

involve

a journal of mathematics

Editorial Board

Kenneth S. Berenhaut, *Managing Editor*

Colin Adams
Arthur T. Benjamin
Martin Bohner
Amarjit S. Budhiraja
Pietro Cerone
Scott Chapman
Joshua N. Cooper
Jem N. Corcoran
Toka Diagana
Michael Dorff
Sever S. Dragomir
Joel Foisy
Errin W. Fulp
Joseph Gallian
Stephan R. Garcia
Anant Godbole
Ron Gould
Sat Gupta
Jim Haglund
Johnny Henderson
Glenn H. Hurlbert
Charles R. Johnson
K. B. Kulasekera
Gerry Ladas
David Larson
Suzanne Lenhart

Chi-Kwong Li
Robert B. Lund
Gaven J. Martin
Mary Meyer
Frank Morgan
Mohammad Sal Moslehian
Zuhair Nashed
Ken Ono
Yuval Peres
Y.-F. S. Pétermann
Jonathon Peterson
Robert J. Plemmons
Carl B. Pomerance
Vadim Ponomarenko
Bjorn Poonen
Józeph H. Przytycki
Richard Rebarber
Robert W. Robinson
Javier Rojo
Filip Saidak
Hari Mohan Srivastava
Andrew J. Sterge
Ann Trenk
Ravi Vakil
Antonia Vecchio
John C. Wierman
Michael E. Zieve



involve

msp.org/involve

INVOLVE YOUR STUDENTS IN RESEARCH

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

MANAGING EDITOR

Kenneth S. Berenhaut Wake Forest University, USA

BOARD OF EDITORS

Colin Adams	Williams College, USA	Robert B. Lund	Clemson University, USA
Arthur T. Benjamin	Harvey Mudd College, USA	Gaven J. Martin	Massey University, New Zealand
Martin Bohner	Missouri U of Science and Technology, USA	Mary Meyer	Colorado State University, USA
Amarjit S. Budhiraja	U of N Carolina, Chapel Hill, USA	Frank Morgan	Williams College, USA
Pietro Cerone	La Trobe University, Australia	Mohammad Sal Moslehian	Ferdowsi University of Mashhad, Iran
Scott Chapman	Sam Houston State University, USA	Zuhair Nashed	University of Central Florida, USA
Joshua N. Cooper	University of South Carolina, USA	Ken Ono	Univ. of Virginia, Charlottesville
Jem N. Corcoran	University of Colorado, USA	Yuval Peres	Microsoft Research, USA
Toka Diagana	University of Alabama in Huntsville, USA	Y.-F. S. Pétermann	Université de Genève, Switzerland
Michael Dorff	Brigham Young University, USA	Jonathon Peterson	Purdue University, USA
Sever S. Dragomir	Victoria University, Australia	Robert J. Plemmons	Wake Forest University, USA
Joel Foisy	SUNY Potsdam, USA	Carl B. Pomerance	Dartmouth College, USA
Errin W. Fulp	Wake Forest University, USA	Vadim Ponomarenko	San Diego State University, USA
Joseph Gallian	University of Minnesota Duluth, USA	Bjorn Poonen	UC Berkeley, USA
Stephan R. Garcia	Pomona College, USA	József H. Przytycki	George Washington University, USA
Anant Godbole	East Tennessee State University, USA	Richard Rebarber	University of Nebraska, USA
Ron Gould	Emory University, USA	Robert W. Robinson	University of Georgia, USA
Sat Gupta	U of North Carolina, Greensboro, USA	Javier Rojo	Oregon State University, USA
Jim Haglund	University of Pennsylvania, USA	Filip Saidak	U of North Carolina, Greensboro, USA
Johnny Henderson	Baylor University, USA	Hari Mohan Srivastava	University of Victoria, Canada
Glenn H. Hurlbert	Virginia Commonwealth University, USA	Andrew J. Sterge	Honorary Editor
Charles R. Johnson	College of William and Mary, USA	Ann Trenk	Wellesley College, USA
K. B. Kulasekera	Clemson University, USA	Ravi Vakil	Stanford University, USA
Gerry Ladas	University of Rhode Island, USA	Antonia Vecchio	Consiglio Nazionale delle Ricerche, Italy
David Larson	Texas A&M University, USA	John C. Wierman	Johns Hopkins University, USA
Suzanne Lenhart	University of Tennessee, USA	Michael E. Zieve	University of Michigan, USA
Chi-Kwong Li	College of William and Mary, USA		

PRODUCTION

Silvio Levy, Scientific Editor

Cover: Alex Scorpan

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2020 is US \$205/year for the electronic version, and \$275/year (+\$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP.

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

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

PUBLISHED BY

 **mathematical sciences publishers**
nonprofit scientific publishing

<http://msp.org/>

© 2020 Mathematical Sciences Publishers

Structured sequences and matrix ranks

Charles Johnson, Yaoxian Qu, Duo Wang and John Wilkes

(Communicated by Stephan Garcia)

We consider infinite sequences from a field and all matrices whose rows consist of distinct consecutive subsequences. We show that these matrices have bounded rank if and only if the sequence is a homogeneous linear recurrence; moreover, it is a k -term linear recurrence if and only if the maximum rank is k . This means, in particular, that the ranks of matrices from the sequence of primes are unbounded. Though not all matrices from the primes have full rank, because of the Green–Tao theorem, we conjecture that square matrices whose entries are a consecutive sequence of primes do have full rank.

1. Introduction

A familiar way to obtain a simple example of a rank-deficient matrix is to array the positive integers as a matrix, wrapping at the end of each row.

For example,

$$\text{rank} \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} = 2.$$

For $m, n \geq 2$, if we start at any integer and array mn consecutive integers as an m -by- n matrix in this way, the rank is 2. In fact, every submatrix of size at least 2-by-2 will have rank 2.

For convenience, we consider *real sequences* a_1, a_2, a_3, \dots , though much of what we say holds over any field. We refer to this sequence as a . The sequence a is a k -term linear recurrence if it is of the form

$$a_{i+k} = b_1 a_i + b_2 a_{i+1} + b_3 a_{i+2} + \cdots + b_k a_{i+k-1},$$

in which b_i is a constant, $i = 1, \dots, k$ [Brousseau 1971]. If we suppress the value of k , we just say “linear recurrence”. Some authors refer to the above as a homogeneous linear recurrence, whereas a nonhomogeneous linear recurrence

MSC2010: primary 15A03; secondary 11B25, 11B37.

Keywords: k -term linear recurrence, prime numbers, row extension, column extension, rank, matrix of a sequence.

allows a constant term. However, a k -term nonhomogeneous linear recurrence is also a $2k$ -term homogeneous one (which also follows from our Theorem 2.3).

A geometric progression is just the case when $k = 1$, and arithmetic progressions are a special case of 2-term linear recurrences. The positive integers are a very simple arithmetic progression.

In general, if a sequence is a k -term (linear) recurrence, it is entirely determined by its first k terms, and the collection of such sequences (with fixed coefficients) is k -dimensional, including degenerate cases.

Definition 1.1. $M = (m_{ij})$ is a *traditional matrix* of the sequence a if its m_{11} entry is a_j and the remaining entries are consecutive:

$$M_{m,n}(j) = \begin{pmatrix} a_j & a_{j+1} & \cdots & a_{j+n-1} \\ a_{j+n} & a_{j+n+1} & \cdots & a_{j+2n-1} \\ \vdots & & & \vdots \\ a_{j+(m-1)+n} & a_{j+m+n} & \cdots & a_{j+mn-1} \end{pmatrix}.$$

Definition 1.2. M is a *nontraditional matrix* of the sequence a if each row is simply a consecutive subsequence and each begins with a distinct element of the sequence.

For example,

$$M = \begin{pmatrix} a_3 & a_4 & a_5 & a_6 \\ a_{10} & a_{11} & a_{12} & a_{13} \\ a_5 & a_6 & a_7 & a_8 \\ a_{20} & a_{21} & a_{22} & a_{23} \end{pmatrix}.$$

In either event, we say that M is a matrix of the sequence. For any infinite sequence a , let $M(a)$ be the set of all matrices of the sequence a .

2. Linear recurrences

We now consider sequences that are linear k -term recurrences and we consider the matrices of such sequences. Sometimes, we suppress the “ k ” and just refer to a “linear recurrence”. All facts about ranks of matrices that we use may be found, for example, in [Lay 2006]. We first note:

Lemma 2.1. *If the sequence a is any k -term linear recurrence, then for any matrix $A \in M(a)$, we have $\text{rank}(A) \leq k$.*

Proof. Let

$$A = \begin{pmatrix} a_j & a_{j+1} & \cdots & a_{j+k-1} & a_{j+k} & \cdots & a_{j+n-1} \\ a_i & a_{i+1} & \cdots & a_{i+k-1} & a_{i+k} & \cdots & a_{i+n-1} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ a_m & a_{m+1} & \cdots & a_{m+k-1} & a_{m+k} & \cdots & a_{m+n-1} \end{pmatrix},$$

where

$$\begin{aligned} a_{j+k} &= b_1 a_j + b_2 a_{j+1} + \cdots + b_k a_{j+k-1}, \\ a_{i+k} &= b_1 a_i + b_2 a_{i+1} + \cdots + b_k a_{i+k-1}, \\ &\vdots \\ a_{m+k} &= b_1 a_m + b_2 a_{m+1} + \cdots + b_k a_{m+k-1}. \end{aligned}$$

So the $(k+1)$ -th column is a linear combination of the first k columns. Similarly, columns after the $(k+1)$ -th column are linear combinations of the k columns before them. Thus at most the first k columns are linearly independent, so $\text{rank}(A) \leq k$. \square

Example 2.2. The Fibonacci sequence is a 2-term linear recurrence. Here is a traditional matrix of the Fibonacci sequence:

$$\text{rank} \begin{pmatrix} 1 & 1 & 2 & 3 \\ 5 & 8 & 13 & 21 \\ 34 & 55 & 89 & 144 \\ 233 & 377 & 610 & 987 \end{pmatrix} = 2.$$

Theorem 2.3. *A sequence a is a linear recurrence if and only if matrices in $M(a)$ are of bounded rank. Moreover, if the largest rank of a matrix in $M(a)$ is k , then the linear recurrence is a k -term one.*

Proof. The forward implication is implied by Lemma 2.1.

For the converse, the bounded-rank hypothesis means that there is a largest rank among matrices in $M(a)$; call it k . Choose a matrix $A \in M(a)$ such that

- $\text{rank}(A) = k$,
- no other matrix in $M(a)$ with rank k has fewer columns,
- given this number of columns, no matrix with rank k has fewer rows.

Suppose that A is m -by- n . Notice that any matrix in $M(a)$ of which A is a submatrix must have rank k . Suppose that $A' \in M(a)$ is m -by- $(n+1)$ and that the first n columns of A' are A , and the last column of A' is a linear combination of the first n columns:

$$a_{i+n} = b_1 a_i + b_2 a_{i+1} + \cdots + b_n a_{i+n-1} \quad (*)$$

(for any row of A' of which a_i is the first entry) and $\dim \text{Nul}(A') = n+1-k$. Now, if we consider any $\bar{A} \in M(a)$ whose first m rows are those of A' , $\text{rank}(\bar{A}) = k$, we have $\dim \text{Nul}(\bar{A}) = n+1-k$, and, since $\text{Nul}(\bar{A}) \subseteq \text{Nul}(A')$, we have $\text{Nul}(\bar{A}) = \text{Nul}(A')$. This means that any allowed row added to A' must satisfy the same linear relations as satisfied collectively by the rows of A' . But, we were allowed to add any row, as long as that row was a consecutive subsequence of a (of length $n+1$) that was not already a row of A' . This means that a satisfies $(*)$ for any $i \geq 1$, and thus a is a linear recurrence, which completes the proof. \square

3. Ranks less than k in k -term sequences

So far we know that if a is a linear recurrence, then it is a k -term (and not less) linear recurrence if and only if the maximum rank among matrices in $M(a)$ is k . We might refer to a “rank- k linear recurrence” to be absolutely unambiguous, as a recurrence presented as a k -term linear recurrence might actually have lower rank and be an h -term recurrence for some $h < k$.

This leaves an interesting question of detail. What about the ranks of m -by- n matrices in $M(a)$, when $k \leq m, n$; can they have rank $< k$, and, if so, under what circumstances? Notice that $M(a)$ is closed under the extraction of submatrices in which the column indices are consecutive and that row indices may as well be taken to be consecutive. So, we are really asking about ranks of contiguous submatrices of matrices in $M(a)$. Are they always full, subject to being no more than k ?

Interestingly, they are, with some modest exceptions; we consider in detail here the case $k = 2$.

Example 3.1. Let a be the sequence which is generated by the 2-term recurrence

$$a_{i+2} = -a_i - \sqrt{2}a_{i+1}.$$

But

$$\begin{pmatrix} a_1 & a_2 \\ a_5 & a_6 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}$$

lies in $M(a)$ and has rank 1.

Suppose we have a 2-term recurrence $C_1a_n + C_2a_{n+1} = a_{n+2}$, and suppose that there exists a contiguous submatrix in $M(a)$ with rank 1:

$$\text{rank} \begin{pmatrix} a_i & a_{i+1} \\ a_j & a_{j+1} \end{pmatrix} = 1, \quad j > i.$$

Lemma 3.2. *If we right-extend the submatrix above by adding columns with subsequent elements or left-extend the submatrix by adding columns with preceding elements in the recurrence, then the rank of the submatrix does not change, and*

$$\begin{pmatrix} a_{i+k} \\ a_{j+k} \end{pmatrix} = r_k \begin{pmatrix} a_{i+k-1} \\ a_{j+k-1} \end{pmatrix}, \quad k \in \mathbb{Z}, \quad r_k \neq 0,$$

where

$$r_{n+1} = \frac{C_1}{r_n} + C_2, \quad r_1 = b,$$

in which r_k is the ratio between entries in one column of the submatrix and the next.

Proof. Since

$$\text{rank} \begin{pmatrix} a_i & a_{i+1} \\ a_j & a_{j+1} \end{pmatrix} = 1,$$

we have

$$\begin{pmatrix} a_{i+1} \\ a_{j+1} \end{pmatrix} = b \begin{pmatrix} a_i \\ a_j \end{pmatrix},$$

and since

$$C_1 \begin{pmatrix} a_i \\ a_j \end{pmatrix} + C_2 \begin{pmatrix} a_{i+1} \\ a_{j+1} \end{pmatrix} = \begin{pmatrix} a_{i+2} \\ a_{j+2} \end{pmatrix},$$

we have

$$\frac{C_1}{b} \begin{pmatrix} a_{i+1} \\ a_{j+1} \end{pmatrix} + C_2 \begin{pmatrix} a_{i+1} \\ a_{j+1} \end{pmatrix} = \begin{pmatrix} a_{i+2} \\ a_{j+2} \end{pmatrix}.$$

Thus

$$\left(\frac{C_1}{b} + C_2 \right) \begin{pmatrix} a_{i+1} \\ a_{j+1} \end{pmatrix} = \begin{pmatrix} a_{i+2} \\ a_{j+2} \end{pmatrix}.$$

Suppose $k = 1$; then

$$\left(\frac{C_1}{b} + C_2 \right) \begin{pmatrix} a_{i+1} \\ a_{j+1} \end{pmatrix} = \begin{pmatrix} a_{i+2} \\ a_{j+2} \end{pmatrix}.$$

Suppose $k = x$; then

$$r_x \begin{pmatrix} a_{i+x-1} \\ a_{j+x-1} \end{pmatrix} = \begin{pmatrix} a_{i+x} \\ a_{j+x} \end{pmatrix}.$$

Since

$$C_1 \begin{pmatrix} a_{i+x-1} \\ a_{j+x-1} \end{pmatrix} + C_2 \begin{pmatrix} a_{i+x} \\ a_{j+x} \end{pmatrix} = \begin{pmatrix} a_{i+x+1} \\ a_{j+x+1} \end{pmatrix},$$

we have

$$\frac{C_1}{r_x} \begin{pmatrix} a_{i+x} \\ a_{j+x} \end{pmatrix} + C_2 \begin{pmatrix} a_{i+x} \\ a_{j+x} \end{pmatrix} = \begin{pmatrix} a_{i+x+1} \\ a_{j+x+1} \end{pmatrix}.$$

Thus

$$r_{x+1} \begin{pmatrix} a_{i+x} \\ a_{j+x} \end{pmatrix} = \begin{pmatrix} a_{i+x+1} \\ a_{j+x+1} \end{pmatrix}$$

when $k = x + 1$. Thus for $k \in \mathbb{N}$, we have

$$\begin{pmatrix} a_{i+k} \\ a_{j+k} \end{pmatrix} = r_k \begin{pmatrix} a_{i+k-1} \\ a_{j+k-1} \end{pmatrix}$$

by mathematical induction. Similarly, we can easily see that for $k \in \mathbb{Z}$,

$$\begin{pmatrix} a_{i+k} \\ a_{j+k} \end{pmatrix} = r_k \begin{pmatrix} a_{i+k-1} \\ a_{j+k-1} \end{pmatrix}.$$

Thus, every added column is linearly dependent to the previous one, which means that the submatrix is still rank-1. \square

Lemma 3.3. *There must exist k such that $r_1 = C_1/r_k + C_2$.*

Proof. Since extending the submatrix does not change its rank, the following matrix is still rank-1:

$$\begin{pmatrix} a_i & a_{i+1} & \cdots & a_j & a_{j+1} \\ a_j & a_{j+1} & \cdots & a_{2j-i} & a_{2j-i+1} \end{pmatrix}. \quad (1)$$

Thus,

$$r_{j-i+1} \begin{pmatrix} a_j \\ a_{2j-i+1} \end{pmatrix} = \begin{pmatrix} a_{j+1} \\ a_{2j-i+1} \end{pmatrix}.$$

Since $a_{j+1}/a_j = b$, we have $r_{j-i+1} = b$. Thus, when $k = j - i + 1$, we have $r_1 = C_1/r_k + C_2$. \square

Since $r_{n+1} = C_1/r_n + C_2$, we know that $\begin{pmatrix} r_{n+1} \\ 1 \end{pmatrix}$ must be parallel to

$$\begin{pmatrix} C_2 & C_1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} r_n \\ 1 \end{pmatrix}.$$

Thus,

$$\begin{pmatrix} r_{n+1} \\ 1 \end{pmatrix} = P_n \begin{pmatrix} C_2 & C_1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} r_n \\ 1 \end{pmatrix}, \quad P_n = \frac{1}{r_n}.$$

Hence

$$\begin{pmatrix} r_n \\ 1 \end{pmatrix} = P_1 P_2 \cdots P_n \begin{pmatrix} C_2 & C_1 \\ 1 & 0 \end{pmatrix}^n \begin{pmatrix} r_1 \\ 1 \end{pmatrix}.$$

According to Lemma 3.3, there must exist k such that

$$\begin{pmatrix} r_1 \\ 1 \end{pmatrix} = \begin{pmatrix} r_k \\ 1 \end{pmatrix} = P_1 P_2 \cdots P_k \begin{pmatrix} C_2 & C_1 \\ 1 & 0 \end{pmatrix}^k \begin{pmatrix} r_1 \\ 1 \end{pmatrix},$$

which means that $\begin{pmatrix} r_1 \\ 1 \end{pmatrix}$ is an eigenvector of

$$\begin{pmatrix} C_2 & C_1 \\ 1 & 0 \end{pmatrix}^k.$$

Theorem 3.4. *If there exists a rank-1 contiguous submatrix in $M(a)$, then a does not degenerate into a geometric recurrence only when*

$$\begin{pmatrix} C_2 & C_1 \\ 1 & 0 \end{pmatrix}^k =: C^k$$

is diagonalizable and has distinct eigenvalues.

Proof. Suppose that C^k is diagonalizable. If C^k does not have distinct eigenvalues, then C must have exactly the same eigenvectors as C^k . Thus $\begin{pmatrix} r_1 \\ 1 \end{pmatrix}$ is an eigenvector of C . Thus

$$\begin{pmatrix} r_1 \\ 1 \end{pmatrix} C = \lambda \begin{pmatrix} r_1 \\ 1 \end{pmatrix},$$

and hence

$$r_1^2 = C_2 r_1 + C_1, \quad r_1 = \frac{C_1}{r_1} + C_2, \quad r_1 = r_2.$$

When $n = 1$, we have $r_n = r_1$ and when $n = 2$, we have $r_n = r_1$; therefore suppose when $n = x$ we have $r_n = r_1$.

Thus $C_1/r_x + C_2 = r_{x+1} = r_1$ when $n = x + 1$, $r_n = r_1$, so $r_n = r_1$ for $n \in \mathbb{N}$. Similarly, we can see that $r_n = r_1$ for $n \in \mathbb{Z}$. Thus the ratio between every pair of consecutive elements in a is b , which means that a is a geometric progression.

When C^k is not diagonalizable, C^k does not have the same eigenvalues since it has only one eigenvalue. Thus C has exactly the same eigenvalue as C^k . Thus $\begin{pmatrix} r_1 \\ 1 \end{pmatrix}$ is an eigenvector of C . Thus $r_1 = r_2$, and as described above, we can see that a is a geometric progression.

Thus, the case that a does not degenerate into a geometric recurrence, illustrated in Example 3.1, can only happen when C^k is diagonalizable and has distinct eigenvalues. \square

4. Prime sequences

We now turn to the sequence of primes and subsequences thereof. Let p be the sequence: 2, 3, 5, 7, 11, 13, 17, 19, 21, The primes are not a linear recurrence and seem arithmetically very unstructured. We generated many examples of traditional matrices of primes (up to 100-by-100) and they were always full rank. This suggests:

Conjecture 4.1. If A is an m -by- n traditional matrix of primes, then $\text{rank } A = \min\{m, n\}$.

This seems difficult to prove, perhaps for the following reason. Suppose that q is a subsequence of p . Again, experiments always turned up full rank for traditional matrices of q .

Observation 4.2. Because of unique factorization, every m -by- n submatrix, with $m, n \geq 2$, of a traditional matrix of q has at least rank 2. Now 2-by-2 minors are of the form $q_i q_j - q_t q_s$ with i, j, t, s distinct and, so, cannot be 0, else $q_i q_j$ and $q_t q_s$ would be distinct prime factorizations of the same integer. However, there exist subsequences and an m -by- n traditional matrix A with $m, n > 2$ such that $\text{rank}(A) = 2$. For instance,

$$\text{rank} \begin{pmatrix} 3 & 5 & 13 \\ 7 & 11 & 29 \\ 17 & 31 & 79 \end{pmatrix} = 2.$$

The above example is not isolated.

Theorem 4.3 [Green and Tao 2008]. *The prime numbers contain infinitely many arithmetic progressions of length k for each positive integer $k > 1$.*

This means that $M(p)$ contains many “large” matrices of “low rank”.

Though we cannot prove Conjecture 4.1, a weaker version does follow from our results about linear recurrences,

Corollary 4.4. *The set of ranks of all matrices of the sequence p of primes is unbounded.*

Proof. Since p is an infinite sequence, according to Theorem 4.3, if the matrices in $M(p)$ had bounded rank, then $M(p)$ would be a linear recurrence. This is not true. Thus, p must include matrices of unbounded rank. \square

References

- [Brousseau 1971] A. Brousseau, *Linear recursion and Fibonacci sequences*, Fibonacci Association, San Jose, CA, 1971.
- [Green and Tao 2008] B. Green and T. Tao, “The primes contain arbitrarily long arithmetic progressions”, *Ann. of Math. (2)* **167**:2 (2008), 481–547. MR Zbl
- [Lay 2006] D. C. Lay, *Linear algebra and its applications*, 3rd ed., Pearson/Addison-Wesley, Boston, MA, 2006.

Received: 2016-05-09 Revised: 2016-09-13 Accepted: 2017-04-09

crjohn@wm.edu	<i>Department of Mathematics, The College of William & Mary, Williamsburg, VA, United States</i>
quyaoxian@gmail.com	<i>Department of Mathematics, University of Notre Dame, South Bend, IN, United States</i>
dwang02@email.wm.edu	<i>Department of Mathematics, The College of William & Mary, Williamsburg, VA, United States</i>
jcwilkes@email.wm.edu	<i>Department of Mathematics, The College of William & Mary, Williamsburg, VA, United States</i>

Analysis of steady states for classes of reaction-diffusion equations with hump-shaped density-dependent dispersal on the boundary

Quinn Morris, Jessica Nash and Catherine Payne

(Communicated by Kenneth S. Berenhaut)

We study a two-point boundary-value problem describing steady states of a population dynamics model with diffusion, logistic growth, and nonlinear density-dependent dispersal on the boundary. In particular, we focus on a model in which the population exhibits hump-shaped density-dependent dispersal on the boundary, and explore its effects on existence, uniqueness and multiplicity of steady states.

1. Introduction

Since the early work of Fisher [1937] and Kolmogorov, Petrovskii, and Piskunov [Kolmogorov et al. 1937], differential equations models have been used to model the dynamics of a population inhabiting a patch. Since their early work, the models have grown in complexity with the goal of understanding long-term persistence of populations through the analysis of such models. In a one-dimensional patch $\Omega = (0, \ell)$ for some $\ell > 0$ such models take the form

$$\begin{aligned} u_t &= Du_{\tilde{x}\tilde{x}} + u\tilde{f}(u), & \tilde{x} \in \Omega, t > 0, \\ B_1u(0, t) &= 0, & t > 0, \\ B_2u(\ell, t) &= 0, & t > 0, \\ u(\tilde{x}, 0) &= u_0(\tilde{x}), \end{aligned} \tag{1}$$

where $u(\tilde{x}, t)$ is the population density at point $\tilde{x} \in \Omega$ at time $t > 0$, D is a diffusion rate within the habitat, $\tilde{f} : (0, \infty) \rightarrow (-\infty, \infty)$ is the per-capita growth rate, B_1 and B_2 are boundary operators (possibly nonlinear), and $u_0 : [0, \ell] \rightarrow [0, \infty)$ is an initial population distribution. Such models are often referred to as “reaction-diffusion models” as the rate of change of the population density (u_t) is determined by both a

MSC2010: 34B18, 34C60, 92D25.

Keywords: differential equation, mathematical ecology, nonlinear dispersal, nonlinear boundary condition, logistic equation, reaction-diffusion equation, density-dependent dispersal.

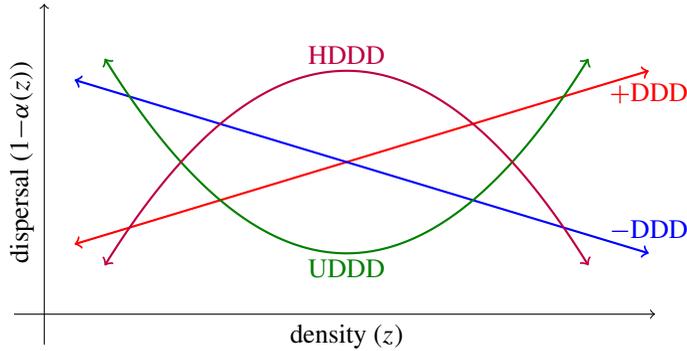


Figure 1. Graphs of possible density-dispersal relations, where z is population density and $1 - \alpha(z)$ is the probability of dispersal from the patch upon reaching a point on the boundary of where the local population density is z .

reaction term ($u\tilde{f}(u)$), which in this case models the growth of the population, and a diffusion term ($Du_{\tilde{x}\tilde{x}}$), which models movement of the population. The diffusion term $Du_{\tilde{x}\tilde{x}}$ for population movement can be derived from a individual random walk model [Skellam 1951] and is a good fit for modeling movement of many different species [Kareiva 1983; Turchin and Thoeny 1993]. See [Cantrell and Cosner 2003] for the derivation and analysis of a number of models of this form.

In this paper, we consider a reaction-diffusion model of a population in a primary patch Ω which is surrounded by a secondary region (called the matrix, in ecology) with hostility $S^* > 0$ ($S^* \approx 0$ indicates that the matrix is relatively not hostile, while $S^* \gg 0$ indicates that the matrix is relatively more hostile). Such scenarios are common in habitats experiencing fragmentation, whereby large regions of primary habitat are broken into smaller fragments either by destruction of parts of the primary habitat or replacement of the primary habitat by a less suitable matrix. Individuals within the primary patch move via diffusion, and if they reach the boundary, choose to remain in the patch with probability $\alpha(z)$, where $\alpha : [0, \infty) \rightarrow (0, \infty)$ is a function which depends on the population density at the boundary. We note that $1 - \alpha(z)$ gives the dispersal rate of the population from the primary patch, a parameter of interest to ecologists.

There are three main types of density-dependent dispersal $1 - \alpha(z)$ that have been studied; see Figure 1. Specifically, species that demonstrate positive density-dependent dispersal (+DDD) have a low dispersal rate for low densities and a high dispersal rate for high densities. Species of this kind are most likely experiencing crowding effects or lack of resources. Comparatively, for species demonstrating negative density-dependent dispersal (-DDD), there is a higher chance for dispersal when there is a smaller population density. This may be because the species is



Figure 2. Examples of species which have been observed to exhibit $-DDD$ at low densities and $+DDD$ at high densities. Left: blue-footed booby, *Sula nebouxii* (see [Kim et al. 2009]). Image by Bernard Gagnon/CC BY-SA 3.0. Right: Glanville fritillary butterfly, *Melitaea cinxia* (see [Kuussaari et al. 1998] for $-DDD$, and see [Enfjäll and Leimar 2005] for $+DDD$). Photo by Christian Fischer/CC BY-SA 3.0.

experiencing mate scarcity or conspecific attraction. There is also some evidence of species that exhibit $-DDD$ at low densities and $+DDD$ at high densities, which we refer to as u-shaped density-dependent dispersal (UDDD). UDDD has been of recent interest in the mathematical literature, see [Cantrell and Cosner 2003; Cantrell et al. 1998; Cronin et al. 2019; Fonseca et al. 2019; Goddard et al. 2018; ≥ 2020], as ecologists find evidence that the blue-footed booby and Glanville fritillary butterfly (see Figure 2) exhibit UDDD in nature. In this paper, we are interested in studying species that demonstrate hump-shaped density-dependent dispersal (HDDD), in which species exhibit $+DDD$ at low densities and $-DDD$ at high densities.

We further assume that individuals in the population exhibit logistic growth. In particular, we assume the per-capita growth rate takes the form

$$\tilde{f}(u) = r \left(1 - \frac{u}{K} \right),$$

where $r > 0$ is the maximum population growth rate and $K > 0$ is the carrying capacity such that $\tilde{f}(u) > 0$ for $u \in (0, K)$ and $\tilde{f}(u) < 0$ for $u \in (K, \infty)$.

Based on these assumptions, the resulting model is [Cronin et al. 2019]

$$\begin{aligned} u_t &= Du_{\tilde{x}\tilde{x}} + ru \left(1 - \frac{u}{K} \right), \quad \tilde{x} \in \Omega, \quad t > 0, \\ -D\alpha(u(0, t))u_{\tilde{x}}(0, t) + S^*[1 - \alpha(u(0, t))]u(0, t) &= 0, \\ D\alpha(u(\ell, t))u_{\tilde{x}}(\ell, t) + S^*[1 - \alpha(u(\ell, t))]u(\ell, t) &= 0, \end{aligned} \tag{2}$$

where $u(\tilde{x}, t)$ is the population density at point $\tilde{x} \in \Omega$ at time $t > 0$ and K is the carrying capacity of the population. Rescaling the spatial variable by taking $x = \tilde{x}/\ell$,

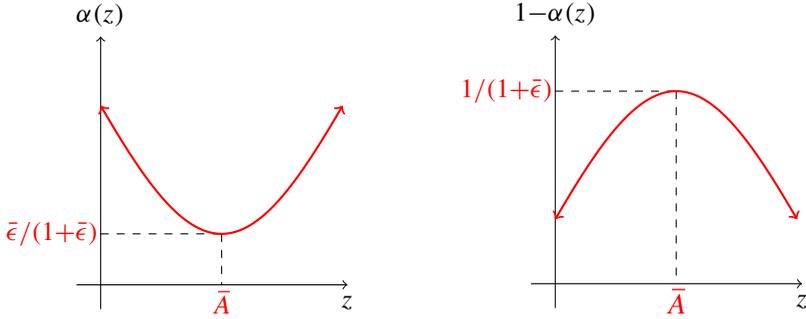


Figure 3. Illustration of relationship between $\alpha(z)$ (left) and dispersal $[1 - \alpha(z)]$ (right). Note that \bar{A} is the density at which the probability of remaining in the patch upon reaching the boundary is at a minimum (in particular, $\alpha(\bar{A}) = \bar{\epsilon}/(1 + \bar{\epsilon})$). Therefore, dispersal on the boundary reaches a maximum of $1 - \alpha(\bar{A}) = 1/(1 + \bar{\epsilon})$ when density is \bar{A} .

we have $u(\tilde{x}, t) = u(x\ell, t) = z(x, t)$, and the model becomes

$$\begin{aligned} z_t &= \frac{D}{\ell^2} z_{xx} + rz \left(1 - \frac{z}{K}\right), \quad x \in (0, 1), \quad t > 0, \\ -\frac{D}{\ell} \alpha(z(0, t)) z_x(0, t) + S^* [1 - \alpha(z(0, t))] z(0, t) &= 0, \\ \frac{D}{\ell} \alpha(z(1, t)) z_x(1, t) + S^* [1 - \alpha(z(1, t))] z(1, t) &= 0, \end{aligned} \quad (3)$$

where $z(x, t)$ is the population density at point $x \in (0, 1)$ at time $t > 0$.

Of particular importance in understanding the dynamics of (3) are the steady states ($\lim_{t \rightarrow \infty} z(x, t)$) of (3). We take $\alpha(z)$ of the form

$$\alpha(z) = \frac{(z - \bar{A})^2 + \bar{\epsilon}}{(z - \bar{A})^2 + (1 + \bar{\epsilon})}$$

for $0 < \bar{A} < K$ and $\bar{\epsilon} > 0$ and examine if steady states of (3) exist. In Figure 3, we note that dispersal $1 - \alpha(z)$ is hump-shaped, as desired, and that the parameter \bar{A} determines the range of population densities on which dispersal is positive density-dependent and negative density-dependent. In particular, if $\bar{A} \approx 0$, then the population exhibits positive density-dependent dispersal for most densities, while if $\bar{A} \approx K$, the population exhibits negative density-dependent dispersal for most densities.

Letting

$$v = \frac{z}{K}, \quad \lambda = \frac{r\ell^2}{D}, \quad \gamma = \frac{S^*}{K^2\sqrt{rD}}, \quad A = \frac{\bar{A}}{K}, \quad \epsilon = \frac{\bar{\epsilon}}{K},$$

we perform a nondimensionalization which yields the one-dimensional problem

$$\begin{aligned} -v'' &= \lambda v(1-v), \quad x \in (0, 1), \\ v'(0) &= \sqrt{\lambda}\gamma \frac{v(0)}{(v(0)-A)^2 + \epsilon}, \\ v'(1) &= -\sqrt{\lambda}\gamma \frac{v(1)}{(v(1)-A)^2 + \epsilon}. \end{aligned} \tag{4}$$

In this paper, we will examine only solutions which are symmetric about $x = \frac{1}{2}$. Note that the following lemma establishes that positive solutions are, in fact, symmetric about $x = \frac{1}{2}$ when $\epsilon > 1 - A^2$.

Lemma 1. *If $\epsilon > 1 - A^2$, then any positive solution v to (4) is symmetric about $x = \frac{1}{2}$; that is, $v(\frac{1}{2} - x) = v(\frac{1}{2} + x)$ for $x \in [0, \frac{1}{2}]$.*

Proof. First, we observe from (4) that any positive solution v of (4) is concave and has exactly one maximum value $x_0 \in (0, 1)$. Furthermore, the solution v of (4) is symmetric about x_0 .

Suppose $x_0 < \frac{1}{2}$, or equivalently $2x_0 < 1$. By symmetry of the solution, let $q_1 = v(2x_0) = v(0) \leq 1$. Then $|v'(2x_0)| = \sqrt{\lambda}\gamma q_1 / ((q_1 - A)^2 + \epsilon)$. Now, let $v(1) = q_2 < q_1$. Then $|v'(1)| = \sqrt{\lambda}\gamma q_2 / ((q_2 - A)^2 + \epsilon)$.

By the assumption that $\epsilon > 1 - A^2$ and $q \leq 1$, we have

$$\frac{d}{dq} \left[\frac{q}{(q-A)^2 + \epsilon} \right] = \frac{A^2 + \epsilon - q^2}{((q-A)^2 + \epsilon)^2} \geq \frac{A^2 + \epsilon - 1}{((q-A)^2 + \epsilon)^2} > 0.$$

Therefore, since $q_2 < q_1$, we must have

$$\frac{q_2}{(q_2 - A)^2 + \epsilon} < \frac{q_1}{(q_1 - A)^2 + \epsilon},$$

which implies $|v'(1)| < |v'(2x_0)|$. But this is a contradiction, since v is concave, and therefore $q_1 = q_2$ and $x_0 = \frac{1}{2}$. \square

For simplicity of notation, we will take $f(s) = s(1-s)$ (the nonlinearity appearing in the differential equation in (4)), and let $F(s) = \int_0^s f(t) dt$. We establish the following theorem:

Theorem 2. *There exists a positive, symmetric solution v to (4) with $\|v\|_\infty = \rho$, $v(1) = q$, and $0 < q < \rho$ if and only if*

$$2 \left(\int_q^\rho \frac{dz}{\sqrt{F(\rho) - F(z)}} \right)^2 = \lambda \tag{5}$$

and

$$\frac{\gamma q}{(q-A)^2 + \epsilon} = \sqrt{2} \sqrt{F(\rho) - F(q)} \tag{6}$$

hold.

In Section 2, we prove Theorem 2 via a quadrature method first introduced in [Laetsch 1971], and recently extended to nonlinear boundary conditions in [Goddard et al. 2018; ≥ 2020]. In Section 3, we provide computationally generated bifurcation curves of (4), and in Section 4, we discuss the biological implications of our numerical results.

Remark 3. We emphasize here that Theorem 2 refers only to positive, symmetric solutions of (4). In the case that $\epsilon > 1 - A^2$, by Lemma 1 all positive solutions are symmetric. If, however, $\epsilon \leq 1 - A^2$, then there may be additional nonsymmetric solutions which are not captured by Theorem 2.

2. Proof of Theorem 2

To prove Theorem 2, we proceed via a quadrature method [Laetsch 1971]. We first show that if v is a positive solution to (4) with $\|v\|_\infty = \rho$ and $v(0) = q = v(1)$, then λ , ρ , and q must satisfy (5) and (6). Multiplying both sides of (4) by $v'(x)$ yields

$$\left(\frac{-[v'(x)]^2}{2} \right)' = \lambda(F(v(x)))'. \quad (7)$$

Since $v(\frac{1}{2}) = \rho$ is a maximum and therefore $v'(\frac{1}{2}) = 0$, integrating (7) from s to $\frac{1}{2}$ and rearranging terms gives

$$v'(s) = \sqrt{2\lambda} \sqrt{F(\rho) - F(v(s))}. \quad (8)$$

Integrating (8) from 0 to x and recalling that $v(0) = q$, we obtain

$$\int_q^{v(x)} \frac{dz}{\sqrt{F(\rho) - F(z)}} = \sqrt{2\lambda} x. \quad (9)$$

When $x = \frac{1}{2}$, we have

$$\int_q^\rho \frac{dz}{\sqrt{F(\rho) - F(z)}} = \sqrt{\frac{\lambda}{2}},$$

and hence (5) is satisfied.

Substituting $x = 0$ into (8) and applying the boundary condition in (4) yields

$$\frac{\gamma q}{(q - A)^2 + \epsilon} = \sqrt{2} \sqrt{F(\rho) - F(q)}. \quad (10)$$

Hence, (6) is also satisfied.

We now prove the reverse implication. Let λ , ρ , and q satisfy (5) and (6). We define $v : [0, 1] \rightarrow [0, \rho]$ by (9) if $x \in (0, \frac{1}{2})$, $v(0) = q$, and $v(x) = v(x - \frac{1}{2})$ if $x \in (\frac{1}{2}, 1]$.

Note that $\sqrt{2\lambda}x$ increases from 0 to

$$\int_q^\rho \frac{dz}{\sqrt{F(\rho) - F(z)}}$$

as x increases from 0 to $\frac{1}{2}$. Similarly,

$$\int_q^v \frac{dz}{\sqrt{F(\rho) - F(z)}}$$

increases from 0 to

$$\int_q^\rho \frac{dz}{\sqrt{F(\rho) - F(z)}}$$

as v increases from q to ρ . Hence, $v(x)$ is well defined for $x \in (0, \frac{1}{2})$.

Defining $H : (0, \frac{1}{2}) \times (q, \rho) \rightarrow \mathbb{R}$ by

$$H(\tau, v) = \int_q^v \frac{dz}{\sqrt{F(\rho) - F(s)}} - \sqrt{2\lambda}\tau,$$

we observe that $H \in C^1((0, \frac{1}{2}) \times (q, \rho))$, $H(x, v(x)) = 0$ for $x \in (0, \frac{1}{2})$, and

$$\left. \frac{\partial H}{\partial v} \right|_{(t, v(t))} = \frac{1}{\sqrt{F(\rho) - F(v(t))}} > 0.$$

By the implicit function theorem, we may therefore conclude that $v \in C^1(0, \frac{1}{2})$. From (9), we may now write

$$v'(x) = \sqrt{2\lambda[F(\rho) - F(v(x))]}, \quad x \in (0, \frac{1}{2}), \quad (11)$$

and since $F \in C^1(q, \rho)$ and $v \in C^1(0, \frac{1}{2})$, we observe that $v' \in C^1(0, \frac{1}{2})$. Differentiating again, we find that

$$-v''(x) = \lambda f(v(x)), \quad x \in (0, \frac{1}{2}).$$

Hence, since f is continuous, $v(\frac{1}{2}) = \rho$ and $v'(\frac{1}{2}) = 0$, we conclude that $v \in C^2(0, 1) \cap C^1[0, 1]$ by our extension of v . Moreover, from (11), we observe that $v'(0) = \sqrt{2\lambda[F(\rho) - F(q)]}$ and from (5), this implies

$$v'(0) = \sqrt{\lambda}\gamma \frac{v(0)}{(v(0) - A)^2 + \epsilon}.$$

Our extension of v guarantees that

$$v'(1) = -\sqrt{\lambda}\gamma \frac{v(1)}{(v(1) - A)^2 + \epsilon}.$$

Hence, v is a positive solution to (4) with $v(0) = q = v(1)$ and $v(\frac{1}{2}) = \rho$ as desired.

3. Numerical results

Using Mathematica, we may numerically solve for (λ, ρ) pairs which simultaneously satisfy (5) and (6). To do so, we implement the following algorithm:

- (1) Choose $\rho \in (0, 1)$.
- (2) For a given ρ , use nonlinear solver to solve (6) for q .
- (3) For a given ρ and its corresponding value for q , use (5) to find λ .
- (4) Plot (λ, ρ) pair.

We select 500 equally spaced values for ρ in $(0, 1)$. For a fixed ρ , we use the FindRoot command in Mathematica, which employs Newton's method, to solve (6) for q with an accuracy of 10^{-4} . Some examples of bifurcation curves with a fixed value of $\epsilon = .1$, and different values for γ and A are provided in Figures 4–7. Parameter values are chosen to illustrate the variety of possible bifurcation curves, but are not otherwise known to be of ecological significance.

We remind the reader, as in Remark 3, that the bifurcation curves provided in Figures 4–7 are the bifurcation curves for positive, symmetric solutions only. Any reference in this section to a solution should be interpreted as a reference to a positive, symmetric solution.

We first illustrate in Figure 4 how the number of solutions to (4) may depend on A . For $\epsilon = .1$ and $\gamma = .1$, we plot in Figure 4 the bifurcation curve for two differing values of A . When $A = .5$, we observe that for sufficiently small λ , there are no solutions, while for λ sufficiently large, there is a unique solution. We notice, however, that when A is decreased to $A = .25$, we now have multiple positive, symmetric solutions for a range of λ .

Secondly, we illustrate in Figures 5 and 6 how multiplicity of solutions may occur. In Figure 5, we first observe how the structure of these ranges of λ for which multiple solutions exist can also vary depending on A . Specifically, in Figure 5, left, for sufficiently small λ there are no solutions, then for the proceeding ranges of λ there is one solution, three solutions, and one solution. Similarly, in Figure 5,

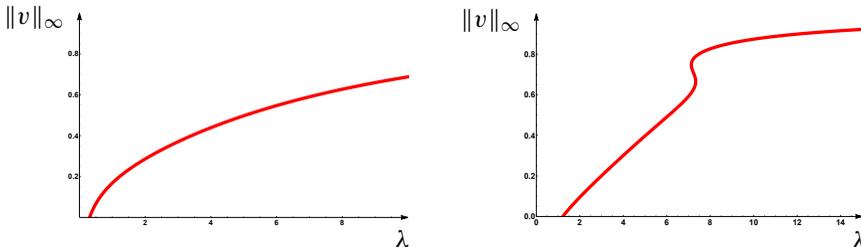


Figure 4. For $\gamma = .1$, we observe that varying values of A introduce multiplicity of solutions: $A = .5$ (left) and $A = .25$ (right).

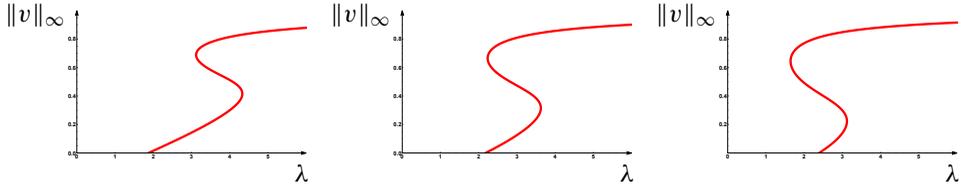


Figure 5. For $\gamma = .1$, we observe that the multiplicity of solutions and the range of λ for which certain multiplicities exist change as A is varied: $A = .15$ (left), $A = .1$ (middle), $A = .05$ (right).

middle, for sufficiently small λ there are no solutions, then for the proceeding values of λ there are three solutions then one solution. In Figure 5, right, for sufficiently small λ there are no solutions, then for the proceeding ranges of λ there are two solutions, three solutions, and one solution.

Building from Figure 5, we illustrate in Figure 6 the multiplicity of solutions that can occur for certain parameter regimes. When $\gamma = \epsilon = A = .1$, we numerically solve (5), (6) for ρ, q when $\lambda = 3$. We find solutions

$$\begin{aligned}
 (\rho_1, q_1) &= (0.12558361681607486, 0.08633543422628163), \\
 (\rho_2, q_2) &= (0.4731562619350849, 0.38016022177071546), \\
 (\rho_3, q_3) &= (0.8091464421614853, 0.7490510239916139).
 \end{aligned}$$

In Figure 6, we illustrate the correspondence between the multiple points $(3, \rho_i)$, $i = 1, 2, 3$, on the bifurcation curve and the multiple solutions v_i , $i = 1, 2, 3$, of (4), with $v_i(0) = q_i = v_i(1)$ and $\|v_i\|_\infty = v_i(\frac{1}{2}) = \rho_i$.

Finally, we illustrate in Figure 7 how large values of γ may force uniqueness of solutions. For $\epsilon = .1$ and $\gamma = 10$, we observe that even a large change in the value of A does not change the shape of the graph. Thus, we conjecture that if $\gamma \gg 1$, there exists $\lambda^* > 0$, so that (4) has no solution for $0 < \lambda < \lambda^*$, has a unique solution for every $\lambda > \lambda^*$, and $\rho(\lambda)$ is strictly increasing.

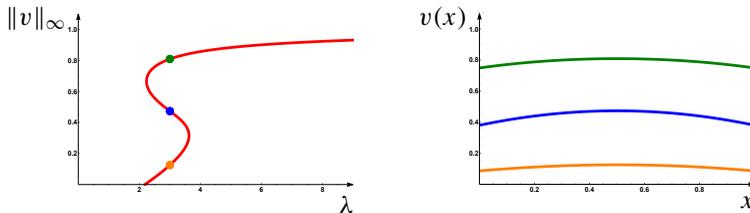


Figure 6. For $\gamma = \epsilon = A = .1$, we observe there are multiple solutions when $\lambda = 3$. Left: bifurcation curve highlighting points $(3, \rho_i)$, $i = 1, 2, 3$. Right: solutions v_i of (4) with $\lambda = 3$.

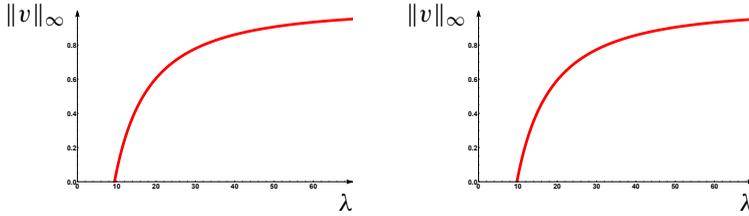


Figure 7. For $\gamma = 10$, we observe that even large perturbations of the parameter A do not introduce multiplicity of solutions: $A = .5$ (left), $A = .000005$ (right).

4. Biological interpretation

Given the numerical results in Section 3, we consider now the biological implications. Recall that $\lambda = r\ell^2/D$ and $\gamma = S^*/\sqrt{rD}$, which shows that λ is proportional to the square of the patch size ℓ and γ is proportional to the hostility of the matrix S^* .

First, we observe that, in all numerical cases explored, regardless of our choices of γ or A , there exist $\lambda^* > 0$ such that (4) has no solution for $\lambda < \lambda^*$. In biological terms, there exists some minimum patch size $\ell^* = \sqrt{D\lambda^*/r}$ that will support a population. Furthermore, in all cases, we observe that for λ sufficiently large, there exists a unique steady state. This illustrates a decreasing role of boundary effects on a population as the habitat grows larger. We note that such behavior is consistent with theoretical results proved in [Fonseka et al. 2019] in the case of UDDD.

Secondly, we observe that when γ is sufficiently small, the parameter A has a large influence on the shape of the bifurcation curve, in particular, by introducing multiple steady states. This phenomena is biologically interesting, as this implies that the population density that induces maximum dispersal on the boundary also determines whether multiple steady states of the population exist, and if so, for what ranges of patch sizes the steady states persist. On the other hand, when γ is sufficiently large, we observe that changes in the parameter A do not influence the shape of the bifurcation curve, and in particular, do not introduce multiple steady states. This suggests that as the outside matrix becomes increasingly hostile, any individual leaving the population dies almost immediately, and thus, the habitat behaves almost identically to one with a lethal boundary.

Acknowledgements

This project was initiated when the authors were at the University of North Carolina at Greensboro under the direction of Professor R. Shivaji and was supported by the NSF grant (DMS-1516560). We thank Shivaji for his guidance throughout this project.

References

- [Cantrell and Cosner 2003] R. S. Cantrell and C. Cosner, *Spatial ecology via reaction-diffusion equations*, John Wiley & Sons, Chichester, 2003. MR Zbl
- [Cantrell et al. 1998] R. S. Cantrell, C. Cosner, and W. F. Fagan, “Competitive reversals inside ecological reserves: the role of external habitat degradation”, *J. Math. Biol.* **37**:6 (1998), 491–533. MR Zbl
- [Cronin et al. 2019] J. T. Cronin, J. Goddard, and R. Shivaji, “Effects of patch–matrix composition and individual movement response on population persistence at the patch level”, *Bull. Math. Biol.* **81**:10 (2019), 3933–3975. MR
- [Enfjäll and Leimar 2005] K. Enfjäll and O. Leimar, “Density-dependent dispersal in the Glanville fritillary, *Melitaea cinxia*”, *Oikos* **108**:3 (2005), 465–473.
- [Fisher 1937] R. A. Fisher, “The wave of advance of advantageous genes”, *Ann. Eugenics* **7**:4 (1937), 355–369. JFM
- [Fonseka et al. 2019] N. Fonseka, J. Goddard, II, Q. Morris, and R. Shivaji, “On the effects of the exterior matrix hostility and a u-shaped density dependent dispersal on a diffusive logistic growth model”, preprint, 2019. To appear in *Discrete Cont. Dyn. Syst. Ser. S*.
- [Goddard et al. 2018] J. Goddard, II, Q. Morris, R. Shivaji, and B. Son, “Bifurcation curves for singular and nonsingular problems with nonlinear boundary conditions”, *Electron. J. Differential Equations* (2018), art. id. 26. MR Zbl
- [Goddard et al. \geq 2020] J. Goddard, II, J. Price, and R. Shivaji, “Analysis of steady states for classes of reaction-diffusion equations with u-shaped density dependent dispersal on the boundary”, in preparation.
- [Kareiva 1983] P. M. Kareiva, “Local movement in herbivorous insects: applying a passive diffusion model to mark-recapture field experiments”, *Oecologia* **57**:3 (1983), 322–327.
- [Kim et al. 2009] S.-Y. Kim, R. Torres, and H. Drummond, “Simultaneous positive and negative density-dependent dispersal in a colonial bird species”, *Ecology* **90**:1 (2009), 230–239.
- [Kolmogorov et al. 1937] A. Kolmogorov, I. Petrovskii, and N. Piskunov, “Study of a diffusion equation that is related to the growth of a quality of matter and its application to a biological problem”, *Byul. Mosk. Gos. Univ. Ser. A Mat. Mekh.* **1**:6 (1937), 1–26. In Russian. Zbl
- [Kuussaari et al. 1998] M. Kuussaari, I. Saccheri, M. Camara, and I. Hanski, “Allee effect and population dynamics in the glanville fritillary butterfly”, *Oikos* **82**:2 (1998), 384–392.
- [Laetsch 1971] T. Laetsch, “The number of solutions of a nonlinear two point boundary value problem”, *Indiana Univ. Math. J.* **20** (1971), 1–13. MR Zbl
- [Skellam 1951] J. G. Skellam, “Random dispersal in theoretical populations”, *Biometrika* **38** (1951), 196–218. MR Zbl
- [Turchin and Thoeny 1993] P. Turchin and W. Thoeny, “Quantifying dispersal of southern pine beetles with mark-recapture experiments and a diffusion model”, *Ecol. Appl.* **3**:1 (1993), 187–198.

Received: 2018-05-14 Revised: 2019-06-13 Accepted: 2019-10-02

morrisqa@appstate.edu *Swarthmore College Swarthmore, PA, United States*

Current address: *Appalachian State University, Boone, NC, United States*

jjnash@uw.edu *The University of North Carolina at Greensboro, Greensboro, NC, United States*

payneca@wssu.edu *Department of Mathematics, Winston-Salem State University, Winston-Salem, NC, United States*

The L -move and Markov theorems for trivalent braids

Carmen Caprau, Gabriel Coloma and Marguerite Davis

(Communicated by Kenneth S. Berenhaut)

The L -move for classical braids extends naturally to trivalent braids. We follow the L -move approach to the Markov theorem to prove a one-move Markov-type theorem for trivalent braids. We also reformulate this L -move Markov theorem and prove a more algebraic Markov-type theorem for trivalent braids. Along the way, we provide a proof of the Alexander theorem analogue for spatial trivalent graphs and trivalent braids.

1. Introduction

The Alexander [1923] and Markov [1936] theorems are fundamental results in classical knot theory. The Alexander theorem states that any oriented link is isotopic to the closure of some braid (which is not unique). The Markov theorem characterizes braids that yield isotopic links via the closure operation. Specifically, the closures of two classical braids represent isotopic links if and only if the braids are related by a finite sequence consisting of braid isotopy and two additional moves, usually referred to as *Markov moves*: conjugation by a crossing and bottom right stabilization.

There is another type of braid move, the so-called L -move, which was introduced by S. Lambropoulou [1993]; see also [Lambropoulou and Rourke 1997]. This move replaces the two moves of the Markov equivalence and yields a one-move Markov-type theorem for oriented links.

As an extension of classical knot theory, spatial graph theory seeks to classify, up to isotopy, embeddings of graphs in three-space. In this paper we focus on oriented spatial trivalent graphs whose vertices are neither sources nor sinks. Just as oriented links can be represented by diagrams that are closures of braids, so can oriented spatial trivalent graphs be represented by closures of trivalent braids (with the same number of top and bottom endpoints). We borrow the L -move approach and show that the L -move can be extended to the setting of trivalent braids. Then we prove

MSC2010: primary 57M15, 57M25; secondary 20F36.

Keywords: L -moves, Markov-type moves, spatial trivalent graphs, trivalent braids.

that this type of move and trivalent braid isotopy are sufficient to prove an L -move Markov-type theorem for trivalent braids. With this theorem at hand, we are able to state and prove an algebraic Markov-type theorem for trivalent braids.

We remark that the L -move was extended to other diagrammatic situations, including virtual braids [Kauffman and Lambropoulou 2006], virtual singular braids [Caprau et al. 2016], and virtual trivalent braids [Caprau et al. 2019].

The paper is organized as follows: We start with a brief discussion about spatial trivalent graphs, trivalent braids, and trivalent braid isotopy in Section 2. In Section 3 we describe our preparation for braiding and a braiding algorithm for spatial trivalent graphs, which implicitly proves the Alexander-type theorem for spatial trivalent graphs. We introduce in Section 4.1 the TL -equivalence among trivalent braids, and use it in Section 4.2 to prove our one-move Markov-type theorem based on the L -move for trivalent braids. Finally, we close with Section 4.3, where we state and prove a more algebraic Markov-type theorem for trivalent braids.

2. Trivalent braids and spatial trivalent graphs

In this section, we briefly review spatial trivalent graphs and trivalent braids.

A trivalent graph is a finite graph whose vertices have valency three and a *spatial trivalent graph* (or shortly *STG*) is a trivalent graph embedded in \mathbb{R}^3 . Two spatial trivalent graphs are called *ambient isotopic* if there exists an orientation-preserving self-homeomorphism on \mathbb{R}^3 taking one graph onto the other.

When studying STGs we work with their diagrams. A *diagram* of a spatial trivalent graph is a projection of a spatial trivalent graph into a plane. It is well known that two spatial trivalent graphs are ambient isotopic if and only if their diagrams are related by planar isotopy and a finite sequence of the moves $R1$ – $R5$ depicted in Figure 1; see for example [Kauffman 1989]. Note that, for the $R3$ move, the sliding strand may be either an understrand or overstrand. We refer to these moves as the *extended Reidemeister moves* for STG diagrams. In addition, if two

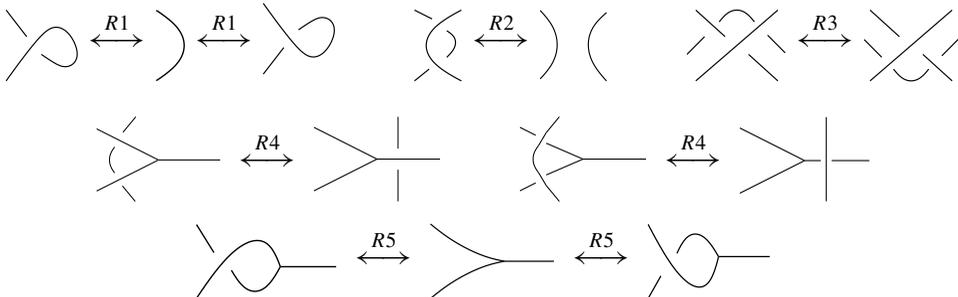


Figure 1. Extended Reidemeister moves for STG diagrams.

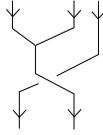


Figure 2. Example of a $(3, 2)$ -braid.

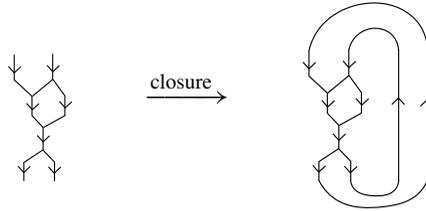


Figure 3. Example of closure operation.

STG diagrams differ by a finite sequence of the extended Reidemeister moves we refer to them as being *isotopic* (or *equivalent*).

A trivalent tangle can be regarded as a local region in a spatial trivalent graph diagram. A *trivalent braid* is a trivalent tangle in braid form. We denote by TB_n^m the set of trivalent braids with m top endpoints and n bottom endpoints, and we refer to an element in TB_n^m as an (m, n) -trivalent braid. See, for instance, the $(3, 2)$ -braid in Figure 2.

The *closure* \bar{b} of an (n, n) -trivalent braid b is the STG diagram obtained by connecting the n top endpoints of b with the corresponding bottom endpoints, drawing n nonintersecting arcs (see Figure 3).

If $b_1 \in TB_n^m$ and $b_2 \in TB_s^n$, then we can compose b_1 with b_2 . The *composition* $b_1 b_2$ is the trivalent braid obtained by placing b_1 on top of b_2 and connecting the bottom endpoints of b_1 with the top endpoints of b_2 . Note that $b_1 b_2 \in TB_s^m$ (see Figure 4).

The *identity* (n, n) -braid, denoted by 1_n , is the braid with n parallel strands free of crossings or vertices. In addition, the braids σ_i, σ_i^{-1} and y_i, λ_i depicted in Figure 5 are called *elementary trivalent braids*; the only restriction for the index i

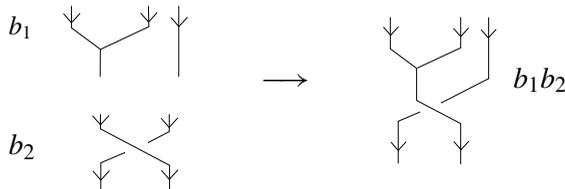


Figure 4. Composition of trivalent braids.

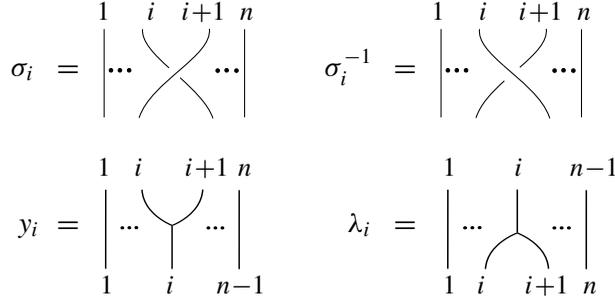
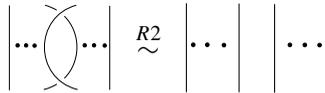


Figure 5. Elementary trivalent braids.

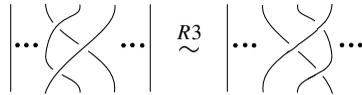
is that it must be less than the number of braid strands at the horizontal level where such elementary trivalent braid is located (note that a trivalent braid in TB_n^m may contain an arbitrary number of strands at some horizontal level). With these at hand, any trivalent braid can be regarded as a composition of elementary trivalent braids, and the corresponding representation of the braid is called a “word”.

Similar to the case of classical braids, trivalent braids are considered up to isotopy. Two (m, n) -trivalent braids are called *isotopic* if they are related by a finite sequence of the following replacements for subwords (note that for each such move below there is also the variant which uses σ_i^{-1}):

- $\sigma_i \sigma_i^{-1} = \sigma_i^{-1} \sigma_i = 1_n$:



- $\sigma_i \sigma_{i+1} \sigma_i = \sigma_{i+1} \sigma_i \sigma_{i+1}$:



- $y_i = \sigma_i y_i,$

$$\lambda_i = \lambda_i \sigma_i:$$



- $\sigma_i \sigma_{i+1} y_i = y_{i+1} \sigma_i,$

$$\sigma_{i+1} \sigma_i y_{i+1} = y_i \sigma_i:$$





Figure 6. Allowed orientations near a vertex.

- $\lambda_i \sigma_{i+1} \sigma_i = \sigma_i \lambda_{i+1}$, $\lambda_{i+1} \sigma_i \sigma_{i+1} = \sigma_i \lambda_i$:



- Commuting relations:

$$\sigma_i b_j = b_j \sigma_i, \quad y_i b_{j-1} = b_j y_i, \quad \lambda_i b_j = b_{j-1} \lambda_i, \quad \text{where } i + 1 < j,$$

$$\text{and } b_i \in \{\sigma_i^{\pm 1}, y_i, \lambda_i\}.$$

In this paper, we work with oriented STG diagrams whose vertices are either zip or unzip vertices (see Figure 6). A *zip vertex* is a trivalent vertex with two of its edges oriented toward it and one edge oriented away from it. On the other hand, an *unzip vertex* has one edge oriented toward it and two edges oriented away from it. We do not allow *sink* or *source* vertices, where all edges are oriented toward or, respectively, away from it.

We say that a spatial trivalent graph is *well-oriented* if it contains only zip and unzip vertices. We remark that any STG can be well-oriented; a proof of this can be found, for example, in [Lebed 2015].

We use the convention that trivalent braids have downward orientation. By the handshaking lemma, every (unoriented) trivalent graph has an even number of vertices (all of degree 3). Similarly, an (n, n) -trivalent braid has an even number of vertices. It is easy to show that for a well-oriented STG and an (n, n) -trivalent braid, each has half of its vertices zipped and half unzipped.

3. Alexander-type theorem for trivalent braids

In this section we shall generalize the Alexander theorem [1923] for classical knots and links to spatial trivalent graphs. We remark that the Alexander theorem for oriented spatial graphs was first proved by K. Kanno and K. Taniyama [2010]; we shall give our own proof here, so that the discussion in the next section on the Markov theorem is simple.

We will first show how to braid an STG diagram. We present a braiding algorithm analogous to the one in [Lambropoulou 1993; Lambropoulou and Rourke 1997]

and adapt it to the setting of spatial trivalent graphs. For the purposes of this paper, we care about how the isotopy moves on diagrams affect the final braids. The conventions introduced in the following braiding process (that is, preparation for braiding and the braiding algorithm) were carefully chosen so as to simplify the examination of the resulting braids.

3.1. Preparation for braiding. We work strictly with well-oriented spatial trivalent graphs, so from now on we assume all STG diagrams are well-oriented. In addition, STG diagrams are assumed to be piecewise linear. This allows us to subdivide an arc into two smaller arcs by marking it with a point. From now on, when we refer to vertices, we strictly mean trivalent vertices, thus, distinguishing vertices from subdivision points. Also, we consider local maxima and minima in arcs to be subdivision points.

Before we begin braiding an arbitrary STG diagram, we need to establish certain requirements for the diagrams. These requirements describe a sort of “general picture” of how an STG diagram must look in order to proceed with the braiding process. A large portion of the preparation for braiding is devoted to explaining such requirements and how to isotope an arbitrary STG diagram so that it satisfies these requirements. As we shall see below, the isotopy needed in order to meet such requirements is local, and, in particular, can be reduced to small changes involving planar isotopy and the $R5$ move.

Now STG diagrams lie in the plane, which is equipped with the top-to-bottom direction. This allows us to impose particular restrictions on the diagrams. For example, we require STG diagrams to contain strictly up-arcs and down-arcs (no horizontal arcs). Furthermore, there should not be pairs of horizontally aligned crossings or vertices, so as to have the vertices and crossings in the corresponding braid lying on different horizontal levels. In addition, vertically aligned vertices, crossings, or subdivision points are not allowed, so as to avoid triple points when creating new pairs of braid strands with the same endpoint. Later, when we show how the braiding process is performed, the justification for rejecting such vertical alignment shall be made clear.

The goal of the braiding process is to preserve the down-arcs in a diagram and replace the up-arcs with pairs of braid strands oriented downwards. An arbitrary up-arc may cross with several other arcs. We subdivide each up-arc into smaller pieces, such that each up-arc between two subdivision points contains at most one crossing. We label all up-arcs with an “ o ” or “ u ” indicating whether it is the over- or under-strand of a crossing in the diagram. Note that for *free up-arcs* (up-arcs that do not contain crossings), we have a choice whether to label them with an “ o ” or “ u ”.

In order to simplify the braiding for vertices, we impose the condition that a subdivision point cannot coincide with a vertex. We want all arcs incident with a



Figure 7. Types of trivalent vertices.

vertex to be oriented downwards. By doing this, we isolate the vertices from the braiding of the up-arcs. We say that a vertex is in *regular position* if, in a small neighborhood, it is incident only with down-arcs.

Now we will introduce conventions for bringing vertices into regular position. Given an STG diagram, every vertex is, roughly speaking, in either one of the four positions illustrated in Figure 7 (there cannot be horizontal arcs). We remark that a vertex in any one of these positions must be oriented so as to satisfy the conventions for either zip or unzip vertices. Note that a vertex incident only with down-arcs must necessarily be either Y - or λ -type, since we do not allow sink or source vertices. We will start by treating the cases for either Y - or λ -type vertices incident with at least one up-arc.

In Figure 8, we consider the various possible orientations for a Y -type vertex and show how to put it in regular position. For a Y -type vertex incident with only one up-arc, we simply perform planar isotopy on the problematic arc (see the first row of Figure 8). Note that the corrected diagram is a λ -type vertex. When correcting a Y -type vertex incident with exactly two up-arcs, we perform an $R5$ move introducing a crossing between the two up-arcs (see the second row of Figure 8). Note that we have a choice for the type of crossing introduced by the $R5$ move, and that the corrected diagram is a Y -type vertex.

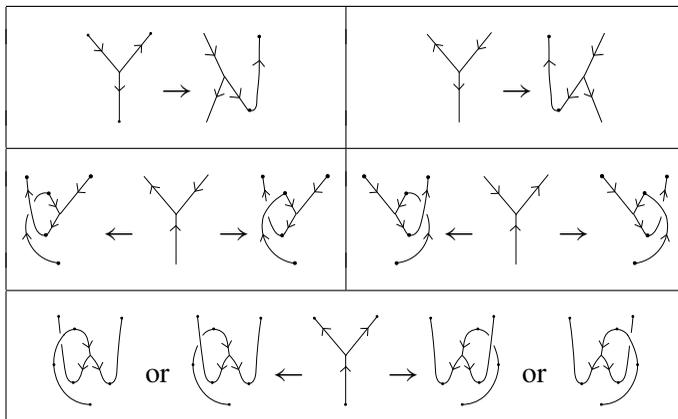


Figure 8. Adjusting Y -type vertices into regular position.

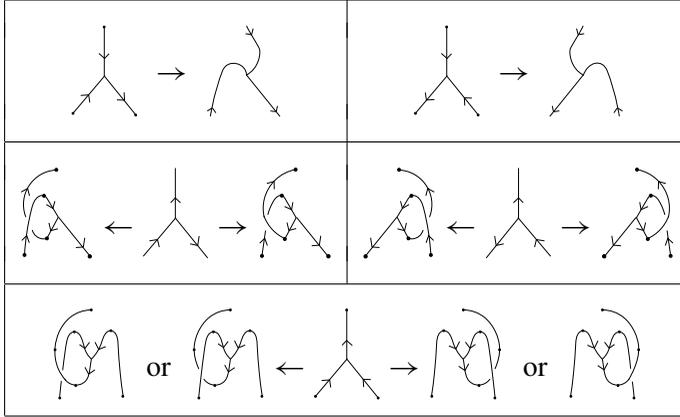


Figure 9. Adjusting λ -type vertices into regular position.

The most interesting case is a Y -type vertex incident with three up-arcs; this reduces to performing planar isotopy on an arc so that, near the vertex, it is oriented downwards, and performing an $R5$ move that introduces a crossing between the remaining two up-arcs (see the bottom row of Figure 8). It is important that the planar isotopy be performed on one of the two arcs oriented away from the vertex. Note that we have a choice concerning which arc the planar isotopy is performed on, and the type of crossing (positive or negative) introduced by the $R5$ move. Note that the resulting diagram is a λ -type vertex.

The process for bringing a λ -type vertex into regular position is similar and is depicted in Figure 9.

Now that we have taken care of Y - and λ -type vertices, only two cases remain to be considered, namely M - and W -type vertices (see the rightmost two diagrams in Figure 7). In fact, these can be reduced to the case of either a Y - or λ -type vertex. For example, given an M -type vertex, we perform planar isotopy on the rightmost arc to obtain a λ -type vertex (see the left-hand side of Figure 10). Now, using the previous conventions for λ -type vertices, we proceed to bring it into regular position. Note that we have not assumed a particular orientation for the arcs incident with the vertex, besides that the vertex is either a zip or unzip vertex. We treat W -type vertices similarly, except that in these cases the isotopy is performed on the leftmost strand, so as to obtain a Y -type vertex (as shown on the right-hand side of Figure 10). Applying these conventions to all vertices in a given STG diagram, we



Figure 10. Spreading arcs around a vertex.

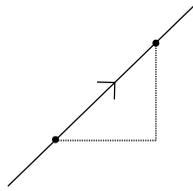


Figure 11. A sliding triangle associated with an up-arc.

obtain a new diagram which is isotopic to the original one and whose vertices are all in regular position.

We now shift our focus to up-arcs which are not incident with a vertex. We shall use the concept of *sliding triangle* introduced in [Lambropoulou and Rourke 1997]. The sliding triangle associated with an up-arc is the right triangle with hypotenuse the up-arc and with right angle lying below the hypotenuse (see Figure 11). As we shall see when we explain the actual braiding, the sliding triangle serves as a guideline for how to arrange the braided outputs of an up-arc. We say a sliding triangle is of type *over* or *under* according to the label of the up-arc it is associated with. Also, we consider sliding triangles to be adjacent whenever the corresponding up-arcs have a common subdivision point.

We introduce the *triangle condition*, see [Lambropoulou and Rourke 1997], which states that nonadjacent sliding triangles can overlap only if they have opposite labels. Later, once we introduce the braiding moves for up-arcs, we will go into detail justifying why the triangle condition is needed. In short, the triangle condition ensures that the braiding moves do not interfere with each other, so that the order in which we eliminate the up-arcs is irrelevant.

Lemma 1. *Let G be an STG diagram with vertices in regular position, and let G' be a subdivision of G . Then there is a refinement of G' such that (for appropriate choices of under/over for free up-arcs) the triangle condition is satisfied.*

Proof. Consider a crossing which contains one up-arc. We want to have the sliding triangle of such up-arc so that it overlaps only with the other strand in the crossing. If the sliding triangle of the up-arc overlaps with any other arc or sliding triangle outside the crossing, then we further subdivide the up-arc such that the resulting up-arc containing the crossing becomes small enough that its sliding triangle covers only a small neighborhood around the crossing.

For the case of a crossing containing two up-arcs, we argue as above so as to have the sliding triangles corresponding to the crossing up-arcs isolated from the rest of the diagram. Note that the corresponding sliding triangles have opposite labels (since they correspond to arcs with opposite labels), so their intersection respects the triangle condition.

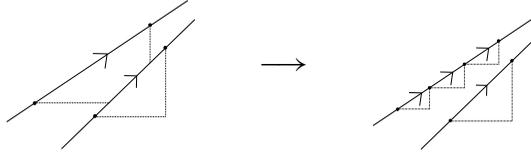


Figure 12. Adjusting sliding triangles.

Similarly, for the case of free up-arcs, we can add subdivision points, if necessary, so that the corresponding sliding triangles are disjoint from the rest of the diagram, as shown in Figure 12. Consequently, we have the liberty to assign any label to the triangles. Note that we do not encounter any problems in regions containing vertices, since the latter are in regular position. \square

From the proof of Lemma 1, we see that given a diagram with a subdivision which satisfies the triangle condition, any refinement of the subdivision (with appropriate labels) satisfies the triangle condition as well.

With this, we conclude the requirements an STG diagram must satisfy so that it is ready to be braided. We summarize the previous discussion in the following definition.

Definition 2. An STG diagram with regular vertices is said to be in *general position* if the following conditions hold:

- (1) There are no horizontal arcs.
- (2) There are no crossings, subdivision points, or vertices that are either horizontally or vertically aligned.
- (3) All nonadjacent sliding triangles must satisfy the triangle condition, and if they intersect, this must be along a common interior (and not a single point).

We refer to *direction-sensitive moves* as the local shifts on an STG diagram in order to put it in general position. These shifts can be in the horizontal or vertical directions. For example, whenever two subdivision points are either vertically or horizontally aligned, we can correct these singularities by performing planar isotopy locally on one of the subdivision points, so that they are no longer horizontally/vertically aligned. A similar argument can be used when correcting alignment of any combination of either subdivision points, crossings, or vertices. In addition, whenever two nonadjacent sliding triangles intersect at a point, we can choose a subdivision point corresponding to one of the participating sliding triangles, and replace it by another subdivision point arbitrarily close to the original one, so that condition (3) of Definition 2 is not violated. On the other hand, if two nonadjacent sliding triangles with common labels intersect, we further subdivide one of the corresponding up-arcs and change labels appropriately. (We refer the reader to the

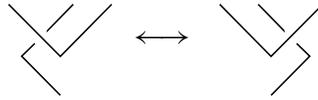


Figure 13. A swing move.

proof of Lemma 3.5 in [Lambropoulou and Rourke 1997] for details.) Note that we allow different choices to be made when shifting a diagram into general position.

Next we shall introduce the last two kinds of the direction-sensitive moves. As we shall see later, these particular moves are crucial for our one-move Markov type theorem to work; it turns out that these moves, together with planar isotopy and $R1$ – $R5$, will allow us to only consider instances of isotopy moves between diagrams in general position (see Lemma 3). The *swing moves* form a special case of the direction-sensitive moves. Figure 13 shows a version of the swing move in which an arc slides across a local minimum; in general a swing move allows an arc to swing over/under an extremum point. It is well known that the swing moves involving arcs with a local maximum can be obtained from those with a minimum together with $R2$ moves. Note that since we regard local extrema points as subdivision points, we cannot have an arc intersecting an extremum point in an STG diagram; a swing move avoids the coincidence of a minimum or maximum and a crossing.

The last of the direction-sensitive moves is the *switch move* for Y - and λ -type vertices (see Figure 14). These moves are intimately related to our requirement that vertices be in regular position. Recall that when bringing a Y - or λ -type vertex into regular position, if it is incident with at least two up-arcs, we have a choice concerning the type of crossing introduced by the $R5$ move (see the last two rows in Figures 8 and 9). Ultimately, all possible choices are related via the switch move, the swing move, and isotopy moves.

Lemma 3. *Let G and G' be isotopic STG diagrams in general position. Then there is a sequence of isotopy moves relating G and G' , all of which are between diagrams in general position.*

Proof. The idea is the following: given a finite sequence of isotopy moves relating G and G' , we can correct the middle stages in such sequence which are not in general position, and, thus obtain an alternative sequence in which all diagrams are



Figure 14. The switch moves for Y - and λ -type vertices.

in general position. It is well known that instances of isotopy in regions free of vertices can be easily achieved while still respecting the general position conditions inside such regions; for a proof of this, see Lemma 3.5 in [Lambropoulou and Rourke 1997]. Thus it remains to be seen that the same is true for regions that contain a vertex. Note first that the extended Reidemeister moves $R1$ – $R4$ applied to an STG diagram in general position yield a diagram still in general position. On the other hand, the $R5$ move and planar isotopy applied near a vertex v in regular position may convert that vertex into nonregular position. Note that since G' is in general position, any such problematic move is countered, later in the sequence, by another move that brings v back into regular position. Therefore, we can readjust the vertex v to regular position according to the rules in Figures 8 and 9, and, thus, complete the sequence with the aid of the direction-sensitive moves, in particular the swing and switch moves. Hence, we can replace the troublesome moves between STG diagrams, where the second diagram is not in regular position, by a sequence of isotopy moves between STG diagrams in general position. \square

Remark 4. From the discussion above, it follows easily that given two isotopic STG diagrams in general position, they differ by a finite sequence of direction-sensitive moves (planar isotopy, swing moves, and switch moves) and the extended Reidemeister moves, where each diagram in such a sequence is in general position as well.

From here on, we shall assume that all STG diagrams are in general position.

3.2. The braiding algorithm. We will now illustrate our braiding algorithm. The idea is to keep the down-arcs, and eliminate the up-arcs by replacing them with pairs of vertically aligned braid strands. The braiding algorithm outlined here is inspired by the one presented in [Lambropoulou and Rourke 1997]. These algorithms share many similarities, namely the way in which up-arcs are braided and the conditions that determine the general position. The main difference is that our set-up requires a careful treatment of arcs incident with a vertex. It is essential that every vertex of a given STG diagram is in regular position prior to shifting the diagram into general position. By doing this, we isolate the regions containing a vertex from the braiding of the up-arcs.

We will first show how to braid crossings. For that, consider an up-arc which is the over-strand of a crossing, and thus labeled “ o ”. The braiding consists of first sliding the up-arc across the sliding triangle, making sure the horizontal arc has a negative slope, so that it does not conflict with the general position requirement. We then cut the vertical segment, and pull the upper cut-point upward and the lower downward (see Figure 15). Note that the new vertical strands are both oriented downwards, and are vertically aligned. In addition, both vertical strands cross *over* all other arcs in the diagram. This is indicated abstractly in diagrams by adding the label “ o ” to both vertical strands (the braid box in Figure 15 indicates a magnified region in the diagram).

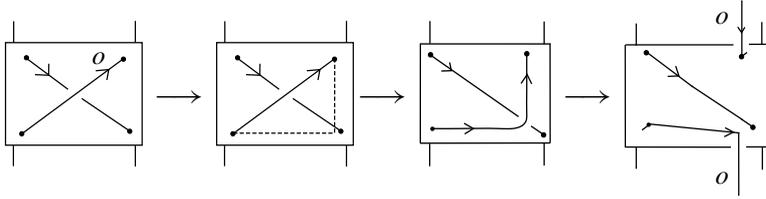


Figure 15. Braiding a crossing.

If the up-arc is the under-strand in the crossing, appropriately labeled “ u ”, then the braiding of it is identical, except that the new pair of vertical braid strands both cross *under* all other arcs in the diagram. By the same reasoning, both vertical braid strands are labeled “ u ”.

The braiding of a free up-arc is done similarly, making sure the new pair of vertical braid strands cross either under or over all other arcs in the diagram, in accordance with the label of the original up-arc (see Figure 16). In essence, by braiding a free up-arc, we simply replace the arc with a pair of vertical braid strands oriented downwards and vertically aligned with the endpoint of the free up-arc. We shall refer to the braiding of an up-arc as a *basic braiding move*.

It is important to remark that by connecting the two newly created pairs of vertical braid strands (outside of the diagram, around the braid axis), we obtain a trivalent tangle isotopic to the original one. This holds when braiding either a free up-arc or a crossing.

Remark 5. We can finally make clear the following:

- (i) It is now clear the reason for rejecting vertical alignment of any combination of vertices, subdivision points, or crossings. The problematic alignments are usually caused by subdivision points which determine the top of an up-arc, because the braiding produces pairs of vertical braid strands aligned with such points. For example, when braiding a diagram that displays vertical alignment between subdivision points, the resulting pairs of vertical braid strands correspond to the same endpoints. In addition, when braiding a diagram that displays vertical alignment between a subdivision point and a crossing or a vertex, we obtain a multiple point (which in a flat projection is a vertex of degree 6 or 5) between the crossing or vertex and one of the new vertical strands.

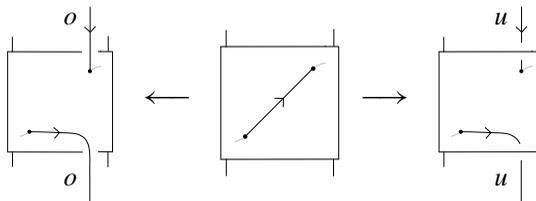


Figure 16. Braiding a free up-arc.

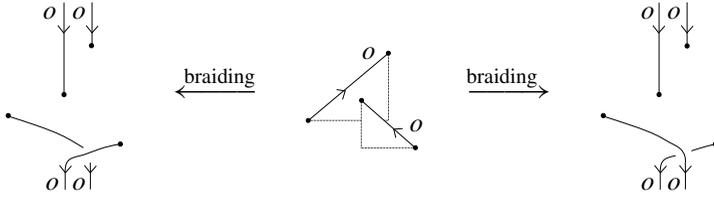


Figure 17. Forbidden sliding triangles.

(ii) We can now give a proper justification as to why we introduced the triangle condition. Let G be an STG diagram equipped with subdivision points and labels, and suppose that G contains two overlapping sliding triangles with the same label. The middle diagram of Figure 17 depicts the magnified region with such overlapping triangles. Now the order in which we braid the up-arcs of G will affect the final braid. The diagram in the left-hand side of Figure 17 is obtained by first braiding the up-arc with left-to-right orientation; conversely, the diagram on the right-hand side of the same figure is obtained by first braiding the up-arc with right-to-left orientation. Note that the two braid diagrams differ by the type of crossing introduced. We avoid this kind of behavior by introducing the triangle condition, which ensures that the braiding moves do not interfere with each other. That is, the order in which we braid the up-arcs does not affect the final braid. This will be important for the proof of the Markov-type theorem.

After braiding every up-arc in a given STG diagram in general position, we obtain a trivalent braid diagram. Recall that well-oriented STG diagrams contain an even number of vertices, half of which are zip and the other half are unzip vertices. This in fact means that by braiding an STG diagram, we indeed obtain an (n, n) -trivalent braid, where n is a positive integer. Hence the closure operation is well-defined. Ultimately, by braiding an STG diagram, we obtain a trivalent braid whose closure is isotopic to the original STG diagram. We state formally the previous discussion in the following theorem.

Theorem 6 (Alexander-type theorem for STGs). *Every well-oriented spatial trivalent graph can be represented as the closure of a trivalent braid.*

Remark 7. In our braiding algorithm, we require that each up-arc is subdivided such that it contains at most one crossing. In fact, we can relax this condition a bit by allowing for up-arcs to cross over any number of strands. The only requirement is that an up-arc contains crossings of only one type. That is, an up-arc must cross either over or under — but not both — any other strand.

Example 8. We provide in Figure 18 an example of the braiding algorithm. We start with a well-oriented spatial trivalent graph diagram representing a spatial version of the Petersen graph. The first step is to put the vertices in regular position

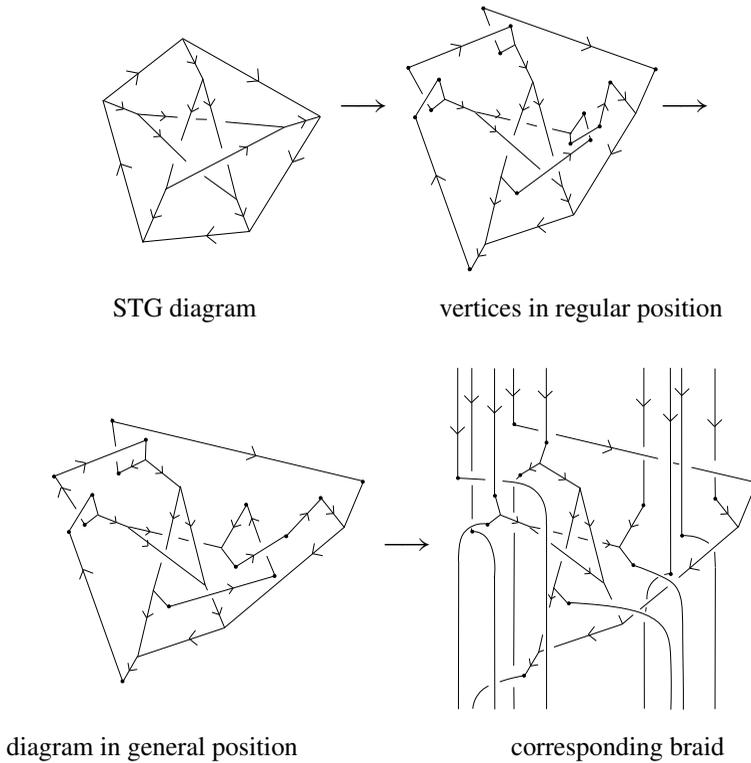


Figure 18. An example of the braiding algorithm.

according to the conventions in Figures 8, 9, and 10. Then we convert the resulting diagram into a diagram in general position (recall Definition 2). Finally, we braid the free up-arcs and all crossings containing up-arcs to arrive at a trivalent braid whose closure is isotopic to the original STG diagram.

4. Markov-type theorems for trivalent braids

Analogous to classical knot theory and in particular to [Markov 1936], we want to classify trivalent braids that, upon the closure operation, yield STG diagrams representing isotopic spatial trivalent graphs.

4.1. Trivalent L -equivalence. The goal of this section is to define an equivalence relation on the set of trivalent braids, which we refer to as *trivalent L -equivalence* (or *TL -equivalence*). This equivalence relation can be seen as an extension of the L -equivalence between classical braids. The L -equivalence is described solely by braid isotopy and the L -moves introduced by Lambropoulou in her Ph.D. thesis [1993], and used to prove the one-move Markov theorem for oriented links; see also

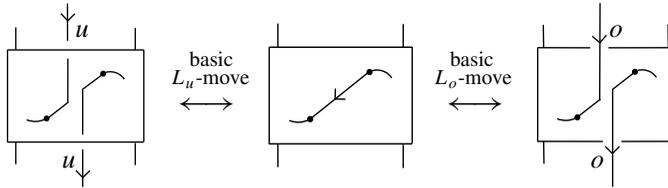


Figure 19. Basic L_u - and L_o -moves.

[Lambropoulou and Rourke 1997, Theorem 2.3]. The L -moves for classical braids extend naturally to trivalent braids, as we shall explain in this section.

Definition 9. A *basic L -move* on a trivalent braid consists of cutting an arc of the braid at a point (such point cannot be a vertex), and pulling the upper cut-point downward and the lower cut-point upward, therefore, creating a new pair of vertical braid strands, as explained in Figure 19. The new pair of braid strands are vertically aligned with the cut-point, and they either cross over or under — but not both — any other arc of the braid. Consequently, there are two types of basic L -moves, namely an *under L -move* (L_u -move) and an *over L -move* (L_o -move).

Using braid isotopy, an L -move may be formulated with a crossing (positive or negative) which can be either to the right or to the left of the cut-point; we refer to these as a *right L -move* or a *left L -move*, respectively. More precisely, we refer to this version of the L -move as either *right/left $+L$ -move* or *right/left $-L$ -move*, in accordance with the type of crossing being created (see Figure 20).

Remark 10. Some comments about the L -moves on trivalent braids are needed.

(i) The effect of the L -move is to stretch an arc around the braid axis, where the arc is being stretched either over or under the braid diagram. Therefore, such a move between trivalent braids yields isotopic closures.

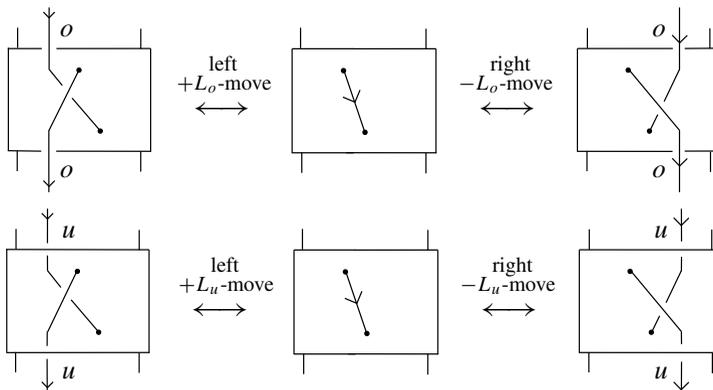


Figure 20. Left and right L -moves.

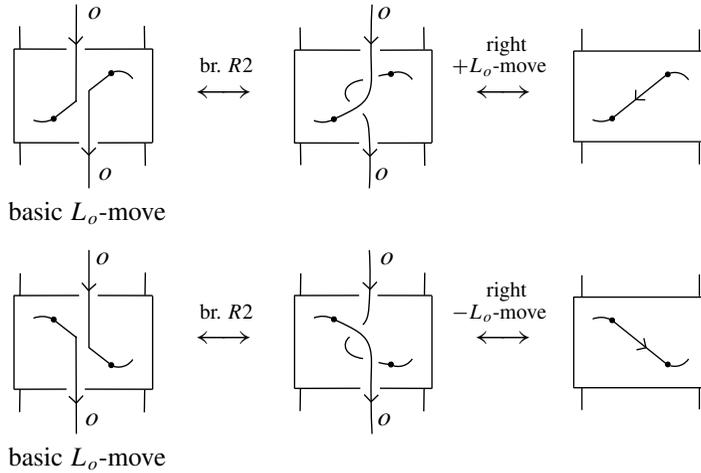


Figure 21. Basic L_o -moves via right $\pm L_o$ -moves.

(ii) Although we defined the L -moves in one direction, we allow for an L -move to be undone in a braid. That is, we allow for the contraction of a pair of vertical strands that correspond to an L -move, so as to obtain a down-arc.

Definition 11. The TL -equivalence is the equivalence relation on the set of trivalent braids determined by (1) braid isotopy and (2) right L -moves.

Note that we did not include the basic L -moves or the left L -moves in the definition of TL -equivalence. In the next lemma, we show that these L -moves follow from the right L -moves and braid isotopy; see also [Lambropoulou and Rourke 1997]. This will give us the freedom to use all versions of the L -moves when comparing TL -equivalent braids.

Lemma 12. *The basic L -moves and left L -moves follow from the right L -moves together with braid isotopy.*

Proof. Figure 21 shows how a basic L_o -move follows from the right L_o -moves together with Reidemeister moves $R2$ in braid form. Then Figure 22 shows that the left L_o -moves follow from basic L_o -moves. Therefore, left L_o -moves follow from right L_o -moves. A similar argument can be used to show that the basic and left L_u -moves follow from the right L_u -moves and braid isotopy. \square

We remark that the L -equivalence for classical braids introduced in [Lambropoulou 1993; Lambropoulou and Rourke 1997] comprises classical braid isotopy and the right L -move for classical braids. Thus, the TL -equivalence, when restricted to classical braids, is the same as the L -equivalence. In other words, the L -equivalence for classical braids extends to trivalent braids (as long as the L -moves are applied away from trivalent vertices, as we explained in the previous discussion).

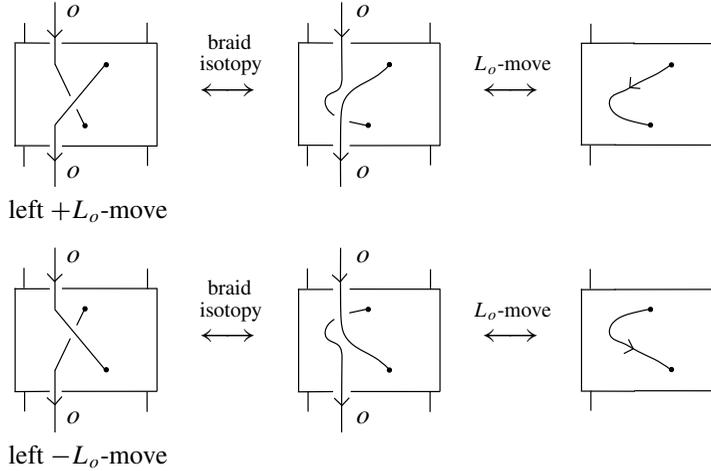


Figure 22. Left $\pm L$ -moves via basic L -moves.

Finally, we define conjugation by elementary trivalent braids σ_i and σ_i^{-1} in TB_n^n . Given a trivalent braid $b \in TB_n^n$, we say that the braids $b\sigma_i^{\pm 1} \sim \sigma_i^{\pm 1}b$, where $1 \leq i \leq n-1$, are related by *elementary conjugation* (see Figure 39). Since σ_i is invertible in TB_n^n with inverse σ_i^{-1} , the elementary conjugation in TB_n^n has the following equivalent form: $b \sim \sigma_i b \sigma_i^{-1}$ or $b \sim \sigma_i^{-1} b \sigma_i$.

The statement of Markov's theorem [1936] for classical braids makes use of conjugation by $\sigma_i^{\pm 1}$. But when employing the L -moves, the elementary conjugation can be dropped from the L -move Markov-type theorem, since elementary conjugation for classical braids follows from L -equivalence; see [Lambropoulou 1993; Lambropoulou and Rourke 1997]. The same holds for trivalent braids, as we now prove.

Lemma 13. *Elementary conjugation in a trivalent braid can be realized by a sequence of L -moves together with braid isotopy.*

Proof. The proof is illustrated in Figure 23; compare with [Häring-Oldenburg and Lambropoulou 2002]. We start with an (n, n) -trivalent braid of the form $\sigma_i^{-1}b$, where $b \in TB_n^n$ and $1 \leq i \leq n-1$, and obtain the trivalent braid $b\sigma_i^{-1}$ through a sequence of L -moves and trivalent braid isotopy. \square

We aim to show that there is a 1-1 correspondence between the isotopy types of STG diagrams and the TL -equivalence classes of trivalent braids. Now, it can be easily seen that different choices when applying the braiding algorithm affect the output of the final braid. In addition, local isotopy changes in an STG diagram may induce a different final braid, upon our braiding algorithm. However, the following theorem asserts that every instance of isotopy between STG diagrams can be translated in terms of TL -equivalence of trivalent braids.

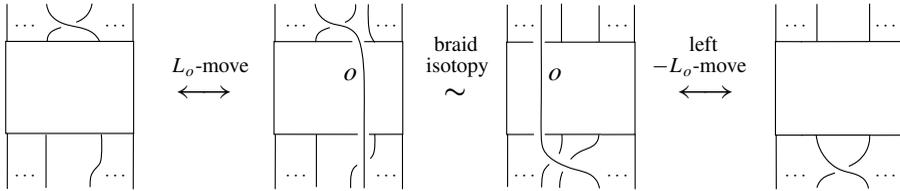


Figure 23. Elementary conjugation in terms of L -moves.

Theorem 14 (*L*-move Markov-type theorem for STGs). *Two well-oriented spatial trivalent graphs are isotopic if and only if any two of their corresponding trivalent braids are TL-equivalent.*

4.2. Proof of Theorem 14. It is clear from the definition of L -moves and braid isotopy that TL -equivalent braids have isotopic closures. Therefore, we only need to show that isotopic STG diagrams yield TL -equivalent braids upon our braiding algorithm. Throughout this proof, diagrams shall be assumed to be in general position. Also, isotopy moves will be considered strictly between diagrams in general position. In order to extend the proof for arbitrary STG diagrams (not necessarily in general position) it is essential to show that different choices when bringing a diagram into general position do not affect the final braid. Therefore, using Lemma 3 will complete the proof that any two isotopic STG diagrams yield TL -equivalent braids, upon our braiding algorithm.

The proof will be divided into two parts: The first part consists of analyzing the different choices made during the braiding process, and showing that each of these yield the same final braid, up to TL -equivalence; these choices amount to how the subdivision points are assigned, and the labels for free up-arcs. The second part will address isotopy between STG diagrams, and thus show it does not affect the final braid; for this, we analyze the direction-sensitive moves and the extended Reidemeister moves for STG diagram.

Our main approach for the proof is the following. For a given STG diagram we shall consider only the local region in which an isotopy move takes place. We assume that all other up-arcs outside such local region have been braided already. By the triangle condition, this choice does not affect the final braid, and, thus, we have the liberty to compare the braided portions corresponding to such local regions and conclude that the final braids are TL -equivalent.

For the first part of the proof, we shall assume that the diagram under consideration is equipped with a choice of subdivision points. To this end, in order to compare the effect of different choices of subdivision points for a given diagram, we need the following lemmas; see Lemmas 4.1 and 4.2 in [Lambropoulou and Rourke 1997] for detailed proofs.

Lemma 15. *If we add an extra subdivision point to an up-arc of an STG diagram, the corresponding braids differ by basic L -moves.*

Lemma 16. *When we braid a free up-arc, which we have the choice of labeling “ u ” or “ o ”, the resulting braid is independent of this choice, up to TL -equivalence.*

Remark 17. The following are consequences of the previous two lemmas:

(i) If we have a chain of overlapping sliding triangles of free up-arcs so that we have a free choice of labeling for the whole chain then, up to TL -equivalence, this choice does not affect the final braid.

(ii) If by adding a subdivision point on an up-arc we have a choice for relabeling the resulting new up-arcs so that the triangle condition is still satisfied, then the resulting braids are TL -equivalent.

Corollary 18. *Given any two subdivisions, S_1 and S_2 , of an STG diagram which will satisfy the triangle condition with appropriate labelings, the resulting braids are TL -equivalent.*

Proof. Recall that whenever a subdivision satisfies the triangle condition, any refinement satisfies it as well. Using Lemmas 15 and 16, the statement follows by considering the subdivision $S_1 \cup S_2$. \square

For the second part of the proof, we shall first examine the choices made when bringing an STG diagram into general position. These amount to different applications of the direction-sensitive moves.

Lemma 19. *Spatial trivalent graph diagrams in general position that differ by direction-sensitive moves correspond to trivalent braids that differ by L -moves.*

Proof. When putting an STG diagram in general position, the first thing we do is shift its vertices into regular position using the conventions introduced in Figures 8 and 9. In the next few paragraphs we check that different choices when applying these conventions do not affect, up to TL -equivalence, the final braid. The only cases where we have different choices are for vertices incident with at least two up-arcs.

Consider either a Y - or λ -type vertex incident with only one up-arc, as shown in the first row of Figures 8 and 9. In any case, there is only one option for shifting such vertex into regular position.

For the case of a Y -type vertex incident with exactly two up-arcs, we apply an $R5$ move between the up-arcs; this is shown in the second row of Figure 8. We have a choice for the type of crossing (positive or negative) to add when the move is applied. Note that the diagrams obtained from either choice differ by a switch move; recall Figure 14. Therefore, we need to verify that the switch move of this type does not affect the final braid, up to TL -equivalence. In Figure 24 we consider the braided portions obtained from a switch move performed on the right-hand side of a vertex.

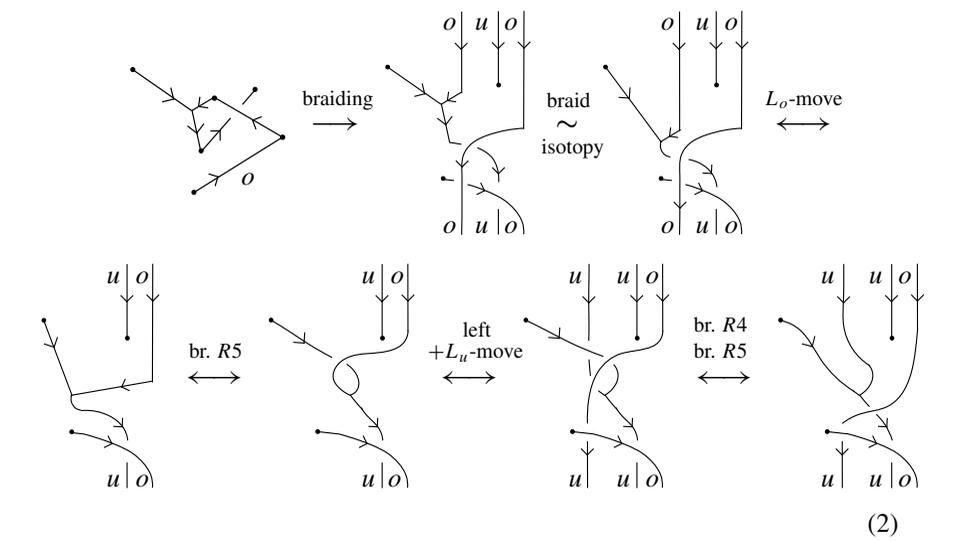
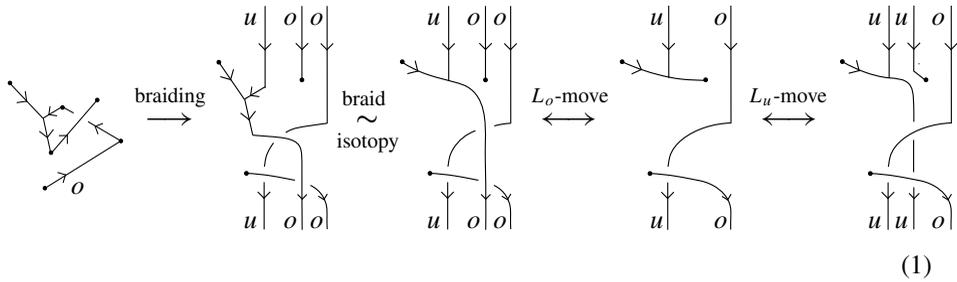


Figure 24. Checking a switch move on a Y -type vertex with two up-arcs.

It is clear that the braids (1) and (2) in Figure 24 differ by planar isotopy. The case where the two up-arcs are on the left-hand side of a Y -type vertex is treated similarly; this can be seen by reflecting the diagrams in Figure 24 across a vertical axis.

For the case of a λ -type vertex incident with two up-arcs we perform an $R5$ move between the up-arcs (as displayed in the second row of Figure 9). Similar to the case of a Y -type vertex, we have a choice for the type of crossing introduced by the $R5$ move. Once again, the braids resulting from either choice differ by a switch move. In Figure 25 we show that this version of the switch move applied on the right-hand side of a λ -type vertex does not affect the final braid up to TL -equivalence; specifically, the braids (1) and (2) in Figure 25 differ by planar isotopy. The case where the two up-arcs lie on the left-hand side of a λ -type vertex follows similarly, and thus we omit it to avoid repetition.

Now consider a Y -type vertex v incident with three up-arcs. In order to shift v into regular position we first perform planar isotopy on one arc and an $R5$ move

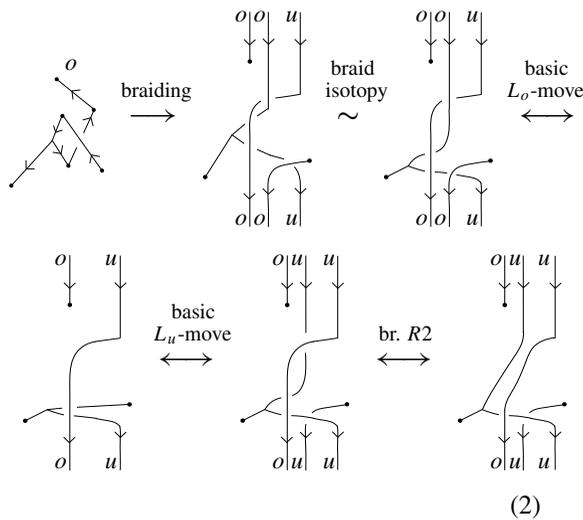
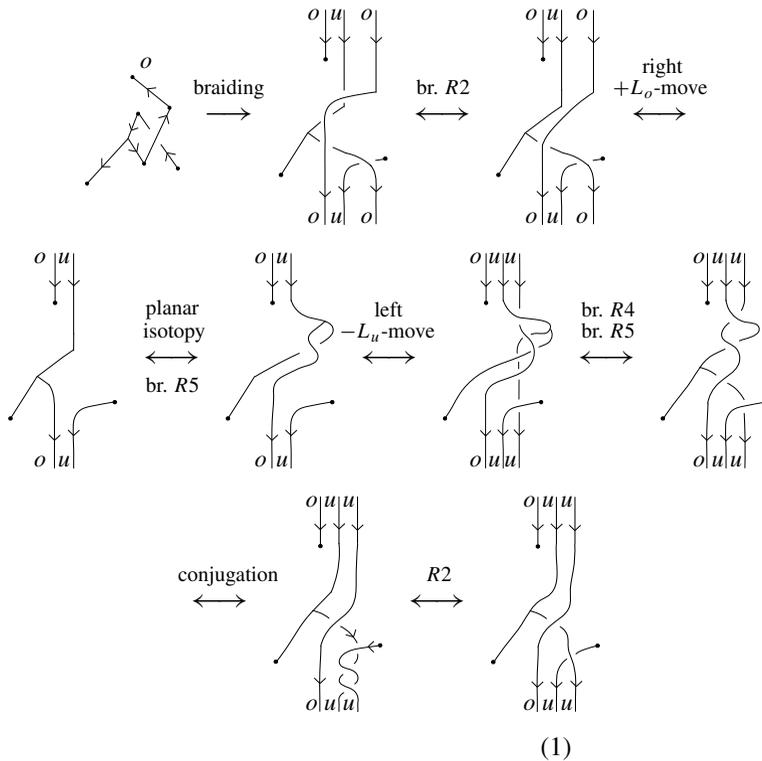


Figure 25. Checking a switch move on a λ -type vertex with two up-arcs.

between the remaining two up-arcs. This means we have four choices for shifting v into regular position (see the last row of Figure 8). We want to compare the braided portions obtained from each choice of shifting v into regular position and show

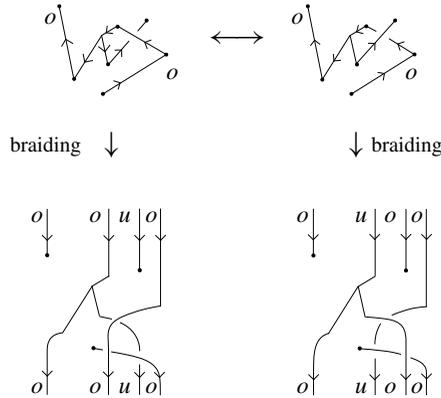


Figure 26

that they are TL -equivalent. In order to do this, we shall first compare choices that result in diagrams which differ by a switch move on the right-hand side of v (see Figure 26). We leave it as an exercise for the reader to show the braids in Figure 26 are TL -equivalent. (One can use a similar approach to that used in Figure 25.) The case for diagrams that differ by a switch move on the left-hand side of v is treated similarly. Finally, we wish to compare a choice where the crossing is on the right-hand side of v to one where the crossing is on the left-hand side of v . Figure 27 illustrates a way to relate such diagrams via $R1$, $R4$, and swing moves (this shall be enough once we check the swing moves and braid isotopy below).

For the last instance of shifting a vertex into regular position, we need to consider a λ -type vertex incident with three up-arcs. Once again, we allow for four different choices to put such vertex into regular position (see the third row of Figure 9). A similar argument to the one in the paragraph above shows that these choices yield TL -equivalent braids.

Now we check the elimination of horizontal arcs. This amounts to planar isotopy between diagrams in general position. For the case of an up-arc, planar isotopy can be treated by subdividing an up-arc; we refer the reader to [Lambropoulou and Rourke 1997] for details. The most interesting case of planar isotopy of a

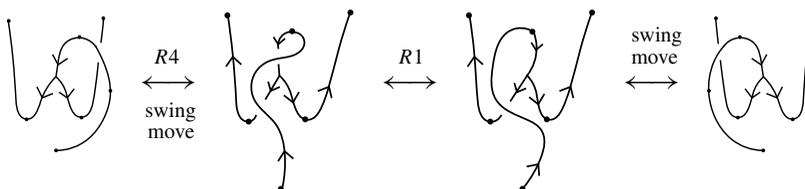


Figure 27

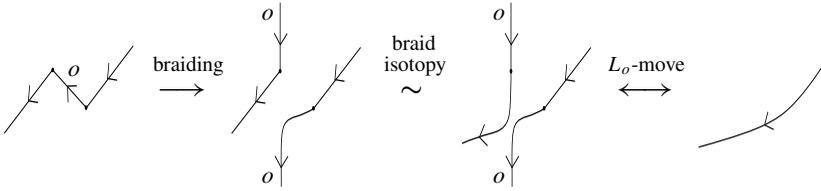


Figure 28. Planar isotopy on a down-arc.

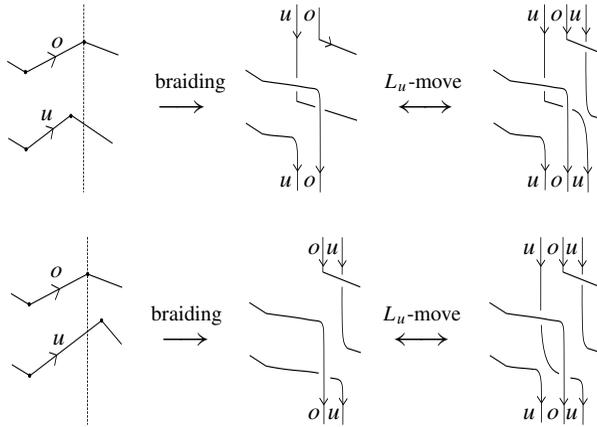


Figure 29. Planar shifts of vertically aligned subdivision points.

down-arc is verified in Figure 28. The remaining cases can be derived easily from the previous one.

Correcting horizontal alignment of either crossings, subdivision points or vertices amounts to small vertical shifts, which yield — up to braid isotopy — the same trivalent braid. In Figure 29 we illustrate the correcting shifts for vertically aligned subdivision points. Note that the final braids are the same, up to planar isotopy. The remaining instances of vertical alignment can be treated similarly.

Finally, in Figures 30 and 31 we show that the swing moves also yield *TL*-equivalent trivalent braids. This completes the proof of the lemma. \square

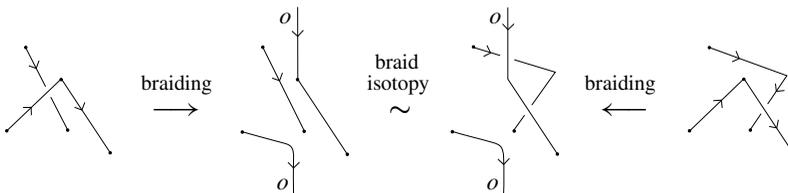


Figure 30. The first case of the swing moves.

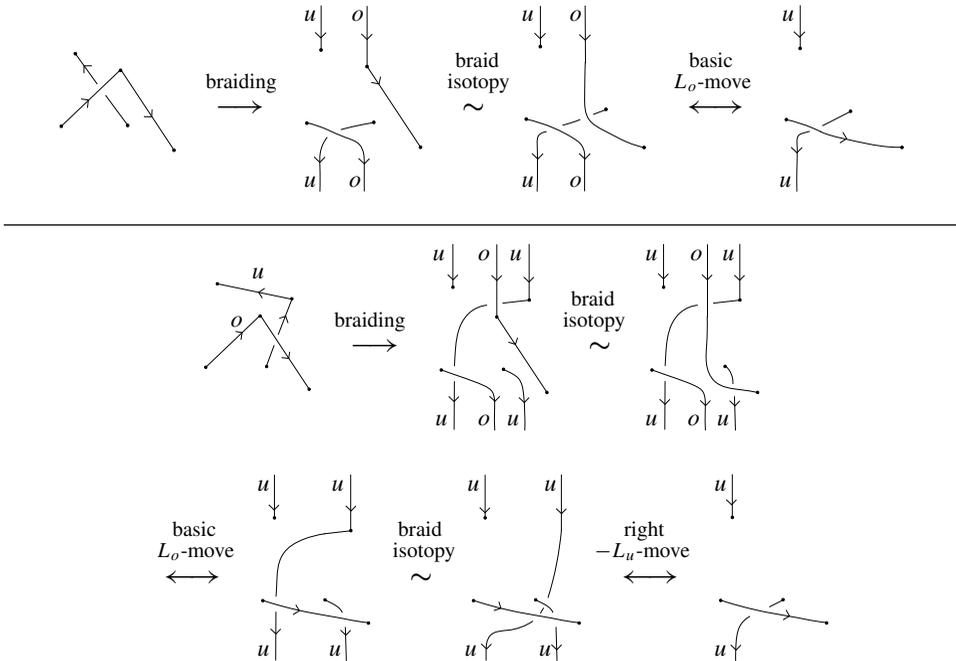


Figure 31. The second case of the swing moves.

We will now show that ambient isotopy does not affect the braiding process.

Lemma 20. *The extended Reidemeister moves yield TL -equivalent braids.*

Proof. In Figure 32, we verify one version of the $R1$ move; the braids corresponding to the two diagrams involved in the move are equivalent up to braid isotopy and L -moves.

The proof of a version of the $R2$ move with one up-arc is given in Figure 33. After braiding the up-arcs in the two diagrams involved in the move, we obtain two braids that differ by an L_u -move and braid isotopy. The more interesting case is

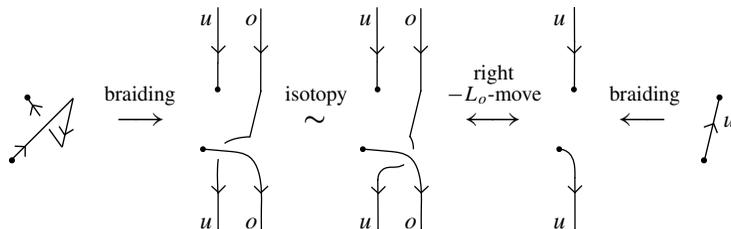


Figure 32. An $R1$ move.

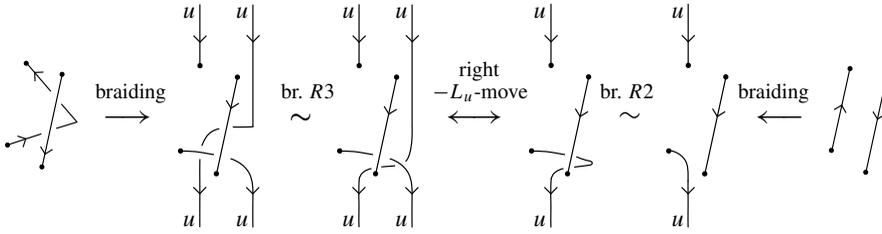


Figure 33. An $R2$ move with one up-arc.

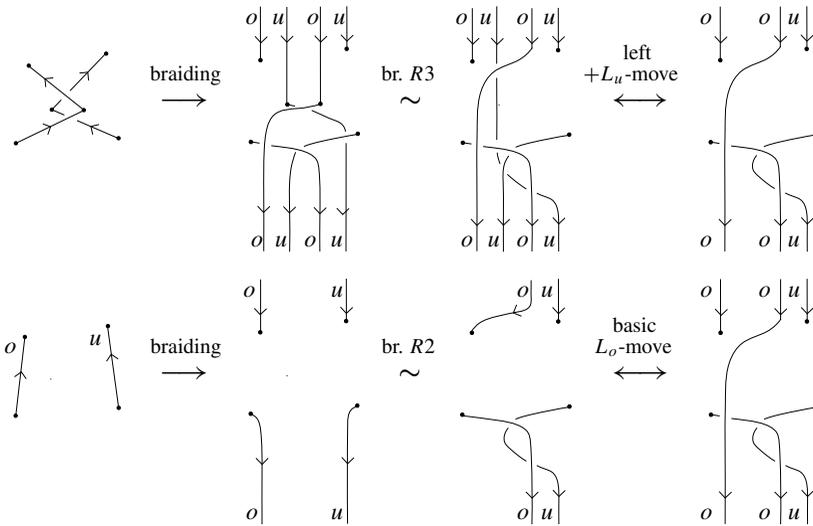


Figure 34. An $R2$ move with two up-arcs.

an $R2$ move with two up-arcs is shown in Figure 34. Again, the trivalent braids associated with the two sides of the move are TL -equivalent.

For the $R3$ move, we rely on the $R2$ moves which have been already verified. In Figure 35 we consider a version of the move with one up-arc. The other oriented versions of the $R3$ move can be verified similarly, by applying braid isotopy (namely the $R2$ move) and then a version of the $R3$ move that was already verified.

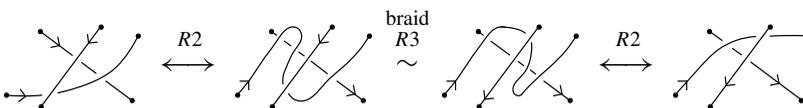


Figure 35. An $R3$ move with one up-arc.

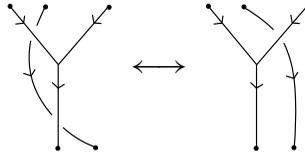


Figure 36. An $R4$ move in braid form and on a Y -type vertex.

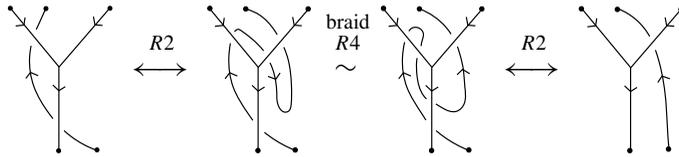


Figure 37. An $R4$ move with an up-arc and on a Y -type vertex.

In addition to the traditional Reidemeister moves, there are two extended Reidemeister moves, namely $R4$ and $R5$.

The two basic versions of the $R4$ move on a regular Y -type vertex with an arc sliding under the vertex are shown in Figures 36 and 37. Figure 36 shows the basic braid isotopy case where all strands in the diagram are oriented downward. Figure 37 shows the move where the sliding strand is an up-arc. We reduce this move to the version of the $R4$ move in braid form by employing first an $R2$ move (in a similar way as we did for the considered version of the $R3$ move). The move with an arc sliding over a Y -type vertex is treated similarly. No other orientations on Y -type vertices need to be considered, since we are assuming our diagram is in general position. The same method can be applied to the various versions of an $R4$ move on a λ -type vertex.

Since we are assuming all vertices in our diagram are in regular position, we see that the only case needed to be verified for the $R5$ move on a Y -type vertex is that shown in Figure 38. However, this move is in braid form and, therefore, it is part of trivalent braid isotopy. This argument applies to the $R5$ move on a λ -type vertex as well.

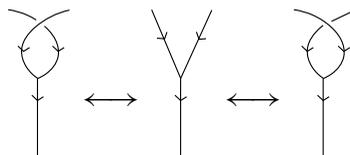


Figure 38. The $R5$ move in braid form and on a Y -type vertex.

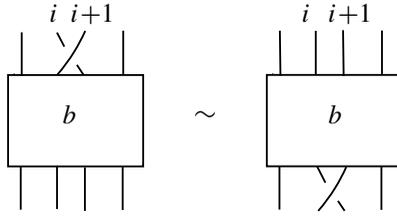


Figure 39. Conjugation by σ_i .

We have shown that our braiding algorithm applied to the two sides of any of the extended Reidemeister moves yield trivalent braids that are TL -equivalent. This concludes the proof of the L -move Markov-type theorem for trivalent braids and spatial trivalent graphs (Theorem 14). \square

4.3. Algebraic Markov theorem. In this section, we state and prove an algebraic Markov-type theorem for trivalent braids and their closures. We use the elementary trivalent braids introduced in Section 2 to define a set of algebraic moves that define an equivalence relation on trivalent braids. This algebraic equivalence relation can replace the geometric TL -equivalence used in Theorem 14.

We use b to represent an arbitrary trivalent braid in TB_n^n . We can also embed b into TB_{n+1}^{n+1} by adding an extra strand to the right of b ; we use the same symbol, b , to denote the resulting braid with an extra strand. Using this operation, we can think of TB_n^n being embedded into TB_{n+1}^{n+1} , so we define $TB := \bigcup_{n=1}^{\infty} TB_n^n$.

Theorem 21 (algebraic Markov-type theorem for STGs). *Two well-oriented spatial trivalent graphs are isotopic if and only if any two corresponding trivalent braids differ by a finite sequence of braid relations in TB and the following moves:*

(i) *Elementary conjugation (conjugation by σ_i and σ_i^{-1} ; see Figure 39):*

$$\sigma_i b \sim b \sigma_i \quad \text{and} \quad \sigma_i^{-1} b \sim b \sigma_i^{-1}, \quad \text{where } b, \sigma_i^{\pm 1} \in TB_n^n, \quad 1 \leq i \leq n-1.$$

(ii) *Right stabilization (see Figure 40):*

$$bc \sim b\sigma_n^{\pm 1}c, \quad \text{where } b, c \in TB_n^n \text{ and } b\sigma_n^{\pm 1}c \in TB_{n+1}^{n+1}.$$

Proof. We note first that braid isotopy is part of both TL -equivalence and the algebraic equivalence of Theorem 21.

It is easy to see that the closures of two trivalent braids that are related by trivalent braid isotopy and a finite sequence of right stabilization and elementary conjugation are isotopic STG diagrams.

For the converse, let b_1 and b_2 be trivalent braids that yield isotopic STG diagrams upon the closure operation. By Theorem 14, we know that b_1 and b_2 are TL -equivalent. Therefore, it suffices to show that the right L -moves for trivalent braids

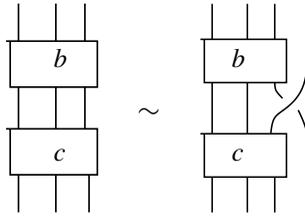


Figure 40. Right stabilization by σ_n .

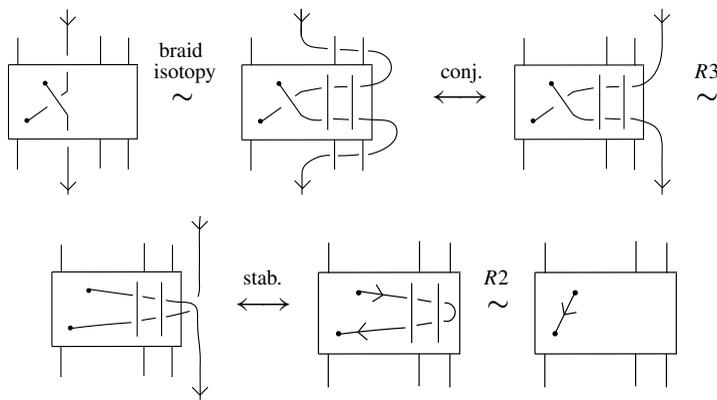


Figure 41. Right L_u -move follows from the algebraic moves.

follow from the algebraic moves of Theorem 21. In Figure 41, we show that the right L_u -move can be obtained from right stabilization, elementary conjugation, and trivalent braid isotopy.

The proof for the right L_o -move follows similarly. This completes the proof of Theorem 21. \square

Final comments. We provided two Markov-type theorems for trivalent braids and spatial trivalent graphs. As in the case of classical braids, our algebraic Markov-type theorem requires two moves, besides braid isotopy: conjugation by the elementary braids σ_i or σ_i^{-1} and right stabilization. In our case, right stabilization is not done in the bottom of a braid but between two trivalent braids. For classical braids, due to conjugation, bottom right stabilization is equivalent to right stabilization between braids. This is not the case for our approach for trivalent braids, since we do not have conjugation by the elementary trivalent braids y_i and λ_i .

We defined the TL -equivalence for trivalent braids as an extension of the L -equivalence for classical braids. TL -equivalence encompasses trivalent braid isotopy and right L -moves. We used the TL -equivalence to prove a one-move Markov-type

theorem for trivalent braids. Then, we used this L -move Markov-type theorem to prove a more algebraic Markov-type theorem for trivalent braids.

Acknowledgements

We gratefully acknowledge support from the NSF grant DMS-1460151 through the Research Experience for Undergraduates (REU) Program at California State University, Fresno. Caprau was also partially supported by Simons Foundation collaboration grant #355640.

References

- [Alexander 1923] J. W. Alexander, “A lemma on systems of knotted curves”, *Proc. Natl. Acad. Sci. USA* **9**:3 (1923), 93–95.
- [Caprau et al. 2016] C. Caprau, A. de la Pena, and S. McGahan, “Virtual singular braids and links”, *Manuscripta Math.* **151**:1-2 (2016), 147–175. MR Zbl
- [Caprau et al. 2019] C. Caprau, A. Dirdak, R. Post, and E. Sawyer, “Alexander- and Markov-type theorems for virtual trivalent braids”, *J. Knot Theory Ramifications* **28**:1 (2019), art. id. 1950003. MR Zbl
- [Häring-Oldenburg and Lambropoulou 2002] R. Häring-Oldenburg and S. Lambropoulou, “Knot theory in handlebodies”, *J. Knot Theory Ramifications* **11**:6 (2002), 921–943. MR Zbl
- [Kanno and Taniyama 2010] K. Kanno and K. Taniyama, “Braid presentation of spatial graphs”, *Tokyo J. Math.* **33**:2 (2010), 509–522. MR Zbl
- [Kauffman 1989] L. H. Kauffman, “Invariants of graphs in three-space”, *Trans. Amer. Math. Soc.* **311**:2 (1989), 697–710. MR Zbl
- [Kauffman and Lambropoulou 2006] L. H. Kauffman and S. Lambropoulou, “Virtual braids and the L -move”, *J. Knot Theory Ramifications* **15**:6 (2006), 773–811. MR Zbl
- [Lambropoulou 1993] S. Lambropoulou, *A study of braids in 3-manifolds*, Ph.D. thesis, University of Warwick, 1993, available at <http://wrap.warwick.ac.uk/73390/>.
- [Lambropoulou and Rourke 1997] S. Lambropoulou and C. P. Rourke, “Markov’s theorem in 3-manifolds”, *Topology Appl.* **78**:1-2 (1997), 95–122. MR Zbl
- [Lebed 2015] V. Lebed, “Qualgebras and knotted 3-valent graphs”, *Fund. Math.* **230**:2 (2015), 167–204. MR Zbl
- [Markov 1936] A. Markov, “Über die freie Äquivalenz der geschlossenen Zöpfe”, *Rec. Math. Moscou [Mat. Sbornik] (N.S.)* **1**:1 (1936), 73–78. Zbl

Received: 2018-07-20

Accepted: 2019-12-28

ccaprau@csufresno.edu

Department of Mathematics, California State University,
Fresno, CA 93740-8001, United States

gabriel.coloma@upr.edu

Department of Mathematical Sciences,
University of Puerto Rico, Mayagüez, Puerto Rico

mdavis7@ithaca.edu

Department of Mathematics, Ithaca College, Ithaca, NY,
United States

Low stages of the Taylor tower for r -immersions

Bridget Schreiner, Franjo Šarčević and Ismar Volić

(Communicated by Józef H. Przytycki)

We study the beginning of the Taylor tower, supplied by manifold calculus of functors, for the space of r -immersions, which are immersions without r -fold self-intersections. We describe the first r layers of the tower and discuss the connectivities of the associated maps. We also prove several results about r -immersions that are of independent interest.

1. Introduction

Let M and N be smooth manifolds. An r -immersion of M in N is an immersion that has no r -fold self-intersections; i.e., no r points of M are mapped to the same point in N . The purpose of this paper is to initiate the study of $\mathrm{rImm}(M, N)$, the space of r -immersions of M in N (see Definition 2.1), using manifold calculus of functors.

This theory, due to [Goodwillie and Weiss 1999; Weiss 1999] (see also [Munson 2010; Munson and Volić 2015, Section 10.2] for overviews), studies contravariant functors $F : \mathcal{O}(M) \rightarrow \mathcal{C}$, where $\mathcal{O}(M)$ is the poset of open subsets of M , and \mathcal{C} is usually Top , the category of topological spaces (but it can be the category of spectra, chain complexes, etc.). The theory produces a *Taylor tower* of functors and natural transformations

$$F(-) \longrightarrow (T_\infty F(-) \rightarrow \cdots \rightarrow T_k F(-) \rightarrow T_{k-1} F(-) \rightarrow \cdots \rightarrow T_0 F(-))$$

whose stages $T_k F(-)$ approximate F in a suitable “polynomial” sense (see Section 5). The hope is that the Taylor tower converges to F , namely that there is an equivalence between F and $T_\infty F$, the inverse limit of the tower.

The main functor to which manifold calculus has been applied with great success is the embedding functor (in fact, this functor is the motivation for the development of the theory). Namely, the space of embeddings $\mathrm{Emb}(M, N)$ can be regarded as a contravariant functor on $\mathcal{O}(M)$ since an inclusion of open subsets of M gives a

MSC2010: primary 57R42; secondary 55R80, 57R40.

Keywords: calculus of functors, manifold calculus, Taylor tower, r -immersions, embeddings, immersions, configuration space, subspace arrangement.

restriction map of embedding spaces. An important and deep result [Goodwillie and Klein 2015; Goodwillie and Weiss 1999] states that

the map $\text{Emb}(M, N) \rightarrow T_k \text{Emb}(M, N)$ is $(k(n-m-2)-m+1)$ -connected, (1)

where m and n are the dimensions of M and N , respectively. (This is restated later as Theorem 5.11 and more details are given there.) Thus if $n - m - 2 > 0$, the Taylor tower converges to $\text{Emb}(M, N)$.

This convergence result has proven to be remarkably fruitful in the study of knot and link spaces [Arone et al. 2008; Dwyer and Hess 2012; Lambrechts et al. 2010; Munson and Volić 2014; Sinha 2006; Songhafouo Tsopméné 2016; Volić 2006], embeddings of long planes [Arone and Turchin 2014; Boavida de Brito and Weiss 2018], more general embedding spaces [Arone et al. 2007; Munson 2005], link maps and their Milnor invariants [Goodwillie and Munson 2010; Munson 2008; 2011], etc. A strong connection to operads has also been established and this point of view pervades much of the current work in manifold calculus. Because of its success with embeddings, the foundation of the theory has also been expanded in various directions [Boavida de Brito and Weiss 2013; Munson and Volić 2012; Tillmann 2019; Turchin 2013].

When $r = 2$, the space of r -immersions is precisely the space of embeddings since the condition is that such an immersion may not have double points, and this defines an embedding (for M compact). Thus $\text{Emb}(M, N)$ is just the beginning of the filtration

$$\text{Emb}(M, N) = 2\text{Imm}(M, N) \subset 3\text{Imm}(M, N) \subset \cdots \subset r\text{Imm}(M, N) \subset \cdots \subset \text{Imm}(M, N).$$

The last space is that of immersions of M in N . All the spaces above are contravariant functors on $\mathcal{O}(M)$ and it is thus natural to try to extend the study of embeddings to r -immersions using manifold calculus. The beginning of such a study, from the operad point of view, already appears in [Dobranskaya and Turchin 2015, Section 11], although that work does not address the convergence question, which is our main concern here.

Another motivation for developing manifold calculus for r -immersions is recent developments in combinatorial topology. One of the goals in this field is to understand and control the self-intersections of maps $K \rightarrow \mathbb{R}^n$ where K is an m -dimensional complex. One of the best-known of this family of questions is the *Tverberg conjecture*: for $r \geq 2$ and $n \geq 1$, any map $f : \Delta^{(r-1)(n+1)} \rightarrow \mathbb{R}^n$ maps points from r disjoint faces to the same point. This conjecture has been disproved by Frick [2015], who uses work of Mabillard and Wagner [2014; 2016] on the generalized Whitney trick that gives a way of resolving self-intersection points and formulates obstructions for doing this.

On the other hand, the Taylor tower for embeddings has been shown to capture and classify obstructions for resolving self-intersections [Goodwillie et al. 2001; Munson 2005] and for turning an immersion into an embedding. The corresponding question of turning an immersion into an r -immersion would then have a home in the Taylor tower for r -immersions, and consequently the Tverberg conjecture might naturally live there as well. Furthermore, the machinery of manifold calculus might provide new context for many related “Tverberg–van Kampen–Flores-type” questions.

However, the tower studied here is still one step removed from Tverberg-type problems due to the condition in such problems that the self-intersections should come from “far away”. Therefore, what is needed after the Taylor tower for r -immersions is understood is a development of the tower for the subspace of r -immersions of a manifold, now with a triangulation, given by those self-intersections that come from disjoint simplices (these are sometimes in the literature called *almost r -immersions*). A recent extension of manifold calculus to simplicial complexes [Tillmann 2019] should be relevant here.

We also expect that the Taylor tower for r -immersions will connect to work of Salikhov [2002] on obstructions to the existence of a homotopy from an immersion to an r -immersion.

1.1. Main results and organization. The results of this paper are summarized in the following diagram:

$$\begin{array}{c}
 T_\infty \mathbf{rImm}(M, N) \\
 \downarrow \vdots \\
 \mathbf{rImm}(M, N) \xrightarrow{\quad} T_r \mathbf{rImm}(M, N) \xleftarrow{\quad} L_r \mathbf{rImm}(M, N); (r-1)n-rm-2 \\
 \downarrow (r-1)n-rm-1 \quad \downarrow (r-1)n-rm-1 \\
 T_{r-1} \mathbf{rImm}(M, N) \xleftarrow{\quad} L_{r-1} \mathbf{rImm}(M, N) \sim * \\
 \downarrow \sim \quad \downarrow \sim \\
 \vdots \\
 \downarrow \sim \\
 T_2 \mathbf{rImm}(M, N) \xleftarrow{\quad} L_2 \mathbf{rImm}(M, N) \sim * \\
 \downarrow \sim \\
 T_1 \mathbf{rImm}(M, N) \simeq \mathbf{Imm}(M, N) \\
 \downarrow \\
 *
 \end{array}
 \tag{2}$$

The spaces $L_k \mathbf{rImm}(M, N)$ are the (homotopy) fibers of the maps

$$T_k \mathbf{rImm}(M, N) \rightarrow T_{k-1} \mathbf{rImm}(M, N).$$

The numbers in the diagram are connectivities of maps and, in the case of $L_r \text{rImm}(M, N)$, the connectivity of that space. The first stage of the Taylor tower is a one-point space by definition (see Definition 5.2). The equivalence between $T_1 \text{rImm}(M, N)$ and $\text{Imm}(M, N)$ is the content of Proposition 6.1. Proposition 6.2 then shows that $L_k \text{rImm}(M, N)$ is contractible for $2 \leq k \leq r - 1$, from which it follows (Theorem 6.3) that there are equivalences

$$T_k \text{rImm}(M, N) \rightarrow T_{k-1} \text{rImm}(M, N) \quad \text{for } 2 \leq k \leq r - 1.$$

We also show in Corollary 2.3 that the map

$$\text{rImm}(M, N) \rightarrow \text{Imm}(M, N)$$

is $((r-1)n-rm-1)$ -connected, and so all the connectivities of the maps

$$\text{rImm}(M, N) \rightarrow T_k \text{rImm}(M, N), \quad 2 \leq k \leq r - 1,$$

follow. Lastly, in Proposition 6.6 we exhibit the connectivity of $L_r \text{rImm}(M, N)$.

There are two main lines of arguments we employ in our proofs. One is general position and transversality, which are well-known topics and we use them in a basic way. The other ingredient is homotopy limits of cubical diagrams; these techniques are central in functor calculus, but are less known so we review them in Section 4.

We also provide a review of manifold calculus and the way it applies to the embedding functor in Section 5. This section is of independent interest since it provides a short path through the theory while supplying many references for further reading.

Also of independent interest are several results that have to do with spaces of r -immersions and r -configuration spaces, such as Theorem 2.2, Proposition 3.3, and Lemma 6.5. We hope that some readers will find these useful regardless of their interest in calculus of functors.

Mentioned above are r -configuration spaces, which are central to our story. These are configuration spaces where up to $r - 1$ points are allowed to be the same. We will say more about them in Section 3 but it is worth noting that these are generally difficult and worthy of investigation in their own right. In fact, the reason why our results stop at the r -th stage of the Taylor tower is that this is the range in which r -configuration spaces are very simple.

Some ideas and future directions of investigation of the higher stages of the Taylor tower for r -immersions are given in Section 7.

2. Spaces of embeddings, immersions, and r -immersions

Suppose M and N are smooth manifolds of dimensions m and n , respectively. We will assume throughout that $m \leq n$. Let $\mathcal{C}^\infty(M, N)$ be the space of smooth maps

from M to N , topologized using the Whitney C^∞ topology. The spaces in the following definition are all topologized as subspaces of $C^\infty(M, N)$.

Definition 2.1. • An *embedding* of M in N is a smooth map $f : M \rightarrow N$ satisfying

- (i) f is a homeomorphism onto its image, and
- (ii) the derivative of f is injective; i.e., the map of tangent spaces $D_x f : T_x M \rightarrow T_{f(x)} N$ is an injection for all $x \in M$.

The space of embeddings of M in N is denoted by $\text{Emb}(M, N)$. A path in the space of embeddings is called an *isotopy*.

- An *immersion* of M in N is a smooth map $M \rightarrow N$ satisfying (ii) above. The space of immersions of M in N is denoted by $\text{Imm}(M, N)$. A path in the space of immersions is called a *regular homotopy*.

- An *r -immersion* of M in N is an immersion of M in N that does not have r -fold intersections; i.e., it satisfies the property that, for any subset of r distinct points $R = \{x_1, \dots, x_r\}$ of M , it is not a constant map when restricted to R . The space of r -immersions of M in N is denoted by $\text{rImm}(M, N)$.

For a compact manifold M , an embedding is an injective immersion. It is then immediate from the above definition that, if M is compact,

$$2 \text{Imm}(M, N) = \text{Emb}(M, N),$$

because the condition that an immersion have no double points is the same as requiring it to be injective. Much of the time, M will for us indeed be a compact manifold and we will indicate when this assumption is being made.

Here is a result that is of independent interest. Its consequence, Corollary 2.3, will turn out to say something about the beginning of the Taylor tower for r -immersions (Corollary 6.4). The reader unfamiliar with the definition of the connectivity of a map should look ahead at Definition 4.1.

Theorem 2.2. *Let M and N be smooth manifolds of dimensions m and n , respectively. Then the inclusion*

$$\text{rImm}(M, N) \rightarrow (r + 1)\text{Imm}(M, N)$$

is $((r-1)n - rm - 1)$ -connected.

Proof. The proof is a standard general position argument. Namely, consider a map

$$\begin{aligned} h : S^k &\rightarrow (r + 1)\text{Imm}(M, N), \\ v &\mapsto f. \end{aligned}$$

We wish to find for which k there exists a lift

$$\begin{array}{ccc}
 & \text{rImm}(M, N) & \\
 & \nearrow & \downarrow \\
 S^k & \xrightarrow{h} & (r+1)\text{Imm}(M, N)
 \end{array} \tag{3}$$

Consider the adjoint H of h , i.e.,

$$\begin{aligned}
 H : M \times S^k &\rightarrow N, \\
 (x, v) &\mapsto f(x),
 \end{aligned}$$

and the map

$$\begin{aligned}
 \tilde{H} : \text{Conf}(r, M) \times S^k &\rightarrow N^r, \\
 (x_1, x_2, \dots, x_r, v) &\mapsto (f(x_1), f(x_2), \dots, f(x_r)).
 \end{aligned} \tag{4}$$

Here $\text{Conf}(r, M)$ is the configuration space of r distinct point in M (see (5) for the precise definition).

Now we want to use the multijet transversality theorem; see, for example, [Munson and Volić 2015, A.2.23]. We are looking at zeroth-order jets, in which case the theorem says the following: Suppose that Z is a submanifold of $\text{Conf}(r, M) \times N^r$. Then any $f : M \rightarrow N$ can be approximated by a map (i.e., it is homotopic to a map by a small homotopy) such that the associated map

$$\begin{aligned}
 \text{Conf}(r, M) &\rightarrow \text{Conf}(r, M) \times N^r, \\
 (x_1, \dots, x_r) &\mapsto (x_1, \dots, x_r, f(x_1), \dots, f(x_r)),
 \end{aligned}$$

is transverse to Z . We are interested in this in the case when Z is the product of $\text{Conf}(r, M)$ and the (thin) diagonal Δ in N^r , i.e., the set of r -tuples $(x_1, \dots, x_r) \in N^r$ such that $x_1 = x_2 = \dots = x_r$. The transversality can then also be expressed by saying that the map

$$\begin{aligned}
 \text{Conf}(r, M) &\rightarrow N^r, \\
 (x_1, \dots, x_r) &\mapsto (f(x_1), \dots, f(x_r)),
 \end{aligned}$$

is transverse to the diagonal.

What we really need is a statement about parametrized families of maps, i.e., maps $M \times P \rightarrow N$ where the parameter manifold P is for us a sphere S^k . This can be done by making the following choice: define a submanifold Z of $\text{Conf}(r, M \times P) \times N^r$ by $((x_1, p_1), \dots, (x_r, p_r), (y_1, \dots, y_r))$ such that $p_1 = \dots = p_r$ and $y_1 = \dots = y_r$. To say that a map $M \times P \rightarrow N$ has the associated map

$$\text{Conf}(r, M \times P) \rightarrow \text{Conf}(r, M \times P) \times N^r$$

which is transverse to Z is the same as saying that the map

$$\text{Conf}(r, M) \times P \rightarrow N^r$$

is transverse to the thin diagonal. This does not mean that for every p in P the map $\text{Conf}(r, M) \rightarrow N^r$ is transverse to the diagonal. But the case we are interested in is when transversality does not hit the diagonal at all. This happens if $\dim(\text{Conf}(r, M) \times P) = rm + \dim(P)$ is less than the codimension of the diagonal in N^r , which is $nr - n = n(r - 1)$. In other words, if $\dim(P) < (r - 1)n - rm$, then arranging, by a small homotopy, for $\text{Conf}(r, M) \times P \rightarrow N^r$ to be transverse to the diagonal means arranging for it to not hit the diagonal at all. Then for each point in P , the map $\text{Conf}(r, M) \rightarrow N^r$ of course does not hit it either.

When $P = S^k$, we thus have that, if $k < (r - 1)n - rm$, the map (4) is transverse to the diagonal, which in turn means that, for all $v \in S^k$, we know $h(v) = f$ is an $(r+1)$ -immersion that does not map any r points x_1, x_2, \dots, x_r in M to the same point in N . But this precisely means that h factors through $\text{rImm}(M, N)$; i.e., the dotted arrow in (3) exists.

This argument also works relatively; i.e., it can be repeated for maps

$$S^k \times I \rightarrow (r + 1)\text{Imm}(M, N).$$

The difference now is that the map induced by the adjoint is

$$\text{Conf}(r, M) \times S^k \times I \rightarrow N^r$$

and so the codimension of the preimage of the diagonal is $rm + k + 1 - (r - 1)n$. This means that the map misses the diagonal if $k < (r - 1)n - rm - 1$.

What we have thus shown is the homotopy classes of maps of S^k to $\text{rImm}(M, N)$ and $(r+1)\text{Imm}(M, N)$ are in bijective correspondence if $k < (r - 1)n - rm - 1$; the bijection is induced by the inclusion $\text{rImm}(M, N) \rightarrow (r + 1)\text{Imm}(M, N)$ and, as we have just shown, lifts of maps $S^k \rightarrow (r + 1)\text{Imm}(M, N)$ and $S^k \times I \rightarrow \text{Imm}(M, N)$ to $\text{rImm}(M, N)$. The inclusion thus induces isomorphisms

$$\pi_k(\text{rImm}(M, N)) \xrightarrow{\cong} \pi_k((r + 1)\text{Imm}(M, N)) \quad \text{for } k < (r - 1)n - rm - 1.$$

In addition, since we have a lift of maps $S^k \rightarrow \text{Imm}(M, N)$ for $k < (r - 1)n - rm$, and in particular for $k = (r - 1)n - rm - 1$, we thus have a surjection

$$\pi_k(\text{rImm}(M, N)) \twoheadrightarrow \pi_k((r + 1)\text{Imm}(M, N)) \quad \text{for } k = (r - 1)n - rm - 1.$$

Putting this together means precisely that the inclusion

$$\text{rImm}(M, N) \rightarrow (r + 1)\text{Imm}(M, N)$$

is $((r - 1)n - rm - 1)$ -connected. □

Corollary 2.3. *With the assumptions as in Theorem 2.2, the inclusion*

$$\mathrm{rImm}(M, N) \rightarrow \mathrm{Imm}(M, N)$$

is $((r-1)n-rm-1)$ -connected.

Proof. The proof of Theorem 2.2 goes through the same way with $(r+1)\mathrm{Imm}(M, N)$ replaced by $\mathrm{Imm}(M, N)$. Alternatively, the connectivity number from Theorem 2.2, $((r-1)n-rm-1)$, does not decrease as r goes to infinity since $m \leq n$ (a standing assumption since otherwise there are no immersions or embeddings of M in N). Since $\mathrm{Imm}(M, N)$ is the limit of the inclusions $\mathrm{rImm}(M, N) \rightarrow (r+1)\mathrm{Imm}(M, N)$ and since the least connectivity of those inclusions is hence $((r-1)n-rm-1)$, it follows by Proposition 4.3 that $\mathrm{rImm}(M, N) \rightarrow \mathrm{Imm}(M, N)$ has the same connectivity. \square

What follows immediately from Corollary 2.3 is that, as long as $(r-1)n-rm-1 \geq 0$, the inclusion $\mathrm{rImm}(M, N) \rightarrow \mathrm{Imm}(M, N)$ is surjective on π_0 . In other words, we recover the following familiar result.

Corollary 2.4. *For $n > (r/(r-1))m$, any immersion of M in N is regular homotopic (homotopic through immersions) to an immersion that has no r -fold self-intersections.*

Remark 2.5. When $r = 2$, namely when $\mathrm{rImm}(M, N)$ is the space of embeddings $\mathrm{Emb}(M, N)$, the previous result says that, for $n > 2m$, any immersion of M in N is regular homotopic to an embedding of M in N , and this is the well-known *Whitney easy embedding theorem*.

3. Configuration and r -configuration spaces

The examples of embedding and r -immersion spaces that are most important for us are those of configuration spaces, which is the case when M is a collection of points. Let $\underline{k} = \{1, 2, \dots, k\}$. We then define the *configuration space of k points in N* to be

$$\mathrm{Conf}(k, N) := \mathrm{Emb}(\underline{k}, N) \cong \{(x_1, x_2, \dots, x_k) \in N^k : x_i \neq x_j \text{ for } i \neq j\}. \quad (5)$$

This space can be thought of as N^k with all the diagonals (i.e., the fat diagonal) removed. A related space, and one that is central in this paper, is the *r -configuration space of k points in N* defined by

$$\mathrm{rConf}(k, N) := \mathrm{rImm}(\underline{k}, N).$$

This is the space of configurations of k points in N where at most $r-1$ of them can be equal. In other words, this is N^k with some, but not all, of the diagonals removed, and is for this reason sometimes called a *partial configuration space*. This

space can also be thought of as the complement of the union of certain diagonals in N^k and is hence an example of a complement of a *subspace arrangement*.

As illustrated by the following example, r -configuration spaces are simple in some cases, and it is precisely this simplicity that will allow us to describe the low stages of the Taylor tower for $\mathrm{rImm}(M, N)$ without too much difficulty.

Example 3.1. If $k < r$, then

$$\mathrm{rConf}(k, N) \cong N^k. \tag{6}$$

This is because there is no restriction on the k points being different, and so no diagonals are removed from N^k . In particular, when $k < r$,

$$\mathrm{rConf}(k, \mathbb{R}^n) \cong (\mathbb{R}^n)^k \simeq *. \tag{7}$$

If $k = r$, then

$$\mathrm{rConf}(r, N) \cong N^r \setminus \Delta, \tag{8}$$

where Δ is as before the thin diagonal. This is because any proper subset of the r points is allowed to be equal in $\mathrm{rImm}(r, N)$, but not all of them. In particular, when $N = \mathbb{R}^n$,

$$\mathrm{rConf}(r, \mathbb{R}^n) \cong (\mathbb{R}^n)^r \setminus \Delta \simeq S^{(r-1)n-1}. \tag{9}$$

The homotopy equivalence is given by retracting $(\mathbb{R}^n)^r \setminus \Delta$ onto the orthogonal complement of $\Delta \setminus \{0\}$, which is $(\mathbb{R}^n)^{r-1} \setminus \{0\}$, and then normalizing to length 1.

The space $\mathrm{rConf}(k, N)$ is in general more difficult and less understood than $\mathrm{Conf}(k, N)$. Its (co-)homology is known (see the Introduction of [Dobrinskaya and Turchin 2015] for an overview of the literature dealing with the (co-)homology of $\mathrm{rConf}(k, N)$), and it is known that its suspension is a wedge of spheres [Dobrinskaya and Turchin 2015, Corollary 3.10]. In addition, the connectivity and homotopy groups of $\mathrm{rConf}(k, N)$ through a range were studied in [Kallel and Saihi 2016]. From the point of view of subspace arrangements, the stable homotopy type of these spaces can be identified as the Spanier–Whitehead dual of

$$\bigvee_{p \in P} (\Delta(P_{<p}) * S^{d(p)-1}),$$

where P is the partition poset associated to the arrangement of the diagonals that have been removed and $d(p)$ is the dimension of the subspace corresponding to $p \in P$ [Ziegler and Živaljević 1993].

Spaces of r -configurations are central to the work here, as they are the building blocks for the Taylor tower for $\mathrm{rImm}(M, N)$. Equally important are the projection maps between them and this is where much of the difficulty lies: For ordinary

configuration spaces, the projection maps

$$\text{Conf}(k+1, N) \rightarrow \text{Conf}(k, N)$$

that forget a point are fibrations [Fadell and Neuwirth 1962]. In the case $N = \mathbb{R}^n$, the fiber is even easily identified as \mathbb{R}^n with k points removed, which is homotopy equivalent to $\bigvee_k S^{n-1}$.

However, the corresponding projections of r -configuration spaces,

$$\text{rConf}(k+1, N) \rightarrow \text{rConf}(k, N), \quad (10)$$

are not fibrations, as illustrated in the following example.

Example 3.2. Consider the projection

$$\begin{aligned} 3\text{Conf}(3, \mathbb{R}^n) &\rightarrow 3\text{Conf}(2, \mathbb{R}^n), \\ (x_1, x_2, x_3) &\mapsto (x_1, x_2). \end{aligned}$$

The fiber over a point $(x_1, x_2) \in 3\text{Conf}(2, \mathbb{R}^n)$ where $x_1 \neq x_2$ is \mathbb{R}^n since, in that fiber, x_3 can be anywhere in \mathbb{R}^n , including at x_1 or at x_2 . The fiber over a point (x_1, x_2) where $x_1 = x_2$ is $\mathbb{R}^n \setminus \{x_1\} \simeq S^{n-1}$ since x_3 can be anywhere except at the point $x_1 = x_2$. Since the fibers over two different points have different homotopy type, the map is not a fibration.

Understanding the connectivity of the projection (10), which is important for understanding the Taylor tower of $\text{rImm}(M, N)$, therefore requires understanding its *homotopy* fiber, and not just its fiber(s). The former is unfortunately an unwieldy space. In the case of $N = \mathbb{R}^n$, however, we at least have a handle on the connectivity of the projection map.

Proposition 3.3. *Suppose $r \leq k < l$. Then the map*

$$\text{rConf}(l, \mathbb{R}^n) \rightarrow \text{rConf}(k, \mathbb{R}^n),$$

given by forgetting $l - k$ configuration points, is $((r-1)n-1)$ -connected.

Proof. First look at the map

$$\text{rConf}(k, \mathbb{R}^n) \rightarrow \text{rConf}(r, \mathbb{R}^n) \simeq S^{(r-1)n-1} \quad (11)$$

given by forgetting $k - r$ points (the equivalence comes from (9)). The space $\text{rConf}(k, \mathbb{R}^n)$ is $((r-1)n-2)$ -connected. This follows from Corollary 2.3 with $M = \underline{k}$ and $N = \mathbb{R}^n$, in which case $\text{Imm}(M, N)$ is contractible; this result also appears as Corollary 4.7 in [Kallel and Saihi 2016]. The sphere $S^{(r-1)n-1}$ has the same connectivity. Now, the connectivity of the space $\text{rConf}(k, \mathbb{R}^n)$, $k \geq r$, does not depend on k , so the same argument applies to the connectivity of the map (11) with l replacing k .

Furthermore, since $\text{rConf}(k, \mathbb{R}^n)$ is a subspace arrangement, one can use the Goresky–MacPherson formulas [1988] to study this cohomology, which is generated precisely by the spherical classes represented by the maps (11). In particular, there is an injection

$$H^{(r-1)n-1}(\text{rConf}(k, \mathbb{R}^n)) \rightarrow H^{(r-1)n-1}(\text{rConf}(l, \mathbb{R}^n))$$

given by inclusion of generators. Since everything is finitely generated, we then have a surjection on homology groups in the same degree, and, by the Hurewicz theorem, a surjection on the first nontrivial homotopy group $\pi_{(r-1)n-1}$. \square

The above in particular provides the connectivity of the projection to one fewer points as in (10), but identifying the homotopy fiber of that map is more difficult. The hope is that it is equivalent to a wedge of spheres, like the fiber for projections of ordinary configurations is. One can show that the *suspension* of the homotopy fiber is indeed a wedge of spheres, which provides evidence that the homotopy fiber is as well.

In the case of r -configurations in arbitrary N , the situation is of course more complicated for various reasons, one of them being that we do not have as good of an understanding of the homology of this space. The hope, however, is that the connectivity from Proposition 3.3 remains the same (it does in the case of ordinary configuration spaces in N).

4. Cubical diagrams and total fibers

We will assume the reader is familiar with the language of homotopy limits, including homotopy fibers and homotopy pullbacks. However, we will almost exclusively require these notions only in the case of cubical diagrams, and a source for that material is [Munson and Volić 2015]; foundational material on the subject can be found in [Goodwillie 1991/92].

Let Top be the category of topological spaces and maps between them. We will also sometimes use the same notation for the category of based spaces and maps and this will not cause confusion.

Definition 4.1. • A nonempty space X is k -connected if $\pi_i(X, x) = 0$ for all $0 \leq i \leq k$ and for all choices of basepoint $x \in X$. An infinitely connected space is *weakly contractible*.

• A map $f : X \rightarrow Y$ is k -connected if its homotopy fiber (over any point $y \in Y$) is $(k-1)$ -connected. Equivalently, if $X \neq \emptyset$, then f is k -connected if, for all $x \in X$, the induced map

$$f_* : \pi_i(X, x) \rightarrow \pi_i(Y, f(x))$$

is an isomorphism for all $i < k$ and a surjection for $i = k$. An infinitely connected map is a *weak equivalence*.

Example 4.2. • Path-connected spaces are 0-connected.

- Simply connected spaces are 1-connected.
- The sphere S^k is $(k-1)$ -connected.
- A map between k -connected spaces is $(k-1)$ -connected.
- A map $X \rightarrow *$ is k -connected if and only if X is $(k-1)$ -connected.

For the proof of the following, see, for example, [Munson and Volić 2015, Proposition 2.6.15].

Proposition 4.3. *Given maps $f : X \rightarrow Y$ and $g : Y \rightarrow Z$:*

- *If f and g are k -connected, then $g \circ f$ is k -connected.*
- *If f is $(k-1)$ -connected and $g \circ f$ is k -connected, then g is k -connected.*
- *If g is $(k+1)$ -connected and $g \circ f$ is k -connected, then f is k -connected.*

Let as before $\underline{k} = \{1, 2, \dots, k\}$ and denote by $\mathcal{P}(\underline{k})$ and $\mathcal{P}_0(\underline{k})$ the set of all subsets of \underline{k} and the set of all nonempty subsets of \underline{k} , respectively. Both of these can be regarded as a category (poset) with inclusions as morphisms.

Definition 4.4. • A k -cube, or a *cubical diagram of dimension k* is a (covariant) functor

$$\mathcal{X} : \mathcal{P}(\underline{k}) \rightarrow \text{Top},$$

$$S \mapsto X_S.$$

- A *punctured k -cube*, or a *punctured cubical diagram of dimension k* is the same except the domain is $\mathcal{P}_0(\underline{k})$.

A 0-cube is a space, a 1-cube is a map of spaces, a 2-cube is a commutative square of spaces, etc. A k -cube, for $k \geq 1$, can be regarded as a map, i.e., a natural transformation, of $(k-1)$ -cubes. So a 1-cube is a map of 0-cubes (spaces), a 2-cube (square) is a map of 1-cubes (maps), a 3-cube is a map of 2-cube (squares), and so on.

Removing the initial space X_\emptyset from a cube of spaces leaves a punctured cube of spaces. Furthermore, because X_\emptyset maps into the rest of the cube, one also has a canonical induced map (see [Munson and Volić 2015, Section 5.4] for details)

$$a(\mathcal{X}) : X_\emptyset \rightarrow \text{holim}_{S \in \mathcal{P}_0(\underline{k})} X_S,$$

where holim denotes the homotopy limit.

Definition 4.5. A k -cube is said to be c -cartesian if $a(\mathcal{X})$ is c -connected. If $c = \infty$, the cube is (*homotopy*) *cartesian*.

A cube is *based* if all the spaces in it are based and basepoints map to basepoints. One way to base a cube is to choose a basepoint in X_\emptyset and let it determine basepoints in the rest of the cube. If a cube is based, so is the punctured cube associated to it, and this produces a natural basepoint in $\text{holim}_{S \in \mathcal{P}_0(\underline{k})} X_S$.

Definition 4.6. The *total (homotopy) fiber of a based cube* \mathcal{X} , $\text{tfiber}(\mathcal{X})$, is

$$\text{tfiber}(\mathcal{X}) = \text{hofiber}(a(\mathcal{X})),$$

where the homotopy fiber is taken over the natural basepoint.

A cube is thus c -cartesian if its total fiber is $(c-1)$ -connected.

There is another convenient description of the total fiber of a cube in terms of *iterated fibers*. For the proof of the following, see [Munson and Volić 2015, Proposition 5.5.4].

Proposition 4.7. Let \mathcal{X} be a based cube. For $\underline{k} = \emptyset$, we have $\text{tfiber}(\mathcal{X}) = X_\emptyset$. For $\underline{k} \neq \emptyset$, regard \mathcal{X} as a map of $(k-1)$ -cubes, $\mathcal{Y} \rightarrow \mathcal{Z}$. Then

$$\text{tfiber}(\mathcal{X}) = \text{hofiber}(\text{tfiber}(\mathcal{Y}) \rightarrow \text{tfiber}(\mathcal{Z})).$$

Example 4.8. Let X_1, \dots, X_k be spaces. Consider the cube $\mathcal{X} : \mathcal{P}(\underline{k}) \rightarrow \text{Top}$ given by $\mathcal{X}(S) = \prod_{i \notin S} X_i$ for $S \neq \underline{k}$ and $\mathcal{X}(\underline{k}) = *$. Then this cube is homotopy cartesian. One way to see this is that each square face is homotopy cartesian, i.e., a homotopy pullback, which implies that the cube is as well. For details, see [Munson and Volić 2015, Example 5.4.21].

Example 4.9. A related example is the k -cube where $\mathcal{X}(\emptyset) = \prod_{i \in \underline{k}} X_i$, $\mathcal{X}(\{i\}) = X_i$, and $\mathcal{X}(S) = *$ for all other $S \subset \underline{k}$. This cube is also homotopy cartesian because regarding it as a map of cubes and taking the fiber produces a $(k-1)$ -cube of the sort from the previous example. Since that cube is homotopy cartesian, so is this one.

For more on total homotopy fibers, see [Munson and Volić 2015, Section 5.4].

5. Manifold calculus of functors and the Taylor tower for embeddings

In this section we review some of the main features of manifold calculus of functors. For details, see [Munson and Volić 2015, Sections 10.2 and 10.3; Munson 2010], in addition to the foundational papers [Goodwillie and Weiss 1999; Weiss 1999]. We will throughout pay attention to the functor $\text{Emb}(M, N)$ since this is the one we wish to emulate in our analysis of $\text{rImm}(M, N)$ in Section 6.

Let M and N be smooth manifolds of dimensions m and n , respectively. Let

$\mathcal{O}(M)$ = category (poset) of open subsets of M with inclusions as morphisms.

Manifold calculus studies contravariant functors

$$F : \mathcal{O}(M) \rightarrow \text{Top}$$

that are

- *finitary*, namely for a sequence of open subsets $U_0 \subset U_1 \subset \dots$ the canonical map $F(\bigcup_i U_i) \rightarrow \text{holim}_i F(U_i)$ is a homotopy equivalence; and
- *isotopy functors*, namely they take isotopy equivalences to homotopy equivalences.

Example 5.1. Functors $\text{Imm}(-, N)$, $\text{rImm}(-, N)$, and $\text{Emb}(-, N)$ are all finitary isotopy functors (see [Weiss 1999, Proposition 1.4] for $\text{Imm}(-, N)$ and $\text{Emb}(-, N)$; the argument for $\text{rImm}(-, N)$ is same as for $\text{Imm}(-, N)$). They are contravariant on $\mathcal{O}(M)$ since an inclusion gives a restriction map “going the other way”.

Now let $\mathcal{O}_k(-)$ be the subcategory $\mathcal{O}(-)$ consisting of open subsets of M diffeomorphic to up to k disjoint balls.

Definition 5.2. Let F be a finitary isotopy functor. For $U \in \mathcal{O}(M)$, the k -th stage of the Taylor tower is defined as

$$T_k F(U) = \text{holim}_{V \in \mathcal{O}_k(U)} F(V). \quad (12)$$

The above homotopy limit is in some sense trying to reconstruct $F(U)$ from information about collections of its open balls (in category theory language, this is a *homotopy right Kan extension*).

We then get the *Taylor tower of F* consisting of functors $T_k F$ with natural transformations between them and admitting a natural transformation from F :

$$F(-) \longrightarrow (T_\infty F(-) \rightarrow \dots \rightarrow T_k F(-) \rightarrow T_{k-1} F(-) \rightarrow \dots \rightarrow T_0 F(-)).$$

The transformations between the stages are induced by inclusions $\mathcal{O}_{k-1}(U) \rightarrow \mathcal{O}_k(U)$ and the transformation from F to the *stages* $T_k F$ of the tower by inclusions $\mathcal{O}_k(U) \rightarrow \mathcal{O}(U)$. The functor $T_\infty F(-)$ is the inverse limit of the tower.

Evaluating this diagram on $U \in \mathcal{O}(M)$, we get a diagram of spaces with maps between the stages that are fibrations.

The definition of $T_k F(-)$ is not easy to work with. But there is an alternative way to think about $T_k F$ in terms of cubical diagrams (at the expense of losing some functoriality properties). We will not need this construction here, but the details of how to go from Definition 5.2 to this cubical model can be found in [Munson and Volić 2015, Example 10.2.18].

The stages $T_k F(-)$ are polynomial in the following sense.

Definition 5.3. Let B_1, \dots, B_k be pairwise disjoint open balls in M . Let \mathcal{X} be the k -cube given by

$$\begin{aligned} \mathcal{X} : \mathcal{P}(k) &\rightarrow \text{Top}, \\ S &\mapsto F\left(\bigcup_{i \notin S} B_i\right). \end{aligned}$$

Then define the k -th derivative of F at the empty set, denoted by $F^{(k)}(\emptyset)$, to be the total fiber of \mathcal{X} .

Definition 5.4. A contravariant functor $F : \mathcal{O}(M) \rightarrow \text{Top}$ is *polynomial of degree $\leq k$* if, for all $U \in \mathcal{O}(M)$ and for all pairwise disjoint nonempty closed subsets A_1, \dots, A_{k+1} of U , the cube

$$\begin{aligned} \mathcal{X} : \mathcal{P}(k+1) &\rightarrow \text{Top}, \\ S &\mapsto F\left(U - \bigcup_{i \in S} A_i\right), \end{aligned}$$

is homotopy cartesian.

The two definitions above are related by setting U to be $k+1$ disjoint open balls and A_i to be its components. What falls out of the definitions then is that a polynomial functor of degree $\leq k$ has contractible derivatives of higher order.

Example 5.5. The immersion functor $\text{Imm}(-, N)$ is polynomial of degree ≤ 1 , or *linear* [Weiss 1999, Example 2.3]. The embedding functor $\text{Emb}(-, N)$ is not polynomial of degree $\leq k$ for any k [Munson 2010, Example 4.7].

Theorem 5.6 [Weiss 1999, Theorem 3.9]. *Let $F : \mathcal{O}(M) \rightarrow \text{Top}$ be a contravariant finitary isotopy functor. The k -th stage $T_k F$ of the Taylor tower for F is polynomial of degree $\leq k$.*

Polynomial functors can be characterized by what they do on balls.

Theorem 5.7 [Weiss 1999, Theorem 5.1]. *Suppose F and G are contravariant finitary isotopy functors from $\mathcal{O}(M)$ to Top that are polynomial of degree $\leq k$. Suppose $T : F \rightarrow G$ is a natural transformation that is an equivalence for all $U \in \mathcal{O}_k(M)$. Then T is an equivalence for all $U \in \mathcal{O}(M)$.*

The proof of the next result combines Theorem 5.7 with the fact that embeddings and immersions agree on a single ball up to homotopy. We will repeat this argument when we deduce that the linearization of the space of r -immersions is also the space of immersions.

Proposition 5.8 [Weiss 1999, bottom of page 97]. *The linearization of the space of embeddings is the space of immersions, namely there is an equivalence*

$$T_1 \text{Emb}(-, N) \simeq \text{Imm}(-, N).$$

There are two natural questions one can ask about the Taylor tower:

- (1) Does the Taylor tower for F converge?
- (2) Does it converge to F ?

The convergence of the tower means that the connectivity of the maps $T_k F \rightarrow T_{k-1} F$ grows to infinity with k . This can be established by looking at the spaces

$$L_k F(-) := \text{hofiber}(T_k F(-) \rightarrow T_{k-1} F(-)) \quad (13)$$

and showing that their connectivity grows with k . Here we need to be working with a based Taylor tower, which can be accomplished by choosing a basepoint in $F(M)$ which in turn bases $T_k(U)$ for all k and U . The functor $L_k F(-)$ is called the k -th layer of the Taylor tower of F and is *homogeneous* in the sense that all its derivatives of degree $< k$ are trivial. One of the main results in manifold calculus of functors is the statement that classifies all homogeneous functors in terms of certain spaces of sections [Weiss 1999, Theorem 8.5]. We will not need this result here, but will use the following consequence; see [Munson 2010, Proposition 5.1.1] for more details.

Proposition 5.9. *Let M be of dimension m and suppose $F : \mathcal{O}(M) \rightarrow \text{Top}$ is a contravariant finitary isotopy functor. If the derivative $F^{(k)}(\emptyset)$ (see Definition 5.3) is c_k -connected, then $L_k F(M)$ is $(c_k - km)$ -connected and, consequently, the map $T_k(M) \rightarrow T_{k-1}(M)$ is $(c_k - km + 1)$ -connected.*

The above is true when M is compact and is replaced by any interior of a compact codimension-zero handlebody, in which case the handle dimension (the highest dimension of a handle necessary to build the handlebody) replaces m . However, we only focus on the case of M itself, and this will remain true for the rest of the paper and for our results. This is fine since, in most applications of manifold calculus, this is the only case of importance.

For the embedding functor in particular, we have the following result.

Theorem 5.10. *The derivative $\text{Emb}(M, N)^{(k)}(\emptyset)$ is $(k-1)(n-2)$ -connected, and hence the connectivity of the map $T_k \text{Emb}(M, N) \rightarrow T_{k-1} \text{Emb}(M, N)$ is*

$$(k-1)(n-2) - km + 1 = (k-1)(n-m-2) - m + 1.$$

For the proof of Theorem 5.10, see, for example, [Munson and Volić 2015, Theorem 10.3.3]. In brief, one first passes from embeddings of balls, which are in the definition of the derivative, to embeddings of points, namely configuration spaces, and this is fine since balls are homotopy equivalent to points. The proof then comes down to showing that the cube of configuration spaces and projections between them is $((k-1)(n-2) + 1)$ -cartesian, and hence the total fiber, i.e., the derivative, has connectivity one less; one reference for this is [Munson and Volić 2015, Example 6.2.9]. The key is that projection maps between configuration spaces are fibrations, so that their (homotopy) fiber is easy to handle. More will be said about this proof, and the ways in which the case of r -immersions is more complicated, in the discussion at the end of the paper.

The second question about the convergence of the Taylor tower, namely whether the tower converges to F , is a more difficult one. By convergence to F we mean that the map $F \rightarrow T_\infty F = \text{holim}_k T_k$ is an equivalence, i.e., infinitely connected. One way to establish this would be to argue that the connectivity of the maps $F \rightarrow T_k F$ grows with k . This is precisely what happens with the embedding functor. Namely, we have the following result from [Goodwillie and Weiss 1999], which builds heavily on the work in [Goodwillie and Klein 2015].

Theorem 5.11. *Suppose M is a smooth closed manifold of dimension m , N is a smooth manifold of dimension n , and $n - m \geq 2$. Then the map*

$$\text{Emb}(M, N) \rightarrow T_k \text{Emb}(M, N)$$

is $(k(n-m-2)-m+1)$ -connected. If $n - m > 2$, then the connectivities grow with k and the Taylor tower therefore converges to $\text{Emb}(M, N)$.

In line with the comments following Proposition 5.9, the above is true when M is replaced by the interior of a codimension-zero handlebody, in which case m has to be replaced by its handle dimension. The proof is difficult and requires various disjunction results for embeddings; for an overview, see [Munson and Volić 2015, Section 10.3.2].

Remark 5.12. The number in Theorem 5.11 is the same as that in Theorem 5.10 (with a shift in k). This is not surprising since the first number can be used to conjecture the second: Suppose we know that the Taylor tower converges, i.e., the connectivities of the maps between the stages increase, and that it converges to F . Suppose $T_{k+1} F(M) \rightarrow T_k F(M)$ is c -connected. Then we have the diagram

$$\begin{array}{ccc} F(M) & \xrightarrow{\sim} & T_\infty F(M) \\ & \searrow & \downarrow c\text{-connected} \\ & & T_k F(M) \end{array}$$

The vertical map in the above diagram is also c -connected, due to Proposition 4.3, because it is the composition of maps for which the least connectivity is c . Again by Proposition 4.3, it then follows that the connectivity of $F(M) \rightarrow T_k F(M)$ must also be c . For $F(M) = \text{Emb}(M, N)$, c is precisely $k(n - m - 2) - m + 1$.

6. First r stages of the Taylor tower for r -immersions

In this section we give the description of the connectivities between the first r stages of the Taylor tower for r -immersions. Much of what we do can be adapted from the case of M to the case of a codimension-zero handlebody in M (see comments following Proposition 5.9 and Theorem 5.11), but we content ourselves with the case of M because that is the most important and useful one. Additionally, the

generalization would require us to venture outside the intended scope and level of difficulty of this paper. The general case will be tackled in future work.

We start with $T_1 \text{rImm}(M, N)$. The following is analogous to Proposition 5.8.

Proposition 6.1. *The linearization of the space of r -immersions is the space of immersions, namely there is an equivalence*

$$T_1 \text{rImm}(M, N) \simeq \text{Imm}(M, N).$$

Proof. We already know from Example 5.5 that $\text{Imm}(M, N)$ is linear and from Theorem 5.6 that $T_1 \text{rImm}(M, N)$ is linear. Therefore by Theorem 5.7, to prove the desired equivalence it suffices to check that the equivalence holds for a single ball. Namely, with B^m the m -dimensional ball, we want to show that

$$\text{rImm}(B^m, N) \simeq \text{Imm}(B^m, N).$$

But the argument for this is the same as showing that spaces of embeddings and immersions agree on a single ball. Namely, one can differentiate an immersion or an r -immersion at, say, the center of the ball, to get a point in the Stiefel manifold of m -frames in the tangent space of N . The fibers of both maps are contractible, which means that $\text{rImm}(B^m, N)$ and $\text{Imm}(B^m, N)$ are equivalent. \square

We next look at the layers of the Taylor tower for $\text{rImm}(M, N)$, which, by Proposition 5.9, will require us to understand the derivatives $\text{rImm}(M, N)^{(k)}(\emptyset)$. Looking back at Definition 5.3, the derivatives are total fibers of cubical diagrams consisting of spaces of r -immersions of unions of balls. But, as alluded to in the discussion following Theorem 5.10, it is possible to replace these with diagrams of r -immersions of points, namely with diagrams of $\text{rConf}(k, N)$. The argument for this is identical to that for the case of embeddings. For details, see the discussion following the statement of [Munson 2010, Theorem 7.2].

Suppose the Taylor tower for $\text{rImm}(M, N)$ has been based; see the comment after (13). Recalling from Definition 4.1 that a weakly contractible space is one that is infinitely connected, we have the following result.

Proposition 6.2. *For $2 \leq k \leq r - 1$, the layer $L_k \text{rImm}(M, N)$ of the Taylor tower for $\text{rImm}(M, N)$ is weakly contractible.*

Proof. It suffices to show that the derivative $\text{rImm}^{(k)}(\emptyset, N)$ is weakly contractible. Then, by Proposition 5.9, it follows that $L_k \text{rImm}(M, N)$ is weakly contractible as well.

By the discussion above, $\text{rImm}^{(k)}(\emptyset, N)$ is the total fiber of the cubical diagram

$$\begin{aligned} \mathcal{X} : \mathcal{P}(k) &\rightarrow \text{Top}, \\ S &\mapsto \text{rConf}(k - |S|, N), \end{aligned}$$

where the maps are projections given by forgetting points in the configuration, namely an inclusion $T \rightarrow S$ gives a map that projects away from those configuration points indexed by $S - T$. However, since $k < r$, as observed in Example 3.1,

$$\mathrm{rConf}(k - |S|, N) = N^{k - |S|}.$$

The cube in question is thus equivalent to the cube

$$\mathcal{X} : \mathcal{P}(\underline{k}) \rightarrow \mathrm{Top}, \quad (14)$$

$$S \mapsto \begin{cases} N^{k - |S|}, & S \neq \underline{k}, \\ *, & S = \underline{k}, \end{cases} \quad (15)$$

with projection maps as before projecting away from those factors indexed by $S - T$. But this cube is homotopy cartesian by Example 4.8. This means precisely that its total fiber is weakly contractible. Equivalent cubes have equivalent total fibers, and so the total fiber of the original cube, namely $\mathrm{rImm}^{(k)}(\emptyset, N)$, is also weakly contractible. \square

Theorem 6.3. *For $2 \leq k \leq r - 1$, we have a weak equivalence*

$$T_k \mathrm{rImm}(M, N) \xrightarrow{\simeq} T_{k-1} \mathrm{rImm}(M, N).$$

Proof. By Proposition 6.2, the fiber $L_k \mathrm{rImm}(M, N)$ of the fibration

$$T_k \mathrm{rImm}(M, N) \rightarrow T_{k-1} \mathrm{rImm}(M, N)$$

is weakly contractible in the given range. By the homotopy long exact sequence of a fibration, it follows that the map

$$T_k \mathrm{rImm}(M, N) \rightarrow T_{k-1} \mathrm{rImm}(M, N)$$

is infinitely connected, or a weak equivalence. \square

Corollary 6.4. *For $1 \leq k \leq r - 1$, the map*

$$\mathrm{rImm}(M, N) \rightarrow T_k \mathrm{rImm}(M, N)$$

is $((r-1)n - rm - 1)$ -connected.

Proof. We know from Corollary 2.3 and Proposition 6.1 that the map

$$\mathrm{rImm}(M, N) \rightarrow T_1 \mathrm{rImm}(M, N) \simeq \mathrm{Imm}(M, N)$$

is $((r-1)n - rm - 1)$ -connected. Inducting up the tower and using Proposition 4.3 along with Theorem 6.3 gives the desired result. \square

Proposition 6.6 will give the connectivity of the next layer, $L_r \mathrm{rImm}(M, N)$. Before we prove it, we need the following result.

Lemma 6.5. *For N a manifold of dimension n , the inclusion*

$$N^r \setminus \Delta \rightarrow N^r,$$

where Δ is the thin diagonal, is $((r-1)n-1)$ -connected.

Proof. This proof is a simpler version of the one from Theorem 2.2. Namely, a map $S^k \rightarrow N^r$ by transversality generically misses the thin diagonal if $k < (r-1)n$, and a map $S^k \times I \rightarrow N^r$ misses it if $k < (r-1)n-1$. This means that, under those dimensional assumptions, the maps are homotopic to maps that lift to $N^r \setminus \Delta$, and this means that the map $N^r \setminus \Delta \rightarrow N^r$ induces isomorphisms on homotopy groups π_k for $k < (r-1)n-1$ and a surjection on $\pi_{(r-1)n-1}$, which is what we wanted to show. \square

Proposition 6.6. *The layer $L_r \mathbf{rImm}(M, N)$ of the Taylor tower for $\mathbf{rImm}(M, N)$ is $((r-1)n-rm-2)$ -connected.*

Remark 6.7. If we knew that the Taylor tower converged, Proposition 6.6 would in fact be immediate from what we know already. As we know from Remark 5.12, convergence would mean that the connectivity of

$$\mathbf{rImm}(M, N) \rightarrow T_r \mathbf{rImm}(M, N)$$

is greater than that of

$$\mathbf{rImm}(M, N) \rightarrow T_{r-1} \mathbf{rImm}(M, N),$$

which is $(r-1)n-rm-1$ from Corollary 6.4. We would thus have the diagram

$$\begin{array}{ccc} \mathbf{rImm}(M, N) & \xrightarrow{>(r-1)n-rm-1} & T_r \mathbf{rImm}(M, N) \\ & \searrow_{(r-1)n-rm-1} & \downarrow \\ & & T_{r-1} \mathbf{rImm}(M, N) \end{array}$$

and it would then follow from Proposition 4.3 that

$$T_r \mathbf{rImm}(M, N) \rightarrow T_{r-1} \mathbf{rImm}(M, N)$$

is $((r-1)n-rm-1)$ -connected. But this is the same as saying that the fiber of this map, namely $L_r \mathbf{rImm}(M, N)$, is $((r-1)n-rm-2)$ -connected.

Proof of Proposition 6.6. We start as in the proof of Proposition 6.2, and so $\mathbf{rImm}^{(r)}(\emptyset, N)$ is the total fiber of the cubical diagram

$$\begin{aligned} \mathcal{X} : \mathcal{P}(\underline{r}) &\rightarrow \text{Top}, \\ S &\mapsto \mathbf{rConf}(r - |S|, N), \end{aligned}$$

with projections as before. The initial space in the cube, by Example 3.1, is

$$\mathbf{rConf}(r - |\emptyset|, N) = \mathbf{rConf}(r, N) = N^r \setminus \Delta.$$

The other spaces are, as in the proof of Proposition 6.2, $\text{rConf}(r - |S|, N) = N^{r-|S|}$ since then $r - |S| < r$. So for example, when $r = 3$, the cube \mathcal{X} is equivalent to

$$\begin{array}{ccccc}
 N^3 \setminus \Delta & \longrightarrow & N^2 & & \\
 \downarrow & \searrow & \downarrow & \searrow & \\
 & & N^2 & \longrightarrow & N \\
 & & \downarrow & \searrow & \downarrow \\
 N^2 & \longrightarrow & N & & \\
 & \searrow & \downarrow & \searrow & \\
 & & N & \longrightarrow & *
 \end{array} \tag{16}$$

Thus

$$\text{rImm}^{(r)}(\emptyset, N) = \text{tfiber } \mathcal{X} \simeq \text{hofiber}(N^r \setminus \Delta \rightarrow \text{holim}_{S \in \mathcal{P}_0(r)} N^{r-|S|}).$$

However, the homotopy limit of the punctured cube on the right is N^r . This follows from Example 4.8 by setting all $X_i = N$ (and $k = r$); that example says precisely that there is an equivalence

$$N^r \xrightarrow{\simeq} \text{holim}_{S \in \mathcal{P}_0(r)} N^{r-|S|}.$$

So now we have reduced the problem to finding the connectivity of the inclusion

$$N^r \setminus \Delta \rightarrow N^r.$$

But by Lemma 6.5, this map is $((r-1)n-1)$ -connected, which means that its homotopy fiber, namely $\text{rImm}^{(r)}(\emptyset, N)$, is $((r-1)n-2)$ -connected. It then follows by Proposition 5.9 that the connectivity of $L_r \text{rImm}(M, N)$ is

$$(r-1)n-2-rm = (r-1)n-rm-2. \quad \square$$

Remark 6.8. When $N = \mathbb{R}^n$, which is often of most interest, then the above proof simplifies as follows: The initial space in the cube is

$$\text{rConf}(r, \mathbb{R}^n) = (\mathbb{R}^n)^r \setminus \Delta \simeq S^{(r-1)n-1},$$

while the other spaces are products of \mathbb{R}^n with itself and hence contractible. When $r = 3$ for example, we want the total fiber of the cube

$$\begin{array}{ccccc}
 S^{2n-1} & \longrightarrow & * & & \\
 \downarrow & \searrow & \downarrow & \searrow & \\
 & & * & \longrightarrow & * \\
 & & \downarrow & \searrow & \downarrow \\
 * & \longrightarrow & * & & \\
 & \searrow & \downarrow & \searrow & \\
 & & * & \longrightarrow & *
 \end{array} \tag{17}$$

But this total fiber is simply S^{2n-1} , which is $(2n-2)$ -connected, or in the general case of $S^{(r-1)n-1}$, $((r-1)n-2)$ -connected, as desired.

7. Some comments about the higher stages

We end by saying a few words about the strategy for finding the connectivities of the higher layers $L_{k+1} \text{rImm}(M, N)$, when $k+1 > r$ (we are indexing by $k+1$ rather than k since the numbers will ultimately come out easier that way). These layers exhibit genuinely different and more difficult behavior. We will focus on the case $N = \mathbb{R}^n$, as more is known about partial configuration spaces in this situation. Details will appear in the thesis work of Šarčević.

As before, to get at the connectivity of $L_{k+1} \text{rImm}(M, \mathbb{R}^n)$, one would first establish the connectivity of the derivative $\text{rImm}^{(k+1)}(\emptyset, \mathbb{R}^n)$. This is the total fiber of the cube

$$\begin{aligned} \mathcal{X} : \mathcal{P}(k+1) &\rightarrow \text{Top}, \\ S &\mapsto X_S = \text{rConf}((k+1) - |S|, \mathbb{R}^n), \end{aligned}$$

with projection maps. The total fiber is

$$\text{rImm}^{(k+1)}(\emptyset, \mathbb{R}^n) = \text{tfiber } \mathcal{X} \simeq \text{hofiber}(X_\emptyset \rightarrow \underset{S \in \mathcal{P}_0(k+1)}{\text{holim}} X_S) \quad (18)$$

Remark 7.1. In the case $k+1 = r+1$, the homotopy limit in the above is simply $(S^{(r-1)n-1})^{r+1}$ (using Example 4.9). The map from (18) whose homotopy fiber we wish to understand is therefore equivalent to the map

$$\text{rConf}(r+1, \mathbb{R}^n) \rightarrow (S^{(r-1)n-1})^{r+1}. \quad (19)$$

One way to get the connectivity of this map is to look at the homology of the two spaces using [Goresky and MacPherson 1988].

The idea now is to emulate the proof in the case of $\text{Emb}(M, \mathbb{R}^n)$ and use the Blakers–Massey Theorem for cubes [Goodwillie 1991/92, Theorem 2.5]; see also [Munson and Volić 2015, Section 6.2]. There are two pieces that would be required:

- (1) The connectivities of the projection maps between r -configuration spaces.
- (2) One would take the homotopy fibers in one direction in \mathcal{X} and then look at each square face of the resulting k -cube; i.e., look at the top square in the diagram

$$\begin{array}{ccccc} Y_\emptyset & \xrightarrow{\quad} & Y_1 & & \\ & \searrow & & \searrow & \\ & & Y_2 & \xrightarrow{\quad} & Y_{12} \\ \text{rConf}(l, \mathbb{R}^n) & \xrightarrow{\quad} & \text{rConf}(l-1, \mathbb{R}^n) & \xrightarrow{\quad} & \text{rConf}(l-2, \mathbb{R}^n) \\ & \searrow & & \searrow & \\ & & \text{rConf}(l-1, \mathbb{R}^n) & \xrightarrow{\quad} & \text{rConf}(l-2, \mathbb{R}^n) \\ \text{rConf}(l-1, \mathbb{R}^n) & \xrightarrow{\quad} & \text{rConf}(l-2, \mathbb{R}^n) & \xrightarrow{\quad} & \text{rConf}(l-3, \mathbb{R}^n) \\ & \searrow & & \searrow & \\ & & \text{Conf}(l-2, \mathbb{R}^n) & \xrightarrow{\quad} & \text{rConf}(l-3, \mathbb{R}^n) \end{array}$$

for various $3 \leq l \leq k + 1$. One that proves that each such square is *homotopy cocartesian*; i.e., the map from the homotopy colimit of the diagram obtained by removing the final space in the square to the final space is an equivalence.

The first item is taken care of by Proposition 3.3, but the second is more difficult. In the case we are trying to mimic, that of $\text{Emb}(M, \mathbb{R}^n)$, one is aided by the fact that the corresponding cube consists of ordinary configuration spaces and that the projection maps between them are fibrations. The homotopy fibers are just fibers, and are easy to understand (they are wedges of spheres). The squares of fibers turn out to be simple, built out of an intersection and a union; see [Munson and Volić 2015, Example 3.7.5]. But as we already discussed in Section 3, the projections of r -configurations are not fibrations, so one has to examine their homotopy fibers directly.

For general N , the situation is even more difficult since we do not know much about the connectivity or the homology of the spaces $\text{rConf}(k, N)$.

Furthermore, without knowing that the tower converges, the connectivity of $\text{rImm}(M, \mathbb{R}^n) \rightarrow T_k \text{rImm}(M, \mathbb{R}^n)$ is harder to obtain. For embeddings, deep *disjunction* results [Goodwillie and Klein 2015] (see also [Munson and Volić 2015, Section 10.3.2] for an overview) are required, and generalizing these to r -immersions is likely difficult.

Acknowledgements

Volić would like to thank Tom Goodwillie, Sadok Kallel, and Rade Živaljević for helpful conversations, as well as the Simons Foundation for its support.

References

- [Arone and Turchin 2014] G. Arone and V. Turchin, “On the rational homology of high-dimensional analogues of spaces of long knots”, *Geom. Topol.* **18**:3 (2014), 1261–1322. MR Zbl
- [Arone et al. 2007] G. Arone, P. Lambrechts, and I. Volić, “Calculus of functors, operad formality, and rational homology of embedding spaces”, *Acta Math.* **199**:2 (2007), 153–198. MR Zbl
- [Arone et al. 2008] G. Arone, P. Lambrechts, V. Turchin, and I. Volić, “Coformality and rational homotopy groups of spaces of long knots”, *Math. Res. Lett.* **15**:1 (2008), 1–14. MR Zbl
- [Boavida de Brito and Weiss 2013] P. Boavida de Brito and M. Weiss, “Manifold calculus and homotopy sheaves”, *Homology Homotopy Appl.* **15**:2 (2013), 361–383. MR Zbl
- [Boavida de Brito and Weiss 2018] P. Boavida de Brito and M. Weiss, “Spaces of smooth embeddings and configuration categories”, *J. Topol.* **11**:1 (2018), 65–143. MR Zbl
- [Dobrinskaya and Turchin 2015] N. Dobrinskaya and V. Turchin, “Homology of non- k -overlapping discs”, *Homology Homotopy Appl.* **17**:2 (2015), 261–290. MR Zbl
- [Dwyer and Hess 2012] W. Dwyer and K. Hess, “Long knots and maps between operads”, *Geom. Topol.* **16**:2 (2012), 919–955. MR Zbl
- [Fadell and Neuwirth 1962] E. Fadell and L. Neuwirth, “Configuration spaces”, *Math. Scand.* **10** (1962), 111–118. MR Zbl

- [Frick 2015] F. Frick, “Counterexamples to the topological Tverberg conjecture”, preprint, 2015. arXiv
- [Goodwillie 1991/92] T. G. Goodwillie, “Calculus, II: Analytic functors”, *K-Theory* **5**:4 (1991/92), 295–332. MR Zbl
- [Goodwillie and Klein 2015] T. G. Goodwillie and J. R. Klein, “Multiple disjunction for spaces of smooth embeddings”, *J. Topol.* **8**:3 (2015), 651–674. MR Zbl
- [Goodwillie and Munson 2010] T. G. Goodwillie and B. A. Munson, “A stable range description of the space of link maps”, *Algebr. Geom. Topol.* **10**:3 (2010), 1305–1315. MR Zbl
- [Goodwillie and Weiss 1999] T. G. Goodwillie and M. Weiss, “Embeddings from the point of view of immersion theory, II”, *Geom. Topol.* **3** (1999), 103–118. MR Zbl
- [Goodwillie et al. 2001] T. G. Goodwillie, J. R. Klein, and M. S. Weiss, “Spaces of smooth embeddings, disjunction and surgery”, pp. 221–284 in *Surveys on surgery theory*, vol. 2, edited by S. Cappell et al., Ann. of Math. Stud. **149**, Princeton Univ. Press, 2001. MR Zbl
- [Goresky and MacPherson 1988] M. Goresky and R. MacPherson, *Stratified Morse theory*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3) **14**, Springer, 1988. MR Zbl
- [Kallel and Saihi 2016] S. Kallel and I. Saihi, “Homotopy groups of diagonal complements”, *Algebr. Geom. Topol.* **16**:5 (2016), 2949–2980. MR Zbl
- [Lambrechts et al. 2010] P. Lambrechts, V. Turchin, and I. Volić, “The rational homology of spaces of long knots in codimension > 2 ”, *Geom. Topol.* **14**:4 (2010), 2151–2187. MR Zbl
- [Mabillard and Wagner 2014] I. Mabillard and U. Wagner, “Eliminating Tverberg points, I: An analogue of the Whitney trick”, pp. 171–180 in *Computational geometry (SoCG’14)*, ACM, New York, 2014. MR Zbl
- [Mabillard and Wagner 2016] I. Mabillard and U. Wagner, “Eliminating higher-multiplicity intersections, II: The deleted product criterion in the r -metastable range”, art. id. 51 in *32nd International Symposium on Computational Geometry*, edited by S. Fekete and A. Lubiw, LIPIcs. Leibniz Int. Proc. Inform. **51**, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2016. MR Zbl
- [Munson 2005] B. A. Munson, “Embeddings in the $\frac{3}{4}$ range”, *Topology* **44**:6 (2005), 1133–1157. MR Zbl
- [Munson 2008] B. A. Munson, “A manifold calculus approach to link maps and the linking number”, *Algebr. Geom. Topol.* **8**:4 (2008), 2323–2353. MR Zbl
- [Munson 2010] B. A. Munson, “Introduction to the manifold calculus of Goodwillie–Weiss”, *Morfismos* **14**:1 (2010), 1–50.
- [Munson 2011] B. A. Munson, “Derivatives of the identity and generalizations of Milnor’s invariants”, *J. Topol.* **4**:2 (2011), 383–405. MR Zbl
- [Munson and Volić 2012] B. A. Munson and I. Volić, “Multivariable manifold calculus of functors”, *Forum Math.* **24**:5 (2012), 1023–1066. MR Zbl
- [Munson and Volić 2014] B. A. Munson and I. Volić, “Cosimplicial models for spaces of links”, *J. Homotopy Relat. Struct.* **9**:2 (2014), 419–454. MR Zbl
- [Munson and Volić 2015] B. A. Munson and I. Volić, *Cubical homotopy theory*, New Mathematical Monographs **25**, Cambridge University Press, 2015. MR Zbl
- [Salikhov 2002] K. Salikhov, “Multiple points of immersions”, preprint, 2002. arXiv
- [Sinha 2006] D. P. Sinha, “Operads and knot spaces”, *J. Amer. Math. Soc.* **19**:2 (2006), 461–486. MR Zbl
- [Songhafou Tsopmnéné 2016] P. A. Songhafou Tsopmnéné, “The rational homology of spaces of long links”, *Algebr. Geom. Topol.* **16**:2 (2016), 757–782. MR Zbl

- [Tillmann 2019] S. Tillmann, “Manifold calculus adapted for simplicial complexes”, *Homology Homotopy Appl.* **21**:1 (2019), 161–186. MR Zbl
- [Turchin 2013] V. Turchin, “Context-free manifold calculus and the Fulton–MacPherson operad”, *Algebr. Geom. Topol.* **13**:3 (2013), 1243–1271. MR Zbl
- [Volić 2006] I. Volić, “Finite type knot invariants and the calculus of functors”, *Compos. Math.* **142**:1 (2006), 222–250. MR Zbl
- [Weiss 1999] M. Weiss, “Embeddings from the point of view of immersion theory, I”, *Geom. Topol.* **3** (1999), 67–101. MR Zbl
- [Ziegler and Živaljević 1993] G. M. Ziegler and R. T. Živaljević, “Homotopy types of subspace arrangements via diagrams of spaces”, *Math. Ann.* **295**:3 (1993), 527–548. MR Zbl

Received: 2018-11-13 Revised: 2019-09-09 Accepted: 2019-11-12

bschrein@nd.edu *Department of Mathematics, University of Notre Dame,
Notre Dame, IN, United States*

franjo.sarcevic@live.de *Department of Mathematics, University of Sarajevo,
Bosnia and Herzegovina*

ivolic@wellesley.edu *Department of Mathematics, Wellesley College,
Wellesley, MA, United States*

A new go-to sampler for Bayesian probit regression

Scott Simmons, Elizabeth J. McGuffey and Douglas VanDerwerken

(Communicated by Jem Noelle Corcoran)

This paper introduces an independent-proposal Metropolis–Hastings sampler for Bayesian probit regression. The Gibbs sampler of Albert and Chib has been the default sampler since its introduction in 1993. We conduct a simulation study comparing the two methods under various combinations of predictor variables and sample sizes. The proposed sampler is shown to outperform that of Albert and Chib in terms of efficiency measured through autocorrelation, effective sample size, and computation time. We then demonstrate performance of the samplers on real data applications with analogous results.

1. Introduction

Modeling binary response data is important in statistics and related fields of biostatistics and econometrics. The classical approach is to fit a logistic or probit regression model via maximum likelihood and to conduct inference based on asymptotic theory. A thorough summary of this approach is given by [Amemiya 1981]. When the sample size n is small, however, reliance on asymptotic theory is dubious; for example, [Griffiths et al. 1987] shows that the maximum likelihood estimator (MLE) exhibits nonnegligible bias in finite samples.

One solution is to take a Bayesian approach. The Bayesian paradigm is based on Bayes' rule, which in our modeling context may be generally expressed as

$$p(\boldsymbol{\theta} | \mathbf{y}) = \frac{\pi(\boldsymbol{\theta})p(\mathbf{y} | \boldsymbol{\theta})}{p(\mathbf{y})} \propto \pi(\boldsymbol{\theta})p(\mathbf{y} | \boldsymbol{\theta}),$$

where $p(\mathbf{y} | \boldsymbol{\theta})$ is the data likelihood; $\pi(\boldsymbol{\theta})$ is the prior distribution assigned to $\boldsymbol{\theta}$, representing one's a priori beliefs about $\boldsymbol{\theta}$; $p(\mathbf{y})$ is the marginal likelihood; and $p(\boldsymbol{\theta} | \mathbf{y})$ is the posterior distribution, describing one's updated beliefs about $\boldsymbol{\theta}$. Interest lies in conducting inference on $\boldsymbol{\theta}$ based on the posterior distribution.

For many data likelihoods, there exists a prior distribution that leads to a posterior distribution of the same family that may be expressed in closed form. There is no

MSC2010: 62C10, 62F15.

Keywords: probit regression, Markov chain Monte Carlo, Gibbs sampling, Metropolis–Hastings.

such conjugate prior for either the logistic or probit likelihoods, so computational techniques must be utilized to summarize and/or sample from the posterior distribution. For instance, [Zellner and Rossi 1984] recommends numerical integration when the number of predictors p is small and importance sampling with a multivariate t instrumental distribution otherwise.

The advent of Gibbs sampling [Geman and Geman 1984; Gelfand and Smith 1990] introduced a whole class of new methods for doing posterior inference. In its simplest form, Gibbs sampling generates random draws from a multivariate posterior distribution by iteratively generating univariate draws from the full conditional distributions $p(\theta_k | \theta_{-k}, \mathbf{y})$, where θ_{-k} is the parameter vector θ excluding the k -th entry. Zeger and Karim [1991] were the first to apply Gibbs sampling to a logistic regression model. However, in this approach not all of the full conditional distributions are available in closed form.

In a now landmark paper, Albert and Chib [1993] introduced a novel Gibbs sampler for sampling from the posterior distribution of a probit regression model with a multivariate normal prior. Thanks to a clever data augmentation technique, the full conditional distributions in the Albert and Chib Gibbs sampler are all named distributions. Albert and Chib's method can be used to obtain *approximate* posterior samples from the logistic regression model, but it took more than a decade for researchers to devise an *exact* (i.e., asymptotically convergent) data-augmentation-based Gibbs sampler for the logistic regression model; several now exist [Holmes and Held 2006; Frühwirth-Schnatter and Frühwirth 2007; Polson et al. 2013]. Holmes and Held [2006], in addition to their contribution to Bayesian logistic regression, also introduced an ostensible improvement to Albert and Chib's probit Gibbs sampler in the form of a joint update of the regression and auxiliary parameters.

In this paper, we draw attention to a Markov chain Monte Carlo (MCMC) sampling algorithm related to the importance sampling approach of [Zellner and Rossi 1984] and demonstrate that it performs favorably in a wide variety of settings when compared to the Albert and Chib sampler. In particular, we recommend that samples from the probit posterior be obtained using an independence Metropolis–Hastings (MH) sampler with a multivariate t proposal distribution centered at the posterior mode and having variance equal to the inverse Hessian evaluated at the mode. Variations of this sampler have been mentioned in the literature, but to our knowledge no one has undertaken to systematically compare it to Albert and Chib's.

Although our algorithm is closely related to the importance sampling scheme of [Zellner and Rossi 1984], one minor difference is that we obtain actual samples from the posterior, allowing for natural comparison to the Albert and Chib sampler, whereas Zellner and Rossi approximated integrals with appropriately weighted draws from the multivariate t -distribution. Perhaps a more notable difference is that we use the inverse Hessian evaluated at the *mode* instead of the (moderately

scaled) inverse Hessian evaluated at the MLE. This difference is nontrivial when the contribution from the prior is considerable.

While the idea of using an independence MH sampler with multivariate t proposal was not explicitly discussed in [Albert and Chib 1993], the authors do suggest that a difficult full conditional may be approximated by a normal or t -distribution centered at the mode and having variance determined by the curvature at the mode; however, normality of the posterior itself is dismissed for small n . In this paper we show that even in the small- n scenario the independence MH sampler is often more efficient.

Polson et al. [2013] discuss the independence MH sampler at length, but in the context of logistic regression, as opposed to probit. They acknowledge that MH was faster than the auxiliary-variable Gibbs sampler in five of eight data sets studied, but warn against its use when the model has a complex prior structure, as is the case for, e.g., mixed models, factor models, and models with spatiotemporal dependence. We realize that similar arguments could be made in favor of the probit Gibbs sampler, but this does not negate the importance of our contribution. We hope to show that the deference which many practicing Bayesians give the Albert and Chib algorithm in the context of “vanilla” probit regression might be better given to the independence MH sampler. To take an extreme example, [Greene 2012, p. 713] goes so far as to all but reject the Bayesian paradigm because he finds the Albert and Chib algorithm too slow—as if no other sampler were available! Even for the more complicated settings alluded to in [Polson et al. 2013], hybrid Metropolis–Hastings–within–Gibbs samplers may yet provide the “best of both worlds”; see, for example, [Gamerman 1997; Geweke and Tanizaki 2001]. Our work offers insight into the kinds of gains that may be attained by such hybrid approaches.

The paper will proceed as follows. In Section 2, we introduce the independence MH sampler we endorse herein and review Albert and Chib’s auxiliary-variable approach. We also provide a brief and candid analysis of the computational efficiency of the adaptation by [Holmes and Held 2006]. Section 3 presents a simulation study that compares the performance of the proposed sampler to Albert and Chib’s across a wide variety of simulated data sets, while Section 4 describes performance of the samplers across several probit regression examples taken from the literature. We close with a discussion of the results. The R code and data sets used for the simulation study and applications are included in the online supplement.

2. MCMC samplers for Bayesian probit regression

2.1. Albert and Chib Gibbs sampler. In the context of binary probit regression, assume that we have n independent binary random variables, $Y_i \stackrel{\text{ind}}{\sim} \text{Bern}(p_i)$. Each $p_i = \Phi(\mathbf{x}'_i \boldsymbol{\beta})$, where \mathbf{x}'_i is the $1 \times (p+1)$ row vector of p covariates (with intercept) for observation i , $\boldsymbol{\beta}$ is a $(p+1) \times 1$ column vector of regression coefficients, and Φ

is the standard normal cdf. We assume a priori that $\boldsymbol{\beta} \sim \pi$, with π usually chosen to be $\mathcal{N}(b_0, B_0)$. (Here, we adopt the common practice of allowing π to represent both the pdf and cdf, as appropriate.) The likelihood is given as the product of Bernoulli pmfs. By Bayes' rule, the posterior distribution is

$$\begin{aligned} p(\boldsymbol{\beta} | \mathbf{y}) &\propto \pi(\boldsymbol{\beta}) p(\mathbf{y} | \boldsymbol{\beta}) = \pi(\boldsymbol{\beta}) \prod_{i=1}^n p(y_i | \boldsymbol{\beta}, \mathbf{x}_i) \\ &= \pi(\boldsymbol{\beta}) \prod_{i=1}^n \Phi(\mathbf{x}'_i \boldsymbol{\beta})^{y_i} (1 - \Phi(\mathbf{x}'_i \boldsymbol{\beta}))^{1-y_i}. \end{aligned}$$

There is no conjugate prior for $\boldsymbol{\beta}$. To facilitate Gibbs sampling, Albert and Chib introduced a latent variable $\mathbf{Z} = (Z_1, Z_2, \dots, Z_n)$, where $Z_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}'_i \boldsymbol{\beta}, 1)$. Letting $Y_i = 1$ if $Z_i > 0$ and $Y_i = 0$ otherwise ensures that each Y_i is independent and marginally $\text{Bern}(p_i)$, as desired. The joint posterior of \mathbf{Z} and $\boldsymbol{\beta}$ is

$$\begin{aligned} p(\mathbf{Z}, \boldsymbol{\beta} | \mathbf{y}) &\propto \pi(\boldsymbol{\beta}) p(\mathbf{Z} | \boldsymbol{\beta}) p(\mathbf{y} | \mathbf{Z}) = \pi(\boldsymbol{\beta}) \prod_{i=1}^n p(Z_i | \boldsymbol{\beta}, \mathbf{x}_i) p(y_i | Z_i) \\ &= \pi(\boldsymbol{\beta}) \prod_{i=1}^n \phi(Z_i | \mathbf{x}'_i \boldsymbol{\beta}, 1) [\mathbf{1}(y_i = 1) \mathbf{1}(Z_i > 0) + \mathbf{1}(y_i = 0) \mathbf{1}(Z_i \leq 0)], \end{aligned}$$

where $\phi(Z_i | \mathbf{x}'_i \boldsymbol{\beta}, 1)$ is the normal pdf with mean $\mathbf{x}'_i \boldsymbol{\beta}$ and variance 1, evaluated at Z_i . This joint posterior cannot be sampled from directly. However, the full conditional distributions are of nice form. In particular, the full conditional distribution of $\boldsymbol{\beta}$ given \mathbf{Z} is

$$p(\boldsymbol{\beta} | \mathbf{Z}, \mathbf{y}) \propto \pi(\boldsymbol{\beta}) \prod_{i=1}^n \phi(Z_i | \mathbf{x}'_i \boldsymbol{\beta}, 1),$$

which is a multivariate normal distribution when π is normal. The full conditional distribution of \mathbf{Z} given $\boldsymbol{\beta}$ is the product of the marginal full conditional distributions, the i -th of which is $p(Z_i | \boldsymbol{\beta}, \mathbf{y})$, where

$$(Z_i | \boldsymbol{\beta}, \mathbf{y}) \sim \begin{cases} \mathcal{TN}(\mathbf{x}'_i \boldsymbol{\beta}, 1, 0, \infty) & \text{if } y_i = 1, \\ \mathcal{TN}(\mathbf{x}'_i \boldsymbol{\beta}, 1, -\infty, 0) & \text{if } y_i = 0. \end{cases}$$

Here $\mathcal{TN}(\mathbf{x}'_i \boldsymbol{\beta}, 1, a, b)$ is the normal distribution truncated to (a, b) . Once the full conditionals are derived, the Gibbs sampling algorithm is straightforward. First, initialize $\boldsymbol{\beta}^{(0)}$, typically at either the maximum likelihood estimate or posterior mode. Then draw $\mathbf{Z}^{(1)}$ from the full conditional distribution of $\mathbf{Z} | \boldsymbol{\beta}^{(0)}, \mathbf{y}$. Next, draw $\boldsymbol{\beta}^{(1)}$ from the full conditional distribution of $\boldsymbol{\beta} | \mathbf{Z}^{(1)}, \mathbf{y}$. This process is iterated until convergence [Geman and Geman 1984].

2.2. Holmes and Held Gibbs sampler. The Albert and Chib Gibbs sampler can be inefficient because of strong posterior correlation between $\boldsymbol{\beta}$ and \mathbf{Z} . Holmes and

Held [2006] adapted the Albert and Chib sampler by performing a joint update of the regression and auxiliary parameters. This is possible because of the factorization

$$p(\boldsymbol{\beta}, \mathbf{Z} | \mathbf{y}) = p(\mathbf{Z} | \mathbf{y})p(\boldsymbol{\beta} | \mathbf{Z}, \mathbf{y}),$$

where $p(\boldsymbol{\beta} | \mathbf{Z}, \mathbf{y})$ is the full conditional given above and $p(\mathbf{Z} | \mathbf{y})$ is the marginal distribution of \mathbf{Z} with $\boldsymbol{\beta}$ integrated out. Holmes and Held show that $p(\mathbf{Z} | \mathbf{y})$ is a multivariate truncated normal, which, while difficult to sample from directly, does allow for straightforward Gibbs sampling. In particular, $p(Z_i | \mathbf{Z}_{-i}, \mathbf{y})$ has a univariate truncated normal distribution. The authors claim that the additional computational burden of the joint sampling step is small compared with the efficiency gained from having less-correlated draws. In particular, they note that the innermost for-loop requires only minor adjustment to sample $p(Z_i | \mathbf{Z}_{-i}, \mathbf{y})$ instead of $p(Z_i | \boldsymbol{\beta}, \mathbf{Z}_{-i}, \mathbf{y})$ [Holmes and Held 2006, p. 160].

However, the authors do not mention the fact that under Albert and Chib’s setup, sampling from the full conditional distributions $p(Z_i | \boldsymbol{\beta}, \mathbf{Z}_{-i}, \mathbf{y})$ can be vectorized, since the Z_i ’s are conditionally independent (rendering the conditioning on \mathbf{Z}_{-i} unnecessary). Thus, there is no need for the extra for-loop that their joint sampling requires. In a compiled language, this may make little difference; but in an interpreted language like R—the language of choice for many if not most practicing statisticians—the difference is tremendous. In the R simulation presented by [Kapourani 2018], for example, the Albert and Chib Gibbs sampler acquires 10,000 samples in about 3 seconds on a standard laptop computer, whereas the Holmes and Held Gibbs sampler acquires 10,000 samples in about 84 seconds. The effective sample size for Holmes and Held is indeed greater (~ 1000 versus ~ 500), but a 2-fold increase in effective sample size is hardly worth a 28-fold increase in computation time.

There are, of course, compiled-language implementations of Bayesian probit regression. One example of such an implementation of Albert and Chib’s Gibbs sampler is the `MCMCprobit` function within the `MCMCpack` R package [Martin et al. 2011], which interfaces with R but outsources the actual sampling to C++. We suspect that a similar implementation of the Holmes and Held Gibbs sampler would mitigate the computational discrepancies noted above but we are not aware of an R package with this functionality. Because we use R for implementing the proposed method, we compare the proposed method only to the Albert and Chib sampler in terms of efficiency. Both samplers are coded using appropriate vectorization in R only.

2.3. Independent-proposal MH. Holmes and Held [2006] have noted that the Albert and Chib Gibbs sampler can be slow to converge when \mathbf{Z} and $\boldsymbol{\beta}$ are highly correlated. The proposed sampler is intended to overcome this inefficiency. It samples from $p(\boldsymbol{\beta} | \mathbf{y})$ using the Metropolis–Hastings algorithm with independent

$t_5(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ proposals,¹ where $\boldsymbol{\mu}$ is the posterior mode and $\boldsymbol{\Sigma}$ is the inverse Hessian evaluated at $\boldsymbol{\mu}$. The mode can be found with the Newton–Raphson algorithm, which requires the gradient and Hessian of the log posterior. Importantly, these are both available in closed form.

The sampler proceeds as follows. We initialize $\boldsymbol{\beta}^{(0)}$ at $\boldsymbol{\mu}$. Then, we draw a proposal, $\boldsymbol{\beta}^*$ from $t_5(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (with pdf denoted by g), and set $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^*$, with probability α , and $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)}$ with probability $1 - \alpha$, where

$$\alpha = \min \left\{ 1, \frac{p(\boldsymbol{\beta}^* | \mathbf{y})g(\boldsymbol{\beta}^{(t)} | \boldsymbol{\beta}^*)}{p(\boldsymbol{\beta}^{(t)} | \mathbf{y})g(\boldsymbol{\beta}^* | \boldsymbol{\beta}^{(t)})} \right\} = \min \left\{ 1, \frac{p(\boldsymbol{\beta}^* | \mathbf{y})g(\boldsymbol{\beta}^{(t)})}{p(\boldsymbol{\beta}^{(t)} | \mathbf{y})g(\boldsymbol{\beta}^*)} \right\}.$$

Note that the second equality is justified by the fact that the proposals are independent of the current state; hence the conditioning is superfluous. Since the proposed jump, $\boldsymbol{\beta}^*$, is independent of the current state, $\boldsymbol{\beta}^{(t)}$, the average jump size is relatively large. This makes for lower autocorrelation between draws and requires fewer draws to attain a desired effective sample size. It is well known that when the sample size n is large compared to the number of predictors p , the posterior is well approximated by a normal distribution, so an independent-proposal MH algorithm with multivariate t proposals will do well. This paper investigates how well the approximation works when n is medium or small. If, for certain combinations of n and p , the $t_5(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a very poor approximation, performance of the proposed sampler will suffer and Albert and Chib’s method may be preferred. The next two sections provide evidence that the proposed sampler is often better, however.

3. Simulation study

3.1. Settings. We compare the Albert and Chib and independent-proposal MH samplers on a variety of simulated data sets which were created by varying the number of predictors (p) and sample size (n). For a given data set, the prior standard deviation (σ) was also varied. Comparisons were made for all $4^3 = 64$ combinations of $p \in \{2, 5, 10, 20\}$, $n \in \{5p, 10p, 25p, 50p\}$, and $\sigma \in \{0.1, 1, 10, 100\}$. Predictors were taken from a single 1000×20 matrix \mathbf{X} with independent rows. Column vectors were sampled and assigned as follows:

$$\begin{aligned} X_1 &\sim \text{Unif}(-1, 1), & X_6 &= X_5^3, & X_{11} &\sim \text{Unif}(0, 2), & X_{16} &\sim \text{Unif}(0, 6), \\ X_2 &= X_1^2, & X_7 &\sim \text{Unif}(-5, 1), & X_{12} &\sim \text{Unif}\left(-1, \frac{1}{2}\right), & X_{17} &\sim \text{Exp}\left(\frac{1}{2}\right), \\ X_3 &\sim \text{Exp}(5), & X_8 &= X_7^2, & X_{13} &= X_{11}^2, & X_{18} &\sim \text{Exp}(2), \\ X_4 &= X_3^2, & X_9 &\sim \text{Exp}\left(\frac{1}{3}\right), & X_{14} &= X_{12}^2, & X_{19} &\sim \text{Unif}(-1, 1), \\ X_5 &\sim \text{Exp}\left(\frac{1}{5}\right), & X_{10} &= X_9^2, & X_{15} &\sim \mathcal{N}(3, 1), & X_{20} &= X_{19}^3; \end{aligned}$$

¹We chose the t -distribution with five degrees of freedom because it approximates the asymptotic normal posterior distribution but has relatively heavy tails, enabling approximation in the small- n case as well.

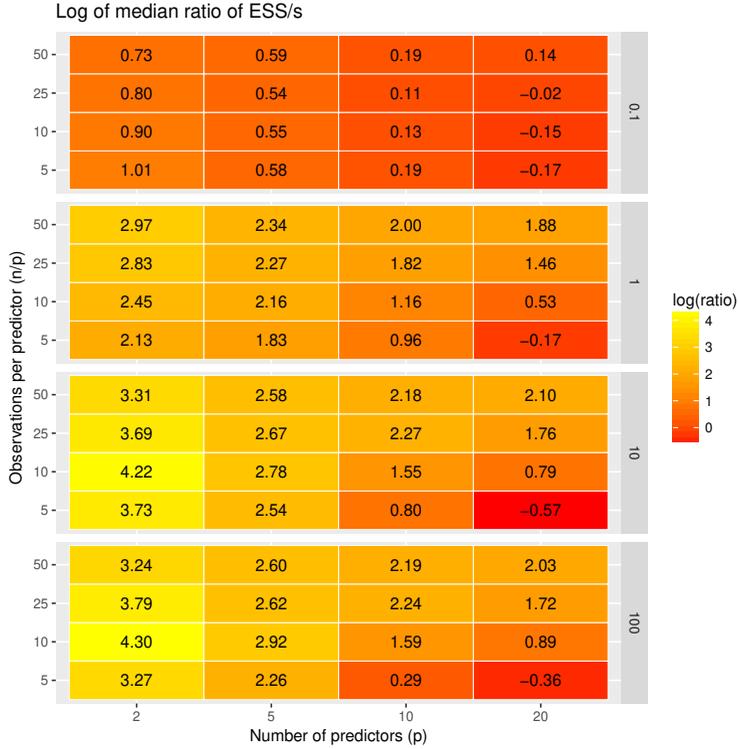


Figure 1. The effective sample size per second of computation time is almost always better for the proposed method than for Albert and Chib Gibbs sampling. The few exceptions correspond to situations with either unrealistically narrow priors or unrealistically few observations per predictor.

and these were then standardized to have mean 0 and variance 1. For a given (p, n) pair, the first p columns and the first n rows of X constituted the predictor matrix, the true parameter vector was the first p elements of the 20-element vector $\beta = (+1, -1, +1, -1, \dots, -1)$, and n binary responses were generated as Bernoulli realizations having expected values $\Phi(X_{n \times p} \beta_{p \times 1})$. MCMC samples were then obtained for each $\sigma \in \{0.1, 1, 10, 100\}$ using the same data with the different priors. Data were generated anew for each of 50 iterations, for a total of $50 \times 4^3 = 3200$ $(n, p, \sigma, \text{iter})$ combinations.

For each $(n, p, \sigma, \text{iter})$ combination, both samplers were initialized at the posterior mode and run for 10,999 iterations, the first 999 of which were discarded as burn-in. The worst-dimension autocorrelation,² the effective sample size (ESS), and

²The worst-dimension autocorrelation is the maximum autocorrelation observed for any univariate β after burn-in.

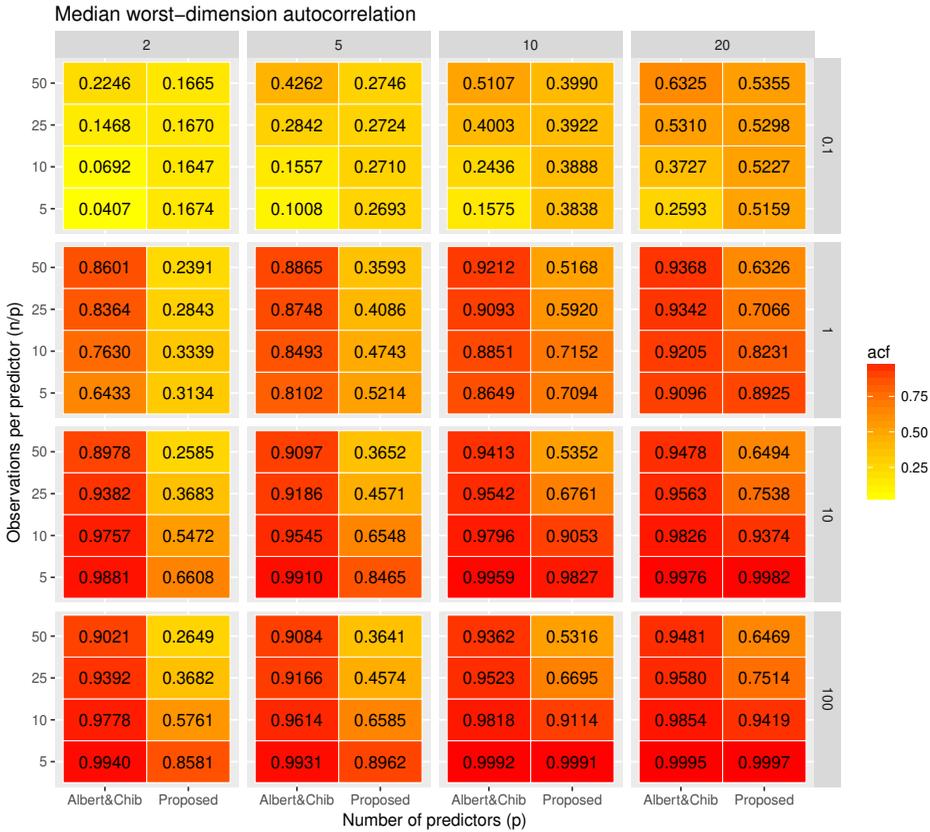


Figure 2. Median worst-dimension lag-1 autocorrelations by sampler.

computation time (in seconds) were recorded for each run. A computation-time-adjusted effective sample size ratio was constructed by dividing the ESS per second of computation time for the independent-proposal MH by the ESS per second of computation time for the Albert and Chib Gibbs sampler. For each (p, n, σ) triplet, the median of these metrics across the 50 iterations was recorded.

3.2. Results. Figure 1 depicts the log of the median value, across 50 iterations, of the computation-time-adjusted ESS ratio. Positive values indicate that the proposed sampler is more efficient, whereas negative values indicate that Albert and Chib is more efficient; values exceeding 2.30 indicate an order-of-magnitude improvement (since $e^{2.30} = 10.0$).

The results reveal that the proposed method outperforms Albert and Chib in most conditions. In particular, in 58 of 64 conditions, the proposed method was more efficient, often by an order of magnitude. In four of the six conditions under which Albert and Chib was better, the improvement was by less than 20%; and all of these

involved small ($\sigma = 1$) or very small ($\sigma = 0.1$) prior standard deviations. Such highly informative priors are in our experience quite rare, as such strong prior knowledge would essentially negate the need to collect data or make posterior inference.

The remaining two conditions, represented by the red cells in the bottom right corners of the $\sigma = 10$ and $\sigma = 100$ subgrids, respectively, reflect situations where sampling was very poor for both methods. In neither of these scenarios was Albert and Chib particularly successful in an absolute sense. Its median worst-dimension autocorrelations were 0.9976 and 0.9995 for $\sigma = 10$ and $\sigma = 100$, respectively (see Figure 2). In short, neither sampler is doing well here, though Albert and Chib is nominally better.

For these two conditions, even if the posterior were well sampled (using a dependent-proposal Metropolis–Hastings algorithm, for example), it is hard to envision it being of much use. These cells correspond to small- n , large- p scenarios where prior knowledge is poor; neither the likelihood nor the prior contains much information, so parameters cannot be estimated reliably. This is why others have cautioned against using probit regression when $n/p < 10$ [Long 1997].

In short, for a wide variety of settings, the proposed method is more efficient than the Albert and Chib Gibbs sampler at producing independent samples from the posterior. The few exceptions correspond to situations with exceptionally narrow priors or exceptionally broad likelihoods, none of which are particularly likely to be encountered in practice.

4. Applications

While the simulation study demonstrated success across a wide range of conditions, the real test is to apply the sampler to actual data. The four applications we consider in this section use previously studied data sets seen as canonical in the probit regression literature. For each application, we compare sampler performance on five outcomes: posterior means of the coefficients, autocorrelation, effective sample size, average jump size (in Euclidean distance), and computation time (in seconds). We also report the computation-time-adjusted ESS ratio, calculated as in the simulation study.

4.1. Prostate cancer. The data for our first application comes from a prostate cancer study on 53 patients and was reported in [Miller et al. 1980]. Chib [1995] used probit regression to predict whether cancer has spread to surrounding lymph nodes using five predictor variables: the patient’s age, the level of serum acid phosphate, positive or negative X-ray results, the size of the tumor, and the pathological grade of the tumor. Results from Albert and Chib’s sampler and the proposed MH sampler are presented in Table 1. The posterior means of the coefficients are almost identical, while the proposed sampler has lower computation time and more than double the effective sample size of the Albert and Chib sampler.

	prostate cancer		Finney data		grade prediction		German health care	
	proposed	Albert and Chib	proposed	Albert and Chib	proposed	Albert and Chib	proposed	Albert and Chib
β_0	1.52	1.53	-6.28	-6.28	-8.41	-8.39	-0.13	-0.13
β_1	-0.04	-0.04	3.10	3.10	1.82	1.82	0.01	0.01
β_2	1.67	1.67	2.58	2.58	0.06	0.06	-0.01	-0.01
β_3	1.29	1.28	—	—	1.58	1.58	-0.01	-0.01
β_4	1.02	1.02	—	—	—	—	-0.15	-0.15
β_5	0.50	0.50	—	—	—	—	0.07	0.07
β_6	—	—	—	—	—	—	0.36	0.36
ACF	0.34	0.60	0.20	0.55	0.36	0.75	0.30	0.41
ESS	45,828	20,113	64,923	28,475	43,798	11,801	2,600	2,013
avg. jump	1.89	1.86	2.18	1.83	2.27	1.75	0.06	0.07
comp. time	2.32	3.86	1.19	2.86	1.17	2.74	49.46	50.13
adj. ESS ratio	3.79		5.48		8.69		1.31	

Table 1. Application results. For each sampler, we report posterior means of β , autocorrelation (ACF), effective sample size (ESS), average jump size, and computation time (in seconds). We compare the two samplers with a computation-time-adjusted ESS ratio; values larger than 1 favor the proposed sampler.

4.2. *Finney data.* In [Albert and Chib 1993], they use data from [Finney 1947] to illustrate the accuracy of their method as the number of simulation iterations increases. As shown in Table 1, the β estimates of the two samplers are identical to the hundredths place. The proposed sampler has nearly double the effective sample size, nearly a third of the autocorrelation, and less than half the computation time.

4.3. *Grade prediction.* The data set for our third application is from an econometric analysis textbook [Greene 2012], an earlier version of which was cited in [Albert and Chib 1993]. The data comes from a study by [Spector and Mazzeo 1980], whose motivation was to measure the effect of a personalized system of instruction on students' grades. As part of their analysis, they modeled grade improvement as predicted by three variables: the student's grade point average, his/her score on an economic literacy test, and whether or not he/she participated in a personalized system of instruction. In Table 1, note the close similarity in the posterior means of the coefficients, despite the small sample size of 32. The efficiency comparison in this application tells a similar story to the previous two, with the proposed sampler outperforming the Albert and Chib sampler in every category.

4.4. *German health care.* The data for our final application is from the same econometric analysis textbook [Greene 2012] as the grade-prediction data. The predictors of age, education, monthly income, marriage status, whether or not children under 16 live in the home, and gender are used to model the binary response of whether or not the subject has visited the doctor in the last three months. The data set contains 27,326 observations on 7,293 German families. Greene uses this model as an example of how slow and inefficient the Albert and Chib method can be. In Table 1 we see that the proposed sampler is only slightly faster; however, the ACF and ESS show that the draws from the proposed sampler are less correlated. Although it took both samplers approximately 50 seconds to run 5,000 iterations, in terms of effective samples, the proposed sampler collected roughly 30% more and thus obtained a similar amount of independent information from the posterior in much less time than Albert and Chib's sampler.

5. Conclusion

Albert and Chib's auxiliary-variable Gibbs sampler [1993] revolutionized Bayesian probit regression and continues to be utilized with wide success in many applications today. As suggested by [Polson et al. 2013], Albert and Chib's sampler is likely the better choice to accommodate more sophisticated modeling frameworks, such as those that involve, for example, dependence among observations. However, many probit regression models used in practice do not require such complexities, and we propose that for this basic modeling context, the independence MH sampler described in this paper should be the new go-to sampler.

We have demonstrated through a simulation study that, in terms of computational efficiency, our proposed sampler outperforms that of Albert and Chib in all but two scenarios, both of which correspond to settings unlikely to be encountered in practice: extremely strong prior knowledge or very diffuse likelihoods. In all other cases, our sampler provides efficiency improvement over the Albert and Chib sampler, and in many cases the scale of improvement is over an order of magnitude.

Beyond simulated settings, we have also demonstrated a striking efficiency advantage provided by our proposed sampler when applied to four real data sets frequently analyzed in probit regression literature. Upon comparing the proposed sampler to Albert and Chib's, we obtain nearly identical coefficient estimates, and our sampler consistently produces higher effective sample sizes in less computation time.

References

- [Albert and Chib 1993] J. H. Albert and S. Chib, "Bayesian analysis of binary and polychotomous response data", *J. Amer. Statist. Assoc.* **88**:422 (1993), 669–679. MR Zbl
- [Amemiya 1981] T. Amemiya, "Qualitative response models: a survey", *J. Econ. Lit.* **19**:4 (1981), 1483–1536.
- [Chib 1995] S. Chib, "Marginal likelihood from the Gibbs output", *J. Amer. Statist. Assoc.* **90**:432 (1995), 1313–1321. MR Zbl
- [Finney 1947] D. J. Finney, "The estimation from individual records of the relationship between dose and quantal response", *Biometrika* **34**:3-4 (1947), 320–334. Zbl
- [Frühwirth-Schnatter and Frühwirth 2007] S. Frühwirth-Schnatter and R. Frühwirth, "Auxiliary mixture sampling with applications to logistic models", *Comput. Statist. Data Anal.* **51**:7 (2007), 3509–3528. MR Zbl
- [Gamerman 1997] D. Gamerman, "Sampling from the posterior distribution in generalized linear mixed models", *Stat. Comput.* **7**:1 (1997), 57–68.
- [Gelfand and Smith 1990] A. E. Gelfand and A. F. M. Smith, "Sampling-based approaches to calculating marginal densities", *J. Amer. Statist. Assoc.* **85**:410 (1990), 398–409. MR
- [Geman and Geman 1984] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images", *IEEE Trans. Pattern Anal. Mach. Intell.* **6**:6 (1984), 721–741. Zbl
- [Geweke and Tanizaki 2001] J. Geweke and H. Tanizaki, "Bayesian estimation of state-space models using the Metropolis–Hastings algorithm within Gibbs sampling", *Comput. Statist. Data Anal.* **37**:2 (2001), 151–170. MR Zbl
- [Greene 2012] W. H. Greene, *Econometric analysis*, 7th ed., Pearson, Upper Saddle River, NJ, 2012.
- [Griffiths et al. 1987] W. E. Griffiths, R. C. Hill, and P. J. Pope, "Small sample properties of probit model estimators", *J. Amer. Statist. Assoc.* **82**:399 (1987), 929–937. MR
- [Holmes and Held 2006] C. C. Holmes and L. Held, "Bayesian auxiliary variable models for binary and multinomial regression", *Bayesian Anal.* **1**:1 (2006), 145–168. MR Zbl
- [Kapourani 2018] A. C. Kapourani, "Gibbs sampling for Bayesian binary probit", online tutorial, 2018, available at <https://rpubs.com/cakapourani/bayesian-binary-probit-model>.
- [Long 1997] J. S. Long, *Regression models for categorical and limited dependent variables*, Advanced quantitative techniques in the social sciences **7**, SAGE, Thousand Oaks, CA, 1997.

- [Martin et al. 2011] A. D. Martin, K. M. Quinna, and J. H. Park, “MCMCpack: Markov chain Monte Carlo in R”, *J. Stat. Softw.* **42** (2011), art. id. 9.
- [Miller et al. 1980] R. G. Miller, B. Efron, B. W. Brown, and L. E. Moses, “Prediction analysis for binary data”, pp. 3–18 in *Biostatistics casebook*, Wiley, New York, 1980.
- [Polson et al. 2013] N. G. Polson, J. G. Scott, and J. Windle, “Bayesian inference for logistic models using Pólya-Gamma latent variables”, *J. Amer. Statist. Assoc.* **108**:504 (2013), 1339–1349. MR Zbl
- [Spector and Mazzeo 1980] L. C. Spector and M. Mazzeo, “Probit analysis and economic education”, *J. Econ. Ed.* **11**:2 (1980), 37–44.
- [Zeger and Karim 1991] S. L. Zeger and M. R. Karim, “Generalized linear models with random effects; a Gibbs sampling approach”, *J. Amer. Statist. Assoc.* **86**:413 (1991), 79–86. MR
- [Zellner and Rossi 1984] A. Zellner and P. E. Rossi, “Bayesian analysis of dichotomous quantal response models”, *J. Econometrics* **25**:3 (1984), 365–393. MR Zbl

Received: 2019-01-28 Accepted: 2019-11-14

scott.thomas.simmons@gmail.com *US Naval Academy, Annapolis, MD, United States*
emcguffe@usna.edu *US Naval Academy, Annapolis, MD, United States*
vanderwe@usna.edu *US Naval Academy, Annapolis, MD, United States*

Characterizing optimal point sets determining one distinct triangle

Hazel N. Brenner, James S. Depret-Guillaume,
Eyvindur A. Palsson and Robert W. Stuckey

(Communicated by Kenneth S. Berenhaut)

We determine the maximum number of points in \mathbb{R}^d which form exactly t distinct triangles, where we restrict ourselves to the case of $t = 1$. We denote this quantity by $F_d(t)$. It is known from the work of Epstein et al. (*Integers* **18** (2018), art. id. A16) that $F_2(1) = 4$. Here we show somewhat surprisingly that $F_3(1) = 4$ and $F_d(1) = d + 1$, whenever $d \geq 3$, and characterize the optimal point configurations. This is an extension of a variant of the distinct distance problem put forward by Erdős and Fishburn (*Discrete Math.* **160**:1-3 (1996), 115–125).

1. Introduction

Erdős [1946] proposed his distinct distance conjecture, which states that any set of n points in the plane will define at least $\Omega(n/\sqrt{\log n})$ distinct distances. Since that time, optimal points sets have been a heavily studied topic within the field of discrete geometry. Guth and Katz [2015] made significant progress towards proving this conjecture when they showed that a set of n points in the plane defined at least $\Omega(n/\log n)$ distinct distances. The analogous problems in dimensions 3 and higher remain open.

Erdős and Fishburn [1996] asked a question related to this: given a positive integer k , what is the maximum number of points which can be embedded in the plane such that precisely k distinct distances are defined, and can all such point configurations be characterized? In their paper, Erdős and Fishburn characterized the optimal configurations for $1 \leq k \leq 4$, and this was extended by Shinohara [2008] for $k = 5$ and by Wei [2012] for $k = 6$. Further, Erdős [1975] conjectured that for sufficiently large values of k , an optimal point configuration exists in the triangular lattice, which continues as an open conjecture. (Figure 1 shows the optimal configurations for k distinct distances in the plane when $2 \leq k \leq 6$.)

MSC2010: primary 52C10; secondary 52C35.

Keywords: one-triangle problem, Erdős problem, optimal configurations, finite point configurations. The work of Palsson was supported in part by Simons Foundation Grant #360560.

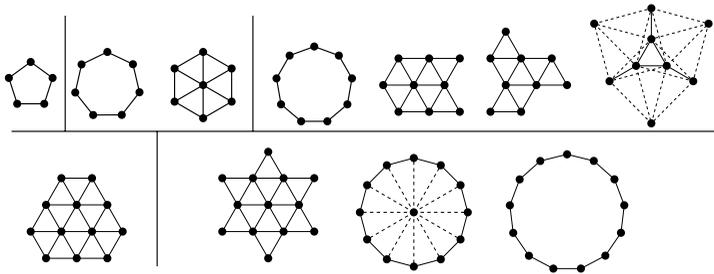


Figure 1. Optimal, or maximal, point set configurations determining exactly k distinct distances in the plane for $2 \leq k \leq 6$ [Brass et al. 2005]. For all k , $2 < k \leq 6$, there exists a configuration in the triangular lattice; Erdős conjectured that this is always true when k is sufficiently large.

Erdős’ distance problem can be extended to consider triangles in place of distances. Since the set of distances generated by a point set can be thought of as being determined by the collection of two-point subsets, we may analogously consider the set of triangles formed by a point set as determined by the collection of three-point subsets. Erdős’ distance problem then becomes: what is the minimum number of distinct triangles formed by a collection of n points in the plane? Hence, the analogue of Erdős and Fishburn’s problem is: given a positive integer t , what is the maximum number of points, n , placed in the plane which define exactly t distinct triangles? Epstein et al. [2018] focused on the latter of these analogues and showed that $n = 4$ for $t = 1$, and $n = 5$ for $t = 2$. Finding maximal point sets in the plane remains an open question for higher values of t . As mentioned above, the higher-dimensional analogues of Erdős’ distance problem are as yet open. Here we concern ourselves with the higher-dimensional analogue of Erdős and Fishburn’s question, rather than with higher values of k . Our main result is the following:

Theorem 1.1. *Suppose $S \subset \mathbb{R}^d$ determines a single distinct triangle T :*

- (1) *If T is equilateral, then S is contained in the set of vertices of a regular d -simplex, and in particular $|S| \leq d + 1$.*
- (2) *If T is not equilateral, then $|S| \leq 4$.*

We can then state the following corollary in the language of optimal point configurations:

Corollary 1.2. *Let $F_d(t)$ denote the maximum number of points which can be placed in \mathbb{R}^d to determine exactly t distinct triangles. Then:*

- (1) *$F_3(1) = 4$ and the only configurations which achieve this are the vertices of a square, a rectangle, or a tetrahedron.*

- (2) $F_d(1) = d + 1$ when $d > 3$ and the only configuration which achieves this is the regular d -simplex.

We also make three observations. First, in \mathbb{R}^d , $d > 3$, the d -simplex is the unique optimal point configuration, yet in dimensions 2 and 3 this is not so. Second, in addition to the above, the 2-simplex fails to be optimal in \mathbb{R}^2 . Third, note that in \mathbb{R}^3 there are two optimal configurations, viz. the tetrahedron (3-simplex) and the rectangle (to include the square), while in \mathbb{R}^2 and \mathbb{R}^d , $d > 3$, there is a single family of solutions (by considering the square to be a special case of the rectangle), and hence, the d -simplex fails to be unique. This transition that happens in \mathbb{R}^3 is surprising and novel. In addition to the above, we have the following notable remark:

Remark 1.3. If $d = 3$, both (1) and (2) of Theorem 1.1 yield optimal configurations. For (1) these configurations are specifically the vertices of the regular tetrahedron, and for (2) they are the vertices of the square, the vertices of a tetrahedron with isosceles faces and the vertices of a tetrahedron with scalene faces. These can be uniquely determined as distance graphs which can be realized in \mathbb{R}^3 in the above ways.

In Section 5 we offer a proof of this remark by way of a construction of said distance graphs and point sets in \mathbb{R}^3 that satisfy them. This remark is particularly interesting as our framework for an upcoming paper in preprint arrives at constructions for optimal configurations determining few distinct triangles by considering the number of distinct distances that can be determined by such configurations [Brenner et al. 2019]. This remark shows that there can exist distinct optimal configurations determining a given number of distinct triangles that determine different numbers of distinct distances. Interestingly, this is not the case for optimal configurations determining two distinct triangles, which may only determine two distinct distances.

2. Definitions and lemmas

We formalize the concepts of a triangle and set out our notation with the following definitions:

Definition 2.1. Given a finite point set $P \subset \mathbb{R}^d$, $d \geq 3$, we say two triples (a, b, c) , $(a', b', c') \in P^3$ are equivalent if there is an isometry mapping one to the other, and we denote this as $(a, b, c) \sim (a', b', c')$.

Definition 2.2. Given a finite point set $P \subset \mathbb{R}^d$, $d \geq 3$, we denote by P_{nc}^3 the set of noncollinear triples $(a, b, c) \in P^3$.

Definition 2.3. Given a finite point set $P \subset \mathbb{R}^2$, we define the set of distinct triangles determined by P as

$$T(P) := P_{\text{nc}}^3 / \sim. \quad (2-1)$$

In this paper when we discuss and count the number of distinct triangles of a finite point set $P \in \mathbb{R}^d$ we are precisely working with the set $T(P)$. Note that this excludes degenerate triangles where all three points lie on a line.

Definition 2.4. Let p and q be points in \mathbb{R}^d for $d \geq 1$. We denote the Euclidean distance between p and q by $d(p, q)$.

Theorem 1.1(1) is a direct consequence of the following lemma:

Lemma 2.5. *Let S be a set of points in \mathbb{R}^d , $d \geq 3$, which defines a single distinct equilateral triangle. Then S has at most $d + 1$ points.*

3. Proof of Theorem 1.1

As stated previously, Theorem 1.1(1) follows directly from Lemma 2.5, so we will omit proof in this section in favor of proving Lemma 2.5 in Section 4. Then, to prove Theorem 1.1(2), we will consider the case where T is isosceles and the case where T is scalene separately. In both cases, we assume towards a contradiction that there exists a point set S containing five points determining such a triangle. Our argument will be made purely on distance graphs and will thus not depend on dimension.

Proof of Theorem 1.1(2). Assume towards a contradiction that there exists a point set S containing five points which determines one distinct nonequilateral triangle. For convenience, we then split into cases, dealing with scalene and isosceles triangles separately.

Scalene: Fix an arbitrary point \mathcal{O} in S and consider the distances from \mathcal{O} to the remaining four points. Note that clearly a point set determining exactly one distinct, scalene triangle determines only three distinct distances. So, by the pigeonhole principle, two of the distances from \mathcal{O} to the remaining points are equal. Without loss of generality, say that the repeated distance is d_1 and specifically $\mathcal{O}A = \mathcal{O}B = d_1$. Then, clearly $\triangle \mathcal{O}AB$ is an isosceles triangle, but we assumed that the only distinct triangle determined by this point set is scalene, so we have the desired contradiction.

Isosceles: Let d_1 denote the repeated edge length of T and d_2 the remaining edge length. Similarly to the above, fix an arbitrary point \mathcal{O} and consider the distances from \mathcal{O} to the remaining points. Clearly, by the same argument as the above, if any two of these distances were d_2 , an isosceles triangle with repeated edge length d_2 would be determined, which would be a contradiction. So, assume that at least three of the four distances are d_1 and label the three points determining them A , B and C . Then consider each of the triangles consisting of two of the points and \mathcal{O} , e.g., $\triangle \mathcal{O}AB$. Clearly this must be congruent to T , and $\mathcal{O}A = \mathcal{O}B = d_1$, so we must have $AB = d_2$. The same holds for all such triangles, so we have $AB = BC = AC = d_2$. But, again, we assumed that the only triangle determined by S was isosceles, so this equilateral triangle of edge length d_2 yields a contradiction. \square

4. Proofs of lemmas

It is well known in the literature that Lemma 2.5 holds, and that in \mathbb{R}^d , $d \geq 2$, a set of points determining a single distinct distance has at most $d + 1$ points. However, in the interest of completeness, we include here a proof of the lemma by induction on the dimension d :

Proof of Lemma 2.5. Base case ($d = 3$): Let $S = \{A_1, A_2, A_3, A_4, A_5\} \subset \mathbb{R}^d = \mathbb{R}^3$ be a point set containing $d + 2 = 5$ points, which defines a single distinct equilateral triangle, call it T . Thus, the triangle $\triangle A_1 A_2 A_3$ must form the triangle T . Define e to be the edge length of T , and let P denote the plane defined by $\{A_1, A_2, A_3\}$.

Since S defines an equilateral triangle, it follows that A_4 and A_5 must be equidistant from $\{A_1, A_2, A_3\}$ and lie upon a line normal to P , which goes through a point $p \in P$, where p is equidistant to $\{A_1, A_2, A_3\}$; p is called the circumcenter of the equilateral triangle $\triangle A_1 A_2 A_3$.

Since p is the circumcenter of the equilateral triangle $\triangle A_1 A_2 A_3$, it follows that

$$d(A_1, p) = d(A_2, p) = d(A_3, p) = \frac{\sqrt{3}}{3}e.$$

Since S defines only the triangle T , it follows that $d(A_4, A_5) = e$. Since A_4, A_5 , and p lie upon the same line, it follows that

$$d(A_4, p) + d(p, A_5) = d(A_4, A_5).$$

Since A_4 and A_5 are equidistant from $\{A_1, A_2, A_3\}$, they are also equidistant from the plane P , and hence, $d(A_4, p) = d(p, A_5) = \frac{1}{2}e$. Applying the Pythagorean theorem we obtain

$$d(A_4, p)^2 + d(A_1, p)^2 = d(A_4, A_1)^2,$$

which yields

$$\left(\frac{1}{2}e\right)^2 + \left(\frac{\sqrt{3}}{3}e\right)^2 = (e)^2.$$

Thus, we obtain

$$\frac{1}{4}e^2 + \frac{1}{3}e^2 = e^2,$$

which implies that $\frac{7}{12} = 1$, a clear contradiction. We note that the vertices of a 3-simplex, the regular tetrahedron, give a configuration in \mathbb{R}^3 which defines a single distinct equilateral triangle and has $3 + 1 = 4$ points. Therefore, a set S defining a single distinct equilateral triangle in \mathbb{R}^3 can have at most $3 + 1 = 4$ points.

Inductive assumption: Suppose that for dimensions $n < d$, a point set S defining a single distinct equilateral triangle can have at most $n + 1$ points.

Inductive step: Now let $S = \{A_1, \dots, A_d, A_{d+1}, A_{d+2}\} \subset \mathbb{R}^d$ be a point set containing $d + 2$ points, which defines a single distinct equilateral triangle, call it T . Thus, the points $\{A_1, \dots, A_d\}$ must be such that any triplet forms the triangle T ,

and so they must form a $(d-1)$ -simplex (this by our inductive assumption). Call this $(d-1)$ -simplex $\triangle A_1 \cdots A_d$. Let e be the edge length of T , and let P be the $(d-1)$ -dimensional hyperplane defined by $\{A_1, \dots, A_d\}$.

Since S defines an equilateral triangle, it follows that A_{d+1} and A_{d+2} must be equidistant to $\{A_1, \dots, A_d\}$ and lie upon a line normal to P , which passes through a point $p \in P$ (viz. the circumcenter of $\triangle A_1 \cdots A_d$), where p is equidistant from $\{A_1, \dots, A_d\}$.

From this point, we can follow the same construction as in the base case, using the points A_1, p, A_{d+1} , and A_{d+2} to arrive at a contradiction. We note here that the d -simplex gives a configuration in \mathbb{R}^d which defines a single distinct equilateral triangle and has $d+1$ points. Thus, a set S in \mathbb{R}^d defining a single distinct equilateral triangle can have at most $d+1$ points, as desired. \square

5. Constructions for Remark 1.3

Clearly the regular tetrahedron is an optimal configuration (of four points) determining one distinct triangle in three dimensions, and it is the unique configuration satisfying Theorem 1.1(1). Then, in the following, we will characterize the distance graphs of the four-point sets satisfying Theorem 1.1(2) (and thus optimal in three dimensions). We then provide constructions of point sets in \mathbb{R}^3 satisfying these distance graphs. For convenience, we will again consider the isosceles and scalene cases separately. Assume in all that follows that S is a set of four points determining one distinct triangle of the respective geometry.

Isosceles: Let d_1 be the repeated edge length of T . Following the framework used in the main proof, fix a point \mathcal{O} in S such that $\mathcal{O}A = d_1$ and $\mathcal{O}B = d_2$ for some A and B in S . Then, notice that given the remaining point C in S , we have $\triangle \mathcal{O}AB \simeq \triangle \mathcal{O}CB \simeq T$, so specifically $\mathcal{O}C = BC = AB = d_1$, and similarly, we must have $AC = d_2$.

Notice that any four noncoplanar points form the vertices of a tetrahedron (if the points are coplanar, this distance graph is clearly uniquely realized by the vertices of the square). And, by the above, a tetrahedron $ABCD$ satisfying the above distance graph must have $AB = d_2$ and $CD = d_2$, with the remaining edges of length d_1 . To construct such a tetrahedron, consider taking points $P = (\frac{1}{2}d_2, 0, 0)$, $Q = (-\frac{1}{2}d_2, 0, 0)$, $R = (0, \frac{1}{2}d_2, 0)$ and $S = (0, -\frac{1}{2}d_2, 0)$. Note that for $d_1 = \frac{\sqrt{2}}{2}d_2$, $PQRS$ satisfies the above distance graph, although it is planar. Then, by arbitrarily translating R and S by the same distance along the z -axis, we can construct any tetrahedron satisfying this distance graph (clearly no tetrahedron with d_1 not satisfying this inequality may exist).

Scalene: Let A, B and C determine $\triangle ABC \simeq T$ in S . Without loss of generality let $AB = d_1$, $BC = d_2$ and $AC = d_3$. Then, notice that each point may determine

each distance exactly once (otherwise they would determine an isosceles triangle, producing a contradiction similar to that used in the main proof). Then, since each point A , B and C determines exactly two of the distances, we may simply fill in that $AD = d_2$, $BD = d_3$ and $CD = d_1$. It is easy to verify that this distance graph determines only one distinct triangle.

Similarly to the isosceles case, we observe that if all four points are coplanar, this distance graph uniquely determines a rectangle. Supposing instead that four points A , B , C and D satisfy the above distance graph and are noncoplanar (thus form a tetrahedron), we clearly must then have that, up to relabeling, $AB = CD = d_1$, $AC = BD = d_2$ and $AD = BC = d_3$. Namely, “opposite” pairs of edges of the tetrahedron are congruent. To construct all such tetrahedra, consider fixing points $P = (\frac{1}{2}d_1, 0, 0)$ and $Q = (-\frac{1}{2}d_1, 0, 0)$. Then, fix $R' = (0, \frac{1}{2}d_1, z')$ and $S' = (0, -\frac{1}{2}d_1, z')$ for arbitrary z' . Consider the unique circle lying in the plane $z = z'$ passing through R' and S' . Then choose R and S as the endpoints of any diameter of this circle such that $R, S \neq R', S'$ and P, Q, R and S are not all coplanar. Clearly, then, $PQRS$ satisfies the above distance graph.

Acknowledgements

We would like to thank an anonymous referee for suggesting the inclusion of Theorem 1.1 in its current form. Their proposed restructuring around this result has significantly improved the paper.

References

- [Brass et al. 2005] P. Brass, W. Moser, and J. Pach, *Research problems in discrete geometry*, Springer, 2005. MR Zbl
- [Brenner et al. 2019] H. N. Brenner, J. S. Depret-Guillaume, E. A. Palsson, and S. Senger, “Uniqueness of optimal point sets determining two distinct triangles”, preprint, 2019. arXiv
- [Epstein et al. 2018] A. Epstein, A. Lott, S. J. Miller, and E. A. Palsson, “Optimal point sets determining few distinct triangles”, *Integers* **18** (2018), art. id. A16. MR Zbl
- [Erdős 1946] P. Erdős, “On sets of distances of n points”, *Amer. Math. Monthly* **53** (1946), 248–250. MR Zbl
- [Erdős 1975] P. Erdős, “On some problems of elementary and combinatorial geometry”, *Ann. Mat. Pura Appl.* (4) **103** (1975), 99–108. MR Zbl
- [Erdős and Fishburn 1996] P. Erdős and P. Fishburn, “Maximum planar sets that determine k distances”, *Discrete Math.* **160**:1-3 (1996), 115–125. MR Zbl
- [Guth and Katz 2015] L. Guth and N. H. Katz, “On the Erdős distinct distances problem in the plane”, *Ann. of Math.* (2) **181**:1 (2015), 155–190. MR Zbl
- [Shinohara 2008] M. Shinohara, “Uniqueness of maximum planar five-distance sets”, *Discrete Math.* **308**:14 (2008), 3048–3055. MR Zbl
- [Wei 2012] X. Wei, “A proof of Erdős–Fishburn’s conjecture for $g(6) = 13$ ”, *Electron. J. Combin.* **19**:4 (2012), art. id. P38. MR Zbl

Received: 2019-02-12 Revised: 2019-09-29 Accepted: 2019-11-11

hazalbrenner@vt.edu *Department of Mathematics, Virginia Tech, Blacksburg, VA,
United States*

jdg@vt.edu *Department of Mathematics, Virginia Tech, Blacksburg, VA,
United States*

palsson@vt.edu *Department of Mathematics, Virginia Tech, Blacksburg, VA,
United States*

rstucke1@kent.edu *Department of Mathematics, Virginia Tech, Blacksburg, VA,
United States*

Current address: *Department of Mathematical Sciences, Kent State University,
Kent, OH, United States*

Solutions of periodic boundary value problems

R. Aadith, Paras Gupta and Jagan Mohan Jonnalagadda

(Communicated by Hari Mohan Srivastava)

We deal with a resonant boundary value problem involving a second-order differential equation with periodic boundary conditions. First, we modify the problem at resonance and consider an equivalent nonresonant boundary value problem. Next, we obtain sufficient conditions for the existence of solutions of the modified boundary value problem, using fixed-point theory. Consequently, these conditions suffice for the existence of solutions of the original boundary value problem. We demonstrate the applicability of established results through examples.

1. Introduction

Researchers have been developing a variety of new methods to establish sufficient conditions on the existence and uniqueness of solutions for boundary value problems (BVPs) at resonance associated with ordinary differential equations. Recently, Han [2007] modified the BVP at resonance and considered a regular BVP (a method referred to as a shift argument) in order to apply a suitable fixed-point theorem. Infante et al. [2016] provided a thorough study of BVPs related to the Neumann boundary conditions using the shift argument. In line with this approach, Almansour and Eloë [2015], Al Mosa and Eloë [2016], and Garcia and Neugebauer [2019] have employed standard fixed-point theorems as well as the monotone method coupled with a method of lower and upper solutions to obtain sufficient conditions on the existence of solutions for nonlinear BVPs at resonance. In continuation to these works, Aljedani and Eloë [2018] and Alshammari et al. [2019] have also developed a quasi-linearization algorithm and constructed sequences of approximate solutions that converge monotonically and quadratically to the unique solution of the resonant BVP.

In this article, we consider the two-point periodic BVP

$$\begin{cases} u'' = f(t, u, u'), & 0 < t < T, \\ u(0) = u(T), & u'(0) = u'(T). \end{cases} \quad (1-1)$$

MSC2010: primary 34B15; secondary 34B27.

Keywords: boundary value problem, resonance, shift, Green's function, fixed point, existence.

Assume

$$f : [0, T] \times \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R} \text{ is continuous in each of its variables.} \quad (\text{H1})$$

The BVP (1-1) is at resonance because the corresponding homogeneous problem

$$\begin{cases} u'' = 0, & 0 < t < T, \\ u(0) = u(T), & u'(0) = u'(T) \end{cases} \quad (\text{1-2})$$

has the nontrivial constant solution. So, we modify (1-1) by employing a shift argument and construct an equivalent nonresonant BVP in order to apply a suitable fixed-point theorem—the Schauder fixed-point theorem or the Leray–Schauder nonlinear alternative. For convenience, we state these theorems as follows:

Theorem 1.1 ([Agarwal et al. 2001] Schauder fixed-point theorem). *If \mathcal{M} is a closed bounded convex subset of a Banach space \mathcal{B} and $A : \mathcal{M} \rightarrow \mathcal{M}$ is completely continuous, then A has a fixed point in \mathcal{M} .*

Theorem 1.2 ([Agarwal et al. 2001] Leray–Schauder nonlinear alternative). *Let V be a closed and bounded subset of a Banach space \mathcal{B} , U be an open subset of V and $0 \in U$. Suppose $A : \bar{U} \rightarrow V$ is a continuous and compact operator. Then, either*

- (1) *A has a fixed point in \bar{U} , or*
- (2) *there exists a point $u \in \partial U$ such that $u = \lambda Au$ for some $\lambda \in (0, 1)$, where ∂U denotes the boundary of U in V .*

2. Existence of solutions of (1-1) using a positive shift argument

Let β be a real number and consider the equivalent BVP

$$\begin{cases} u'' + \beta^2 u = f(t, u, u') + \beta^2 u = g(t, u, u'), & 0 < t < T, \\ u(0) = u(T), & u'(0) = u'(T). \end{cases} \quad (\text{2-1})$$

The BVP (2-1) is not at resonance because the unique solution of the corresponding homogeneous problem

$$\begin{cases} u'' + \beta^2 u = 0, & 0 < t < T, \\ u(0) = u(T), & u'(0) = u'(T) \end{cases} \quad (\text{2-2})$$

is $u \equiv 0$. Assume

$$0 < \beta < \frac{\pi}{T}. \quad (\text{H2})$$

In Lemmas 2.1 and 2.2, we construct the Green's function associated with (2-2) and state its properties. We omit the proofs of these lemmas as they are trivial.

Lemma 2.1. *Assume $k \in C[0, T]$. Then, the linear boundary value problem*

$$\begin{cases} u'' + \beta^2 u = k(t), & 0 < t < T, \\ u(0) = u(T), & u'(0) = u'(T) \end{cases} \quad (\text{2-3})$$

has a unique solution

$$u(t) = \int_0^T G(t, s)k(s) ds, \quad 0 \leq t \leq T, \tag{2-4}$$

where

$$G(t, s) = \frac{1}{2\beta \sin(\frac{1}{2}\beta T)} \begin{cases} \cos \beta(t - s - \frac{1}{2}T), & 0 \leq s \leq t \leq T, \\ \cos \beta(t - s + \frac{1}{2}T), & 0 \leq t \leq s \leq T \end{cases} \tag{2-5}$$

is the Green's function corresponding to (2-2).

Lemma 2.2. *G satisfies the following properties:*

- (1) $G \in C[0, T] \times C[0, T]$.
- (2) $G(t, s) > 0$ for all $(t, s) \in [0, T] \times [0, T]$.
- (3) $\max_{(t,s) \in [0,T] \times [0,T]} G(t, s) = 1/(2\beta \sin(\frac{1}{2}\beta T))$.
- (4) $\max_{(t,s) \in [0,T] \times [0,T]} \left| \frac{\partial}{\partial t} G(t, s) \right| = \frac{1}{2}$.
- (5) $\max_{t \in [0,T]} \int_0^T G(t, s) ds = 1/\beta^2$.
- (6) $\max_{t \in [0,T]} \int_0^T \left| \frac{\partial}{\partial t} G(t, s) \right| ds = (2/\beta) \tan(\frac{1}{4}\beta T)$.

Theorem 2.3. *Assume (H1) and (H2) hold. Then, u is a solution of (2-1) if and only if $u \in C'[0, 1]$ and*

$$u(t) = \int_0^T G(t, s)g(s, u(s), u'(s)) ds, \quad 0 \leq t \leq T. \tag{2-6}$$

Note that u is a solution of (2-1) if and only if it is a solution of (1-1).

Let $\mathcal{B} = C'[0, 1]$ be the Banach space of functions whose first derivatives are continuous on $[0, 1]$ endowed with the norm

$$\|u\| = \max\{\|u\|_0, \|u'\|_0\},$$

where

$$\|u\|_0 = \max_{t \in [0,1]} |u(t)| \quad \text{and} \quad \|u'\|_0 = \max_{t \in [0,1]} |u'(t)|.$$

Define the operator $A : \mathcal{B} \rightarrow \mathcal{B}$ by

$$(Au)(t) = \int_0^T G(t, s)g(s, u(s), u'(s)) ds, \quad 0 \leq t \leq T. \tag{2-7}$$

Clearly, u is a fixed point of A if and only if u is a solution of (2-1).

Theorem 2.4. *Assume (H1) and (H2) hold and $g(t, u, u') = f(t, u, u') + \beta^2 u$ is bounded on $[0, T] \times \mathbb{R} \times \mathbb{R}$. Then, there exists a solution of (1-1).*

Proof. Define

$$M = \sup\{|g(t, u, u')| : (t, u, u') \in [0, T] \times \mathbb{R} \times \mathbb{R}\}$$

and

$$\mathcal{M} = \{u \in \mathcal{B} : \|u\| \leq LM\},$$

where

$$L = \max\left\{\frac{1}{\beta^2}, \frac{2}{\beta} \tan\left(\frac{\beta T}{4}\right)\right\} > 0. \quad (2-8)$$

Clearly, \mathcal{M} is a closed bounded convex subset of \mathcal{B} . We claim that $A : \mathcal{M} \rightarrow \mathcal{M}$. To see this, let $u \in \mathcal{M}$ and consider

$$\begin{aligned} |(Au)(t)| &= \left| \int_0^T G(t, s)g(s, u(s), u'(s)) ds \right| \\ &\leq \int_0^T G(t, s)|g(s, u(s), u'(s))| ds \\ &\leq M \int_0^T G(t, s) ds \leq \frac{M}{\beta^2}, \end{aligned}$$

implying

$$\|(Au)\|_0 \leq \frac{M}{\beta^2}.$$

Also, consider

$$\begin{aligned} |(Au)'(t)| &= \left| \int_0^T \frac{\partial}{\partial t} G(t, s)g(s, u(s), u'(s)) ds \right| \\ &\leq \int_0^T \left| \frac{\partial}{\partial t} G(t, s) \right| |g(s, u(s), u'(s))| ds \\ &\leq M \int_0^T \left| \frac{\partial}{\partial t} G(t, s) \right| ds \leq \frac{2M}{\beta} \tan\left(\frac{\beta T}{4}\right), \end{aligned}$$

implying

$$\|(Au)'\|_0 \leq \frac{2M}{\beta} \tan\left(\frac{\beta T}{4}\right).$$

Thus, we have

$$\|(Au)\| = \max\{\|(Au)\|_0, \|(Au)'\|_0\} \leq LM.$$

Consequently, $A : \mathcal{M} \rightarrow \mathcal{M}$. A standard application of the Arzelà–Ascoli theorem gives us that A is completely continuous. So, it follows by Theorem 1.1 that the operator A has a fixed point in \mathcal{M} . \square

Theorem 2.5. Assume (H1) and (H2) hold. Assume there exists a continuous function $\phi : [0, T] \rightarrow \mathbb{R}^+$ and a nondecreasing function $\psi : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ such that

$$|g(t, u, u')| \leq \phi(t)\psi(\|u\|), \quad (t, u, u') \in [0, T] \times \mathbb{R} \times \mathbb{R}$$

and

$$\|\phi\|\psi(L) \leq 1. \tag{2-9}$$

Then, there exists a solution of (1-1).

Proof. Define

$$U = \{u \in \mathcal{B} : \|u\| < L\}.$$

Clearly, U is an open subset of \mathcal{B} with $0 \in U$ and $A : \bar{U} \rightarrow \mathcal{B}$. Now, suppose there exist a $u \in \partial U$ and $\lambda \in (0, 1)$ such that

$$u(t) = \lambda(Au)(t), \quad 0 \leq t \leq T. \tag{2-10}$$

Consider

$$\begin{aligned} |(Au)(t)| &= \left| \int_0^T G(t, s)g(s, u(s), u'(s)) ds \right| \\ &\leq \int_0^T G(t, s)|g(s, u(s), u'(s))| ds \\ &\leq \int_0^T G(t, s)\phi(s)\psi(\|u\|) ds \\ &\leq \|\phi\|\psi(\|u\|) \int_0^T G(t, s) ds \leq \frac{\|\phi\|\psi(L)}{\beta^2}, \end{aligned}$$

implying

$$\|(Au)\|_0 \leq \frac{\|\phi\|\psi(L)}{\beta^2}.$$

Also, consider

$$\begin{aligned} \|(Au)'(t)| &= \left| \int_0^T \frac{\partial}{\partial t} G(t, s)g(s, u(s), u'(s)) ds \right| \\ &\leq \int_0^T \left| \frac{\partial}{\partial t} G(t, s) \right| |g(s, u(s), u'(s))| ds \\ &\leq \int_0^T \left| \frac{\partial}{\partial t} G(t, s) \right| \phi(s)\psi(\|u\|) ds \\ &\leq \|\phi\|\psi(\|u\|) \int_0^T \left| \frac{\partial}{\partial t} G(t, s) \right| ds \leq \frac{2\|\phi\|\psi(L)}{\beta} \tan\left(\frac{\beta T}{4}\right), \end{aligned}$$

implying

$$\|(Au)'\|_0 \leq \frac{2\|\phi\|\psi(L)}{\beta} \tan\left(\frac{\beta T}{4}\right).$$

Thus, we have

$$\|(Au)\| = \max\{\|(Au)\|_0, \|(Au)'\|_0\} \leq \|\phi\|\psi(L)L. \tag{2-11}$$

Further, it follows from (2-10) and (2-11) that

$$\|\phi\|\psi(L) > 1.$$

This is a contradiction to (2-9). Thus, by Theorem 1.2, the operator A has a fixed point in U . □

3. Existence of solutions of (1-1) using a negative shift argument

Let β be a real number and consider the equivalent BVP

$$\begin{cases} u'' - \beta^2 u = f(t, u, u') - \beta^2 u = h(t, u, u'), & 0 < t < T, \\ u(0) = u(T), \quad u'(0) = u'(T). \end{cases} \tag{3-1}$$

The BVP (3-1) is not at resonance because the unique solution of the corresponding homogeneous problem

$$\begin{cases} u'' - \beta^2 u = 0, & 0 < t < T, \\ u(0) = u(T), \quad u'(0) = u'(T) \end{cases} \tag{3-2}$$

is $u \equiv 0$. In Lemmas 3.1 and 3.2, we construct the Green’s function associated with (3-2) and state its properties. We omit the proofs of these lemmas as they are trivial.

Lemma 3.1. *Assume $l \in C[0, T]$. Then, the linear boundary value problem*

$$\begin{cases} u'' - \beta^2 u = l(t), & 0 < t < T, \\ u(0) = u(T), \quad u'(0) = u'(T) \end{cases} \tag{3-3}$$

has a unique solution

$$u(t) = \int_0^T H(t, s)l(s) ds, \quad 0 \leq t \leq T, \tag{3-4}$$

where

$$H(t, s) = -\frac{1}{2\beta \sinh(\frac{1}{2}\beta T)} \begin{cases} \cosh \beta(t - s - \frac{1}{2}T), & 0 \leq s \leq t \leq T, \\ \cosh \beta(t - s + \frac{1}{2}T), & 0 \leq t \leq s \leq T \end{cases} \tag{3-5}$$

is the Green’s function corresponding to (3-2).

Lemma 3.2. *G satisfies the following properties:*

- (1) $H \in C[0, T] \times C[0, T]$.
- (2) $H(t, s) < 0$ for all $(t, s) \in [0, T] \times [0, T]$.
- (3) $\max_{(t,s) \in [0, T] \times [0, T]} |H(t, s)| = \coth(\frac{1}{2}\beta T)/(2\beta)$.
- (4) $\max_{(t,s) \in [0, T] \times [0, T]} |\frac{\partial}{\partial t} H(t, s)| = \frac{1}{2}$.

$$(5) \max_{t \in [0, T]} \int_0^T |H(t, s)| ds = 1/\beta^2.$$

$$(6) \max_{t \in [0, T]} \int_0^T \left| \frac{\partial}{\partial t} H(t, s) \right| ds = (2/\beta) \tanh\left(\frac{1}{4}\beta T\right).$$

Theorem 3.3. *Assume (H1) and (H2) hold. Then, u is a solution of (3-1) if and only if $u \in C'[0, 1]$ and*

$$u(t) = \int_0^T H(t, s)h(s, u(s), u'(s)) ds, \quad 0 \leq t \leq T. \tag{3-6}$$

Note that u is a solution of (3-1) if and only if it is a solution of (1-1).

Define the operator $B : \mathcal{B} \rightarrow \mathcal{B}$ by

$$(Bu)(t) = \int_0^T H(t, s)h(s, u(s), u'(s)) ds, \quad 0 \leq t \leq T. \tag{3-7}$$

Clearly, u is a fixed point of B if and only if u is a solution of (3-1).

Theorem 3.4. *Assume (H1), (H2) hold and $h(t, u, u') = f(t, u, u') - \beta^2 u$ is bounded on $[0, T] \times \mathbb{R} \times \mathbb{R}$. Then, there exists a solution of (1-1).*

Proof. Define

$$N = \sup\{|h(t, u, u')| : (t, u, u') \in [0, T] \times \mathbb{R} \times \mathbb{R}\}$$

and

$$\mathcal{N} = \{u \in \mathcal{B} : \|u\| \leq N\Lambda\},$$

where

$$\Lambda = \max\left\{\frac{1}{\beta^2}, \frac{2}{\beta} \tanh\left(\frac{\beta T}{4}\right)\right\} > 0. \tag{3-8}$$

Clearly, \mathcal{N} is a closed bounded convex subset of \mathcal{B} . Proceeding in the same way as in Theorem 2.4, we obtain

$$\|(Bu)\|_0 \leq \frac{N}{\beta^2}$$

and

$$\|(Bu)'\|_0 \leq \frac{2N}{\beta} \tanh\left(\frac{\beta T}{4}\right).$$

Thus, we obtain

$$\|(Bu)\| = \max\{\|(Bu)\|_0, \|(Bu)'\|_0\} \leq N\Lambda,$$

implying that $B : \mathcal{N} \rightarrow \mathcal{N}$. A standard application of the Arzelà–Ascoli theorem gives us that B is completely continuous. So, it follows by Theorem 1.1 that the operator B has a fixed point in \mathcal{N} . □

Theorem 3.5. Assume (H1) and (H2) hold. Assume there exists a continuous function $\phi : [0, T] \rightarrow \mathbb{R}^+$ and a nondecreasing function $\psi : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ such that

$$|h(t, u, u')| \leq \phi(t)\psi(\|u\|), \quad (t, u, u') \in [0, T] \times \mathbb{R} \times \mathbb{R},$$

and

$$\|\phi\|\psi(\Lambda) \leq 1. \quad (3-9)$$

Then, there exists a solution of (1-1).

Proof. Define

$$U = \{u \in \mathcal{B} : \|u\| < \Lambda\}.$$

Clearly, U is an open subset of \mathcal{B} with $0 \in U$ and $B : \bar{U} \rightarrow \mathcal{B}$. Now, suppose there exist a $u \in \partial U$ and $\lambda \in (0, 1)$ such that

$$u(t) = \lambda(Bu)(t), \quad 0 \leq t \leq T. \quad (3-10)$$

Proceeding in the same way as in Theorem 2.5, we obtain

$$\|(Bu)\|_0 \leq \frac{\|\phi\|\psi(\Lambda)}{\beta^2}$$

and

$$\|(Bu)'\|_0 \leq \frac{2\|\phi\|\psi(\Lambda)}{\beta} \tanh\left(\frac{\beta T}{4}\right).$$

Thus, we have

$$\|(Bu)\| = \max\{\|(Bu)\|_0, \|(Bu)'\|_0\} \leq \|\phi\|\psi(\Lambda)\Lambda. \quad (3-11)$$

Using (3-10) and (3-11), we get

$$\|\phi\|\psi(\Lambda) > 1.$$

This is a contradiction to (3-9). So, it follows by Theorem 1.2 that the operator B has a fixed point in U . \square

4. Examples

Example 1. Consider the two-point periodic BVP

$$\begin{cases} u'' = \sin t - \frac{1}{16}u, & 0 < t < 2\pi, \\ u(0) = u(2\pi), & u'(0) = u'(2\pi). \end{cases} \quad (4-1)$$

Take $\beta = \frac{1}{4} \in (0, \frac{1}{2})$. Here $T = 2\pi$ and $f(t, u, u') = \sin t - \frac{1}{16}u$ such that $g(t, u, u') = f(t, u, u') + \beta^2 u = \sin t$ is bounded on $[0, 2\pi] \times \mathbb{R} \times \mathbb{R}$. Then, by Theorem 2.4, the BVP (4-1) has a solution.

Example 2. Consider the two-point periodic BVP

$$\begin{cases} u'' = \sin t + \frac{1}{16}u, & 0 < t < 2\pi, \\ u(0) = u(2\pi), & u'(0) = u'(2\pi). \end{cases} \quad (4-2)$$

Take $\beta = \frac{1}{4} \in (0, \frac{1}{2})$. Here $T = 2\pi$ and $f(t, u, u') = \sin t + \frac{1}{16}u$ such that $h(t, u, u') = f(t, u, u') - \beta^2 u = \sin t$ is bounded on $[0, 2\pi] \times \mathbb{R} \times \mathbb{R}$. Then, by Theorem 3.4, the BVP (4-2) has a solution.

References

- [Agarwal et al. 2001] R. P. Agarwal, M. Meehan, and D. O'Regan, *Fixed point theory and applications*, Cambridge Tracts in Mathematics **141**, Cambridge University Press, 2001. MR Zbl
- [Al Mosa and Eloe 2016] S. Al Mosa and P. Eloe, "Upper and lower solution method for boundary value problems at resonance", *Electron. J. Qual. Theory Differ. Equ.* **2016** (2016), art. id. 40. MR Zbl
- [Aljedani and Eloe 2018] J. Aljedani and P. Eloe, "Uniqueness of solutions of boundary value problems at resonance", *Adv. Theory Nonlinear Anal. Appl.* **2:3** (2018), art. id. 2018:15. Zbl
- [Almansour and Eloe 2015] A. Almansour and P. Eloe, "Fixed points and solutions of boundary value problems at resonance", *Ann. Polon. Math.* **115:3** (2015), 263–274. MR Zbl
- [Alshammari et al. 2019] M. Alshammari, K. Alanazi, and P. Eloe, "Quasilinearization and boundary value problems at resonance", preprint, 2019. To appear in *Georgian Math. J.*
- [Garcia and Neugebauer 2019] A. E. Garcia and J. T. Neugebauer, "Solutions of boundary value problems at resonance with periodic and antiperiodic boundary conditions", *Involve* **12:1** (2019), 171–180. MR Zbl
- [Han 2007] X. Han, "Positive solutions for a three-point boundary value problem at resonance", *J. Math. Anal. Appl.* **336:1** (2007), 556–568. MR Zbl
- [Infante et al. 2016] G. Infante, P. Pietramala, and F. A. F. Tojo, "Non-trivial solutions of local and non-local Neumann boundary-value problems", *Proc. Roy. Soc. Edinburgh Sect. A* **146:2** (2016), 337–369. MR Zbl

Received: 2019-03-21 Revised: 2019-04-05 Accepted: 2019-09-20

f20150671@hyderabad.bits-pilani.ac.in

*Department of Mathematics, Birla Institute of Technology
and Science Pilani, Hyderabad, India*

f20160518@hyderabad.bits-pilani.ac.in

*Department of Mathematics, Birla Institute of Technology
and Science Pilani, Hyderabad, India*

j.jaganmohan@hotmail.com

*Department of Mathematics, Birla Institute of Technology
and Science Pilani, Hyderabad, India*

A few more trees the chromatic symmetric function can distinguish

Jake Huryn and Sergei Chmutov

(Communicated by Joel Foisy)

A well-known open problem in graph theory asks whether Stanley's *chromatic symmetric function*, a generalization of the chromatic polynomial of a graph, distinguishes between any two nonisomorphic trees. Previous work has proven the conjecture for a class of trees called *spiders*. This paper generalizes the class of spiders to *n-spiders*, where normal spiders correspond to $n = 1$, and verifies the conjecture for $n = 2$.

1. The chromatic symmetric function and its coefficients

Let $G = (V, E)$ be a graph, and let $\mathcal{K}(F)$ denote the set of connected components of the graph (V, F) for any $F \subseteq E$. Then the *chromatic symmetric function* of G can be expressed as [Stanley 1995, Theorem 2.5]

$$X_G = \sum_{F \subseteq E} \left((-1)^{|F|} \prod_{K \in \mathcal{K}(F)} p_{|V(K)|} \right). \quad (1)$$

This is the so-called *subsets of edges* formulation of X_G . Here the symbols p_k represent the power sum symmetric functions, but will be treated as formal commuting indeterminants. Up to the sign $(-1)^{|F|}$ this polynomial was introduced in [Chmutov et al. 1994, Section 1.3], motivated by problems in knot theory, and its relationship with Stanley's definition was observed in [Noble and Welsh 1999, Theorem 6.1]. The question of whether this invariant distinguishes all nonisomorphic trees has been studied for many specific classes of trees, for example in [Martin et al. 2008; Loebl and Sereni 2019], and has been computationally verified for all trees with up to 29 vertices [Heil and Ji 2019].

Recall that a *tree* is a connected graph lacking cycles. The removal of any edge of a tree disconnects the graph into a disjoint union of two trees, and so the removal of n edges produces a disjoint union of $n + 1$ trees. A *leaf* of a tree is a vertex of degree 1.

MSC2010: 05C05, 05C31, 05E05.

Keywords: graph theory, combinatorics, chromatic symmetric function.

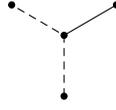


Figure 1. Dashed lines represent deletions, i.e., edges not in the chosen subset. The figure shows a 2-cut corresponding to the product $p_1^2 p_2$ and contributing to the coefficient $c_{1,1}$. Note that there are three distinct ways to make a 2-cut of this tree which yields the same partition, so $c_{1,1} = 3$ and X_T has a $-3p_1^2 p_2$ term. Considering all subsets of edges, we find that $X_T = p_1^4 - 3p_1^2 p_2 + 3p_1 p_3 - p_4$, where each term corresponds to successively larger subsets.

Let $T = (V, E)$ be a tree. We will discuss subsets $F \subseteq E$ by how many edges are removed from E , which is one fewer than the number of connected components of (V, F) , and thus one fewer than the number of factors in the product

$$\prod_{K \in \mathcal{K}(F)} p_{|V(K)|}.$$

This means that no cancellation will occur in X_T , since if terms have opposite signs, they must have a different number of p_k factors. (This is false if T is not a tree, and indeed, general graphs are not distinguished by their chromatic symmetric functions; see [Martin et al. 2008] for examples.) This lack of cancellation is important because it allows us to give, for trees, a coherent combinatorial interpretation of (1).

Moreover, each $F \subseteq E$ induces a partition $\lambda = (\lambda_1, \dots, \lambda_\ell)$ of $|V|$. The parts of λ are the numbers $|V(K)|$ for all $K \in \mathcal{K}(F)$, since the connected components of (V, F) partition V itself. We see that in (1), the term in the sum corresponding to F will be $(-1)^{|F|} p_{\lambda_1} \cdots p_{\lambda_\ell}$. Given a partition λ of $|V|$, denote the absolute value of the coefficient on $p_{\lambda_1} \cdots p_{\lambda_\ell}$ in X_T (now considering all subsets of E) by $c_\lambda(T) = c_{\lambda_2, \dots, \lambda_\ell}(T)$. Thus $c_\lambda(T)$ is the number of subsets of the edge set which induce the partition λ . We will write the subscripts in weakly decreasing order, omitting the largest part, and often simply write c_λ or $c_{\lambda_2, \dots, \lambda_\ell}$ when working with a specific tree.

We have come to the combinatorial interpretation for c_λ (and thus for (1)), which represents the number of ways we can remove $\ell - 1$ edges from T (an $(\ell - 1)$ -cut of T) to partition T , via the spanning subgraph, into connected components of order $\lambda_1, \dots, \lambda_\ell$ (Figure 1). In particular, the reader is encouraged to check that c_1 is the number of leaves of T . Intuitively, X_T contains all information about the sizes of the components we can get when we remove any subset of the tree's edges. Stanley's conjecture is exactly that this data is sufficient to completely determine any tree's isomorphism class.

The difficulty of Stanley’s conjecture, in an informal sense, comes from the fact that the information contained in each coefficient of X_T is nonlocal; for example, we can determine from X_T the number of vertices, number of leaves, the degree sequence, and the path sequence [Martin et al. 2008, Corollary 5]. The problem is how we can piece together this information, along with the other information in the coefficients, to see the small-scale shapes of the tree and how these structures are connected.

This paper attempts to develop a framework which allows for this to be done, and uses this framework to prove that a particular infinite family of trees is distinguished by the chromatic symmetric function.

2. Rooted subtrees and X_T

Of central importance in the following results is the use of *rooted trees*, that is, trees with one vertex designated as the root. Let $r(n)$ be the number of rooted tree isomorphism classes having order n . We will name each rooted tree isomorphism class $R_{n,i}$ (giving the same name to any rooted tree in this class), where n is the order of the rooted tree and $i \in \{1, \dots, r(n)\}$ is a semiarbitrary indexing of all rooted trees of order n . In particular, let $R_{n,1}$ denote a path with the root at a leaf, and, for $n \geq 2$, let $R_{n+1,2}$ be the same, with the addition of a single vertex appended to the second outwardmost vertex from the root (Figure 2).

We also define $c_\lambda(R)$ for any rooted tree R . Given a partition λ , define $c_\lambda(R)$ as the number of ways to cut R into $|\lambda| + 1$ connected components such that the orders of those components not containing the root correspond to the parts of λ . In this case, no parts of λ will be omitted from the subscript of c_λ , since the connected component containing the root is already omitted from λ . For example, the root of R may also be a leaf, but it will not contribute to $c_1(R)$. In particular, we have $c_1(R_{n,1}) = 1$ and $c_1(R_{n,2}) = 2$ for all n .

In practice, these rooted trees will be employed as *rooted subtrees* of a tree T . Suppose we remove an edge from T , splitting T into two connected components. Then each component forms a rooted subtree of T , in which the root is the vertex

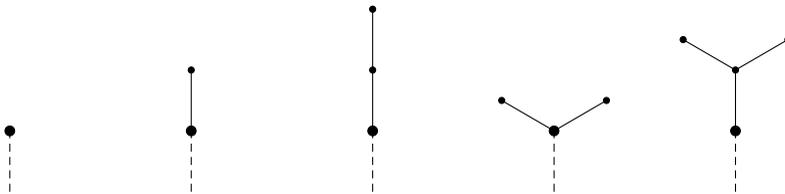


Figure 2. Rooted subtrees $R_{1,1}$, $R_{2,1}$, $R_{3,1}$, $R_{3,2}$, and $R_{4,2}$, respectively, with the roots enlarged. The dashed lines are the unique edges connecting the rooted subtrees to the rest of the tree.

incident to the edge we removed. This process produces every rooted subtree of T . Let the number of rooted subtrees of T isomorphic to $R_{n,i}$ be denoted $\rho_{n,i}(T)$, or simply $\rho_{n,i}$. The method of this paper hinges upon the calculation of these numbers $\rho_{n,i}$, which essentially represent the small-scale shapes of the tree, from the coefficients of X_T .

Let T have order d . It follows from our definitions that we have the equations

$$c_n = \sum_{i=1}^{r(n)} \rho_{n,i} = \sum_{i=1}^{r(d-n)} \rho_{d-n,i}$$

if $n \neq d/2$, and

$$c_{d/2} = \frac{1}{2} \sum_{i=1}^{r(d/2)} \rho_{d/2,i}.$$

The following result is a similar but more complicated equation for $c_{n,1}$. The reader may check that $c_{1,1} = \binom{c_1}{2} + c_2$, but this gives us no more information about T .

Lemma 1. *Let $T = (V, E)$ be a tree of order d , and let $T_{n,i}$ denote the tree obtained from $R_{n,i}$ by forgetting the distinction of the root. Then if $2 \leq n < (d - 1)/2$,*

$$c_{n,1} = \sum_{i=1}^{r(n)} (c_1 - c_1(R_{n,i})) \rho_{n,i} + \sum_{j=1}^{r(n+1)} c_1(T_{n+1,j}) \rho_{n+1,j}. \tag{2}$$

Also, if $d \geq 6$, then

$$c_{1,1,1,1} = \binom{c_1}{4} + \binom{c_1-1}{2} c_2 + \binom{c_2}{2} + (c_1 - 1) \rho_{3,1} + (c_1 - 2) \rho_{3,2} + c_4. \tag{3}$$

Proof. The coefficient $c_{n,1}$ tells us the number of ways we can cut T in two places to get a connected component of order 1 and another of order n , and of course a third of order $d - n - 1$. There are three ways of joining these components with two edges as illustrated by Figure 3.

The derivation of (2) should then be clear from the figure. In particular, observe that the second picture requires that the isolated vertex not originally be the root of the rooted subtree of order $n + 1$, while the third requires exactly the opposite. Thus, both terms combined negate the distinction of the root.

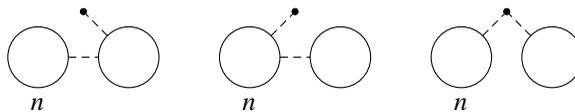


Figure 3. The three ways of joining the connected components of a partition contributing to $c_{n,1}$. The first picture corresponds to the first term of (2). The labeled components have order n .

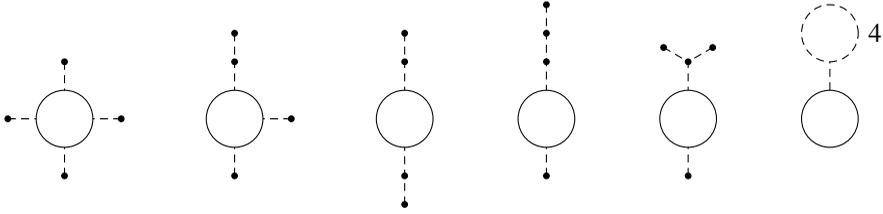


Figure 4. The ways of joining the connected components of the partition contributing to $c_{1,1,1,1}$. The pictures correspond to the respective terms in (3).

Now let $d \geq 6$. From Figure 4 we can see that (3) holds; we require $d \geq 6$ so that the nondashed circles drawn in Figure 4 are distinguished from a vertex. \square

Thus the coefficients $c_{1,1,1,1}$ and $c_3 = \rho_{3,1} + \rho_{3,2}$ (or $c_3 = \rho_{3,1}/2 + \rho_{3,2}/2$ if $d = 6$) form a system of equations which allows us to solve for $\rho_{3,1}$ and $\rho_{3,2}$.

3. Distinguishing spiders and their generalizations

A *spider* is a tree with exactly one vertex of degree at least 3. That vertex is called the *torso* of the spider, and the other vertices form paths extending from the torso called *legs*. It is shown in [Martin et al. 2008, Section 3] that the chromatic symmetric function completely distinguishes spiders.

Call a tree a *2-spider* if it is a modification of a spider in which any leg may be appended with a single vertex joined to the second outwardmost vertex (with respect to the torso) of that leg. Call such modified legs the *2-legs*. The reasoning behind these names is that one can think of these modified legs as being copies of the rooted subtree $R_{n,2}$, while normal legs are copies of $R_{n,1}$. We may thus call normal spiders *1-spiders* and normal legs *1-legs*. The order of a leg is defined to not include the root, so that the order of a (2-)spider is one more than the sum of the orders of each leg (Figure 5).

We may describe these 2-spiders up to isomorphism by a pair of positive integer sequences $(\lambda; \mu) = (\lambda_1, \dots, \lambda_\ell; \mu_1, \dots, \mu_m)$, where λ lists the orders of the 1-legs and μ lists the orders of the 2-legs. With this in mind, we can now state and prove the main result.

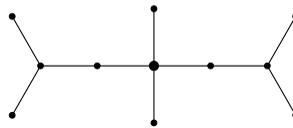


Figure 5. A 2-spider with two 2-legs, satisfying (ii) of Theorem 2, described by $(\lambda; \mu) = (1, 1; 4, 4)$ and $(\lambda^*; \mu^*) = (2; 2, 2, 2, 2)$. The torso is enlarged.

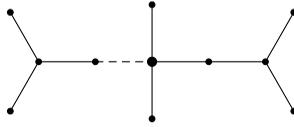


Figure 6. The dashed edge represents the edge we cut. Note that we have split the tree into two subtrees, one of which is of the form $R_{4,2}$, and the other, containing the torso, which is not of the form $R_{n,1}$ or $R_{n,2}$ since the tree does not satisfy (i). Thus the leg we cut contributes to $\rho_{4,2}$, and of course this means it must be a 2-leg with order at least 4. Note that we can also cut this same leg on different edges to contribute to $\rho_{3,2}$ and twice to $\rho_{1,1}$.

Theorem 2. *If T_1 and T_2 are nonisomorphic 2-spiders, then $X_{T_1} \neq X_{T_2}$.*

Proof. Let T be a 2-spider with order d ; the smallest 2-spider which is not also a 1-spider has order 6, so assume $d \geq 6$. We will show that T , knowing that it is a 2-spider, can be reconstructed from X_T .

We would like to determine, for $1 \leq n \leq d/2$, the number λ_n^* of 1-legs with order at least n and the number μ_n^* of 2-legs with order at least n (although necessarily $\mu_1^* = \mu_2^* = \mu_3^*$, since any 2-leg must have order at least 3). If we consider λ and μ to be partitions, then λ^* and μ^* are their respective conjugate partitions, so with this information we can recover T .

The proof is separated based on the following (neither mutually exclusive nor exhaustive) possibilities:

- (i) T has only three legs, two of order 1.
- (ii) If a leg of T has order n , then $n \leq d/2$.

First, suppose T satisfies (i). We find that either $\rho_{3,2} = 1$ and $c_1 = 3$, or $\rho_{3,2} = 2$ and $c_1 = 4$. This is sufficient to distinguish this case, and we also know that the number of 2-legs of T must be $\rho_{3,2} - 1$. From this it is easy to reconstruct T since we know d .

Now suppose T satisfies (ii) but not (i). Suppose we cut an edge of some leg L to split T into two rooted subtrees. Then the connected component containing the torso will not be of the form $R_{n,1}$ or $R_{n,2}$ for any n (note that this is false if T satisfies (i)). However, the other connected component will be of the form $R_{n,1}$ if L is a 1-leg or $R_{n,2}$ if L is a 2-leg, and L must have order at least n . Thus if $n \geq 2$, $\lambda_n^* = \rho_{n,1}$ and $\mu_n^* = \rho_{n,2}$ (Figure 6). However, if $n = 1$ then $\rho_{1,1}$ also counts both leaves on each 2-leg, so to correct for this double counting we instead use

$$\lambda_1^* = \rho_{1,1} - 2\rho_{3,2}.$$

We know $\rho_{1,1} = c_1$ and $\rho_{2,1} = c_2$, and moreover we know $\rho_{3,1}$ and $\rho_{3,2}$ in terms of the coefficients c_λ from (3) and $c_3 = \rho_{3,1} + \rho_{3,2}$. Now let $4 \leq n < d/2$. It follows from the length restriction (ii) that any rooted subtree of T containing the torso must have order at least $d/2$. Thus since any rooted subtree not containing the torso is either of the form $R_{n,1}$ or $R_{n,2}$, we know that if i is not 1 or 2, then $\rho_{n,i} = 0$. Then

$$c_n = \rho_{n,1} + \rho_{n,2},$$

and, by Lemma 1,

$$c_{n-1,1} = (c_1 - 1)\rho_{n-1,1} + (c_1 - 2)\rho_{n-1,2} + 2\rho_{n,1} + 3\rho_{n,2}.$$

Thus we may solve inductively for all such $\rho_{n,1}$ and $\rho_{n,2}$ with $n = 3$ as the base case.

Now suppose $n = d/2$; there can be at most one leg of this order. From (2) we can tell if there is a leg of order n by examining $c_{n-1,1}$, since if there were not, the second sum in the equation would be zero. Note then that exactly one of $\rho_{n-1,1}$ and $\rho_{n-1,2}$ can be nonzero, since a 2-spider cannot have a leg of order $d/2 - 1$ if it also has one of order $d/2$. Thus we know the type of that largest leg.

Notice that in case (ii), $c_i \geq c_j$ if $3 \leq i < j < d/2$ since

$$c_i = \rho_{i,1} + \rho_{i,2} = \lambda_i^* + \mu_i^* \geq \lambda_j^* + \mu_j^* = c_j.$$

Finally, suppose T satisfies none of the cases. The smallest such 2-spider has order 9, so suppose $d \geq 9$. Let the length of the (unique) largest leg have order $k > d/2$. A quick calculation verifies that we cannot also have a leg of order $d - k - 1$ or larger, and that

$$3 \leq d - k - 1 < \frac{d}{2} - 1.$$

Thus $c_{d-k-1} = 1$, but $c_{d-k} = 2$, which distinguishes this case from the second case (Figure 7). Moreover, we only need to find λ_n^* and μ_n^* up to $n = d - k - 1$, which will give the types of all legs, and their respective lengths, excluding the length of the longest leg. To do this, we simply solve for $\rho_{3,1}$ and $\rho_{3,2}$ and do the same process as in the second case up to $c_{d-k-2,1}$. Then we can find the length of the longest leg, thus reconstructing T , by looking at d . □

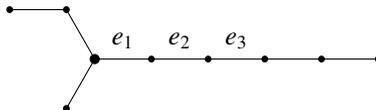


Figure 7. A 2-spider which satisfies neither (i) nor (ii), with torso enlarged; here $d = 9$ and $k = 5$. The edge e_3 is the only one we can cut to get a connected component of order 3, but both e_1 and e_2 give connected components of order 4, that is, $c_3 = 1$ but $c_4 = 2$.

It turns out that solving Stanley’s conjecture can be done by simply solving for all $\rho_{n,i}$ in terms of the coefficients of X_T . In Section 2 we did this for $n = 3$, by examining c_3 and $c_{1,1,1,1}$, which by itself proves Stanley’s conjecture for all trees of order at most 7. This avoids the unpleasant case-checking which allows Theorem 2 to verify Stanley’s conjecture for an infinite family. While there are some interesting heuristics for determining which c_λ can yield which coefficients on the desired $\rho_{n,i}$ in its expansion, even for $n = 4$ this is very difficult to do.

Even worse, if we were attempting to distinguish trees of order 20, we would need to solve for each variable $\rho_{10,i}$, of which there would be more than coefficients of X_T ! What this seems to imply, then, is that there must be some “hidden” information about the values $\rho_{n,i}$ can take, as a consequence of the definition of a tree, that is not expressed directly in the chromatic symmetric function.

4. Acknowledgements

This work was done as a part of the Summer 2018 undergraduate research working group “Knots and Graphs” at the Ohio State University, which can be found at <http://www.math.ohio-state.edu/~chmutov/wor-gr-su18/wor-gr.htm>. We are grateful to all participants of the group for valuable discussions and to the OSU Honors Program Research Fund for the students’ financial support. Chmutov was supported by the grant NSF-DMS #1547357 RTG: Algebraic Topology and Its Applications.

References

- [Chmutov et al. 1994] S. V. Chmutov, S. V. Duzhin, and S. K. Lando, “Vassiliev knot invariants, III: Forest algebra and weighted graphs”, pp. 135–145 in *Singularities and bifurcations*, edited by V. I. Arnol’d, Adv. Soviet Math. **21**, Amer. Math. Soc., Providence, RI, 1994. MR Zbl
- [Heil and Ji 2019] S. Heil and C. Ji, “On an algorithm for comparing the chromatic symmetric functions of trees”, *Australas. J. Combin.* **75** (2019), 210–222. MR
- [Loebl and Sereni 2019] M. Loebl and J.-S. Sereni, “Isomorphism of weighted trees and Stanley’s isomorphism conjecture for caterpillars”, *Ann. Inst. Henri Poincaré D* **6**:3 (2019), 357–384. MR Zbl
- [Martin et al. 2008] J. L. Martin, M. Morin, and J. D. Wagner, “On distinguishing trees by their chromatic symmetric functions”, *J. Combin. Theory Ser. A* **115**:2 (2008), 237–253. MR Zbl
- [Noble and Welsh 1999] S. D. Noble and D. J. A. Welsh, “A weighted graph polynomial from chromatic invariants of knots”, *Ann. Inst. Fourier (Grenoble)* **49**:3 (1999), 1057–1087. MR Zbl
- [Stanley 1995] R. P. Stanley, “A symmetric function generalization of the chromatic polynomial of a graph”, *Adv. Math.* **111**:1 (1995), 166–194. MR Zbl

Received: 2019-03-30 Revised: 2019-08-02 Accepted: 2019-11-04

huryn.5@osu.edu

The Ohio State University, Columbus, OH, United States

chmutov@math.ohio-state.edu

The Ohio State University, Columbus, OH, United States

One-point hyperbolic-type metrics

Marina Borovikova, Zair Ibragimov,
Miguel Jimenez Bravo and Alexandro Luna

(Communicated by Kenneth S. Berenhaut)

We study basic properties of one-parametric families of the \tilde{j} -metric, the scale-invariant Cassinian metric and the half-Apollonian metric on locally compact, noncomplete metric spaces. We first establish basic properties of these metrics on once-punctured general metric spaces and obtain sharp estimates between these metrics, and then we show that all these properties, except for δ -hyperbolicity, extend to the settings of locally compact noncomplete metric spaces. We also show that these metrics are δ -hyperbolic only if the underlying space is a once-punctured metric space.

1. Introduction

The hyperbolic metric and hyperbolic geometry have become powerful tools in geometric function theory on Euclidean spaces; see [Beardon and Minda 2007]. In recent years, various *one-point hyperbolic-type* metrics have been introduced as a substitute for the hyperbolic metric in more general metric spaces. Such metrics can be classified into two groups, namely, *one-point length metrics* and *one-point distance-function metrics*. Examples of one-point length metrics include the quasihyperbolic metric, the Ferrand metric and the Kulkarni–Pinkall metric. These metrics have been extensively studied by many authors; see [Ferrand 1988; Gehring and Palka 1976; Herron 2016; Herron and Julian 2013; Kulkarni and Pinkall 1994]. In this paper, we study one-point distance-function metrics such as the \tilde{j} -metric, the half-Apollonian metric, and the scale-invariant Cassinian metric in general metric spaces. These metrics were originally introduced and studied in Euclidean spaces; see, for example, [Hästö 2006; Hästö et al. 2006; Hästö and Lindén 2004; Ibragimov 2003; 2016; Lindén 2007; Vuorinen 1988].

One of the key features of the one-point distance-function metrics is their δ -hyperbolicity property, which was introduced by M. Gromov [1987] as an extension

MSC2010: primary 30F45; secondary 51F99, 30C99.

Keywords: semimetric spaces, metric spaces, Ptolemaic spaces, δ -hyperbolic spaces, half-Apollonian metric, \tilde{j} -metric, scale-invariant Cassinian metric.

of the concept of negative curvature to general metric spaces. The notion of δ -hyperbolicity has found numerous applications in many areas of mathematics and is widely used in geometric function theory, geometric group theory, and analysis on metric spaces. For more discussions on δ -hyperbolic spaces the reader is referred to [Beardon and Minda 2007; Bonk 2006; Bonk et al. 2001; Bonk and Schramm 2000; Gromov 1987; Väisälä 2005].

We recall the following general approach to constructing one-point distance-function metrics in Euclidean spaces. Let $D \subset \mathbb{R}^n$ be any domain with a nonempty boundary ∂D . To construct a one-point distance-function metric d_D on D , one first constructs a δ -hyperbolic metric d_p on the once-punctured space $\mathbb{R}^n \setminus \{p\}$ for each $p \in \mathbb{R}^n$ and then defines d_D by $d_D(x, y) = \sup\{d_p(x, y) : p \in \partial D\}$. Taking the supremum in this context is very natural since the boundary ∂D is usually uncountable. Clearly, the supremum of metrics is again a metric. However, as it was shown in [Aksoy et al. 2018], the δ -hyperbolicity property is not, in general, preserved when taking the supremum. Metrics obtained in this manner are usually referred to as one-point hyperbolic-type metrics.

In Euclidean spaces, the \tilde{j} -metric, the half-Apollonian, and the scale-invariant Cassinian metrics are shown to be *roughly similar* to each other [Ibragimov 2016, Theorems 3.3 and 3.5]. Moreover, these metrics are δ -hyperbolic if and only if the underlying domain has only one boundary point. In other words, if the domain has more than one boundary point, then these metrics, which are defined as the supremums over all boundary points, are not δ -hyperbolic. The main purpose of this paper is to show that similar results hold in general metric spaces.

We note that even though some results in this paper are similar to those in [Ibragimov 2016], the methods used in this paper differ significantly. The main reason for this is the fact that the underlying Euclidean metric, which was used in [Ibragimov 2016], satisfies both the triangle inequality and Ptolemy's inequality. Another reason is the fact that Euclidean spaces support full groups of Möbius transformations. For example, the one-point scale-invariant Cassinian metric, $\tilde{\tau}_p$, is invariant under Möbius transformations, see [Ibragimov 2016, Lemma 2.1], which was used in the proof of [Ibragimov 2016, Lemma 2.2]. Because of these reasons, the distance function $\tilde{\tau}_{c,a}$ considered in Section 4 is a metric with constant $c = 1$ over Euclidean spaces. Using another method, the same result was proved in general metric spaces provided the underlying metric satisfies Ptolemy's inequality; see [Aksoy et al. 2018, Theorem 2.2]. In that paper, it was also shown that $\tilde{\tau}_{c,a}$ is a metric, over any general metric space, when one takes $c = 2$. One of the main results of this paper states that $\tilde{\tau}_{c,a}$ is still a metric, over general metric spaces, with $c = \frac{3}{2}$ (see Theorem 4.3), and that this constant is sharp, as is shown in the example preceding Theorem 4.3. We note that a similar metric was introduced in [Dovgoshey et al. 2016], which coincides with $\tilde{\tau}_{c,a}$ in once-punctured metric spaces;

however, this metric is different when considering more than one boundary point. In particular, the sharpness of $\frac{3}{2}$ in this paper improves the result of Theorem 1.1 in [Dovgoshey et al. 2016], over once-punctured metric spaces.

Moreover, for the families of one-point hyperbolic-type metrics discussed in this paper, Ptolemy's inequality turns out to be irrelevant for the underlying distance function ρ , but the triangle inequality is essential (see Lemmas 3.2 and 4.2). Lastly, in this paper we also consider the sharpness of c for the family of one-point \tilde{j} -metrics $\tilde{j}_{c,a}$, which has not been considered before. In particular, we show that the distance function $\tilde{j}_{c,a}$ is a metric when $c = 1$, and that this constant is sharp as is shown in the example following Lemma 3.2. Ultimately, our goal is to study the metrics considered in this paper in locally compact noncomplete metric spaces. In Section 6, we establish preliminary results of the metrics in this setting, including sharp estimates between them (see Theorem 6.5).

2. Basic concepts

A function $\rho : \mathcal{X} \times \mathcal{X} \rightarrow [0, \infty)$ on a set \mathcal{X} is called a *distance function* on \mathcal{X} if, for all $x, y \in \mathcal{X}$, we have $\rho(x, x) = 0$ and $\rho(x, y) = \rho(y, x)$. A distance function ρ is called a *semimetric* if $\rho(x, y) = 0$ implies $x = y$. A distance function ρ is called a *pseudometric* if it satisfies the *triangle inequality*. That is, for all $x, y, z \in \mathcal{X}$, we have

$$\rho(x, y) \leq \rho(x, z) + \rho(z, y).$$

A distance function ρ on a set \mathcal{X} is called δ -*hyperbolic* ($\delta \geq 0$) if ρ satisfies the δ -*hyperbolicity condition*: for all $x, y, z, w \in \mathcal{X}$, we have

$$\rho(x, y) + \rho(z, w) \leq [\rho(x, z) + \rho(y, w)] \vee [\rho(x, w) + \rho(y, z)] + 2\delta.$$

Here, and in what follows, we set $s \vee t = \max\{s, t\}$ and $s \wedge t = \min\{s, t\}$ for nonnegative numbers s and t . A semimetric ρ is called a *metric* if it satisfies the triangle inequality. A semimetric ρ is called *Ptolemaic* if ρ satisfies *Ptolemy's inequality*. That is, for all $x, y, z, w \in \mathcal{X}$, we have

$$\rho(x, y)\rho(z, w) \leq \rho(x, z)\rho(y, w) + \rho(x, w)\rho(y, z).$$

The pair (\mathcal{X}, ρ) , where ρ is semimetric (pseudometric, metric, Ptolemaic or δ -hyperbolic metric) on \mathcal{X} is called a *semimetric space* (*pseudometric space*, *metric space*, *Ptolemaic space* or *δ -hyperbolic metric space*).

Two metrics ρ_1 and ρ_2 on a set \mathcal{X} are called *bilipschitz equivalent* if there exist $L_1 > 0$ and $L_2 > 0$ such that, for all $x, y \in \mathcal{X}$, we have

$$L_1\rho_1(x, y) \leq \rho_2(x, y) \leq L_2\rho_1(x, y).$$

The metrics ρ_1 and ρ_2 on a set \mathcal{X} are called *roughly similar* if there exist $\lambda > 0$ and $k \geq 0$ such that, for all $x, y \in \mathcal{X}$, we have

$$\lambda\rho_1(x, y) - k \leq \rho_2(x, y) \leq \lambda\rho_1(x, y) + k.$$

We say that the metrics ρ_1 and ρ_2 on a set \mathcal{X} are *geometrically equivalent* if they are both bilipschitz equivalent and roughly similar. We do not distinguish metrics that are geometrically equivalent. We will make use of the following lemma.

Lemma 2.1. *Let (\mathcal{X}, ρ) be a semimetric space. Then*

$$\begin{aligned} & [\rho(x, a) \vee \rho(y, a)][\rho(z, a) \vee \rho(w, a)] \\ & \leq ([\rho(x, a) \vee \rho(z, a)][\rho(y, a) \vee \rho(w, a)]) \vee ([\rho(x, a) \vee \rho(w, a)][\rho(y, a) \vee \rho(z, a)]) \end{aligned}$$

for all $a, x, y, z, w \in \mathcal{X}$.

Proof. We have

$$\begin{aligned} & [\rho(x, a) \vee \rho(y, a)][\rho(z, a) \vee \rho(w, a)] \\ & = \rho(x, a)\rho(z, a) \vee \rho(x, a)\rho(w, a) \vee \rho(y, a)\rho(z, a) \vee \rho(y, a)\rho(w, a) \\ & \leq [\rho(x, a)\rho(y, a) \vee \rho(x, a)\rho(w, a) \vee \rho(z, a)\rho(y, a) \vee \rho(z, a)\rho(w, a)] \\ & \quad \vee [\rho(x, a)\rho(y, a) \vee \rho(x, a)\rho(z, a) \vee \rho(w, a)\rho(y, a) \vee \rho(w, a)\rho(z, a)] \\ & = ([\rho(x, a) \vee \rho(z, a)][\rho(y, a) \vee \rho(w, a)]) \vee ([\rho(x, a) \vee \rho(w, a)][\rho(y, a) \vee \rho(z, a)]), \end{aligned}$$

as required. \square

3. A family of one-point \tilde{j} -metrics

Let (\mathcal{X}, ρ) be an arbitrary semimetric space and let $a \in \mathcal{X}$ be an arbitrary point. For convenience, we put $\mathcal{X}_a = \mathcal{X} \setminus \{a\}$. Fix a parameter $c > 0$ and consider the family $\{\tilde{j}_{c,a}\}$ of distance functions, defined by

$$\tilde{j}_{c,a}(x, y) = \log\left(1 + c \frac{\rho(x, y)}{\rho(x, a) \wedge \rho(y, a)}\right), \quad x, y \in \mathcal{X}_a. \quad (3.1)$$

Clearly, $\tilde{j}_{c,a}(x, y) \geq 0$, $\tilde{j}_{c,a}(x, y) = \tilde{j}_{c,a}(y, x)$, and $\tilde{j}_{c,a}(x, y) = 0$ if and only if $x = y$ so that $\tilde{j}_{c,a}$ is a semimetric. Our aim in this section is to show that the semimetric $\tilde{j}_{c,a}$ is a δ -hyperbolic metric provided ρ is a metric and $c \geq 1$.

We begin by showing that $\tilde{j}_{c,a}$ does not, in general, satisfy the triangle inequality even if ρ is assumed to be Ptolemaic.

Lemma 3.2. *For each $c > 0$, there exists a four-point Ptolemaic space (\mathcal{X}, ρ) , $\mathcal{X} = \{a, x, y, z\}$, such that $\tilde{j}_{c,a}(x, y) > \tilde{j}_{c,a}(x, z) + \tilde{j}_{c,a}(z, y)$.*

Proof. Let \mathcal{X} be a four-point set, say $\mathcal{X} = \{a, x, y, z\}$, and let ρ be a semimetric on \mathcal{X} with

$$\rho(x, y) = 3 + c, \quad \rho(y, a) = 2 + c, \quad \rho(x, z) = \rho(x, a) = \rho(z, a) = \rho(z, y) = 1.$$

Then

$$\rho(x, y)\rho(z, a) = 3 + c, \quad \rho(x, z)\rho(y, a) = 2 + c, \quad \rho(x, a)\rho(y, z) = 1$$

so that (\mathcal{X}, ρ) is a Ptolemaic space. We have

$$\tilde{j}_{c,a}(x, y) = \log(1 + 3c + c^2), \quad \tilde{j}_{c,a}(x, z) = \log(1 + c), \quad \tilde{j}_{c,a}(z, y) = \log(1 + c)$$

so that $\tilde{j}_{c,a}(x, y) > \tilde{j}_{c,a}(x, z) + \tilde{j}_{c,a}(z, y)$, completing the proof. \square

The next example shows that if we require that $\tilde{j}_{c,a}$ satisfy the triangle inequality for arbitrary metric spaces (\mathcal{X}, ρ) , then we must have $c \geq 1$. Indeed, let $\mathcal{X} = \mathbb{S}$ be the unit circle in the complex plane \mathbb{C} and let ρ be the path metric on \mathcal{X} . Put $a = 1$, $x = e^{(\pi/3)i}$, $z = e^{(2\pi/3)i}$, and $y = -1$. Then

$$\rho(x, a) = \rho(x, z) = \rho(z, y) = \frac{\pi}{3}, \quad \rho(x, y) = \rho(z, a) = \frac{2\pi}{3}, \quad \rho(y, a) = \pi$$

so that

$$\tilde{j}_{c,a}(x, y) = \log(1 + 2c), \quad \tilde{j}_{c,a}(x, z) = \log(1 + c), \quad \tilde{j}_{c,a}(z, y) = \log(1 + c/2).$$

Then the triangle inequality $\tilde{j}_{c,a}(x, y) \leq \tilde{j}_{c,a}(x, z) + \tilde{j}_{c,a}(z, y)$ is equivalent to $1 + 2c \leq (1 + c)(1 + c/2)$. The latter implies $c \geq 1$.

Theorem 3.3. *Let (\mathcal{X}, ρ) be any metric space and let $c \geq 1$. Then for any $a \in \mathcal{X}$ the space $(\mathcal{X}_a, \tilde{j}_{c,a})$ is a δ -hyperbolic metric space with $\delta = \log(2c + 1)$.*

Proof. First, we show that the semimetric $\tilde{j}_{c,a}$ satisfies the triangle inequality. Let $x, y, z \in \mathcal{X}_a$ be arbitrary points. Then the triangle inequality $\tilde{j}_{c,a}(x, y) \leq \tilde{j}_{c,a}(x, z) + \tilde{j}_{c,a}(z, y)$ is equivalent to

$$\frac{\rho(x, y)}{\rho(x, a) \wedge \rho(y, a)} \leq \frac{\rho(x, z)}{\rho(x, a) \wedge \rho(z, a)} + \frac{\rho(z, y)}{\rho(z, a) \wedge \rho(y, a)} + c \frac{\rho(x, z)\rho(z, y)}{[\rho(x, a) \wedge \rho(z, a)][\rho(z, a) \wedge \rho(y, a)]}.$$

Since $c \geq 1$ it suffices to show that

$$\frac{\rho(x, y)}{\rho(x, a) \wedge \rho(y, a)} \leq \frac{\rho(x, z)}{\rho(x, a) \wedge \rho(z, a)} + \frac{\rho(z, y)}{\rho(z, a) \wedge \rho(y, a)} + \frac{\rho(x, z)\rho(z, y)}{[\rho(x, a) \wedge \rho(z, a)][\rho(z, a) \wedge \rho(y, a)]}. \quad (3.4)$$

Due to symmetry we can assume that $\rho(x, a) \leq \rho(y, a)$.

Case 1: $\rho(z, a) \leq \rho(x, a)$. Then inequality (3.4) is equivalent to

$$\frac{\rho(x, y)}{\rho(x, a)} \leq \frac{\rho(x, z)}{\rho(z, a)} + \frac{\rho(z, y)}{\rho(z, a)} + \frac{\rho(x, z)\rho(z, y)}{[\rho(z, a)]^2}.$$

Since

$$\frac{\rho(x, y)}{\rho(x, a)} \leq \frac{\rho(x, z) + \rho(z, y)}{\rho(x, a)} \leq \frac{\rho(x, z)}{\rho(z, a)} + \frac{\rho(z, y)}{\rho(z, a)},$$

the latter follows.

Case 2: $\rho(x, a) \leq \rho(z, a) \leq \rho(y, a)$. Then inequality (3.4) is equivalent to

$$\frac{\rho(x, y)}{\rho(x, a)} \leq \frac{\rho(x, z)}{\rho(x, a)} + \frac{\rho(z, y)}{\rho(z, a)} + \frac{\rho(x, z)\rho(z, y)}{\rho(x, a)\rho(z, a)}$$

or, equivalently,

$$\rho(x, y)\rho(z, a) \leq \rho(x, z)\rho(z, a) + \rho(z, y)[\rho(x, a) + \rho(x, z)].$$

Since $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$ and $\rho(z, a) \leq \rho(x, a) + \rho(x, z)$, the inequality follows.

Case 3: $\rho(y, a) \leq \rho(z, a)$. Then inequality (3.4) is equivalent to

$$\frac{\rho(x, y)}{\rho(x, a)} \leq \frac{\rho(x, z)}{\rho(x, a)} + \frac{\rho(z, y)}{\rho(y, a)} + \frac{\rho(x, z)\rho(z, y)}{\rho(x, a)\rho(y, a)}$$

or, equivalently,

$$\rho(x, y)\rho(y, a) \leq \rho(x, z)\rho(y, a) + \rho(z, y)[\rho(x, a) + \rho(x, z)].$$

Since $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$ and $\rho(x, a) + \rho(x, z) \geq \rho(z, a) \geq \rho(y, a)$, the inequality follows, completing the proof that $\tilde{j}_{c,a}$ is a metric on \mathcal{X}_a .

Next, we show that the metric $\tilde{j}_{c,a}$ is δ -hyperbolic with $\delta = \log(2c + 1)$. Given $x, y \in \mathcal{X}_a$, by the triangle inequality, we have

$$\rho(x, a) \vee \rho(y, a) \leq \rho(x, a) \wedge \rho(y, a) + c\rho(x, y) \leq (2c + 1)[\rho(x, a) \vee \rho(y, a)]$$

so that

$$\frac{\rho(x, a) \vee \rho(y, a)}{\rho(x, a) \wedge \rho(y, a)} \leq \frac{\rho(x, a) \wedge \rho(y, a) + c\rho(x, y)}{\rho(x, a) \wedge \rho(y, a)} \leq \frac{(2c + 1)[\rho(x, a) \vee \rho(y, a)]}{\rho(x, a) \wedge \rho(y, a)}.$$

Hence

$$T_a(x, y) \leq \exp(\tilde{j}_{c,a}(x, y)) \leq (2c + 1)T_a(x, y), \quad (3.5)$$

where $\exp(t) = e^t$ is the exponential function and

$$T_a(x, y) = \frac{\rho(x, a) \vee \rho(y, a)}{\rho(x, a) \wedge \rho(y, a)} = \left[\frac{\rho(x, a) \vee \rho(y, a)}{\sqrt{\rho(x, a)\rho(y, a)}} \right]^2.$$

Now, let $x, y, z, w \in \mathcal{X}_a$ be arbitrary points. Lemma 2.1 implies

$$T_a(x, y)T_a(z, w) \leq T_a(x, z)T_a(y, w) \vee T_a(x, w)T_a(y, z).$$

Consequently, using (3.5) we have

$$\begin{aligned} \exp(\tilde{j}_{c,a}(x, y) + \tilde{j}_{c,a}(z, w)) &\leq (2c + 1)^2 [T_a(x, y)T_a(z, w)] \\ &\leq (2c + 1)^2 [T_a(x, z)T_a(y, w) \vee T_a(x, w)T_a(y, z)] \\ &\leq (2c + 1)^2 [\exp(\tilde{j}_{c,a}(x, z) + \tilde{j}_{c,a}(y, w)) \vee \exp(\tilde{j}_{c,a}(x, w) + \tilde{j}_{c,a}(y, z))] \end{aligned}$$

or, equivalently,

$$\begin{aligned} \tilde{j}_{c,a}(x, y) + \tilde{j}_{c,a}(z, w) &\leq [\tilde{j}_{c,a}(x, z) + \tilde{j}_{c,a}(y, w)] \vee [\tilde{j}_{c,a}(x, w) + \tilde{j}_{c,a}(y, z)] + 2 \log(2c + 1), \end{aligned}$$

completing the proof. \square

Theorem 3.3 suggests that a natural category of spaces on which to consider the family of semimetrics $\tilde{j}_{c,a}$ is the category of metric spaces. Moreover, in such a setting one should consider the parameter c to be greater than or equal to 1. Now, if (\mathcal{X}, ρ) is a metric space, $a \in \mathcal{X}$ and $c \geq 1$, then the metrics $\tilde{j}_{1,a}$ and $\tilde{j}_{c,a}$ are geometrically equivalent. Indeed, the latter follows from the facts that

$$\log(1 + x) \leq \log(1 + cx) \leq c \log(1 + x) \tag{3.6}$$

and

$$\log(1 + x) \leq \log(1 + cx) \leq \log(1 + x) + \log c \tag{3.7}$$

for all $x \geq 0$ and $c \geq 1$. Therefore, and since we do not distinguish geometrically equivalent metrics, we shall consider only the metric $\tilde{j}_{1,a}$, denoted simply by \tilde{j}_a ,

$$\tilde{j}_a(x, y) = \log\left(1 + \frac{\rho(x, y)}{\rho(x, a) \wedge \rho(y, a)}\right). \tag{3.8}$$

We shall refer to \tilde{j}_a as a *one-point \tilde{j} -metric*. For subdomains D of the Euclidean space \mathbb{R}^n the metric \tilde{j}_D , called the *\tilde{j} -metric* and defined by

$$\tilde{j}_D(x, y) = \sup\{\tilde{j}_a(x, y) : a \in \partial D\},$$

was introduced by M. Vuorinen [1988]. The following theorem is an immediate consequence of Theorem 3.3.

Theorem 3.9. *Let (\mathcal{X}, ρ) be any metric space. Then for each $a \in \mathcal{X}$, the space $(\mathcal{X} \setminus \{a\}, \tilde{j}_a)$ is a δ -hyperbolic metric space with $\delta = \log 3$.*

4. A family of one-point, scale-invariant Cassinian metrics

Let (\mathcal{X}, ρ) be a semimetric space and let $a \in \mathcal{X}$ be an arbitrary point. For convenience, we put $\mathcal{X}_a = \mathcal{X} \setminus \{a\}$. Fix a parameter $c > 0$ and consider the family $\{\tilde{\tau}_{c,a}\}$ of distance functions, defined by

$$\tilde{\tau}_{c,a}(x, y) = \log\left(1 + c \frac{\rho(x, y)}{\sqrt{\rho(x, a)\rho(y, a)}}\right), \quad x, y \in \mathcal{X}_a. \quad (4.1)$$

We have $\tilde{\tau}_{c,a}(x, y) \geq 0$, $\tilde{\tau}_{c,a}(x, y) = \tilde{\tau}_{c,a}(y, x)$, and $\tilde{\tau}_{c,a}(x, y) = 0$ if and only if $x = y$ so that $\tilde{\tau}_{c,a}$ is a semimetric. Our immediate goal is to show that the semimetric $\tilde{\tau}_{c,a}$ is a metric provided ρ is a metric and $c \geq \frac{3}{2}$.

We begin by showing that $\tilde{\tau}_{c,a}$ does not, in general, satisfy the triangle inequality even if ρ is assumed to be Ptolemaic.

Lemma 4.2. *For each $c > 0$, there exists a four-point Ptolemaic space (\mathcal{X}, ρ) , $\mathcal{X} = \{a, x, y, z\}$, such that $\tilde{\tau}_{c,a}(x, y) > \tilde{\tau}_{c,a}(x, z) + \tilde{\tau}_{c,a}(z, y)$.*

Proof. Let \mathcal{X} be a four-point set, say $\mathcal{X} = \{a, x, y, z\}$, and let ρ be a semimetric on \mathcal{X} with

$$\rho(x, y) = 4c + 2, \quad \rho(y, a) = 4c + 1, \quad \rho(x, z) = \rho(x, a) = \rho(z, a) = \rho(z, y) = 1.$$

Then

$$\rho(x, y)\rho(z, a) = 4c + 2, \quad \rho(x, z)\rho(y, a) = 4c + 1, \quad \text{and} \quad \rho(x, a)\rho(y, z) = 1$$

so that (\mathcal{X}, ρ) is a Ptolemaic space. We have

$$\tilde{\tau}_{c,a}(x, y) = \log\left(1 + c \frac{4c + 2}{\sqrt{4c + 1}}\right), \quad \tilde{\tau}_{c,a}(z, y) = \log\left(1 + c \frac{1}{\sqrt{4c + 1}}\right)$$

and $\tilde{\tau}_{c,a}(x, z) = \log(1 + c)$ so that $\tilde{\tau}_{c,a}(x, y) > \tilde{\tau}_{c,a}(x, z) + \tilde{\tau}_{c,a}(z, y)$, as required. \square

The next example shows that if we require that $\tilde{\tau}_{c,a}$ satisfy the triangle inequality for arbitrary metric spaces (\mathcal{X}, ρ) , then we must have $c \geq \frac{3}{2}$. Indeed, let $\mathcal{X} = \mathbb{S}$ be the unit circle in the complex plane \mathbb{C} and let ρ be the path metric on \mathcal{X} . Put $x = 1$, $y = -1$, $a = e^{i\theta}$, and $z = -e^{i\theta}$, where $0 < \theta < \pi$. Then

$$\rho(x, y) = \rho(z, a) = \pi, \quad \rho(x, a) = \rho(y, z) = \theta, \quad \rho(x, z) = \rho(y, a) = \pi - \theta$$

so that

$$\begin{aligned} \tilde{\tau}_{c,a}(x, y) &= \log\left(1 + c \frac{\pi}{\sqrt{\theta(\pi - \theta)}}\right), \\ \tilde{\tau}_{c,a}(x, z) &= \log\left(1 + c \frac{\pi - \theta}{\sqrt{\pi\theta}}\right) \quad \text{and} \quad \tilde{\tau}_{c,a}(z, y) = \log\left(1 + c \frac{\theta}{\sqrt{\pi(\pi - \theta)}}\right). \end{aligned}$$

Then the triangle inequality $\tilde{\tau}_{c,a}(x, y) \leq \tilde{\tau}_{c,a}(x, z) + \tilde{\tau}_{c,a}(z, y)$ is equivalent to

$$\frac{\pi}{\sqrt{\theta(\pi - \theta)}} \leq \frac{\pi - \theta}{\sqrt{\pi\theta}} + \frac{\theta}{\sqrt{\pi(\pi - \theta)}} + c \frac{\theta(\pi - \theta)}{\pi\sqrt{\theta(\pi - \theta)}}.$$

Equivalently,

$$\frac{\sqrt{\pi}[\pi^{3/2} - (\pi - \theta)^{3/2} - \theta^{3/2}]}{\theta(\pi - \theta)} \leq c.$$

Since

$$\lim_{\theta \rightarrow 0} \frac{\sqrt{\pi}[\pi^{3/2} - (\pi - \theta)^{3/2} - \theta^{3/2}]}{\theta(\pi - \theta)} = \lim_{\theta \rightarrow 0} \frac{\sqrt{\pi}[\frac{3}{2}(\pi - \theta)^{1/2} - \frac{3}{2}\theta^{1/2}]}{\pi - 2\theta} = \frac{3}{2},$$

we conclude that $c \geq \frac{3}{2}$.

Theorem 4.3. *Let (\mathcal{X}, ρ) be any metric space and let $c \geq \frac{3}{2}$. Then for each $a \in \mathcal{X}$, the space $(\mathcal{X}_a, \tilde{\tau}_{c,a})$ is a metric space.*

Proof. We need to show that

$$\tilde{\tau}_{c,a}(x, y) \leq \tilde{\tau}_{c,a}(x, z) + \tilde{\tau}_{c,a}(z, y). \tag{4.4}$$

Due to the scale invariance of $\tilde{\tau}_{c,a}$ we can assume that $\rho(z, a) = 1$. Then inequality (4.4) is equivalent to

$$\left(1 + c \frac{\rho(x, y)}{\sqrt{\rho(x, a)\rho(y, a)}}\right) \leq \left(1 + c \frac{\rho(x, z)}{\sqrt{\rho(x, a)}}\right) \left(1 + c \frac{\rho(z, y)}{\sqrt{\rho(y, a)}}\right)$$

or, equivalently,

$$\frac{\rho(x, y)}{\sqrt{\rho(x, a)\rho(y, a)}} \leq \frac{\rho(x, z)}{\sqrt{\rho(x, a)}} + \frac{\rho(z, y)}{\sqrt{\rho(y, a)}} + c \frac{\rho(x, z)\rho(z, y)}{\sqrt{\rho(x, a)\rho(y, a)}}. \tag{4.5}$$

Due to symmetry in x and y , we can assume that $\rho(x, a) \leq \rho(y, a)$. Since $c \geq \frac{3}{2}$, it suffices to show that

$$\frac{\rho(x, y)}{\sqrt{\rho(x, a)\rho(y, a)}} \leq \frac{\rho(x, z)}{\sqrt{\rho(x, a)}} + \frac{\rho(z, y)}{\sqrt{\rho(y, a)}} + \frac{3}{2} \cdot \frac{\rho(x, z)\rho(z, y)}{\sqrt{\rho(x, a)\rho(y, a)}} \tag{4.6}$$

or, equivalently,

$$\frac{\rho(x, y)}{\rho(x, z)\rho(y, z)} - \frac{\sqrt{\rho(y, a)}}{\rho(y, z)} - \frac{\sqrt{\rho(x, a)}}{\rho(x, z)} \leq \frac{3}{2}. \tag{4.7}$$

If $\rho(x, y) \leq \rho(x, z)$, then since $1 = \rho(z, a) \leq \rho(y, a) + \rho(y, z)$, we have

$$\frac{\rho(x, y)}{\rho(x, z)\rho(y, z)} - \frac{\sqrt{\rho(y, a)}}{\rho(y, z)} \leq \frac{1 - \sqrt{\rho(y, a)}}{\rho(y, z)} \leq \frac{1 - \sqrt{\rho(y, a)}}{1 - \rho(y, a)} = \frac{1}{1 + \sqrt{\rho(y, a)}} \leq 1$$

so that inequality (4.7) holds. Hence we can further assume that $\rho(x, y) > \rho(x, z)$.

Since $\rho(x, y) \leq \rho(x, z) + \rho(y, z)$, it suffices to show that

$$\frac{1}{\rho(x, z)} + \frac{1}{\rho(y, z)} - \frac{\sqrt{\rho(y, a)}}{\rho(y, z)} - \frac{\sqrt{\rho(x, a)}}{\rho(x, z)} \leq \frac{3}{2}$$

or, equivalently,

$$\frac{1 - \sqrt{\rho(x, a)}}{\rho(x, z)} + \frac{1 - \sqrt{\rho(y, a)}}{\rho(y, z)} \leq \frac{3}{2}. \quad (4.8)$$

Since $\rho(x, z) + \rho(x, a) \geq \rho(z, a) = 1$ and $\rho(y, z) + \rho(y, a) \geq \rho(z, a) = 1$ we have

$$\rho(x, z) \geq 1 - \rho(x, a) \quad \text{and} \quad \rho(y, z) \geq 1 - \rho(y, a).$$

Consequently, we have

$$\frac{1 - \sqrt{\rho(x, a)}}{\rho(x, z)} + \frac{1 - \sqrt{\rho(y, a)}}{\rho(y, z)} \leq \frac{1}{1 + \sqrt{\rho(x, a)}} + \frac{1}{1 + \sqrt{\rho(y, a)}}.$$

Since

$$\frac{1}{1 + \sqrt{\rho(x, a)}} + \frac{1}{1 + \sqrt{\rho(y, a)}} = \frac{2 + \sqrt{\rho(x, a)} + \sqrt{\rho(y, a)}}{1 + \sqrt{\rho(x, a)} + \sqrt{\rho(y, a)} + \sqrt{\rho(x, a)\rho(y, a)}},$$

it suffices to show that

$$\frac{2 + \sqrt{\rho(x, a)} + \sqrt{\rho(y, a)}}{1 + \sqrt{\rho(x, a)} + \sqrt{\rho(y, a)} + \sqrt{\rho(x, a)\rho(y, a)}} \leq \frac{3}{2}$$

or, equivalently,

$$\sqrt{\rho(x, a)} + \sqrt{\rho(y, a)} + 3\sqrt{\rho(x, a)\rho(y, a)} \geq 1. \quad (4.9)$$

Therefore, it suffices to show that (4.9) holds. We shall prove it by way of contradiction. That is, suppose

$$\sqrt{\rho(x, a)} + \sqrt{\rho(y, a)} + 3\sqrt{\rho(x, a)\rho(y, a)} < 1. \quad (4.10)$$

Clearly, $\rho(y, a) < 1$. Inequality (4.10) implies $2\sqrt{\rho(x, a)} + 3\rho(x, a) < 1$ or, equivalently, $\rho(x, a) < \frac{1}{9}$. Put $\rho(x, a) = t$. Since $1 = \rho(z, a) \leq \rho(x, z) + \rho(x, a)$, we have $\rho(x, z) \geq 1 - t$. Since

$$1 - t \leq \rho(x, z) < \rho(x, y) \leq \rho(x, a) + \rho(y, a),$$

we obtain $\rho(y, a) > 1 - 2t$. Since

$$\begin{aligned} 1 &> \sqrt{\rho(x, a)} + \sqrt{\rho(y, a)} + 3\sqrt{\rho(x, a)\rho(y, a)} \\ &> \rho(x, a) + \rho(y, a) + 3\sqrt{1 - 2t}\sqrt{\rho(x, a)} \\ &\geq \rho(x, y) + 3\sqrt{1 - 2t}\sqrt{\rho(x, a)} > 1 - t + 3\sqrt{1 - 2t}\sqrt{\rho(x, a)}, \end{aligned}$$

we obtain

$$t = \rho(x, a) < \frac{t^2}{9(1-2t)} \leq \frac{t}{7},$$

which is a required contradiction. \square

Theorem 4.11. *Let (\mathcal{X}, ρ) be any metric space and let $c \geq 1$. Then for each $a \in \mathcal{X}$, the semimetric $\tilde{\tau}_{c,a}$ is δ -hyperbolic with $\delta = \log(2c + 1)$.*

Proof. Given $x, y \in \mathcal{X}_a$, by the triangle inequality, we have

$$\sqrt{\rho(x, a)\rho(y, a)} + c\rho(x, y) \geq \rho(x, a) \wedge \rho(y, a) + \rho(x, y) \geq \rho(x, a) \vee \rho(y, a)$$

and

$$\begin{aligned} \sqrt{\rho(x, a)\rho(y, a)} + c\rho(x, y) &\leq \rho(x, a) \vee \rho(y, a) + 2c[\rho(x, a) \vee \rho(y, a)] \\ &= (2c + 1)[\rho(x, a) \vee \rho(y, a)]. \end{aligned}$$

Hence

$$P_a(x, y) \leq \exp(\tilde{\tau}_{c,a}(x, y)) \leq (2c + 1)P_a(x, y), \quad (4.12)$$

where

$$P_a(x, y) = \frac{\rho(x, a) \vee \rho(y, a)}{\sqrt{\rho(x, a)\rho(y, a)}}.$$

Now let $x, y, z, w \in \mathcal{X}_a$ be arbitrary points. Lemma 2.1 implies

$$P_a(x, y)P_a(z, w) \leq P_a(x, z)P_a(y, w) \vee P_a(x, w)P_a(y, z).$$

Consequently, using (4.12) we have

$$\begin{aligned} &\exp(\tilde{\tau}_{c,a}(x, y) + \tilde{\tau}_{c,a}(z, w)) \\ &\leq (2c + 1)^2 [P_a(x, y)P_a(z, w)] \\ &\leq (2c + 1)^2 [P_a(x, z)P_a(y, w) \vee P_a(x, w)P_a(y, z)] \\ &\leq (2c + 1)^2 [\exp(\tilde{\tau}_{c,a}(x, z) + \tilde{\tau}_{c,a}(y, w)) \vee \exp(\tilde{\tau}_{c,a}(x, w) + \tilde{\tau}_{c,a}(y, z))] \end{aligned}$$

or, equivalently,

$$\begin{aligned} &\tilde{\tau}_{c,a}(x, y) + \tilde{\tau}_{c,a}(z, w) \\ &\leq [\tilde{\tau}_{c,a}(x, z) + \tilde{\tau}_{c,a}(y, w)] \vee [\tilde{\tau}_{c,a}(x, w) + \tilde{\tau}_{c,a}(y, z)] + 2\log(2c + 1), \end{aligned}$$

completing the proof. \square

Theorem 4.3 suggests that a natural category of spaces on which to consider the family of semimetrics $\tilde{\tau}_{c,a}$ is the category of metric spaces. Moreover, in such a setting one should consider the parameter c to be greater than or equal to $\frac{3}{2}$. Now, if (\mathcal{X}, ρ) is a metric space and $c \geq \frac{3}{2}$, then for all $a \in \mathcal{X}$ and $x, y \in \mathcal{X}_a$, we have

$$\tilde{\tau}_{3/2,a}(x, y) \leq \tilde{\tau}_{c,a}(x, y) \leq \frac{2c}{3} \tilde{\tau}_{3/2,a}(x, y)$$

and

$$\tilde{\tau}_{3/2,a}(x, y) \leq \tilde{\tau}_{c,a}(x, y) \leq \tilde{\tau}_{3/2,a}(x, y) + \log \frac{2c}{3}.$$

That is, the metrics $\tilde{\tau}_{3/2,a}$ and $\tilde{\tau}_{c,a}$ are geometrically equivalent. Therefore, we shall consider only the metric $\tilde{\tau}_{3/2,a}$, denoted simply by $\tilde{\tau}_a$,

$$\tilde{\tau}_a(x, y) = \log \left(1 + \frac{3}{2} \frac{\rho(x, y)}{\sqrt{\rho(x, a)\rho(y, a)}} \right). \quad (4.13)$$

We shall refer to $\tilde{\tau}_a$ as a *one-point, scale-invariant Cassinian metric*. For subdomains D of the Euclidean space \mathbb{R}^n the metric $\tilde{\tau}_D$, called the *scale-invariant Cassinian metric* and defined by

$$\tilde{\tau}_D(x, y) = \sup\{\tilde{\tau}_{1,a}(x, y) : a \in \partial D\},$$

was introduced in [Ibragimov 2016]. For a general metric space (\mathcal{X}, ρ) , the semimetric $\tilde{\tau}_{1,a}$ was shown to be a metric provided the metric ρ also satisfies Ptolemy's inequality; see [Aksoy et al. 2018, Theorem 2.3]. The metric $\tilde{\tau}_{2,a}$ was first considered in [Aksoy et al. 2018], where it was shown to be δ -hyperbolic with $\delta = \log 6$; see Lemma 4.1 therein. Theorem 4.11 implies that $\tilde{\tau}_{2,a}$ is, in fact, δ -hyperbolic with $\delta = \log 5$.

The following theorem is an immediate consequence of Theorems 4.3 and 4.11.

Theorem 4.14. *Let (\mathcal{X}, ρ) be any metric space. Then for each $a \in \mathcal{X}$, the space $(\mathcal{X} \setminus \{a\}, \tilde{\tau}_a)$ is a δ -hyperbolic metric space with $\delta = \log 4$.*

We end this section by showing that the metrics \tilde{j}_a and $\tilde{\tau}_a$ are geometrically equivalent.

Theorem 4.15. *Let (\mathcal{X}, ρ) be any metric space and let $a \in \mathcal{X}$. Then for all $x, y \in \mathcal{X}_a$ we have*

$$\frac{1}{2}\tilde{j}_a(x, y) \leq \tilde{\tau}_a(x, y) \leq \frac{3}{2}\tilde{j}_a(x, y) \quad (4.16)$$

and

$$\frac{1}{2}\tilde{j}_a(x, y) \leq \tilde{\tau}_a(x, y) \leq \frac{1}{2}\tilde{j}_a(x, y) + \log \frac{4}{\sqrt{3}}. \quad (4.17)$$

The inequalities are sharp.

Proof. By the triangle inequality, we have

$$\frac{\rho(x, a)\rho(y, a)}{\rho(x, a) \wedge \rho(y, a)} = \rho(x, a) \vee \rho(y, a) \leq \sqrt{\rho(x, a)\rho(y, a)} + \rho(x, y),$$

which implies

$$\frac{\rho(x, y)}{\rho(x, a) \wedge \rho(y, a)} \leq \frac{\rho(x, y)}{\sqrt{\rho(x, a)\rho(y, a)}} + \frac{\rho(x, y)^2}{\rho(x, a)\rho(y, a)}.$$

Hence

$$\tilde{j}_a(x, y) \leq \log \left(1 + \frac{\rho(x, y)}{\sqrt{\rho(x, a)\rho(y, a)}} + \frac{\rho(x, y)^2}{\rho(x, a)\rho(y, a)} \right) \leq 2\tilde{\tau}_a(x, y)$$

so that the first inequality in (4.16) holds. The latter is sharp when $\rho(x, y) = \rho(x, a) + \rho(y, a) = 1$ and $\rho(x, a) \rightarrow 0$.

Since $\rho(x, a) \wedge \rho(y, a) \leq \sqrt{\rho(x, a)\rho(y, a)}$, we have

$$\tilde{\tau}_a(x, y) \leq \log \left(1 + \frac{3}{2} \frac{\rho(x, y)}{\rho(x, a) \wedge \rho(y, a)} \right) \leq \frac{3}{2} \log \left(1 + \frac{\rho(x, y)}{\rho(x, a) \wedge \rho(y, a)} \right) = \frac{3}{2} \tilde{j}_a(x, y),$$

so that the second inequality in (4.16) also holds. The latter is sharp when $\rho(x, a) = \rho(y, a)$, $\rho(x, y) = \rho(x, a)^2$ and $\rho(x, a) \rightarrow 0$.

To show the second inequality in (4.17) we can assume, without loss of generality, that $\rho(x, a) \leq \rho(y, a)$. Using the triangle inequality one can verify that

$$\left(\frac{3}{2} \rho(x, y) \right)^2 \leq \frac{13}{3} \rho(x, a)\rho(y, a) + \frac{7}{3} \rho(x, y)\rho(y, a). \quad (4.18)$$

Indeed, (4.18) is equivalent to

$$27\rho^2(x, y) \leq 52\rho(x, a)\rho(y, a) + 28\rho(x, y)\rho(y, a).$$

To prove this, we note that $\rho(x, y) \leq \rho(x, a) + \rho(y, a) \leq 2\rho(y, a)$ and also that $\rho(x, y)\rho(x, a) \leq \rho(x, y)\rho(y, a)$. Combining these inequalities gives

$$27\rho(x, y)\rho(x, a) \leq 52\rho(x, a)\rho(y, a) + \rho(x, y)\rho(y, a),$$

and hence

$$\begin{aligned} 27\rho^2(x, y) &\leq 27\rho(x, y)\rho(x, a) + 27\rho(x, y)\rho(y, a) \\ &\leq 52\rho(x, a)\rho(y, a) + \rho(x, y)\rho(y, a) + 27\rho(x, y)\rho(x, a), \end{aligned}$$

so that (4.18) holds.

Now, inequality (4.18) is equivalent to

$$\frac{9}{4} \frac{\rho(x, y)^2}{\rho(x, a)\rho(y, a)} \leq \frac{13}{3} + \frac{7}{3} \frac{\rho(x, y)}{\rho(x, a)} = \frac{13}{3} + \frac{7}{3} \frac{\rho(x, y)}{\rho(x, a) \wedge \rho(y, a)}.$$

Since $\rho(x, a) \leq \rho(y, a)$, we also have

$$3 \frac{\rho(x, y)}{\sqrt{\rho(x, a)\rho(y, a)}} \leq 3 \frac{\rho(x, y)}{\rho(x, a) \wedge \rho(y, a)}.$$

Hence

$$1 + 3 \frac{\rho(x, y)}{\sqrt{\rho(x, a)\rho(y, a)}} + \frac{9}{4} \frac{\rho(x, y)^2}{\rho(x, a)\rho(y, a)} \leq \frac{16}{3} \left(1 + \frac{\rho(x, y)}{\rho(x, a) \wedge \rho(y, a)} \right),$$

implying the second inequality in (4.17). The latter is sharp when $\rho(x, a) = \rho(y, a)$ and $\rho(x, y) = \rho(x, a) + \rho(y, a)$. \square

5. A one-point, half-Apollonian pseudometric

Let (\mathcal{X}, ρ) be a semimetric space and let $a \in \mathcal{X}$ be an arbitrary point. Set $\mathcal{X}_a = \mathcal{X} \setminus \{a\}$. For $x, y \in \mathcal{X}_a$, define $\tilde{\alpha}_a$ by

$$\tilde{\alpha}_a(x, y) = \left| \log \frac{\rho(x, a)}{\rho(y, a)} \right| = \log \frac{\rho(x, a) \vee \rho(y, a)}{\rho(x, a) \wedge \rho(y, a)} = 2 \log \frac{\rho(x, a) \vee \rho(y, a)}{\sqrt{\rho(x, a)\rho(y, a)}}. \quad (5.1)$$

We have $\tilde{\alpha}_a(x, y) \geq 0$, $\tilde{\alpha}_a(x, y) = \tilde{\alpha}_a(y, x)$, and $\tilde{\alpha}_a(x, x) = 0$ so that $\tilde{\alpha}_a$ is a distance function. The next theorem shows that $\tilde{\alpha}_a$ satisfies the δ -hyperbolicity condition with $\delta = 0$. Consequently, it also satisfies the triangle inequality.

Theorem 5.2. *Let (\mathcal{X}, ρ) be any semimetric space and let $a \in \mathcal{X}$. Then for all $x, y, z, w \in \mathcal{X}_a$, we have*

$$\tilde{\alpha}_a(x, y) + \tilde{\alpha}_a(z, w) \leq [\tilde{\alpha}_a(x, z) + \tilde{\alpha}_a(y, w)] \vee [\tilde{\alpha}_a(x, w) + \tilde{\alpha}_a(y, z)]. \quad (5.3)$$

In particular, for all $x, y, z \in \mathcal{X}_a$, we have

$$\tilde{\alpha}_a(x, y) \leq \tilde{\alpha}_a(x, z) + \tilde{\alpha}_a(z, y). \quad (5.4)$$

Proof. Using (5.1) we see that (5.3) is equivalent to

$$\begin{aligned} & [\rho(x, a) \vee \rho(y, a)][\rho(z, a) \vee \rho(w, a)] \\ & \leq ([\rho(x, a) \vee \rho(z, a)][\rho(y, a) \vee \rho(w, a)]) \vee ([\rho(x, a) \vee \rho(w, a)][\rho(y, a) \vee \rho(z, a)]). \end{aligned}$$

The latter follows from Lemma 2.1, completing the proof of (5.3). Putting $w = z$ in (5.3), we obtain (5.4). \square

Remark 5.5. Note that if $\rho(x, a) = \rho(y, a)$ for some $x \neq y$, then $\tilde{\alpha}_a(x, y) = 0$. Hence $\tilde{\alpha}_a$ is, in general, a pseudometric on \mathcal{X}_a . Also, we have

$$\tilde{\alpha}_a(x, y) = \tilde{\alpha}_a(x, z) + \tilde{\alpha}_a(z, y)$$

if and only if $\rho(x, a) \leq \rho(z, a) \leq \rho(y, a)$ or $\rho(y, a) \leq \rho(z, a) \leq \rho(x, a)$.

We shall refer to $\tilde{\alpha}_a$ as a *one-point, half-Apollonian pseudometric*. For proper subdomains D of the Euclidean space \mathbb{R}^n , the (pseudo-)metric $\tilde{\alpha}_D$, called the *half-Apollonian metric* and defined by

$$\tilde{\alpha}_D(x, y) = \sup\{\tilde{\alpha}_a(x, y) : a \in \partial D\},$$

was introduced by P. Hästö and H. Lindén [2004].

Next, we show that $\tilde{\alpha}_a$ is roughly similar to both the one-point \tilde{j} -metric \tilde{j}_a and the one-point, scale-invariant Cassinian metric $\tilde{\tau}_a$ when (\mathcal{X}, ρ) is a metric space.

Theorem 5.6. *Let (\mathcal{X}, ρ) be any metric space and let $a \in \mathcal{X}$. Then for all $x, y \in \mathcal{X}_a$ we have*

$$\tilde{\alpha}_a(x, y) \leq \tilde{j}_a(x, y) \leq \tilde{\alpha}_a(x, y) + \log 3 \quad (5.7)$$

and

$$\frac{1}{2}\tilde{\alpha}_a(x, y) \leq \tilde{\tau}_a(x, y) \leq \frac{1}{2}\tilde{\alpha}_a(x, y) + \log 4. \quad (5.8)$$

The inequalities are sharp.

Proof. By the triangle inequality, we have

$$\rho(x, a) \vee \rho(y, a) \leq \rho(x, a) \wedge \rho(y, a) + \rho(x, y) \leq 3[\rho(x, a) \vee \rho(y, a)]$$

so that

$$\begin{aligned} \tilde{\alpha}_a(x, y) &= \log \frac{\rho(x, a) \vee \rho(y, a)}{\rho(x, a) \wedge \rho(y, a)} \leq \log \left(1 + \frac{\rho(x, y)}{\rho(x, a) \wedge \rho(y, a)} \right) \\ &= \tilde{j}_a(x, y) = \log \left(\frac{\rho(x, a) \wedge \rho(y, a) + \rho(x, y)}{\rho(x, a) \wedge \rho(y, a)} \right) \leq \tilde{\alpha}_a(x, y) + \log 3, \end{aligned}$$

completing the proof of (5.7). The first inequality in (5.7) is sharp when $\rho(x, a) = \rho(y, a) + \rho(x, y)$. The second inequality in (5.7) is sharp when $\rho(x, a) = \rho(y, a) = \rho(x, y)/2$.

To prove (5.8), we use the triangle inequality to obtain

$$\begin{aligned} \rho(x, a) \vee \rho(y, a) &\leq \rho(x, a) \wedge \rho(y, a) + \rho(x, y) \\ &\leq \sqrt{\rho(x, a)\rho(y, a)} + \frac{3}{2}\rho(x, y). \end{aligned}$$

Hence

$$\frac{1}{2}\tilde{\alpha}_a(x, y) = \log \frac{\rho(x, a) \vee \rho(y, a)}{\sqrt{\rho(x, a)\rho(y, a)}} \leq \log \left(1 + \frac{3}{2} \frac{\rho(x, y)}{\sqrt{\rho(x, a)\rho(y, a)}} \right) = \tilde{\tau}_a(x, y).$$

Also, since $\sqrt{\rho(x, a)\rho(y, a)} = \sqrt{\rho(x, a) \vee \rho(y, a)} \sqrt{\rho(x, a) \wedge \rho(y, a)}$ and

$$\begin{aligned} \sqrt{\rho(x, a)\rho(y, a)} + \frac{3}{2}\rho(x, y) &\leq \rho(x, a) \vee \rho(y, a) + 3[\rho(x, a) \vee \rho(y, a)] \\ &= 4[\rho(x, a) \vee \rho(y, a)], \end{aligned}$$

we have

$$\begin{aligned} \tilde{\tau}_a(x, y) &= \log \frac{\sqrt{\rho(x, a)\rho(y, a)} + \frac{3}{2}\rho(x, y)}{\sqrt{\rho(x, a)\rho(y, a)}} \\ &\leq \log 4 \left(\frac{\rho(x, a) \vee \rho(y, a)}{\rho(x, a) \wedge \rho(y, a)} \right)^{1/2} = \frac{1}{2}\tilde{\alpha}_a(x, y) + \log 4, \end{aligned}$$

completing the proof of (5.8). The first inequality in (5.8) is sharp when $\rho(x, a) = \rho(y, a) + \rho(x, y)$ and $\rho(x, y) \rightarrow 0$. The second inequality in (5.8) is sharp when $\rho(x, a) = \rho(y, a) = \rho(x, y)/2$. \square

6. One-point hyperbolic-type metrics on locally compact noncomplete metric spaces

In this section we define the \tilde{j} -metric, the scale-invariant Cassinian metric and the half-Apollonian metric on locally compact, noncomplete metric spaces. Recall that a metric space (\mathcal{X}, ρ) is called locally compact if, for each $x \in \mathcal{X}$, there exist an open set \mathcal{U} and a compact set \mathcal{V} such that $x \in \mathcal{V} \subset \mathcal{U} \subset \mathcal{X}$. Suppose that (\mathcal{X}, ρ) is a locally compact, noncomplete metric space. Let $\bar{\mathcal{X}}$ be its metric completion and let $\partial\mathcal{X} = \bar{\mathcal{X}} \setminus \mathcal{X}$. For $x, y \in \mathcal{X}$, we define

$$\tilde{j}_{\mathcal{X}}(x, y) = \sup\{\tilde{j}_a(x, y) : a \in \partial\mathcal{X}\} = \sup_{a \in \partial\mathcal{X}} \log\left(1 + \frac{\rho(x, y)}{\rho(x, a) \wedge \rho(y, a)}\right), \quad (6.1)$$

$$\tilde{\tau}_{\mathcal{X}}(x, y) = \sup\{\tilde{\tau}_a(x, y) : a \in \partial\mathcal{X}\} = \sup_{a \in \partial\mathcal{X}} \log\left(1 + \frac{3}{2} \frac{\rho(x, y)}{\sqrt{\rho(x, a) \rho(y, a)}}\right), \quad (6.2)$$

$$\tilde{\alpha}_{\mathcal{X}}(x, y) = \sup\{\tilde{\alpha}_a(x, y) : a \in \partial\mathcal{X}\} = \sup_{a \in \partial\mathcal{X}} \log \frac{\rho(x, a) \vee \rho(y, a)}{\rho(x, a) \wedge \rho(y, a)}. \quad (6.3)$$

Observe that since \mathcal{X} is locally compact, we have $\inf\{\rho(x, a) : a \in \partial\mathcal{X}\} > 0$ for each $x \in \mathcal{X}$ and, consequently, the quantities $\tilde{j}_{\mathcal{X}}(x, y)$, $\tilde{\tau}_{\mathcal{X}}(x, y)$, and $\tilde{\alpha}_{\mathcal{X}}(x, y)$ are finite for each $x, y \in \mathcal{X}$. It is also a simple observation that all the functions $\tilde{j}_{\mathcal{X}}$, $\tilde{\tau}_{\mathcal{X}}$ and $\tilde{\alpha}_{\mathcal{X}}$ are distance functions and, moreover, the distance functions $\tilde{j}_{\mathcal{X}}$ and $\tilde{\tau}_{\mathcal{X}}$ are semimetrics. Next, we show that each of the distance functions satisfies the triangle inequality. Let $x, y, z \in \mathcal{X}$ be any points. Given arbitrary $\epsilon > 0$, there exists $a \in \partial\mathcal{X}$ such that $\tilde{j}_{\mathcal{X}}(x, y) < \tilde{j}_a(x, y) + \epsilon$. Hence

$$\tilde{j}_{\mathcal{X}}(x, y) < \tilde{j}_a(x, y) + \epsilon \leq \tilde{j}_a(x, z) + \tilde{j}_a(z, y) + \epsilon \leq \tilde{j}_{\mathcal{X}}(x, z) + \tilde{j}_{\mathcal{X}}(z, y) + \epsilon.$$

Since ϵ is arbitrary, we obtain $\tilde{j}_{\mathcal{X}}(x, y) \leq \tilde{j}_{\mathcal{X}}(x, z) + \tilde{j}_{\mathcal{X}}(z, y)$. In a similar fashion, we show that $\tilde{\tau}_{\mathcal{X}}(x, y) \leq \tilde{\tau}_{\mathcal{X}}(x, z) + \tilde{\tau}_{\mathcal{X}}(z, y)$ and $\tilde{\alpha}_{\mathcal{X}}(x, y) \leq \tilde{\alpha}_{\mathcal{X}}(x, z) + \tilde{\alpha}_{\mathcal{X}}(z, y)$, as required. Hence we obtain the following theorem.

Theorem 6.4. *Let (\mathcal{X}, ρ) be a locally compact, noncomplete metric space. Then the distance functions $\tilde{j}_{\mathcal{X}}$ and $\tilde{\tau}_{\mathcal{X}}$ are metrics on \mathcal{X} and the distance function $\tilde{\alpha}_{\mathcal{X}}$ is a pseudometric on \mathcal{X} .*

We shall refer to $\tilde{j}_{\mathcal{X}}$, $\tilde{\tau}_{\mathcal{X}}$ and $\tilde{\alpha}_{\mathcal{X}}$ as the \tilde{j} -metric, the scale-invariant Cassinian metric and the half-Apollonian metric on \mathcal{X} , respectively. Note that even though the distance function $\tilde{\alpha}_{\mathcal{X}}$ is, in general, a pseudometric, we shall refer to it as a metric since it is a true metric under a mild condition on \mathcal{X} , namely, when \mathcal{X} has a *one-point separation property*. We say that a locally compact noncomplete metric space (\mathcal{X}, ρ) has a one-point separation property if for each $x, y \in \mathcal{X}$ there exists $a \in \partial\mathcal{X}$ such that $\rho(x, a) \neq \rho(y, a)$. It follows that if \mathcal{X} has the one-point

separation property, then the distance function $\tilde{\alpha}_{\mathcal{X}}$ is a metric, justifying the term “half-Apollonian metric”.

Next, we show that the same relations between the one-point distance functions \tilde{j}_a , $\tilde{\tau}_a$ and $\tilde{\alpha}_a$ established in Theorems 4.15 and 5.6 continue to hold for the distance functions $\tilde{j}_{\mathcal{X}}$, $\tilde{\tau}_{\mathcal{X}}$ and $\tilde{\alpha}_{\mathcal{X}}$.

Theorem 6.5. *Let (\mathcal{X}, ρ) be a locally compact, noncomplete metric space. Then for each $x, y \in \mathcal{X}$ we have*

$$\frac{1}{2}\tilde{j}_{\mathcal{X}}(x, y) \leq \tilde{\tau}_{\mathcal{X}}(x, y) \leq \frac{3}{2}\tilde{j}_{\mathcal{X}}(x, y), \tag{6.6}$$

$$\frac{1}{2}\tilde{j}_{\mathcal{X}}(x, y) \leq \tilde{\tau}_{\mathcal{X}}(x, y) \leq \frac{1}{2}\tilde{j}_{\mathcal{X}}(x, y) + \log \frac{4}{\sqrt{3}}, \tag{6.7}$$

$$\tilde{\alpha}_{\mathcal{X}}(x, y) \leq \tilde{j}_{\mathcal{X}}(x, y) \leq \tilde{\alpha}_{\mathcal{X}}(x, y) + \log 3, \tag{6.8}$$

$$\frac{1}{2}\tilde{\alpha}_{\mathcal{X}}(x, y) \leq \tilde{\tau}_{\mathcal{X}}(x, y) \leq \frac{1}{2}\tilde{\alpha}_{\mathcal{X}}(x, y) + \log 4. \tag{6.9}$$

Moreover, all the inequalities above are sharp.

Proof. Let $\epsilon > 0$ be arbitrary. Then there exists $a \in \partial \mathcal{X}$ such that

$$\frac{1}{2}\tilde{j}_{\mathcal{X}}(x, y) < \frac{1}{2}\tilde{j}_a(x, y) + \epsilon.$$

By Theorem 4.15 we have

$$\frac{1}{2}\tilde{j}_{\mathcal{X}}(x, y) < \frac{1}{2}\tilde{j}_a(x, y) + \epsilon \leq \tilde{\tau}_a(x, y) + \epsilon \leq \tilde{\tau}_{\mathcal{X}}(x, y) + \epsilon.$$

Since ϵ is arbitrary we have $\frac{1}{2}\tilde{j}_{\mathcal{X}}(x, y) \leq \tilde{\tau}_{\mathcal{X}}(x, y)$, completing the proof of the first inequality in (6.6). The remaining inequalities are proved in a similar fashion. Finally, sharpness of the inequalities follows from the sharpness parts in Theorems 4.15 and 5.6. □

Remark 6.10. The \tilde{j} -metric $\tilde{j}_{\mathcal{X}}$ is not δ -hyperbolic if the set $\partial \mathcal{X}$ contains more than one point [Hästö 2006]. Inequalities (6.7) and (6.8) imply that the same is true for the scale-invariant Cassinian metric $\tilde{\tau}_{\mathcal{X}}$ and the half-Apollonian metric $\tilde{\alpha}_{\mathcal{X}}$.

We end the paper by establishing relations between the metrics $\tilde{j}_{\mathcal{X}}$, $\tilde{\tau}_{\mathcal{X}}$, $\tilde{\alpha}_{\mathcal{X}}$ and the metric $u_{\mathcal{X}}$ and the j -metric $j_{\mathcal{X}}$. The latter metrics were introduced in the context of locally compact, noncomplete metric spaces in [Ibragimov 2011]. Let (\mathcal{X}, ρ) be a locally compact, noncomplete metric space and let $\partial \mathcal{X} = \bar{\mathcal{X}} \setminus \mathcal{X}$. Then

$$u_{\mathcal{X}}(x, y) = 2 \log \frac{\rho(x, y) + \text{dist}(x, \partial \mathcal{X}) \vee \text{dist}(y, \partial \mathcal{X})}{\sqrt{\text{dist}(x, \partial \mathcal{X}) \text{dist}(y, \partial \mathcal{X})}}$$

and

$$j_{\mathcal{X}}(x, y) = \frac{1}{2} \log \left(1 + \frac{\rho(x, y)}{\text{dist}(x, \partial \mathcal{X})} \right) \left(1 + \frac{\rho(x, y)}{\text{dist}(y, \partial \mathcal{X})} \right).$$

Here $\text{dist}(x, \partial \mathcal{X}) = \inf\{\rho(x, a) : a \in \partial \mathcal{X}\}$ for $x \in \mathcal{X}$ [Ibragimov 2011]. We have

$$2j_{\mathcal{X}}(x, y) \leq u_{\mathcal{X}}(x, y) \leq 2j_{\mathcal{X}}(x, y) + 2 \log 2. \quad (6.11)$$

In particular, the metric $u_{\mathcal{X}}$ is δ -hyperbolic with $\delta = \log 4$ and the metric $j_{\mathcal{X}}$ is δ -hyperbolic with $\delta = \frac{5}{2} \log 2$ [Ibragimov 2011, Theorems 2.7 and 3.2].

The j -metric was first introduced in [Gehring and Osgood 1979; Gehring and Palka 1976] for domains in the Euclidean space \mathbb{R}^n . It is well known that the j -metric and the \tilde{j} -metric are bilipschitz equivalent. More precisely, if D is a proper subdomain of \mathbb{R}^n , then

$$\frac{1}{2} \tilde{j}_D(x, y) \leq j_D(x, y) \leq \tilde{j}_D(x, y) \quad (6.12)$$

for all $x, y \in D$; see, for example, [Vuorinen 1988]. Next, we show that inequality (6.12) extends to the settings of locally compact, noncomplete metric spaces.

Theorem 6.13. *Let (\mathcal{X}, ρ) be a locally compact, noncomplete metric space. Then for all $x, y \in \mathcal{X}$, we have*

$$\frac{1}{2} \tilde{j}_{\mathcal{X}}(x, y) \leq j_{\mathcal{X}}(x, y) \leq \tilde{j}_{\mathcal{X}}(x, y). \quad (6.14)$$

Moreover,

$$\tilde{j}_{\mathcal{X}}(x, y) \leq u_{\mathcal{X}}(x, y) \leq 2\tilde{j}_{\mathcal{X}}(x, y) + 2 \log 2, \quad (6.15)$$

$$\frac{1}{3} \tilde{\tau}_{\mathcal{X}}(x, y) \leq j_{\mathcal{X}}(x, y) \leq 2\tilde{\tau}_{\mathcal{X}}(x, y), \quad (6.16)$$

$$\frac{2}{3} \tilde{\tau}_{\mathcal{X}}(x, y) \leq u_{\mathcal{X}}(x, y) \leq 4\tilde{\tau}_{\mathcal{X}}(x, y) + 2 \log 2, \quad (6.17)$$

$$\frac{1}{2} \tilde{\alpha}_{\mathcal{X}}(x, y) \leq j_{\mathcal{X}}(x, y) \leq \tilde{\alpha}_{\mathcal{X}}(x, y) + \log 3, \quad (6.18)$$

$$\tilde{\alpha}_{\mathcal{X}}(x, y) \leq u_{\mathcal{X}}(x, y) \leq 2\tilde{\alpha}_{\mathcal{X}}(x, y) + 2 \log 6. \quad (6.19)$$

Proof. First, we show that inequality (6.14) holds. Given $x, y \in \mathcal{X}$, let $a, b \in \partial \mathcal{X}$ be such that $\text{dist}(x, \partial \mathcal{X}) = \rho(x, a)$ and $\text{dist}(y, \partial \mathcal{X}) = \rho(y, b)$. Without loss of generality, we can assume that $\rho(x, a) \leq \rho(y, b)$. Then $\rho(x, a) \leq \rho(y, b) \leq \rho(y, a)$ and hence

$$\begin{aligned} \log \left(1 + \frac{\rho(x, y)}{\rho(x, a) \wedge \rho(y, a)} \right) &= \log \left(1 + \frac{\rho(x, y)}{\rho(x, a)} \right) \leq \log \left(1 + \frac{\rho(x, y)}{\rho(x, a)} \right) \left(1 + \frac{\rho(x, y)}{\rho(y, b)} \right) \\ &= \log \left(1 + \frac{\rho(x, y)}{\text{dist}(x, \partial \mathcal{X})} \right) \left(1 + \frac{\rho(x, y)}{\text{dist}(y, \partial \mathcal{X})} \right) \\ &= 2j_{\mathcal{X}}(x, y). \end{aligned}$$

Hence

$$\tilde{j}_{\mathcal{X}}(x, y) = \sup_{a \in \partial \mathcal{X}} \log \left(1 + \frac{\rho(x, y)}{\rho(x, a) \wedge \rho(y, a)} \right) \leq 2j_{\mathcal{X}}(x, y),$$

proving the first inequality in (6.14). Again, since $\rho(x, a) \leq \rho(y, b) \leq \rho(y, a)$, we have

$$\begin{aligned} j_{\mathcal{X}}(x, y) &= \frac{1}{2} \log \left(1 + \frac{\rho(x, y)}{\text{dist}(x, \partial \mathcal{X})} \right) \left(1 + \frac{\rho(x, y)}{\text{dist}(y, \partial \mathcal{X})} \right) \\ &= \frac{1}{2} \log \left(1 + \frac{\rho(x, y)}{\rho(x, a)} \right) \left(1 + \frac{\rho(x, y)}{\rho(y, b)} \right) \leq \log \left(1 + \frac{\rho(x, y)}{\rho(x, a)} \right) \\ &= \log \left(1 + \frac{\rho(x, y)}{\rho(x, a) \wedge \rho(y, a)} \right) \leq \sup_{a \in \partial \mathcal{X}} \log \left(1 + \frac{\rho(x, y)}{\rho(x, a) \wedge \rho(y, a)} \right) \\ &= \tilde{j}_{\mathcal{X}}(x, y), \end{aligned}$$

proving the second inequality in (6.14). The remaining inequalities immediately follow from Theorem 6.5 and inequalities (6.11) and (6.14). \square

References

- [Aksoy et al. 2018] A. G. Aksoy, Z. Ibragimov, and W. Whiting, “Averaging one-point hyperbolic-type metrics”, *Proc. Amer. Math. Soc.* **146**:12 (2018), 5205–5218. MR Zbl
- [Beardon and Minda 2007] A. F. Beardon and D. Minda, “The hyperbolic metric and geometric function theory”, pp. 9–56 in *Quasiconformal mappings and their applications*, edited by S. Ponnusamy et al., Narosa, New Delhi, 2007. MR Zbl
- [Bonk 2006] M. Bonk, “Quasiconformal geometry of fractals”, pp. 1349–1373 in *International Congress of Mathematicians, Vol. II*, edited by M. Sanz-Solé et al., Eur. Math. Soc., Zürich, 2006. MR Zbl
- [Bonk and Schramm 2000] M. Bonk and O. Schramm, “Embeddings of Gromov hyperbolic spaces”, *Geom. Funct. Anal.* **10**:2 (2000), 266–306. MR Zbl
- [Bonk et al. 2001] M. Bonk, J. Heinonen, and P. Koskela, *Uniformizing Gromov hyperbolic spaces*, Astérisque **270**, Société Mathématique de France, Paris, 2001. MR Zbl
- [Dovgoshey et al. 2016] O. Dovgoshey, P. Hariri, and M. Vuorinen, “Comparison theorems for hyperbolic type metrics”, *Complex Var. Elliptic Equ.* **61**:11 (2016), 1464–1480. MR Zbl
- [Ferrand 1988] J. Ferrand, “A characterization of quasiconformal mappings by the behaviour of a function of three points”, pp. 110–123 in *Complex analysis, Joensuu 1987*, edited by I. Laine et al., Lecture Notes in Math. **1351**, Springer, 1988. MR Zbl
- [Gehring and Osgood 1979] F. W. Gehring and B. G. Osgood, “Uniform domains and the quasihyperbolic metric”, *J. Analyse Math.* **36** (1979), 50–74. MR Zbl
- [Gehring and Palka 1976] F. W. Gehring and B. P. Palka, “Quasiconformally homogeneous domains”, *J. Analyse Math.* **30** (1976), 172–199. MR Zbl
- [Gromov 1987] M. Gromov, “Hyperbolic groups”, pp. 75–263 in *Essays in group theory*, edited by S. M. Gersten, Math. Sci. Res. Inst. Publ. **8**, Springer, 1987. MR Zbl
- [Hästö 2006] P. A. Hästö, “Gromov hyperbolicity of the j_G and \tilde{j}_G metrics”, *Proc. Amer. Math. Soc.* **134**:4 (2006), 1137–1142. MR Zbl
- [Hästö and Lindén 2004] P. Hästö and H. Lindén, “Isometries of the half-Apollonian metric”, *Complex Var. Theory Appl.* **49**:6 (2004), 405–415. MR Zbl
- [Hästö et al. 2006] P. Hästö, Z. Ibragimov, and H. Lindén, “Isometries of relative metrics”, *Comput. Methods Funct. Theory* **6**:1 (2006), 15–28. MR Zbl

- [Herron 2016] D. A. Herron, “Universal convexity for quasihyperbolic type metrics”, *Conform. Geom. Dyn.* **20** (2016), 1–24. MR Zbl
- [Herron and Julian 2013] D. A. Herron and P. K. Julian, “Ferrand’s Möbius invariant metric”, *J. Anal.* **21** (2013), 101–121. MR Zbl
- [Ibragimov 2003] Z. Ibragimov, “On the Apollonian metric of domains in $\overline{\mathbb{R}^n}$ ”, *Complex Var. Theory Appl.* **48**:10 (2003), 837–855. MR Zbl
- [Ibragimov 2011] Z. Ibragimov, “Hyperbolizing metric spaces”, *Proc. Amer. Math. Soc.* **139**:12 (2011), 4401–4407. MR Zbl
- [Ibragimov 2016] Z. Ibragimov, “A scale-invariant Cassinian metric”, *J. Anal.* **24**:1 (2016), 111–129. MR Zbl
- [Kulkarni and Pinkall 1994] R. S. Kulkarni and U. Pinkall, “A canonical metric for Möbius structures and its applications”, *Math. Z.* **216**:1 (1994), 89–129. MR Zbl
- [Lindén 2007] H. Lindén, “Hyperbolic-type metrics”, pp. 151–164 in *Quasiconformal mappings and their applications*, edited by S. Ponnusamy et al., Narosa, New Delhi, 2007. MR Zbl
- [Väisälä 2005] J. Väisälä, “Gromov hyperbolic spaces”, *Expo. Math.* **23**:3 (2005), 187–231. MR Zbl
- [Vuorinen 1988] M. Vuorinen, *Conformal geometry and quasiregular mappings*, Lecture Notes in Mathematics **1319**, Springer, 1988. MR Zbl

Received: 2019-06-17 Revised: 2019-10-03 Accepted: 2019-11-04

mborovikova@fullerton.edu	<i>Department of Mathematics, California State University, Fullerton, CA, United States</i>
zibragimov@fullerton.edu	<i>Department of Mathematics, California State University, Fullerton, CA, United States</i>
mjim92@csu.fullerton.edu	<i>Department of Mathematics, California State University, Fullerton, CA, United States</i>
capitala677@csu.fullerton.edu	<i>Department of Mathematics, California State University, Fullerton, CA, United States</i>

Some generalizations of the ASR search algorithm for quasitwisted codes

Nuh Aydin, Thomas H. Guidotti, Peihan Liu,
Armiya S. Shaikh and Robert O. VandenBerg

(Communicated by Kenneth S. Berenhaut)

One of the most important and challenging problems in coding theory is explicit construction of linear codes with the best possible parameters. It is well known that the class of quasitwisted (QT) codes is asymptotically good and contains many linear codes with best known parameters (BKLCs). A search algorithm (ASR) on QT codes has been particularly effective to construct such codes. Recently, the ASR algorithm was generalized based on the notion of code equivalence. In this work, we introduce a new generalization of the ASR algorithm to include a broader scope of QT codes. As a result of implementing this algorithm, we have found eight new linear codes over the field \mathbb{F}_5 . Furthermore, we have found seven additional new codes from the standard constructions of puncturing, shortening or Construction X. We also introduce a new search algorithm that can be viewed as a further generalization of ASR into the class multitwisted (MT) codes. Using this method, we have found many codes with best known parameters with more direct and desirable constructions than what is currently available in the database of BKLCs.

1. Introduction

A linear code C of length n over a finite field \mathbb{F}_q is a vector subspace of \mathbb{F}_q^n with three basic parameters: the length n , the dimension k , and the minimum (Hamming) distance/weight d . Such a code is referred to as an $[n, k, d]_q$ code. Given an alphabet \mathbb{F}_q , a length n , and a dimension k , a key problem in coding theory is to construct a code with the highest possible minimum distance $d_q(n, k)$. There are theoretical upper bounds on $d_q(n, k)$, but upper bounds usually do not give a way of constructing codes and it is not guaranteed that they may be attained. Determining $d_q(n, k)$ with an explicit construction of a code attaining it is a challenging problem with many instances that are open. There are databases which keep records of the best known linear codes (BKLC), providing information about upper and lower

MSC2010: 94B15, 94B60.

Keywords: quasitwisted codes, multitwisted codes, best known linear codes, ASR search algorithm.

bounds, and their constructions. One of the databases is maintained by M. Grassl [2019] and another is available within the Magma software.¹ Both of these databases cover information about BKLCs over finite fields \mathbb{F}_q for $q = 2, 3, 4, 5, 7, 8, 9$ up to a certain value of n for each alphabet.

Computers are very useful in searching for new codes with best possible parameters. The two main challenges in this search come from the fact that computing the minimum distance of a linear code is NP-hard [Vardy 1997] and the number of linear codes for a given length n and dimension k grows very fast. Therefore, exhaustive searches for arbitrary linear codes are not possible for all but small parameters. In order to combat these computational challenges, researchers choose to focus on certain classes of codes that have ample mathematical structure and are known to contain many good codes. One well-known example of such a class of codes is the class of quasitwisted (QT) codes, which are a generalization of cyclic, constacyclic, and quasicyclic (QC) codes. For the last several decades, hundreds of new linear codes have been found by computer searches from the class of QC and QT codes; see, e.g., [Chen 1994; Daskalov and Gulliver 2000; Gulliver and Bhargava 1996]. In particular, a specific search algorithm called ASR [Aydin et al. 2001] on 1-generator QT codes has been effective in discovering record-breaking codes; see, e.g., [Daskalov and Hristov 2003a; 2003b; 2004; Aydin and Siap 2002]. The ASR algorithm searches for QT codes with generator polynomials of the form

$$(g(x), f_2(x)g(x), f_3(x)g(x), \dots, f_\ell(x)g(x)), \quad (1)$$

where ℓ is the number of blocks in the QT code and each $f_i(x) \in \mathbb{F}_q[x]$ has degree less than $k = m - \deg(g(x))$ and is relatively prime with $h(x)$, where $h(x)$ is the check polynomial of the constacyclic code generated by $g(x)$, that is, $x^m - a = g(x)h(x)$ [Aydin et al. 2001]. Such a code is a 1-generator QT code of index ℓ with parameters $[m\ell, k, \geq d\ell]$, if the constacyclic code generated by $g(x)$ has parameters $[m, k, d]$.

We have two main contributions to this research. Our first contribution is to generalize this search algorithm to generate 1-generator QT codes with generators of the form $(f_1(x)g_1(x), f_2(x)g_2(x), \dots, f_\ell(x)g_\ell(x))$, where polynomials $g_i(x)$ generate nonequivalent constacyclic codes with length m and dimension k , and each of the scrambling polynomials $f_i(x)$ fulfills the conditions $\gcd(h_i(x), f_i(x)) = 1$ and $\deg(f_i(x)) < k$. If there are ℓ blocks and r generators for a specific m and k , then there are $\binom{\ell+r-1}{r-1}$ possible combinations of generator polynomials.² Meaning, for a given m, k , and ℓ we have $\binom{\ell+r-1}{r-1}$ different searches. This generalized search

¹<http://magma.maths.usyd.edu.au/magma/>

²This follows from the following well-known formula from combinatorics: the number of distinct nonnegative integer-valued vectors (x_1, x_2, \dots, x_k) satisfying $x_1 + x_2 + \dots + x_k = n$ is $\binom{n+k-1}{n}$, which is the same as $\binom{n+k-1}{k-1}$. In our situation, we take $k = r$ and $n = \ell$.

algorithm builds upon the work of previous researchers who generalized the search by partitioning generator polynomials using code equivalence [Aydin et al. 2019]. Our second contribution is to introduce a new search algorithm to find linear codes with generator matrices of the form $[I_k, C]$, where I_k is the $k \times k$ identity matrix over \mathbb{F}_q and C is a circulant matrix corresponding to the generator of a constacyclic code of length $n - k$.

We first give some basic information on the structure of cyclic and QT codes, and the background of our generalized and new search methods. We then discuss some implementation details regarding both methods. Finally, we present new codes and their generators. We used Magma software along with C++ programs in executing our search algorithms.

2. Basic definitions

Cyclic codes are an important class of codes that connect coding theory to algebra [Prange 1957; 1958]. A cyclic code C is a linear code that is closed under the cyclic shift operation, meaning that if $c = (c_0, c_1, \dots, c_{n-1})$ is a codeword then $\pi(c) = (c_{n-1}, c_0, \dots, c_{n-2})$ must be as well. Representing a codeword $c = (c_0, c_1, \dots, c_{n-1})$ as the polynomial $c(x) = c_0 + c_1x + \dots + c_{n-1}x^{n-1}$ is the basis of the link between coding theory and algebra. It is well known that cyclic codes of length n over \mathbb{F}_q are precisely ideals in the quotient ring $\mathbb{F}_q[x]/\langle x^n - 1 \rangle$, which is a principal ideal ring. There is a one-to-one correspondence between divisors of $x^n - 1$ over $\mathbb{F}_q[x]$ and cyclic codes of length n over \mathbb{F}_q . Every cyclic code C has a unique standard generator polynomial $g(x)$ that divides $x^n - 1$. If we write $x^n - 1 = g(x)h(x)$ then the dimension of C is $\deg(h(x)) = n - \deg(g(x))$, and $h(x)$ is called the check polynomial of C .

An important generalization of cyclic codes is the class of constacyclic codes. A linear code C is called a constacyclic code if it is closed under the constacyclic shift operator π_a , where $a \in \mathbb{F}_q^* = \mathbb{F}_q \setminus \{0\}$. This means if c is a codeword then so is $\pi_a(c) = (ac_{n-1}, c_0, \dots, c_{n-2})$. Note that, the case $a = 1$ gives the cyclic codes. Similarly to cyclic codes, every constacyclic code C has a standard generator polynomial $g(x)$ that divides $x^n - a$. Any other generator of $C = \langle g(x) \rangle$ has the form $g(x)f(x)$, where $\gcd(g(x), h(x)) = 1$. Algebraically, constacyclic codes of length n over \mathbb{F}_q with shift constant a are precisely the ideals in $\mathbb{F}_q[x]/\langle x^n - a \rangle$. Given a generator $p(x) = p_0 + p_1x + \dots + p_{n-1}x^{n-1}$ of a constacyclic code, it has a generator matrix of the form

$$\begin{bmatrix} p_0 & p_1 & p_2 & \cdots & p_{n-1} \\ ap_{n-1} & p_0 & p_1 & \cdots & p_{n-2} \\ ap_{n-2} & ap_{n-1} & p_0 & \cdots & p_{n-3} \\ \vdots & \vdots & \vdots & & \vdots \\ ap_{n-k+1} & ap_{n-k+2} & ap_{n-k+2} & \cdots & p_{n-k} \end{bmatrix}.$$

Such a matrix is called an a -circulant matrix (also called a twistulant matrix). The special case $a = 1$ gives us the class of cyclic codes and its circulant matrices. We obtain quasicyclic (QC) and quasitwisted (QT) codes as further generalizations of cyclic and constacyclic codes.

A quasitwisted code is closed under a constacyclic shift by more than one position. A linear code C is said to be ℓ -QT (or a QT code of index ℓ) if for a positive integer ℓ , whenever $(c_0, c_1, \dots, c_{n-1}) \in C$, then $(ac_{n-\ell}, \dots, ac_{n-1}, c_0, c_1, \dots, c_{n-\ell-1}) \in C$. If we take $a = 1$, then we have the class of QC codes. Thus the class of QT codes is a generalization of QC codes. A generator matrix of an r -generator QT code has the form

$$\begin{bmatrix} G_{11} & G_{12} & \cdots & G_{1l} \\ G_{21} & G_{22} & \cdots & G_{2l} \\ \vdots & \vdots & & \vdots \\ G_{r1} & G_{r2} & \cdots & G_{rl} \end{bmatrix},$$

where each G_{ij} is a circulant matrix corresponding to a constacyclic code [Aydin et al. 2001]. Thus we can think of QT codes as codes made up of “blocks” of constacyclic codes. Throughout this paper we will be working with 1-generator QT codes, that is, codes whose generator matrices have the form

$$[G_1 \ G_2 \ \cdots \ G_l].$$

A QT code of length $n = m\ell$ and index ℓ is an R -submodule of R^ℓ , where $R = \mathbb{F}_q[x]/\langle x^m - a \rangle$.

Multitwisted (MT) codes are a more recent generalization of QT codes [Aydin and Halilović 2017]. A linear code C is multitwisted if for any codeword

$$\vec{c} = (c_{1,0}, \dots, c_{1,m_1-1}; c_{2,0}, \dots, c_{2,m_2-1}; \cdots; c_{\ell,0}, \dots, c_{\ell,m_\ell-1}) \in C$$

its MT shift

$$(a_1 c_{1,m_1-1}, c_{1,0}, \dots, c_{1,m_1-2}; a_2 c_{2,m_2-1}, c_{2,0}, \dots, c_{2,m_2-2}; \cdots; a_\ell c_{\ell,m_\ell-1}, \dots, c_{\ell,m_\ell-2})$$

is also a codeword, where $a_1, a_2, \dots, a_\ell \in \mathbb{F}_q^*$. Under the usual representation of a codeword \vec{c} as the corresponding polynomial (in this case a tuple of polynomials) $C(x) = (c_1(x), c_2(x), \dots, c_\ell(x))$, where $c_i(x) = c_{i,0} + c_{i,1}x + \cdots + c_{i,m_i-1}x^{m_i-1}$, we observe that the MT shift corresponds to the operation

$$xC(x) = (xc_1(x) \bmod x^{m_1} - a_1, \dots, xc_\ell(x) \bmod x^{m_\ell} - a_\ell)$$

in the ring

$$V = \prod_{i=1}^{\ell} \mathbb{F}_q[x]/\langle x^{m_i} - a_i \rangle,$$

where $a_i \in \mathbb{F}_q^*$ and m_i are (possibly distinct) positive integers.

3. A new generalization of the ASR algorithm

Recently, a generalization of the ASR algorithm was introduced in [Aydin et al. 2019] that made the algorithm more comprehensive based on the notion of code equivalence. Given a block length m , dimension k , alphabet size q and shift constant $a \in \mathbb{F}_q^*$, the original ASR algorithm used a code of largest minimum weight among all constacyclic codes of length m and dimension k over \mathbb{F}_q as the building block of a QT code with a generator of the form (1) (in the case of multiple such codes, one was arbitrarily chosen). Each such generator $g(x)$ is a divisor of $x^m - a$, with check polynomial $h(x)$, i.e., $x^m - a = g(x)h(x)$. The generalization introduced in [Aydin et al. 2019] first partitions all constacyclic codes of length m and dimension k over \mathbb{F}_q into equivalence classes. It keeps the (standard) generator of one code for each equivalence class, and uses each one of these polynomials as the building block of a QT search. As a result of this more comprehensive approach, a number new codes were found in [Aydin et al. 2019] that would have been missed by the original algorithm.

Past researchers have written Magma code that outputs generator polynomials for all nonequivalent cyclic and constacyclic codes for a fixed n over a finite field \mathbb{F}_q using cyclotomic cosets [Aydin et al. 2019]. Our first step in generalizing their search method was to take the list of nonequivalent generators and sort them by degree. The end result of the program is a set of text files each containing generator polynomials of a fixed degree.

In the past, researchers have automated the process of searching for new QT codes. They wrote C++ programs that take in the list of nonequivalent generator polynomials and generate Magma code in output files, and then they wrote a runner script that is able to run all of the Magma files simultaneously with one command. In order to consider all $\binom{\ell+r-1}{r-1}$ possible combinations of the distinct generator polynomials, we made significant changes to this program. To accomplish this we wrote a function that runs through each number $1, \dots, (\ell+1)^r$ and converts it to a number base $\ell+1$. To give a concrete example, if we have $\ell = 2$ and $r = 6$, the decimal number 10 will be represented in base 3 as 000101. We interpret this number in base $\ell+1$ as the number of times each of the r generators appear. In this instance, this means that g_4 and g_6 will appear as the generators of the QT code. In order for this number in base $\ell+1$ to be a valid combination of generator polynomials, it is clear that the components must sum to the number of blocks. Further, this method avoids equivalent codes because each number in base $\ell+1$ represents a unique combination of the r generator polynomials.

Our further generalization of the algorithm in [Aydin et al. 2019] does not require the generator polynomial $g(x)$ in (1) to be the same. Instead, we consider QT codes with generators of the form

$$(f_1(x)g_1(x), f_2(x)g_2(x), \dots, f_\ell(x)g_\ell(x)),$$

where each $g_i(x)$ is the standard generator of a constacyclic code of length m and dimension k , hence a divisor of $x^m - a$, for distinct equivalence classes. After storing the generator polynomials in an array, for each $g_i(x)$ we generate a random polynomial f_i over \mathbb{F}_q of degree $< k$ such that $\gcd(f_i(x), h_i(x)) = 1$. For each fixed set of polynomials $(g_1(x), g_2(x), \dots, g_\ell(x))$, we generate a large number of codes, where the $f_i(x)$ vary (generated by computer), with a generator of the form $(f_1(x)g_1(x), f_2(x)g_2(x), \dots, f_\ell(x)g_\ell(x))$.

For such a code the dimension is equal to $m - \deg(\gcd(g_1(x), g_2(x), \dots, g_\ell(x)))$. Since $\deg(\gcd(g_1(x), g_2(x), \dots, g_\ell(x))) \leq \deg(g_1(x))$, it is often the case that $m - \deg(\gcd(g_1(x), g_2(x), \dots, g_\ell(x))) > m - \deg(g_1(x))$. As such, truncating the generator matrix so that it has $m - \deg(g_1(x))$ rows often produces a code with higher minimum distance.

Note that when $g_1(x) = g_2(x) = \dots = g_\ell(x) = g(x)$, we obtain the form of the generator given by the algorithm ASR that has been used in previous searches and produced a large number of record-breaking codes over many alphabets. Hence, our method is more general. In the implementation of the ASR algorithm, the polynomial $f_1(x)$ is often taken to be 1, meaning codes are obtained by generators of the form $(g(x), f_2(x)g(x), \dots, f_\ell(x)g(x))$. We observe however that this may cause the search to miss some codes with potentially higher minimum distances. We illustrate this by a concrete example for $\ell = 2$.

Let S be the set of all polynomials of degree $< k$ over \mathbb{F}_q that are relatively prime with the check polynomial $h(x)$. If $\ell = 2$ and $f_1(x) = 1$, then the size of the search space is $|S|$. If we do not make the simplification $f_1(x) = 1$, then the size of the search space is increased to $|S|^2$. The computational cost in the increased size of the search space comes with a benefit, however. Through exhaustive computer search, we found that it is not possible to obtain every code of the form $(g(x)f_1(x), g(x)f_2(x))$ (or its equivalent) as a code of the form $(g(x), g(x)p(x))$. In particular, we have found cases in which an identical $g(x)$ can generate a code with a higher minimum distance if the generator is in the form $(g(x)f_1(x), g(x)f_2(x))$ as opposed to the form $(g(x), g(x)p(x))$. One such example is over \mathbb{F}_3 . If we take $g(x) = x^9 + x^7 + x + 1$, $n = 14$, $\ell = 2$, and $a = 1$, the QC code generated by $(g(x)f_1(x), g(x)f_2(x))$, where $f_1(x) = x^2 + 2x + 2$ and $f_2(x) = x^5 + 2x^3 + x + 1$ has minimum distance 16. On the other hand, the largest minimum distance of any QC code with a generator of the form $(g(x), f(x)g(x))$ is 14. Because of this oversight in the previous searches, we reran the usual ASR algorithm with $f_1(x)$ not fixed to 1. We obtained several new linear codes that are QC with this search. They are listed in Table 1 below.

In addition to trying cases where the degrees of the generator polynomials are equal, we also tried cases where the degrees are not equal. Say that we are working with $\ell = 2$ and we have $m - \deg(g_1(x)) = k_1$ and $m - \deg(g_2(x)) = k_2$, where

$k_1 < k_2$. In such a case it is clear that the left-hand block, corresponding to the $k_1 - 1$ constacyclic shifts of $g_1(x)$, will have fewer linearly independent rows than the right-hand block. In this case, without truncating we will have the dimension $m - \deg(\gcd(g_1(x), g_2(x)))$. In general this is greater than k_1 or k_2 , so the first block will have a number of rows that are linearly dependent. As a result when we row reduce the generator matrix there will be extra rows of zeros, which is detrimental to the minimum distance of the code. In order to avoid this issue, we ran a search in which the blocks are truncated to $k = \min(k_1, k_2)$. With this method, we found codes whose minimum distances are within 2 units of the BKLCs.

4. A new search algorithm: ICY

It is well known that every linear code has an equivalent linear code with a generator matrix of the form $G = [I_k \mid A]$, where I_k is the $k \times k$ identity matrix and A is a $k \times n - k$ matrix. This is known as the standard form of G . Our new search method is motivated by this fact as well as the form of the generator matrices of 1-generator MT codes. It combines these two forms. Consider linear codes with generator matrices of the form $G = [I_k \mid C_p]$, where C_p is the circulant generator matrix of a constacyclic code of length $n - k$ and dimension k defined by a generator polynomial (not necessarily the standard generator polynomial) $p(x) = p_0 + p_1x + \cdots + p_{n-k-1}x^{n-k-1}$,

$$\begin{bmatrix} p_0 & p_1 & p_2 & \cdots & p_{n-k-1} \\ ap_{n-k-1} & p_0 & p_1 & \cdots & p_{n-k-2} \\ ap_{n-k-2} & ap_{n-k-1} & p_0 & \cdots & p_{n-k-3} \\ \vdots & \vdots & \vdots & & \vdots \end{bmatrix},$$

where each row is the constacyclic shift of the row above it.

We can view such a code as a 1-generator MT code generated by $(1, p(x))$, where each component is a constacyclic code of length k and $n - k$ respectively. We modify the algorithm ASR to search for 1-generator MT codes with generators of the form $(1, g(x)f(x))$, where $g(x)$ is the standard generator of a constacyclic code of length $n - k$ with shift constant a , $x^m - a = g(x)h(x)$, and $f(x)$ is coprime with $h(x)$. We impose one further condition on $f(x)$. Since we fix k and n in advance, we know the minimum weight, d_{tar} of BKLC of length n and dimension k . In order for a code with generator of the form $(1, g(x)f(x))$ to have a greater minimum weight than the BKLC, it is necessary that the weight of the polynomial $f(x)g(x)$ be $\geq d_{\text{tar}}$. Without fulfilling this condition the resulting code would definitely have a minimum distance $\leq d_{\text{tar}}$. Our search uses the previous partition program from [Aydin et al. 2019] in Magma to find all nonequivalent generator polynomials (divisors of $x^m - a$) using cyclotomic cosets. Then, we

	$[n, k, d]_q$	α	polynomials
1	$[36, 14, 15]_5$	1	$g = [14014],$ $f_1 = [40041120014234],$ $f_2 = [41304123120031]$
2	$[42, 19, 15]_5$	1	$g = [111],$ $f_1 = [2103021100133],$ $f_2 = [404410144313]$
3	$[42, 13, 20]_5$	1	$g = [123333321],$ $f_1 = [2103021100133],$ $f_2 = [404410144313]$
4	$[56, 19, 23]_5$	1	$g = [1123421123],$ $f_1 = [2024244022144132123],$ $f_2 = [2141002343343311431]$
5	$[66, 16, 33]_5$	1	$g = [142141404012120124],$ $f_1 = [3442111233311333],$ $f_2 = [1313324003444113]$
6	$[66, 15, 34]_5$	1	$g = [1334323141114131121],$ $f_1 = [212410113224012],$ $f_2 = [230103303344443]$
7	$[76, 20, 35]_5$	1	$g = [1413203213422033444],$ $f_1 = [2343300200010024341],$ $f_2 = [32340321321200233421]$
8	$[76, 18, 37]_5$	1	$g = [140412123134331311011],$ $f_1 = [201430342434231303],$ $f_2 = [123443413204300424]$

Table 1. Record breaking QT codes.

run a C++ program that generates a search in Magma for each distinct generator polynomial along with a runner script.

We call this search method “ICY” from the form of the generator matrix $[I, C]$. The following result is a special case of Theorem 5.3 in [Aydin and Halilović 2017] and gives a lower bound on the minimum weight of codes generated by the ICY method.

Proposition 1. *Let $a \in \mathbb{F}_q^*$, $n \in \mathbb{Z}^+$, be such that $x^n - a = g(x)h(x)$, with $k = \deg(h(x))$. Let $C_2 = \langle g(x) \rangle$ be a constacyclic code with shift constant a and parameters $[n, k, d_2]$, with generator matrix G_2 . Let $G = [I_k, G_2]$. Then the MT code generated by G has parameters $[n + k, k, d]$, where $d \geq d_2 + 1$.*

	$[n, k, d]_q$	method
1	$[41, 18, 15]_5$	shortening of $[42, 19, 15]_5$ by coordinate 1
2	$[55, 18, 23]_5$	shortening of $[56, 19, 23]_5$ by coordinate 1
3	$[65, 16, 32]_5$	puncturing of $[66, 16, 33]_5$ by coordinate 1
4	$[65, 15, 33]_5$	shortening of $[66, 16, 33]_5$ by coordinate 1
5	$[75, 20, 34]_5$	puncturing of $[76, 20, 35]_5$ by coordinate 1
6	$[75, 17, 37]_5$	shortening of $[76, 18, 37]_5$ by coordinate 1
7	$[80, 20, 37]_5$ *	applying Construction X to $[76, 20, 34]_5$ and $[76, 18, 37]_5$ along with a $[4, 2, 3]$ -code

Table 2. Additional new codes; the starred construction is due to Grassl [2019].

Using this method have found many codes with the same parameters as the BKLCs over \mathbb{F}_2 and \mathbb{F}_5 . One such example is a code with parameters $[44, 14, 19]_5$ with shift constant $a = 2$,

$$g(x) = x^{16} + 4x^{15} + 4x^{14} + 3x^{13} + 3x^{12} + 2x^{11} + 2x^{10} + 2x^9 + 3x^8 + 2x^7 + 2x^6 + 2x^5 + 3x^4 + 3x^3 + 4x^2 + 4x + 1,$$

$$f(x) = 3x^{13} + 3x^{11} + 2x^{10} + 4x^9 + 2x^8 + 3x^6 + 4x^5 + 4x^4 + 4x^2 + 4x + 3.$$

The constacyclic code generated by $g(x)$ has parameters $[30, 14, 5]_5$, so the minimum distance of the resulting $[I, C]$ code is far greater than $d_2 + 1$. Furthermore, this code has a much simpler and more elegant construction than the current BKLC given in [Grassl 2019]. Moreover, according to the database of QC and QT codes, there does not exist a QT code with these parameters over \mathbb{F}_5 . Although this code is MT, not QT, its structure is very close to the structure of a QT code. The fact that this code has the parameters of BKLC with a more desirable and simpler construction, as well as having better parameters than known QT codes makes it an excellent code. We have found a number of similar codes. They are listed in Table 3 below.

5. Multitwisted searches

Multitwisted (MT) codes [Aydin and Halilović 2017] are a recent generalization of QT codes, which also generalize previously introduced classes of double cyclic codes [Borges et al. 2018; Gao et al. 2016], QCT codes [Aydin et al. 2007], and GQC codes [Siap and Kulhan 2005]. We conducted a computer search based on Theorem 5.7 from [Aydin and Halilović 2017] and our earlier algorithms.

We fix n_2 and a_2 over a given finite field \mathbb{F}_q . Our cyclic partition program gives all nonequivalent generators $g(x)$ (divisors of $x^{n_2} - a_2$). Then we sort all polynomials by degree. Once we have all of the generators sorted by degree we loop through many values of $k = n_1$. We chose $10 \leq n_1 \leq n_2 - 5$. For each n_1 we

	$[n, k, d]_q$	ℓ	α	polynomials
1	$[49, 19, 12]_2 \star$	2	1	$g = [100100101101],$ $f = [1101101001010100101]$
2	$[50, 20, 12]_2$	2	1	$g = [11100011011],$ $f = [110101111100011111]$
3	$[51, 21, 12]_2 \star$	2	1	$g = [1100001111],$ $f = [1011110001000111]$
4	$[52, 22, 12]_2 \star$	2	1	$g = [100000101],$ $f = [10001111111110110101]$
5	$[55, 25, 12]_2 \star$	2	1	$g = [11111],$ $f = [1101111101001001001011101]$
6	$[67, 25, 16]_2 \star$	2	1	$g = [111110011000100001],$ $f = [111001111011010111011]$
7	$[68, 26, 16]_2$	2	1	$g = [11010000101110101],$ $f = [110010010110011111010111]$
8	$[69, 27, 16]_2 \star$	2	1	$g = [1001011110111011],$ $f = [11101000100010001010110111]$
9	$[96, 36, 20]_2 \star$	2	1	$g = [1101001001001011110010101],$ $f = [11000010011111010000111010000110101]$
10	$[98, 38, 20]_2 \star$	2	1	$g = [10000111101001111110111],$ $f = [1000110000100000001010010111110001]$
11	$[109, 47, 20]_2 \star$	2	1	$g = [1011101000110101],$ $f = [10011101101110101011010001010000011100001]$
12	$[87, 24, 24]_2 \star$	2	1	$g = [1010001010010010000101011110000100001001],$ $f = [10101111001011100110001]$
13	$[88, 25, 24]_2 \star$	2	1	$g = [110010110000001011001010001011110011011],$ $f = [1100001110000111101000001]$
14	$[45, 18, 16]_5$	2	2	$g = [1000000004],$ $f = [133011120220340341]$
15	$[44, 14, 19]_5 \star$	2	2	$g = [14433222322233441],$ $f = [30324203440443]$

Table 3. Good codes from ICY method. Here the star denotes a code with better parameters than those in the online database of QT codes [Chen].

find all generator polynomials of degree $n_2 - n_1$ and create a search program for each. We also loop through all $a \in \mathbb{F}_q^*$ of distinct orders. We chose these bounds to limit the search space to a manageable size.

Our search produced a number of ties for BKLCs over \mathbb{F}_5 and \mathbb{F}_2 . One such example is the code with parameters $[51, 17, 21]_5$, where

$$\begin{aligned} n_1 &= 17, & n_2 &= 34, & a_1 &= 4, & a_2 &= 1, & g(x) &= x^{17} + 4, \\ f_1(x) &= 2x^{16} + 2x^{15} + 4x^{14} + 2x^{13} + 3x^{12} + 2x^{11} \\ &\quad + 2x^{10} + 2x^9 + 2x^8 + 2x^7 + 3x^6 + 4x^5 + 2x^4 + 4x^3 + 2x^2 + 1, \\ f_2(x) &= x^{16} + 4x^{15} + x^{14} + 2x^{13} + 4x^{12} + 2x^{11} + 2x^{10} \\ &\quad + 2x^9 + 2x^8 + x^7 + 2x^6 + 4x^5 + 3x^4 + 4x^3 + 2x^2 + x + 1. \end{aligned}$$

While our tie is an equally similar construction to the one recorded in the database [Grassl 2019], we have found a number of ties with simpler constructions. One such code has the parameters $[44, 16, 17]_5$, where

$$\begin{aligned} n_2 &= 28, & n_1 &= 16, & a_2 &= 1, & a_1 &= 1, \\ g(x) &= 2x^{27} + x^{26} + 3x^{25} + 3x^{24} + 4x^{23} + 2x^{21} + 4x^{19} + 3x^{18} + 4x^{17} + 4x^{16} + 4x^{15} \\ &\quad + 3x^{14} + 3x^{13} + 3x^{12} + 3x^{11} + 3x^9 + 2x^8 + 3x^7 + 4x^6 + 3x^5 + 3x^4 + 3x^3 + x^2 + x, \\ f_1(x) &= 2x^{15} + 2x^{14} + 2x^{12} + 2x^{11} + 4x^{10} + 3x^9 + 3x^8 + 3x^4 + 4x^3 + x^2 + 4x + 4, \\ f_2(x) &= 2x^{15} + 4x^{14} + 4x^{13} + 2x^{12} + x^{11} + 4x^{10} + 3x^9 + 2x^7 + x^6 + x^4 + 2x^3 + x^2 + x. \end{aligned}$$

6. Computational results

First, we present our new record-breaking codes from the QT search in Table 1. We present the polynomials with the highest-degree coefficients on the left. Each of these codes has $\ell = 2$. Table 2 gives new codes obtained from these QT codes through the standard procedures of shortening and puncturing. Finally, in Table 3 we present codes that have the parameters of the BKLCs but with simpler and more desirable constructions obtained by the ICY method. Most of these codes have better parameters than the codes in the database of QC and QT codes [Chen]. Each such code is marked with a \star .

Acknowledgement

This work was supported by Kenyon Summer Science Scholars program.

References

- [Aydin and Halilović 2017] N. Aydin and A. Halilović, “A generalization of quasi-twisted codes: multi-twisted codes”, *Finite Fields Appl.* **45** (2017), 96–106. MR Zbl
- [Aydin and Siap 2002] N. Aydin and I. Siap, “New quasi-cyclic codes over \mathbb{F}_5 ”, *Appl. Math. Lett.* **15**:7 (2002), 833–836. MR Zbl

- [Aydin et al. 2001] N. Aydin, I. Siap, and D. K. Ray-Chaudhuri, “The structure of 1-generator quasi-twisted codes and new linear codes”, *Des. Codes Cryptogr.* **24**:3 (2001), 313–326. MR Zbl
- [Aydin et al. 2007] N. Aydin, T. Asamov, and T. A. Gulliver, “Some open problems on quasi-twisted and related code constructions and good quaternary codes”, pp. 856–860 in *Proceedings of the 2007 IEEE International Symposium on Information Theory* (Nice, France, 2007), IEEE, Piscataway, NJ, 2007.
- [Aydin et al. 2019] N. Aydin, J. Lambrinos, and O. Vandenberg, “On equivalence of cyclic codes, generalization of a quasi-twisted search algorithm, and new linear codes”, *Des. Codes Cryptogr.* **87**:10 (2019), 2199–2212. MR Zbl
- [Borges et al. 2018] J. Borges, C. Fernández-Córdoba, and R. Ten-Valls, “ \mathbb{Z}_2 -double cyclic codes”, *Des. Codes Cryptogr.* **86**:3 (2018), 463–479. MR Zbl
- [Chen 1994] Z. Chen, “Six new binary quasi-cyclic codes”, *IEEE Trans. Inform. Theory* **40**:5 (1994), 1666–1667. MR Zbl
- [Chen] E. Chen, “Online database of quasi-twisted codes”, website, <http://www.tec.hkr.se/~chen/research/codes/>.
- [Daskalov and Gulliver 2000] R. N. Daskalov and T. A. Gulliver, “New quasi-twisted quaternary linear codes”, *IEEE Trans. Inform. Theory* **46**:7 (2000), 2642–2643. MR Zbl
- [Daskalov and Hristov 2003a] R. Daskalov and P. Hristov, “New binary one-generator quasi-cyclic codes”, *IEEE Trans. Inform. Theory* **49**:11 (2003), 3001–3005. MR Zbl
- [Daskalov and Hristov 2003b] R. Daskalov and P. Hristov, “New quasi-twisted degenerate ternary linear codes”, *IEEE Trans. Inform. Theory* **49**:9 (2003), 2259–2263. MR Zbl
- [Daskalov et al. 2004] R. Daskalov, P. Hristov, and E. Metodieva, “New minimum distance bounds for linear codes over $\text{GF}(5)$ ”, *Discrete Math.* **275**:1-3 (2004), 97–110. MR Zbl
- [Gao et al. 2016] J. Gao, M. Shi, T. Wu, and F.-W. Fu, “On double cyclic codes over \mathbb{Z}_4 ”, *Finite Fields Appl.* **39** (2016), 233–250. MR Zbl
- [Grassl 2019] M. Grassl, “Code tables: bounds on the parameters of various types of codes”, website, 2019, <http://www.codetables.de/>.
- [Gulliver and Bhargava 1996] T. A. Gulliver and V. K. Bhargava, “New good rate $(m-1)/pm$ ternary and quaternary quasi-cyclic codes”, *Des. Codes Cryptogr.* **7**:3 (1996), 223–233. MR Zbl
- [Prange 1957] E. Prange, “Cyclic error-correcting codes in two symbols”, technical report TN-57-103, Air Force Cambridge Research Center, 1957.
- [Prange 1958] E. Prange, “Some cyclic error-correcting codes with simple decoding algorithm”, technical report TN-58-156, Air Force Cambridge Research Center, 1958.
- [Siap and Kulhan 2005] I. Siap and N. Kulhan, “The structure of generalized quasi cyclic codes”, *Appl. Math. E-Notes* **5** (2005), 24–30. MR Zbl
- [Vardy 1997] A. Vardy, “The intractability of computing the minimum distance of a code”, *IEEE Trans. Inform. Theory* **43**:6 (1997), 1757–1766. MR Zbl

Received: 2019-08-02 Revised: 2019-12-08 Accepted: 2019-12-27

aydinn@kenyon.edu Kenyon College, Gambier, OH, United States

guidotti1@kenyon.edu Kenyon College, Gambier, OH, United States

liu4@kenyon.edu Kenyon College, Gambier, OH, United States

shaikh1@kenyon.edu Kenyon College, Gambier, OH, United States

vandenberg1@kenyon.edu Kenyon College, Gambier, OH, United States

Continuous factorization of the identity matrix

Yuying Dai, Ankush Hore, Siqi Jiao, Tianxu Lan and Pavlos Motakis

(Communicated by Kenneth S. Berenhaut)

We investigate conditions under which the identity matrix I_n can be continuously factorized through a continuous $N \times N$ matrix function A with domain in \mathbb{R} . We study the relationship of the dimension N , the diagonal entries of A , and the norm of A to the dimension n and the norms of the matrices that witness the factorization of I_n through A .

1. Introduction

The problem from which this paper draws motivation concerns the relation between the magnitude of the diagonal entries a_{ii} of an $N \times N$ matrix A , the norm of A , and the dimension n of a vector space that A preserves in a satisfying manner, as precisely described below.

Problem 1. Given $N \in \mathbb{N}$ and $\delta > 0$ find the largest $n \in \mathbb{N}$ with the following property: for every $N \times N$ matrix $A = (a_{ij})$, with $\|A\| \leq 1$, the diagonal entries of which satisfy $|a_{ii}| \geq \delta$ for $1 \leq i \leq N$, there exist $n \times N$ and $N \times n$ matrices L and R so that $LAR = I_n$ and $\|L\|\|R\| \leq 2/\delta$.

The upper bound imposed on the quantity $\|L\|\|R\|$ must necessarily be at least $1/\delta$ (see Remark 2.12). We use elementary combinatorics and linear algebra to study Problem 1. Subsequently, we allow the entries of A to vary continuously and study the corresponding problem in the solution of which it is additionally required that the preserved vector spaces vary continuously as well. In this article we are mainly concerned with the following.

Problem 2. Given $N \in \mathbb{N}$ and $\delta > 0$ find the largest $n \in \mathbb{N}$ with the following property: for every $N \times N$ continuous matrix function $A : \mathbb{R} \rightarrow M_N(\mathbb{R})$ with $\|A(t)\| \leq 1$ and $|a_{ii}(t)| \geq \delta$ for $1 \leq i \leq N$ and all $t \in \mathbb{R}$, there exist continuous matrix functions $L : \mathbb{R} \rightarrow M_{n \times N}(\mathbb{R})$ and $R : \mathbb{R} \rightarrow M_{N \times n}(\mathbb{R})$ so that $L(t)A(t)R(t) = I_n$ and $\|L(t)\|\|R(t)\| \leq 2/\delta$ for all $t \in \mathbb{R}$.

MSC2010: 15A23, 46B07.

Keywords: factorization of the identity, matrices with large diagonal.

We provide lower bounds for n in Problems 1 and 2. In particular, we show that in both cases the order of magnitude of n is at least $\delta^{4/3}N^{1/3}$ (see Theorems 2.10 and 3.9). In the continuous case, this is achieved by using the proof of our estimate for Problem 1 pointwise. In this fashion, we obtain an open cover of \mathbb{R} consisting of intervals on each of which there are continuous matrix functions L and R factoring I_n through A . In the final step, we use these local solutions as building blocks to construct a continuous solution defined on the entire real line.

Although our approach is entirely Euclidean and finite-dimensional, this topic has origins that fit neither description. On a (generally infinite-dimensional) Banach space X with a coordinate system $(e_i)_i$ (e.g., a Schauder basis) every bounded linear operator $A : X \rightarrow X$ can be identified with an infinite matrix (a_{ij}) . If this matrix has large diagonal, in the sense that $\inf_i |a_{ii}| > 0$, one may ask whether there exist bounded linear operators $L, R : X \rightarrow X$ so that $LAR = I_X$. A. D. Andrew [1979] first showed that the answer is yes if $X = L_p$, $1 < p < \infty$, and the coordinate system under consideration is the Haar system. Since then, a number of papers have contributed to the study of this general problem in a variety of infinite-dimensional Banach spaces X ; see, e.g., [Laustsen et al. 2018; Lechner 2017; 2018a; 2018b; 2019c; Lechner et al. 2018]. The source of the finite-dimensional version of this problem can be traced to J. Bourgain and L. Tzafriri [1987]. Their paper, among other results, provides an estimate for n in Problem 1 which is of the order $\delta^2 N$ (see Remark 2.11). Within this context, other finite-dimensional non-Euclidean spaces have been studied by R. Lechner [2019a; 2019b]. To the best of our knowledge, the continuous matrix function case has not been considered before.

The paper is divided into two sections. In Section 2 we provide necessary estimates for the norm of a matrix, as well as estimates for the size of families of columns of a given matrix A with the property of being almost orthogonal to one another. Subsequently, we proceed to give an estimate of n for Problem 1 by defining matrices L and R . In Section 3 we explicitly use the definition of L and R of the constant case to find for each t in the domain of the matrix function A $L(t)$ and $R(t)$ as desired. We then extend these solutions continuously on a small interval around t . From there on, we synthesize these local solutions by taking appropriate convex combinations of them and we observe that the desired conclusion is satisfied.

In the sequel, for an $N \times N$ matrix $A = (a_{i,j}) = [a_1 \cdots a_N]$ we will consider the quantity $\theta = \min_i \|a_i\|$, instead of $\delta = \min_i |a_{i,i}|$. As $\delta \leq \theta$, our results are slightly more general than already advertised. We have included proofs of some well-known facts and estimates in an effort to make this paper as self-contained as possible. Although all results are stated and proved for matrices with real entries, obvious modifications make them valid for matrices with complex entries as well.

2. The constant case

We use elementary counting tools and tools from linear algebra to factorize the identity matrix through a square matrix with large diagonal. The section is organized into three subsections. The first one includes simple estimates of the norm of a matrix, the second one presents combinatorial arguments that are used to find collections of columns of a matrix that are almost orthogonal to one another, and in the third one we present the construction of the factors L and R and prove their desired properties.

Let us recall some necessary notions used in this section. We identify \mathbb{R}^n with the collection of $n \times 1$ matrices. Thus when we write $x = (x_1, \dots, x_n)$ in reality we mean $x = [x_1 \cdots x_n]^\top$. For $1 \leq i \leq n$ we denote by e_i the vector in \mathbb{R}^n that has 1 in the i -th entry and 0 in all others. Recall that for a vector $x = (x_1, \dots, x_n)$ in \mathbb{R}^n we define its Euclidean norm to be the quantity

$$\|x\| = \left(\sum_{i=1}^n |x_i|^2 \right)^{1/2}.$$

For two vectors $x = (x_1, \dots, x_n)$, $y = (y_1, \dots, y_n)$ in \mathbb{R}^n their inner product is the quantity

$$\langle x, y \rangle = \sum_{i=1}^n x_i y_i.$$

The Cauchy–Schwarz inequality states that for such x and y we have $|\langle x, y \rangle| \leq \|x\| \|y\|$; see, e.g., [Meckes and Meckes 2018, Theorem 4.6]. For an $m \times n$ matrix $A = (a_{i,j})$ when we write $A = [a_1 \cdots a_n]$ we mean that for each $1 \leq j \leq n$ the entries of the j -th column of A form a_j , i.e., the vector $(a_{1,j}, \dots, a_{m,j})$ in \mathbb{R}^m (a similar notation can be used for writing A with respect to its rows $\alpha_1^\top, \dots, \alpha_m^\top$). Then, for $n \in \mathbb{N}$ the $n \times n$ identity matrix I_n is the matrix $[e_1 \cdots e_n]$. Recall, if A is an $m \times n$ matrix with columns a_1, \dots, a_n and B is a $k \times m$ matrix with rows $\beta_1^\top, \dots, \beta_k^\top$, then the i, j -th entry of the product matrix BA is $\langle \beta_i, a_j \rangle$. For an $m \times n$ matrix A we define its norm to be the quantity

$$\|A\| = \sup\{\|Ax\| : x \in \mathbb{R}^n, \|x\| \leq 1\}.$$

It is easy to see that for A and x of appropriate dimensions we have $\|Ax\| \leq \|A\| \|x\|$. Similarly, by the association property of matrix multiplication, see, e.g., [Meckes and Meckes 2018, Theorem 2.10], for matrices A and B of appropriate dimensions we have $\|AB\| \leq \|A\| \|B\|$. Finally, recall that a function $f : \mathbb{R} \rightarrow \mathbb{R}$ is called convex if for every $0 \leq \lambda \leq 1$ and $s, t \in \mathbb{R}$ we have

$$f(\lambda s + (1 - \lambda)t) \leq \lambda f(s) + (1 - \lambda)f(t).$$

A direct computation can be used to show that the square function $f(t) = t^2$ is a convex function.

2A. Upper bounds of matrix norms. The estimates in this subsection are elementary and well known, yet we include the simple proofs for completeness.

Proposition 2.1. *Let $m, n \in \mathbb{N}$ and $A = [a_1 \cdots a_n]$ be an $m \times n$ matrix. Set*

$$\Lambda = \max_{1 \leq i \leq n} \|a_i\| \quad \text{and} \quad \lambda = \max_{1 \leq i \neq j \leq n} |\langle a_i, a_j \rangle|.$$

Then $\|A\| \leq (\Lambda^2 + (n-1)\lambda)^{1/2}$.

Proof. Let $x = (x_1, \dots, x_m)$ be a vector of norm 1. By convexity of the square function we have

$$\left(\sum_{i=1}^n \frac{1}{n} |x_i| \right)^2 \leq \frac{1}{n} \sum_{i=1}^n |x_i|^2,$$

or

$$\sum_{i=1}^n |x_i| \leq n^{1/2} \|x\|.$$

Then,

$$\begin{aligned} \|Ax\|^2 &= \langle Ax, Ax \rangle = \sum_{i=1}^m x_i^2 \|a_i\|^2 + \sum_{i \neq j} x_i x_j \langle a_i, a_j \rangle \\ &\leq \Lambda^2 \|x\|^2 + \lambda \sum_{i \neq j} |x_i x_j| = \Lambda^2 + \lambda \left(\sum_{i=1}^n |x_i| \sum_{i=j}^n |x_j| - \sum_{i=1}^n |x_i|^2 \right) \\ &\leq \Lambda^2 + \lambda(n-1). \quad \square \end{aligned}$$

Corollary 2.2. *Let $n \in \mathbb{N}$ and $A = (a_{i,j})$ be an $m \times n$ matrix. Set $d = \max_{i,j} |a_{i,j}|$. Then $\|A\| \leq dm^{1/2}n^{1/2}$.*

Proof. Every column of A has norm at most $dm^{1/2}$ and any two different columns have inner product with absolute value at most md^2 . A direct application of Proposition 2.1 yields the desired bound. \square

Corollary 2.3. *Let $N, n \in \mathbb{N}$ and $A = [a_1 \cdots a_n]$ be an $N \times n$ matrix. Set*

$$\lambda = \max_{1 \leq i \neq j \leq n} |\langle a_i, a_j \rangle| \quad \text{and} \quad \Delta = \max_{1 \leq i \leq n} \|a_i\|^2 - 1|.$$

Then $\|A^T A - I_n\| \leq n \max\{\lambda, \Delta\}$.

Proof. The i, j entry of the matrix $A^T A - I$ is $\langle a_i, a_j \rangle$ if $i \neq j$ and $\|a_i\|^2 - 1$ if $i = j$. The result follows from applying Corollary 2.2. \square

2B. Counting arguments. In this section we estimate the maximal number of columns of a norm-1 matrix that can have large inner product with a fixed column. This estimate is then used to find collections of columns which are almost orthogonal to one another.

Proposition 2.4. *Let $A = [a_1 \cdots a_M]$ be an $N \times M$ matrix and let $\varepsilon > 0$. Then for every $1 \leq i \leq M$ the set*

$$B_i^\varepsilon = \{1 \leq j \leq M : |\langle a_i, a_j \rangle| \geq \varepsilon\}$$

has at most $\|A\|^4/\varepsilon^2$ elements.

Proof. If a_i is the zero vector then the conclusion is obvious and we may therefore assume that it is not. Recall that for any matrix A we have $\|A\| = \|A^\top\|$. Indeed, if x is a norm-1 vector with $\|A\| = \|Ax\|$ then

$$\begin{aligned} \|A\|^2 &= \langle Ax, Ax \rangle = \langle x, A^\top Ax \rangle \\ &\leq \|x\| \|A^\top Ax\| \leq \|A^\top\| \|A\| \|x\|^2 = \|A^\top\| \|A\| \end{aligned}$$

and hence $\|A\| \leq \|A^\top\|$. By symmetry of the argument we also have $\|A^\top\| \leq \|A\|$. We calculate

$$\|A\|^2 = \|A^T\|^2 \geq \frac{1}{\|a_i\|^2} \|A^T a_i\|^2 = \frac{1}{\|A e_i\|^2} \sum_{k=1}^M |\langle a_k, a_i \rangle|^2 \geq \frac{1}{\|A\|^2} \varepsilon^2 \#B_i^\varepsilon. \quad \square$$

Corollary 2.5. *Let $n \in \mathbb{N}$ with $n \geq 2$, $0 < \varepsilon < 1/(n-1)^{1/2}$, and $N \geq n/\varepsilon^2$. Then for any $L \in \mathbb{N}$ and $L \times N$ matrix $A = [a_1 \cdots a_N]$, with $\|A\| \leq 1$, there exists $F \subset \{1, \dots, N\}$, with $\#F = n$, so that for $i \neq j \in F$ we have $|\langle a_i, a_j \rangle| < \varepsilon$.*

Proof. Set $i_1 = 1$ and inductively pick i_2, \dots, i_n so that for $2 \leq k \leq n$

$$i_k \in \{1, \dots, N\} \setminus (\{i_1, \dots, i_{k-1}\} \cup (\bigcup_{m=1}^{k-1} B_{i_m}^\varepsilon)).$$

This is possible because, by Proposition 2.4, in every inductive step $2 \leq k \leq n$ the set $\{1, \dots, N\} \setminus (\{i_1, \dots, i_{k-1}\} \cup (\bigcup_{m=1}^{k-1} B_{i_m}^\varepsilon))$ has at least

$$N - \left(k - 1 + \frac{(k-1)}{\varepsilon^2}\right) \geq \frac{n}{\varepsilon^2} - (n-1) \left(1 + \frac{1}{\varepsilon^2}\right) = \frac{1}{\varepsilon^2} - (n-1) > 0$$

elements. □

The following estimate will be used in Section 3. We include it here for consistency.

Corollary 2.6. *Let $n \in \mathbb{N}$ with $n \geq 2$, $0 < \varepsilon < 1/(n-1)^{1/2}$, and $N \geq 5n/\varepsilon^2$. Let $A = [a_1 \cdots a_N]$ be an $N \times N$ matrix with $\|A\| \leq 1$. Then for every $F_1, F_2 \subset \{1, \dots, N\}$ with $\#F_1 = \#F_2 = n$ there exists $F_3 \subset \{1, \dots, N\}$ with $\#F_3 = n$ so that the following hold:*

- (i) F_3 is disjoint from $F_1 \cup F_2$.
- (ii) For any $i \neq j \in F_3$ we have $|\langle a_i, a_j \rangle| < \varepsilon$.
- (iii) For any $i \in F_3, j \in F_1 \cup F_2$, we have $|\langle a_i, a_j \rangle| < \varepsilon$.

Proof. Define $G = \{1, \dots, N\} \setminus ((F_1 \cup F_2) \cup (\bigcup_{i \in F_1 \cup F_2} B_i^\varepsilon))$. Then

$$\#G \geq \frac{5n}{\varepsilon^2} - 2n - \frac{2n}{\varepsilon^2} = \frac{3n}{\varepsilon^2} - 2n \geq \frac{n}{\varepsilon^2}.$$

We now follow the exact same argument as in the proof of Corollary 2.5 to find $F_3 \subset G$ with $\#F_3 = n$ so that for all $i \neq j \in F_3$ we have $|\langle a_i, a_j \rangle| < \varepsilon$. The fact that $F_3 \subset G$ also yields (i) and (iii). \square

2C. The matrices L and R . We next explicitly define the matrices L and R with the property $LAR = I_n$. For the definition of L and R we use the results from Section 2B. We then use the estimates provided in Section 2A to estimate the quantity $\|L\| \|R\|$.

We now introduce the matrices $L_{(A,F)}$, $R_{(A,F)}$ that are defined using A and a subset F of the columns of A . This dependence on F will also be important in the next section.

Definition 2.7. Let $n \leq N \in \mathbb{N}$, $A = [a_1 \cdots a_N]$ be an $N \times N$ matrix, and $F = \{i_1 < \cdots < i_n\}$ be a subset of $\{1, \dots, N\}$ with $\|a_i\| > 0$ for $i \in F$. For $k = 1, \dots, n$ set $r_{(A,F)}^k = e_{i_k} / \|a_{i_k}\|$; i.e., $r_{(A,F)}^k$ is the N -dimensional vector that has $1/\|a_{i_k}\|$ in the i_k -th entry and zero everywhere else. Define the $N \times n$ and $n \times N$ matrices

$$R_{(A,F)} = [r_{(A,F)}^1 \cdots r_{(A,F)}^n] \quad \text{and} \quad L_{(A,F)} = (AR_{(A,F)})^T.$$

Remark 2.8. Observe that for $1 \leq k \leq n$ we have $Ar_{(A,F)}^k = a_{i_k} / \|a_{i_k}\|$ and thus

$$AR_{(A,F)} = \begin{bmatrix} \frac{a_{i_1}}{\|a_{i_1}\|} & \cdots & \frac{a_{i_n}}{\|a_{i_n}\|} \end{bmatrix}.$$

Here, we give estimates for the norms of the matrices $L_{(A,F)}$, $R_{(A,F)}$, and $L_{(A,F)}AR_{(A,F)} - I_n$.

Proposition 2.9. Let $n \leq N \in \mathbb{N}$, A be an $N \times N$ matrix, and $F = \{i_1 < \cdots < i_n\}$ be a subset of $\{1, \dots, N\}$ with $\|a_i\| > 0$ for $i \in F$. Set

$$\theta = \min_{i \in F} \|a_i\| \quad \text{and} \quad \varepsilon = \max_{i \neq j \in F} |\langle a_i, a_j \rangle|.$$

Then we have

$$\|R_{(A,F)}\| \leq \theta^{-1}, \quad \|L_{(A,F)}\| \leq 1 + \frac{(n-1)^{1/2} \varepsilon^{1/2}}{\theta},$$

and

$$\|L_{(A,F)}AR_{(A,F)} - I_n\| \leq \frac{n\varepsilon}{\theta^2}.$$

Proof. The first two estimates follow from Proposition 2.1, whereas the third is a consequence of Corollary 2.3. For the first one observe that the columns of $R_{(A,F)}$

all have norm at most $1/\theta$ and they are all orthogonal to one another. For the second one, if we set $b_k = a_{i_k}/\|a_{i_k}\|$ for $1 \leq k \leq n$ then by Remark 2.8

$$AR_{(A,F)} = [b_1 \cdots b_n].$$

That is, all columns of $AR_{(A,F)}$ have norm 1 and for $1 \leq k \neq m \leq n$ we have $|\langle b_k, b_m \rangle| \leq \varepsilon/\theta^2$. Recall that for $x \geq 0$ we have $(1+x)^{1/2} \leq 1+x^{1/2}$. Thus,

$$\|L_{(A,F)}\| = \|L_{(A,F)}^T\| \leq \left(1 + \frac{(n-1)\varepsilon}{\theta^2}\right)^{1/2} \leq 1 + \frac{(n-1)^{1/2}\varepsilon^{1/2}}{\theta}.$$

The last estimate follows from Corollary 2.3 directly applied to the matrix $AR_{(A,F)} = [b_1 \cdots b_n]$. \square

The following is the main result of this section.

Theorem 2.10. *Let $N \in \mathbb{N}$ and let $A = [a_1 \cdots a_N]$ be an $N \times N$ matrix with $\|A\| \leq 1$. If $\theta = \min_{1 \leq i \leq N} \|a_i\| > 0$ then for every $1 \leq n \leq \frac{1}{5}\theta^{4/3}N^{1/3}$ there exist $n \times N$ and $N \times n$ matrices L and R respectively so that $LAR = I_n$ and $\|L\|\|R\| \leq 2/\theta$.*

Proof. If $n = 1$ the result easily follows by picking any column a_i and defining $R = e_i/\|a_i\|$ and $L = a_i^T/\|a_i\|$. We will therefore assume that $2 \leq n \leq \frac{1}{5}\theta^{4/3}N^{1/3}$. Define $\varepsilon = \theta^2/(9(n-1))$. This choice of ε ensures that

$$\frac{(n-1)^{1/2}\varepsilon^{1/2}}{\theta} = \frac{1}{3} \quad \text{and} \quad \frac{n\varepsilon}{\theta^2} \leq \frac{1}{4}. \quad (1)$$

The two estimates above will be used as assumptions to apply Proposition 2.9; however, we will first use Corollary 2.5. For that purpose, the choice of ε ensures that

$$\frac{n}{\varepsilon^2} = \frac{81n(n-1)^2}{\theta^4} \leq \frac{81}{\theta^4}n^3 \leq \frac{81}{\theta^4} \frac{\theta^4 N}{125} \leq N,$$

i.e., $N \geq n/\varepsilon^2$. It is also easily checked that $\varepsilon < 1/(n-1)^{1/2}$ (because $0 < \theta \leq 1$). Thus, by Corollary 2.5, there exists $F \subset \{1, \dots, N\}$ with $\#F = n$ so that for $i \neq j \in F$ we have $|\langle a_i, a_j \rangle| < \varepsilon$.

Consider now the matrices $L_{(A,F)}$ and $R_{(A,F)}$ given by Definition 2.7. By Proposition 2.9 and (1) we deduce

$$\|R_{(A,F)}\| \leq \theta^{-1}, \quad \|L_{(A,F)}\| \leq \frac{4}{3}, \quad \text{and} \quad \|L_{(A,F)}AR_{(A,F)} - I_n\| \leq \frac{1}{4}. \quad (2)$$

Set $R = R_{(A,F)}$. To define L , recall that if S is an $n \times n$ matrix with $\|S - I_n\| = c < 1$ then S^{-1} exists and $\|S^{-1}\| \leq 1/(1-c)$. One way to see this is to observe that $S^{-1} = \sum_{k=0}^{\infty} (I - S)^k$. Therefore, the matrix $(L_{(A,F)}AR_{(A,F)})^{-1}$ is well-defined and has norm at most $1/(1 - \frac{1}{4}) = \frac{4}{3}$. Finally, set $L = (L_{(A,F)}AR_{(A,F)})^{-1}L_{(A,F)}$ and observe that $LAR = I_n$, $\|R\| \leq 1/\theta$, and $\|L\| \leq \frac{16}{9} \leq 2$. \square

Remark 2.11. The theorem above may also be stated for an $N \times N$ matrix A without restrictions on $\|A\|$ as follows: if $\theta = \min_{1 \leq i \leq N} \|a_i\| > 0$ then for every $1 \leq n \leq \frac{1}{5}(\theta/\|A\|)^{4/3}N^{1/3}$ there exist $n \times N$ and $N \times n$ matrices L and R respectively so that $LAR = I_n$ and $\|L\|\|R\| \leq 2\|A\|/\theta$. This estimate can be compared to [Bourgain and Tzafriri 1987, Theorem 1.2], which yields a similar result: there exist universal constants $c, C > 0$ so that if N, A , and θ are as above then for every $1 \leq n \leq c(\theta/\|A\|)^2N$ there exist $n \times N$ and $N \times n$ matrices L and R respectively so that $LAR = I_n$ and $\|L\|\|R\| \leq C\|A\|/\theta$. We observe that the result from [Bourgain and Tzafriri 1987] gives a better relation between the dimension n and N , whereas our result gives a better relation between n and the quantity $\theta/\|A\|$.

Remark 2.12. In Theorem 2.10 whenever $n \geq 2$, the quantity $\|L\|\|R\|$ cannot be demanded to be below $1/\theta$. To see this fix $0 < \theta \leq 1$ and consider the $N \times N$ diagonal matrix A with first diagonal entry 1 and all other diagonal entries θ . If $n \geq 2$ and we assume that L, R are matrices with $LAR = I_n$, consider the subspace X of \mathbb{R}^n of all vectors orthogonal to $R^\top e_1$. Then X has codimension at most 1 and in particular it is nontrivial; i.e., we may pick $x \in X$ with $\|x\| = 1$. Then,

$$Rx = \sum_{i=1}^n \langle e_i, Rx \rangle e_i = \sum_{i=2}^n \langle e_i, Rx \rangle e_i$$

and thus we can compute that $ARx = \sum_{i=2}^n \theta \langle e_i, Rx \rangle e_i = \theta Rx$. By assumption, $LAR = I_n$ and so $\|x\| = \|LARx\| = \theta \|LRx\| \leq \theta \|L\|\|R\|\|x\|$. We conclude $\|L\|\|R\| \geq 1/\theta$.

3. The continuous case

In this section we present the main result of our paper. We demonstrate how the estimates from the previous section can be utilized to continuously factor the identity matrix through a continuous matrix function $A = A(t)$ with large diagonal entries. The idea behind the argument is to first obtain continuous factors $L(t), R(t)$ on small intervals that cover the real line and then stitch the different solutions together in a continuous manner.

Let us recall the notion of a matrix function. We denote by $M_{m \times n}(\mathbb{R})$ the set consisting of all $m \times n$ matrices with real entries. We will write $M_N(\mathbb{R})$ instead of $M_{N \times N}(\mathbb{R})$. A matrix function A is a function with some domain D and range in some $M_{m \times n}(\mathbb{R})$; i.e., it maps every $t \in D$ to some $m \times n$ matrix $A(t) = (a_{i,j}(t))$. Whenever the domain D is equipped with a topology (e.g., when D is a subset of \mathbb{R} with the usual distance) then we say that a matrix function A is continuous whenever all its entries $a_{i,j}$, viewed as scalar functions with domain D , are continuous. It is straightforward that for continuous matrix functions A, B with appropriate dimensions and common domain D the product AB is a continuous matrix function.

The first proposition of this section infers that to prove the main result it is enough to find continuous factors $L(t)$, $R(t)$ so that $L(t)A(t)R(t)$ is sufficiently close to the identity matrix for all t . We begin with two well-known lemmas, which we prove for the sake of completeness.

Lemma 3.1. *Let I be an interval of \mathbb{R} , $m, n \in \mathbb{N}$, and $A : I \rightarrow M_{m \times n}(\mathbb{R})$ be a matrix function. For any t_0 in I the matrix function A is continuous at t_0 if and only if $\lim_{t \rightarrow t_0} \|A(t) - A(t_0)\| = 0$.*

Proof. Note that for any $m \times n$ matrix $B = (b_{i,j})$ and any $1 \leq i_0 \leq m$, $1 \leq j_0 \leq n$ we have $|b_{i_0, j_0}(t)| = |\langle e_{i_0}, B e_{j_0} \rangle| \leq \|B\|$. By Corollary 2.2 we also have $\|B\| \leq m^{1/2} n^{1/2} \max_{i,j} |b_{i,j}|$. For $t \in I$ we apply our observation to the matrix $B = A(t) - A(t_0)$ to obtain that for any $1 \leq i_0 \leq m$, $1 \leq j_0 \leq n$ we have

$$|a_{i_0, j_0}(t) - a_{i_0, j_0}(t_0)| \leq \|A(t) - A(t_0)\| \leq m^{1/2} n^{1/2} \max_{i,j} |a_{i,j}(t) - a_{i,j}(t_0)|.$$

The desired conclusion immediately follows. \square

Lemma 3.2. *Let $N \in \mathbb{N}$, I be an interval of \mathbb{R} , and $A : I \rightarrow M_N(\mathbb{R})$ be a continuous matrix function such that $A(t)$ is invertible for all $t \in I$. Then $A^{-1} : I \rightarrow M_N(\mathbb{R})$ is a continuous matrix function.*

Proof. We fix t_0 in I and estimate $\|A^{-1}(t) - A^{-1}(t_0)\|$ for t close to t_0 . Observe that $A^{-1}(t) - A^{-1}(t_0) = A^{-1}(t)(A(t_0) - A(t))A^{-1}(t_0)$. We deduce

$$\|A^{-1}(t) - A^{-1}(t_0)\| \leq \|A^{-1}(t)\| \|A(t_0) - A(t)\| \|A^{-1}(t_0)\| \quad (3)$$

and

$$\|A^{-1}(t)\| \leq \|A^{-1}(t_0)\| + \|A^{-1}(t)\| \|A(t_0) - A(t)\| \|A^{-1}(t_0)\|,$$

which, solving for $\|A^{-1}(t)\|$, yields

$$\|A^{-1}(t)\| \leq \frac{\|A^{-1}(t_0)\|}{1 - \|A(t_0) - A(t)\| \|A^{-1}(t_0)\|}. \quad (4)$$

The quantity on the right-hand side of the inequality above is well-defined for t sufficiently close to t_0 . We plug (4) into (3) to get rid of the term $\|A^{-1}(t)\|$:

$$\|A^{-1}(t) - A^{-1}(t_0)\| \leq \frac{\|A^{-1}(t_0)\|^2 \|A(t_0) - A(t)\|}{(1 - \|A(t_0) - A(t)\| \|A^{-1}(t_0)\|)}.$$

This estimate, in conjunction with Lemma 3.1, yields that the continuity of $A : I \rightarrow M_N(\mathbb{R})$ at t_0 implies the continuity of $A^{-1} : I \rightarrow M_N(\mathbb{R})$ at t_0 . \square

Proposition 3.3. *Let $n \leq N \in \mathbb{N}$, I be an interval of \mathbb{R} , and $A : I \rightarrow M_N(\mathbb{R})$ be a continuous matrix function. Assume that $0 < C < 1$, $\Delta \geq 0$, and $L : I \rightarrow M_{n \times n}(\mathbb{R})$, $R : I \rightarrow M_{N \times n}(\mathbb{R})$ are continuous matrix functions so that for all $t \in I$ we have $\|L(t)A(t)R(t) - I_n\| \leq C$ and $\|L(t)\| \|R(t)\| \leq \Delta$. Then there exist continuous*

matrix functions $\tilde{L} : I \rightarrow M_{n \times N}(\mathbb{R})$, $\tilde{R} : I \rightarrow M_{N \times n}(\mathbb{R})$ so that for all $t \in I$ we have $\tilde{L}(t)A(t)\tilde{R}(t) = I_n$ and $\|\tilde{L}\|\|\tilde{R}\| \leq \Delta/(1-C)$.

Proof. For each $t \in I$, because we have that $\|L(t)A(t)R(t) - I_n\| \leq C$, the matrix $L(t)A(t)R(t)$ is invertible, and in particular $\|(L(t)A(t)R(t))^{-1}\| \leq 1/(1-C)$. By Lemma 3.2 the matrix function $(LAR)^{-1} : I \rightarrow M_n(\mathbb{R})$ is continuous. We define $\tilde{L} : I \rightarrow M_{n \times N}(\mathbb{R})$ as $\tilde{L}(t) = (L(t)A(t)R(t))^{-1}L(t)$ and just set $\tilde{R} = R$. Both \tilde{L} and \tilde{R} are continuous and clearly for all $t \in I$ we have $\tilde{L}(t)A(t)\tilde{R}(t) = I_n$. Additionally, for $t \in I$ we have $\|\tilde{L}(t)\|\|\tilde{R}\| \leq \|(L(t)A(t)R(t))^{-1}\|\|L\|\|R\| \leq \Delta/(1-C)$. \square

Recall the matrices $L_{(A,F)}$ and $R_{(A,F)}$ from Definition 2.7. In the sequel we will start with two versions of pairs $L_{(A,F_1)}$, $R_{(A,F_1)}$, $L_{(A,F_2)}$ and $R_{(A,F_2)}$, and a scalar $0 \leq \lambda \leq 1$. We will combine them into a new pair $L_{(A,F_1,F_2)}^\lambda$ and $R_{(A,F_1,F_2)}^\lambda$.

Definition 3.4. Let $n \leq N \in \mathbb{N}$, let $A = [a_1 \cdots a_N]$ be an $N \times N$ matrix, let $F_1 = \{i_1 < \cdots < i_n\}$, $F_2 = \{j_1 < \cdots < j_n\}$ be disjoint subsets of $\{1, \dots, N\}$, and let $0 \leq \lambda \leq 1$. We assume that $\|a_i\| > 0$ for $i \in F_1 \cup F_2$. Define the $N \times n$ and $n \times N$ matrices

$$\begin{aligned} R_{(A,F_1,F_2)}^\lambda &= \lambda^{1/2}R_{(A,F_1)} + (1-\lambda)^{1/2}R_{(A,F_2)}, \\ L_{(A,F_1,F_2)}^\lambda &= \lambda^{1/2}L_{(A,F_1)} + (1-\lambda)^{1/2}L_{(A,F_2)}. \end{aligned}$$

Remark 3.5. The matrices $R_{(A,F_1,F_2)}^\lambda$, $L_{(A,F_1,F_2)}^\lambda$ lie “between” $R_{(A,F_1)}$, $R_{(A,F_2)}$ and $L_{(A,F_1)}$, $L_{(A,F_2)}$ respectively. Clearly, if $\lambda = 1$ then

$$R_{(A,F_1,F_2)}^1 = R_{(A,F_1)}, \quad L_{(A,F_1,F_2)}^1 = L_{(A,F_1)}$$

and if $\lambda = 0$ then

$$R_{(A,F_1,F_2)}^0 = R_{(A,F_2)}, \quad L_{(A,F_1,F_2)}^0 = L_{(A,F_2)}.$$

Remark 3.6. Recall that for $k = 1, \dots, n$, we have $R_{(A,F_1)}e_k = e_{i_k}/\|a_{i_k}\|$ and $R_{(A,F_2)}e_k = e_{j_k}/\|a_{j_k}\|$, which means that

$$R_{(A,F_1,F_2)}^\lambda e_k = \lambda^{1/2}e_{i_k}/\|a_{i_k}\| + (1-\lambda)^{1/2}e_{j_k}/\|a_{j_k}\|.$$

Therefore

$$AR_{(A,F_1,F_2)}^\lambda = \left[\left(\lambda^{1/2} \frac{a_{i_1}}{\|a_{i_1}\|} + (1-\lambda)^{1/2} \frac{a_{j_1}}{\|a_{j_1}\|} \right) \cdots \left(\lambda^{1/2} \frac{a_{i_n}}{\|a_{i_n}\|} + (1-\lambda)^{1/2} \frac{a_{j_n}}{\|a_{j_n}\|} \right) \right].$$

Remark 3.7. It will be important to note for the sequel the following: if $n \leq N \in \mathbb{N}$, I is an interval of \mathbb{R} , $\lambda : I \rightarrow [0, 1]$ is a continuous scalar function, $A = [a_1 \cdots a_N] : I \rightarrow M_N(\mathbb{R})$ is a continuous matrix function, and F_1, F_2 are disjoint subsets of $\{1, \dots, N\}$ with $\#F_1 = \#F_2 = n$ so that $\|a_i(t)\| > 0$ for all $i \in F_1 \cup F_2$ and $t \in I$, then the matrix functions $R_{(F_1,F_2,A(t))}^{\lambda(t)} : I \rightarrow M_{N \times n}(\mathbb{R})$, $L_{(F_1,F_2,A(t))}^{\lambda(t)} : I \rightarrow M_{n \times N}(\mathbb{R})$ are both continuous.

The following proposition basically states that if we have appropriately picked $L_{(A, F_1)}$, $R_{(A, F_1)}$, $L_{(A, F_2)}$ and $R_{(A, F_2)}$ then for any scalar $0 \leq \lambda \leq 1$ the new pair $L_{(A, F_1, F_2)}^\lambda$, $R_{(A, F_1, F_2)}^\lambda$ satisfies a conclusion similar to that of Proposition 2.9.

Proposition 3.8. *Let $n \leq N \in \mathbb{N}$, let $A = [a_1 \cdots a_N]$ be an $N \times N$ matrix, let $F_1 = \{i_1 < \cdots < i_n\}$, $F_2 = \{j_1 < \cdots < j_n\}$ be disjoint subsets of $\{1, \dots, N\}$ and let $0 \leq \lambda \leq 1$. Set*

$$\theta = \min_{i \in F_1 \cup F_2} \|a_i\| \quad \text{and} \quad \varepsilon = \max_{i \neq j \in F_1 \cup F_2} |\langle a_i, a_j \rangle|.$$

If $\theta > 0$ then we have

$$\|R_{(A, F_1, F_2)}^\lambda\| \leq \theta^{-1}, \quad \|L_{(A, F_1, F_2)}^\lambda\| \leq 1 + \frac{(2n)^{1/2} \varepsilon^{1/2}}{\theta},$$

and

$$\|L_{(A, F_1, F_2)}^\lambda A R_{(A, F_1, F_2)}^\lambda - I_n\| \leq \frac{2n\varepsilon}{\theta^2}.$$

Proof. This proof is very similar in spirit to that of Proposition 2.9. We examine for $1 \leq k \leq n$ column k of $R_{(A, F_1, F_2)}^\lambda$, i.e., the vector $R_{(A, F_1, F_2)}^\lambda e_k$:

$$\|R_{(A, F_1, F_2)}^\lambda e_k\|^2 = \frac{\lambda}{\|a_{i_k}\|^2} + \frac{(1-\lambda)}{\|a_{j_k}\|^2} \leq \frac{1}{\theta^2}.$$

It is also easy to see that for $k_1 \neq k_2$ the columns of $R_{(A, F_1, F_2)}^\lambda$ are orthogonal. Therefore, by Proposition 2.1 we have $\|R_{(A, F_1, F_2)}^\lambda\| \leq 1/\theta$.

For the second estimate, we define, for $1 \leq k \leq n$,

$$b_k = \frac{\lambda^{1/2} a_{i_k}}{\|a_{i_k}\|} + \frac{(1-\lambda)^{1/2} a_{j_k}}{\|a_{j_k}\|}.$$

By Remark 3.6 we have

$$(L_{(A, F_1, F_2)}^\lambda)^T = A R_{(A, F_1, F_2)}^\lambda = [b_1 \cdots b_n].$$

We calculate, for $1 \leq k \leq n$, the norm of column k :

$$\begin{aligned} \|b_k\|^2 &= \left\langle \frac{\lambda^{1/2}}{\|a_{i_k}\|} a_{i_k} + \frac{(1-\lambda)^{1/2}}{\|a_{j_k}\|} a_{j_k}, \frac{\lambda^{1/2}}{\|a_{i_k}\|} a_{i_k} + \frac{(1-\lambda)^{1/2}}{\|a_{j_k}\|} a_{j_k} \right\rangle \\ &= \lambda + (1-\lambda) + 2\lambda^{1/2}(1-\lambda)^{1/2} \left\langle \frac{a_{i_k}}{\|a_{i_k}\|}, \frac{a_{j_k}}{\|a_{j_k}\|} \right\rangle; \end{aligned}$$

that is,

$$\|b_k\|^2 - 1 \leq 2\lambda^{1/2}(1-\lambda)^{1/2} \frac{\varepsilon}{\theta^2} \leq \frac{\varepsilon}{\theta^2} \quad \text{for } 1 \leq k \leq n, \quad (5)$$

where we used $0 \leq 2\lambda^{1/2}(1-\lambda)^{1/2} \leq 1$ for $0 \leq \lambda \leq 1$. In particular, we have

$$\|b_k\| \leq \left(1 + \frac{\varepsilon}{\theta^2}\right)^{1/2} \quad \text{for } 1 \leq k \leq n. \quad (6)$$

Next, we will show that

$$\text{for } 1 \leq k_1 \neq k_2 \leq n, \quad |\langle b_{k_1}, b_{k_2} \rangle| \leq 2 \frac{\varepsilon}{\theta^2}. \quad (7)$$

We have

$$\begin{aligned} |\langle b_{k_1}, b_{k_2} \rangle| &\leq \lambda \left\| \left\langle \frac{a_{i_{k_1}}}{\|a_{i_{k_1}}\|}, \frac{a_{i_{k_2}}}{\|a_{i_{k_2}}\|} \right\rangle \right\| + (1-\lambda) \left\| \left\langle \frac{a_{j_{k_1}}}{\|a_{j_{k_1}}\|}, \frac{a_{j_{k_2}}}{\|a_{j_{k_2}}\|} \right\rangle \right\| \\ &\quad + \lambda^{1/2}(1-\lambda)^{1/2} \left(\left\| \left\langle \frac{a_{i_{k_1}}}{\|a_{i_{k_1}}\|}, \frac{a_{j_{k_2}}}{\|a_{j_{k_2}}\|} \right\rangle \right\| + \left\| \left\langle \frac{a_{j_{k_1}}}{\|a_{j_{k_1}}\|}, \frac{a_{i_{k_2}}}{\|a_{i_{k_2}}\|} \right\rangle \right\| \right) \\ &\leq \frac{\varepsilon}{\theta^2} + 2\lambda^{1/2}(1-\lambda)^{1/2} \frac{\varepsilon}{\theta^2} \leq 2 \frac{\varepsilon}{\theta^2}. \end{aligned}$$

We now apply Proposition 2.1, which by (6) and (7), gives that

$$\begin{aligned} \|L_{(A, F_1, F_2)}^\lambda\| &= \|AR_{(A, F_1, F_2)}^\lambda\| \leq \left(1 + \frac{\varepsilon}{\theta^2} + (n-1)2 \frac{\varepsilon}{\theta^2} \right)^{1/2} \\ &\leq 1 + (2n-1)^{1/2} \frac{\varepsilon^{1/2}}{\theta} \leq 1 + (2n)^{1/2} \frac{\varepsilon^{1/2}}{\theta}. \end{aligned}$$

The final estimate follows from Corollary 2.3 directly applied to the matrix $AR_{(A, F_1, F_2)}^\lambda = [b_1 \cdots b_k]$ and (5), (7). \square

We are finally ready to state and prove the main result of this paper.

Theorem 3.9. *Let $N \in \mathbb{N}$, let I be an interval of \mathbb{R} and let $A = [a_1 \cdots a_N] : I \rightarrow M_N(\mathbb{R})$ be a continuous function so that the following hold:*

- (i) *For $t \in I$ we have $\|A(t)\| \leq 1$.*
- (ii) *$\theta = \inf_{t \in I} \min_{1 \leq i \leq N} \|a_i(t)\| > 0$.*

Then for every $1 \leq n \leq \frac{1}{12} \theta^{4/3} N^{1/3}$ there exist continuous functions $L : I \rightarrow M_n \times N(\mathbb{R})$ and $R : I \rightarrow M_{N \times n}(\mathbb{R})$ so that for all $t \in I$ we have $L(t)A(t)R(t) = I_n$ and $\|L(t)\| \|R(t)\| \leq 2/\theta$.

Proof. By Proposition 3.3 it is sufficient to find continuous $L(t)$, $R(t)$ so that for all $t \in I$ we have $\|L(t)A(t)R(t) - I_n\| \leq \frac{1}{4}$ and $\|L(t)\| \|R(t)\| \leq 4/(3\theta)$.

The case $n = 1$ is treated easily by taking an arbitrary $1 \leq i \leq N$ and defining $R(t) = e_i / \|a_i(t)\|$ and $L(t) = a_i(t) / \|a_i(t)\|$; thus we assume that $2 \leq n \leq \frac{1}{12} \theta^{4/3} N^{1/3}$. Define $\varepsilon = \theta^2 / (18n)$. This choice of ε is related to the estimates from Proposition 3.8 and also Corollaries 2.5 and 2.6. Let us note that we have

$$\frac{(2n)^{1/2} \varepsilon^{1/2}}{\theta} = \frac{1}{3} \quad \text{and} \quad \frac{2n\varepsilon}{\theta^2} \leq \frac{1}{4} \quad (8)$$

and also

$$5 \frac{n}{\varepsilon^2} = 5 \frac{18^2 n^3}{\theta^4} \leq 5 \frac{18^2 \theta^4 N}{\theta^4 12^3} \leq N. \quad (9)$$

Let us assume henceforth that $I = [0, \infty)$. The case $I = \mathbb{R}$ is treated by performing the same argument on both sides of 0. Other cases are treated similarly. Otherwise they can be deduced from the previous two cases by using, e.g., that any open interval is homeomorphic to \mathbb{R} and every half-open interval is homeomorphic to $[0, +\infty)$, and any continuous function on a closed bounded interval $[t_1, t_2]$ can be continuously extended to \mathbb{R} by assigning the value $A(t_1)$ to each $t \leq t_1$ and the value $A(t_2)$ to each $t \geq t_2$.

We start by finding a strictly increasing sequence $0 = t_0 < t_1 < t_2 < \dots$ with $\lim_m t_m = \infty$ so that for all $m \in \mathbb{N}$ there exists $F_m \subset \{1, \dots, N\}$ with

- (a) $\#F_m = n$ and
- (b) for all $i \neq j \in F_m$ and $t_{m-1} \leq t \leq t_m$ we have $|\langle a_i(t), a_j(t) \rangle| < \varepsilon$.

This is achieved as follows. For each $r \in [0, 1]$ we use Corollary 2.5 to find $F_r \subset \{1, \dots, N\}$ so that for all $i \neq j \in F_r$ we have $|\langle a_i(t), a_j(t) \rangle| < \varepsilon$. Because A is continuous, we may find a small open interval I_r containing r (half-open if $r = 0$) so that for all $i \neq j \in F_r$ and $t \in I_r$ we still have $|\langle a_i(t), a_j(t) \rangle| < \varepsilon$. Because $[0, 1] \subset \bigcup_{r \in [0, 1]} I_r$ and the interval $[0, 1]$ is compact there must exist $r_1 < \dots < r_{m_1}$ so that $[0, 1] \subset \bigcup_{i=1}^{m_1} I_{r_i}$. By perhaps getting rid of a few intervals we may assume that none of them is contained in the union of the others. Then, by perhaps making some of the intervals a little shorter we may assume that $\sup(I_{r_i}) \leq r_{i+1}$ for $1 \leq i < m_1 - 1$ and $r_{i-1} \leq \inf(I_{r_i})$ for $1 < i \leq m_1$. In other words, for $i = 1, \dots, m_1 - 1$ we have $\emptyset \neq I_{r_i} \cap I_{r_{i+1}} \subset (r_i, r_{i+1})$. Define $t_0 = 0$, $t_{m_1} = 1$ and for $1 \leq i < m_1$ pick $t_i \in (r_i, r_{i+1})$. If we then set $F_i = F_{r_i}$ for $1 \leq i \leq m_1$ we obtain that (a) and (b) are satisfied up to $m = m_1$. For $k = 2, 3, \dots$ repeat the same argument on $[k-1, k]$ to find $(t_i)_{i=m_{k-1}+1}^{m_k}$ and $(F_i)_{i=m_{k-1}+1}^{m_k}$ that satisfy (a) and (b).

The next step is to apply for each $m = 1, 2, \dots$ Corollary 2.6 to the matrix $A(t_m)$ and the sets F_m, F_{m+1} . By doing so we find a set $G_m \subset \{1, \dots, N\} \setminus (F_m \cup F_{m+1})$ with $\#G_m = n$ so that for all $i \neq j$ with $i \in G_m$ and $j \in G_m \cup F_m \cup F_{m+1}$ we have $|\langle a_i(t_m), a_j(t_m) \rangle| < \varepsilon$. We now use the continuity of A once more to find $s_m < t_m < u_m$ so that for all $t \in (s_m, u_m)$ the above hold as well. By perhaps moving s_m, u_m a bit closer to t_m we have the following situation:

- (c) $0 = t_0 < s_1 < t_1 < u_1 < s_2 < t_2 < u_2 < s_3 < t_3 < u_3 < \dots$.
- (d) For $m = 1, 2, \dots$ we have $G_m \subset \{1, \dots, N\} \setminus (F_m \cup F_{m+1})$ with $\#G_m = n$ so that for all $t \in (s_m, u_m)$, $i \neq j$, with $i \in G_m$ and $j \in G_m \cup F_m \cup F_{m-1}$, we have $|\langle a_i(t), a_j(t) \rangle| < \varepsilon$.

We are finally ready to define $L(t)$ and $R(t)$. Set $0 = u_0$. For each $m = 1, 2, \dots$ take a continuous $\lambda_m : [s_m, u_m] \rightarrow [0, 1]$ with $\lambda_m(s_m) = \lambda_m(u_m) = 1$ and $\lambda_m(t_m) = 0$:

- (A) For $m = 0, 1, \dots$ and $t \in [u_m, s_{m+1}]$ set $R(t) = R_{(A(t), F_{m+1})}$.

(B) For $m = 1, 2, \dots$ and $t \in [s_m, t_m]$ define

$$R(t) = R_{A(t), F_m, G_m}^{\lambda_m(t)}.$$

We point out that, by Remark 3.5,

$$R(s_m) = R_{A(s_m), F_m, G_m}^1 = R_{(A(s_m), F_m)} \quad \text{and} \quad R(t_m) = R_{A(t_m), F_m, G_m}^0 = R_{A(t_m), G_m}.$$

(C) For $m = 1, 2, \dots$ and $t \in [t_m, u_m]$ define

$$R(t) = R_{A(t), F_{m+1}, G_m}^{\lambda_m(t)}.$$

Once more, by Remark 3.5,

$$R(t_m) = R_{A(t_m), F_{m+1}, G_m}^0 = R_{(A(t_m), G_m)} \quad \text{and} \quad R(u_m) = R_{A(u_m), F_{m+1}, G_m}^1 = R_{A(u_m), F_{m+1}}.$$

By Remark 3.7, in each case (A), (B), and (C) the function R is continuous and the values at the endpoints of the corresponding intervals match. Thus R defines a continuous function on I and thus so does $L = (AR)^T$.

We next wish to show that for $t \geq 0$ we have $\|L(t)A(t)R(t) - I_n\| \leq \frac{1}{4}$ and $\|L(t)\| \|R(t)\| \leq 4/(3\theta)$ and the proof will be complete. If $t \in [u_m, s_{m+1}]$, for some $m \in \mathbb{N}$, then this follows from definition (A) above and (8) applied to Proposition 2.9. If $t \in [s_m, u_m]$ for some $m \in \mathbb{N}$ then this follows from definition (B) or (C), property (d), and (8) applied to Proposition 3.8. \square

We conclude with some open questions regarding the topic of the paper.

Question 1. As was pointed out in Remark 2.11, [Bourgain and Tzafriri 1987] implies a version of Theorem 2.10 (in which $2/\theta$ is replaced by C/θ and C is a nonexplicit finite constant) with an estimate $n \gtrsim \theta^2 N$. This is better than our estimate $n \gtrsim \theta^{3/4} N^{1/3}$, provided that $N \gtrsim 1/\theta$. Can the probabilistic technique from [Bourgain and Tzafriri 1987] be used to obtain a similar version of the continuous Theorem 3.9 with an estimate $n \gtrsim \theta^2 N$?

Question 2. For the theorem in the continuous case, we considered $A : I \rightarrow M_N(\mathbb{R})$, where I is an interval of \mathbb{R} . We conjecture that a version of Theorem 3.9 is also true for a continuous matrix function $A : \mathbb{R}^d \rightarrow M_N(\mathbb{R})$. What is the relation between d , N , θ , and the dimension n in the conclusion of such a theorem?

For $1 \leq p \leq \infty$ and an $N \times N$ matrix A let $\|A\|_p$ denote the quantity $\max\{\|Ax\|_p : \|x\|_p \leq 1\}$. In particular, $\|A\| = \|A\|_2$.

Question 3. The methods used in this paper rely heavily on properties of the Euclidean norm. In the statement of Theorem 3.9 we may replace condition (i) with $\|A(t)\|_p \leq 1$. It would be interesting to prove a version of this theorem, as different methods might be necessary.

Acknowledgements

We would like to thank the anonymous referee for recommending the inclusion of Question 1.

This work was done in the research project “Continuous factorization of the identity matrix” at the Illinois Geometry Lab in Spring 2019. Dai, Jiao, and Lan participated as undergraduate scholars, Hore served as graduate student team leader, and Motakis as faculty mentor. The project was supported by the National Science Foundation under grant number DMS-1449269. Motakis was supported by the National Science Foundation under grant number DMS-1912897.

References

- [Andrew 1979] A. D. Andrew, “Perturbations of Schauder bases in the spaces $C(K)$ and L^p , $p > 1$ ”, *Studia Math.* **65**:3 (1979), 287–298. MR Zbl
- [Bourgain and Tzafriri 1987] J. Bourgain and L. Tzafriri, “Invertibility of “large” submatrices with applications to the geometry of Banach spaces and harmonic analysis”, *Israel J. Math.* **57**:2 (1987), 137–224. MR Zbl
- [Laustsen et al. 2018] N. J. Laustsen, R. Lechner, and P. F. X. Müller, “Factorization of the identity through operators with large diagonal”, *J. Funct. Anal.* **275**:11 (2018), 3169–3207. MR Zbl
- [Lechner 2017] R. Lechner, “Direct sums of finite dimensional SL_n^∞ spaces”, preprint, 2017. arXiv
- [Lechner 2018a] R. Lechner, “Factorization in mixed norm Hardy and BMO spaces”, *Studia Math.* **242**:3 (2018), 231–265. MR Zbl
- [Lechner 2018b] R. Lechner, “Factorization in SL^∞ ”, *Israel J. Math.* **226**:2 (2018), 957–991. MR Zbl
- [Lechner 2019a] R. Lechner, “Dimension dependence of factorization problems: biparameter Hardy spaces”, *Proc. Amer. Math. Soc.* **147**:4 (2019), 1639–1652. MR Zbl
- [Lechner 2019b] R. Lechner, “Dimension dependence of factorization problems: Hardy spaces and SL_n^∞ ”, *Israel J. Math.* **232**:2 (2019), 677–693. MR Zbl
- [Lechner 2019c] R. Lechner, “Subsymmetric weak* Schauder bases and factorization of the identity”, *Studia Math.* **248**:3 (2019), 295–319. MR Zbl
- [Lechner et al. 2018] R. Lechner, P. Motakis, P. F. X. Müller, and T. Schlumprecht, “Strategically reproducible bases and the factorization property”, preprint, 2018. arXiv
- [Meckes and Meckes 2018] E. S. Meckes and M. W. Meckes, *Linear algebra*, Cambridge University Press, 2018. Zbl

Received: 2019-08-31

Revised: 2019-10-07

Accepted: 2019-10-14

yuyingd2@illinois.edu

Department of Mathematics, University of Illinois at Urbana-Champaign, Urbana, IL, United States

ahore2@illinois.edu

Department of Mathematics, University of Illinois at Urbana-Champaign, Urbana, IL, United States

sjiao2@illinois.edu

Department of Mathematics, University of Illinois at Urbana-Champaign, Urbana, IL, United States

tl2971@columbia.edu

*Department of Mathematics, University of Illinois at
Urbana-Champaign, Urbana, IL, United States*

pmotakis@illinois.edu

*Department of Mathematics, University of Illinois at
Urbana-Champaign, Urbana, IL, United States*

Almost excellent unique factorization domains

Sarah M. Fleming and Susan Loepp

(Communicated by Scott T. Chapman)

Let (T, \mathfrak{m}) be a complete local (Noetherian) domain such that $\text{depth } T > 1$. In addition, suppose T contains the rationals, $|T| = |T/\mathfrak{m}|$, and the set of all principal height-1 prime ideals of T has the same cardinality as T . We construct a universally catenary local unique factorization domain A such that the completion of A is T and such that there exist uncountably many height-1 prime ideals \mathfrak{q} of A such that $(T/(\mathfrak{q} \cap A)T)_{\mathfrak{q}}$ is a field. Furthermore, in the case where T is a normal domain, we can make A “close” to excellent in the following sense: the formal fiber at every prime ideal of A of height not equal to 1 is geometrically regular, and uncountably many height-1 prime ideals of A have geometrically regular formal fibers.

1. Introduction

One important area of study in commutative algebra is completions of local rings. Many mathematicians have worked to characterize the completions of certain classes of rings. For example, Heitmann [1993] found surprisingly weak necessary and sufficient conditions for a ring to be the completion of a local unique factorization domain (UFD), and Loepp [2003] characterized the completions of local excellent integral domains of characteristic zero. The question of when a ring is the completion of an excellent UFD, the natural extension of these two results, is still open. Bryk et al. [2005] obtained a partial result by constructing what they term “almost excellent UFDs”. Boocher et al. [2010] also attempted to characterize the completions of excellent UFDs and arrived at several necessary conditions for the construction.

One technique for studying the relationship between a ring and its completion is to use the formal fibers, which encode information about the prime ideals of the ring.

Definition 1.1. Let A be a local (Noetherian) ring and let \hat{A} denote the completion of A with respect to its maximal ideal. For a prime ideal \mathfrak{p} of A , the *formal fiber* of A at \mathfrak{p} is given by the set $\text{Spec}(\hat{A} \otimes_A \kappa(\mathfrak{p}))$, where $\kappa(\mathfrak{p})$ is the residue field $A_{\mathfrak{p}}/\mathfrak{p}A_{\mathfrak{p}}$. For a local domain A , the *generic formal fiber* of A is the formal fiber of A at the zero ideal, which is the set $\text{Spec}(\hat{A} \otimes_A K)$, where K is the quotient field of A .

MSC2010: primary 13F15, 13F40; secondary 13B35, 13J10.

Keywords: completions of local rings, excellent rings, unique factorization domains.

It is also possible to characterize the formal fibers in terms of the prime ideals of A and \hat{A} . Let $\varphi : \text{Spec } \hat{A} \rightarrow \text{Spec } A$ be the map that sends a prime ideal $\mathfrak{q} \in \text{Spec } \hat{A}$ to $\mathfrak{q} \cap A$. There is a one-to-one correspondence between the inverse image of an ideal \mathfrak{p} under this morphism and the elements of $\text{Spec}(\hat{A} \otimes_A \kappa(\mathfrak{p}))$: the prime ideals of $\text{Spec}(\hat{A} \otimes_A \kappa(\mathfrak{p}))$ are of the form $\mathfrak{q} \otimes_A \kappa(\mathfrak{p})$ where $\mathfrak{q} \cap A = \mathfrak{p}$.

Our ultimate goal is to characterize the completions of excellent UFDs, but in order to break this problem down into a more manageable one, we consider what it means for a ring to be “close” to excellent. In order for a ring A to be excellent, all of its formal fibers must be geometrically regular. What if there were a way to get “most” of the formal fibers of A to be geometrically regular?

We first work towards getting a large number of the formal fibers of A at height-1 prime ideals to be geometrically regular—in fact, uncountably many. More specifically, we pose the following question, letting \mathcal{S} denote the set of principal height-1 prime ideals of T .

Question 1.2. Let T be a complete local domain that contains the rationals. When does there exist a local universally catenary UFD A such that $\hat{A} \cong T$ and there is an uncountable subset \mathcal{S}' of \mathcal{S} such that

- (1) if $\mathfrak{q} \in \mathcal{S}'$, then $\mathfrak{q} \cap A \neq (0)$,
- (2) if $\mathfrak{q}, \mathfrak{q}' \in \mathcal{S}'$ with $\mathfrak{q} \neq \mathfrak{q}'$, then $\mathfrak{q} \cap A \neq \mathfrak{q}' \cap A$, and
- (3) if $\mathfrak{q} \in \mathcal{S}'$, then the formal fiber of A at $\mathfrak{q} \cap A$ is geometrically regular?

Letting \mathfrak{m} denote the maximal ideal of T , we prove that, if $\dim T = 2$, $\text{depth } T > 1$, and $|T| = |T/\mathfrak{m}| = |\mathcal{S}|$, then there does exist a universally catenary UFD A whose completion is T that satisfies the above conditions. If $\dim T > 2$, then we prove that there exists a universally catenary UFD A with $\hat{A} = T$ that satisfies conditions (1) and (2) and a weaker version of (3): if \mathfrak{q} is in \mathcal{S}' , then $(T/(\mathfrak{q} \cap A)T)_{\mathfrak{q}}$ is a field. Our construction is a first step towards constructing an excellent UFD; however, we have no control over (i) the generic formal fiber or (ii) the formal fibers at prime ideals of A of height greater than 1.

If we assume T is a complete local normal domain containing the rationals with $\text{depth } T > 1$ and $|T| = |T/\mathfrak{m}| = |\mathcal{S}|$, then we show there exists a local universally catenary UFD A where $\hat{A} = T$ and such that A satisfies conditions (1), (2), and (3), along with the additional condition that the formal fibers of A at every prime ideal of height not equal to 1 is geometrically regular.

Thus, given a complete local normal domain T containing the rationals where $\text{depth } T > 1$ and $|T| = |T/\mathfrak{m}| = |\mathcal{S}|$, we construct a local universally catenary UFD A with completion T that is “close” to excellent; “close” in the sense that A is universally catenary and the formal fiber at every prime ideal—except perhaps uncountably many height-1 prime ideals—is geometrically regular. Further, we

construct A so that the formal fibers at uncountably many of the height-1 prime ideals of A are geometrically regular. This is a new result, as previously, mathematicians have only been able to construct UFDs with countably many of the height-1 prime ideals having geometrically regular formal fibers; see [Boocher et al. 2010].

2. Preliminaries

In this paper, all rings are assumed to be commutative with unity and local rings are assumed to be Noetherian. If a ring has exactly one maximal ideal, but is not necessarily Noetherian, we say it is quasilocal. Often, we denote a quasilocal ring R with maximal ideal \mathfrak{m} by (R, \mathfrak{m}) . We recall the definition of geometrically regular and what it means for a local ring to be excellent.

Definition 2.1. Let k be a field, and let B be a Noetherian k -algebra. Then B is called *geometrically regular over k* if, for every finite field extension $k \subseteq L$, the ring $B \otimes_k L$ is regular.

Definition 2.2. A local ring A is *excellent* if it satisfies the following conditions:

- (1) A is universally catenary.
- (2) The formal fibers of A are geometrically regular; that is, for every $\mathfrak{p} \in \text{Spec } A$, the ring $\hat{A} \otimes_A L$ is regular for all finite field extensions L of $\kappa(\mathfrak{p})$.

Theorem 31.7 in [Matsumura 1986] states that, if A is a local ring, then it is formally catenary if and only if it is universally catenary. In addition, if A is a domain, formally equidimensional and universally catenary are equivalent, so a domain A is universally catenary if and only if its completion T is equidimensional by Theorem 31.6 in [Matsumura 1986]. Therefore, in order for a ring T to be the completion of a universally catenary domain, T must be equidimensional.

To show that the formal fibers of A are geometrically regular, it is enough to consider purely inseparable finite field extensions L of $\kappa(\mathfrak{p})$ for every $\mathfrak{p} \in \text{Spec } A$. If A contains the rationals, then $\kappa(\mathfrak{p})$ is a field of characteristic zero for every $\mathfrak{p} \in \text{Spec } A$, so the only purely inseparable finite field extension of $\kappa(\mathfrak{p})$ is $\kappa(\mathfrak{p})$ itself. Thus, if A contains the rationals, we need show only that the ring $\hat{A} \otimes_A \kappa(\mathfrak{p})$ is geometrically regular for every $\mathfrak{p} \in \text{Spec } A$.

Furthermore, if \hat{A} contains the rationals, it is sufficient to show that, for every $\mathfrak{p} \in \text{Spec } A$, if $\mathfrak{q} \in \text{Spec } \hat{A}$ and $\mathfrak{q} \cap A = \mathfrak{p}$, then $(\hat{A}/\mathfrak{p}\hat{A})_{\mathfrak{q}}$ is a regular local ring. This is a much easier definition to work with than the tensor product definition, and it is the definition we use since our \hat{A} contains the rationals.

We now establish several necessary conditions for a ring T to be the completion of an excellent local UFD A . For the rest of this section, let A denote an excellent local UFD. We begin with:

Theorem 2.3 [Heitmann 1993, Theorem 1]. *Let R be an integrally closed local domain. Then no element of the prime subring of R is a zero divisor in \widehat{R} . Moreover, \widehat{R} is either a field, a discrete valuation ring, or a ring with depth at least 2.*

In order for a complete local ring T to be the completion of an excellent local UFD, T , of course, must be the completion of a UFD. Note that unique factorization domains are integrally closed. Thus, if T has Krull dimension at least 2, then, by Theorem 2.3, it must have depth at least 2. In addition, T must satisfy the condition that no element of the prime subring of T is a zero divisor.

Since we know that the completion map is faithfully flat and regular if A is excellent, A is normal if and only if \widehat{A} is normal. Since A is a UFD, it is normal. Thus, if T is the completion of an excellent local UFD, it must be normal.

Furthermore, we have already established that a domain is universally catenary if and only if it is formally equidimensional; thus if A is an excellent domain, \widehat{A} must be equidimensional. So for T to be the completion of an excellent local UFD, T must be equidimensional.

To summarize, suppose T has Krull dimension at least 2 and is the completion of an excellent local UFD. Then the following statements must be true:

- (1) No element of the prime subring of T is a zero divisor in T and $\text{depth } T > 1$.
- (2) T is normal.
- (3) T is equidimensional.

Because of these necessary conditions, we often assume that T satisfies the above three conditions.

Given such a complete local ring T , we now discuss how we construct an “almost excellent” local UFD A whose completion is T . Our general construction is as follows: we start with a localization of the prime subring of T and then build up intermediate subrings, adjoining generators of prime ideals in \mathcal{S} , the set of principal height-1 prime ideals of T , along the way. Here we present, without proof, several lemmas we use throughout our construction.

First, we must show that the completion of the subring we construct is in fact T . In order to show that the constructed subring is Noetherian and has completion T , we use the following lemma.

Lemma 2.4 [Heitmann 1994, Proposition 1]. *Let $(R, \mathfrak{m} \cap R)$ be a quasilocal subring of a complete local ring (T, \mathfrak{m}) . If $R \rightarrow T/\mathfrak{m}^2$ is surjective, and if $\mathfrak{a}T \cap R = \mathfrak{a}$ for every finitely generated ideal \mathfrak{a} of R , then R is Noetherian and the natural homomorphism $\widehat{R} \rightarrow T$ is an isomorphism.*

Lemma 2.5 yields restrictions on the cardinality of the constructed subrings.

Lemma 2.5 [Charters and Loepp 2004, Lemma 2.2]. *Let (T, \mathfrak{m}) be a complete local ring of dimension at least 1. Let \mathfrak{p} be a nonmaximal prime ideal of T . Then $|T/\mathfrak{p}| = |T| \geq 2^{\aleph_0}$.*

Throughout the construction, we adjoin infinitely many elements of T to intermediate subrings while preserving cardinality properties and properties of certain prime ideals. (We elaborate on these properties in the following section.) Lemmas 2.6 and 2.7 from [Heitmann 1993] allow us to do so by avoiding certain cosets of certain prime ideals. Lemma 2.6 deals with the situation where the number of prime ideals to avoid is countable, whereas Lemma 2.7 is for avoiding uncountably many prime ideals.

Lemma 2.6 [Heitmann 1993, Lemma 2]. *Let (T, \mathfrak{m}) be a complete local ring and let D be a subset of T . Suppose C is a subset of T such that $\mathfrak{m} \notin C$, and suppose \mathfrak{a} is an ideal of T such that $\mathfrak{a} \not\subseteq \mathfrak{p}$ for all $\mathfrak{p} \in C$. If C and D are countable, then*

$$\mathfrak{a} \not\subseteq \bigcup \{t + \mathfrak{p} \mid t \in D, \mathfrak{p} \in C\},$$

where $t + \mathfrak{p}$ denotes the coset of \mathfrak{p} in T ; that is, $t + \mathfrak{p} = \{t + x \mid x \in \mathfrak{p}\}$.

Lemma 2.7 [Heitmann 1993, Lemma 3]. *Let (T, \mathfrak{m}) be a local ring and let D be a subset of T . Suppose C is a subset of T and \mathfrak{a} is an ideal of T such that $\mathfrak{a} \not\subseteq \mathfrak{p}$ for all $\mathfrak{p} \in C$. If $|C \times D| < |T/\mathfrak{m}|$, then*

$$\mathfrak{a} \not\subseteq \bigcup \{t + \mathfrak{p} \mid t \in D, \mathfrak{p} \in C\}.$$

The following lemma from [Loepp 1997] allows us to adjoin generators of principal prime ideals of T to intermediate subrings. (We require only a weaker version of this statement.)

Lemma 2.8 [Loepp 1997, Lemma 4]. *Suppose (T, \mathfrak{m}) is a local ring with infinite residue field. Let C_1 and C_2 be subsets of T and let $u, w \in T$ such that $u \notin \mathfrak{p}$ for every $\mathfrak{p} \in C_1$ and $w \notin \mathfrak{q}$ for every $\mathfrak{q} \in C_2$. Also, suppose D_1 and D_2 are subsets of T . If $|C_1 \times D_1| < |T/\mathfrak{m}|$ and $|C_2 \times D_2| < |T/\mathfrak{m}|$, then there exists a unit $x \in T$ such that*

$$ux \notin \bigcup \{t + \mathfrak{p} \mid t \in D_1, \mathfrak{p} \in C_1\} \quad \text{and} \quad wx^{-1} \notin \bigcup \{t + \mathfrak{p} \mid t \in D_2, \mathfrak{p} \in C_2\}.$$

3. The construction

We follow closely the construction of a UFD in [Heitmann 1993], so several lemmas and definitions are taken directly from that work.

Definition 3.1. Let (T, \mathfrak{m}) be a complete local ring, and let $(R, R \cap \mathfrak{m})$ be a quasiloocal unique factorization domain contained in T satisfying the following:

- (i) $|R| \leq \sup(\aleph_0, |T/\mathfrak{m}|)$ with equality only if T/\mathfrak{m} is countable.
- (ii) $\mathfrak{q} \cap R = (0)$ for all $\mathfrak{q} \in \text{Ass } T$.
- (iii) If $t \in T$ is regular and $\mathfrak{q} \in \text{Ass}(T/tT)$, then $\text{ht}(\mathfrak{q} \cap R) \leq 1$.

Then R is called an N -subring of T .

We construct “good” ring extensions of N -subrings inside T by adjoining elements, using the techniques and terminology from [Heitmann 1993], so that the eventual result is a UFD.

Definition 3.2. If R and S are N -subrings of T with $R \subseteq S$, we say that S is an A -extension of R if

- (i) prime elements of R are prime in S , and
- (ii) $|S| \leq \sup(\aleph_0, |R|)$.

We now mimic Heitmann’s construction of a UFD that has completion T , adding on additional steps that guarantee that our UFD satisfies the additional conditions (1)–(3) of Question 1.2. We start with a lemma from [Heitmann 1993]. Since several of our later proofs will be based on the proof of Lemma 3.3, we include a detailed proof of the lemma here.

Lemma 3.3 [Heitmann 1993, Lemma 5]. *Let (T, \mathfrak{m}) be a complete local ring with depth $T > 1$, let R be an N -subring of T , and let $t \in T$. Then there exists an infinite A -extension S of R such that $t + \mathfrak{m}^2 \in \text{Image}(S \rightarrow T/\mathfrak{m}^2)$.*

Proof. Note that this proof is taken almost directly from the proof of Lemma 3.5 in [Fleming et al. 2019]. Let

$$C = \{\mathfrak{p} \in \text{Spec } T \mid \mathfrak{p} \in \text{Ass}(T/rT) \text{ with } 0 \neq r \in R\} \cup \text{Ass } T.$$

Since depth $T > 1$, we have $\mathfrak{m} \notin C$, so $\mathfrak{m}^2 \not\subseteq \mathfrak{p}$ for every $\mathfrak{p} \in C$. For every $\mathfrak{p} \in C$, define a subset $D_{(\mathfrak{p})}$ of T : $D_{(\mathfrak{p})}$ is to be a full set of distinct coset representatives u in T for those cosets $u + \mathfrak{p}$ in T/\mathfrak{p} such that $(u + t) + \mathfrak{p} \in T/\mathfrak{p}$ is algebraic over $R/(\mathfrak{p} \cap R)$. (Note that the map $R/(\mathfrak{p} \cap R) \rightarrow T/\mathfrak{p}$ is an injection, so this language makes sense.) Set $D = \bigcup_{\mathfrak{p} \in C} D_{(\mathfrak{p})}$. If R is countable, then C is countable, and therefore D is also countable. Otherwise, $|R| < |T/\mathfrak{m}|$ and so $|C \times D| < |T/\mathfrak{m}|$. By Lemma 2.6 if R is countable and by Lemma 2.7 otherwise, choose $x \in \mathfrak{m}^2$ such that $(x + t) + \mathfrak{p}$ is transcendental over $R/(\mathfrak{p} \cap R)$ for all $\mathfrak{p} \in C$. Define $R' = R[x + t]_{\mathfrak{m} \cap R[x+t]}$. We claim that R' is an N -subring. In fact, R' is an A -extension of R .

Suppose $\mathfrak{p} \in C$. We show that $\mathfrak{p} \cap R' = (\mathfrak{p} \cap R)R'$. Since $R' = R[x + t]_{\mathfrak{m} \cap R[x+t]}$, elements of $\mathfrak{p} \cap R'$ are of the form uf , where $u \in R'^{\times}$ is a unit and $f \in R[x + t]$. Since u is a unit, this gives us that $f \in \mathfrak{p}$. We can treat f as a polynomial in $x + t$ over R . Because $x + t + \mathfrak{p}$ is transcendental over $R/(\mathfrak{p} \cap R)$, each of the coefficients of f must be in $\mathfrak{p} \cap R$. Then $f \in (\mathfrak{p} \cap R)R[x + t]$ and $uf \in (\mathfrak{p} \cap R)R'$. Therefore, $\mathfrak{p} \cap R' \subseteq (\mathfrak{p} \cap R)R'$. The opposite containment is clear; thus, we have equality.

Since R is a UFD and $x + t$ is transcendental over R , it follows that $R[x + t]$ is a UFD. Any localization of a UFD is a UFD, so R' is also a UFD. In creating R' , we are simply adjoining a transcendental element and localizing; thus, $|R'| = \sup(\aleph_0, |R|)$ and R' satisfies condition (i) of Definition 3.1. Suppose $\mathfrak{q} \in \text{Ass } T$. Then, since R

is an N -subring, we have $\mathfrak{q} \cap R' = (\mathfrak{q} \cap R)R' = (0)$, so R' satisfies condition (ii) of Definition 3.1.

Let $\mathfrak{p} \in \text{Ass}(T/rT)$ for some regular $r \in R$. First suppose $\mathfrak{p} \cap R = (0)$. Then, in $R'_{\mathfrak{p} \cap R'} = R[x+t]_{\mathfrak{p} \cap R[x+t]}$, all nonzero elements of R are units, so $R'_{\mathfrak{p} \cap R'}$ is isomorphic to $k[X]$ with additional elements inverted where k is a field and X is an indeterminate. Therefore, $\dim R'_{\mathfrak{p} \cap R'} \leq 1$ and we have $\text{ht}(\mathfrak{p} \cap R') \leq 1$.

Now suppose $\mathfrak{p} \cap R = aR$ for some nonzero $a \in R$. By Theorem 6.2 of [Matsumura 1986], we know that $\mathfrak{p} \in \text{Ass}(T/rT)$ if and only if $\mathfrak{p}T_{\mathfrak{p}} \in \text{Ass}(T_{\mathfrak{p}}/rT_{\mathfrak{p}})$. Thus $\mathfrak{p}T_{\mathfrak{p}}$ consists of zero divisors of $T_{\mathfrak{p}}/rT_{\mathfrak{p}}$, and $T_{\mathfrak{p}}/rT_{\mathfrak{p}}$ consists only of zero divisors and units. Therefore, $T_{\mathfrak{p}}/rT_{\mathfrak{p}}$ has depth zero, and since $a \in R$ is regular and in \mathfrak{p} , we know $T_{\mathfrak{p}}/aT_{\mathfrak{p}}$ must also have depth zero. Then $\mathfrak{p}T_{\mathfrak{p}}$ consists only of zero divisors of $T_{\mathfrak{p}}/aT_{\mathfrak{p}}$, so $\mathfrak{p}T_{\mathfrak{p}} \in \text{Ass}(T_{\mathfrak{p}}/aT_{\mathfrak{p}})$. Thus $\mathfrak{p} \in \text{Ass}(T/aT)$ and $\mathfrak{p} \in C$, so $\mathfrak{p} \cap R' = (\mathfrak{p} \cap R)R' = aR'$.

We now show that any principal prime ideal of R' must have height at most 1. Suppose that there exists some prime ideal \mathfrak{q} of R' such that $(0) \subsetneq \mathfrak{q} \subsetneq aR'$. We show that $\mathfrak{q} = (0)$. Let $y \in \mathfrak{q}$. Then $y = ar_1$ for some $r_1 \in R'$. Since $\mathfrak{q} \subsetneq aR'$, we must have $a \notin \mathfrak{q}$. Thus $r_1 \in \mathfrak{q} \subsetneq aR'$, so we can write r_1 as $r_1 = ar_2$ for some $r_2 \in R'$. Again, since $a \notin \mathfrak{q}$ but $r_1 \in \mathfrak{q}$, we must have $r_2 \in \mathfrak{q} \subsetneq aR'$, so we can write r_2 as $r_2 = ar_3$. This gives us that $y = ar_1 = a^2r_2 = a^3r_3$. Continue this process. Looking in T , we have $y \in \bigcap_{i=1}^{\infty} a^i T$; however, since T is Noetherian, we know that $\bigcap_{i=1}^{\infty} a^i T = (0)$. Therefore, $y = 0$ and $\mathfrak{q} = (0)$, so the only prime ideal that aR' contains is the zero ideal. Thus $\text{ht}(aR') \leq 1$.

We have thus shown that R' is an N -subring; to show it is an A -extension of R , we must simply demonstrate that the cardinality condition is satisfied and that prime elements of R remain prime in R' . The cardinality condition is satisfied since $|R'| = \sup(\aleph_0, |R|)$. Furthermore, since $x+t$ is transcendental over R , prime elements of R remain prime in R' , and $S = R'$ is our desired A -extension. \square

We have shown that, for an element $t \in T$, there exists an A -extension S of an N -subring R such that $t + \mathfrak{m}^2$ is in the image of the map from S to T/\mathfrak{m}^2 . In our construction, we apply this lemma infinitely often in order to make the map from our final subring A to T/\mathfrak{m}^2 a surjection.

For the stronger case of our theorem, we make the formal fibers geometrically regular at prime ideals \mathfrak{q} of T of height greater than or equal to 2. To do so, we make the map from A to T/\mathfrak{q} a surjection, and we use a similar lemma with the change that we consider \mathfrak{q} rather than \mathfrak{m}^2 .

Lemma 3.4. *Let (T, \mathfrak{m}) be a complete local normal domain, let R be an N -subring of T , and let $t \in T$. Suppose \mathfrak{q} is an ideal of T such that $\text{ht } \mathfrak{q} \geq 2$. Then there exists an infinite A -extension S of R such that $t + \mathfrak{q} \in \text{Image}(S \rightarrow T/\mathfrak{q})$. Moreover, if $t \in \mathfrak{q}$, we have $\mathfrak{q} \cap S \neq (0)$.*

Proof. The proof of this lemma is similar to the proof of the previous lemma. Let

$$C = \{\mathfrak{p} \in \text{Spec } T \mid \mathfrak{p} \in \text{Ass}(T/rT) \text{ with } 0 \neq r \in R\} \cup \text{Ass } T,$$

and for all $\mathfrak{p} \in C$, let $D_{(\mathfrak{p})}$ be a set of coset representatives of the cosets $u + \mathfrak{p} \in T/\mathfrak{p}$ that make $u + t + \mathfrak{p}$ algebraic over $R/(\mathfrak{p} \cap R)$. Let $D = \bigcup_{\mathfrak{p} \in C} D_{(\mathfrak{p})}$.

We note that, since T is a domain, the only associated prime ideal of T is the zero ideal. Thus $\mathfrak{q} \notin \text{Ass } T$. Since T is normal, it satisfies Serre's (S_2) condition, and furthermore, T is a normal domain, so (S_2) is equivalent to every prime divisor of a nonzero principal ideal having height 1. Therefore, all elements of $\text{Ass}(T/rT)$ for all $0 \neq r \in R$ have height 1. Since \mathfrak{q} has height strictly greater than 1, \mathfrak{q} cannot be contained in any element of C . We thus apply Lemma 2.6 if R is countable and Lemma 2.7 otherwise to choose $x \in \mathfrak{q}$ such that $x + t + \mathfrak{p}$ is transcendental over $R/(\mathfrak{p} \cap R)$ for all $\mathfrak{p} \in C$. Define $S = R[x + t]_{\mathfrak{m} \cap R[x+t]}$. Using the same proof as in the previous lemma, S is an A -extension of R .

We now must show that, if $t \in \mathfrak{q}$, we have $\mathfrak{q} \cap S \neq (0)$. Suppose $t \in \mathfrak{q}$. Then $x + t \in S$ for some $x \in \mathfrak{q}$ that makes $x + t + \mathfrak{p}$ transcendental over $R/(\mathfrak{p} \cap R)$ for all $\mathfrak{p} \in C$. Since $x + t + \mathfrak{p}$ is transcendental, $x + t \neq 0$. Therefore, $x + t$ is a nonzero element of \mathfrak{q} , so $0 \neq x + t \in \mathfrak{q} \cap S$ and $\mathfrak{q} \cap S \neq (0)$. \square

Given a complete local domain T , we want to adjoin the generators of infinitely many principal height-1 prime ideals of T to an N -subring R of T . To do so, we suppose that a prime ideal pT of T has zero intersection with an N -subring R and show that there exists an A -extension of R that contains a generator of pT .

Lemma 3.5. *Let (T, \mathfrak{m}) be a complete local domain of dimension at least 1 such that $|T/\mathfrak{m}| = |T|$. Let R be an N -subring of T , and let p be a nonzero prime element of T such that $pT \cap R = (0)$. Then there exists an A -extension S of R such that $pT \cap S = puS$ for some unit $u \in T$.*

Proof. Let

$$C = \{\mathfrak{p} \in \text{Spec } T \mid \mathfrak{p} \in \text{Ass}(T/rT) \text{ with } 0 \neq r \in R\} \cup \text{Ass } T.$$

For this proof, as in the proof of Lemma 3.3, each $\mathfrak{p} \in C$ corresponds to a subset $D_{(\mathfrak{p})}$ of T . Here $D_{(\mathfrak{p})}$ is a full set of distinct coset representatives u in T for those cosets $u + \mathfrak{p}$ such that $(pu) + \mathfrak{p} \in T/\mathfrak{p}$ is algebraic over $R/(\mathfrak{p} \cap R)$. Define $D = \bigcup_{\mathfrak{p} \in C} D_{(\mathfrak{p})}$. By Lemma 2.5, T is uncountable, and so T/\mathfrak{m} is also uncountable. It follows that $|C \times D| < |T/\mathfrak{m}|$. Furthermore, we know that $pT \notin C$: pT cannot be an associated prime ideal of T because it has height greater than zero, and pT cannot be an associated prime ideal of T/rT for some nonzero $r \in R$ because $pT \cap R = (0)$. Therefore, we can use Lemma 2.8 to find a unit $u \in T$ such that

$$pu \notin \bigcup \{t + \mathfrak{p} \mid t \in D, \mathfrak{p} \in C\}.$$

We claim that $S = R[pu]_{\mathfrak{m} \cap R[pu]}$ is an N -subring. Parts of this proof mimic the proof of Lemma 3.3.

Suppose $\mathfrak{p} \in C$. As in the proof of Lemma 3.3, elements of $\mathfrak{p} \cap S$ are of the form vf where $v \in S^\times$ is a unit and $f \in R[pu]$. Since v is a unit, this gives us that $f \in \mathfrak{p}$. Note that f is a polynomial in pu over R . Because $pu + \mathfrak{p}$ is transcendental over $R/(\mathfrak{p} \cap R)$, each of the coefficients of f must be in $\mathfrak{p} \cap R$. Then $f \in (\mathfrak{p} \cap R)R[pu]$ and $vf \in (\mathfrak{p} \cap R)S$. Therefore, $\mathfrak{p} \cap S \subseteq (\mathfrak{p} \cap R)S$. The opposite containment is clear; thus, we have equality.

We now show that S is an N -subring. Since R is a UFD and pu is transcendental over R , we have that $R[pu]$ is a UFD; furthermore, any localization of a UFD is a UFD, so S is also a UFD. The ring S satisfies the cardinality condition by the same argument as in Lemma 3.3: since we are adjoining a transcendental element and localizing, we have $|S| = \sup(\aleph_0, |R|)$. Furthermore, if $\mathfrak{q} \in \text{Ass } T$, then $\mathfrak{q} \in C$ so we have $\mathfrak{q} \cap S = (\mathfrak{q} \cap R)S = (0)$.

It remains to show that, for all regular $r \in R$ we have that $\mathfrak{p} \in \text{Ass}(T/rT)$ satisfies $\text{ht}(\mathfrak{p} \cap S) \leq 1$. This part of the proof follows exactly the proof of Lemma 3.3. Thus S is an N -subring. Furthermore, since we are adjoining a single transcendental element, prime elements of R remain prime in S . Thus S is in fact an A -extension of R .

Clearly, $puS \subseteq pT \cap S$. Suppose $f \in pT \cap R[pu]$. Then

$$f = r_n(pu)^n + \cdots + r_1(pu) + r_0$$

for some $r_i \in R$. Since $f \in pT$, we have $r_0 \in pT$ and so $r_0 \in pT \cap R = (0)$. It follows that $f \in (pu)R[pu]$ and we have $pT \cap R[pu] \subseteq (pu)R[pu]$. Therefore, $pT \cap S \subseteq (pu)S$ and we have $pT \cap S = puS$ as desired. \square

In order to ensure that our final subring A is Noetherian and has completion T , we must close up all finitely generated ideals. In other words, if \mathfrak{a} is a finitely generated ideal of A , we must show that $\mathfrak{a}T \cap A = \mathfrak{a}$. This will allow us to use Lemma 2.4. The next two lemmas will help us close up finitely generated ideals.

Lemma 3.6 [Heitmann 1993, Lemma 4]. *Let (T, \mathfrak{m}) be a complete local ring, and let R be an N -subring of T . Suppose $c \in R$, and let \mathfrak{a} be a finitely generated ideal of R with $c \in \mathfrak{a}T$. Then there exists an A -extension S of R such that $c \in \mathfrak{a}S$.*

Once we have our chain of intermediate subrings, we take their union to produce our final subring A . The following lemma helps us control this union.

Lemma 3.7 [Heitmann 1993, Lemma 6]. *Let (T, \mathfrak{m}) be a complete local ring and let R_0 be an N -subring of T . Let Ω be a well-ordered set with least element 0 and assume either Ω is countable or, for all $\alpha \in \Omega$, we have $|\{\beta \in \Omega \mid \beta < \alpha\}| < |T/\mathfrak{m}|$. Suppose $\{R_\alpha \mid \alpha \in \Omega\}$ is an ascending collection of rings such that, if α is a limit*

ordinal, then $R_\alpha = \bigcup_{\beta < \alpha} R_\beta$, while if $\alpha = \beta + 1$ is a successor ordinal, then R_α is an A -extension of R_β .

Then $S = \bigcup_{\beta \in \Omega} R_\beta$ satisfies all the conditions to be an N -subring of T except perhaps the cardinality condition. Instead, $|S| \leq \sup(\aleph_0, |R_0|, |\Omega|)$. Furthermore, elements which are prime in some R_β remain prime in S .

Now we prove that, given an N -subring R of T , there exists an A -extension S of R that satisfies all of the conditions we need our final subring A to satisfy.

Lemma 3.8. *Let (T, \mathfrak{m}) be a complete local domain with depth $T > 1$ such that $|T/\mathfrak{m}| = |T|$, and let R be an N -subring of T . Let p be a nonzero prime element of T such that $pT \cap R = (0)$, and let $t + \mathfrak{m}^2 \in T/\mathfrak{m}^2$. Then there exists an infinite A -extension S of R such that $t + \mathfrak{m}^2 \in \text{Image}(S \rightarrow T/\mathfrak{m}^2)$, $pT \cap S = puS$ for some unit $u \in T$, and, for every finitely generated ideal \mathfrak{a} of S , we have $\mathfrak{a}T \cap S = \mathfrak{a}$.*

Proof. First, use Lemma 3.5 to obtain an A -extension R' of R such that $pT \cap R' = puR'$ for some unit $u \in T$. Then use Lemma 3.3 to obtain an infinite A -extension R_0 of R' such that $t + \mathfrak{m}^2 \in \text{Image}(R_0 \rightarrow T/\mathfrak{m}^2)$. We must now prove the statement about finitely generated ideals. Let

$$\Omega = \{(\mathfrak{a}, c) \mid \mathfrak{a} \text{ is a finitely generated ideal of } R_0 \text{ and } c \in \mathfrak{a}T \cap R_0\}.$$

Then $|\Omega| = |R_0|$, so since $|T/\mathfrak{m}| = |T|$ and T is uncountable, we have $|\Omega| < |T/\mathfrak{m}|$. Let 0 designate the initial element of Ω , and well-order Ω in a such a way that it does not have a maximal element. Then Ω satisfies the hypotheses of Lemma 3.7. We define an increasing chain of intermediate subrings recursively with one subring for each element of Ω . We begin with R_0 . If $\alpha = \beta + 1$ is a successor ordinal and $\beta = (\mathfrak{a}, c)$ then we choose R_α to be an A -extension of R_β given by Lemma 3.6 such that $c \in \mathfrak{a}R_\alpha$. If α is a limit ordinal, define $R_\alpha = \bigcup_{\beta < \alpha} R_\beta$ and set $S_0 = \bigcup_{\alpha \in \Omega} R_\alpha$. Then by Lemma 3.7, S_0 is an A -extension of R_0 and if \mathfrak{a} is a finitely generated ideal of R_0 with $c \in \mathfrak{a}T \cap R_0$, then $(\mathfrak{a}, c) = \beta$ for some $\beta \in \Omega$. Then for some $\alpha > \beta$, we have $c \in \mathfrak{a}R_\alpha \subseteq \mathfrak{a}S_0$. Thus $\mathfrak{a}T \cap R_0 \subseteq \mathfrak{a}S_0$.

We repeat this process to obtain an A -extension S_1 of S_0 such that $\mathfrak{a}T \cap S_0 \subseteq \mathfrak{a}S_1$ for every finitely generated ideal \mathfrak{a} of S_0 . We continue recursively to obtain an ascending chain $S_0 \subseteq S_1 \subseteq \dots$ such that $\mathfrak{a}T \cap S_n \subseteq \mathfrak{a}S_{n+1}$ for every finitely generated ideal \mathfrak{a} of S_n . Let $S = \bigcup S_i$. By Lemma 3.7, S is an A -extension of S_0 (hence also of R). Since we have $t + \mathfrak{m}^2 \in \text{Image}(R_0 \rightarrow T/\mathfrak{m}^2)$ and $R_0 \subset S$, we must have $t + \mathfrak{m}^2 \in \text{Image}(S \rightarrow T/\mathfrak{m}^2)$. If \mathfrak{a} is a finitely generated ideal of S , then there must exist some S_n that contains a generating set $\{a_1, \dots, a_k\}$ for \mathfrak{a} . If $c \in \mathfrak{a}T \cap S$, then there exists some $m \geq n$ such that $c \in S_m$. Then

$$c \in (a_1, \dots, a_k)T \cap S_m \subseteq (a_1, \dots, a_k)S_{m+1} \subseteq \mathfrak{a}.$$

Thus $\mathfrak{a}T \cap S \subseteq \mathfrak{a}$. Since the reverse containment is clear, we have equality, and $\mathfrak{a}T \cap S = \mathfrak{a}$. Furthermore, puS is a finitely generated ideal of S , so by what we have just shown, we must have $(puS)T \cap S = puS$. But we know that $(puS)T = puT = pT$ since u is a unit in T . Thus $pT \cap S = puS$, and S satisfies all the properties in the statement of the theorem. \square

We now present the analogous lemma for the stronger case of our theorem.

Lemma 3.9. *Let (T, \mathfrak{m}) be a complete normal local domain with $\text{depth } T > 1$ such that $|T/\mathfrak{m}| = |T|$, and let R be an N -subring of T . Let p be a nonzero prime element of T such that $pT \cap R = (0)$, and let $t + \mathfrak{q} \in T/\mathfrak{q}$ for some ideal \mathfrak{q} of T of height at least 2. Then there exists an infinite A -extension S of R such that*

- (1) $t + \mathfrak{q} \in \text{Image}(S \rightarrow T/\mathfrak{q})$ with $\mathfrak{q} \cap S \neq (0)$ if $t \in \mathfrak{q}$,
- (2) $pT \cap S = puS$ for some unit $u \in T$, and,
- (3) for every finitely generated ideal \mathfrak{a} of S , we have $\mathfrak{a}T \cap S = \mathfrak{a}$.

Proof. Follow the proof of Lemma 3.8 replacing the use of Lemma 3.3 with Lemma 3.4. \square

4. The main theorems

We are now ready to prove our main theorems.

Theorem 4.1. *Let (T, \mathfrak{m}) be a complete local domain such that $\text{depth } T > 1$. Let \mathcal{S} be the set of principal height-1 prime ideals of T . Suppose that $|T| = |T/\mathfrak{m}| = |\mathcal{S}|$ and $\mathbb{Q} \subseteq T$. Then there exists a local unique factorization domain A such that $\hat{A} \cong T$ and A satisfies the following conditions:*

- (1) A is universally catenary.
- (2) There is an uncountable subset \mathcal{S}' of \mathcal{S} such that
 - (a) if $\mathfrak{q} \in \mathcal{S}'$, then $\mathfrak{q} \cap A \neq (0)$,
 - (b) if $\mathfrak{q}, \mathfrak{q}' \in \mathcal{S}'$ with $\mathfrak{q} \neq \mathfrak{q}'$, then $\mathfrak{q} \cap A \neq \mathfrak{q}' \cap A$, and
 - (c) if $\mathfrak{q} \in \mathcal{S}'$, then $(T/(\mathfrak{q} \cap A)T)_{\mathfrak{q}}$ is a field.

Proof. Let $\Omega = T/\mathfrak{m}^2$ be well-ordered so that 0 is its initial element and each element of Ω has fewer than $|\Omega|$ predecessors. Then Ω satisfies the hypotheses of Lemma 3.7 since $|T| \geq |T/\mathfrak{m}^2| \geq |T/\mathfrak{m}|$ implies $|T/\mathfrak{m}^2| = |T/\mathfrak{m}|$. We now recursively define a family of rings $\{R_\alpha \mid \alpha \in \Omega\}$ satisfying the hypotheses of Lemma 3.7.

Let R_0 be \mathbb{Q} , and note that R_0 is an N -subring of T .

Whenever $\alpha = \beta + 1$ is a successor ordinal, use Lemma 3.8 to construct an A -extension R_α of R_β such that $\beta = t + \mathfrak{m}^2 \in \text{Image}(R_\alpha \rightarrow T/\mathfrak{m}^2)$, $pT \cap R_\alpha = puR_\alpha$ for some unit $u \in T$ and some nonzero prime element p of T , and, for every

finitely generated ideal \mathfrak{a} of R_α , we have $\mathfrak{a}T \cap R_\alpha = \mathfrak{a}$. In order to do so, we must first justify why, at each step, there exists some nonzero prime element p of T such that $pT \cap R_\beta = (0)$. Since $|T| = |T/\mathfrak{m}|$, we have $|T/\mathfrak{m}| = |T/\mathfrak{m}^2|$. Thus $|R_\beta| < |T/\mathfrak{m}^2| = |T/\mathfrak{m}| = |\mathcal{S}|$. Since each element of R_β can be contained in only finitely many height-1 prime ideals of T , there are at most $|R_\beta|$ height-1 prime ideals of T having nonzero intersection with R_β . Furthermore, we know that $|R_\beta| \geq \aleph_0$ because $\mathbb{Q} \subset R_\beta$. Therefore, there are at most $\max(\aleph_0, |R_\beta|) = |R_\beta|$ elements of \mathcal{S} having nonzero intersection with R_β . Since $|\mathcal{S}| > |R_\beta|$, there must exist at least one nonzero element of \mathcal{S} whose intersection with R_β is (0) . Thus, at every step, there is a nonzero prime element p of T with the property that $pT \cap R_\beta = (0)$, and so the hypotheses of Lemma 3.8 are satisfied. Therefore, we can in fact use Lemma 3.8 to construct an A -extension R_α of R_β with our desired properties.

If α is a limit ordinal, let $R_\alpha = \bigcup_{\beta < \alpha} R_\beta$. We claim $A = \bigcup_{\alpha \in \Omega} R_\alpha$ is our desired subring.

First, we claim that A is Noetherian and $\hat{A} \cong T$. By construction, the map $A \rightarrow T/\mathfrak{m}^2$ is surjective. Let $\mathfrak{a} = (a_1, \dots, a_n)$ be a finitely generated ideal of A , and let $x \in \mathfrak{a}T \cap A$. Then there exists an $\alpha \in \Omega$ such that α is a successor ordinal and $x, a_1, \dots, a_n \in R_\alpha$. Then

$$x \in (a_1, \dots, a_n)T \cap R_\alpha = (a_1, \dots, a_n)R_\alpha \subseteq \mathfrak{a}.$$

It follows that $\mathfrak{a}T \cap A = \mathfrak{a}$ for all finitely generated ideals \mathfrak{a} of A . By Lemma 2.4, we have that A is Noetherian and $\hat{A} \cong T$.

By Lemma 3.7, A is a UFD, and since T is a domain, it is equidimensional, and so A is universally catenary. Define

$$\mathcal{S}' = \{pT \in \mathcal{S} \mid pT \cap A = puA \text{ for some unit } u \in T\}.$$

Then, by construction, \mathcal{S}' is uncountable, if $\mathfrak{q} \in \mathcal{S}'$, then $\mathfrak{q} \cap A \neq (0)$, and if $\mathfrak{q}, \mathfrak{q}' \in \mathcal{S}'$ with $\mathfrak{q} \neq \mathfrak{q}'$, then $\mathfrak{q} \cap A \neq \mathfrak{q}' \cap A$. Now we show that, if $\mathfrak{q} \in \mathcal{S}'$, then $(T/(\mathfrak{q} \cap A)T)_{\mathfrak{q}}$ is a field. Suppose $\mathfrak{q} = pT$ and $\mathfrak{q} \cap A = puA$ for some unit $u \in T$. Then

$$(T/(\mathfrak{q} \cap A)T)_{\mathfrak{q}} \cong (T/puT)_{\mathfrak{q}} = (T/pT)_{\mathfrak{q}} = (T/\mathfrak{q})_{\mathfrak{q}} = T_{\mathfrak{q}}/\mathfrak{q}T_{\mathfrak{q}}$$

is a field. Thus the A we construct does indeed satisfy the desired properties stated in the theorem. □

Corollary 4.2. *Let (T, \mathfrak{m}) be a complete local domain such that $\dim T = 2$ and $\text{depth } T = 2$. Let \mathcal{S} be the set of principal height-1 prime ideals of T . Suppose that $|T| = |T/\mathfrak{m}| = |\mathcal{S}|$ and $\mathbb{Q} \subseteq T$. Then there exists a local universally catenary unique factorization domain A such that $\hat{A} \cong T$ and A satisfies conditions (1)–(3) of Question 1.2.*

Proof. By Theorem 4.1, there exists a local UFD A that satisfies conditions (1) and (2) of Question 1.2 and a modified version of condition (3). Thus we need only show that condition (3) holds. Suppose $\mathfrak{q} \in \mathcal{S}'$ with $\mathfrak{q} = pT$. Then, $\mathfrak{q} \cap A = puA$ for some unit u of T . Since T has dimension 2, if Q is a prime ideal of T with $Q \cap A = puA$, then Q has height 1 and $Q = pT = \mathfrak{q}$. Therefore, in order to show that the formal fiber of A at $\mathfrak{q} \cap A$ is geometrically regular, it suffices to show that the ring $(T/(\mathfrak{q} \cap A)T)_{\mathfrak{q}}$ is a regular local ring. By Theorem 4.1, we know that $(T/(\mathfrak{q} \cap A)T)_{\mathfrak{q}}$ is a field, and hence a regular local ring. Thus A satisfies conditions (1)–(3) as desired. \square

We now impose the condition that T is a normal domain so we can make the generic formal fiber of A and all formal fibers at prime ideals of height greater than 1 be geometrically regular.

Theorem 4.3. *Let (T, \mathfrak{m}) be a normal complete local domain and depth $T > 1$. Let \mathcal{S} be the set of principal height-1 prime ideals of T . Suppose that $|\mathcal{S}| = |T| = |T/\mathfrak{m}|$ and $\mathbb{Q} \subseteq T$. Then there exists a local unique factorization domain A such that $\hat{A} \cong T$ and A satisfies the following conditions:*

- (1) A is universally catenary.
- (2) If $\mathfrak{p} \in \text{Spec } A$ with the height of \mathfrak{p} not equal to 1, the formal fiber of A at \mathfrak{p} is geometrically regular.
- (3) There is an uncountable subset \mathcal{S}' of \mathcal{S} such that
 - (a) if $\mathfrak{q} \in \mathcal{S}'$, then $\mathfrak{q} \cap A \neq (0)$,
 - (b) if $\mathfrak{q}, \mathfrak{q}' \in \mathcal{S}'$ with $\mathfrak{q} \neq \mathfrak{q}'$, then $\mathfrak{q} \cap A \neq \mathfrak{q}' \cap A$, and
 - (c) if $\mathfrak{q} \in \mathcal{S}'$, then the formal fiber of A at $\mathfrak{q} \cap A$ is geometrically regular.

Proof. Let

$$\Omega = \{t + \mathfrak{q} \in T/\mathfrak{q} \mid \mathfrak{q} \text{ is an ideal of } T \text{ with } \text{ht } \mathfrak{q} \geq 2\}$$

be well-ordered so that 0 is its initial element and each element of Ω has fewer than $|\Omega|$ predecessors. Then Ω satisfies the hypotheses of Lemma 3.7. Also, note that the height of \mathfrak{m}^2 is at least 2. We now recursively define a family of rings $\{R_\alpha \mid \alpha \in \Omega\}$ satisfying the hypotheses of Lemma 3.7.

Let R_0 be \mathbb{Q} , and note that R_0 is an N -subring of T .

Whenever $\alpha = \beta + 1$ is a successor ordinal, we use Lemma 3.9 to construct an A -extension R_α of R_β such that $\beta = t + \mathfrak{q} \in \text{Image}(R_\alpha \rightarrow T/\mathfrak{q})$ with $\mathfrak{q} \cap R_\alpha \neq (0)$ if $t \in \mathfrak{q}$, $pT \cap R_\alpha = puR_\alpha$ for some unit $u \in T$ and some nonzero prime element p in T with $pT \cap R_\beta = (0)$, and, for every finitely generated ideal \mathfrak{a} of R_α , we have $\mathfrak{a}T \cap R_\alpha = \mathfrak{a}$. In order to do so, we must first justify why at each step, there exists a nonzero prime element p of T such that $pT \cap R_\beta = (0)$. The justification for this is the same argument that is used in the proof of Theorem 4.1.

If α is a limit ordinal, let $R_\alpha = \bigcup_{\beta < \alpha} R_\beta$. We claim $A = \bigcup_{\alpha \in \Omega} R_\alpha$ is our desired subring.

The proof that A is Noetherian and $\hat{A} \cong T$ is the same argument that is used in the proof of Theorem 4.1.

By Lemma 3.7, A is a UFD. Since T is a domain, it is equidimensional; hence, A is formally equidimensional and therefore universally catenary. In addition, because the map $A \rightarrow T/\mathfrak{q}$ is surjective for all ideals \mathfrak{q} of T of height at least 2, we claim that the formal fibers at all prime ideals of A of height at least 2 are geometrically regular. Let \mathfrak{p} be a prime ideal of A with height at least 2. Then, by the going down theorem, $\mathfrak{p}T$ is an ideal of T of height at least 2. Because the map $A \rightarrow T/\mathfrak{q}$ is surjective for all ideals \mathfrak{q} of T of height at least 2, we have that the map $A \rightarrow T/\mathfrak{p}T$ is a surjection; hence by the first isomorphism theorem, we obtain $A/\mathfrak{p} \cong T/\mathfrak{p}T$. Since \mathfrak{p} is a prime ideal of A , we know A/\mathfrak{p} is a domain, so $T/\mathfrak{p}T$ is a domain. Therefore, $\mathfrak{p}T$ is a prime ideal of T . In fact, as A/\mathfrak{p} is complete, $\mathfrak{p}T$ is the only prime ideal of T whose intersection with A is \mathfrak{p} , so, in order to prove that the formal fiber of A at \mathfrak{p} is geometrically regular, we need only prove that the ring $(T/\mathfrak{p}T)_{\mathfrak{p}T}$ is a regular local ring. Now, $(T/\mathfrak{p}T)_{\mathfrak{p}T} \cong T_{\mathfrak{p}}/\mathfrak{p}T_{\mathfrak{p}}$ is a field, and hence a regular local ring. This implies that the formal fiber of A at \mathfrak{p} is geometrically regular.

We now show that the generic formal fiber of A is geometrically regular. Since $\mathbb{Q} \subseteq T$, it is enough to show that $(T/(0)T)_{\mathfrak{q}}$ is a regular local ring for all $\mathfrak{q} \in \text{Spec } T$ such that $\mathfrak{q} \cap A = (0)$. This is equivalent to showing that $T_{\mathfrak{q}}$ is a regular local ring for all $\mathfrak{q} \in \text{Spec } T$ such that $\mathfrak{q} \cap A = (0)$. Recall that a ring T satisfies Serre's (R_1) condition if T_P is regular for all prime ideals P of T with height at most 1. By Theorem 23.8 in [Matsumura 1986], since T is normal, T satisfies Serre's (R_1) condition. Thus, for all $\mathfrak{q} \in \text{Spec } T$ with $\text{ht } \mathfrak{q} \leq 1$, $T_{\mathfrak{q}}$ is a regular local ring. We claim that this is enough; in other words, we claim that all elements of the generic formal fiber of A have height at most 1. By our use of Lemma 3.9 in the construction of A , for all prime ideals \mathfrak{q} of T with height at least 2, $\mathfrak{q} \cap A \neq (0)$; thus \mathfrak{q} cannot be in the generic formal fiber of A . Therefore, all elements of the generic formal fiber of A have height at most 1 and the generic formal fiber of A is geometrically regular.

Define

$$\mathcal{S}' = \{pT \in \mathcal{S} \mid pT \cap A = puA \text{ for some unit } u \in T\}.$$

Then, by construction, \mathcal{S}' is uncountable, if $\mathfrak{q} \in \mathcal{S}'$, then $\mathfrak{q} \cap A \neq (0)$ and if $\mathfrak{q}, \mathfrak{q}' \in \mathcal{S}'$ with $\mathfrak{q} \neq \mathfrak{q}'$, then $\mathfrak{q} \cap A \neq \mathfrak{q}' \cap A$. Let $\mathfrak{q} = pT \in \mathcal{S}'$. Then $\mathfrak{q} \cap A = puA$ for some unit u in T . Suppose Q is a prime ideal of T such that $Q \cap A = puA$. We claim $Q = pT = \mathfrak{q}$. If the height of Q is at least 2, then, by construction, $A \rightarrow T/Q$ is surjective, and so $A/puA \cong T/Q$. Since A is catenary, the dimension of A/puA is $\dim A - 1$. But the dimension of T/Q is less than $\dim T - 1$, a contradiction. It follows that Q is a height-1 prime ideal of T . Hence, as $pT \subseteq Q$, we have

$Q = pT = \mathfrak{q}$. Therefore, to show that the formal fiber of A at $\mathfrak{q} \cap A$ is geometrically regular, it suffices to show that $(T/(\mathfrak{q} \cap A)T)_{\mathfrak{q}}$ is a field. This argument is given in the proof of Theorem 4.1. \square

We present an example of a class of rings that satisfies the hypotheses of Theorem 4.3.

Example 4.4. Let $T = \mathbb{C}[[x_1, x_2, \dots, x_n]]/(x_1x_2 - x_3^2)$, where $n \geq 3$. Then T is the completion of a local UFD A satisfying the properties in Theorem 4.3.

Proof. We must show that T satisfies the hypotheses of Theorem 4.3. Note that T is a complete local ring. Since $(x_1x_2 - x_3^2)$ is a prime ideal of $\mathbb{C}[[x_1, x_2, \dots, x_n]]$, T is a domain. One can check that T is normal and $\text{depth } T > 1$. Now, $|T| = |T/\mathfrak{m}| = |\mathbb{C}|$, and we must show that $|T| = |\mathcal{S}|$. Let $\mathfrak{q}_\alpha = (x_1 + \alpha x_2)$. We claim that for all nonzero $\alpha \in \mathbb{C}$, the ideal \mathfrak{q}_α is prime. To prove this, consider the ring

$$T/\mathfrak{q}_\alpha = \mathbb{C}[[x_1, x_2, \dots, x_n]]/(x_1x_2 - x_3^2, \mathfrak{q}_\alpha) = \mathbb{C}[[x_1, x_2, \dots, x_n]]/(-\alpha x_2^3 - x_3^2).$$

Since $-\alpha x_2^3 - x_3^2$ is an irreducible polynomial in $\mathbb{C}[[x_1, x_2, \dots, x_n]]$, the ideal $(-\alpha x_2^3 - x_3^2)$ is prime in $\mathbb{C}[[x_1, x_2, \dots, x_n]]$; hence T/\mathfrak{q}_α is a domain. Therefore, \mathfrak{q}_α is a prime ideal of T , and one can verify that $\mathfrak{q}_\alpha = \mathfrak{q}_\beta$ if and only if $\alpha = \beta$. We then consider the set of prime ideals $\mathcal{Q} = \{\mathfrak{q}_\alpha \mid \alpha \in \mathbb{C} \text{ and } \alpha \neq 0\} \subseteq \mathcal{S}$. We thus have the following string of inequalities:

$$|\mathbb{C}| = |\mathcal{Q}| \leq |\mathcal{S}| \leq |T| = |\mathbb{C}|.$$

Therefore, we obtain that $|\mathcal{S}| = |T|$, and so $|\mathcal{S}| = |T| = |T/\mathfrak{m}|$. We can thus apply Theorem 4.3 to T . \square

5. Acknowledgements

We would like to thank Cory Colbert for his suggestions and corrections. Fleming also thanks the Clare Boothe Luce Program of the Henry Luce Foundation for their generous support of her work. Finally, we thank the referee for taking the time to carefully read this paper and to make detailed comments and suggestions for improvements. We especially appreciate the referee's suggestions for improving the exposition of the paper.

References

- [Boocher et al. 2010] A. Boocher, M. Daub, R. K. Johnson, H. Lindo, S. Loepp, and P. A. Woodard, "Formal fibers of unique factorization domains", *Canad. J. Math.* **62**:4 (2010), 721–736. MR Zbl
- [Bryk et al. 2005] J. Bryk, S. Mapes, C. Samuels, and G. Wang, "Constructing almost excellent unique factorization domains", *Comm. Algebra* **33**:5 (2005), 1321–1336. MR Zbl
- [Charters and Loepp 2004] P. Charters and S. Loepp, "Semilocal generic formal fibers", *J. Algebra* **278**:1 (2004), 370–382. MR Zbl

- [Fleming et al. 2019] S. M. Fleming, L. Ji, S. Loepp, P. M. McDonald, N. Pande, and D. Schwein, “Completely controlling the dimensions of formal fiber rings at prime ideals of small height”, *J. Commut. Algebra* **11**:3 (2019), 363–388. MR Zbl
- [Heitmann 1993] R. C. Heitmann, “Characterization of completions of unique factorization domains”, *Trans. Amer. Math. Soc.* **337**:1 (1993), 379–387. MR Zbl
- [Heitmann 1994] R. C. Heitmann, “Completions of local rings with an isolated singularity”, *J. Algebra* **163**:2 (1994), 538–567. MR Zbl
- [Loepp 1997] S. Loepp, “Constructing local generic formal fibers”, *J. Algebra* **187**:1 (1997), 16–38. MR Zbl
- [Loepp 2003] S. Loepp, “Characterization of completions of excellent domains of characteristic zero”, *J. Algebra* **265**:1 (2003), 221–228. MR Zbl
- [Matsumura 1986] H. Matsumura, *Commutative ring theory*, Cambridge Studies in Advanced Mathematics **8**, Cambridge University Press, 1986. MR Zbl

Received: 2019-09-02 Revised: 2019-12-10 Accepted: 2019-12-10

fleming.sarah.m@gmail.com *Williams College, Williamstown, MA, United States*

sloepp@williams.edu *Department of Mathematics and Statistics, Williams College,
Williamstown, MA, 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

2020 vol. 13 no. 1

Structured sequences and matrix ranks	1
CHARLES JOHNSON, YAOXIAN QU, DUO WANG AND JOHN WILKES	
Analysis of steady states for classes of reaction-diffusion equations with hump-shaped density-dependent dispersal on the boundary	9
QUINN MORRIS, JESSICA NASH AND CATHERINE PAYNE	
The L -move and Markov theorems for trivalent braids	21
CARMEN CAPRAU, GABRIEL COLOMA AND MARGUERITE DAVIS	
Low stages of the Taylor tower for r -immersions	51
BRIDGET SCHREINER, FRANJO ŠARČEVIĆ AND ISMAR VOLIĆ	
A new go-to sampler for Bayesian probit regression	77
SCOTT SIMMONS, ELIZABETH J. MCGUFFEY AND DOUGLAS VANDERWERKEN	
Characterizing optimal point sets determining one distinct triangle	91
HAZEL N. BRENNER, JAMES S. DEPRET-GUILLAUME, EYVINDUR A. PALSSON AND ROBERT W. STUCKEY	
Solutions of periodic boundary value problems	99
R. AADITH, PARAS GUPTA AND JAGAN MOHAN JONNALAGADDA	
A few more trees the chromatic symmetric function can distinguish	109
JAKE HURYN AND SERGEI CHMUTOV	
One-point hyperbolic-type metrics	117
MARINA BOROVIKOVA, ZAIR IBRAGIMOV, MIGUEL JIMENEZ BRAVO AND ALEXANDRO LUNA	
Some generalizations of the ASR search algorithm for quasitwisted codes	137
NUH AYDIN, THOMAS H. GUIDOTTI, PEIHAN LIU, ARMIYA S. SHAIKH AND ROBERT O. VANDENBERG	
Continuous factorization of the identity matrix	149
YUYING DAI, ANKUSH HORE, SIQI JIAO, TIANXU LAN AND PAVLOS MOTAKIS	
Almost excellent unique factorization domains	165
SARAH M. FLEMING AND SUSAN LOEPP	

