases} -b_1^2 \pmod{p} & \text{if } n \equiv 0 \pmod{2}, \\ b_1^2 \pmod{p} & \text{if } n \equiv 1 \pmod{2}. \end{cases} \quad \square$$

For the convenience of the reader, we restate Theorem 1.1 here.

Theorem 1.1. *Let $p \geq 3$ be prime, and let $c, d \in \mathbb{Z}$, with $c \neq 0$, $d > 0$ and $d \not\equiv 0 \pmod{p}$. Let*

$$\begin{aligned} f(x) &= x(x+d)(x+2d) \cdots (x+(p-1)d) + c \\ &= x^p + a_{p-1} x^{p-1} + \dots + a_1 x + c. \end{aligned}$$

- (1) *If $c \not\equiv 0 \pmod{p}$, then $f(x)$ is irreducible. If, in addition, $c \neq -z^2$ for any $z \in \mathbb{Z}$, then $F(x)$ is irreducible.*
- (2) *Let k be a fixed positive integer, and suppose that $|c| = kp^w$, where*

$$p^w > k^{p-1} + a_{p-1} k^{p-2} + \dots + a_2 k + a_1.$$

Then both $f(x)$ and $F(x)$ are irreducible if one of the sets of conditions below holds:

- (a) $c > 0$.
- (b) $c < 0$, $w \equiv 1 \pmod{2}$ and $k \not\equiv 0 \pmod{p}$.
- (c) $c < 0$ and $p \equiv 3 \pmod{4}$.

Proof. Since $d \not\equiv 0 \pmod{p}$, we have that

$$f(x) \equiv x(x-1)(x-2)\cdots(x-(p-1)) + c \equiv x^p - x + c \pmod{p}.$$

Hence, since $c \not\equiv 0 \pmod{p}$, we have from Corollary 2.5 that $f(x)$ is irreducible. If, in addition, $c \neq -z^2$ for any $z \in \mathbb{Z}$, then $F(x)$ is irreducible by Theorem 2.3(2).

To prove (2), note that since $d \not\equiv 0 \pmod{p}$, we have

$$a_1 = d^{p-1}(p-1)! \equiv -1 \pmod{p} \not\equiv 0 \pmod{p} \quad (4-1)$$

by Fermat's little theorem and Wilson's theorem. Hence, $f(x)$ is irreducible by Corollary 2.7.

To establish parts (2a), (2b) and (2c), first note that $\deg(f(x)) = p$ is odd. Thus, if $c = kp^w > 0$, then $F(x)$ is irreducible by Theorem 2.3(2), which resolves (2a). For (2b), observe that kp^w is not a square since $w \equiv 1 \pmod{2}$ and $k \not\equiv 0 \pmod{p}$. Thus, again it follows from Theorem 2.3(2) that $F(x)$ is irreducible. Finally, for (2c), since $p \equiv 3 \pmod{4}$, we have from (4-1) that $a_1 \not\equiv z^2 \pmod{p}$ for any $z \in \mathbb{F}_p$. Therefore, $F(x)$ is irreducible by Theorem 4.1(2). \square

Acknowledgements

The authors thank the referee for the very careful reading of the manuscript, the many excellent suggestions, and most of all, the very timely manner in which the report was received.

References

- [Bennett 1998] M. A. Bennett, "On the number of solutions of simultaneous Pell equations", *J. Reine Angew. Math.* **498** (1998), 173–199. MR 1629862 Zbl 1044.11011
- [Bremner et al. 2000] A. Bremner, J. H. Silverman, and N. Tzanakis, "Integral points in arithmetic progression on $y^2 = x(x^2 - n^2)$ ", *J. Number Theory* **80**:2 (2000), 187–208. MR 1740510 Zbl 1009.11035
- [Bugeaud et al. 2011] Y. Bugeaud, C. Levesque, and M. Waldschmidt, "Équations de Fermat–Pell–Mahler simultanées", *Publ. Math. Debrecen* **79**:3-4 (2011), 357–366. MR 2907971 Zbl 1249.11053
- [Flügel 1909] W. Flügel, "Solution to problem 226", *Archiv. der Math. und Physik* **15** (1909), 271.
- [Győry et al. 2011] K. Győry, L. Hajdu, and R. Tijdeman, "Irreducibility criteria of Schur-type and Pólya-type", *Monatsh. Math.* **163**:4 (2011), 415–443. MR 2820371 Zbl 1232.11112
- [Koblitz 1993] N. Koblitz, *Introduction to elliptic curves and modular forms*, 2nd ed., Graduate Texts in Mathematics **97**, Springer, New York, 1993. MR 1216136 Zbl 0804.11039

- [Schinzel 1982] A. Schinzel, *Selected topics on polynomials*, University of Michigan Press, Ann Arbor, MI, 1982. MR 649775 Zbl 0487.12002
- [Schinzel 2000] A. Schinzel, *Polynomials with special regard to reducibility*, Encyclopedia of Mathematics and its Applications **77**, Cambridge University Press, 2000. MR 1770638 Zbl 0956.12001
- [Schur 1908] I. Schur, “Problem 226”, *Archiv Math. Physik* **13**:3 (1908), 367.
- [Seres 1956] I. Seres, “Lösung und Verallgemeinerung eines Schurschen Irreduzibilitätsproblems für Polynome”, *Acta Math. Acad. Sci. Hungar.* **7** (1956), 151–157. MR 0082952 Zbl 0071.01801
- [Serret 1992] J.-A. Serret, *Cours d’algèbre supérieure, II*, 4th ed., Éditions Jacques Gabay, Sceaux, 1992. MR 1190472
- [Silverman 2009] J. H. Silverman, *The arithmetic of elliptic curves*, 2nd ed., Graduate Texts in Mathematics **106**, Springer, Dordrecht, 2009. MR 2514094 Zbl 1194.11005
- [Weisner 1934] L. Weisner, “Criteria for the irreducibility of polynomials”, *Bull. Amer. Math. Soc.* **40**:12 (1934), 864–870. MR 1562990 Zbl 0010.29001
- [Westlund 1909] J. Westlund, “On the irreducibility of certain polynomials”, *Amer. Math. Monthly* **16**:4 (1909), 66–67. MR 1517192 JFM 40.0123.01

Received: 2015-03-14 Revised: 2015-05-18 Accepted: 2015-06-17

lkjone@ship.edu

*Department of Mathematics, Shippensburg University,
Shippensburg, PA 17257-2299, United States*

al5903@ship.edu

*Department of Mathematics, Shippensburg University,
Shippensburg, PA 17257-2299, United States*

Oscillation of solutions to nonlinear first-order delay differential equations

James P. Dix and Julio G. Dix

(Communicated by Kenneth S. Berenhaut)

In this article, we present sufficient conditions for the oscillation of all solutions to the delay differential equation

$$x'(t) + \sum_{i=1}^n f_i(t, x(\tau_i(t))) = 0, \quad t \geq t_0.$$

In particular, we extend known results from linear to nonlinear equations, and improve the bounds of previous criteria.

1. Introduction

In this article, we study the delay differential equation

$$x'(t) + \sum_{i=1}^n f_i(t, x(\tau_i(t))) = 0, \quad t \geq t_0, \quad (1-1)$$

where $f_i : [t_0, \infty) \times \mathbb{R} \rightarrow \mathbb{R}$ and $\tau_i : [t_0, \infty) \rightarrow \mathbb{R}$ are continuous functions satisfying conditions stated below. We establish sufficient conditions for all solutions to oscillate.

When $f_i(t, x) = p_i(t)x$, equation (1-1) becomes linear and it is easy to show that all solutions oscillate or tend to zero, under the assumption

$$\int_{t_0}^{\infty} \sum_{i=1}^n p_i(s) ds = \infty. \quad (1-2)$$

This result has been extended to delay equations of several types: nonlinear, nonhomogeneous, higher order, neutral equations, etc.; see, for example, [Dix et al. 2008; Elbert and Stavroulakis 1995; Erbe et al. 1995; Gil' 2014; Győri and Ladas 1991; Hale 1977; Ladde et al. 1987; Zhou 2011]. Since we want to ensure oscillation, we impose conditions stronger than the one above.

MSC2010: 34K11, 34C10.

Keywords: oscillation of solutions, first-order delay differential equation.

For $n = 1$ and $f_1(t, x) = p_1(t)x$, there are two well-known conditions for the oscillation of all solutions: [Ladde et al. 1987, Theorem 2.1.3],

$$\limsup_{t \rightarrow \infty} \int_{\tau_1(t)}^t p_1(s) ds > 1, \tag{1-3}$$

and [Ladde et al. 1987, Theorem 2.1.1],

$$\liminf_{t \rightarrow \infty} \int_{\tau_1(t)}^t p_1(s) ds > \frac{1}{e}. \tag{1-4}$$

Some authors try to narrow the gap between these two lower bounds, while others extended the above criteria for covering more general equations. In this article, we try both of these tasks.

Braverman and Karpuz [2011] showed that when applying (1-3), the conditions that $\tau_1(t) < t$ and τ_1 be nondecreasing are necessary. They also modified (1-3) by using Grönwall’s inequality. Chatzarakis and Öcalan [2015] applied the modified condition to multiple delay equations. We extend these results to nonlinear equations.

For $f_i(t, x) = p_i(t)x$, Grammatikopoulos et al. [2003] assumed that τ_i is monotonic. We do not use the monotonicity assumption. Györi and Ladas [1991] stated conditions using a nondecreasing upper bound for the delayed arguments, similar to our σ defined below. Hunt and Yorke [1984] proved oscillation of solutions assuming that

$$\liminf_{t \rightarrow \infty} \sum_{i=1}^n p_i(s)(t - \tau_i(t)) > \frac{1}{e}$$

and that $t - \tau_i(t)$ is bounded. They did not assume monotonicity of τ_i , and used an inequality of differentials in their proof. We extend their result to nonlinear equations; see Theorem 4.6 below. Li [1996] used a logarithmic inequality to obtain a condition weaker than (1-4) for constant delays. We use the same logarithmic inequality for variable delays in nonlinear equations. Fukagai and Kusano [1984] considered retarded and advanced nonlinear equations with $f_i(t, x) = p_i(t)g_i(x)$, where g_i satisfies conditions similar to those in (H2) below. We assume that $f_i(t, x) \geq p_i(t)g_i(x)$, and then apply the Grönwall and logarithmic inequalities.

In this article, we use the hypotheses

(H1) $\tau_i(t) < t$ for $t \geq t_0$, and $\lim_{t \rightarrow \infty} \tau_i(t) = \infty$ for $i = 1, 2, \dots, n$.

(H2) $x f_i(t, x) \geq 0$, and there exist continuous functions $p_i(t) \geq 0$ and $g_i(x)$ such that

$$|f_i(t, x)| \geq p_i(t)|g_i(x)| \quad \forall x \in \mathbb{R}, t \geq t_0,$$

where $x g_i(x) > 0$ for $x \neq 0$ and $\limsup_{x \rightarrow 0} x/g_i(x) < \infty$. Without loss of generality, we assume that

$$\limsup_{x \rightarrow 0} \frac{x}{g_i(x)} < 1. \tag{1-5}$$

If $\limsup_{x \rightarrow 0} x/g_i(x) = M_1 \geq 1$, we multiply p_i by a constant greater than M_1 , and divide g_i by the same constant; so the assumption is satisfied without modifying f_i .

We define the functions

$$\tau_0(t) = \max_{1 \leq i \leq n} \tau_i(t), \quad \sigma(t) = \max_{t_0 \leq s \leq t} \tau_0(s).$$

Then σ is nondecreasing. Also by (H1), we have $\tau_i(t) \leq \tau_0(t) \leq \sigma(t) < t$, and

$$\lim_{t \rightarrow \infty} \tau_0(t) = \infty, \quad \lim_{t \rightarrow \infty} \sigma(t) = \infty.$$

Let $t_{-1} = \min_{1 \leq i \leq n} \inf_{t_0 \leq t} \tau_i(t)$. Then the initial condition for (1-1) is

$$x(t) = \phi(t) \quad \text{for } t \in [t_{-1}, t_0], \tag{1-6}$$

where $\phi : [t_{-1}, t_0] \rightarrow \mathbb{R}$ is a continuous function.

By a solution we mean a function that is continuous on $[t_{-1}, \infty)$, differentiable on $[t_0, \infty)$, and satisfies (1-1) and (1-6).

A unique solution x can be obtained by the method of steps: Using the information on $[t_{-1}, t_0]$, define x by integrating (1-1) for $t \in [t_0, t_1]$, where t_1 is the largest value such that $\tau_i(t) \leq t_0$ for all $t \leq t_1$, where $i = 1, 2, \dots, n$. Then we repeat the process for $[t_1, t_2)$ and so on.

A function is said to be oscillatory if it has arbitrarily large zeros; otherwise it is called nonoscillatory. A function x is said to be eventually positive if there exists t^* such that $x(t) > 0$ for all $t \geq t^*$. We define eventually negative similarly.

Lemma 1.1. *Under assumptions (H1), (H2) and (1-2), if x is an eventually positive solution of (1-1), then there exists $t_1 \geq t_0$ such that $x(t) > 0$, $x(\tau_i(t)) > 0$, x is nonincreasing, and $|x(\tau_i(t))| \leq |g_i(x(\tau_i(t)))|$ for $t \geq t_1$ and $i = 1, 2, \dots, n$.*

Proof. Since x is eventually positive, there exists $t^* \geq t_0$ such that $x(t) > 0$ for $t \geq t^*$. Since $\lim_{t \rightarrow \infty} \tau_i(t) = \infty$ for $i = 1, 2, \dots, n$, there exists $t^{**} \geq t^*$ such that $\tau_i(t) \geq t^*$; thus $x(\tau_i(t)) > 0$ for $t \geq t^{**}$ and $i = 1, 2, \dots, n$.

From (H2), $f_i(t, x(\tau_i(t))) \geq 0$, and from (1-1), $x'(t) \leq 0$. Therefore, x is nonincreasing. Since x is nonnegative and nonincreasing, it must converge to a number $\alpha \geq 0$ as $t \rightarrow \infty$. We claim that $\alpha = 0$. To reach a contradiction, assume that $\lim_{t \rightarrow \infty} x(t) = \alpha > 0$. Then $0 < \alpha \leq x \leq x_{\max}$. Since g_i is continuous and positive on $[\alpha, x_{\max}]$, there exists $\gamma_i > 0$ such that $\gamma_i \leq g(x(\tau_i(t)))$ for all $t \geq t^{**}$. By (1-1) and (H2),

$$0 \geq x'(t) + \sum_{i=1}^n p_i(t)g_i(x(\tau_i(t))) \geq x'(t) + \sum_{i=1}^n p_i(t)\gamma_i.$$

Integrating from t^{**} to t ,

$$\alpha - x(t^{**}) \leq x(t) - x(t^{**}) \leq - \int_{t^{**}}^t \sum_{i=1}^n p_i(s)\gamma_i ds.$$

Note that as $t \rightarrow \infty$, by (1-2), the right-hand side approaches $-\infty$, while the left-hand side is constant. This contradiction implies $\lim_{t \rightarrow \infty} x(t) = 0$. From (1-5) and $\lim_{t \rightarrow \infty} \tau_i(t) = \infty$, there exists $t_1 \geq t^{**}$ such that $x(\tau_i(t)) \leq g_i(x(\tau_i(t)))$ for all $t \geq t_1$. \square

Under the assumptions of Lemma 1.1, from the definitions of τ_0 and σ , for all $t \geq t_1$, we have the inequalities

$$0 = x'(t) + \sum_{i=1}^n f_i(t, x(\tau_i(t))) \geq x'(t) + \sum_{i=1}^n p_i(t)x(\tau_i(t)) \tag{1-7}$$

$$\geq x'(t) + x(\tau_0(t)) \sum_{i=1}^n p_i(t) \tag{1-8}$$

$$\geq x'(t) + x(\sigma(t)) \sum_{i=1}^n p_i(t) \tag{1-9}$$

$$\geq x'(t) + x(t) \sum_{i=1}^n p_i(t). \tag{1-10}$$

For the rest of this article, we reserve the symbol t_1 for the value obtained in Lemma 1.1. Note that a similar value t_1 can be obtained for eventually negative solutions. In such case, inequalities (1-7)–(1-10) need to be reversed.

2. Conditions using the limit superior

A direct application of [Ladde et al. 1987, Theorem 2.1.3] to (1-9) states that

$$\limsup_{t \rightarrow \infty} \int_{\sigma(t)}^t \sum_{i=1}^n p_i(s) ds > 1 \tag{2-1}$$

implies the oscillation of all solutions to (1-1). This corresponds to [Ladde et al. 1987, Remark 2.7.3], where the assumption that τ_i is nondecreasing needs to be added.

Regarding the necessity of σ being monotonic and $\sigma(t) < t$, Braverman and Karpuz [2011] considered the single delay equation

$$x'(t) + p_1 x(\tau_1(t)) = 0, \tag{2-2}$$

with the assumption

$$\limsup_{t \rightarrow \infty} \int_{\tau_1(t)}^t p_1 ds > A, \tag{2-3}$$

where A and p_1 are positive constants. They showed that for every A , there exists a p_1 and a nonmonotonic delay τ_1 , with $\tau_1(t) = t$ on some intervals, such that (2-3) is satisfied, but (2-2) has a nonoscillatory solution. We shall show a similar result for (1-1), when τ_1 remains monotonic; see Theorem 2.5 below.

As in [Braverman and Karpuz 2011, Corollary 1] and [Chatzarakis and Öcalan 2015, Theorem 1], we use Grönwall’s inequality to obtain a condition weaker than (2-1).

Lemma 2.1. *Assume that (H1), (H2) and (1-2) hold and that x is an eventually positive solution of (1-1). Then*

$$\int_{\sigma(t)}^t \sum_{i=1}^n p_i(s) \exp\left(\int_{\tau_i(s)}^{\sigma(t)} \sum_{j=1}^n p_j(r) dr\right) ds < 1 \quad \forall t \geq t_1, \tag{2-4}$$

where t_1 is defined by Lemma 1.1

Proof. Grönwall’s inequality applied to (1-10) with $x > 0$ and $\tau_i(s) \leq \sigma(t)$ yields

$$x(\tau_i(s)) \geq x(\sigma(s)) \exp\left(\int_{\tau_i(s)}^{\sigma(t)} \sum_{j=1}^n p_j(r) dr\right). \tag{2-5}$$

Integrating (1-1) from $\sigma(t)$ to t and using (H2) and (2-5) yields

$$\begin{aligned} 0 &\geq x(t) - x(\sigma(t)) + \int_{\sigma(t)}^t \sum_{i=1}^n p_i(s) x(\tau_i(s)) ds \\ &\geq x(t) - x(\sigma(t)) + x(\sigma(t)) \int_{\sigma(t)}^t \sum_{i=1}^n p_i(s) \exp\left(\int_{\tau_i(s)}^{\sigma(t)} \sum_{j=1}^n p_j(r) dr\right) ds. \end{aligned} \tag{2-6}$$

Denoting the outer integral by $\mathbb{P}(t)$,

$$0 < x(t) \leq x(\sigma(t))(1 - \mathbb{P}(t)) \quad \forall t \geq t_1. \tag{2-7}$$

Therefore, $\mathbb{P}(t) < 1$ for all $t \geq t_1$, which completes the proof. □

Theorem 2.2. *Assume (H1), (H2) and (1-2). If there exists a sequence $\{u_k\} \rightarrow \infty$ such that*

$$\int_{\sigma(u_k)}^{u_k} \sum_{i=1}^n p_i(s) \exp\left(\int_{\tau_i(s)}^{\sigma(u_k)} \sum_{j=1}^n p_j(r) dr\right) ds \geq 1 \quad \forall k, \tag{2-8}$$

then all solutions of (1-1) are oscillatory.

Proof. To reach a contradiction, assume that there is a nonoscillatory solution x , and initially assume x is eventually positive. Let t_1 be defined by Lemma 1.1. Then by Lemma 2.1, inequality (2-4) is satisfied, which contradicts (2-8). Therefore x cannot be eventually positive.

When x is eventually negative, we prove a variation of Lemma 1.1 in which $x(t) < 0, x(\tau_i(t)) < 0, x$ is nondecreasing, and $|x(\tau_i(t))| \leq |g_i(x(\tau_i(t)))|$ for $t \geq t_1$. Then we show that Lemma 2.1 still holds. In its proof, we need to reverse inequalities (2-5), (2-6) and (2-7). With these two lemmas, we obtain again a contradiction to (2-8), which implies that x cannot be eventually negative. □

Remark 2.3. Note that (2-8) is implied by

$$\limsup_{t \rightarrow \infty} \int_{\sigma(t)}^t \sum_{i=1}^n p_i(s) \exp\left(\int_{\tau_i(s)}^{\sigma(t)} \sum_{j=1}^n p_j(r) dr\right) ds > 1. \tag{2-9}$$

Since the exponent in (2-9) is not negative, it follows that (2-9) is implied by (2-1). In summary, (2-8) is less restrictive than (2-2).

Remark 2.4. When the equal sign in (1-1) is replaced by \leq , the new equation cannot have eventually positive solutions under assumption (2-8). Similarly when the equal sign in (1-1) is replaced by \geq , the new equation cannot have eventually negative solutions under assumption (2-8).

Regarding the necessity of the hypothesis $\sigma(t) < t$ in Theorem 2.2, we consider the single delay equation

$$x'(t) + p_1 x(\tau_1(t)) = 0 \tag{2-10}$$

with the assumption

$$\limsup_{t \rightarrow \infty} \int_{\tau_1(t)}^t p_1 \exp\left(\int_{\tau_1(s)}^{\sigma(t)} p_1 dr\right) ds > A, \tag{2-11}$$

where A and p_1 are positive constants.

Theorem 2.5. For each p_1 and each $A < e$, there exists a monotonic delay with $\tau_1(t) = t$ on certain intervals such that (2-11) is satisfied, but (2-10) has a nonoscillatory solution.

Proof. Since the continuous mapping $y \mapsto ye^y$ is strictly increasing and maps zero to zero and 1 to e , there exists $\beta < 1$ such that $\beta e^\beta = A$. Since for positive integers, $\frac{m-1}{m} < 1$ and $\lim_{m \rightarrow \infty} \frac{m-1}{m} = 1$, there exists m such that $\beta < \frac{m-1}{m} < 1$. Then

$$\frac{m\beta}{(m-1)p_1} < \frac{1}{p_1}.$$

By the completeness of the real numbers, there exists α such that

$$\frac{m\beta}{(m-1)p_1} < \alpha < \frac{1}{p_1}.$$

In summary, for some integer m , we have

$$\alpha p_1 < 1 \quad \text{and} \quad \beta < \frac{(m-1)\alpha p_1}{m}. \tag{2-12}$$

As a delayed argument, we define the piecewise linear function

$$\tau_1(t) = \begin{cases} t & \text{if } 0 \leq t \leq \alpha, \\ \alpha & \text{if } \alpha < t < \frac{2m-1}{m}\alpha, \\ 2\alpha + m(t - 2\alpha) & \text{if } \frac{2m-1}{m}\alpha \leq t \leq 2\alpha. \end{cases}$$

For $t \in (2\alpha, 4\alpha]$, we use the formula $\tau_1(t) = 2\alpha + \tau_1(t - 2\alpha)$, and a similar formula for $t \in (4\alpha, 6\alpha]$, etc. Note that τ_1 is continuous, nondecreasing, $\lim_{t \rightarrow \infty} \tau_1(t) = \infty$, and $\tau_1(t) = \tau_0(t) = \sigma(t)$. To define a solution to (2-10), we use an initial condition $x(t) = x_0 > 0$ for $t \leq 0$.

On the interval $[0, \alpha]$, equation (2-10) becomes an ordinary differential equation whose solution is $x(t) = x_0 e^{-p_1 t}$, which is positive and decreasing.

On the interval $[\alpha, \frac{2m-1}{m}\alpha]$, the delayed argument is $\tau_1(t) = \alpha$. Then (2-10) has the solution

$$x(t) = x(\alpha) - p_1 x(\alpha)(t - \alpha) = x(\alpha)(1 - (t - \alpha)p_1), \tag{2-13}$$

which is decreasing. From the inequality $t \leq \frac{2m-1}{m}\alpha < 2\alpha$, we obtain the lower bound

$$x(t) > x(\alpha)(1 - \alpha p_1),$$

which is positive because of (2-12).

So far the solution is positive on $[0, \frac{2m-1}{m}\alpha]$. Next we show that the solution cannot have zeros in $(\frac{2m-1}{m}\alpha, 2\alpha]$. To reach a contradiction, let t_2 be the smallest zero in $(\frac{2m-1}{m}\alpha, 2\alpha]$. By the mean value theorem, there exists t^* in $(\frac{2m-1}{m}\alpha, t_2)$ such that

$$x'(t^*) = \frac{x(\frac{2m-1}{m}\alpha) - 0}{\frac{2m-1}{m}\alpha - t_2}.$$

From $t_2 < 2\alpha$, it follows that

$$x'(t^*) < \frac{x(\frac{2m-1}{m}\alpha)}{-\frac{\alpha}{m}}. \tag{2-14}$$

Note that for $t \leq t_2$, we have $\tau_1(t) < t_2$. Since $x(t) \geq 0$ for all $t \leq t_2$, by (2-10), $x'(t) \leq 0$ so that x is nonincreasing for all $t \leq t_2$. Because x is nonincreasing and $\alpha \leq \tau_1(t^*)$, we have $x(\tau_1(t^*)) \leq x(\alpha)$. This and (2-14) imply

$$0 = x'(t^*) + p_1 x(\tau_1(t^*)) < \frac{x(\frac{2m-1}{m}\alpha)}{-\frac{\alpha}{m}} + p_1 x(\alpha). \tag{2-15}$$

From (2-13),

$$x\left(\frac{2m-1}{m}\alpha\right) = x(\alpha)\left(1 - \left(\frac{2m-1}{m}\alpha - \alpha\right)p_1\right) = x(\alpha)\left(1 - \frac{m-1}{m}\alpha p_1\right).$$

Substituting this value in (2-15) yields

$$x(\alpha)\left(1 - \frac{m-1}{m}\alpha p_1\right) < \frac{\alpha}{m} p_1 x(\alpha),$$

which implies $1 - \frac{m-1}{m}\alpha p_1 < \frac{\alpha}{m} p_1$. This in turn implies $1 < \alpha p_1$, and contradicts (2-11). Therefore, $x(t) > 0$ on $[0, 2\alpha]$.

Next we set $x(2\alpha)$ as the initial value, and solve (2-10) on $[2\alpha, 4\alpha]$. Repeating this process, we have a positive solution on $[0, \infty)$.

It remains to show that (2-11) is satisfied. From the definition of τ_1 , when $t = u_1 = \frac{2m-1}{m}\alpha$, we have $\sigma(u_1) = \frac{2m-1}{m}\alpha$. For $\alpha \leq s \leq \frac{2m-1}{m}\alpha$, we have that $\tau_1(s) = \alpha$. Then (2-11) becomes

$$\int_{\alpha}^{\frac{2m-1}{m}\alpha} p_1 \exp\left(\int_{\alpha}^{\frac{2m-1}{m}\alpha} p_1 dr\right) ds = \frac{m-1}{m}\alpha p_1 \exp\left(\frac{m-1}{m}\alpha p_1\right).$$

Since the mapping $y \mapsto e^y$ is increasing, by (2-12),

$$\frac{m-1}{m}\alpha p_1 \exp\left(\frac{m-1}{m}\alpha p_1\right) > \beta e^{\beta} > A.$$

Repeating this process at $u_k = 2k\alpha + \frac{2m-1}{m}\alpha$, we obtain a sequence at which the above inequality holds. The presence of this sequence implies (2-8) and (2-11) are satisfied. □

3. Conditions using the limit inferior

A direct application of [Ladde et al. 1987, Theorem 2.1.1] to (1-8) states that

$$\liminf_{t \rightarrow \infty} \int_{\sigma(t)}^t \sum_{i=1}^n p_i(s) ds > \frac{1}{e} \tag{3-1}$$

implies the oscillation of all solutions of (1-1). Also note that (3-1) implies (1-2).

Grammatikopoulos et al. [2003] showed that for (1-1) with $f_i(t, x) = p_i(t)x$, all solutions are oscillatory when the τ_i are nondecreasing, and

$$\int_0^{\infty} |p_i(s) - p_j(s)| ds < \infty, \tag{3-2}$$

$$\liminf_{t \rightarrow \infty} \int_{\tau_i(t)}^t p_i(s) ds = \beta_i > 0, \quad \sum_{i=1}^n \liminf_{t \rightarrow \infty} \int_{\tau_i(t)}^t p_i(s) ds > \frac{1}{e}.$$

As in the previous part, we use Grönwall’s inequality for finding a condition less restrictive than (3-1).

Lemma 3.1. *Assume (H1), (H2). If x is an eventually positive solution of (1-1), and*

$$\liminf_{t \rightarrow \infty} \int_{\sigma(t)}^t \sum_{i=1}^n p_i(s) \exp\left(\int_{\tau_i(s)}^{\sigma(s)} \sum_{j=1}^n p_j(r) dr\right) ds > \frac{1}{e}, \tag{3-3}$$

then $\lim_{t \rightarrow \infty} x(\sigma(t))/x(t) = \infty$.

Proof. By a contrapositive argument, we can show that (3-3) implies (1-2), so we let t_1 be defined by Lemma 1.1. Applying Grönwall’s inequality to (1-10) yields

(2-5), which is substituted in (1-7) to obtain

$$0 \geq x'(t) + \sum_{i=1}^n p_i(t)x(\sigma(t)) \exp\left(\int_{\tau_i(t)}^{\sigma(t)} \sum_{j=1}^n p_j(r) dr\right) \quad \forall t \geq t_1. \quad (3-4)$$

Dividing by $x(t)$ and integrating from $\sigma(t)$ to t , we obtain

$$\ln\left(\frac{x(t)}{x(\sigma(t))}\right) + \int_{\sigma(t)}^t \sum_{i=1}^n p_i(s) \frac{x(\sigma(s))}{x(s)} \exp\left(\int_{\tau_i(s)}^{\sigma(s)} \sum_{j=1}^n p_j(r) dr\right) ds \leq 0. \quad (3-5)$$

From (3-3), there exist constants $t_2 \geq t_1$ and α such that

$$\int_{\sigma(t)}^t \sum_{i=1}^n p_i(s) \exp\left(\int_{\tau_i(s)}^{\sigma(s)} \sum_{j=1}^n p_j(r) dr\right) ds \geq \alpha > \frac{1}{e} \quad \forall t \geq t_2.$$

Since $\sigma(s) < s$ and x is nonincreasing, $x(\sigma(s))/x(s) \geq 1$. Then (3-5) and the above inequality yield

$$\ln\left(\frac{x(t)}{x(\sigma(t))}\right) + \alpha \leq 0.$$

Since $\alpha e \leq e^\alpha$ for all α ,

$$\alpha e \leq e^\alpha \leq \frac{x(\sigma(t))}{x(t)} \quad \forall t \geq t_2. \quad (3-6)$$

Since $\lim_{t \rightarrow \infty} \sigma(t) = \infty$, there exists $t_3 \geq t_2$ such that $\sigma(t) \geq t_2$ for all $t \geq t_3$. Using (3-6) in (3-5), we obtain

$$(\alpha e)^2 \leq \frac{x(\sigma(s))}{x(s)} \quad \forall t \geq t_3.$$

Repeating this process, we obtain

$$(\alpha e)^k \leq \frac{x(\sigma(s))}{x(s)}$$

for all t sufficiently large. Since $\alpha e > 1$, the assertion of the lemma follows. \square

Theorem 3.2. *Under assumptions (H1), (H2) and (3-3), all solutions to (1-1) are oscillatory.*

Proof. To reach a contradiction, assume that there is a nonoscillatory solution x , which initially is assumed to be eventually positive. By a contrapositive argument, we can show that (3-3) implies (1-2), so we let t_1 be defined by Lemma 1.1. To simplify notation, we define

$$\mathbb{P}(s) = \sum_{i=1}^n p_i(s) \exp\left(\int_{\tau_i(s)}^{\sigma(s)} \sum_{i=1}^n p_i(r) dr\right).$$

Then from (3-3), there exist constants $t_2 \geq t_1$ and α such that

$$\int_{\sigma(t)}^t \mathbb{P}(s) ds \geq \alpha > \frac{1}{e} \quad \forall t \geq t_2.$$

Using the intermediate value theorem, we can show that there exists $t^* \in (\sigma(t), t)$ such that

$$\int_{\sigma(t)}^{t^*} \mathbb{P}(s) ds \geq \frac{\alpha}{2} \quad \text{and} \quad \int_{t^*}^t \mathbb{P}(s) ds \geq \frac{\alpha}{2}. \tag{3-7}$$

Integrating (1-7) from $\sigma(t)$ to t^* and using (2-5) yield

$$x(t^*) - x(\sigma(t)) + x(\sigma(t^*)) \int_{\sigma(t)}^{t^*} \mathbb{P}(s) ds \leq 0.$$

Using that $x(t^*) > 0$ and (3-7), we obtain

$$x(\sigma(t^*)) \leq \frac{2}{\alpha} x(\sigma(t)). \tag{3-8}$$

Integrating (1-7) from t^* to t and using (2-5) yield

$$x(\sigma(t)) - x(t^*) + x(\sigma(t)) \int_{t^*}^t \mathbb{P}(s) ds \leq 0.$$

Using that $x(\sigma(t)) > 0$ and (3-7), we obtain

$$x(\sigma(t)) \leq \frac{2}{\alpha} x(t^*).$$

Using this inequality in (3-8) yields

$$\frac{x(\sigma(t^*))}{x(t^*)} \leq \left(\frac{2}{\alpha}\right)^2.$$

Because $\sigma(t) \leq t^* \leq t$ and $\lim_{t \rightarrow \infty} \sigma(t) = \infty$, the above inequality contradicts Lemma 3.1; so the solution x cannot be eventually positive.

When x is eventually negative, as in Lemma 1.1, there exist $t_1 \geq t_0$ such that $x(t) < 0$, $x(\tau_i(t)) < 0$, $x(t)$ is nondecreasing and $|x(\tau_i(t))| \leq |g_i(x(\tau_i(t)))|$ for $t \geq t_1$. Then Lemma 3.1 holds, but in its proof we need to reverse inequality (3-4). Again we reach a contradiction indicating that x cannot be eventually negative. \square

Remark 3.3. Note that the exponent in (3-3) is nonnegative; therefore, condition (3-1) is more restrictive than (3-3). Also the statements in Remark 2.4 apply to condition (3-3).

4. Estimates using a logarithmic inequality

Li [1996] used the inequality $e^{rx} \geq x + \frac{1}{r}(1 + \ln r)$ to show that all solutions to (1-1) are oscillatory when $f_i(t, x) = p_i(t)x$ and the delays have the form $\tau_i(t) = t - k_i$

with positive constants k_i . There, the key assumption is

$$\int_{t_0}^{\infty} \sum_{i=1}^n p_i(s) \left(1 + \ln \left(\int_s^{s+k_i} \sum_{j=1}^n p_j(r) dr \right) \right) ds = \infty. \tag{4-1}$$

We want to extend the result in [Li 1996] to (1-1) that are nonlinear and have variable delays. The variable delays cause some difficulties when obtaining a condition similar to (4-1).

First we define a function that is the inverse of σ almost everywhere. Under assumption (H1), the function σ is continuous; thus for each s , the set $\sigma^{-1}(s)$ is closed. Since σ is monotonic and $\lim_{t \rightarrow \infty} \sigma(t) = \infty$, the set $\sigma^{-1}(s)$ is a closed and bounded interval. There are at most countably many of those closed intervals that do not consist of a single point. Let

$$\sigma_{\text{inv}}(s) = \max\{t : \sigma(t) = s\}.$$

Note that σ_{inv} is strictly increasing and has at most countably many discontinuities. Also $s < \sigma_{\text{inv}}(s)$, and σ_{inv} is bounded on bounded intervals. Under these conditions, σ_{inv} is Riemann integrable, and expressions of the form $\int_a^b p(s) \int_s^{\sigma_{\text{inv}}(s)} \lambda(r) dr ds$ are well-defined for all continuous functions λ, p . Also the value of this integral remains the same when $\sigma_{\text{inv}}(s)$ is replaced by any t as long as $\sigma(t) = s$. This happens because the integrand would change only at countably many points.

Lemma 4.1. *Under assumption (H1), for $a \leq \sigma(b)$ and any continuous nonnegative functions λ and p , we have*

$$\int_a^b p(s) \int_{\sigma(s)}^s \lambda(r) dr ds \geq \int_a^{\sigma(b)} \lambda(s) \int_s^{\sigma_{\text{inv}}(s)} p(r) dr ds. \tag{4-2}$$

Proof. Interchanging the order of integration on the left-hand side of (4-2) gives

$$\begin{aligned} \int_a^b p(s) \int_{\sigma(s)}^s \lambda(r) dr ds &= \int_{\sigma(a)}^a \lambda(r) \int_a^{\sigma_{\text{inv}}(r)} p(s) ds dr \\ &\quad + \int_a^{\sigma(b)} \lambda(r) \int_r^{\sigma_{\text{inv}}(r)} p(s) ds dr + \int_{\sigma(b)}^b \lambda(r) \int_r^b p(s) ds dr. \end{aligned}$$

Since all these integrals are nonnegative, we use the second integral in the right-hand side as a lower bound. Renaming the variables r and s , we obtain the assertion of the lemma. □

Lemma 4.2. *Under assumptions (H1), (H2) and (1-2), if (1-1) has an eventually positive solution, then*

$$\int_t^{\sigma_{\text{inv}}(t)} \sum_{i=1}^n p_i(s) ds < 1 \quad \forall t \geq t_1,$$

where t_1 is defined by Lemma 1.1

Proof. Let x be an eventually positive solution of (1-1). Recall that x is nonincreasing, σ is nondecreasing, and $t < \sigma_{\text{inv}}(t)$. Integrating (1-9) from t to $\sigma_{\text{inv}}(t)$, we have

$$x(\sigma_{\text{inv}}(t)) - x(t) + x(t) \int_t^{\sigma_{\text{inv}}(t)} \sum_{i=1}^n p_i(s) ds \leq 0. \tag{4-3}$$

Then

$$0 < x(\sigma_{\text{inv}}(t)) \leq x(t) \left(1 - \int_t^{\sigma_{\text{inv}}(t)} \sum_{i=1}^n p_i(s) ds \right) \quad \forall t \geq t_1. \tag{4-4}$$

The assertion of the lemma follows. □

Lemma 4.3. *Under assumptions (H1), (H2) and (1-2), if x is an eventually positive solution of (1-1) and*

$$\limsup_{t \rightarrow \infty} \int_t^{\sigma_{\text{inv}}(t)} \sum_{i=1}^n p_i(s) ds > 0, \tag{4-5}$$

then $\liminf_{t \rightarrow \infty} x(\sigma(t))/x(t) < \infty$.

Proof. Let t_1 be defined by Lemma 1.1. From (4-5), there exist a constant α and a sequence $\{t_k\}_{k=2}^\infty \rightarrow \infty$ such that

$$\int_{t_k}^{\sigma_{\text{inv}}(t_k)} \sum_{i=1}^n p_i(s) ds \geq \alpha > 0 \quad \forall k \geq 2.$$

Using the intermediate value theorem, we can show that there exists t_k^* in the interval $(t_k, \sigma_{\text{inv}}(t_k))$ such that

$$\int_{t_k}^{t_k^*} \sum_{i=1}^n p_i(s) ds \geq \frac{\alpha}{2} \quad \text{and} \quad \int_{t_k^*}^{\sigma_{\text{inv}}(t_k)} \sum_{i=1}^n p_i(s) ds \geq \frac{\alpha}{2}. \tag{4-6}$$

Integrating (1-9) from t_k to t_k^* , and using that σ is nondecreasing while x is nonincreasing, yields

$$x(t_k^*) - x(t_k) + x(\sigma(t_k^*)) \int_{t_k}^{t_k^*} \sum_{i=1}^n p_i(s) ds \leq 0.$$

Using that $x(t_k^*) > 0$ and (4-6), we have

$$x(\sigma(t_k^*)) \leq \frac{2}{\alpha} x(t_k). \tag{4-7}$$

Integrating (1-9) from t_k^* to $\sigma_{\text{inv}}(t_k)$ yields

$$x(\sigma_{\text{inv}}(t_k)) - x(t_k^*) + x(t_k) \int_{t_k^*}^{\sigma_{\text{inv}}(t_k)} \sum_{i=1}^n p_i(s) ds \leq 0.$$

Using that $x(\sigma_{\text{inv}}(t_k)) > 0$ and (4-6), we have

$$x(t_k) \leq \frac{2}{\alpha} x(t_k^*). \tag{4-8}$$

Using (4-8) in (4-7), it follows that

$$\frac{x(\sigma(t_k^*))}{x(t_k^*)} \leq \left(\frac{2}{\alpha}\right)^2 \quad \forall k \geq 2.$$

The assertion of the lemma follows by calculating the limit inferior as $k \rightarrow \infty$. \square

Theorem 4.4. *Assume (H1), (H2), and*

$$\int_s^{\sigma_{\text{inv}}(s)} \sum_{j=1}^n p_j(r) dr > 0 \quad \forall s \geq t_0, \tag{4-9}$$

$$\int_{t_0}^{\infty} \sum_{i=1}^n p_i(s) \left(1 + \ln \left(\int_s^{\sigma_{\text{inv}}(s)} \sum_{j=1}^n p_j(r) dr \right) \right) ds = \infty. \tag{4-10}$$

Then every solution of (1-1) is oscillatory.

Proof. To reach a contradiction, assume that there is a nonoscillatory solution x , which initially is assumed to be eventually positive. By a contrapositive argument, we can show that (4-10) implies (1-2), so we let t_1 be defined by Lemma 1.1. Let

$$\lambda(t) = \frac{-x'(t)}{x(t)} \quad \text{for } t \geq t_1.$$

Then λ is a continuous and nonnegative function. Integrating λ from a value t^* to t , we have $x(t) = x(t^*) \exp\left(-\int_{t^*}^t \lambda(s) ds\right)$. Then

$$x'(t) = -\lambda(t)x(t^*) \exp\left(-\int_{t^*}^t \lambda(s) ds\right).$$

Substituting this expression in (1-1) yields

$$\lambda(t) = \frac{1}{x(t^*)} \sum_{i=1}^n f_i(t, x(\tau_i(t))) \exp\left(\int_{t^*}^t \lambda(s) ds\right).$$

For $t^* = \sigma(t) < t$, using (H2) and $x(\sigma(t)) \leq x(\tau_i(t))$, we obtain

$$\lambda(t) \geq \sum_{i=1}^n p_i(t) \exp\left(\int_{\sigma(t)}^t \lambda(r) dr\right). \tag{4-11}$$

Note that the corresponding inequality on [Li 1996, page 3734] is incorrect, but it does not affect their proof of Theorem 1. Next as in [Li 1996], we use the inequality

$$e^{\gamma\beta} \geq \gamma + \frac{1}{\beta}(1 + \ln(\beta)) \quad \forall \beta > 0, \tag{4-12}$$

which can be shown by fixing β and minimizing $e^{\gamma\beta} - \gamma - \frac{1}{\beta}(1 + \ln(\beta))$ with respect to γ . Let

$$\beta(s) = \int_s^{\sigma_{\text{inv}}(s)} \sum_{i=1}^n p_i(r) dr,$$

which is positive. Then by (4-11) and (4-12),

$$\begin{aligned} \lambda(s) &\geq \sum_{j=1}^n p_j(s) \exp\left(\frac{1}{\beta(s)} \int_{\sigma(s)}^s \lambda(r) dr \beta(s)\right) \\ &\geq \sum_{i=1}^n p_i(s) \frac{1}{\beta(s)} \left(\int_{\sigma(s)}^s \lambda(r) dr + (1 + \ln(\beta(s)))\right). \end{aligned}$$

Multiplying by $\beta(s)$ and integrating from t_1 to t ,

$$\int_{t_1}^t \lambda(s)\beta(s) ds \geq \int_{t_1}^t \sum_{i=1}^n p_i(s) \int_{\sigma(s)}^s \lambda(r) dr ds + \int_{t_1}^t \sum_{i=1}^n p_i(s)(1 + \ln(\beta(s))) ds.$$

By Lemma 4.1, with $a = t_1$ and $b = t$, we have

$$\begin{aligned} &\int_{t_1}^t \lambda(s)\beta(s) ds \\ &\geq \int_{t_1}^{\sigma(t)} \lambda(s) \int_s^{\sigma_{\text{inv}}(s)} \sum_{i=1}^n p_i(r) dr ds + \int_{t_1}^t \sum_{i=1}^n p_i(s)(1 + \ln(\beta(s))) ds. \end{aligned}$$

Substituting $\beta(s)$ by its value on the left-hand side, and combining integrals, gives

$$\int_t^{\sigma(t)} \lambda(s) \int_s^{\sigma_{\text{inv}}(s)} \sum_{i=1}^n p_i(r) dr ds \geq \int_{t_1}^t \sum_{i=1}^n p_i(s)(1 + \ln(\beta(s))) ds.$$

By Lemma 4.2, the coefficient of $\lambda(s)$ is at most 1. Then

$$\ln\left(\frac{x(\sigma(t))}{x(t)}\right) = \int_t^{\sigma(t)} \lambda(s) ds \geq \int_{t_1}^t \sum_{i=1}^n p_i(s)(1 + \ln(\beta(s))) ds.$$

In the limit as $t \rightarrow \infty$, the right-hand side approaches ∞ because of (4-10). Therefore, $\lim_{t \rightarrow \infty} x(\sigma(t))/x(t) = \infty$, which contradicts Lemma 4.3. This shows that the solution cannot be eventually positive.

When x is eventually negative, as in Lemma 1.1, we obtain a $t_1 \geq t_0$ such that $x(t) < 0$, $x(\tau_i(t)) < 0$, x is nondecreasing, and $|x(\tau_i(t))| \leq |g_i(x(\tau_i(t)))|$ for $t \geq t_1$. Lemma 4.1 holds; it is independent of x . Lemma 4.2 holds, but in its proof we need to reverse the inequalities in (4-3) and (4-4). Lemma 4.3 holds, but in its proof we need to reverse the inequalities in (4-6), (4-7) and (4-8). In the first part of this proof, we need to reverse inequality (4-10). Again, we reach a contradiction indicating that the solution cannot be eventually negative. \square

Remark 4.5. If $t - \tau_i(t) = k_i$, a positive constant, then (3-1) implies (4-10). In general, conditions (2-1), (3-1) and (4-10) are independent of each other. Here we present an example where (4-10) is satisfied, but (2-1) and (3-1) are not satisfied.

Consider (1-1) with only one delay, $f_1(t, x) = p_1(t)x$, $\tau_1(t) = t - \frac{1}{e}$, and

$$p_1(t) = \begin{cases} 4et & \text{if } 0 \leq t \leq \frac{1}{2e}, \\ 2 & \text{if } \frac{1}{2e} < t < 1 - \frac{1}{2e}, \\ -4e(t - 1) & \text{if } 1 - \frac{1}{2e} \leq t \leq 1. \end{cases}$$

For $t \geq 1$, extend p_1 with period 1. Then

$$\frac{1}{e} \leq \int_s^{s+\frac{1}{e}} p_1(r) dr \leq \frac{2}{e}.$$

Note that the lower bound is attained when s is an integer minus $\frac{1}{2e}$; therefore

$$\liminf_{t \rightarrow \infty} \int_{t-\frac{1}{e}}^t p_1(r) dr = \frac{1}{e},$$

and (2-1) is not satisfied. The upper bound is attained when s equals an integer plus $\frac{1}{2e}$; thus

$$\limsup_{t \rightarrow \infty} \int_{t-\frac{1}{e}}^t p_1(r) dr = \frac{2}{e} < 1,$$

and (3-1) is not satisfied. Condition (4-10) is satisfied, because

$$\int_0^\infty p_1(s) \left(1 + \ln \left(\int_s^{s+\frac{1}{e}} p_1(r) dr \right) \right) ds \geq \sum_{k=0}^\infty \int_{k+\frac{1}{2e}}^{k+1-\frac{3}{2e}} 2 \left(1 + \ln \frac{2}{e} \right) = \infty.$$

Now we extend the results in [Hunt and Yorke 1984] from the linear to the nonlinear case of equation (1-1). However, the Grönwall and the logarithmic inequalities cannot be applied in this case.

Theorem 4.6. Assume (H1), (H2) and that there exists a constant β such that

$$0 < t - \tau_i(t) \leq \beta \quad \forall t \geq t_0, 1 \leq i \leq n, \tag{4-13}$$

$$\liminf_{t \rightarrow \infty} \sum_{i=1}^n p_i(t)(t - \tau_i(t)) > \frac{1}{e}. \tag{4-14}$$

Then all solutions of (1-1) are oscillatory.

Proof. To reach a contradiction, assume that there is a nonoscillatory solution x , which initially is assumed to be eventually positive. First we show that (4-14)

implies (1-2), which allows us to use Lemma 1.1. From (4-14), there exist $t^* \geq t_0$ such that

$$\sum_{i=1}^n p_i(t)(t - \tau_i(t)) \geq \frac{1}{e}$$

for all $t \geq t^*$. Then by (4-13),

$$\beta \int_{t^*}^{\infty} \sum_{i=1}^n p_i(t) dt \geq \int_{t^*}^{\infty} \sum_{i=1}^n p_i(t)(t - \tau_i(t)) dt \geq \int_{t^*}^{\infty} \frac{1}{e} dt = \infty.$$

Let t_1 be defined by Lemma 1.1.

From (4-14), there exist constants $t_2 \geq t_1$ and α such that

$$\sum_{i=1}^n p_i(t)(t - \tau_i(t)) \geq \alpha > \frac{1}{e} \quad \forall t \geq t_2.$$

Let $y(t) = -\ln(x(t))$. Then $x(t) = \exp(-y(t))$ and from (1-7), we have

$$y'(t) \geq \sum_{i=1}^n p_i(t) \exp(y(t) - y(\tau_i(t))) \quad \forall t \geq t_2. \tag{4-15}$$

As in [Hunt and Yorke 1984], we construct a solution u to a delay differential equation such that $u(t) \leq y(t)$ and u blows up in finite time. Let u be the solution to the delay equation

$$u'(t) = \alpha \inf_{t-\beta \leq r < t} \frac{1}{t-r} \exp(u(t) - u(r)) \quad \forall t \geq t_2 + \beta, \tag{4-16}$$

with the constant initial condition

$$u(t) = u(t_2 + \beta) \leq \min_{t_2 \leq s \leq t_2 + \beta} y(s) \quad \text{for } t \leq t_2 + \beta.$$

The rest of the proof is the same as that of [Hunt and Yorke 1984, Theorem 1]; so we just outline the steps. First justify the existence of the solution to (4-16), and denote by $r(t)$ the value at which the infimum is attained. Then show that u and u' are increasing, and that, $r(t)$, being a minimizer, satisfies either $t - r(t) = 1/u'(t)$ or $(t - \beta) \leq 1/u'(\beta)$ when $r(t) = \beta$. Then construct a recurrence sequence $\{t_n\}$ increasing to a value t^* , while $u(t_n) \rightarrow \infty$. This implies $\lim_{t \rightarrow t^*} -\ln(x(t)) = \infty$ and $x(t^*) = 0$, which contradicts x being eventually positive.

When x is eventually negative, as in Lemma 1.1, we obtain $t_1 \geq t_0$ such that $x(t) < 0$, $x(\tau_i(t)) < 0$, x is nonincreasing, and $|x(\tau_i(t))| \leq |g_i(x(\tau_i(t)))|$ for $t \geq t_1$. We redefine $y(t) = -\ln(-x(t))$; thus $-x(t) = \exp(-y(t))$. From (1-7) with the inequality reversed, we obtain (4-15). The rest of the proof is as for the eventually positive case. □

Remark 4.7. Note that the integral in (3-1) satisfies

$$\int_{\sigma(t)}^t \sum_{i=1}^n p_i(s) ds \leq \sum_{i=1}^n \int_{\tau_i(t)}^t p_i(s) ds,$$

and that for $p_i(t)$ constant, the right-hand side of this inequality is $p_i(t)(t - \tau_i(t))$, which is used in (4-14). Therefore when $p_i(t)$ is constant, (3-1) implies (4-14). When $p_i(t)$ is constant and $\tau_i(t) = \beta$, conditions (3-1) and (4-14) are the same. In general, (4-14) is independent of both (3-1) and (3-3).

The above conditions are only sufficient for the oscillation of all solutions; finding necessary conditions may be a direction for future research.

References

- [Braverman and Karpuz 2011] E. Braverman and B. Karpuz, “On oscillation of differential and difference equations with non-monotone delays”, *Appl. Math. Comput.* **218**:7 (2011), 3880–3887. MR 2012h:34145 Zbl 1256.39013
- [Chatzarakis and Öcalan 2015] G. E. Chatzarakis and Ö. Öcalan, “Oscillation of differential equations with several non-monotonic advanced arguments”, *Dyn. Syst.* **30**:3 (2015), 310–323. MR 3373715 Zbl 06514202
- [Dix et al. 2008] J. G. Dix, N. Misra, L. Padhy, and R. Rath, “Oscillatory and asymptotic behaviour of a neutral differential equation with oscillating coefficients”, *Electron. J. Qual. Theory Differ. Equ.* **2008**:19 (2008), 1–10. MR 2009d:34198 Zbl 1183.34107
- [Elbert and Stavroulakis 1995] Á. Elbert and I. P. Stavroulakis, “Oscillation and nonoscillation criteria for delay differential equations”, *Proc. Amer. Math. Soc.* **123**:5 (1995), 1503–1510. MR 95f:34099 Zbl 0828.34057
- [Erbe et al. 1995] L. H. Erbe, Q. Kong, and B. G. Zhang, *Oscillation theory for functional-differential equations*, Monographs and Textbooks in Pure and Applied Mathematics **190**, Marcel Dekker, New York, 1995. MR 96c:34147 Zbl 0821.34067
- [Fukagai and Kusano 1984] N. Fukagai and T. Kusano, “Oscillation theory of first order functional-differential equations with deviating arguments”, *Ann. Mat. Pura Appl.* (4) **136** (1984), 95–117. MR 86b:34135 Zbl 0552.34062
- [Gil’ 2014] M. I. Gil’, *Stability of neutral functional differential equations*, Atlantis Studies in Differential Equations **3**, Atlantis Press, Paris, 2014. MR 3289984
- [Grammatikopoulos et al. 2003] M. K. Grammatikopoulos, R. Koplatadze, and I. P. Stavroulakis, “On the oscillation of solutions of first order differential equations with retarded arguments”, *Georgian Math. J.* **10**:1 (2003), 63–76. MR 2004c:34199 Zbl 1051.34051
- [Győri and Ladas 1991] I. Győri and G. Ladas, *Oscillation theory of delay differential equations: With applications*, Clarendon Press, Oxford, 1991. MR 93m:34109 Zbl 0780.34048
- [Hale 1977] J. Hale, *Theory of functional differential equations*, 2nd ed., Applied Mathematical Sciences **3**, Springer, New York, 1977. MR 58 #22904 Zbl 0352.34001
- [Hunt and Yorke 1984] B. R. Hunt and J. A. Yorke, “When all solutions of $x' = -\sum q_i(t)x(t - T_i(t))$ oscillate”, *J. Differential Equations* **53**:2 (1984), 139–145. MR 85k:34169 Zbl 0571.34057

- [Ladde et al. 1987] G. S. Ladde, V. Lakshmikantham, and B. G. Zhang, *Oscillation theory of differential equations with deviating arguments*, Monographs and Textbooks in Pure and Applied Mathematics **110**, Marcel Dekker, New York, 1987. MR 90h:34118 Zbl 0832.34071
- [Li 1996] B. Li, “Oscillation of first order delay differential equations”, *Proc. Amer. Math. Soc.* **124**:12 (1996), 3729–3737. MR 97b:34078 Zbl 0865.34057
- [Zhou 2011] Q. Zhou, “Asymptotic behavior of solutions to a first-order non-homogeneous delay differential equation”, *Electron. J. Differential Equations* **2011**:103 (2011), 1–8. MR 2012g:34163 Zbl 1230.34061

Received: 2015-04-08 Revised: 2015-05-22 Accepted: 2015-06-13

james.p.dix@gmail.com

University of Texas, Austin, TX 78703, United States

jd01@txstate.edu

*Department of Mathematics, Texas State University,
601 University Drive, San Marcos, TX 78666, United States*

A variational approach to a generalized elastica problem

C. Alex Safsten and Logan C. Tatham

(Communicated by Frank Morgan)

In this paper, we apply the calculus of variations to solve the elastica problem. We examine a more general elastica problem in which the material under consideration need not be uniformly rigid. Using, the Euler–Lagrange equations, we derive a system of nonlinear differential equations whose solutions are given by these generalized elastica curves. We consider certain simplifying cases in which we can solve the system of differential equations. Finally, we use novel numerical techniques to approach solutions to the problem in full generality.

1. Introduction

Historically, it has been of much interest to find the shape to which a material conforms when it is bent. This is known as the elastica problem, and has been studied by mathematicians including the Bernoullis, Euler, and Laplace [Euler 1786; Levien 2008]. Elastica problems date back to the thirteenth century mathematician Jordanus de Nemore, who mentioned them in his book, *De ratione ponderis* (see [Tartaglia 1565]). The problems were explored analytically by James Bernoulli in the seventeenth century, though he did not utilize variational methods, which is curious, as his brother Johann was instrumental in the development of the calculus of variations [Goldstine 1980].

Today, the elastica problem is still of great interest, as it has applications in many fields, including engineering, animation, and even industrial design. For example, at Brigham Young University, Professor David Morgan bends sheet metal to construct bowls. Professor Morgan employs a trial and error technique to find a shape which, when cut from a metal sheet, bends into an aesthetically pleasing bowl whose base is flush with the surface, as shown in Figure 1. Currently, Professor Morgan and other design professionals who utilize extreme deflections in their designs use a “guess and

MSC2010: primary 49M30; secondary 49S05.

Keywords: calculus of variations, elastica, evolutionary algorithm, paper bending, Jacobi elliptic functions.



Figure 1. Left: An unfolded shape cut from sheet metal. Right: The metal shape folds to a bowl.

check” algorithm to determine how to cut and bend their material. An analytic expression for the bent shapes would facilitate the construction of industrial designers’ art. Deflections are especially of interest to those in the field of engineering known as compliant mechanisms. Compliant mechanisms are devices which exhibit movement through bending the material of the device itself, as opposed to with hinges.

There are two ways of looking at the extreme deflection problem. First, given a material and boundary conditions representing the bending, what shape does it make? Second, given specific boundary conditions, can one vary the rigidity of the material to ensure it will bend to a specified shape. Euler, Bernoulli, and others answered the first of these questions in the case that the material is uniformly rigid, but the answer is the solution to a system of differential equations that is rather difficult to solve. We reproduce their results, and show how to find solutions to these differential equations in a few special cases, using a novel approach to fit boundary conditions. Furthermore, we present an algorithm capable of finding a more general class of solutions. We also explore the consequences for a material that is not uniformly rigid. Finally, we answer the second question and present a closed-form solution for a rigidity function that will make a material conform to the desired shape.

In this paper, we apply the calculus of variations and numerical methods for answering both of these questions. We analyze problems involving a “strip” of material, in which bending varies in direction, as opposed to a “sheet” of material, in which bending varies in several directions. We propose a simple model of the energy stored in a strip of flexible material and utilize this model to approach an analytic answer to these problems.

2. Preliminary

The calculus of variations. The calculus of variations is a theory of optimization. Optimization using variational calculus is analogous to optimization using differentiation in elementary calculus. However, rather than optimizing a function with

respect to a single real variable, optimization methods in the calculus of variations seek to optimize a functional. A functional is a mapping $\mathcal{F} \rightarrow \mathbb{R}$, where \mathcal{F} is a set of functions. Generally, we only consider functions that are n -times continuously differentiable on a closed interval; that is, $\mathcal{F} \subset C^n([a, b])$. The mapping usually takes the form

$$J = \int_a^b \mathcal{F}(x, f(x), f'(x), f''(x), \dots, f^{(n)}(x)) dx, \tag{1}$$

where

- $f \in \mathcal{F}$,
- \mathcal{F} , when treated as a function only of x , is n -times differentiable,
- \mathcal{F} is continuously differentiable in $f, f', \dots, f^{(n)}$,
- J is the value to be optimized with respect to f .

Our primary assumption is that flexible materials in extreme deflection problems form shapes that minimize their stored energy. Therefore extreme deflection problems can be understood by letting f be a function whose plot in the xy -plane gives the shape of the material and

$$J = \int_a^b \mathcal{F}(x, f(x), \dots) dx$$

be a functional that calculates the energy stored in the material from its shape. We endeavor to use the calculus of variations to find the function f that gives the minimum energy J . The following information on the calculus of variations comes from the excellent book [Goldstine 1980], which goes into greater depth for the interested reader.

It can be shown that if an extremum of the functional

$$J = \int_a^b \mathcal{F}(x, f(x), f'(x)) dx$$

occurs at f , then the following identity, known as the *Euler–Lagrange equation*, holds:

$$0 = \frac{\partial \mathcal{F}}{\partial f} - \frac{d}{dx} \frac{\partial \mathcal{F}}{\partial f'}. \tag{2}$$

For the more general functional

$$J = \int_a^b \mathcal{F}(x, f(x), f'(x), f''(x), \dots, f^{(n)}(x)) dx,$$

the Euler–Lagrange equation extends as

$$0 = \frac{\partial \mathcal{F}}{\partial f} - \frac{d}{dx} \frac{\partial \mathcal{F}}{\partial f'} + \frac{d^2}{dx^2} \frac{\partial \mathcal{F}}{\partial f''} - \frac{d^3}{dx^3} \frac{\partial \mathcal{F}}{\partial f'''} + \dots + (-1)^n \frac{d^n}{dx^n} \frac{\partial \mathcal{F}}{\partial f^{(n)}}. \tag{3}$$

If we are optimizing over a set of vector-valued functions

$$f(x) = (f_1(x), \dots, f_m(x)),$$

one can show that the Euler–Lagrange equation must be true for each component function. In our problem, we are examining shapes made by flexible material as functions mapping to \mathbb{R}^2 . We find it easier to consider two independent functions $x = x(t)$ and $y = y(t)$. Our functional thus takes the form,

$$J = \int_a^b \mathcal{F}(t, x(t), y(t), x'(t), y'(t), \dots, x^{(n)}(t), y^{(n)}(t)) dt.$$

The Euler–Lagrange equation becomes a system of equations, where (3) is satisfied for both component functions; that is,

$$\begin{aligned} 0 &= \frac{\partial \mathcal{F}}{\partial x} - \frac{d}{dt} \frac{\partial \mathcal{F}}{\partial x'} + \dots + (-1)^n \frac{d^n}{dt^n} \frac{\partial \mathcal{F}}{\partial x^{(n)}}, \\ 0 &= \frac{\partial \mathcal{F}}{\partial y} - \frac{d}{dt} \frac{\partial \mathcal{F}}{\partial y'} + \dots + (-1)^n \frac{d^n}{dt^n} \frac{\partial \mathcal{F}}{\partial y^{(n)}}. \end{aligned}$$

Note that the Euler–Lagrange equation is a necessary but not sufficient condition for an extremum, similar to solving $f'(x) = 0$ in elementary calculus optimization problems. The condition that gives sufficiency is known as the second variation, but it is a common practice to omit it in variational problems. It is very complex and in this paper, we will not need to address this problem, as it is obvious when solutions are minimizers. We also mention that while the Euler–Lagrange equation is satisfied by the extremum of the functional, it does not identify which local minima and maxima are global minima and maxima. This is similar to calculus in that solving $f'(x) = 0$ or $\nabla f(\mathbf{x}) = \mathbf{0}$ does not identify local extrema as global extrema. The Euler–Lagrange equation gives differential equations; extrema are found by solving these equations subject to boundary conditions.

Constrained optimization. One also may wish to minimize the functional subject to a constraint. For clarity, allow us to draw an analogy. In multivariable calculus, we may wish to optimize the function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ subject to a constraint function $g(\mathbf{x}) = 0$. To do this, we solve the equations

$$\begin{aligned} \nabla f(\mathbf{x}) &= \lambda \nabla g(\mathbf{x}), \\ g(\mathbf{x}) &= 0, \end{aligned} \tag{4}$$

where λ is a constant.

This is known as the method of Lagrange multipliers. In the calculus of variations, one may wish to optimize the functional

$$J = \int_a^b \mathcal{F}(t, x(t), y(t), x'(t), y'(t)) dt \quad \text{subject to} \quad \Phi(x, y) = 0, \tag{5}$$

where $\Phi(x, y)$ is an ordinary function of $(x(t), y(t))$. To do this, we define

$$\mathcal{L} = \mathcal{F} - \lambda(t)\Phi$$

and then solve the system of equations

$$\frac{\partial \mathcal{L}}{\partial x} - \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial x'} = 0, \quad \frac{\partial \mathcal{L}}{\partial y} - \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial y'} = 0, \quad \Phi(x, y) = 0. \tag{6}$$

So that the analogy becomes clear, note that solving (4) is equivalent to setting $L(\mathbf{x}) = f(\mathbf{x}) - \lambda g(\mathbf{x})$ and solving the unconstrained problem

$$\begin{aligned} \nabla L(\mathbf{x}) &= \mathbf{0}, \\ g(\mathbf{x}) &= 0, \end{aligned}$$

which is

$$\frac{\partial L}{\partial x} = 0, \quad \frac{\partial L}{\partial y} = 0, \quad g(\mathbf{x}) = 0.$$

There are other types of constrained optimization problems, such as the *isoparametric problem*

$$J = \int_a^b \mathcal{F}(x, f(x), f''(x)) dx \quad \text{subject to} \quad \int_a^b \sqrt{1 + (f')^2} dx = 1.$$

However, our analysis focuses on constrained problems having the same form as (5).

3. Results

Model. In our problem, the natural state of the system subject to an extreme deflection will be that for which stored energy is minimized relative to other curves that conform to the same boundary conditions. This invites the question: how does one measure the energy in a bent strip of material? The model that we propose is that potential energy stored in an infinitesimal section of deflected material is proportional to the curvature squared. This is the same model used for linear springs and thus it treats the material as if it were constructed from infinitesimal springs throughout. Our model then is

$$dU = \alpha(t)\kappa^2(t) ds, \tag{7}$$

where

- U is potential energy,
- $\alpha(t)$ is a positive, continuously differentiable spring coefficient,
- $\kappa(t)$ is the curvature,
- s is the arc length.

One can think of this as our definition of “flexible material”. For parametric curves in the plane given by $\gamma(t) = (x(t), y(t))$, the curvature $\kappa(t)$ is given by

$$\kappa(t) = \frac{|x'(t)y''(t) - y'(t)x''(t)|}{(x'(t)^2 + y'(t)^2)^{3/2}}. \quad (8)$$

We will plot the shape that the deflected material makes in the xy -plane. We seek parametric equations of the form $x = x(t)$, $y = y(t)$, where $0 \leq t \leq 1$, whose plot represents the shape to which the deflected material conforms. We also introduce the arc-length parametrization constraint

$$(x')^2 + (y')^2 = 1 \quad \text{for all } t \in [0, 1]. \quad (9)$$

Observe that this greatly simplifies the expression for curvature in (8). Because any curve in the xy -plane with derivatives not both vanishing at the same point can be arc-length parametrized, this is not a restrictive assumption. Since integrating (7) gives the potential energy, which we seek to minimize, we are left to solve the problem

$$\min_{(x(t), y(t), \alpha(t))} \int_0^1 \alpha(t) \kappa^2(t) dt. \quad (10)$$

Now that we have a good understanding of the calculus of variations, we may apply it to (10). Specifically, we may apply the constrained Euler–Lagrange condition, equation (6), where¹

$$\mathcal{F} = \alpha \kappa^2 \quad \text{and} \quad \Phi = (x')^2 + (y')^2 - 1 = 0.$$

This will give a system of differential equations, which are necessary conditions for their solutions to be minimizers of (10).

Proposition 3.1. *If $x(t)$ and $y(t)$ are solutions of the general problem*

$$\min_{(x(t), y(t), \alpha(t))} \int_0^1 \kappa^2 dt,$$

then they satisfy

$$c_1 y' + \alpha' x'' y' + \alpha x''' y' = c_2 x' + \alpha' x' y'' + \alpha x' y'''$$

and

$$(x')^2 + (y')^2 = 1$$

on $[0, 1]$ for some constants c_1, c_2 .

Proof. Since our curves are arc-length parametrized, curvature squared reduces to

$$\kappa^2 = (x'')^2 + (y'')^2.$$

¹For convenience, we stop writing explicit dependence on t .

Thus, we may apply the constrained Euler–Lagrange condition, (6), where

$$\begin{aligned} \mathcal{F} &= \alpha\kappa^2 = \alpha(x'')^2 + \alpha(y'')^2, \\ \Phi &= (x')^2 + (y')^2 - 1 = 0. \end{aligned}$$

Applying (6) to the x parameter gives

$$\begin{aligned} 0 &= \frac{d^2}{dt^2} \frac{\partial(\mathcal{F} - \lambda\Phi)}{\partial x''} - \frac{d}{dt} \frac{\partial(\mathcal{F} - \lambda\Phi)}{\partial x'} + \frac{\partial(\mathcal{F} - \lambda\Phi)}{\partial x} \\ &= \lambda'x' + \lambda x'' + \alpha''x'' + 2\alpha'x''' + \alpha x^{(iv)}, \end{aligned}$$

which simplifies to

$$-(\lambda x')' = (\alpha'x'' + \alpha x''')'.$$

Integrating gives

$$-\lambda x' = c_1 + \alpha'x'' + \alpha x'''$$

for some constant c_1 . A similar computation in y yields

$$-\lambda y' = c_2 + \alpha'y'' + \alpha y'''$$

for some constant c_2 .

Thus we must solve

$$c_1 + \alpha'x'' + \alpha x''' = -\lambda x' \quad \text{and} \quad c_2 + \alpha'y'' + \alpha y''' = -\lambda y'.$$

Multiplying the first by y' and the second by x' allows us to eliminate λ :

$$c_1 y' + \alpha'x''y' + \alpha x'''y' = -\lambda x'y' = c_2 x' + \alpha'x'y'' + \alpha x'y'''.$$

Recalling the constraint, we will solve the system

$$\begin{aligned} c_1 y' + \alpha'x''y' + \alpha x'''y' &= c_2 x' + \alpha'x'y'' + \alpha x'y''', \\ (x')^2 + (y')^2 &= 1. \end{aligned} \tag{11}$$

This completes the proof. □

This is a very difficult system of differential equations to solve. Fortunately, we can simplify it significantly. Since $(x')^2(t) + (y')^2(t) = 1$ for all t , we can define a function $\theta(t)$ such that

$$x'(t) = \cos(\theta(t)) \quad \text{and} \quad y'(t) = \sin(\theta(t)).$$

Then

$$\begin{aligned} x''(t) &= -\sin(\theta(t))\theta'(t), \\ x'''(t) &= -\cos(\theta(t))(\theta')^2(t) - \sin(\theta(t))\theta''(t), \\ y''(t) &= \sin(\theta(t))\theta'(t), \\ y'''(t) &= -\sin(\theta(t))(\theta')^2(t) + \cos(\theta(t))\theta''(t). \end{aligned}$$

The interested reader may verify that when applying this substitution to (11), one arrives at the much simpler single differential equation

$$0 = c_2 \sin \theta - c_1 \cos \theta + (\alpha \theta')'. \quad (12)$$

This equation was first derived for $\alpha = 1$ by Gustav Kirchhoff in 1859. He observed that it is equivalent to the equation of motion for a pendulum, which is the basis for a multitude of analogies from elastica curves to pendulum dynamics [Love 1906].

Equation (12) contains two parameters, c_1 and c_2 , and the rigidity function $\alpha(t)$, each of which is completely arbitrary (with the stipulation that α is positive and continuously differentiable). If c_1 and c_2 are both zero, (12) is linear, so we call solutions associated with this case the linear solutions. We can also simplify (12) by assuming a constant rigidity function such that $\alpha'(t) = 0$. We call solutions in this case constant rigidity solutions. We examine solutions in all four cases.

Linear, constant rigidity solutions. The easiest case is the linear, constant rigidity case. In this case, $c_1 = c_2 = \alpha'(t) = 0$. Equation (12) reduces to $\theta''(t) = 0$, the solutions to which are

$$\theta(t) = c_3 t + c_4. \quad (13)$$

If $c_3 = 0$, we get

$$\begin{aligned} x(t) &= \cos(c_4)t + c_5, \\ y(t) &= \sin(c_4)t + c_6. \end{aligned} \quad (14)$$

If $c_3 \neq 0$, then

$$\begin{aligned} x(t) &= \frac{1}{c_3} \sin(c_3 t + c_4) + c_5, \\ y(t) &= -\frac{1}{c_3} \cos(c_3 t + c_4) + c_6. \end{aligned} \quad (15)$$

Thus, after fitting boundary conditions by identifying c_3 , c_4 , c_5 , and c_6 , we find that linear constant rigidity solutions are either line segments as in (14), or arcs of circles as in (15).

Linear, nonconstant rigidity solutions. Also relatively easy to solve is the linear, nonconstant rigidity case. Here, we have $c_1 = c_2 = 0$. Equation (12) reduces to

$$\alpha(t)\theta'(t) = c_3.$$

To have physical meaning, $\alpha(t)$ must be nonnegative. If it is also nonvanishing, we know that $c_3/\alpha(t)$ is continuous, so it is integrable. Thus, by direct integration, we can solve for $\theta(t)$:

$$\theta(t) = c_4 + \int_0^t \frac{c_3}{\alpha(t')} dt'. \quad (16)$$

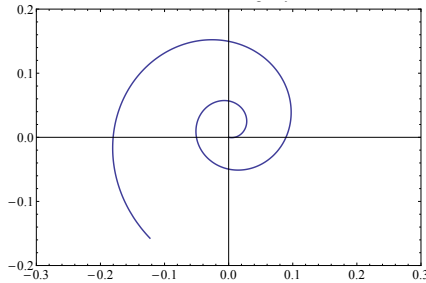


Figure 2. A solution to the linear, nonconstant rigidity case. The rigidity, in this case, is lower at the end near the origin. In regions of lower rigidity, curvature requires less energy, resulting in a spiral which grows tighter with lower rigidity.

This is as far as we can go without specifying $\alpha(t)$. As an example, suppose $\alpha(t) = at + b$, that describes a material which is more rigid on one end than the other. Then,

$$\theta(t) = c_4 + \frac{c_3}{a} \ln(at + b),$$

and

$$\begin{aligned} x'(t) &= \cos\left(c_4 + \frac{c_3}{a} \ln(at + b)\right), \\ y'(t) &= \sin\left(c_4 + \frac{c_3}{a} \ln(at + b)\right). \end{aligned} \tag{17}$$

By integrating once more, we arrive at

$$\begin{aligned} x(t) &= \frac{(at + b)\left(a \sin\left(\frac{c_3 \log(at+b)}{a} + c_4\right) + c_4\right) - c_3 \cos\left(\frac{c_3 \log(at+b)}{a} + c_4\right)}{a^2 + c_3^2} + c_5, \\ y(t) &= \frac{(at + b)\left(c_3 \sin\left(\frac{c_3 \log(at+b)}{a} + c_4\right) + a \cos\left(\frac{c_3 \log(at+b)}{a} + c_4\right)\right)}{a^2 + c_3^2} + c_6. \end{aligned} \tag{18}$$

We have provided a plot of this solution in Figure 2 for $a = 1, b = 0.1$, and $c_3 = c_4 = 5$. The rigidity is lower at the end near the origin. Since that cost of curvature in terms of energy is lower in regions of less rigidity, the spiral is tighter there.

Nonlinear, constant rigidity solutions. In the case of nonlinear, constant rigidity solutions, c_1 and c_2 are not both zero, and $\alpha(t)$ is a constant function. This is perhaps the most interesting case because it is possible to find exact solutions that are far more general than any linear solutions. We first show how to determine solutions by using special functions, and then show how to apply boundary conditions to these solutions. Finally, we present some numerical techniques for approximating solutions when applying boundary conditions proves too difficult. In this section, we suppose $\alpha = 1$.

Jacobi elliptic and amplitude functions. The Jacobi elliptic functions are a class of special functions, sn, cn, and dn (along with others we will not discuss here)[Prasolov and Solovyev 1997]. They are often written as functions of the variable u with respect to a parameter k as $\text{sn}(u, k)$, $\text{cn}(u, k)$, and $\text{dn}(u, k)$. The Jacobi amplitude function, usually given by $\text{am}(u, k)$, is related to the Jacobi elliptic functions by

$$\begin{aligned} \text{sn}(u, k) &= \sin(\text{am}(u, k)), \\ \text{cn}(u, k) &= \cos(\text{am}(u, k)), \\ \text{dn}(u, k) &= \frac{d}{du} \text{am}(u, k) = \sqrt{1 - k^2 \sin^2(\text{am}(u, k))}. \end{aligned}$$

It is easy to take a second derivative to demonstrate

$$\frac{d^2}{du^2} \text{am}(u, k) = -\frac{k^2}{2} \sin(2 \text{am}(u, k)).$$

These functions and their properties will be useful in finding an exact solution to (11).

In order to use Jacobi functions to solve (12), we will apply some transformations. We define $R = \sqrt{c_1^2 + c_2^2}$ and $\phi = -\arctan(c_1/c_2)$ so that (12) can be written as

$$\theta''(t) = -R \sin(\theta(t) + \phi).$$

Next, we make the transformation

$$\tau(t) = \frac{1}{2}(\theta(t) + \phi),$$

which gives

$$\tau''(t) = -\frac{R}{2} \sin(2\tau(t)). \tag{19}$$

Proposition 3.2. *Given R and ϕ , the function*

$$\tau(t) = \text{am}\left(c_3 t + c_4, \frac{\sqrt{R}}{c_3}\right) \tag{20}$$

is a solution to (19), where c_3 and c_4 are constants.

Proof. Let $u = c_3 t + c_4$. Then letting $\tau(t)$ be defined as in (20), we have

$$\tau''(t) = \frac{d^2\tau}{dt^2} = \frac{d^2\tau}{du^2} \left(\frac{du}{dt}\right)^2 = -c_3^2 \frac{R}{2c_3^2} \sin 2\tau(t) = -\frac{R}{2} \sin(2\tau(t)). \quad \square$$

Working backwards through our substitutions, we can see that solutions for $x'(t)$ and $y'(t)$ are

$$\begin{aligned} x'(t) &= \cos\left(2 \text{am}\left(c_3(t + c_4), \frac{\sqrt{R}}{c_3}\right) - \phi\right), \\ y'(t) &= \sin\left(2 \text{am}\left(c_3(t + c_4), \frac{\sqrt{R}}{c_3}\right) - \phi\right). \end{aligned} \tag{21}$$

Remark 3.3. Using trigonometric identities, it is not hard to also express these in terms of sn and cn. However, the expressions are very long and not very enlightening, so we do not present them here.

Though it may be possible to find closed-form expressions for $x(t)$ and $y(t)$, we will not find it necessary to do so. We simply define the solutions as

$$\begin{aligned} x(t) &= c_5 + \int_0^t \cos\left(2 \operatorname{am}\left(c_3(t' + c_4), \frac{\sqrt{R}}{c_3}\right) - \phi\right) dt', \\ y(t) &= c_6 + \int_0^t \sin\left(2 \operatorname{am}\left(c_3(t' + c_4), \frac{\sqrt{R}}{c_3}\right) - \phi\right) dt'. \end{aligned} \tag{22}$$

Fitting boundary conditions. Now we have found the solutions, but it can still be very difficult to fit boundary conditions. At this point, we find it easiest to apply computational techniques to find a solution. For example, one of the simplest nontrivial solutions we may wish to find is the shape formed when one bends a strip of paper holding both ends together, as if to fold the paper in half, but not creasing the paper. We call this a “teardrop” shape. Such a shape has the following boundary conditions:

$$x(0) = 0, \quad y(0) = 0, \quad x'(0) = 1. \tag{23}$$

We also apply natural symmetries of the problem,

$$\begin{aligned} x(t) &= x(1 - t), \\ y(t) &= -y(1 - t), \end{aligned} \tag{24}$$

from which we obtain boundary conditions for $x(1)$, $y(1)$, and $x'(1)$. Furthermore, the $y'(0)$ and $y'(1)$ boundary conditions are determined by the arc-length constraint. We have six free variables. Four of these are apparent in (21), namely R , ϕ , c_3 , and c_4 . The other two are constants of integration that come from integrating (21), which we call c_5 and c_6 . We will usually set both of these to zero so that one end of our solution is at the origin. In the following claim, we argue that $\phi = 0$.

Claim 3.4. *In the equations*

$$\begin{aligned} x'''y' + c_1y' &= x'y''' + c_2x', \\ (x')^2 + (y')^2 &= 1, \end{aligned}$$

with symmetries (24), we have $c_1 = 0$.

Proof. This will use some facts of even and odd functions. First, the product of any two functions with the same parity is even and the product of two functions with opposite parity is odd. Furthermore, the sets of even and odd functions are both closed under addition and scalar multiplication.

Now, we note that odd order derivatives of $x(t)$ and $y(t)$ are odd and even (respectively) about $t = \frac{1}{2}$. One can write (11) with $\alpha = 1$ as

$$c_1y' = x'y''' - x'''y' + c_2x'.$$

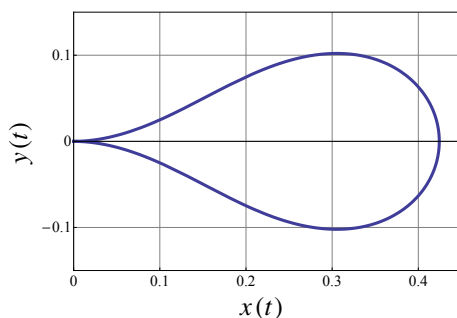


Figure 3. A solution to the teardrop problem, in which the ends of a strip of paper are pinched together.

On one hand, since y' is even, the left-hand side is even. On the other hand, since x', x''' are odd and y', y''' are even, all terms on the right side are odd. Because odd functions are closed under subtraction, the right-hand side is odd. Thus, $c_1 y$ is both even and odd, meaning

$$c_1 y'(t) \equiv 0 \quad \text{for all } t.$$

Since $y(t)$ is not a constant function, one may conclude that $c_1 = 0$. \square

Since $c_1 = 0$, and $\phi = -\arctan(c_1/c_2)$, we have $\phi = 0$.

We now present our process for finding a solution to the “teardrop problem”.

(1) With $c_1 = 0$, we know $\phi = 0$, and $R = c_2$. We make a guess as to the value of c_2 . After several trials, we find that $c_2 = 137$ gives reasonable results. We will later change the value of R to be exact.

(2) The functions given by (21) are periodic, and represent the derivatives of x and y . Two conditions that must be met are $x(1) = x(0)$ and $y(1) = y(0)$. We find the period T such that $x(t) = x(t \pm T)$ and $y(t) = y(t \pm T)$. We find a value for c_3 such that $\int_0^T y'(t) dt = 0$ via numeric integration. (The numeric integrator is not exact, but is correct to five decimal places).

(3) The constant c_4 merely shifts $x'(t)$ and $y'(t)$ in time. We use a root finder to find a value for c_4 such that $x'(0) = 1$ and $y'(0) = 0$.

(4) We finally fit all the boundary conditions by rescaling the parameter c_3 so that the period is 1. We do this by making the substitution $c_3 \rightarrow c_3/T$ and $R \rightarrow R/T^2$ so that the value of \sqrt{R}/c_3 is unchanged.

(5) Finally, we plot the result, as shown in Figure 3.

Through a similar process to the steps above, but with different boundary conditions, we arrive at a solution for the “bump problem”, where the ends of a flexible material on a flat surface are pushed in, keeping the ends flat. This solution is shown in Figure 4. Table 1 gives the values of the various constants for both solutions. We

Curve	Bound. cond. (goal)	Bound. cond. (actual)	Parameter values
teardrop, Figure 3	$x(0) = 0$	$x(0) = 0$	$R = 54.2257$ $\phi = 0$ $c_3 = 5.38408$ $c_4 = 0.174661$ $c_5 = 0$ $c_6 = 0$
	$x'(0) = 1$	$x'(0) = 1$	
	$x(1) = 0$	$x(1) = 0$	
	$x'(1) = -1$	$x(1) = -1$	
	$y(0) = 0$	$y(0) = 0$	
	$y'(0) = 0$	$y'(0) = 0$	
	$y(1) = 0$	$y(1) = -2.676 \times 10^{-5}$	
	$y'(1) = 0$	$y'(1) = 0$	
bump, Figure 4	$x(0) = 0$	$x(0) = 0$	$R = 209.804$ $\phi = 0$ $c_3 = 3.06405$ $c_4 = 0$ $c_5 = 0$ $c_6 = 0$
	$x'(0) = 1$	$x'(0) = 1$	
	$x(1) = .7822$	$x(1) = .7822$	
	$x'(1) = 1$	$x(1) = 1$	
	$y(0) = 0$	$y(0) = 0$	
	$y'(0) = 0$	$y'(0) = 0$	
	$y(1) = 0$	$y(1) = 0$	
	$y'(1) = 0$	$y'(1) = 0$	

Table 1. The constants derived in order to force (22) to meet the specified boundary conditions.

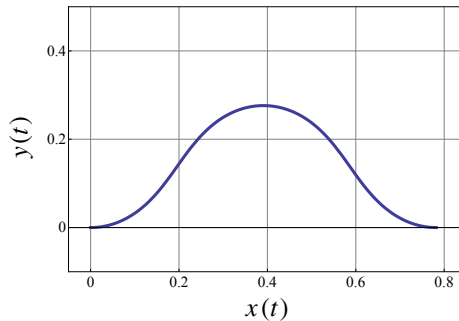


Figure 4. A solution to the bump problem, in which the ends of a strip of a flexible material are pushed together along a flat surface.

do not assert that the teardrop or bump solutions we found here match boundary conditions exactly, nor do we insist that these are unique. We merely emphasize that they are exact solutions to (11) which *nearly* match specified boundary conditions.

These problems lend themselves to analytic solutions, but others are not so easy. To solve more difficult problems, we found standard numeric differential equation solving techniques to be fruitless. But, in harmony with the variational nature of this problem, there is an optimization technique we can apply: evolutionary algorithms.

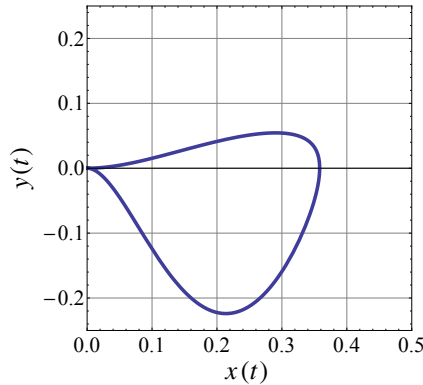


Figure 5. A curve generated by random-coefficient polynomials with boundary conditions enforced.

Evolutionary algorithms. As the name suggests, evolutionary algorithms utilize the same principles as evolutionary biology for selective optimization. The purpose of our evolutionary algorithm is to find a polynomial approximation to a shape that satisfies given boundary conditions and minimizes stored energy.

We assume that the solutions to (11) are analytic, and therefore have a Taylor series expansion. Then, we generate a fixed number N of (uniformly distributed) random-coefficient polynomial pairs, called a *generation*. Each pair forms a parametric curve that satisfies our boundary conditions and has length 1. (See Figure 5). We do not, however, require the resulting parametric curve to have constant arc-length. This is because the arc-length constraint is impossible to enforce with nonlinear polynomials, and the image generated by a pair of polynomials can also be traced out by a constant arc-length function.

We then arrange the N members of the generation, ordering the members by their stored energy calculated via numeric integration of (7). The worst 90% (meaning those with the highest energy) are then removed from the population. Only the most fit members of a generation remain. Here, “most fit” is deemed to mean those with the least stored energy. We introduce a mutation factor by perturbing the coefficients of the remaining polynomials by multiplying them by a randomly determined factor. Finally, we “breed” these polynomials to generate a new generation of N curves. The breeding process consists of randomly selecting two distinct curves and constructing a new polynomial by taking a randomly weighted average of every pair of coefficients. We also constrain the new curve to satisfy the specified boundary conditions and normalize its arc-length.

The members of the new generation have characteristics of their predecessors. Since only the lowest-energy curves are selected from each generation, generally the average energy per curve from each generation is no greater than that of the previous generation. Thus, repeating the process for many generations gives us increasingly

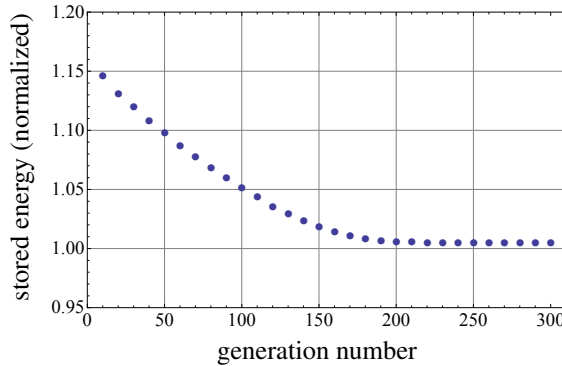


Figure 6. The energy of the best curve in the population for each generation. The scale of the energies is normalized so that the energy of the exact solution is 1.

accurate approximations to the lowest-energy curve conforming to our specifications. Figure 6 shows how the energy of the best curve in the population decreases with each generation. While the evolutionary algorithm does not provide an exact solution to our differential equation, it does provide a polynomial approximation to the image of a curve that does solve the differential equation. As an example, we have applied the evolutionary algorithm to the teardrop problem, and obtained the approximation in Figure 7 after more than 1000 generations. We can calculate the stored energy in this approximation by integrating (7) and compare it to the stored energy in the exact solution, as presented in Table 2.

The polynomials to obtain the best approximation curve are given by

$$x(t) = \sum_{n=0}^6 a_n t^n, \quad y(t) = \sum_{n=0}^9 b_n t^n.$$

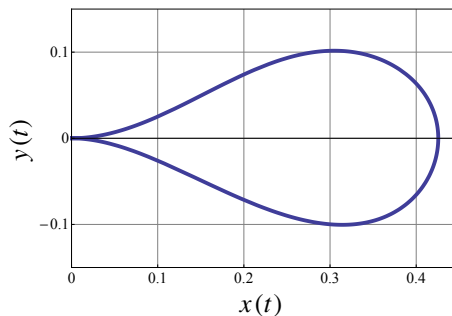


Figure 7. An approximate solution to the teardrop problem.

Curve	Figure	Stored energy (normalized)
random curve	5	2.2781
best approximation	7	1.0054
exact solution	3	1

Table 2. The total stored energy of the exact solution is slightly lower than the stored energy of the best approximation. We also include the energy of the random curve to show how close the approximation is.

The coefficients a_i and b_i are

$$\begin{aligned}
 a_0 &= 0, & a_1 &= 1.34525, & a_2 &= 0.184519, & a_3 &= -2.8019, \\
 a_4 &= -0.568997, & a_5 &= 4.0875, & a_6 &= -2.24638, \\
 b_0 &= 0, & b_1 &= 0, & b_2 &= -6.64108, & b_3 &= 28.9681, \\
 b_4 &= -36.8535, & b_5 &= 3.26174, & b_6 &= 1.8281, & b_7 &= 60.7485, \\
 b_8 &= -83.0815, & b_9 &= 31.7697.
 \end{aligned}$$

Nonlinear, nonconstant rigidity solutions. While we can find nonlinear solutions in the constant rigidity case, we have been unable to find exact solutions in the nonconstant rigidity case. Nevertheless, we can find some interesting numerical solutions. The evolutionary algorithm abandons the arc-length constraint, while the rigidity function relies on arc-length parametrization. To use the evolutionary algorithm, therefore, requires a reparametrization of $\alpha(t)$, which is a hard problem in general. Alternatively, we could calculate the value of $\alpha(t)$ for every iteration of the numeric integrator, but such a process is computationally slow. Therefore, we are content with solving (12) subject to initial conditions, with no guarantee as to the resulting boundary conditions. We provide two examples. Figure 8 shows a nonconstant rigidity function that takes the form of an inverted Gaussian curve.

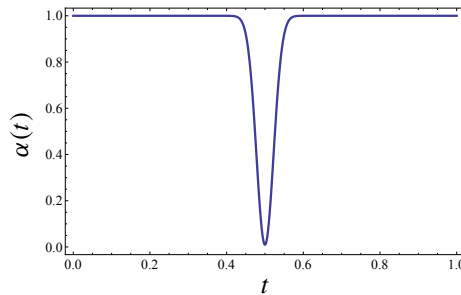


Figure 8. A rigidity function $\alpha(t)$ designed to model a crease in the material.

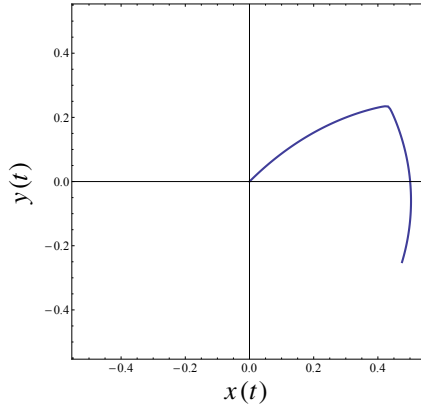


Figure 9. A solution to (12) with respect to the rigidity function shown in Figure 8.

This is meant to model a material with a crease. The resulting solution is given in Figure 9. Figure 10 shows a nonconstant rigidity function that takes the form of a hyperbolic tangent curve. This is meant to model a material that is loose on one end and stiff on the other, with a very sharp transition (imagine a piece of rubber attached to a steel bar). The resulting solution is given in Figure 11.

Finding rigidity. We now consider the second question that was posed in the introduction: given specific boundary conditions, can one vary the rigidity of the material to ensure it will bend to a specified shape? We may rewrite (12) as

$$(\alpha\theta')' = c_1 \cos \theta - c_2 \sin \theta = c_1x' + c_2y',$$

where c_1 and c_2 may be chosen freely (as long as they are both nonzero). We may integrate and solve for α as

$$\alpha = \frac{1}{\theta'}(c_1x - c_2y). \tag{25}$$

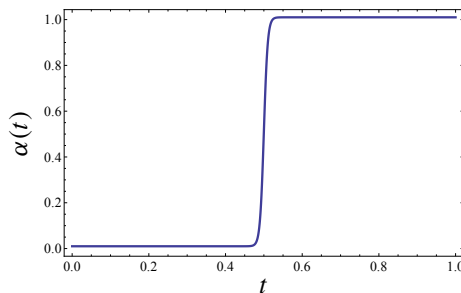


Figure 10. A rigidity function $\alpha(t)$ deigned to model a material with one loose end and one stiff end.

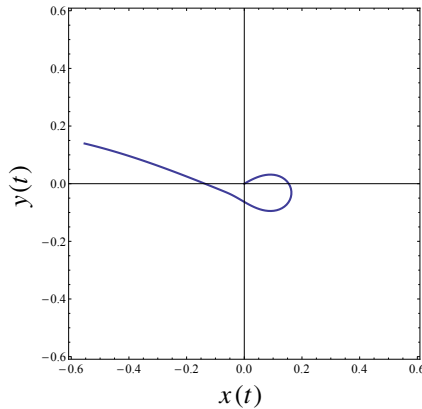


Figure 11. A solution to (12) with respect to the rigidity function shown in Figure 10.

Recall that $\theta(t)$ is defined to be the function satisfying

$$x'(t) = \cos(\theta(t)) \quad \text{and} \quad y'(t) = \sin(\theta(t)).$$

But taking derivatives of these equations yields

$$x'' = -\theta' \sin \theta = -\theta' y' \quad \text{and} \quad y'' = \theta' \cos \theta = \theta' x'$$

so that ²

$$\theta' = -\frac{x''}{y'} = \frac{y''}{x'}.$$

Substituting this back into (25) gives the nice identity³

$$\alpha(t) = \frac{x'(t)}{y''(t)}(c_1 x(t) - c_2 y(t)) = -\frac{y'(t)}{x''(t)}(c_1 x(t) - c_2 y(t)). \quad (26)$$

Thus, if one wants a material with shape given by $(x(t), y(t))$, equation (26) gives the variable rigidity function that will make the material conform to that shape.⁴ This method, of course, may not work if $x''(t) = 0$ or $y''(t) = 0$ at a point. A careful choice of c_1 and c_2 may mitigate this problem in some cases.

4. Conclusion

In this paper, we have explored analytic and numeric methods for solving the extreme deflection problem on a strip of flexible material. We used the calculus

²There is consistency between this equation and differentiating the arc-length constraint $(x')^2 + (y')^2 = 1$.

³This may be recast into several forms, but these are the simplest.

⁴Note that this equation is only valid if $\alpha(t) > 0$.

of variations to derive a system of differential equations (11), which has solutions in several forms. In particular, we confirmed results from Euler and others by deriving (21). We provided a technique for applying boundary conditions to these solutions. Often, fitting boundary conditions is difficult, so we have also explored numeric methods for approximating solutions. We also solved for what rigidity function a material would need in order for it to conform to a given shape.

Next steps. There is still much research that can be done on the elastica problem. Currently, we are only able to solve (11) subject to boundary conditions to numerical accuracy. Is there a way to find an exact solution for general boundary conditions? Can we prove that these solutions are unique?

One could also explore the consequences of varying the bending in another direction. That is, rather than examining a strip of flexible material bent in one direction, one could examine a sheet bent in two directions. The curvature for these surfaces is much more complicated, but the applications to solutions for this problem are more abundant.

References

- [Euler 1786] L. Euler, “Methodus facilis omnia symptomata linearum curvarum non in eodem plano sitarum investigandi”, *Acta Acad. Sci. Imp. Petro.* **1782**:1 (1786), 19–57.
- [Goldstine 1980] H. H. Goldstine, *A history of the calculus of variations from the 17th through the 19th century*, Studies in the History of Mathematics and Physical Sciences **5**, Springer, New York, 1980. MR 601774 Zbl 0452.49002
- [Levien 2008] R. Levien, “The elastica: a mathematical history”, technical report, University of California, Berkeley, 2008, available at <http://www.eecs.berkeley.edu/Pubs/TechRpts/2008/EECS-2008-103.pdf>.
- [Love 1906] A. E. H. Love, *A treatise on the mathematical theory of elasticity*, 2nd ed., Cambridge University Press, 1906.
- [Prasolov and Solov'yev 1997] V. Prasolov and Y. Solov'yev, *Elliptic functions and elliptic integrals*, Translations of Mathematical Monographs **170**, American Mathematical Society, Providence, RI, 1997. MR 1469740 Zbl 0946.11001
- [Tartaglia 1565] N. Tartaglia, *De ratione ponderis*, Venice, 1565.

Received: 2015-04-23 Revised: 2015-06-24 Accepted: 2015-07-01

alex.safsten@gmail.com *Mathematics Department, Brigham Young University,
295 TMCB, Provo, UT 84602, United States*

logantatham@gmail.com *Mathematics Department, Brigham Young University,
295 TMCB, Provo, UT 84602, United States*

When is a subgroup of a ring an ideal?

Sunil K. Chebolu and Christina L. Henry

(Communicated by Kenneth S. Berenhaut)

Let R be a commutative ring. When is a subgroup of $(R, +)$ an ideal of R ? We investigate this problem for the rings \mathbb{Z}^d and $\prod_{i=1}^d \mathbb{Z}_{n_i}$. In the cases of $\mathbb{Z} \times \mathbb{Z}$ and $\mathbb{Z}_n \times \mathbb{Z}_m$, our results give, for any given subgroup of these rings, a computable criterion for the problem under consideration. We also compute the probability that a randomly chosen subgroup from $\mathbb{Z}_n \times \mathbb{Z}_m$ is an ideal.

1. Introduction

Let R be a commutative ring. The object of this paper is to determine necessary and sufficient conditions for a given subgroup of $(R, +)$ to be an ideal of R . Our motivation for asking this question arose from some problems on Mathieu subspaces (more is explained in the next paragraph). To begin, consider the ring \mathbb{Z} of integers. Every subgroup of \mathbb{Z} is of the form $k\mathbb{Z}$ for some integer k , and each of these subgroups is clearly also an ideal. In fact, the same is true also for the ring \mathbb{Z}_n (the ring of integers modulo n). It turns out that these are the only rings R in which every subgroup of $(R, +)$ is also an ideal of R ; see Proposition 2.1. In particular, when we consider product rings, we get some subgroups that are not ideals. For instance, the diagonal $\{(x, x) \mid x \in \mathbb{Z}\}$ in $\mathbb{Z} \times \mathbb{Z}$ is clearly a subgroup of $(\mathbb{Z} \times \mathbb{Z}, +)$ but not an ideal in the ring $\mathbb{Z} \times \mathbb{Z}$. In this paper, we consider the product rings \mathbb{Z}^d (in Section 3) and $\prod_{i=1}^d \mathbb{Z}_{n_i}$ (in Section 4), and for various subgroups of these rings, we give necessary and sufficient conditions for a given subgroup to be an ideal. In the cases of $\mathbb{Z} \times \mathbb{Z}$ and $\mathbb{Z}_n \times \mathbb{Z}_m$, our necessary and sufficient conditions are also computable for any given subgroup of these rings. As one would expect, our results show that in general an arbitrary subgroup of a ring is seldom an ideal. In fact, we make this statement precise in Theorem 5.4, where we compute explicitly the probability that a randomly chosen subgroup from $\mathbb{Z}_n \times \mathbb{Z}_m$ is an ideal. For instance, when p is a prime and the ring is $\mathbb{Z}_p \times \mathbb{Z}_p$, this probability is only $4/(p+3)$. We will use several basic facts and tools from abstract algebra, which can be found in

MSC2010: primary 13AXX; secondary 20KXX.

Keywords: ring, subgroup, ideal, Mathieu subspace, Goursat.

Chebolu is supported by an NSA grant (H98230-13-1-0238).

[Dummit and Foote 2004]. We also use a theorem in group theory due to Goursat; a good exposition of this theorem can be found in [Petrillo 2011], and we review it in Theorem 4.4. Although we focus mainly on the rings $\mathbb{Z} \times \mathbb{Z}$ and $\mathbb{Z}_n \times \mathbb{Z}_m$, where possible we offer some generalizations. By a subgroup of a ring R , we always mean a subgroup of the additive group $(R, +)$.

This problem came up naturally when Chebolu and his collaborators (Yamskulna and Zhao) were recently working on some problems involving Mathieu subspaces in some rings. A Mathieu subspace is a generalization of an ideal: for a commutative ring R , a \mathbb{Z} -submodule M of R is said to be a Mathieu subspace of R if whenever a^n belongs to M (for all $n \geq 1$), then ra^n belongs to M for all n sufficiently large. Every ideal is a Mathieu subspace, but the converse is not necessarily true. The notion of a Mathieu subspace was introduced by Wenhua Zhao [2010], and it proved to be a central idea in the research on several landmark conjectures in algebra and geometry, including the Jacobian conjecture. As a result, Mathieu subspaces received serious attention and extensive writing; see [Zhao 2012] and the references therein. Chebolu and his collaborators were led to the problem of determining when a subgroup of a ring is a Mathieu subspace. Since ideals are important and relatively well-understood classes of Mathieu subspaces, it was natural to investigate the same question for ideals. Thus the problem we study in this paper is an interesting offshoot of our Mathieu subspaces project.

2. Generators

In the introduction, we noted that the rings \mathbb{Z} and \mathbb{Z}_n have the property that every subgroup in them is also an ideal. It is not hard to show that these are the only rings with this property.

Proposition 2.1. *Let R be a unital commutative ring, i.e., a commutative ring with a multiplicative identity. If every subgroup of $(R, +)$ is also an ideal, then R is isomorphic to either \mathbb{Z} or \mathbb{Z}_n for some positive integer n .*

Proof. Since R is a unital ring, there is a natural map $\phi: \mathbb{Z} \rightarrow R$ that sends 1 to 1_R , the multiplicative identity of R . The image of this homomorphism is exactly the subgroup of $(R, +)$ that is generated by 1_R . If every subgroup of $(R, +)$ is an ideal, then, in particular, the subgroup generated by 1_R is also an ideal. However, the only ideal that contains 1_R is the entire ring R . This means ϕ is surjective. From the first isomorphism theorem, we have $\mathbb{Z}/\ker \phi \cong R$. It follows that R is isomorphic to \mathbb{Z} or \mathbb{Z}_n for some integer n . (In the former case, R has characteristic 0, and in the latter, R has characteristic n .) \square

We will now show that every subgroup of \mathbb{Z}^d or $\prod_{i=1}^d \mathbb{Z}_{n_i}$ is generated by at most d elements. We will recall some standard results from abstract algebra, which can be found in [Dummit and Foote 2004].

Theorem 2.2. *Let R be a PID and let M be a free R -module of rank r . Then every submodule of M is also free and has rank at most r .*

This theorem takes care of \mathbb{Z}^d . For $\prod_{i=1}^d \mathbb{Z}_{n_i}$, we need the following corollary, which can be derived easily from the above theorem.

Corollary 2.3. *Let R be a PID and let M be a finitely generated R -module. If M is generated by r elements, then every submodule of M is generated by at most r elements.*

Corollary 2.4. *Every subgroup of $(\prod_{i=1}^d \mathbb{Z}_{n_i}, +)$ or of $(\mathbb{Z}^d, +)$ is generated by at most d elements.*

Proof. The ring $\prod_{i=1}^d \mathbb{Z}_{n_i}$ is a \mathbb{Z} -module that is clearly generated by d elements; the standard basis forms a generating set. Therefore by the above corollary, every subgroup of $\prod_{i=1}^d \mathbb{Z}_{n_i}$ is generated by at most d elements. The corresponding statement for \mathbb{Z}^d is a special case of the above theorem. \square

This corollary gives a natural stratification of the class of all nonsubgroups of these rings, which is based on the minimal number of generators of a given subgroup. This stratification will be helpful in our analysis.

3. The ring $\mathbb{Z} \times \mathbb{Z}$

In this section, we determine when a given additive subgroup of the ring \mathbb{Z}^d is an ideal. The trivial subgroup, which consists of the single element $(0, 0, \dots, 0)$, is also trivially an ideal, so we will consider nonzero subgroups. As explained in the previous section, a nonzero subgroup of \mathbb{Z}^d is free of rank at most d . We will begin with rank-1 subgroups, where the problem is straightforward.

Proposition 3.1. *Let L be a subgroup of \mathbb{Z}^d generated by (a_1, \dots, a_d) . Then L is an ideal if and only if all but one of the a_i are zero.*

Proof. If all but one of the a_i are zero, then L is clearly an ideal in one of the factors of \mathbb{Z}^d . On the other hand, if we have more than one nonzero a_i , say a_i and a_j , then consider $e_i = (0, \dots, 0, 1, 0, \dots, 0)$, which has 1 at the i -th spot. If L is an ideal, then $e_i(a_1, \dots, a_d) = (0, \dots, 0, a_i, 0, \dots, 0)$ should belong to L . This is a contradiction, so we are done. \square

More generally, the following is true.

Lemma 3.2. *Let R be an integral domain. A subgroup of $(R, +)$ generated by a nonzero element a is an ideal of R if and only if R is isomorphic to \mathbb{Z} or \mathbb{Z}_p for some prime p .*

Proof. Let $\langle a \rangle$ be the additive subgroup of $(R, +)$ generated by a ($\neq 0$). Let r be an arbitrary element of R . If $\langle a \rangle$ is an ideal, then we should have $ra = na$ for some integer n . This equation implies that $(r - n1_R)a = 0$. Since we are working in an integral

domain and a is nonzero, we get $r - n1_R = 0$, or $r = n1_R$. Since r is arbitrary, this implies that $(R, +)$ is a cyclic group generated by 1_R . This means R is isomorphic to \mathbb{Z} or \mathbb{Z}_n for some n . But since R is an integral domain, n has to be a prime. \square

Now we move on to subgroups of rank at least 2 in \mathbb{Z}^d , where the problem is more interesting. We begin with an example to show the subtlety in the problem.

Example 3.3. Consider the ring $\mathbb{Z} \times \mathbb{Z}$ and let S and T denote the following rank-2 subgroups of $(\mathbb{Z} \times \mathbb{Z}, +)$:

$$S = \langle (2, 0), (3, 1) \rangle,$$

$$T = \langle (2, 0), (2, 1) \rangle.$$

We claim that S is not an ideal but T is. If S is an ideal, then the element $(0, 1)$ ($= (0, 1)(3, 1)$) should belong to it. That means the pair of equations $2x + 3y = 0$ and $y = 1$ have to be consistent over \mathbb{Z} . However, it is easy to see that this is not the case. On the other hand, T is an ideal in $\mathbb{Z} \times \mathbb{Z}$. In fact, $T = 2\mathbb{Z} \times \mathbb{Z}$. See Theorem 3.8 for the general result.

We begin by classifying ideals of \mathbb{Z}^d whose additive groups are free of rank k .

Proposition 3.4. *Let I be an ideal in \mathbb{Z}^d . Then I is free of rank k ($1 \leq k \leq n$) if and only if I is of the form $\prod_{i=1}^d d_i \mathbb{Z}$, where exactly k of the numbers d_i are nonzero.*

Proof. Recall that every ideal in \mathbb{Z}^d is of the form $\prod_{i=1}^d d_i \mathbb{Z}$, where the d_i are integers. The rank of $\prod_{i=1}^d d_i \mathbb{Z}$ is exactly the number of d_i that are nonzero. \square

In view of this proposition, to determine when a subgroup of rank k in \mathbb{Z}^d is an ideal, it is enough (after deleting the zero coordinates) to consider the problem when $d = k$. The latter is addressed in the next two theorems. We begin with a lemma that we will need in these theorems. Recall that an integer matrix A is said to be unimodular if it is invertible over the ring of integers. This statement is equivalent (as can be seen by Cramer's formula for the inverse) to saying that the determinant of A is either 1 or -1 . In the following lemma, a subgroup of \mathbb{Z}^n of rank n will be called a lattice of \mathbb{Z}^n .

Lemma 3.5. *Let A and B be two $n \times n$ matrices over the integers that are invertible over the rationals. The columns of A and those of B form two bases for a lattice L if and only if there exists a unimodular matrix X such that $AX = B$.*

Proof. Since the columns of A and B form a basis for L , there exist integer square matrices X and Y such that $AX = B$ and $BY = A$. Multiplying the first equation on the right-hand side by Y , we get $AXY = BY$. But $BY = A$, so we get $AXY = A$. Since A is invertible over the rationals, we multiply the inverse (over the rationals) of A on both sides to conclude that $XY = I$. This means X is invertible over \mathbb{Z} (i.e, it is unimodular) and $AX = B$. For the other direction, let Y be the inverse

of X over \mathbb{Z} , so we have $AX = B$ and $BY = A$. The first equation tells us that the column space of B is contained in that of A , and the second equation says that the column space of A is contained in that of B . \square

Theorem 3.6. *Let H be a subgroup of rank k in \mathbb{Z}^k . Let the columns of a $k \times k$ matrix A be a \mathbb{Z} -basis for H . Then the following are equivalent:*

- (1) H is an ideal in \mathbb{Z}^k .
- (2) There exists a unimodular matrix U such that AU is a diagonal matrix.
- (3) There is a sequence of elementary row operations (over \mathbb{Z}) that can convert A into a diagonal matrix.

Proof. Let H (as in the statement of the theorem) be an ideal in \mathbb{Z}^k . Then by Proposition 3.4, H is of the form $\prod_{i=1}^k d_i \mathbb{Z}$ for some integers d_i . Since H has rank k , all these integers have to be nonzero. H can be written in this form if and only if the columns of A and those of the diagonal matrix $D = \text{Diagonal}(d_1, \dots, d_k)$ form a basis for H . By the above lemma, this happens if and only if there is a unimodular matrix U such that $AU = D$. Hence we have the equivalence of statements (1) and (2). The equivalence of (2) and (3) for the field of real numbers is well-known (the famous reduced row echelon form of an invertible matrix). The reader can verify that the proof works over \mathbb{Z} when properly interpreted. For instance, the role played by nonzero real numbers in the world of \mathbb{Z} are the units ± 1 . This gives the equivalence of statements (2) and (3). \square

Since \mathbb{Z} is a Euclidean domain where we can talk about gcds, we can take the above theorem one step further. Let A^* denote the adjoint matrix of A . Recall that the formula for the inverse of A (an invertible matrix) is given by

$$A^{-1} = \frac{1}{\det(A)} A^* = \frac{1}{\det(A)} ((a_{ij}^*)).$$

Theorem 3.7. *Let H be a subgroup of rank k in \mathbb{Z}^k . Let the columns of a $k \times k$ matrix A be a \mathbb{Z} -basis for H . Then the following are equivalent:*

- (1) H is an ideal in \mathbb{Z}^k .
- (2) There exists a unimodular matrix U such that AU is a diagonal matrix.
- (3) There is a sequence of k nonzero integers d_1, d_2, \dots, d_k such that
 - (a) $\det(A) = \pm d_1 d_2 \cdots d_k$,
 - (b) $\det(A)/d_i$ divides $\gcd(a_{1i}^*, \dots, a_{ki}^*)$ for all i .

Proof. We already saw the equivalence of (1) and (2) in Theorem 3.6. Now we will show that (2) and (3) are equivalent. Let H and A be as in the statement of the theorem. There exists a unimodular matrix U such that AU is a diagonal matrix if

and only if for some diagonal matrix $D = \text{Diagonal}(d_1, \dots, d_k)$, the matrix $A^{-1}D$ is unimodular. Using Cramer's formula for the inverse, we can equivalently say that

$$X = \frac{1}{\det(A)} A^* D$$

is unimodular. Since X is unimodular, its determinant is ± 1 . Taking determinants of both sides of the above matrix equation will give (a). Moreover, the entries of X should be all integers. For that to happen, $\det(A)$ should divide all the entries in each of the columns $d_i(a_{1i}^*, \dots, a_{ki}^*)^T$, or equivalently $\det(A)/d_i$ should divide all the entries in each of the columns $(a_{1i}^*, \dots, a_{ki}^*)^T$. Since \mathbb{Z} is a Euclidean domain, the last statement is equivalent to (b). \square

We can tell exactly when condition (2) of Theorem 3.7 holds in the case of $\mathbb{Z} \times \mathbb{Z}$. That gives the following result, which along with the rank-1 result proved earlier, gives a full answer to our problem for the ring $\mathbb{Z} \times \mathbb{Z}$.

Theorem 3.8. *Let L be a rank-2 subgroup of $\mathbb{Z} \times \mathbb{Z}$ that is generated by vectors (a, b) and (c, d) . Then L is an ideal in $\mathbb{Z} \times \mathbb{Z}$ if and only if $ad - bc$ divides $\gcd(a, c) \gcd(b, d)$.*

Proof. Let L be a rank-2 subgroup of $\mathbb{Z} \times \mathbb{Z}$ that is generated by vectors (a, b) and (c, d) , and let A be the 2×2 matrix with these two columns. From the above theorems, and using the formula for the inverse of a 2×2 matrix, we conclude that L is an ideal if and only if there exist nonzero integers d_1 and d_2 such that

- (1) $ad - bc = \pm d_1 d_2$,
- (2) $(ad - bc)/d_1$ divides $\gcd(b, d)$ and $(ad - bc)/d_2$ divides $\gcd(a, c)$.

We claim that nonzero integers d_1 and d_2 exist with these properties if and only if $ad - bc$ divides $\gcd(a, c) \gcd(b, d)$. If d_1 and d_2 exist such that (1) and (2) hold, then from (2) we get $(ad - bc)^2/(d_1 d_2)$ divides $\gcd(a, c) \gcd(b, d)$, but $(ad - bc)^2/(d_1 d_2) = ad - bc$. This proves one direction. For the other, direction, suppose $ad - bc$ divides $\gcd(a, c) \gcd(b, d)$. Then an elementary number theory fact tells us we can write $ad - bc$ as $d_1 d_2$, where d_1 divides $\gcd(a, c)$ and d_2 divides $\gcd(b, d)$. \square

We now explain how one can arrive at Theorem 3.8 more directly by solving linear equations over \mathbb{Z} . Recall that our problem boils down to the following question. *Given an integer matrix A with nonzero determinant, when does there exist a unimodular matrix X such that AX is a diagonal matrix?* To address this, we let $X = (x_{ij})$ and consider the matrix equation

$$\begin{bmatrix} a & c \\ b & d \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} = \begin{bmatrix} u & 0 \\ 0 & v \end{bmatrix}.$$

This gives us the set of equations

$$ax_{12} + cx_{22} = 0, \quad (3-1)$$

$$bx_{11} + dx_{21} = 0, \quad (3-2)$$

$$x_{11}x_{22} - x_{12}x_{21} = 1. \quad (3-3)$$

(X is unimodular, so its determinant is either 1 or -1 . However, by swapping the columns of A if necessary, we may assume that the determinant of X is 1, which gives us the third equation.) L is an ideal if and only if the above system of equations has a solution in integers x_{ij} . Let us begin with (3-1): $ax_{12} + cx_{22} = 0$ if and only if $ax_{12} = -cx_{22}$. Then

$$x_{12} = \frac{-c}{\gcd(a, c)}\alpha, \quad x_{22} = \frac{a}{\gcd(a, c)}\alpha \quad \text{for some integer } \alpha.$$

Similarly, using (3-2), we get

$$x_{11} = \frac{-d}{\gcd(b, d)}\beta, \quad x_{21} = \frac{b}{\gcd(b, d)}\beta \quad \text{for some integer } \beta.$$

Substituting these values in the determinant condition (3-3), we get

$$\begin{aligned} 1 &= x_{11}x_{22} - x_{12}x_{21} \\ &= \frac{-d}{\gcd(b, d)}\beta \frac{a}{\gcd(a, c)}\alpha - \frac{-c}{\gcd(a, c)}\alpha \frac{b}{\gcd(b, d)}\beta \\ &= \alpha\beta \left(\frac{-ad}{\gcd(a, c)\gcd(b, d)} - \frac{-bc}{\gcd(a, c)\gcd(b, d)} \right). \end{aligned}$$

Hence,

$$\gcd(a, c)\gcd(b, d) = -\alpha\beta(ad - bc). \quad (3-4)$$

Thus we see from (3-4) that the set of equations (3-1)–(3-3) is consistent over \mathbb{Z} if and only if $\det(A) = ad - bc$ divides $\gcd(a, c)\gcd(b, d)$ in \mathbb{Z} . In that case, we can take $\alpha = -1$ and

$$\beta = \frac{\gcd(a, c)\gcd(b, d)}{ad - bc}.$$

This completes the alternative proof of Theorem 3.8.

The following corollary follows immediately from Theorem 3.8.

Corollary 3.9. *Let (a, b) and (c, d) be two vectors in $\mathbb{Z} \times \mathbb{Z}$ and L be the lattice generated by these two vectors.*

- (1) *If $ad - bc = \pm 1$, then L is an ideal in $\mathbb{Z} \times \mathbb{Z}$.*
- (2) *If $ad - bc$ is a prime, then L is an ideal if and only if $ad - bc$ divides either $\gcd(a, c)$ or $\gcd(b, d)$.*

4. The ring $\mathbb{Z}_n \times \mathbb{Z}_m$

Let n and m be positive integers and consider the ring $\mathbb{Z}_n \times \mathbb{Z}_m$. Our problem is to determine when a subgroup of $(\mathbb{Z}_n \times \mathbb{Z}_m, +)$ is an ideal. We have seen that a nonzero subgroup of $\mathbb{Z}_n \times \mathbb{Z}_m$ is generated by either one or two elements, so we have two cases to consider. First, consider a subgroup L in the ring $\mathbb{Z}_n \times \mathbb{Z}_m$ that is generated by (a, b) . If either $a = 0$ in \mathbb{Z}_n or $b = 0$ in \mathbb{Z}_m , the problem is trivial because L is simply an ideal in one of the components of $\mathbb{Z}_n \times \mathbb{Z}_m$. So let us assume that both a and b are nonzero in their respective component rings. Then we have the following theorem.

Theorem 4.1. *Let $1 \leq a < n$ and $1 \leq b < m$. The subgroup generated by (a, b) in the ring $\mathbb{Z}_n \times \mathbb{Z}_m$ is a ideal if and only if*

$$\gcd\left(\frac{n}{\gcd(a, n)}, \frac{m}{\gcd(b, m)}\right) = 1.$$

Proof. Since our rings are principal ideal rings, every ideal in $\mathbb{Z}_n \times \mathbb{Z}_m$ is of the form $d_1\mathbb{Z}_n \times d_2\mathbb{Z}_m$, where d_1 and d_2 are some integers. For brevity, we will denote this ideal by $\langle d_1 \rangle \times \langle d_2 \rangle$.

Returning to our problem, let us assume that the line L generated by (a, b) is an ideal of $\mathbb{Z}_n \times \mathbb{Z}_m$. From above, we have

$$L = \langle d_1 \rangle \times \langle d_2 \rangle.$$

Consider the restrictions to L of the natural projection maps: $\pi_1: \mathbb{Z}_n \times \mathbb{Z}_m \rightarrow \mathbb{Z}_n$ and $\pi_2: \mathbb{Z}_n \times \mathbb{Z}_m \rightarrow \mathbb{Z}_m$. We will compute $\pi_1(L)$ in two different ways. On the one hand, since $L = \langle d_1 \rangle \times \langle d_2 \rangle$, we have $\pi_1(L) = \langle d_1 \rangle$. On the other hand, L is generated by (a, b) , so the first components of the elements of L pick up all multiples of a . Therefore $\pi_1(L) = \langle a \rangle$. This shows that $\langle a \rangle = \langle d_1 \rangle$. Similarly, working with the second projection map, we conclude that $\langle b \rangle = \langle d_2 \rangle$.

To summarize, L spanned by (a, b) is an ideal if and only if

$$\langle (a, b) \rangle = \langle a \rangle \times \langle b \rangle.$$

The inclusion $\langle (a, b) \rangle \subseteq \langle a \rangle \times \langle b \rangle$ is obvious. Therefore, equality holds if and only if both sides have the same cardinality. These cardinalities are given by the following formulas ($\text{ord } x$ denotes the additive order of x):

$$|\langle (a, b) \rangle| = \text{lcm}(\text{ord } a, \text{ord } b) = \frac{\text{ord } a \text{ ord } b}{\gcd(\text{ord } a, \text{ord } b)},$$

$$|\langle a \rangle \times \langle b \rangle| = \text{ord } a \text{ ord } b.$$

Equating these two expressions, clearly L spanned by (a, b) in $\mathbb{Z}_n \times \mathbb{Z}_m$ is an ideal if and only if $\gcd(\text{ord } a, \text{ord } b) = 1$. The theorem now follows from the fact that the order of an element c in $(\mathbb{Z}_s, +)$ is given by $s/\gcd(c, s)$. \square

Remark 4.2. When m and n are relatively prime, Theorem 4.1 implies that every line in $\mathbb{Z}_n \times \mathbb{Z}_m$ is an ideal. This is indeed the case because for relatively prime integers m and n , we have $\mathbb{Z}_n \times \mathbb{Z}_m \cong \mathbb{Z}_{nm}$.

More generally, the following theorem is true:

Theorem 4.3. *The subgroup generated by the element (a_1, a_2, \dots, a_k) in the ring $\mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2} \times \dots \times \mathbb{Z}_{n_k}$ is an ideal if and only if*

$$\prod_{1 \leq i < j \leq n} \gcd\left(\frac{n_i}{\gcd(a_i, n_i)}, \frac{n_j}{\gcd(a_j, n_j)}\right) = 1.$$

Proof. From the proof of Theorem 4.1, it follows that the subgroup generated by the element (a_1, a_2, \dots, a_k) in $\mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2} \times \dots \times \mathbb{Z}_{n_k}$ is an ideal if and only if

$$\prod_i \text{ord } a_i = \text{lcm}_i \text{ ord } a_i.$$

Showing that this last equation holds if and only if

$$\prod_{1 \leq i < j \leq n} \gcd(\text{ord } a_i, \text{ord } a_j) = 1$$

can be done as an exercise. Then using the formula mentioned above for the order of an element in \mathbb{Z}_s , we now get the condition given in the statement of the theorem. \square

We now investigate when a subgroup of $\mathbb{Z}_n \times \mathbb{Z}_m$ generated by two elements is an ideal. To this end, the following theorem from group theory, due to Goursat, will be useful. We will also use this theorem in the next section, where we compute some probabilities.

Theorem 4.4 (Goursat [Petrillo 2011]). *Let G_1 and G_2 be any two groups. There exists a bijection between the set S of all subgroups of $G_1 \times G_2$ and the set T of all 5-tuples $(A_1, B_1, A_2, B_2, \phi)$, where A_i is a subgroup of G_i , B_i is a normal subgroup of A_i , and ϕ is a group isomorphism from A_1/B_1 to A_2/B_2 .*

Let $\pi_i: G_1 \times G_2 \rightarrow G_i$ denote the projection homomorphisms. The desired bijection in this theorem is given as follows. For a subgroup U of $G_1 \times G_2$, we define a 5-tuple $(A_{U_1}, B_{U_1}, A_{U_2}, B_{U_2}, \phi_U)$, where

$$\begin{aligned} A_{U_1} &= \text{Im}(\pi_1|_U), \\ B_{U_1} &= \pi_1(\ker(\pi_2|_U)), \\ A_{U_2} &= \text{Im}(\pi_2|_U), \\ B_{U_2} &= \pi_2(\ker(\pi_1|_U)), \\ \phi_U(a_1 B_{U_1}) &= a_2 B_{U_2}, \quad \text{when } (a_1, a_2) \in U. \end{aligned}$$

Conversely, given a 5-tuple $(A_1, B_1, A_2, B_2, \phi)$, the corresponding subgroup U of $G_1 \times G_2$ is given by

$$U_\phi = \{(a_1, a_2) \in A_1 \times A_2 \mid \phi(a_1 B_1) = a_2 B_2\}.$$

Corollary 4.5. *Let $G_1 \times G_2$ be a finite group and let $(A_{U_1}, B_{U_1}, A_{U_2}, B_{U_2}, \phi_U)$ correspond to the subgroup U of $G_1 \times G_2$. Then we have*

$$|U| = |A_{U_1}| |B_{U_2}|.$$

Proof. It is clear from the correspondence in Goursat’s theorem that

$$|U| = |A_{U_1}/B_{U_1}| |B_{U_1}| |B_{U_2}| = |A_{U_1}| |B_{U_2}|. \quad \square$$

Given elements α and β in \mathbb{Z}_n , consider the linear map $\phi_{\alpha,\beta}: \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Z}_n$ defined by $\phi_{\alpha,\beta}(x, y) = \alpha x + \beta y$. Then we have the following theorem.

Theorem 4.6. *The subgroup of $\mathbb{Z}_n \times \mathbb{Z}_m$ generated by (a, b) and (c, d) is an ideal of $\mathbb{Z}_n \times \mathbb{Z}_m$ if and only if*

$$(\ker \phi_{a,c})(\ker \phi_{b,d}) = \mathbb{Z} \times \mathbb{Z}.$$

Proof. Let H denote the subgroup generated by (a, b) and (c, d) in $\mathbb{Z}_n \times \mathbb{Z}_m$. Suppose H is an ideal in $\mathbb{Z}_n \times \mathbb{Z}_m$. Then there exists α in \mathbb{Z}_n and β in \mathbb{Z}_m such that $H = \langle \alpha \rangle \times \langle \beta \rangle$. Taking projection maps, we can see that $\alpha = \gcd(a, c) \pmod n$ and $\beta = \gcd(b, d) \pmod m$. Thus H is an ideal if and only if

$$\langle (a, b), (c, d) \rangle = \langle \gcd(a, c) \rangle \times \langle \gcd(b, d) \rangle.$$

As in Theorem 4.1, the left-hand side is easily seen to be contained in the right-hand side, and we have equality if and only if both sides have the same cardinality. The cardinality of the right-hand side is $\text{ord}(\gcd(a, c)) \text{ord}(\gcd(b, d))$. The cardinality of the left-hand side can be computed using Corollary 4.5: it is given by $\text{ord}(\gcd(a, c) | \pi_2(\ker \pi_1 |_H))$. Equating these two expressions, we conclude that H is an ideal if and only if $\text{ord}(\gcd(b, d)) = | \pi_2(\ker \pi_1 |_H) |$. The left-hand side of this equation is the cardinality of the set

$$S = \{bx + dy \mid x, y \in \mathbb{Z}\} \subseteq \mathbb{Z}_m,$$

and the right-hand side is the cardinality of the set

$$T = \{bx + dy \mid x, y \in \mathbb{Z} \text{ such that } ax + cy = 0 \in \mathbb{Z}_n\} \subseteq \mathbb{Z}_m.$$

S and T have the same cardinality precisely when the image of $\phi_{b,d}: \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Z}_m$ is the same as the image of $\phi_{b,d}$ restricted to the kernel of $\phi_{a,c}: \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Z}_n$. That happens exactly when $\ker(\phi_{a,c})$ intersects every coset in $\mathbb{Z} \times \mathbb{Z} / \ker(\phi_{b,d})$, which is true if and only if $(\ker \phi_{a,c})(\ker \phi_{b,d}) = \mathbb{Z} \times \mathbb{Z}$. \square

We can get a finite-type condition that is equivalent to the one in Theorem 4.6. To get this, set $l = \text{lcm}(m, n)$. Then given elements α and β in \mathbb{Z}_n , define the linear map $\psi_{\alpha, \beta}: \mathbb{Z}_l \times \mathbb{Z}_l \rightarrow \mathbb{Z}_n$ as $\psi_{\alpha, \beta}(x, y) = \alpha x + \beta y$. We now have the following corollary.

Corollary 4.7. *The subgroup of $\mathbb{Z}_n \times \mathbb{Z}_m$ generated by (a, b) and (c, d) is an ideal of $\mathbb{Z}_n \times \mathbb{Z}_m$ if and only if*

$$|(\ker \psi_{a,c})(\ker \psi_{b,d})| = nm.$$

Proof. This follows from the proof of the previous theorem. Note that the maps $\phi_{a,c}$ and $\phi_{b,d}$ factor through $\psi_{a,c}$ and $\psi_{b,d}$ respectively. \square

Goursat's theorem for more than two components [Bauer et al. 2011] has a very complicated structure, and in particular, it is not helpful to solve our problem.

5. Probability that a subgroup is an ideal

As one would expect, the above results suggest that a subgroup of a ring is rarely an ideal. Now we will make this precise by computing explicitly the probability that a randomly chosen subgroup of $\mathbb{Z}_n \times \mathbb{Z}_m$ is an ideal using the approach and results from [Petrillo 2011]. Let P_R denote the probability that a randomly chosen subgroup of a finite ring R is an ideal. This probability is given by

$$P_R = \frac{\text{total number of ideals in } R}{\text{total number of subgroups in } (R, +)}.$$

Our interest is in the ring $\mathbb{Z}_n \times \mathbb{Z}_m$. If either n or m is 1, then clearly $P_R = 1$. So we will assume that $n > 1$ and $m > 1$. Let $S = \{p_1, \dots, p_k\}$ denote the set of all distinct primes which divide mn . Then the prime factorizations of m and n are

$$m = p_1^{r_1} \cdots p_k^{r_k} \quad \text{and} \quad n = p_1^{s_1} \cdots p_k^{s_k},$$

where the exponents are nonnegative integers, and the Chinese remainder theorem gives the decomposition

$$\mathbb{Z}_n \times \mathbb{Z}_m = (\mathbb{Z}_{p_1^{r_1}} \times \mathbb{Z}_{p_1^{s_1}}) \times \cdots \times (\mathbb{Z}_{p_k^{r_k}} \times \mathbb{Z}_{p_k^{s_k}}).$$

Lemma 5.1.
$$P_{\mathbb{Z}_n \times \mathbb{Z}_m} = \prod_{i=1}^k P_{\mathbb{Z}_{p_i^{r_i}} \times \mathbb{Z}_{p_i^{s_i}}}.$$

Proof. This follows from two facts. First, note that every ideal I in $\mathbb{Z}_n \times \mathbb{Z}_m$ is of the form $I = \prod_{i=1}^k I_i$, where I_i is an ideal of the ring $\mathbb{Z}_{p_i^{r_i}} \times \mathbb{Z}_{p_i^{s_i}}$. Next we use a theorem of Suzuki [1951] that says if G_1 and G_2 are two finite groups with relatively prime orders, then every subgroup of $G_1 \times G_2$ is of the form $H_1 \times H_2$,

where H_i is a subgroup of G_i . In particular, every subgroup H of $(\mathbb{Z}_n \times \mathbb{Z}_m, +)$ is of the form $\prod_{i=1}^k H_i$, where H_i is a subgroup of $\mathbb{Z}_{p_i}^{r_i} \times \mathbb{Z}_{p_i}^{s_i}$. Then we have

$$\begin{aligned} P_{\mathbb{Z}_n \times \mathbb{Z}_m} &= \frac{\text{total number of ideals in } \mathbb{Z}_n \times \mathbb{Z}_m}{\text{total number of subgroups in } (\mathbb{Z}_n \times \mathbb{Z}_m, +)} \\ &= \prod_{i=1}^k \frac{\text{total number of ideals in } \mathbb{Z}_{p_i}^{r_i} \times \mathbb{Z}_{p_i}^{s_i}}{\text{total number of subgroups in } (\mathbb{Z}_{p_i}^{r_i} \times \mathbb{Z}_{p_i}^{s_i}, +)} \\ &= \prod_{i=1}^k P_{\mathbb{Z}_{p_i}^{r_i} \times \mathbb{Z}_{p_i}^{s_i}}. \end{aligned} \quad \square$$

In view of Lemma 5.1, it is enough to compute

$$P_{\mathbb{Z}_{p_i}^{r_i} \times \mathbb{Z}_{p_i}^{s_i}}.$$

We do this in the next two lemmas, beginning by computing the number of ideals.

Lemma 5.2. *The number of ideals in $\mathbb{Z}_{p^r} \times \mathbb{Z}_{p^s}$ is equal to $(r + 1)(s + 1)$.*

Proof. Every ideal in $\mathbb{Z}_{p^r} \times \mathbb{Z}_{p^s}$ is of the form $a\mathbb{Z}_{p^r} \times b\mathbb{Z}_{p^s}$, where a is a divisor of p^r and b is a divisor of p^s . This gives $(r + 1)(s + 1)$ for the total number of ideals. \square

Next we have to compute the number of subgroups in $\mathbb{Z}_{p^r} \times \mathbb{Z}_{p^s}$. This number can be obtained using the above-mentioned Goursat’s theorem.

Lemma 5.3 [Petrillo 2011]. *The total number of subgroups of $\mathbb{Z}_{p^r} \times \mathbb{Z}_{p^s}$ ($r \leq s$) is*

$$\frac{p^{r+1}((s - r + 1)(p - 1) + 2) - ((s + r + 3)(p - 1) + 2)}{(p - 1)^2}.$$

Proof sketch. Goursat’s theorem can be greatly simplified in the case under consideration. There is a unique subgroup of order p^k in \mathbb{Z}_{p^r} for any $0 \leq k \leq r$ and these subgroups form a linear chain. Moreover, the group of automorphisms of \mathbb{Z}_{p^k} corresponds to the units in this ring, and we have $p^k - p^{k-1}$ of them. We now have to count the 5-tuples $(A_1, B_1, A_2, B_2, \phi)$ that correspond to subgroups in Goursat’s theorem. If $|A_i/B_i| = 1$, the number of subgroups is $(r + 1)(s + 1)$ because we have $r + 1$ choices for A_1/B_1 and $s + 1$ choices for A_2/B_2 (clearly ϕ is trivial). If $|A_i/B_i| = p^k$ for $1 \leq k \leq r$, we have $r - k + 1$ choices for A_1/B_1 and $s - k + 1$ choices for A_2/B_2 , and finally $p^k - p^{k-1}$ choices for ϕ , so in this case we have $(r - k + 1)(s - k + 1)(p^k - p^{k-1})$ subgroups. In total we have

$$(r + 1)(s + 1) + \sum_{k=1}^r (r - k + 1)(s - k + 1)(p^k - p^{k-1})$$

subgroups. The rest is straightforward algebra; see [Petrillo 2011]. \square

Combining the above lemmas, we get our formulas for $P_{\mathbb{Z}_{p^r} \times \mathbb{Z}_{p^s}}$ and $P_{\mathbb{Z}_n \times \mathbb{Z}_m}$.

Theorem 5.4. *Let p be a prime and r, s, n, m be positive integers, with $r \leq s$. Then*

$$P_{\mathbb{Z}_{p^r} \times \mathbb{Z}_{p^s}} = \frac{(r+1)(s+1)(p-1)^2}{p^{r+1}((s-r+1)(p-1)+2) - ((s+r+3)(p-1)+2)},$$

$$P_{\mathbb{Z}_n \times \mathbb{Z}_m} = \prod_{i=1}^k \frac{(r_i+1)(s_i+1)(p_i-1)^2}{p_i^{r_i+1}(|s_i-r_i|+1)(p_i-1)+2 - ((s_i+r_i+3)(p_i-1)+2)}.$$

We now record two special cases, which can be derived from Theorem 5.4 using routine algebra.

Corollary 5.5. *Let p be a prime and let r be a positive integer. Then*

$$P_{\mathbb{Z}_{p^r} \times \mathbb{Z}_{p^r}} = \frac{(r+1)^2(p-1)^2}{p^{r+1}(p+1) - 2r(p-1) - 3p+1} \quad \text{and} \quad P_{\mathbb{Z}_p \times \mathbb{Z}_p} = \frac{4}{p+3}.$$

It is clear from the above expressions that these probabilities are small, as expected. For instance, by choosing a large prime, the value of $P_{\mathbb{Z}_p \times \mathbb{Z}_p}$ can be made arbitrarily small. Similarly for a fixed prime p , the numerator of $P_{\mathbb{Z}_{p^r} \times \mathbb{Z}_{p^r}}$ is a polynomial function in r , whereas the denominator is an exponential function in r . Thus $\lim_{r \rightarrow \infty} P_{\mathbb{Z}_{p^r} \times \mathbb{Z}_{p^r}} = 0$.

The main obstruction in generalizing these formulas to the rings $R = \prod_{i=1}^k \mathbb{Z}_{n_i}$ is the lack of a closed formula for the number of subgroups in $(\prod_{i=1}^k \mathbb{Z}_{p^i}, +)$ when $k \geq 3$. However, when the integers n_i are all square-free, one can compute P_R easily. This is because Lemma 5.1 helps us to reduce the problem of computing P_R to the problem of computing P_S , where $S = \prod_{i=1}^r \mathbb{Z}_p$ for some prime p and positive integer $r (\leq k)$. The latter is a vector space over \mathbb{F}_p , where subgroups are same as vector subspaces. The number of subspaces in $(S, +)$ is given by the well-known formula

$$\sum_{i=1}^r \binom{r}{i}_p,$$

where $\binom{r}{i}_p$ is the Gaussian binomial coefficient, which counts the number of i -dimensional subspaces of \mathbb{F}_p^r . Explicitly its value is given by

$$\binom{r}{i}_p = \frac{(p^r-1)(p^r-p)\cdots(p^r-p^{i-1})}{(p^i-1)(p^i-p)\cdots(p^i-p^{i-1})}.$$

Since the number of ideals in S is 2^r , we get this formula:

Proposition 5.6.
$$P_{\mathbb{Z}_p^r} = \frac{2^r}{\sum_{i=1}^r \binom{r}{i}_p}.$$

Acknowledgement

We would like to thank the referee for comments and suggestions, which we used to improve the exposition of this paper.

References

- [Bauer et al. 2011] K. Bauer, D. Sen, and P. Zvengrowski, “A generalized Goursat Lemma”, preprint, 2011. arXiv 1109.0024
- [Dummit and Foote 2004] D. S. Dummit and R. M. Foote, *Abstract algebra*, 3rd ed., Wiley, Hoboken, NJ, 2004. MR 2007h:00003 Zbl 1037.00003
- [Petrillo 2011] J. Petrillo, “Counting subgroups in a direct product of finite cyclic groups”, *College Math. J.* **42**:3 (2011), 215–222. MR 2012f:20074 Zbl 1272.97033
- [Suzuki 1951] M. Suzuki, “On the lattice of subgroups of finite groups”, *Trans. Amer. Math. Soc.* **70** (1951), 345–371. MR 12,586b Zbl 0043.02502
- [Zhao 2010] W. Zhao, “Generalizations of the image conjecture and the Mathieu conjecture”, *J. Pure Appl. Algebra* **214**:7 (2010), 1200–1216. MR 2011e:33032 Zbl 1205.33017
- [Zhao 2012] W. Zhao, “Mathieu subspaces of associative algebras”, *J. Algebra* **350** (2012), 245–272. MR 2859886 Zbl 1255.16018

Received: 2015-05-15 Revised: 2015-06-02 Accepted: 2015-06-17

schebol@ilstu.edu

*Department of Mathematics, Illinois State University,
Normal, IL 61790, United States*

clhenry@ilstu.edu

*Department of Mathematics, Illinois State University,
Normal, IL 61790, United States*

Explicit bounds for the pseudospectra of various classes of matrices and operators

Feixue Gong, Olivia Meyerson, Jeremy Meza,
Mihai Stoiciu and Abigail Ward

(Communicated by Stephan Garcia)

We study the ε -pseudospectra $\sigma_\varepsilon(A)$ of square matrices $A \in \mathbb{C}^{N \times N}$. We give a complete characterization of the ε -pseudospectra of 2×2 matrices and describe the asymptotic behavior (as $\varepsilon \rightarrow 0$) of $\sigma_\varepsilon(A)$ for every square matrix A . We also present explicit upper and lower bounds for the ε -pseudospectra of bidiagonal matrices, as well as for finite-rank operators.

1. Introduction

The pseudospectra of matrices and operators is an important mathematical object that has found applications in various areas of mathematics: linear algebra, functional analysis, numerical analysis, and differential equations. An overview of the main results on pseudospectra can be found in [Trefethen and Embree 2005].

In this paper we describe the asymptotic behavior of the ε -pseudospectrum of all $n \times n$ matrices. We apply this asymptotic bound to several classes of matrices and operators, including 2×2 matrices, bidiagonal matrices, and finite-rank operators, and additionally provide explicit bounds on their ε -pseudospectra.

The paper is organized as follows: in Section 2, we give the three standard equivalent definitions for the pseudospectrum and present the “classical” results on ε -pseudospectra of normal and diagonalizable matrices (the Bauer–Fike theorems). Section 3 contains a detailed analysis of the ε -pseudospectrum of 2×2 matrices, including both the nondiagonalizable and the diagonalizable cases. The asymptotic behavior (as $\varepsilon \rightarrow 0$) of the ε -pseudospectrum of each $n \times n$ matrix is described in Section 4, where we show (in Theorem 4.2) that, for every square matrix, the ε -pseudospectrum converges, as $\varepsilon \rightarrow 0$, to a union of disks. We apply the main result of Section 4 to several classes of matrices: matrices with a simple eigenvalue, matrices with an eigenvalue with geometric multiplicity 1, 2×2 matrices, and Jordan blocks.

MSC2010: 15A18, 15A60, 47A10, 65F15.

Keywords: spectrum, pseudospectrum, bidiagonal matrices, perturbation of eigenvalues.

Section 5 is dedicated to the analysis of arbitrary periodic bidiagonal matrices A . We derive explicit formulas (in terms the coefficients of A) for the asymptotic radii, given by Theorem 4.2, of the ε -pseudospectrum of A as $\varepsilon \rightarrow 0$. In the last section (Section 6), we consider finite-rank operators and show that the ε -pseudospectrum of an operator of rank m is at most as big as $C\varepsilon^{1/m}$ as $\varepsilon \rightarrow 0$.

2. Pseudospectra

Motivation and definitions. The concept of the spectrum of a matrix $A \in \mathbb{C}^{N \times N}$ provides a fundamental tool for understanding the behavior of A . As is well known, a complex number $z \in \mathbb{C}$ is in the spectrum of A (denoted $\sigma(A)$) whenever $zI - A$ (which we will denote as $z - A$) is not invertible, i.e., the characteristic polynomial of A has z as a root. As slightly perturbing the coefficients of A will change the roots of the characteristic polynomial, the property of “membership in the set of eigenvalues” is not well-suited for many purposes, especially those in numerical analysis. We thus want to find a characterization of when a complex number is close to an eigenvalue, and we do this by considering the set of complex numbers z such that $\|(z - A)^{-1}\|$ is large, where the norm here is the usual operator norm induced by the Euclidean norm, i.e.,

$$\|A\| = \sup_{\|v\|=1} \|Av\|.$$

The motivation for considering this question comes from the observation that if z_n is a sequence of complex numbers converging to an eigenvalue λ of A , then $\|(z_n - A)^{-1}\| \rightarrow \infty$ as $n \rightarrow \infty$. We call the operator $(z - A)^{-1}$ the *resolvent* of A . The observation that the norm of the resolvent is large when z is close to an eigenvalue of A leads us to the first definition of the ε -pseudospectrum of an operator.

Definition 2.1. Let $A \in \mathbb{C}^{N \times N}$, and let $\varepsilon > 0$. The ε -pseudospectrum of A is the set of $z \in \mathbb{C}$ such that

$$\|(z - A)^{-1}\| > 1/\varepsilon.$$

Note that the boundary of the ε -pseudospectrum is exactly the $1/\varepsilon$ level curve of the function $z \mapsto \|(z - A)^{-1}\|$. Figure 1 depicts the behavior of this function near the eigenvalues.

The resolvent norm has singularities in the complex plane, and as we approach these points, the resolvent norm grows to infinity. Conversely, if $\|(z - A)^{-1}\|$ approaches infinity, then z must approach some eigenvalue of A [Trefethen and Embree 2005, Theorem 2.4].

(It is also possible to develop a theory of pseudospectrum for operators on Banach spaces, and it is important to note that this converse does not necessarily hold for such operators; that is, there are operators [Davies 1999a; 1999b] such that $\|(z - A)^{-1}\|$ approaches infinity, but z does not approach the spectrum of A .)

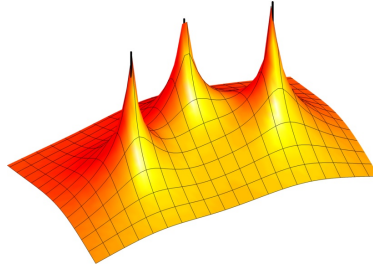


Figure 1. Contour plot of the resolvent norm.

The second and third definitions of the ε -pseudospectrum arise from eigenvalue perturbation theory [Kato 1995].

Definition 2.2. Let $A \in \mathbb{C}^{N \times N}$. The ε -pseudospectrum of A is the set of $z \in \mathbb{C}$ such that

$$z \in \sigma(A + E)$$

for some E with $\|E\| < \varepsilon$.

Definition 2.3. Let $A \in \mathbb{C}^{N \times N}$. The ε -pseudospectrum of A is the set of $z \in \mathbb{C}$ such that

$$\|(z - A)v\| < \varepsilon$$

for some unit vector v .

The third definition is similar to our first definition in that it quantifies how close z is to an eigenvalue of A . In addition to this, it also gives us the notion of an ε -pseudoeigenvector.

Theorem 2.4 (equivalence of the definitions of pseudospectra). *For every matrix $A \in \mathbb{C}^{N \times N}$, the three definitions above are equivalent.*

The proof of this theorem is given in [Trefethen and Embree 2005, Section 2]. As all three definitions are equivalent, we can unambiguously denote the ε -pseudospectrum of A as $\sigma_\varepsilon(A)$.

While the investigation of the set $\sigma_\varepsilon(A)$ can be motivated by questions in numerical analysis, the main impetus for the in-depth study of the ε -pseudospectrum is the study of the size and behavior of the norms $\|e^{tA}\|$ (with $t \in [0, \infty)$) and $\|A^k\|$ (with $k \in \{0, 1, 2, \dots\}$), where A is a matrix or an operator that defines the differential equation $x' = Ax$ or the difference equation $x_{k+1} = Ax_k$.

As explained in [Trefethen and Embree 2005, Part IV] and in [Böttcher 2006], the ε -pseudospectra of A can be used to define and compute the Kreiss constant $\mathcal{K}(A)$, which in turn can be used, via the Kreiss matrix theorem [Trefethen and Embree 2005, Section 18], to find upper and lower bounds for $\sup_{t \geq 0} \|e^{tA}\|$ and $\sup_{k \geq 0} \|A^k\|$. Thus, while the spectrum and the numerical range of A only provide information on $\|e^{tA}\|$ and $\|A^k\|$ in the limits $t \rightarrow \infty$ and $t \rightarrow 0$, as well as for $k \rightarrow \infty$, the

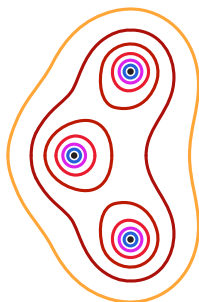


Figure 2. The curves bounding the ε -pseudospectra of a matrix A for different values of ε .

pseudospectrum of A gives information on the size of these norms (and therefore on the size of the solutions of the corresponding linear differential/difference equations) for all values of t and k .

Figure 2 depicts an example of ε -pseudospectra for a specific matrix and for various ε . We see that the boundaries of ε -pseudospectra for a matrix are curves in the complex plane around the eigenvalues of the matrix. We are interested in understanding geometric and algebraic properties of these curves.

Several properties of pseudospectra are proven in [Trefethen and Embree 2005, Section 2]. One of which is that if $A \in \mathbb{C}^{N \times N}$, then $\sigma_\varepsilon(A)$ is nonempty, open, and bounded, with at most N connected components, each containing one or more eigenvalues of A . This leads us to the following notation:

Notation. For $\lambda \in \sigma(A)$, we write $\sigma_\varepsilon(A) \upharpoonright \lambda$ to be the connected component of $\sigma_\varepsilon(A)$ that contains λ .

Another property, which follows straight from the definitions of pseudospectra, is that $\bigcap_{\varepsilon > 0} \sigma_\varepsilon(A) = \sigma(A)$. From these properties, it follows that there is ε small enough so that $\sigma_\varepsilon(A)$ consists of exactly $|\sigma(A)|$ connected components, each an open set around a distinct eigenvalue. In particular, there is ε small enough so that $\sigma(A) \cap \sigma_\varepsilon(A) \upharpoonright \lambda = \{\lambda\}$.

When a matrix A is the direct sum of smaller matrices, we can look at the pseudospectra of the smaller matrices to understand the ε -pseudospectrum of A . We get the following theorem from [Trefethen and Embree 2005]:

Theorem 2.5.
$$\sigma_\varepsilon(A_1 \oplus A_2) = \sigma_\varepsilon(A_1) \cup \sigma_\varepsilon(A_2).$$

Normal matrices. Recall that a matrix A is *normal* if $AA^* = A^*A$, or equivalently, if A can be diagonalized with an orthonormal basis of eigenvectors.

The pseudospectra of these matrices are particularly well-behaved: Theorem 2.6 shows that the ε -pseudospectrum of a normal matrix is exactly the union of disks of radius ε around each eigenvalue, as in shown in Figure 3. This is clear for diagonal

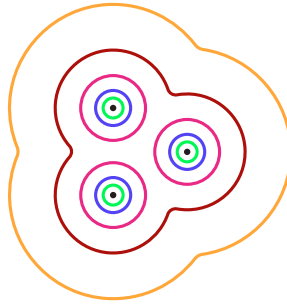


Figure 3. The curves bounding the ε -pseudospectra of a normal matrix for different values of ε . Note that the boundaries are unions of perfect circles around each eigenvalue.

matrices; it follows for normal matrices since, as we shall see, the ε -pseudospectrum of a matrix is invariant under a unitary change of basis.

Theorem 2.6. *Let $A \in \mathbb{C}^{N \times N}$. Then,*

$$\sigma(A) + B(0, \varepsilon) \subseteq \sigma_\varepsilon(A) \quad \text{for all } \varepsilon > 0. \tag{2-1}$$

Furthermore, A is a normal matrix if and only if

$$\sigma_\varepsilon(A) = \sigma(A) + B(0, \varepsilon) \quad \text{for all } \varepsilon > 0. \tag{2-2}$$

The proof of this theorem can be found in [Trefethen and Embree 2005, Section 2].

Nonnormal diagonalizable matrices. Now suppose A is diagonalizable but not normal, i.e., we cannot diagonalize A by an isometry of \mathbb{C}^N . In this case, we do not expect to get an exact characterization of the ε -pseudospectra as we did previously. That is, there exist matrices with pseudospectra larger than the union of disks of radius ε . Regardless, we can still characterize the behavior of nonnormal, diagonalizable matrices.

Theorem 2.7 (Bauer–Fike). *Let $A \in \mathbb{C}^{N \times N}$ be diagonalizable with $A = V D V^{-1}$. Then for each $\varepsilon > 0$,*

$$\sigma(A) + B(0, \varepsilon) \subseteq \sigma_\varepsilon(A) \subseteq \sigma(A) + B(0, \varepsilon \kappa(V)),$$

where

$$\kappa(V) = \|V\| \|V^{-1}\| = \frac{s_{\max}(V)}{s_{\min}(V)},$$

and $s_{\max}(V), s_{\min}(V)$ are the maximum and minimum singular values of V .

Here, $\kappa(V)$ is known as the condition number of V . Note that $\kappa(V) \geq 1$, with equality attained if and only if A is normal. Thus, $\kappa(V)$ can be thought of as a measure of the normality of a matrix. However, there is some ambiguity when we define $\kappa(V)$, as V is not uniquely determined. If the eigenvalues are distinct, then $\kappa(V)$ becomes unique if the eigenvectors are normalized by $\|v_j\| = 1$.

Nondiagonalizable matrices. So far we have considered normal matrices, and more generally, diagonalizable matrices. We now relax our constraint that our matrix be diagonalizable, and provide similar bounds on the pseudospectra. While not every matrix is diagonalizable, every matrix can be put in Jordan normal form. Below we give a brief review of the Jordan form.

Let $A \in \mathbb{C}^{N \times N}$ and suppose A has only one eigenvalue, λ with geometric multiplicity 1. Writing A in Jordan form, there exists a matrix V such that $AV = VJ$, where J is a single Jordan block of size N . Write

$$V = (v_1 \ v_2 \ \dots \ v_n).$$

Then,

$$AV = (Av_1 \ Av_2 \ \dots \ Av_n) = (\lambda v_1 \ v_1 + \lambda v_2 \ \dots \ v_{n-1} + \lambda v_n) = VJ,$$

and hence v_1 is a right eigenvector associated with λ and v_2, \dots, v_n are *generalized right eigenvectors*, that is, right eigenvectors for $(A - \lambda I)^k$ for $k > 1$. Similarly, there exists a matrix U such that $U^*A = JU^*$, where now the rows of U^* are left generalized eigenvectors associated with λ .

We can also quantify the normality of an eigenvalue in the same way $\kappa(V)$ quantifies the normality of a matrix.

Definition 2.8. For each simple eigenvalue λ_j of a matrix A , the *condition number* of λ_j is defined as

$$\kappa(\lambda_j) = \frac{\|u_j\| \|v_j\|}{|u_j^* v_j|},$$

where v_j and u_j^* are the right and left eigenvectors associated with λ_j , respectively.

Note. The Cauchy–Schwarz inequality implies $|u_j^* v_j| \leq \|u_j\| \|v_j\|$, so $\kappa(\lambda_j) \geq 1$, with equality when u_j and v_j are collinear. An eigenvalue for which $\kappa(\lambda_j) = 1$ is called a normal eigenvalue; a matrix A with all simple eigenvalues is normal if and only if $\kappa(\lambda_j) = 1$ for all eigenvalues.

With this definition, we can find finer bounds for the pseudospectrum of a matrix; in particular, we can find bounds for the components of the pseudospectrum centered around each eigenvalue. The following theorem can be found, for example, in [Baumgärtel 1985].

Theorem 2.9 (asymptotic pseudospectra inclusion regions). *Suppose $A \in \mathbb{C}^{N \times N}$ has N distinct eigenvalues. Then, as $\varepsilon \rightarrow 0$,*

$$\sigma_\varepsilon(A) \subseteq \bigcup_{j=1}^N B(\lambda_j, \varepsilon \kappa(\lambda_j) + \mathcal{O}(\varepsilon^2)).$$

We can drop the $\mathcal{O}(\varepsilon^2)$ term, for which we get an increase in the radius of our inclusion disks by a factor of N [Bauer and Fike 1960, Theorem 4].

Theorem 2.10 (Bauer–Fike theorem based on $\kappa(\lambda_j)$). *Suppose $A \in \mathbb{C}^{N \times N}$ has N distinct eigenvalues. Then for all $\varepsilon > 0$,*

$$\sigma_\varepsilon(A) \subseteq \bigcup_{j=1}^N B(\lambda_j, \varepsilon N \kappa(\lambda_j)).$$

The above two theorems give us upper bounds on the pseudospectra of A only when A has N distinct eigenvalues. These results can be generalized for matrices that do not have distinct eigenvalues. The following is proven in [Trefethen and Embree 2005, Section 52].

Theorem 2.11 (asymptotic formula for the resolvent norm). *Let $\lambda_j \in \sigma(A)$ be an eigenvalue of A with k_j the size of the largest Jordan block associated to λ_j . For every $z \in \sigma_\varepsilon(A)$, for small enough ε ,*

$$|z - \lambda_j| \leq (C_j \varepsilon)^{1/k_j},$$

where $C_j = \|V_j T_j^{k_j-1} U_j^*\|$ and $T = J - \lambda I$.

We extend these results by providing lower bounds for arbitrary matrices, as well as explicit formulas for the ε -pseudospectra of 2×2 matrices.

3. Pseudospectra of 2×2 matrices

The following section presents a complete characterization of the ε -pseudospectra of all 2×2 matrices. We classify matrices by whether they are diagonalizable or nondiagonalizable and determine the ε -pseudospectra for each class. We begin with an explicit formula for computing the norm of a 2×2 matrix.

Let A be a 2×2 matrix with complex coefficients and let s_{\max} denote the largest singular value of A .

Then,

$$\|A\|^2 = s_{\max}^2 = \frac{\text{Tr}(A^*A) + \sqrt{\text{Tr}(A^*A)^2 - 4 \det(A^*A)}}{2}. \tag{3-1}$$

Nondiagonalizable 2×2 matrices. Every nondiagonalizable 2×2 matrix must have exactly one eigenvalue of geometric multiplicity 1. In this case, we can Jordan-decompose the matrix and use the first definition of pseudospectra to show that $\sigma_\varepsilon(A)$ must be a perfect disk.

Proposition 3.1. *Let A be a nondiagonalizable 2×2 matrix, and let λ denote the eigenvalue of A . Write $A = V J V^{-1}$, where*

$$V = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad J = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}. \tag{3-2}$$

Given $\varepsilon > 0$,

$$\sigma_\varepsilon(A) = B(\lambda, |k|), \tag{3-3}$$

where

$$|k| = \sqrt{C\varepsilon + \varepsilon^2} \quad \text{and} \quad C = \frac{|a|^2 + |c|^2}{|ad - bc|}. \tag{3-4}$$

Proof. Let $z = \lambda + k$, where $k \in \mathbb{C}$. Then we have $(z - A)^{-1} = V(z - J)^{-1}V^{-1}$.

Taking the norm, this yields

$$\|(z - A)^{-1}\| = \frac{\|M\|}{|k^2(ad - bc)|}, \quad \text{where } M = \begin{pmatrix} adk - ac - bck & a^2 \\ -c^2 & -bck + ac + adk \end{pmatrix}.$$

From (3-1), we obtain that

$$\varepsilon^{-1} < \|(z - A)^{-1}\| = \frac{\sqrt{\text{Tr}(M^*M)} + \sqrt{\text{Tr}(M^*M)^2 - 4 \det(M^*M)}}{|k|^2 |ad - bc| \sqrt{2}}.$$

Note that this function depends only on $|k| = |z - \lambda|$; thus for every $\varepsilon > 0$, we have that $\sigma_\varepsilon(A)$ will be a disk. Solving for k to find the curve bounding the pseudospectrum, we obtain

$$|k| = \sqrt{\frac{|a|^2 + |c|^2}{|ad - bc|} \varepsilon + \varepsilon^2}. \quad \square$$

Diagonalizable 2×2 matrices. Diagonalizable 2×2 matrices must have two distinct eigenvalues or be a multiple of the identity matrix. In either case, the pseudospectra can be described by the following proposition.

Proposition 3.2. *Let A be a diagonalizable 2×2 matrix and let λ_1, λ_2 be the eigenvalues of A and v_1, v_2 be the eigenvectors associated with the eigenvalues. Then the boundary of $\sigma_\varepsilon(A)$ is the set of points z that satisfy the equation*

$$(\varepsilon^2 - |z - \lambda_1|^2)(\varepsilon^2 - |z - \lambda_2|^2) - \varepsilon^2 |\lambda_1 - \lambda_2|^2 \cot^2(\theta) = 0, \tag{3-5}$$

where θ is the angle between the two eigenvectors.

Proof. Since A is diagonalizable, we can write $A = VDV^{-1}$, where

$$V = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad D = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}. \tag{3-6}$$

Without loss of generality, let $z = \lambda_1 + k$.

Let $\gamma = \lambda_1 - \lambda_2$ and $r = ad - bc$. Then,

$$\|(z - A)^{-1}\| = \|V(z - D)^{-1}V^{-1}\| = \frac{\|M\|}{|rk(\gamma + k)|},$$

where

$$M = \begin{pmatrix} ad\gamma + rk & -ab\gamma \\ cd\gamma & -bc\gamma + rk \end{pmatrix}.$$

Calculating $\text{Tr}(M^*M)$, we obtain

$$\text{Tr}(M^*M) = |r|^2(|\gamma + k|^2 + |k|^2 + |\gamma|^2 \cot^2 \theta), \tag{3-7}$$

where θ is the angle between the two eigenvectors, which are exactly the columns of V . For the determinant, we have

$$\det(M^*M) = |r|^4 |k|^2 |k + \gamma|^2. \tag{3-8}$$

Plugging the above into (3-1), we get

$$\varepsilon^{-1} = \|(z - A)^{-1}\| = \sqrt{\frac{\text{Tr}(M^*M) + \sqrt{\text{Tr}(M^*M)^2 - 4 \det(M^*M)}}{2|r|^2 |k(\gamma + k)|^2}}.$$

Rewriting and simplifying, we obtain the curve describing the boundary of the pseudospectrum:

$$(\varepsilon^2 - |k|^2)(\varepsilon^2 - |k + \gamma|^2) - \varepsilon^2 |\gamma|^2 \cot^2 \theta = 0. \quad \square$$

Note that for normal matrices, the eigenvectors are orthogonal. Therefore the equation above reduces to

$$(\varepsilon^2 - |k|^2)(\varepsilon^2 - |k + \gamma|^2) = 0, \tag{3-9}$$

which describes two disks of radius ε centered around λ_1, λ_2 , as we expect.

When the matrix only has one eigenvalue and is still diagonalizable (i.e., when it is a multiple of the identity), we obtain

$$(\varepsilon^2 - |k|^2)^2 = 0,$$

which is a disk of radius ε centered around the eigenvalue.

One consequence of Proposition 3.2 to note is that the shape of $\sigma_\varepsilon(A)$ is dependent on both the eigenvalues *and* the eigenvectors of the matrix A . Another less obvious consequence is that the pseudospectrum of a 2×2 matrix approaches a union of disks as ε tends to 0.

Proposition 3.3. *Let A be a diagonalizable 2×2 matrix with two distinct eigenvalues, λ_1, λ_2 . Then, $\sigma_\varepsilon(A) \upharpoonright_{\lambda_i}$ asymptotically tends toward a disk. In particular,*

$$\frac{r_{\max}(\lambda_i)}{r_{\min}(\lambda_i)} = 1 + \mathcal{O}(\varepsilon),$$

where $r_{\max}(\lambda_i), r_{\min}(\lambda_i)$ are the maximum and minimum distances from λ_i to $\partial\sigma_\varepsilon(A) \upharpoonright_{\lambda_i}$. Moreover, for A diagonalizable but not normal, $\sigma_\varepsilon(A) \upharpoonright_{\lambda_i}$ is never a perfect disk.

Proof. Let ε be small enough so that the ε -pseudospectrum is disconnected. Without loss of generality, we will consider $\sigma_\varepsilon(A) \upharpoonright_{\lambda_1}$.

Let $z_{\max} \in \partial\sigma_\varepsilon(A)$ such that $|z_{\max} - \lambda_1|$ is a maximum. Set $r_{\max}(\lambda) = |z_{\max} - \lambda|$. Consider the line joining λ_1 and λ_2 . Suppose for contradiction that z_{\max} does not lie on this line. Then, rotate z_{\max} in the direction of λ_2 so that it is on this line, and call this new point z' . Note that $|z' - \lambda_2| < |z_{\max} - \lambda_2|$, but $|z' - \lambda_1| = |z_{\max} - \lambda_1|$. As such, we get that

$$\begin{aligned} (|z' - \lambda_1|^2 - \varepsilon^2)(|z' - \lambda_2|^2 - \varepsilon^2) &< (|z_{\max} - \lambda_1|^2 - \varepsilon^2)(|z_{\max} - \lambda_2|^2 - \varepsilon^2) \\ &= \varepsilon^2 |\lambda_1 - \lambda_2|^2 \cot^2 \theta. \end{aligned}$$

Thus, from Proposition 3.2, we have that $z' \in \sigma_\varepsilon(A)$ but z' is not on the boundary of $\sigma_\varepsilon(A)$. Starting from z' and traversing the line joining λ_1 and λ_2 , we can find $z'' \in \partial\sigma_\varepsilon(A)$ such that $|z'' - \lambda_1| > |z' - \lambda_1| = |z_{\max} - \lambda_1|$. This contradicts our choice of z_{\max} and so z_{\max} must be on the line joining λ_1 and λ_2 . A similar argument shows that z_{\min} must also be on this line, where $z_{\min} \in \partial\sigma_\varepsilon(A)$ such that $r_{\min} = |z_{\min} - \lambda_1|$ is a minimum.

Since z_{\max} is on the line joining λ_1 and λ_2 , we have the exact equality

$$|z_{\max} - \lambda_2| = |z_{\max} - \lambda_1| + |\lambda_2 - \lambda_1|.$$

Let $y = |\lambda_2 - \lambda_1|$. The equation describing $r_{\max}(\lambda_1)$ becomes

$$(r_{\max}(\lambda_1)^2 - \varepsilon^2)((y - r_{\max}(\lambda_1))^2 - \varepsilon^2) = \varepsilon^2 y^2 \cot^2 \theta.$$

Similarly, we can obtain the equation for $r_{\min}(\lambda_1)$. Solving for $r_{\max}(\lambda_1)$ and $r_{\min}(\lambda_1)$, we get

$$r_{\max}(\lambda_1) = \frac{1}{2}(y - \sqrt{y^2 + 4\varepsilon^2 - 4y\varepsilon \csc \theta}), \tag{3-10}$$

$$r_{\min}(\lambda_1) = \frac{1}{2}(\sqrt{y^2 + 4\varepsilon^2 + 4y\varepsilon \csc \theta} - y). \tag{3-11}$$

For ε small, we can use the approximation $(1 + \varepsilon)^p = 1 + p\varepsilon + \mathcal{O}(\varepsilon^2)$. Then,

$$\begin{aligned} \frac{r_{\max}(\lambda_1)}{r_{\min}(\lambda_1)} &= \frac{1 - \sqrt{1 + 4(\varepsilon/y)^2 - 4(\varepsilon/y) \csc \theta}}{\sqrt{1 + 4(\varepsilon/y)^2 + 4(\varepsilon/y) \csc \theta} - 1} \\ &= \frac{1 + \eta\varepsilon + \mathcal{O}(\varepsilon^2)}{1 - \eta\varepsilon + \mathcal{O}(\varepsilon^2)}, \end{aligned} \tag{3-12}$$

where $\eta = (\cos \theta \cot \theta)/y$. Using the geometric series approximation $1/(1 - x) = 1 + x + \mathcal{O}(x^2)$, we find that

$$\frac{r_{\max}}{r_{\min}} = 1 + \frac{(2 \cos \theta \cot \theta)\varepsilon}{y} + \mathcal{O}(\varepsilon^2). \tag{3-13}$$

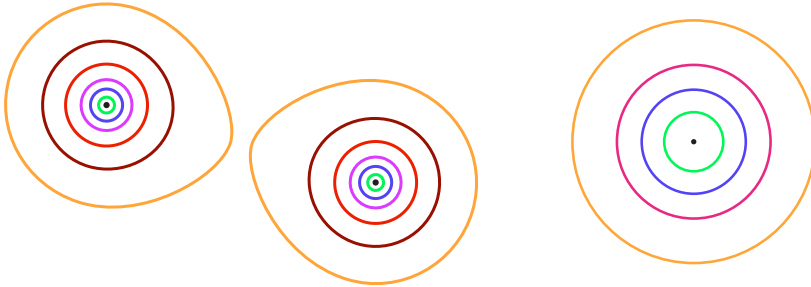


Figure 4. The ε -pseudospectra of a diagonalizable 2×2 matrix.

Thus, $\sigma_\varepsilon(A)$ tends towards a disk. Moreover, if A is diagonalizable but not normal, then the eigenvectors are linearly independent but not orthogonal, so θ is not a multiple of $\pi/2$ or π , and therefore $\cos \theta \cot \theta \neq 0$ and $(r_{\max}(\lambda))/r_{\min}(\lambda) \neq 1$. \square

This result can be observed by looking at plots of the pseudospectra of diagonalizable 2×2 matrices.

The image on the left in Figure 4 shows the pseudospectra of a particular 2×2 matrix. One can see that for large enough values of ε , the pseudospectra around either eigenvalue are not perfect disks. The image on the right is the pseudospectra of the same matrix (restricted to one eigenvalue), with smaller values of epsilon. Here, the pseudospectra appear to converge to disks. We find that this result holds in general for every $N \times N$ matrix and this is proven in the following section.

4. Asymptotic union of disks theorem

In Propositions 3.1 and 3.2, we showed that the ε -pseudospectra for all 2×2 matrices are disks or asymptotically converge to a union of disks. We now explore whether this behavior holds in the general case. It is possible to find matrices whose ε -pseudospectra exhibit pathological properties for large ε ; for example, the nondiagonalizable matrix given in Figure 5 has, for larger ε , an ε -pseudospectrum that is not convex and not simply connected.

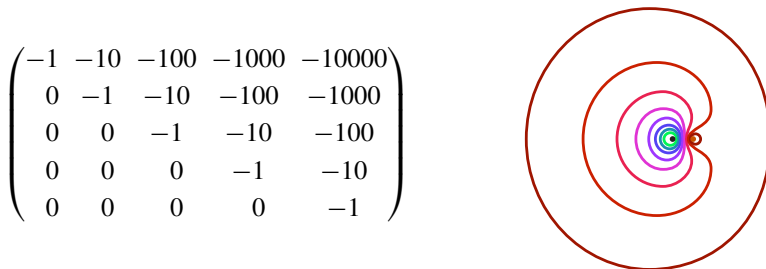


Figure 5. A Toeplitz matrix and its pseudospectra.

Thus, pseudospectra may behave poorly for large enough ε ; however, in the limit as $\varepsilon \rightarrow 0$, these properties disappear and the pseudospectra behave as disks centered around the eigenvalues with well-understood radii. In order to understand this asymptotic behavior, we use the following set-up (which follows [Moro et al. 1997]).

Let $A \in \mathbb{C}^{N \times N}$ and fix $\lambda \in \sigma(A)$. Write the Jordan decomposition of A as

$$\begin{pmatrix} J & \\ & \hat{J} \end{pmatrix} = \begin{pmatrix} Q \\ \hat{Q} \end{pmatrix} A (P \ \hat{P}), \quad \begin{pmatrix} Q \\ \hat{Q} \end{pmatrix} (P \ \hat{P}) = I,$$

where J consists of Jordan blocks J_1, \dots, J_m corresponding to the eigenvalue λ , and \hat{J} consists of Jordan blocks corresponding to the other eigenvalues of A .

Let n be the size of the largest Jordan block corresponding to λ , and suppose there are ℓ Jordan blocks corresponding to λ of size $n \times n$. Arrange the Jordan blocks in J in weakly decreasing order, according to size. That is,

$$\dim(J_1) = \dots = \dim(J_\ell) > \dim(J_{\ell+1}) \geq \dots \geq \dim(J_m),$$

where J_1, \dots, J_ℓ are $n \times n$.

Further partition P ,

$$P = (P_1 \ \dots \ P_\ell \ \dots \ P_m),$$

in a way that agrees with the above partition of J , so that the first column, x_j , of each P_j is a right eigenvector of A associated with λ . We also partition Q likewise,

$$Q = \begin{pmatrix} Q_1 \\ \vdots \\ Q_\ell \\ \vdots \\ Q_m \end{pmatrix}.$$

The last row, y_j , of each Q_j is a left eigenvector of A corresponding to λ .

We now build the matrices

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_\ell \end{pmatrix}, \quad X = (x_1 \ x_2 \ \dots \ x_\ell),$$

where X and Y are the matrices of right and left eigenvectors, respectively, corresponding to the Jordan blocks of maximal size for λ .

The following theorem is presented by Moro, Burke, and Overton [Moro et al. 1997] and due to Lidskiĭ [1966].

Theorem 4.1 [Lidskiĭ 1966]. *Given ℓ, n as defined above corresponding to the matrix A , there are ℓn eigenvalues of the perturbed matrix $A + \varepsilon E$ admitting a first-order expansion*

$$\lambda_{j,k}(\varepsilon) = \lambda + (\gamma_j \varepsilon)^{1/n} + o(\varepsilon^{1/n})$$

for $j = 1, \dots, \ell$, and $k = 1, \dots, n$, where γ_j are the eigenvalues of YEX and the different values of $\lambda_{j,k}(\varepsilon)$ for $k = 1, \dots, n$ are defined by taking the distinct n -th roots of γ_j .

Lidskii’s result can be interpreted in terms of the ε -pseudospectrum of a matrix A in order to understand the radii of $\sigma_\varepsilon(A)$ as $\varepsilon \rightarrow 0$.

Theorem 4.2. *Let $A \in \mathbb{C}^{N \times N}$. Let $\varepsilon > 0$. Given $\lambda \in \sigma(A)$, for ε small enough, there exists a connected component $U \subseteq \sigma_\varepsilon(A)$ such that $U \cap \sigma(A) = \lambda$; denote this component of the ε -pseudospectrum by $\sigma_\varepsilon(A) \upharpoonright_\lambda$.*

Then, as $\varepsilon \rightarrow 0$,

$$B(\lambda, (C\varepsilon)^{1/n} + o(\varepsilon^{1/n})) \subseteq \sigma_\varepsilon(A) \upharpoonright_\lambda \subseteq B(\lambda, (C\varepsilon)^{1/n} + o(\varepsilon^{1/n})),$$

where $C = \|XY\|$, with X, Y defined above, and n is the size of the largest Jordan block corresponding to λ .

Proof. Lower bound: Give $E \in \mathbb{C}^{N \times N}$, and let $\gamma_{\max}(E)$ be the largest eigenvalue of YEX . It is shown [Moro et al. 1997, Theorem 4.2] that

$$\alpha := \max_{\|E\| \leq 1} \gamma_{\max}(E) = \|XY\|.$$

Moreover, the E that maximizes γ is given by $E = vu$, where v and u are the right and left singular vectors of the largest singular value of XY , normalized so $\|v\| = \|u\| = 1$. We claim that $B(\lambda, (|\alpha|\varepsilon)^{1/n} + o(\varepsilon^{1/n})) \subseteq \sigma_\varepsilon(A) \upharpoonright_\lambda$.

Fix $E = vu$, with v, u defined above, fix $\theta \in [0, 2n\pi]$, and define $\tilde{E} = e^{i\theta} E$. Note that γ is an eigenvalue of YEX if and only if $e^{i\theta}\gamma$ is an eigenvalue of $Y\tilde{E}X$. Since α is an eigenvalue of E , we know that $e^{i\theta}\alpha$ is an eigenvalue of $Y\tilde{E}X$. Considering the perturbed matrix $A + \varepsilon\tilde{E}$, Theorem 4.1 implies that there is a perturbed eigenvalue $\lambda(\varepsilon)$ of the form

$$\lambda(\varepsilon) = \lambda + (e^{i\theta}\alpha\varepsilon)^{1/n} + o(\varepsilon^{1/n}),$$

and thus $\lambda(\varepsilon) \in \overline{\sigma_\varepsilon(A) \upharpoonright_\lambda}$. Ranging θ from 0 to $2n\pi$, we get the desired result.

Upper bound: Using the proof of [Trefethen and Embree 2005, Theorem 52.3], we know that asymptotically

$$\sigma_\varepsilon(A) \upharpoonright_\lambda \subseteq B(\lambda, (\beta\varepsilon)^{1/n} + o(\varepsilon^{1/n})),$$

where $\beta = \|PD^{n-1}Q\|$ and $J = \lambda I + D$. We claim $\beta = \|XY\| = \alpha$.

Note that $D^{n-1} = \text{diag}[\Gamma_1, \dots, \Gamma_\ell, 0]$, where Γ_k is an $n \times n$ matrix with a 1 in the top right entry and zeros elsewhere. We find

$$PD^{n-1} = (\Omega_1 \ \Omega_2 \ \cdots \ \Omega_\ell \ 0),$$

where Ω_j is a matrix whose last column is x_j with zeros elsewhere. This then gives

$$PD^{n-1}Q = \begin{pmatrix} XY & 0 \\ 0 & 0 \end{pmatrix}.$$

Thus $\beta = \|PD^{n-1}Q\| = \|XY\| = \alpha$. □

We present special cases of matrices to explore the consequences of Theorem 4.2.

Special cases.

(1) λ is simple: Then, $n = 1$ and X and Y become the right and left eigenvectors x and y^* for λ , respectively. Hence, $C = \|XY\| = \|xy^*\| = \|x\|\|y\| = \kappa(\lambda)$, where we normalize so that $|y^*x| = 1$. Then, Theorem 4.2 becomes

$$\sigma_\varepsilon(A) \upharpoonright_\lambda \approx B(\lambda, \kappa(\lambda)\varepsilon),$$

which matches with Theorem 2.9.

(2) λ has geometric multiplicity 1: In this case, we obtain the same result as when λ is simple, except n may not equal 1. In other words,

$$\sigma_\varepsilon(A) \upharpoonright_\lambda \approx B(\lambda, (\kappa(\lambda)\varepsilon)^{1/n}).$$

(3) $A \in \mathbb{C}^{2 \times 2}$: There are two cases, as in Section 3.

First, assume A is nondiagonalizable. In this case, A only has one eigenvalue, λ . Writing $A = VJV^{-1}$, where V and J are as defined in (3-2), we have that,

$$X = (a \ c)^T, \quad Y = \frac{1}{ad - bc} (-c \ a).$$

From Theorem 4.2, we then have that as $\varepsilon \rightarrow 0$,

$$\sigma_\varepsilon(A) \approx B\left(\lambda, \left(\frac{|a|^2 + |c|^2}{|ad - bc|}\varepsilon\right)^{1/2} + o(\varepsilon^{1/2})\right).$$

This agrees asymptotically with (3-4); however, (3-4) gives an explicit formula for $\sigma_\varepsilon(A)$.

In the case where A is diagonalizable, A has two eigenvalues, λ_1 and λ_2 . Again, we write $A = DVD^{-1}$, where V and D are as defined in (3-6). From this, we have

$$\|XY\| = \frac{(|a|^2 + |c|^2)(|b|^2 + |d|^2)}{|ad - bc|} = \csc \theta.$$

Thus, as $\varepsilon \rightarrow 0$, we have from Theorem 4.2 that

$$B(\lambda, (\csc \theta)\varepsilon + o(\varepsilon)) \subseteq \sigma_\varepsilon(A) \subseteq B(\lambda, (\csc \theta)\varepsilon + o(\varepsilon)).$$

So,

$$\frac{r_{\max}}{r_{\min}} = \frac{(\csc \theta)\varepsilon + o(\varepsilon)}{(\csc \theta)\varepsilon + o(\varepsilon)} = 1 + o(1).$$

This agrees with the ratio we obtain from the explicit formula for diagonalizable 2×2 matrices; however, (3-13) gives us more information on the $o(1)$ term.

(4) A is a Jordan block: From [Trefethen and Embree 2005, p. 470], we know that the ε -pseudospectrum of the Jordan block is exactly a disk about the eigenvalue of J of some radius. An explicit formula for the radius remains unknown; however, we can use Theorem 4.2 to find the asymptotic behavior.

Proposition 4.3 (asymptotic Bound). *Let J be an $N \times N$ Jordan block. Then*

$$\sigma_\varepsilon(J) = B(\lambda, \varepsilon^{1/N} + o(\varepsilon^{1/N})).$$

Proof. The $N \times N$ Jordan block has left and right eigenvectors u_j and v_j , where $\|u_j\| = 1$ and $\|v_j\| = 1$. So, from Theorem 4.2, we find $C = \|XY\| = \|v_j u_j\| = 1$. Thus,

$$\sigma_\varepsilon(J) \approx B(\lambda, \varepsilon^{1/N} + o(\varepsilon^{1/N})). \quad \square$$

By a simple computation, we can also get a better explicit lower bound on the ε -pseudospectra of an $N \times N$ Jordan block that agrees with our asymptotic bound.

Proposition 4.4. *Let J be an $N \times N$ Jordan block. Then,*

$$B(\lambda, \sqrt[N]{\varepsilon(1 + \varepsilon)^{N-1}}) \subseteq \sigma_\varepsilon(J).$$

Proof. We use the second definition for $\sigma_\varepsilon(J)$. Let

$$E = \begin{pmatrix} 0 & k & & & \\ & 0 & k & & \\ & & \ddots & \ddots & \\ & & & \ddots & k \\ k & & & & 0 \end{pmatrix},$$

where $|k| < \varepsilon$, and note that $\|E\| < \varepsilon$. We take $\det(J + E - zI)$ and set it equal to zero to find the eigenvalues of $J + E$:

$$\begin{aligned} 0 &= \det(J + E - zI) \\ &= \det \begin{pmatrix} \lambda - z & k + 1 & & & \\ & \lambda - z & k + 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & k + 1 \\ k & & & & \lambda - z \end{pmatrix} \\ &= (\lambda - z)^N + (-1)^{N-1} k (1 + k)^{N-1} \\ &= (-1)^{N-1} ((z - \lambda)^N + k(1 + k)^{N-1}), \end{aligned}$$

and we know that

$$\begin{aligned}
 0 = (-1)^{N-1}((z - \lambda)^N + k(1 + k)^{N-1}) &\iff (z - \lambda)^N = k(1 + k)^{N-1} \\
 &\iff z - \lambda = \sqrt[N]{k(1 + k)^{N-1}}.
 \end{aligned}$$

So, $B(\lambda, \sqrt[N]{\varepsilon(1 + \varepsilon)^{N-1}}) \subseteq \sigma_\varepsilon(J)$. □

5. Pseudospectra of bidiagonal matrices

In this section we consider bidiagonal matrices, a class of matrices with important applications in spectral theory and mathematical physics. We investigate the pseudospectra of periodic bidiagonal matrices and show that the powers n and the coefficients C in Theorem 4.2 can be computed explicitly. We consider the coefficients $\{a_k\}_{k=1}^N$ and $\{b_k\}_{k=1}^{N-1}$, which define the bidiagonal matrix

$$A = \text{bidiag}(\{a_k\}_{k=1}^N, \{b_k\}_{k=1}^{N-1}) = \begin{pmatrix} a_1 & b_1 & & & & \\ & a_2 & b_2 & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & \\ & & & & a_{N-1} & b_{N-1} \\ & & & & & a_N \end{pmatrix}.$$

Note that if $b_i = 0$ for some i , then the matrix A “decouples” into the direct sum

$$A = \text{bidiag}(\{a_k\}_{k=1}^i, \{b_k\}_{k=1}^{i-1}) \oplus \text{bidiag}(\{a_k\}_{k=i+1}^N, \{b_k\}_{k=i+1}^{N-1}),$$

and by Theorem 2.5 the pseudospectrum of A is the union of pseudospectra of smaller bidiagonal matrices. Therefore we can assume, without loss of generality, that $b_i \neq 0$ for every $i \in \{1, 2, \dots, N - 1\}$.

Note also that the eigenvalues of A are $\{a_1, a_2, \dots, a_n\}$ and some eigenvalues may be repeated in the list. In order to apply Theorem 4.2, we have to find the dimension of the largest Jordan block associated to each eigenvalue of the matrix A . The following proposition addresses this question:

Proposition 5.1. *Let*

$$A = \text{bidiag}(\{a_k\}_{k=1}^N, \{b_k\}_{k=1}^{N-1})$$

with $b_i \neq 0$ for every i and suppose that a is an eigenvalue of A . Then

$$\dim N(A - aI) = 1,$$

where $N(A - aI)$ is the eigenspace corresponding to the eigenvalue a of the matrix A .

Proof. Suppose $a = a_{i_1} = a_{i_2} = \dots = a_{i_m}$, where $1 \leq i_1 < i_2 < \dots < i_m \leq N$ and $a \neq a_k$ for every $k \in \{1, 2, \dots, n\} \setminus \{i_1, i_2, \dots, i_m\}$. We have

the right and left eigenvectors, respectively. We will give the explicit expressions for v_j and u_j .

We will begin by introducing ε -pseudospectra for simple special cases that lead to the most general case.

The cases will be presented as follows:

- **Case 1:** Let A be a $kn \times kn$ matrix with a_1, \dots, a_k distinct.
- **Case 2:** Let A be an $N \times N$ matrix with a_1, \dots, a_k distinct.
- **General case:** Let A be an $N \times N$ matrix with a_1, \dots, a_k not distinct.

To shorten notation for the rest of this section, we define

$$f(x) = \begin{cases} x & x \neq 0, \\ 1 & x = 0. \end{cases}$$

Case 1: The size of A is $kn \times kn$ and the a_i are distinct.

We write the elements of the superdiagonal as b_1, b_2, \dots, b_{N-1} , and we let $p = k(n - 1) + j$.

We have that

$$v_j = \begin{pmatrix} \frac{b_1 \cdots b_{j-1}}{f(a_j - a_1) \cdots f(a_j - a_{j-1})} \\ \frac{b_2 \cdots b_{j-1}}{f(a_j - a_2) \cdots f(a_j - a_{j-1})} \\ \vdots \\ \frac{b_{j-1}}{f(a_j - a_{j-1})} \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad u_j^* = \frac{1}{(f(a_1 - a_j) \cdots f(a_k - a_j))^{n-1}} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \frac{b_j \cdots b_p}{f(a_{j+1} - a_j)} \\ \vdots \\ \frac{b_j \cdots b_{N-2}}{f(a_{j+1} - a_j) \cdots f(a_{k-1} - a_j)} \\ \frac{b_j \cdots b_{N-1}}{f(a_{j+1} - a_j) \cdots f(a_k - a_j)} \end{pmatrix}^T.$$

Direct computation will show that these are indeed left and right eigenvectors associated with each eigenvalue a_j .

Case 2: The size of A is $N \times N$, and the a_i are distinct.

We relax our assumption that the size of our matrix is $kn \times kn$, for period k on the diagonal. Let n, r be such that $N = kn + r$, where $0 < r \leq k$. In other words, a_r is the last entry on the main diagonal, so the period does not necessarily complete.

For a_j , the right eigenvector is given by

$$v_j = \left(\frac{b_1 \cdots b_{j-1}}{f(a_j - a_1) \cdots f(a_j - a_{j-1})} \quad \frac{b_2 \cdots b_{j-1}}{f(a_j - a_2) \cdots f(a_j - a_{j-1})} \quad \cdots \quad \frac{b_{j-1}}{f(a_j - a_{j-1})} \quad 1 \quad 0 \quad \cdots \quad 0 \right)^T.$$

We split up the formula for the left eigenvectors into two cases: (i) $1 \leq j \leq r$ and (ii) $r < j \leq k$.

(i) $1 \leq j \leq r$. On the main diagonal, there are n complete blocks with entries a_1, \dots, a_k , and one partial block at the end with entries a_1, \dots, a_r . When $1 \leq j \leq r$, we have that a_j is in this last partial block. In this case, let $p = kn + j$.

We have that

$$u_j = \mu_j \begin{pmatrix} 0 \\ \vdots \\ 0 \\ (b_j \cdots b_{p-1}) \cdot f(a_{j+1} - a_j) \cdots f(a_r - a_j) \\ (b_j \cdots b_p) \cdot f(a_{j+2} - a_j) \cdots f(a_r - a_j) \\ \vdots \\ (b_j \cdots b_{N-2}) \cdot f(a_r - a_j) \\ b_j \cdots b_{N-1} \end{pmatrix}^T,$$

where

$$\mu_j = \frac{f(a_1 - a_j) \cdots f(a_{j-1} - a_j)}{(f(a_1 - a_j) \cdots f(a_k - a_j))^n f(a_1 - a_j) \cdots f(a_r - a_j)}.$$

(ii) $r < j \leq k$. In this case, a_j is in the last complete block. Now, let $p = k(n - 1) + j$.

We have that

$$u_j = \mu_j \begin{pmatrix} 0 \\ \vdots \\ 0 \\ (b_j \cdots b_{p-1}) \cdot f(a_1 - a_j) \cdots f(a_r - a_j) f(a_{j+1} - a_j) \cdots f(a_k - a_j) \\ (b_j \cdots b_p) \cdot f(a_1 - a_j) \cdots f(a_r - a_j) f(a_{j+2} - a_j) \cdots f(a_k - a_j) \\ \vdots \\ (b_j \cdots b_{p+k-j-1}) \cdot f(a_1 - a_j) \cdots f(a_r - a_j) \\ (b_j \cdots b_{p+k-j}) \cdot f(a_2 - a_j) \cdots f(a_r - a_j) \\ \vdots \\ (b_j \cdots b_{N-2}) \cdot f(a_r - a_j) \\ b_j \cdots b_{N-1} \end{pmatrix},$$

where

$$\mu_j = \frac{f(a_1 - a_j) \cdots f(a_{j-1} - a_j)}{(f(a_1 - a_j) \cdots f(a_k - a_j))^n f(a_1 - a_j) \cdots f(a_r - a_j)}.$$

Case 3: General case. The size of A is $N \times N$ and the a_i are not distinct for $1 \leq i \leq k$.

Let A be an $N \times N$ periodic bidiagonal matrix with period k on the main diagonal. Let n, r be such that $N = kn + r$, where $0 < r \leq k$. Write a_1, \dots, a_k for the entries on the main diagonal (the a_i not distinct) and b_1, \dots, b_{N-1} for the entries on the superdiagonal. Let a_r be the last entry on the main diagonal.

We can explicitly find the left and right eigenvectors for each eigenvalue α . Suppose α first appears in position ℓ of the period k . Then the corresponding right eigenvector for α is the same form as v_ℓ in Case 2. That is,

$$v_\ell = \left(\frac{b_1 \cdots b_{\ell-1}}{f(a_\ell - a_1) \cdots f(a_\ell - a_{\ell-1})} \quad \frac{b_2 \cdots b_{\ell-1}}{f(a_\ell - a_2) \cdots f(a_\ell - a_{\ell-1})} \quad \cdots \quad \frac{b_{\ell-1}}{f(a_\ell - a_{\ell-1})} \quad 1 \quad 0 \quad \cdots \quad 0 \right)^T.$$

The corresponding left eigenvector for α depends on the first and last positions of α . Let $k(n - 1) = \ell q + s$ and set $q \equiv m \pmod k$. We split up the formula for the left eigenvector of α into two cases, which again mirror the formulas given in Case 2:

(i) $1 \leq \ell \leq r$ and (ii) $r < \ell \leq k$.

For both of these two cases, we define

$$g(b_i) = \begin{cases} b_i & i \geq p, \\ 1 & i < p. \end{cases}$$

(i) $1 \leq \ell \leq r$. In this case, α appears in the partial block. Let $p = kn + \ell$. We have

$$u_\ell = \mu_\ell \begin{pmatrix} 0 \\ \vdots \\ 0 \\ g(b_{p+m-\ell-1})f(a_{m+1} - a_\ell) \cdots f(a_r - a_\ell) \\ g(b_{p+m-\ell-1})g(b_{p+m-l})f(a_{m+2} - a_\ell) \cdots f(a_r - a_\ell) \\ \vdots \\ g(b_{p+m-\ell-1}) \cdots g(b_{N-2})f(a_r - a_\ell) \\ g(b_{p+m-\ell-1}) \cdots g(b_{N-1}) \end{pmatrix},$$

where

$$\mu_\ell = \frac{b_\ell \cdots b_{p-1} f(a_1 - a_\ell) \cdots f(a_{\ell-1} - a_\ell)}{(f(a_1 - a_\ell) \cdots f(a_k - a_\ell))^n f(a_1 - a_\ell) \cdots f(a_r - a_\ell)}.$$

(ii) $r < \ell \leq k$. In this case, α is in the last complete block. Here, we let $p = k(n - 1) + \ell$.

Now, we have

$$u_\ell = \mu_\ell \begin{pmatrix} 0 \\ \vdots \\ 0 \\ g(b_{p+m-\ell-1})f(a_1 - a_\ell) \cdots (a_r - a_\ell)f(a_{m+1} - a_\ell) \cdots f(a_k - a_\ell) \\ g(b_{p+m-\ell-1})g(b_{p+m-l})f(a_1 - a_\ell) \cdots f(a_r - a_\ell)f(a_{m+2} - a_\ell) \cdots f(a_k - a_\ell) \\ \vdots \\ g(b_{p+m-\ell-1}) \cdots g(b_{p+k-\ell-1})f(a_1 - a_\ell) \cdots f(a_r - a_\ell) \\ g(b_{p+m-\ell-1}) \cdots g(b_{p+k-\ell})f(a_2 - a_\ell) \cdots f(a_r - a_\ell) \\ \vdots \\ g(b_{p+m-\ell-1}) \cdots g(b_{N-2})f(a_r - a_\ell) \\ g(b_{p+m-\ell-1}) \cdots g(b_{N-1}) \end{pmatrix},$$

where

$$\mu_\ell = \frac{(b_\ell \cdots b_{p-1}) \cdot f(a_1 - a_\ell) \cdots f(a_{\ell-1} - a_\ell)}{(f(a_1 - a_\ell) \cdots f(a_k - a_\ell))^n f(a_1 - a_\ell) \cdots f(a_r - a_\ell)}.$$

From these formulas, we can find the eigenvectors, and hence the asymptotic behavior of the ε -pseudospectrum for every bidiagonal matrix A :

$$\sigma_\varepsilon(A) \approx \bigcup_{j=1}^k B(a_j, (C_j \varepsilon)^{1/n_j}),$$

where $C_j = \|v_j\| \|u_j\|$ and n_j is the size of the Jordan block corresponding to a_j .

Note. Let A be a periodic, bidiagonal matrix and suppose $b_i = 0$ for some i . Then the matrix decouples into the direct sum of smaller matrices; call them A_1, \dots, A_n . To find the ε -pseudospectrum of A , apply the same analysis to these smaller matrices, and from Theorem 2.5, we have that

$$\sigma_\varepsilon(A) = \bigcup_{i=1}^n \sigma_\varepsilon(A_i).$$

6. Finite-rank operators

The majority of this paper has focused on both explicit and asymptotic characterizations of ε -pseudospectra for various classes of finite-dimensional linear operators. A natural next step is to consider finite rank operators on an infinite-dimensional space.

In Section 2 we defined ε -pseudospectra for matrices, although our definitions are exactly the same in the infinite-dimensional case. For our purposes, the only noteworthy difference between matrices and operators is that the spectrum of an operator is no longer defined as the collection of eigenvalues, but rather

$$\sigma(A) = \{\lambda \mid \lambda I - A \text{ does not have a bounded inverse}\}.$$

As a result, we do not get the same properties for pseudospectra as we did previously; in particular, $\sigma_\varepsilon(A)$ is not necessarily bounded.

That being said, the following theorem shows that finite-rank operators behave similarly to matrices in that asymptotically the radii of ε -pseudospectra are bounded by powers of epsilon. The following theorem makes this precise.

Theorem 6.1. *Let V be a Hilbert space and $A : V \rightarrow V$ a finite-rank operator on V . Then there exists C such that for sufficiently small ε ,*

$$\sigma_\varepsilon(A) \subseteq \sigma(A) + B(0, C\varepsilon^{1/(m+1)}),$$

where m is the rank of A . Furthermore, this bound is sharp in the sense that there exists a rank- m operator A and a constant c such that

$$\sigma_\varepsilon(A) \supseteq \sigma(A) + B(0, c\varepsilon^{1/(m+1)})$$

for sufficiently small ε .

Proof. Since A has finite rank, there exists a finite-dimensional subspace U such that $V = U \oplus W$ and $A(U) \subseteq U$ and $A(W) = \{0\}$. Choosing an orthonormal basis for A which respects this decomposition, we can write $A = A' \oplus 0$. Then the spectrum of A is $\sigma(A') \cup \{0\}$, and we know that for every $\varepsilon > 0$,

$$\sigma_\varepsilon(A) = \sigma_\varepsilon(A') \cup \sigma_\varepsilon(0).$$

The ε -pseudospectrum of the zero operator is well-understood since this operator is normal; for all ε , it is precisely the ball of radius ε . It thus suffices to consider the ε -pseudospectrum of the finite-rank operator $A' : U \rightarrow U$, where U is finite-dimensional. The ε -pseudospectrum of this operator goes like $\varepsilon^{1/j}$, where j is the dimension of the largest Jordan block; we will prove that $j \leq m + 1$. Note that the rank of the $n \times n$ Jordan block given by

$$A = \begin{pmatrix} \lambda & 1 & 0 & 0 & \dots \\ 0 & \lambda & 1 & 0 & \dots \\ \vdots & \vdots & \ddots & \ddots & \dots \\ \vdots & \vdots & \vdots & \ddots & 1 \\ 0 & 0 & 0 & 0 & \lambda \end{pmatrix}$$

is n if $\lambda \neq 0$ and $n - 1$ if $\lambda = 0$. Since we know that the rank of A is larger than or equal to the rank of the largest Jordan block, we have an upper bound on the dimension of the largest Jordan block: it is of size $m + 1$, with equality attained when $\lambda = 0$. By Theorem 4.2, we then know that $\sigma_\varepsilon(A)$ is contained, for small enough ε , in the set $\sigma(A) + C\varepsilon^{1/(m+1)}$.

Note that this bound is sharp; we can see this by taking V to be \mathbb{R}^{m+1} and considering the rank- m operator

$$A_m = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ \vdots & \vdots & \ddots & \ddots & \dots \\ \vdots & \vdots & \vdots & \ddots & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

whose pseudospectrum will contain the ball of radius $\varepsilon^{1/(m+1)}$ by Proposition 4.4. \square

Open Questions. The natural question to ask now is whether we can extend this result to more arbitrary operators on Hilbert spaces. In particular, for a bounded operator A , we would like to establish if there exists a continuous function $r_A(\varepsilon)$ such that for sufficiently small ε ,

$$\sigma_\varepsilon(A) \subseteq \sigma(A) + B(0, r_A(\varepsilon)).$$

For a matrix A , we proved in Theorem 4.2 that $r_A(\varepsilon) = C\varepsilon^{1/n}$, where n is the size of the largest Jordan block associated to A , and C is a constant that depends on

the left and right eigenvectors associated to a certain eigenvalue. For a finite-rank operator A , we proved in Theorem 6.1 that $r_A(\varepsilon) = C\varepsilon^{1/(m+1)}$, where m is the rank of the operator and C is as above.

For closed but not necessarily bounded operators, the picture is more complex, as the spectrum need not be bounded or even nonempty. For example, the operator $A : u \mapsto u'$ in $L^2[0, 1]$ with domain $D(A)$ being the set of absolutely continuous functions on $[0, 1]$ satisfying $u(1) = 0$ has empty spectrum. When $D(A)$ is the entire space, the spectrum of A is the entire complex plane. Davies [1999a] also provides an example of an unbounded operator with unbounded pseudospectrum.

Given these examples, we can see that Theorem 6.1 will not generalize to unbounded operators, as the pseudospectrum of an unbounded operator may be unbounded for all ε .

Nonetheless, we do still have a certain convergence of the ε -pseudospectrum to the spectrum [Trefethen and Embree 2005, Section 4], namely $\bigcap_{\varepsilon>0} \sigma_\varepsilon(A) = \sigma(A)$. Also, while the ε -pseudospectrum may be unbounded, each bounded component of it necessarily contains a component of the spectrum. These results imply that the bounded components of the ε -pseudospectrum must converge to the spectrum. Therefore, if we restrict our attention to these bounded components, we can attempt to generalize Theorems 4.2 and 6.1 by asking whether the *bounded* components of $\sigma_\varepsilon(A)$ converge to the spectrum as a union of disks.

Acknowledgements

Support for this project was provided by the National Science Foundation REU Grant DMS-0850577 and DMS-1347804, the Clare Boothe Luce Program of the Henry Luce Foundation, and the SMALL REU at Williams College.

References

- [Bauer and Fike 1960] F. L. Bauer and C. T. Fike, “Norms and exclusion theorems”, *Numer. Math.* **2** (1960), 137–141. MR 0118729 Zbl 0101.25503
- [Baumgärtel 1985] H. Baumgärtel, *Analytic perturbation theory for matrices and operators*, Operator Theory: Advances and Applications **15**, Birkhäuser, Basel, 1985. MR 878974 Zbl 0591.47013
- [Böttcher 2006] A. Böttcher, “Review of “Spectra and pseudospectra: The behavior of nonnormal matrices and operators” by L. N. Trefethen and M. Embree”, *Linear Algebra Appl.* **416**:2-3 (2006), 1098–1101.
- [Davies 1999a] E. B. Davies, “Pseudo-spectra, the harmonic oscillator and complex resonances”, *R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci.* **455**:1982 (1999), 585–599. MR 1700903 Zbl 0931.70016
- [Davies 1999b] E. B. Davies, “Semi-classical states for non-self-adjoint Schrödinger operators”, *Comm. Math. Phys.* **200**:1 (1999), 35–41. MR 1671904 Zbl 0921.47060
- [Kato 1995] T. Kato, *Perturbation theory for linear operators*, Springer, Berlin, 1995. MR 1335452 Zbl 0836.47009

[Lidskiĭ 1966] V. B. Lidskiĭ, “On the theory of perturbations of nonselfadjoint operators”, *Ž. Vyčisl. Mat. i Mat. Fiz.* **6**:1 (1966), 52–60. In Russian; translated in *{USSR} Computational Mathematics and Mathematical Physics* **6**:1 (1966), 73–85. MR 0196930 Zbl 0166.40501

[Moro et al. 1997] J. Moro, J. V. Burke, and M. L. Overton, “On the Lidskiĭ–Vishik–Lyusternik perturbation theory for eigenvalues of matrices with arbitrary Jordan structure”, *SIAM J. Matrix Anal. Appl.* **18**:4 (1997), 793–817. MR 1471994 Zbl 0889.15016

[Trefethen and Embree 2005] L. N. Trefethen and M. Embree, *Spectra and pseudospectra: The behavior of nonnormal matrices and operators*, Princeton Univ. Press, 2005. MR 2155029 Zbl 1085.15009

Received: 2015-05-25 Revised: 2015-06-22 Accepted: 2015-06-23

fgong@mit.edu *Department of Mathematics and Statistics, Williams College,
Williamstown, MA 01267, United States*

olivia.s.meyerson@gmail.com *Department of Mathematics and Statistics, Williams College,
Williamstown, MA 01267, United States*

jdmeza@alumni.cmu.edu *Department of Mathematics, Carnegie Mellon University,
Pittsburgh, PA 15213, United States*

mstoiciu@williams.edu *Department of Mathematics and Statistics, Williams College,
Williamstown, MA 01267, United States*

abigailward@uchicago.edu *Department of Mathematics, The University of Chicago,
Chicago, IL 60637, 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

2016

vol. 9

no. 3

A combinatorial proof of a decomposition property of reduced residue systems	361
YOTSANAN MEEMARK AND THANAKORN PRINYASART	
Strong depth and quasigeodesics in finitely generated groups	367
BRIAN GAPINSKI, MATTHEW HORAK AND TYLER WEBER	
Generalized factorization in $\mathbb{Z}/m\mathbb{Z}$	379
AUSTIN MAHLUM AND CHRISTOPHER PARK MOONEY	
Cocircular relative equilibria of four vortices	395
JONATHAN GOMEZ, ALEXANDER GUTIERREZ, JOHN LITTLE, ROBERTO PELAYO AND JESSE ROBERT	
On weak lattice point visibility	411
NEIL R. NICHOLSON AND REBECCA RACHAN	
Connectivity of the zero-divisor graph for finite rings	415
REZA AKHTAR AND LUCAS LEE	
Enumeration of m -endomorphisms	423
LOUIS RUBIN AND BRIAN RUSHTON	
Quantum Schubert polynomials for the G_2 flag manifold	437
RACHEL E. ELLIOTT, MARK E. LEWERS AND LEONARDO C. MIHALCEA	
The irreducibility of polynomials related to a question of Schur	453
LENNY JONES AND ALICIA LAMARCHE	
Oscillation of solutions to nonlinear first-order delay differential equations	465
JAMES P. DIX AND JULIO G. DIX	
A variational approach to a generalized elastica problem	483
C. ALEX SAFSTEN AND LOGAN C. TATHAM	
When is a subgroup of a ring an ideal?	503
SUNIL K. CHEBOLU AND CHRISTINA L. HENRY	
Explicit bounds for the pseudospectra of various classes of matrices and operators	517
FEIXUE GONG, OLIVIA MEYERSON, JEREMY MEZA, MIHAI STOICIU AND ABIGAIL WARD	

