involve

a journal of mathematics

Editorial Board

Kenneth S. Berenhaut, *Managing Editor*

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

msp

2014 vol. 7, no. 5

involve

msp.org/involve

EDITORS

MANAGING EDITOR

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

BOARD OF EDITORS

Colin Adams	Williams College, USA colin.c.adams@williams.edu
John V. Baxley	Wake Forest University, NC, USA baxley@wfu.edu
Arthur T. Benjamin	Harvey Mudd College, USA benjamin@hmc.edu
Martin Bohner	Missouri U of Science and Technology, USA bohner@mst.edu
Nigel Boston	University of Wisconsin, USA boston@math.wisc.edu
Amarjit S. Budhiraja	U of North Carolina, Chapel Hill, USA budhiraj@email.unc.edu
Pietro Cerone	La Trobe University, Australia P.Cerone@latrobe.edu.au
Scott Chapman	Sam Houston State University, USA scott.chapman@shsu.edu
Joshua N. Cooper	University of South Carolina, USA cooper@math.sc.edu
Jem N. Corcoran	University of Colorado, USA corcoran@colorado.edu
Toka Diagana	Howard University, USA tdiagana@howard.edu
Michael Dorff	Brigham Young University, USA mdorff@math.byu.edu
Sever S. Dragomir	Victoria University, Australia sever@matilda.vu.edu.au
Behrouz Emamizadeh	The Petroleum Institute, UAE bemamizadeh@pi.ac.ae
Joel Foisy	SUNY Potsdam foisyjs@potsdam.edu
Errin W. Fulp	Wake Forest University, USA fulp@wfu.edu
Joseph Gallian	University of Minnesota Duluth, USA jgallian@d.umn.edu
Stephan R. Garcia	Pomona College, USA stephan.garcia@pomona.edu
Anant Godbole	East Tennessee State University, USA godbole@etsu.edu
Ron Gould	Emory University, USA rg@mathcs.emory.edu
Andrew Granville	Université Montréal, Canada andrew@dms.umontreal.ca
Jerrold Griggs	University of South Carolina, USA griggs@math.sc.edu
Sat Gupta	U of North Carolina, Greensboro, USA sngupta@uncg.edu
Jim Haglund	University of Pennsylvania, USA jhaglund@math.upenn.edu
Johnny Henderson	Baylor University, USA johnny_henderson@baylor.edu
Jim Hoste	Pitzer College jhoste@pitzer.edu
Natalia Hritonenko	Prairie View A&M University, USA nahritonenko@pvamu.edu
Glenn H. Hurlbert	Arizona State University,USA hurlbert@asu.edu
Charles R. Johnson	College of William and Mary, USA crjohnso@math.wm.edu
K. B. Kulasekera	Clemson University, USA kk@ces.clemson.edu
Gerry Ladas	University of Rhode Island, USA gladas@math.uri.edu
David Larson	Texas A&M University, USA larson@math.tamu.edu
Suzanne Lenhart	University of Tennessee, USA lenhart@math.utk.edu
Chi-Kwong Li	College of William and Mary, USA ckli@math.wm.edu
Robert B. Lund	Clemson University, USA lund@clemson.edu
Gaven J. Martin	Massey University, New Zealand g.j.martin@massey.ac.nz
Mary Meyer	Colorado State University, USA meyer@stat.colostate.edu
Emil Minchev	Ruse, Bulgaria eminchev@hotmail.com
Frank Morgan	Williams College, USA frank.morgan@williams.edu
Mohammad Sal Moslehian	Ferdowsi University of Mashhad, Iran moslehian@ferdowsi.um.ac.ir
Zuhair Nashed	University of Central Florida, USA znashed@mail.ucf.edu
Ken Ono	Emory University, USA ono@mathcs.emory.edu
Timothy E. O'Brien	Loyola University Chicago, USA tobrie1@luc.edu
Joseph O'Rourke	Smith College, USA orourke@cs.smith.edu
Yuval Peres	Microsoft Research, USA peres@microsoft.com
Y.-F. S. Pétermann	Université de Genève, Switzerland petermann@math.unige.ch
Robert J. Plemmons	Wake Forest University, USA plemmons@wfu.edu
Carl B. Pomerance	Dartmouth College, USA carl.pomerance@dartmouth.edu
Vadim Ponomarenko	San Diego State University, USA vadim@sciences.sdsu.edu
Bjorn Poonen	UC Berkeley, USA poonen@math.berkeley.edu
James Propp	U Mass Lowell, USA jpropp@cs.uml.edu
Józeph H. Przytycki	George Washington University, USA przytyck@gwu.edu
Richard Rebarber	University of Nebraska, USA rrebarbe@math.unl.edu
Robert W. Robinson	University of Georgia, USA rwr@cs.uga.edu
Filip Saidak	U of North Carolina, Greensboro, USA f_saidak@uncg.edu
James A. Sellers	Penn State University, USA sellersj@math.psu.edu
Andrew J. Sterge	Honorary Editor andy@ajsterge.com
Ann Trenk	Wellesley College, USA atrenk@wellesley.edu
Ravi Vakil	Stanford University, USA vakil@math.stanford.edu
Antonia Vecchio	Consiglio Nazionale delle Ricerche, Italy antonia.vecchio@cnr.it
Ram U. Verma	University of Toledo, USA verma99@msn.com
John C. Wierman	Johns Hopkins University, USA wierman@jhu.edu
Michael E. Zieve	University of Michigan, USA zieve@umich.edu

PRODUCTION

Silvio Levy, Scientific Editor

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2014 is US $120/year for the electronic version, and $165/year (+$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to MSP.

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

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

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

© 2014 Mathematical Sciences Publishers

INVOLVE 7:5 (2014)
dx.doi.org/10.2140/involve.2014.7.585

Infinite cardinalities in the Hausdorff metric geometry

Alexander Zupan

(Communicated by Józef H. Przytycki)

The Hausdorff metric measures the distance between nonempty compact sets in $\mathbb{R}^n$, the collection of which is denoted $\mathcal{H}(\mathbb{R}^n)$. Betweenness in $\mathcal{H}(\mathbb{R}^n)$ can be defined in the same manner as betweenness in Euclidean geometry. But unlike betweenness in $\mathbb{R}^n$, for some elements A and B of $\mathcal{H}(\mathbb{R}^n)$ there can be many elements between A and B at a fixed distance from A. Blackburn et al. ("A missing prime configuration in the Hausdorff metric geometry", *J. Geom.*, **92**:1–2 (2009), pp. 28–59) demonstrate that there are infinitely many positive integers k such that there exist elements A and B having exactly k different elements between A and B at each distance from A while proving the surprising result that no such A and B exist for $k = 19$. In this vein, we prove that there do not exist elements A and B with exactly a countably infinite number of elements at any location between A and B.

1. Introduction

The Hausdorff metric provides a means to measure distance in the family $\mathcal{H}(\mathbb{R}^n)$ of nonempty compact sets in n-dimensional Euclidean space. There is a natural embedding of $\mathbb{R}^n$ into $\mathcal{H}(\mathbb{R}^n)$ that takes $x \in \mathbb{R}^n$ to $\{x\} \in \mathcal{H}(\mathbb{R}^n)$. The notion of betweenness in $\mathbb{R}^n$ extends naturally to $\mathcal{H}(\mathbb{R}^n)$. However, in Euclidean space, there is a unique point between a and b at a given distance less than $d(a, b)$ from a, while in $\mathcal{H}(\mathbb{R}^n)$ there can be many distinct elements between elements A and B at a given distance from A. For instance, for infinitely many numbers k we can find A and B with exactly k elements between A and B at a given distance from A, and we can also find A and B such that this number of elements between A and B is infinite. Blackburn et al. [2009] proved the surprising result that there exist no two elements A and B in $\mathcal{H}(\mathbb{R}^n)$ with the property that A and B have exactly 19 elements of $\mathcal{H}(\mathbb{R}^n)$ between them at a given distance from A. In this paper, we will prove that there is another cardinality that is missing; namely, there exist no

MSC2010: primary 51F99; secondary 54B20.
Keywords: Hausdorff metric, betweenness, metric geometry.

two elements $A, B \in \mathcal{H}(\mathbb{R}^n)$ with exactly a countably infinite number of elements between them at any location. The argument uses a different approach than that of [Blackburn et al. 2009]: the proof there is exhaustive, and while it is succinct enough to prove the absence of 19, the method may be too unwieldy to show the existence of larger conjectured missing numbers. It is our hope that the idea of too many removable points forcing a larger cardinality of sets between A and B might be adapted to the finite case.

2. Preliminaries

Let $\mathcal{H}(\mathbb{R}^n)$ denote the collection of nonempty compact subsets of $\mathbb{R}^n$. We will refer to these compact sets as "elements" of $\mathcal{H}(\mathbb{R}^n)$. For any $a \in \mathbb{R}^n$ and $B \in \mathcal{H}(\mathbb{R}^n)$, let $d(a, B) = \min_{b \in B} d_E(a, b)$, where d_E denotes the Euclidean metric on $\mathbb{R}^n$. The Hausdorff metric is then defined as follows:

Definition 2.1. Let $A, B \in \mathcal{H}(\mathbb{R}^n)$. The Hausdorff distance between A and B is given by

$$h(A, B) = \max\{d(A, B), d(B, A)\},$$

where $d(A, B) = \max_{a \in A} d(a, B)$.

In other words, the distance from A to B is the maximum of the distances between points in A to the set B, and the Hausdorff distance between A and B is the maximum of the distance from A to B and the distance from B to A. Note the maximum and minimum in the definitions above are well-defined since both A and B are compact sets. To verify that the Hausdorff distance defines a metric on $\mathcal{H}(\mathbb{R}^n)$, see, for instance, [Edgar 1990].

Example 2.2. Let $n = 2$ and consider the sets shown on the left in Figure 1. Let $S^1(r)$ denote the circle of radius r centered at the origin, so that $A = \{(0, 0)\} \cup S^1(4)$, $B = S^1(2)$, and $C = S^1(1) \cup S^1(3)$. Then, for any $a \in A$ and $b \in B$, we have $d_E(a, b) \geq 2$. Further, for such a there exists a point $b \in B$ such that $d(a, b) = 2$, which implies that $d(a, B) = 2$ for all $a \in A$; hence, $d(A, B) = 2$. Similarly, for every $b \in B$, we have $d_E(b, a_0) = 2$ where a_0 is the origin, so $d(b, A) = 2$ for all $b \in B$, which shows $d(B, A) = 2$ as well. It follows that $h(A, B) = 2$. A similar verification shows that $h(A, C) = h(B, C) = 1$.

The set C' pictured on the right in Figure 1 is a compact subset of C. Here we see that for every $a \in A$ and $c \in C'$, $d_E(a, c) \geq 1$. Additionally, for every $a \in A$, there exists $c \in C'$ such that $d_E(a, c) = 1$, so $d(a, C') = 1$ for all such A and $d(A, C') = 1$. Likewise, for all $c \in C'$, there is some $a \in A$ such that $d_E(c, a) = 1$, so $d(C', A) = 1$ and $h(A, C') = 1$. A similar computation shows that $d(B, C') = 1$.

In $\mathbb{R}^n$ we say that c is between a and b at a distance $t \in \mathbb{R}$ from a (where $0 < t < d_E(a, b)$) if $d_E(a, b) = d_E(a, c) + d_E(c, b)$ and $d_E(a, c) = t$. If $\{a\}, \{b\} \in \mathcal{H}(\mathbb{R}^n)$

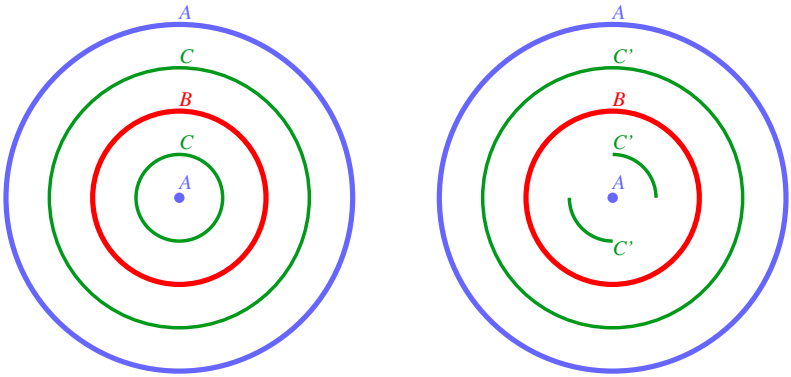


Figure 1. Two elements C and C' between sets A and B.

are single point sets, it is easy to see that $d_E(a, b) = h(\{a\}, \{b\})$. Thus, we can naturally extend betweenness in $\mathbb{R}^n$ to $\mathcal{H}(\mathbb{R}^n)$.

Definition 2.3. Let $A, B, C \in \mathcal{H}(\mathbb{R}^n)$ and $0 < t < h(A, B)$. We say that C is between A and B at a distance t from A if

$$h(A, B) = h(A, C) + h(C, B) \quad \text{and} \quad h(A, C) = t.$$

Thus, betweenness in $\mathbb{R}^n$ is preserved under the natural embedding of $\mathbb{R}^n$ into $\mathcal{H}(\mathbb{R}^n)$. To see that there can be multiple elements at some location between compact sets A and B, recall the previous example:

Example 2.4. As computed above, we have

$$h(A, B) = 2 = 1 + 1 = h(A, C) + h(C, B).$$

However, we also have

$$h(A, B) = 2 = 1 + 1 = h(A, C') + h(C', B)$$

for the same sets A and B, so C' and C lie between A and B one unit from A. In fact, if C'' is the union of $S^1(3)$ and any nonempty compact subset of $S^1(1)$, C'' is also between A and B one unit from A, so for this example there are uncountably many elements of $\mathcal{H}(\mathbb{R}^2)$ between A and B at a distance one from A.

Lemma 2.5 of [Blackburn et al. 2009] says that if $A, B \in \mathcal{H}(\mathbb{R}^n)$ and there exists some $a \in A$ or $b \in B$ such that $d(a, B) \neq h(A, B)$ or $d(b, A) \neq h(A, B)$, then there are infinitely many elements $C \in \mathcal{H}(\mathbb{R}^n)$ between A and B at any location. We can improve this result. Under the hypotheses, the authors find an injective map from the open interval $(0, \epsilon)$ to the collection of elements in $\mathcal{H}(\mathbb{R}^n)$ that lie between A and B at a given location to conclude that there are infinitely many such elements, but this implies further that under the assumptions of the lemma, then there are, in

fact, uncountably many elements of $\mathcal{H}(\mathbb{R}^n)$ between A and B at any location, since $(0, \epsilon)$ is uncountable.

In light of this observation, we employ the following definition:

Definition 2.5 [Blackburn et al. 2009]. A configuration $[A, B]$ is a pair of sets $A, B \in \mathcal{H}(\mathbb{R}^n)$ with $A \neq B$ such that

$$h(A, B) = d(b, A) = d(a, B) \quad \text{for all } a \in A \text{ and } b \in B.$$

It follows that if the pair $A, B \in \mathcal{H}(\mathbb{R}^n)$ is not a configuration, then the number of elements at any location between A and B is uncountable. Hence, if there are countably many elements at each location between A and B, then the pair must be a configuration $[A, B]$. Note that the elements A and B described in the example above constitute a configuration, and so being a configuration is necessary but not sufficient for there to be countably many elements of $\mathcal{H}(\mathbb{R}^n)$ at a given location between A and B.

We adopt the notation $\#([A, B])_t$ to represent the cardinality of the collection of elements between A and B in a configuration at a distance $0 < t < h(A, B)$ from A. Blackburn et al. [2009] demonstrated that when A and B are finite sets, $\#([A, B])_t$ is finite and $\#([A, B])_s = \#([A, B])_t$ for every s, t satisfying $0 < s, t < h(A, B)$. In this case, $\#([A, B])_t$ is simply denoted $\#([A, B])$.

3. An alternative characterization of $\#([A, B])_t$

Let $(A)_t$ denote the dilation of $A \in \mathcal{H}(\mathbb{R}^n)$ by t; that is, $(A)_t = \{x \in \mathbb{R}^n : d(x, A) \leq t\}$. In addition, for elements $A, B \in \mathcal{H}(\mathbb{R}^n)$ with $t + s = h(A, B)$, $t, s > 0$, define $C(t)$ to be the set $(A)_t \cap (B)_s$. Lemma 3.6 of [Bogdewicz 2000] shows that for such A, B, t, and s, the set $C(t)$ is between A and B at a distance t from A. In [Braun et al. 2005] it is shown that any element $C \in \mathcal{H}(\mathbb{R}^n)$ with $h(A, C) = t$ satisfies $C \subset (A)_t$. It follows that if C is any element between A and B at a distance t from A, then $C \subset C(t)$. Thus, we can think of $C(t)$ as the largest element between A and B at a distance t from A. From this point forward, we set the convention that for any configuration $[A, B]$ with $0 < t < h(A, B)$, we have $s = h(A, B) - t$ and $C(t) = (A)_t \cap (B)_s$.

Example 3.1. Using our previous example with $t = s = 1$, we can see that $(A)_t$ is the union of the unit disk with an annulus with inner and outer radii of three and five, while $(B)_s$ is an annulus with inner and outer radii of one and three, so that $C(t) = (A)_t \cap (B)_s = S^1(1) \cup S^1(3) = C$, where C is the set pictured on the left side of Figure 1.

In general, one way to determine $\#([A, B])_t$ is to count the number of elements of $\mathcal{H}(\mathbb{R}^n)$ at a location t from A between A and B. Alternatively, we can count the

number of ways to remove subsets $U \subset \mathbb{R}^n$ from the largest set $C(t)$ between A and B to get another element $C(t) \setminus U \in \mathcal{H}(\mathbb{R}^n)$ between A and B at a distance t from A. We note immediately that if $C(t) \setminus U$ is to be compact, U must be open in $C(t)$.

Recall that in a configuration $[A, B]$, we have that $h(A, B) = d(a, B) = d(b, A)$ for every $a \in A$ and $b \in B$. Thus, by the compactness of B, for every $a \in A$ there must be at least one $b \in B$ such that $d_E(a, b) = h(A, B)$, and likewise for each $b \in B$. This relation between pairs of points in A and B proves to be especially important, and so we make the following definition:

Definition 3.2. Let $A, B \in \mathcal{H}(\mathbb{R}^n)$. We say that $a \in A$ and $b \in B$ are *adjacent*, and write $a \Subset b$, if $d_E(a, b) = h(A, B)$. The adjacency set of a in B, $[a]_B$, is defined to be $[a]_B = \{b \in B : a \Subset b\}$.

Note that under the definition of adjacency, it is not necessary for the sets A and B to form a configuration, but for a configuration $[A, B]$, we have that for any $a \in A$ and $b \in B$, both $[a]_B$ and $[b]_A$ are nonempty. Referring back to our original example, we have for the origin $a_0 \in A$ that $[a_0]_B = B$, since every point $b \in B$ satisfies $d_E(a_0, b) = 2$. On the other hand, for any $b \in B$, we can write $b = 2e^{i\theta}$, and $[b]_A$ consists of the origin a_0 and $4e^{i\theta}$.

Suppose a configuration $[A, B]$ has largest set $C(t)$ between A and B. Lemma 3.1 of [Blackburn et al. 2009] says that for every $c \in C(t)$, there is precisely one $a \in A$ and $b \in B$ such that $c \Subset a$ and $c \Subset b$. Also, $[a]_{C(t)}$ and $[b]_{C(t)}$ are nonempty. Thus, the functions $q_A : C(t) \to A$ and $q_B : C(t) \to B$ that map c to these unique points a and b, respectively, are both well-defined and onto.

Now, we return to the idea of deciding which sets U we can remove from $C(t)$ to get some element $C(t) \setminus U$ between A and B at the same location. Observe that if we remove every point in $[a]_{C(t)}$ for some $a \in A$, then $d(a, C(t) \setminus U) > d(a, C(t)) = t$, and thus $C(t) \setminus U$ cannot be between A and B at the same location. Similarly, we cannot remove every point in $C(t)$ adjacent to some $b \in B$. Thus, we define a new collection of sets Υ_t, which will turn out to be the collection of removable sets described above:

Given $[A, B]$ and $0 < t < h(A, B)$, define Υ_t to be the collection of sets U open in $C(t)$ such that

- for every $a \in q_A(U)$, $[a]_{C(t)} \setminus U \neq \varnothing$, and
- for every $b \in q_B(U)$, $[b]_{C(t)} \setminus U \neq \varnothing$.

These two conditions ensure that for any $U \in \Upsilon_t$, we have $C(t) \setminus U$ is between A and B at a distance t from A. Note that $\varnothing$ is always an element of Υ_t, and $C(t)$ is never such an element. We set one more convention, that if $[A, B]$ is a configuration and $0 < t < h(A, B)$, then $\mathcal{H}_t$ is the collection of all elements of $\mathcal{H}(\mathbb{R}^n)$ between A and B at a distance t from A. More precisely, we have:

Theorem 3.3 [Blackburn et al. 2009]. *For any configuration* $[A, B]$ *and any* t *satisfying* $0 < t < h(A, B)$, *the function* $f : \Upsilon_t \to \mathcal{K}_t$ *defined by* $f(U) = C(t) \setminus U$ *is a bijection.*

From the theorem it follows that $\#([A, B])_t = |\mathcal{K}_t| = |\Upsilon_t|$. This is the exact tool we will need to show that no configuration $[A, B]$ and $0 < t < h(A, B)$ satisfies $\#([A, B])_t = |\mathbb{Z}|$. In our example, with $t = 1$, Υ_1 is the collection of all sets $U \neq S^1(1)$ that are open in $S^1(1)$. We observe that in this case Υ_1 is uncountable, verifying that there are uncountably many elements of $\mathcal{H}(\mathbb{R}^n)$ between A and B at a distance one from A.

4. Orders of infinity between sets in a configuration

Recall from above that if two elements A, B do not form a configuration, then there are uncountably many elements between A and B at every location. Thus, we may restrict our search for pairs of sets $A, B \in \mathcal{H}(\mathbb{R}^n)$ with countably many such elements to configurations. We use the fact that if $[A, B]$ is a configuration with $0 < t < h(A, B)$, then as stated above, $\#([A, B])_t = |\Upsilon_t|$.

We will need a definition and two lemmas to prove our main result.

Definition 4.1. A point w contained in a set W is a cluster point of W if for every $\epsilon > 0$, $B_\epsilon(w) \cap (W \setminus \{w\}) \neq \varnothing$. If w is not a cluster point, it is isolated.

The first lemma follows directly from our definition of Υ_t.

Lemma 4.2. *For a configuration* $[A, B]$ *with* $0 < t < h(A, B)$, *let* $U \in \Upsilon_t$. *If* $V \subset U$ *such that* V *is open in* C_t, *then* $V \in \Upsilon_t$ *as well.*

Proof. By definition of Υ_t, we have that

- for all $a \in q_A(U)$ there exists $c \in [a]_{C(t)}$ such that $c \notin U$ and
- for all $b \in q_B(U)$ there exists $c \in [b]_{C(t)}$ such that $c \notin U$.

This means that for all $a \in q_A(V)$, we must have that $a \in q_A(U)$ as $V \subset U$. Thus, there exists $c \in [a]_{C(t)}$ such that $c \notin U$, and so $c \notin V$. Similarly, for every $b \in q_B(V)$, there exists $c \in [b]_{C(t)}$ such that $b \notin V$. It follows that $V \in \Upsilon_t$. □

In other words, if we can remove some set of points U from $C(t)$ to get an element of $\mathcal{K}_t$, then certainly we can remove some relatively open subset V of U from $C(t)$ to get another element of $\mathcal{K}_t$. The next lemma takes a bit more work and lies at the core of our argument.

Lemma 4.3. *For a configuration* $[A, B]$ *with* $0 < t < h(A, B)$, *if* $\#([A, B])_t = \infty$, *then there exists* $W \in \Upsilon_t$ *such that* $|W| = \infty$.

Proof. Suppose by way of contradiction that $|U|$ is finite for every $U \in \Upsilon_t$. Let $U \in \Upsilon_t$, and choose a point $x \in U$. As $|U|$ is finite, we can find some $\epsilon_x > 0$ such that $B_{\epsilon_x}(x) \cap U = \{x\}$. Further, since U is relatively open in $C(t)$, we can choose ϵ_x to be small enough so that $B_{\epsilon_x}(x) \cap C \subset U$. Hence, if $V_x = B_{\epsilon_x}(x) \cap C$, then $V_x = \{x\}$ and certainly V_x is open in C and $V_x \subset U$, so by Lemma 4.2, $V_x \in \Upsilon_t$. Since $|\Upsilon_t| = \infty$ and every element $U \in \Upsilon_t$ can be written as a union of sets V_x, we must have infinitely many such singleton point sets as well.

Define

$$V = \bigcup_{\substack{x \in U \\ U \in \Upsilon_t}} V_x.$$

We split the proof into two cases. Suppose first that there exists some $a \in q_A(V)$ such that $|[a]_V| = \infty$. This means that a is adjacent to infinitely many points in V. Note that if $v_1, v_2 \in V$ satisfy $v_1 \mathrel{\underset{\sim}{\sim}} a_0$, $v_2 \mathrel{\underset{\sim}{\sim}} a_0$, $v_1 \mathrel{\underset{\sim}{\sim}} b_0$, and $v_2 \mathrel{\underset{\sim}{\sim}} b_0$ for some $a_0 \in A$ and $b_0 \in B$, then $v_1 = v_2$ by the uniqueness of betweenness in Euclidean geometry. Thus, every pair of distinct points v_1 and v_2 in $[a]_V$ must be adjacent to distinct points in B, or equivalently, q_B is injective on $[a]_V$. Fix a point $v^* \in [a]_V$, and let $W = [a]_V \setminus \{v^*\}$.

It is clear that $|W| = \infty$. We claim that $W \in \Upsilon_t$. First, note that W is the union of singleton point sets, each of which is open in $C(t)$, so W is open in C. We have established that $q_A(W) = \{a\}$, where $v^* \in [a]_{C(t)}$ but $v^* \notin W$. Now, let $b \in q_B(W)$. By the argument above, we have that b is adjacent to exactly one point $w \in W$. Further, $\{w\} \in \Upsilon$, so there exists some $c \in [b]_C$ such that $c \notin \{w\}$; that is, $c \neq w$. Since b is adjacent to no other points in W, it follows that $c \notin W$, as desired. We conclude that $W \in \Upsilon$, which is a contradiction to the assumption that every set in Υ_t is finite. A similar proof holds if there exists some $b \in q_B(V)$ such that $|[b]_V| = \infty$.

In the second case suppose that $[a]_V$ and $[b]_V$ are finite for every $a \in q_A(V)$ and $b \in q_B(V)$. Choose a point $v_1 \in V$ and let $a_1 = q_A(v_1)$ and $b_1 = q_B(v_1)$. Since $[a_1]_V$ and $[b_1]_V$ are finite while V is infinite, we can choose $v_2 \neq v_1 \in V$ such that v_2 is adjacent to neither a_1 nor b_1. Continuing in this manner, we can construct three infinite sequences of distinct points $\{a_i\}_i$, $\{b_i\}_i$, and $\{v_i\}_i$ such that v_m is adjacent to a_m and b_m but v_m is not adjacent to any of the points $a_1, \ldots, a_{m-1}, b_1, \ldots, b_{m-1}$.

Let $W = \bigcup v_i$. Then certainly $|W| = \infty$ and W is open in C as the union of open singleton sets. We claim that $W \in \Upsilon_t$. For any $a \in q_A(W)$, we know that $a = a_m$ for some integer m. Now $a_m \mathrel{\underset{\sim}{\sim}} v_m$, and since $\{v_m\} \in \Upsilon_t$, we know that there exists some $c \in [a]_C$ such that $c \neq v_m$. But by definition of the sequence $\{v_i\}_i$, $c \neq v_i$ for any $i \neq m$, and thus $c \notin W$. A similar argument shows that for every $b \in q_B(W)$, there exists $c \in [b]_C$ such that $c \notin W$. We conclude that $W \in \Upsilon_t$, which is a contradiction, completing the proof. □

Finally, we are in a position to prove our main theorem.

Theorem 4.4. *There exist no two sets A and B that have a countably infinite number of elements at any given location between A and B.*

Proof. Suppose by way of contradiction there exists a configuration $[A, B]$ and some t, $0 < t < h(A, B)$, such that $\#([A, B])_t = |\Upsilon_t| = |\mathbb{Z}|$. Thus, by Lemma 4.3 there exists some element $W \in \Upsilon_t$ such that $|W| = \infty$. We will find an infinite family of nonempty disjoint open subsets of C contained in W. There are two cases to consider. Suppose first that W contains infinitely many points which are isolated in W, and call a countably infinite subset of these points $\{w_i\}_i$. By definition $w_m \in W$ is isolated if there exists a ball $B_{\epsilon_m}(w_m)$ such that $B_{\epsilon_m}(w_m) \cap W = \{w_m\}$, and by choosing ϵ small enough, we can guarantee that $B_{\epsilon_m}(w_m) \cap C(t) \subset W$. Thus, if $W_m = B_{\epsilon_m}(w_m) \cap C = \{w_m\}$, then $\{W_i\}_i$ is a family of infinite disjoint open subsets of $C(t)$ contained in W.

In the second case, suppose that W contains finitely many isolated points, so that W contains infinitely many cluster points. Choose some cluster point $w_1 \in W$. Since there are finitely many isolated points in W, we can choose $\epsilon_1 > 0$ such that $B_{\epsilon_1}(w_1) \cap W$ contains only cluster points. Since w_1 is itself a cluster point in W, we know $|B_{\epsilon_1}(w_1) \cap W|$ must be infinite, and so if we shrink ϵ_1 further we can find a cluster point $w_2 \in W$ such that $w_2 \notin \overline{B_{\epsilon_1}(w_1)}$. Now, we choose ϵ_2 such that $B_{\epsilon_2}(w_2) \cap W$ consists of infinitely many cluster points and $B_{\epsilon_2}(w_2) \cap B_{\epsilon_1}(w_1) = \varnothing$. Shrinking ϵ_2 slightly yields a cluster point $w_3 \in W$ such that w_3 is in neither $\overline{B_{\epsilon_1}(w_1)}$ nor $\overline{B_{\epsilon_2}(w_2)}$. Continuing in this fashion, we can find an infinite sequence $\{w_i\}_i$ in W with corresponding radii $\{\epsilon_i\}_i$. Let $W_m = B_{\epsilon_m}(w_m) \cap W$. Then $\{W_i\}_i$ is a family of infinite disjoint open subsets of C contained in W.

In either case, we find a family $\{W_i\}_i$ of infinite pairwise disjoint open subsets of C contained in W. Let $2^{\mathbb{Z}}$ be the power set of $\mathbb{Z}$ and define a map $g : 2^{\mathbb{Z}} \to \Upsilon_t$ by

$$g(S) = \bigcup_{i \in S} W_i.$$

First, we note that for any $S \subset \mathbb{Z}$, we have $g(S) \in \Upsilon$ by Lemma 4.2, using the fact that each W_i is an open subset of $W \in \Upsilon$, so $\bigcup_{i \in S} W_i$ is also an open subset of W. Next, we claim that g is injective. But this is clear from the fact that the sets in $\{W_i\}_i$ are disjoint: if $S \neq S'$, then without loss of generality there is some $m \in S$ such that $n \notin S'$, so $W_m \subset g(S)$ whereas $W_m \cap g(S') = \varnothing$, and thus $g(S) \neq g(S')$. It follows directly that $|\Upsilon| \geq |2^{\mathbb{Z}}| = |\mathbb{R}|$. We conclude that $\#([A, B])_t \geq |\mathbb{R}|$, proving the theorem. □

Acknowledgments

We thank Steve Schlicker for his help with the revision of this paper. This research was partially supported by National Science Foundation grant DMS-0451254.

References

[Blackburn et al. 2009] C. C. Blackburn, K. Lund, S. Schlicker, P. Sigmon, and A. Zupan, "A missing prime configuration in the Hausdorff metric geometry", *J. Geom.* **92**:1–2 (2009), 28–59. MR 2010d:51020 Zbl 1171.54008

[Bogdewicz 2000] A. Bogdewicz, "Some metric properties of hyperspaces", *Demonstratio Math.* **33**:1 (2000), 135–149. MR 1759874 Zbl 0948.54015

[Braun et al. 2005] D. Braun, J. Mayberry, A. Powers, and S. Schlicker, "A singular introduction to the Hausdorff metric geometry", *Pi Mu Epsilon Journal* **12**:3 (2005), 129–138.

[Edgar 1990] G. A. Edgar, *Measure, topology, and fractal geometry*, Springer, New York, 1990. MR 92a:54001 Zbl 0727.28003

Received: 2010-09-22 Revised: 2014-04-23 Accepted: 2014-05-11

zupan@math.utexas.edu *Department of Mathematics, University of Texas at Austin, 1 University Station C1200, Austin, TX 78712, United States*

INVOLVE 7:5 (2014)
dx.doi.org/10.2140/involve.2014.7.595

Computing positive semidefinite minimum rank for small graphs

Steven Osborne and Nathan Warnberg

(Communicated by Chi-Kwong Li)

The positive semidefinite minimum rank of a simple graph G is defined to be the smallest possible rank over all positive semidefinite real symmetric matrices whose ij-th entry (for $i \neq j$) is nonzero whenever $\{i, j\}$ is an edge in G and is zero otherwise. The computation of this parameter directly is difficult. However, there are a number of known bounding parameters and techniques which can be calculated and performed on a computer. We programmed an implementation of these bounds and techniques in the open-source mathematical software Sage. The program, in conjunction with the orthogonal representation method, establishes the positive semidefinite minimum rank for all graphs of order 7 or less.

1. Introduction

Define a *graph* $G = (V, E)$ with *vertex set* $V = V(G)$ and *edge set* $E = E(G)$. The graphs discussed herein are simple (no loops or multiple edges) and undirected. The *order* of G, $|G|$, is the cardinality of $V(G)$. Two vertices v and w of a graph G are *neighbors* if $\{v, w\} \in E(G)$. If H is a graph with $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$ we call H a *subgraph* of G. H is an *induced subgraph* of G if H is a subgraph of G and if for all pairs $v, w \in V(H)$, $\{v, w\} \in E(H)$ if $\{v, w\} \in E(G)$. Given a set of vertices $S \subseteq V(G)$, $G - S$ is the induced subgraph of G with vertices $V(G) \setminus S$.

A graph $P = (V, E)$, where $V(P) = \{v_1, v_2, \dots, v_n\}$, is called a *path* if the edges of the graph are exactly $\{v_i, v_{i+1}\}$ for $i = 1, 2, \dots, n-1$. A *cycle* is a path that also has the edge $\{v_n, v_1\}$. A graph G is *chordal* if every induced cycle has length no greater than 3. A graph is *connected* if for any two vertices, v_1, v_2, there exists a path with endpoints v_1 and v_2. A connected graph with no cycles is a *tree*. An induced graph that is a tree is an *induced tree*. A graph with n vertices in which there is an edge between every vertex is called a *complete* graph and is denoted K_n. See Figure 1 for examples.

MSC2010: primary 05C50; secondary 15A03.
Keywords: zero forcing number, maximum nullity, minimum rank, positive semidefinite, zero forcing, graph, matrix.

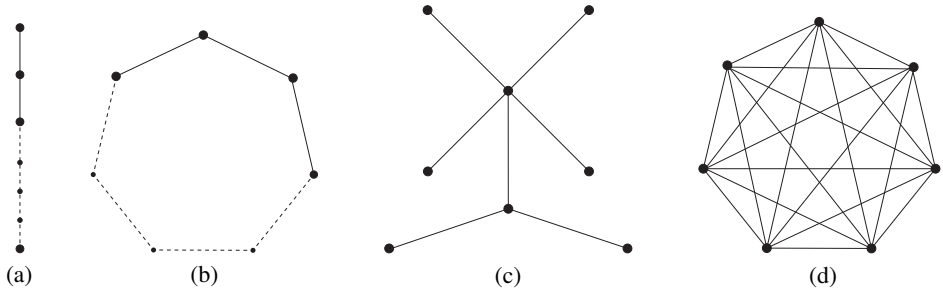


Figure 1. Examples of graphs: (a) a path; (b) a cycle; (c) a tree; (d) the complete graph on 7 vertices.

Let $S_n(\mathbb{R})$ denote the set of real symmetric $n \times n$ matrices. For $A = [a_{ij}] \in S_n(\mathbb{R})$, the *graph of* A, denoted $\mathcal{G}(A)$, is the graph with vertices $\{1, 2, \dots, n\}$ and edges $\big\{\{i, j\} : a_{ij} \neq 0 \text{ and } i \neq j\big\}$.

The *positive semidefinite maximum nullity* of G is

$$\mathrm{M}_+(G) = \max\{\text{null } A : A \in S_n(\mathbb{R}) \text{ is positive semidefinite and } \mathcal{G}(A) = G\}$$

and the *positive semidefinite minimum rank* of G is

$$\mathrm{mr}_+(G) = \min\{\text{rank } A : A \in S_n(\mathbb{R}) \text{ is positive semidefinite and } \mathcal{G}(A) = G\}.$$

Clearly $\mathrm{mr}_+(G) + \mathrm{M}_+(G) = |G|$.

The following concept was introduced in [Barioli et al. 2010]: in a graph G where all vertices in some vertex set $S \subseteq V(G)$ are colored black and the remaining vertices are colored white, the *positive semidefinite color change rule* is: If $W_1, W_2, \dots, W_k$ are the sets of vertices of the k connected components of $G - S$ ($k = 1$ is a possibility), $w \in W_i$, $u \in S$, and w is the only white neighbor of u in the subgraph of G induced by $V(W_i \cup S)$, then change the color of w to black, written as $u \to w$. Given an initial set B of black vertices, the *final coloring* of B is the set of vertices colored black as result of applying the positive semidefinite color change rule iteratively until no more vertices may be colored black. If the final coloring of B is $V(G)$, B is called a *positive semidefinite zero forcing set* of G. The *positive semidefinite zero forcing number* of a graph G, denoted $\mathrm{Z}_+(G)$, is the minimum of $|B|$ for all B positive semidefinite zero forcing sets of G. In [Barioli et al. 2010] it was shown that if G is a graph then $\mathrm{M}_+(G) \leq \mathrm{Z}_+(G)$.

Example 1.1. Consider the graph G in Figure 2(a) with the set $B = \{v_4\}$ initially colored black. When the positive semidefinite color change rule is applied, the connected component W_1 of $G - B$ is the induced subgraph of G on the vertices $\{v_1, v_2, v_3\}$. Since v_3 is the only white neighbor of v_4 in the subgraph of G induced by $W_1 \cup B$ (this is actually all of G), $v_4 \to v_3$ as demonstrated in Figure 2(b). For the

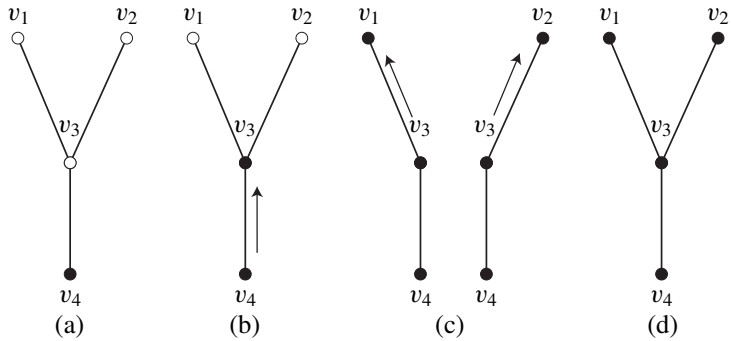


Figure 2. Illustrating Example 1.1.

next iteration, the set of black vertices is $B' = \{v_3, v_4\}$. The connected components of $G - B'$ are W_1', induced by $\{v_1\}$, and W_2', induced by $\{v_2\}$. Vertex v_1 is the only white neighbor of vertex v_3 in the subgraph of G induced by $W_1' \cup B'$ and v_2 is the only white neighbor of vertex v_3 in the subgraph of G induced by $W_2' \cup B'$. Therefore, $v_3 \to v_1$ and $v_3 \to v_2$; see Figure 2(c). Now, the entire graph has been forced black, as shown in Figure 2(d), and since the process was started by a single black vertex, $\mathrm{Z}_+(G) \le 1$. However, at least one vertex must be colored to begin the zero forcing process. Therefore, $\mathrm{Z}_+(G) = 1$.

Let G be a graph and S the smallest subset of $V(G)$ such that $G - S$ is disconnected. Then $|S| = \kappa(G)$ is called the *vertex connectivity* of G. A *clique covering* of G is a set of induced subgraphs $\{S_i\}$ of G such that each S_i is complete and $E(G) = \bigcup E(S_i)$. The *clique cover number* of a graph G, denoted $\mathrm{cc}(G)$, is the minimum of $|\{S_i\}|$ over all $\{S_i\}$ clique coverings of G.

In [Booth et al. 2008] $\mathrm{M}_+(G)$ was determined for every graph G of order at most 6. Use of published software (Zq.py; see [Butler and Grout 2011]) for computing $\mathrm{Z}_+(G)$ establishes $\mathrm{M}_+(G) = \mathrm{Z}_+(G)$ for $|G| \le 6$. We developed a program (see [Osborne and Warnberg 2011a]) in the open-source computer mathematics software system Sage (sagemath.org) to compute bounds for positive semidefinite maximum nullity. The program uses Zq.py [Butler and Grout 2011] and known results for computing positive semidefinite maximum nullity. These results are summarized in Section 2. A detailed description of the program may be found in Appendix A. Sections 2 and 3 provide a survey of techniques for computing positive semidefinite minimum rank.

In Section 3 we determine $\mathrm{M}_+(G)$ for $|G| \le 7$ and show $\mathrm{M}_+(G) = \mathrm{Z}_+(G)$ for all such graphs. For all but 13 graphs of order 7, $\mathrm{M}_+(G)$ can be computed by the program. We then established $\mathrm{M}_+(G)$ for the remaining 13 graphs by utilizing orthogonal representation to find a positive semidefinite matrix A with $\mathcal{G}(A) = G$ and nullity of $A = \mathrm{Z}_+(G)$. This establishes that $\mathrm{M}_+(G) = \mathrm{Z}_+(G)$ for each graph G of order at most 7. These matrices are listed in Appendix B.

2. Known results used by the program to establish positive semidefinite minimum rank/maximum nullity

Note that all of our parameters sum over the connected components of a disconnected graph. Given its relation to the positive semidefinite zero forcing number, the following results are given in terms of positive semidefinite maximum nullity. However, given a graph G, $\mathrm{M}_+(G) + \mathrm{mr}_+(G) = |G|$, so all of the following results may easily be translated to positive semidefinite minimum rank.

Theorem 2.1 [Ekstrand et al. 2013]. *Let G be a graph.*

(i) $\mathrm{Z}_+(G) = 1$ *if and only if* $\mathrm{M}_+(G) = 1$.

(ii) $\mathrm{Z}_+(G) = 2$ *if and only if* $\mathrm{M}_+(G) = 2$.

(iii) $\mathrm{Z}_+(G) = 3$ *implies* $\mathrm{M}_+(G) = 3$.

Corollary 2.2. *If* $\mathrm{Z}_+(G) \geq 3$, *then* $\mathrm{M}_+(G) \geq 3$.

Observation 2.3 [Ekstrand et al. 2013]. $\mathrm{Z}_+(G) = |G| - 1$ *if and only if* $\mathrm{M}_+(G) = |G| - 1$.

Note that the only graph G having $\mathrm{Z}_+(G) = |G| - 1$ is K_n, the complete graph on n vertices.

For a chordal graph G, it was shown in [Booth et al. 2008] that $\mathrm{cc}(G) = \mathrm{mr}_+(G)$, in [Hackney et al. 2009] it was shown that $OS(G) = \mathrm{cc}(G)$, and in [Barioli et al. 2010] it was shown that $\mathrm{Z}_+(G) + OS(G) = |G|$, where $OS(G)$ is the ordered subgraph number of G (see [Mitchell et al. 2010] for the definition of $OS(G)$). Thus $\mathrm{Z}_+(G) = \mathrm{M}_+(G)$, which gives the next theorem.

Theorem 2.4 [Barioli et al. 2010; Booth et al. 2008; Hackney et al. 2009]. *If G is chordal, then* $\mathrm{M}_+(G) = \mathrm{Z}_+(G)$.

Example 2.5. Consider graph $G551$ in Figure 3, left. Sets of vertices of size 1 or 2 are clearly not positive semidefinite zero forcing sets, so $\mathrm{Z}_+(G551) \geq 3$. Notice that choosing an initial set of 3 black vertices that are all nonadjacent does not force anything. By symmetry this reduces to two cases. In the first case we choose $\{1, 2\}$ as our adjacent black vertices and as our third we choose any of the remaining vertices and notice that the graph will not be forced. Similarly, choosing $\{1, 3\}$ as our adjacent black vertices and any of the remaining vertices as our third also fails to force the graph. Thus, $\mathrm{Z}_+(G551) \geq 4$. Observe that $\{1, 3, 4, 5\}$ forms a positive semidefinite zero forcing set meaning $\mathrm{Z}_+(G551) \leq 4$, hence $\mathrm{Z}_+(G551) = 4$. However, $G551$ is chordal as its largest cycle is size 3. Therefore, by Theorem 2.4 $\mathrm{M}_+(G551) = 4$.

Theorem 2.6 [Lovász et al. 1989; 2000]. *For every graph G, $\kappa(G) \leq \mathrm{M}_+(G)$.*

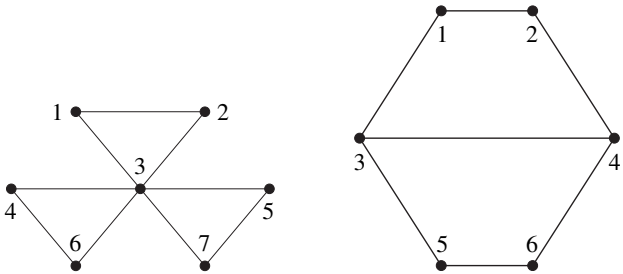


Figure 3. Graphs $G551$ (left) and $G128$ (right).

Example 2.7. By inspection, removing any one vertex from graph $G128$ (see Figure 3, right) will not result in a disconnected graph. Therefore, $\kappa(G) \geq 2$. Further, $\{3, 4\}$ forms a positive semidefinite zero forcing set for $G128$. Thus, $\mathrm{Z}_+(G) \leq 2$. This gives $2 \leq \kappa(G) \leq \mathrm{M}_+(G) \leq \mathrm{Z}_+(G) \leq 2$.

For a graph G the *neighborhood* of $v \in V(G)$ is

$$N_G(v) = \{w \in V(G) \mid v \text{ is adjacent to } w\}.$$

Vertices v and w are called *duplicate vertices* if $N_G(v) \cup \{v\} = N_G(w) \cup \{w\}$.

Proposition 2.8 [Ekstrand et al. 2013]. *If v and w are duplicate vertices in a connected graph G with $|G| \geq 3$, then $\mathrm{Z}_+(G - v) = \mathrm{Z}_+(G) - 1$.*

Proposition 2.9 [Booth et al. 2008]. *If v and w are duplicate vertices in a connected graph G with $|G| \geq 3$, then $\mathrm{mr}_+(G - v) = \mathrm{mr}_+(G)$.*

Recall that for any graph G, $\mathrm{mr}_+(G) + \mathrm{M}_+(G) = |G|$, which gives the following corollary.

Corollary 2.10. *If v and w are duplicate vertices in a connected graph G with $|G| \geq 3$, then $\mathrm{M}_+(G - v) = \mathrm{M}_+(G) - 1$.*

Example 2.11. In graph $G1196$ (see Figure 4, left) notice that 2 and 4 are duplicate vertices, as are vertices 3 and 5. Removal of vertices 2 and 3 results in a graph that is isomorphic to graph $G43$ (see Figure 4, right). $\mathrm{Z}_+(G43) = 2$ thus $\mathrm{M}_+(G43) = 2$ by Theorem 2.1. Therefore, $\mathrm{M}_+(G1196) = 4$ by Corollary 2.10.

Cut-vertex reduction is a standard technique in the study of minimum rank. A vertex v of a connected graph G is a *cut-vertex* if $G - v$ is disconnected. Suppose G_i, $i = 1, \ldots, h$, are graphs of order at least two, there is a vertex v such that for all $i \neq j$, $G_i \cap G_j = \{v\}$, and $G = \cup_{i=1}^{h} G_i$ (if $h \geq 2$, then clearly v is a cut-vertex of G). Then it is observed in [van der Holst 2009] that

$$\mathrm{mr}_+(G) = \sum_{i=1}^{h} \mathrm{mr}_+(G_i).$$

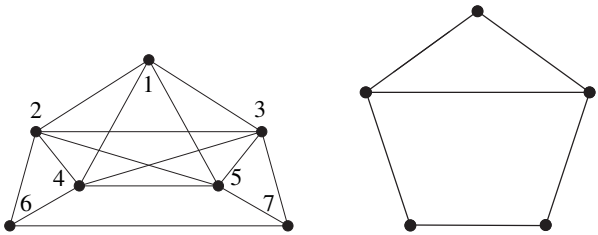


Figure 4. Graphs $G1196$ (left) and $G43$ (right).

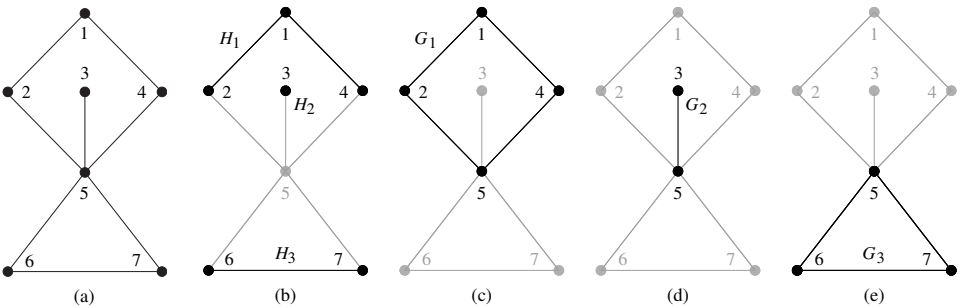


Figure 5. Graph $G419$.

Because $\mathrm{mr}_+(G) + \mathrm{M}_+(G) = |G|$, this is equivalent to

$$\mathrm{M}_+(G) = \left(\sum_{i=1}^{h} \mathrm{M}_+(G_i)\right) - h + 1. \tag{1}$$

It is shown in [Mitchell et al. 2010] that

$$OS(G) = \sum_{i=1}^{h} OS(G_i).$$

Since $OS(G) + \mathrm{Z}_+(G) = |G|$ (shown in [Barioli et al. 2010]), this is equivalent to

$$\mathrm{Z}_+(G) = \left(\sum_{i=1}^{h} \mathrm{Z}_+(G_i)\right) - h + 1. \tag{2}$$

Example 2.12. Equation (2) can be used to compute $\mathrm{Z}_+(G419)$ and $\mathrm{M}_+(G419)$ (see Figure 5(a)). Notice that vertex 5 is a cut vertex of the graph since removing it results in a disconnected graph with 3 components, namely H_1, H_2 and H_3. When vertex 5 is reconnected to each of our components it is easy to see that $G_i \cap G_j = \{5\}$ for $i, j \in \{1, 2, 3\}$ with $i \neq j$, as illustrated by Figures 5(c)–(e). It is also clear that $\cup_{i=1}^{3} G_i = G419$, $\mathrm{Z}_+(G_1) = 2$, $\mathrm{Z}_+(G_2) = 1$, and $\mathrm{Z}_+(G_3) = 2$. Thus, by

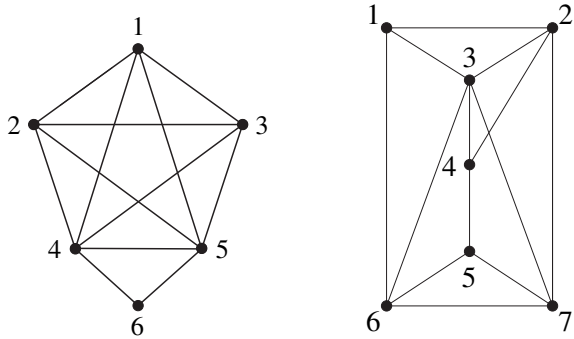


Figure 6. Graphs $G200$ (left) and $G1090$ (right).

Equation (2), $\mathrm{Z}_+(G419) = 2+1+2-3+1 = 3$. A similar argument shows that $\mathrm{M}_+(G419) = 3$.

Observe that if $\kappa(G) = 1$, there exists a cut vertex. The next result is an immediate consequence of the cut-vertex reduction Equations (1) and (2).

Observation 2.13 [Ekstrand et al. 2013]. *Suppose G_i, $i = 1, \dots, h$ are graphs, there is a vertex v such that for all $i \neq j$, $G_i \cap G_j = \{v\}$, and $G = \bigcup_{i=1}^{h} G_i$. If $\mathrm{M}_+(G_i) = \mathrm{Z}_+(G_i)$ for all $i = 1, \dots, h$, then $\mathrm{M}_+(G) = \mathrm{Z}_+(G)$.*

Observation 2.14 [Hackney et al. 2009]. *If G is a graph then $\mathrm{cc}(G) \geq \mathrm{mr}_+(G)$.*

Corollary 2.15. $|G| - \mathrm{cc}(G) \leq \mathrm{M}_+(G)$.

Example 2.16. In Figure 6, left, notice that graph $G200$ is not complete so

$$\mathrm{mr}_+(G200) \geq 2.$$

Also, note that the subgraphs induced by $S_1 = \{1, 2, 3, 4, 5\}$ and $S_2 = \{4, 5, 6\}$ are complete and $E(G200) = E(S_1) \cup E(S_2)$ so $\mathrm{cc}(G200) \leq 2$, hence $\mathrm{mr}_+(G200) = 2$.

In [Booth et al. 2008] the *tree size* of a graph G, denoted $\mathrm{ts}(G)$, is defined to be the number of vertices in a maximum induced tree of G. Also from [Booth et al. 2008], if T is a maximum induced tree and w is a vertex not belonging to T, denote by $\mathscr{E}(w)$ the set of all edges of all paths in T between every pair of vertices of T that are adjacent to w. The following theorem was established by Booth et al. [2008].

Theorem 2.17 [Booth et al. 2008]. *For a connected graph G,*

$$\mathrm{mr}_+(G) = \mathrm{ts}(G) - 1 \tag{3}$$

if the following condition holds: there exists a maximum induced tree T such that for u and w not on T, $\mathscr{E}(u) \cap \mathscr{E}(w) \neq \varnothing$ if and only if u and w are adjacent in G.

Note that Equation (3) may be rewritten as $\mathrm{M}_+(G) = |G| - \mathrm{ts}(G) + 1$.

Example 2.18. To illustrate the previous theorem we consider graph $G1090$ (see Figure 6, right). To find $\mathrm{ts}(G1090)$ notice that $G1090$ has two disjoint, induced K_3's, namely the graphs induced by vertex sets $\{1, 2, 3\}$ and $\{5, 6, 7\}$. This means in order to find an induced tree, removal of one vertex from each K_3 is required. By inspection, removal of any of the nine pairs $\big\{\{1, 5\}, \{1, 6\}, \{1, 7\}, \{2, 5\}, \dots, \{3, 7\}\big\}$ results in a graph with a cycle, thus $\mathrm{ts}(G1090) \le 4$. However, the subgraph induced by $\{1, 4, 5, 6\}$ is a tree (call it T), hence $\mathrm{ts}(G1090) = 4$. We show T satisfies the condition of Theorem 2.17. The vertices not in $G1090 - T$ are 2, 3, and 7, which are all adjacent in $G1090$.

$$\mathscr{E}(2) = \big\{(1, 6), (5, 6), (4, 5)\big\} = \mathscr{E}(3) \quad \text{and} \quad \mathscr{E}(7) = \big\{(5, 6)\big\}.$$

Therefore, $\mathscr{E}(2) \cap \mathscr{E}(3) \cap \mathscr{E}(7) \neq \varnothing$ and the condition holds because $\{2, 3, 7\}$ are pairwise adjacent. Thus $\mathrm{M}_+(G1090) = 4$.

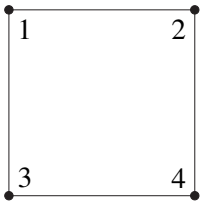
3. Computation of positive semidefinite maximum nullity of graphs of order 7 or less

The program developed by Osborne and Warnberg [2011a] implements the results from Section 2. Running the program on all graphs of order 7 or less yielded positive semidefinite maximum nullity for 1239 of 1252 graphs. It may be noted that the positive semidefinite maximum nullity was already known for the 208 graphs of order 6 or less (see [Booth et al. 2008]). However, the program was able to successfully compute the positive semidefinite maximum nullity for these graphs without referencing this information. For the remaining 13 graphs, the method of orthogonal representations was used to construct a matrix representation exhibiting nullity equal to the positive semidefinite zero forcing number. These matrices are shown in Appendix B.

A set $\vec{V} = \{\vec{v}_1, \dots, \vec{v}_n\}$ in $\mathbb{R}^d$ is an *orthogonal representation* of the graph G if for $i \neq j$, the dot product of $\vec{v}_i$ with $\vec{v}_j$ is nonzero if the vertices i and j are adjacent, and zero otherwise. If $\vec{V} = \{\vec{v}_1, \dots, \vec{v}_n\}$ is an orthogonal representation of the graph G in $\mathbb{R}^d$ and $B = [\vec{v}_1 \dots \vec{v}_n]$, then $B^T B \in \mathcal{S}_+(G)$ and rank $B^T B \le d$. Thus, if a representation is found in $\mathbb{R}^d$ then $\mathrm{mr}_+(G) \le d$ and $\mathrm{M}_+(G) \ge |G| - d$.

Example 3.1. Consider graph $G17$ in Figure 7, left. Note that when we refer to a graph in the form $G17$ we are using notation from [Read and Wilson 1998]. To start constructing an orthogonal representation for $G17$ let $v_1, v_2, v_3, v_4 \in \mathbb{R}^2$ correspond to vertices 1, 2, 3 and 4 respectively. Choose as many disjoint vertices as possible, say 1 and 4. By definition $v_1 \cdot v_4 = 0$ so let $v_1 = \left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$ and $v_4 = \left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$. To find v_2 and v_3, set

$$v_2 = \begin{bmatrix} ca_2 \\ b_2 \end{bmatrix} \quad \text{and} \quad v_3 = \begin{bmatrix} a_3 \\ b_3 \end{bmatrix}.$$



$$\begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \\ 1 & 0 & 2 & -1 \\ 0 & 1 & -1 & 1 \end{pmatrix}$$

Figure 7. Graph $G17$ (left); A, a matrix representation of $G17$ (right).

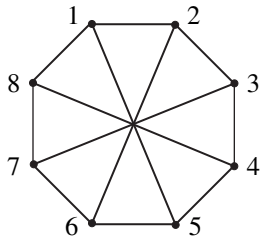


Figure 8. Möbius ladder on 8 vertices.

Now, v_2 is adjacent to v_1 and v_4 so $v_1 \cdot v_2 \neq 0$ and $v_2 \cdot v_4 \neq 0$. Thus $a_2 \neq 0 \neq b_2$. Similarly, $a_3 \neq 0 \neq b_3$. Since v_2 and v_3 are not adjacent, we know $v_2 \cdot v_3 = a_2a_3 + b_2b_3 = 0$. With these restrictions it is clear that $a_2 = a_3 = b_2 = 1$ and $b_3 = -1$ is a solution and an orthogonal representation construction is complete. This gives

$$B = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & -1 & 1 \end{bmatrix} \quad \text{and} \quad B^T B = A$$

(see Figure 7, right). By construction, $\operatorname{rank}(A) = 2$. Thus $\mathrm{mr}_+(G17) \leq 2$ and $\mathrm{M}_+(G17) \geq |G| - 2 = 2$. Observe that $\{1, 2\}$ forms a positive semidefinite zero forcing set for graph $G17$ hence $\mathrm{Z}_+(G17) \leq 2$. Finally, $2 \leq \mathrm{M}_+(G17) \leq \mathrm{Z}_+(G17) \leq 2$.

In every case, positive semidefinite maximum nullity was found to equal the positive semidefinite zero forcing number. This has established the next result.

Theorem 3.2. *If G is a graph with 7 or fewer vertices, then* $\mathrm{M}_+(G) = \mathrm{Z}_+(G)$.

See [Osborne and Warnberg 2011b] for a complete spreadsheet containing positive semidefinite maximum nullity and zero forcing number for all graphs with 7 or fewer vertices.

Corollary 3.3. *Suppose G_i, $i = 1, \ldots, h$, are graphs with $|G_i| \leq 7$, there is a vertex v such that for all $i \neq j$, $G_i \cap G_j = \{v\}$, and $G = \bigcup_{i=1}^{h} G_i$. Then* $\mathrm{M}_+(G) = \mathrm{Z}_+(G)$.

Proof. Apply Theorem 3.2 to Observation 2.13. □

Note that Theorem 3.2 cannot be extended to graphs with more than 7 vertices as $\mathrm{Z}_+(V_8) = 4$ and $\mathrm{M}_+(V_8) = 3$ (shown in [Mitchell et al. 2010]), where V_8 is the Möbius ladder on 8 vertices (see Figure 8).

Appendix A: Method used by the program

The program uses the following general method:

(1) Separate the graph into its connected components and work on each component separately. Results will be summed before reporting.

(2) Compute $\mathrm{Z}_+(G)$.

(a) If $\mathrm{Z}_+(G) \leq 3$, apply the results of Theorem 2.1.

(b) Else, use Corollary 2.2 to establish a lower bound for $\mathrm{M}_+(G)$.

(3) If $\mathrm{Z}_+(G) = |G| - 1$, apply the results of Observation 2.3.

(4) If G is chordal, apply Theorem 2.4.

(5) Compute the vertex connectivity of G ($\kappa(G)$).

(a) If $\kappa(G) = \mathrm{Z}_+(G)$, apply Theorem 2.6.

(b) Else, if $\kappa(G)$ is a tighter bound for $\mathrm{M}_+(G)$, improve the lower bound.

(6) If there are duplicate vertices in the graph, discard all but one copy by applying Corollary 2.10 and returning to step 2.

(7) Apply the cut-vertex formula iteratively by applying Equation (1) and returning to step 2 for each component.

(8) Compute the clique cover number of G.

(a) If $|G| - \mathrm{cc}(G) = \mathrm{Z}_+(G)$, apply Corollary 2.15.

(b) Else, if $\mathrm{cc}(G)$ is a tighter bound for $\mathrm{M}_+(G)$, improve the lower bound.

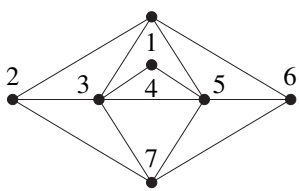
(9) Apply Theorem 2.17 to determine if $\mathrm{M}_+(G) = |G| - \mathrm{ts}(G) + 1$.

Appendix B: Matrix representations

Each of the following thirteen matrices satisfies $\mathrm{null}(A) = 4 = \mathrm{Z}_+(G)$.

$$\begin{bmatrix} 2 & -1 & -1 & 0 & 1 & 1 & 0 \\ -1 & 1 & 1 & 0 & 0 & 0 & 1 \\ -1 & 1 & 2 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 2 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 & 2 \end{bmatrix}$$

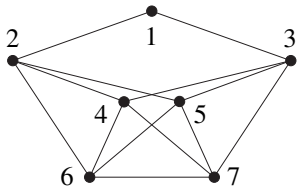
G 1060



(continued on next page)

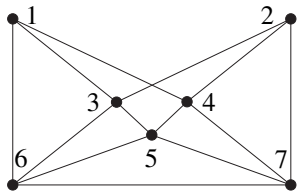
$$\begin{bmatrix} 1 & -1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 3 & 0 & -1 & 3 & 1 & 0 \\ 1 & 0 & 2 & -2 & 1 & 0 & -1 \\ 0 & -1 & -2 & 5 & 0 & 1 & 3 \\ 0 & 3 & 1 & 0 & 5 & 2 & 1 \\ 0 & 1 & 0 & 1 & 2 & 1 & 1 \\ 0 & 0 & -1 & 3 & 1 & 1 & 2 \end{bmatrix}$$

G 1075



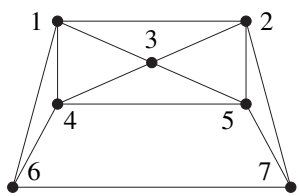
$$\begin{bmatrix} 1 & 0 & -1 & 4 & 0 & -1 & 0 \\ 0 & 1 & 4 & 2 & 0 & 0 & 1 \\ -1 & 4 & 33 & 0 & -4 & -15 & 0 \\ 4 & 2 & 0 & 21 & 1 & 0 & 3 \\ 0 & 0 & -4 & 1 & 1 & 4 & 1 \\ -1 & 0 & -15 & 0 & 4 & 17 & 4 \\ 0 & 1 & 0 & 3 & 1 & 4 & 2 \end{bmatrix}$$

G 1100



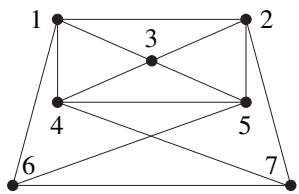
$$\begin{bmatrix} 1 & 1 & 1 & 2 & 0 & 3 & 0 \\ 1 & 6 & 7 & 0 & -1 & 0 & 1 \\ 1 & 7 & 10 & -1 & -3 & 0 & 0 \\ 2 & 0 & -1 & 5 & 1 & 7 & 0 \\ 0 & -1 & -3 & 1 & 2 & 0 & 1 \\ 3 & 0 & 0 & 7 & 0 & 11 & -1 \\ 0 & 1 & 0 & 0 & 1 & -1 & 1 \end{bmatrix}$$

G 1104



$$\begin{bmatrix} 1 & 1 & 1 & 1 & 0 & -1 & 0 \\ 1 & 3 & 2 & 0 & 1 & 0 & 1 \\ 1 & 2 & 2 & 2 & 1 & 0 & 0 \\ 1 & 0 & 2 & 6 & 1 & 0 & -2 \\ 0 & 1 & 1 & 1 & 1 & 1 & 0 \\ -1 & 0 & 0 & 0 & 1 & 2 & 0 \\ 0 & 1 & 0 & -2 & 0 & 0 & 1 \end{bmatrix}$$

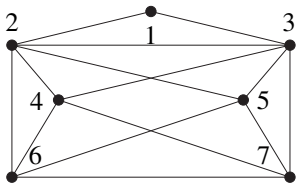
G 1105



(continued on next page)

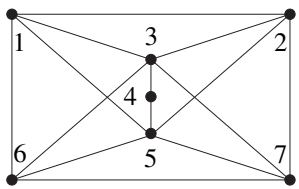
$$\begin{bmatrix} 1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 3 & -2 & 1 & 1 & 3 & 0 \\ -1 & -2 & 6 & -2 & 1 & 0 & 3 \\ 0 & 1 & -2 & 1 & 0 & 1 & -1 \\ 0 & 1 & 1 & 0 & 1 & 2 & 1 \\ 0 & 3 & 0 & 1 & 2 & 5 & 1 \\ 0 & 0 & 3 & -1 & 1 & 1 & 2 \end{bmatrix}$$

G 1135



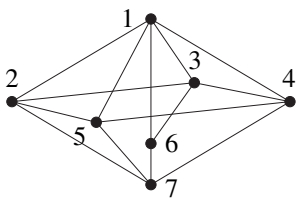
$$\begin{bmatrix} 2 & 1 & -3 & 0 & 3 & -1 & 0 \\ 1 & 1 & -2 & 0 & 2 & 0 & 1 \\ -3 & -2 & 30 & 5 & 0 & 1 & -1 \\ 0 & 0 & 5 & 1 & 1 & 0 & 0 \\ 3 & 2 & 0 & 1 & 6 & -1 & 1 \\ -1 & 0 & 1 & 0 & -1 & 1 & 1 \\ 0 & 1 & -1 & 0 & 1 & 1 & 2 \end{bmatrix}$$

G 1137



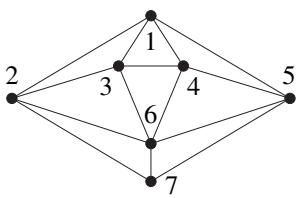
$$\begin{bmatrix} 3 & 1 & -3 & 1 & 3 & 1 & 0 \\ 1 & 1 & 2 & 0 & 2 & 0 & 1 \\ -3 & 2 & 21 & -4 & 0 & -1 & 0 \\ 1 & 0 & -4 & 1 & 1 & 0 & 1 \\ 3 & 2 & 0 & 1 & 5 & 0 & 3 \\ 1 & 0 & -1 & 0 & 0 & 1 & -2 \\ 0 & 1 & 0 & 1 & 3 & -2 & 6 \end{bmatrix}$$

G 1165



$$\begin{bmatrix} 1 & 2 & 1 & 1 & 1 & 0 & 0 \\ 2 & 6 & 1 & 0 & 0 & 2 & 1 \\ 1 & 1 & 2 & 3 & 0 & -1 & 0 \\ 1 & 0 & 3 & 5 & -1 & -2 & 0 \\ 1 & 0 & 0 & -1 & 11 & -2 & -3 \\ 0 & 2 & -1 & -2 & -2 & 2 & 1 \\ 0 & 1 & 0 & 0 & -3 & 1 & 1 \end{bmatrix}$$

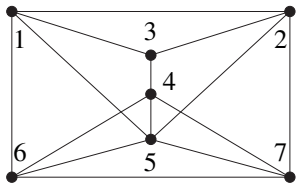
G 1167



(continued on next page)

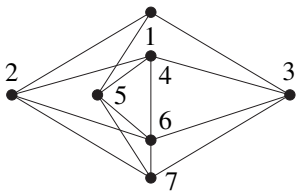
$$\begin{bmatrix} 2 & -3 & 1 & 0 & 1 & 1 & 0 \\ -3 & 6 & -1 & 0 & -1 & 0 & 1 \\ 1 & -1 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 3 & 2 & 3 & 1 \\ 1 & -1 & 0 & 2 & 2 & 3 & 1 \\ 1 & 0 & 0 & 3 & 3 & 5 & 2 \\ 0 & 1 & 0 & 1 & 1 & 2 & 1 \end{bmatrix}$$

G 1168



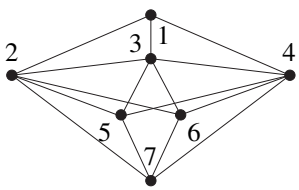
$$\begin{bmatrix} 1 & 1 & 3 & 0 & 2 & 0 & 0 \\ 1 & 6 & 0 & -2 & 0 & -1 & 1 \\ 3 & 0 & 14 & 2 & 0 & 3 & 1 \\ 0 & -2 & 2 & 1 & -1 & 1 & 0 \\ 2 & 0 & 0 & -1 & 21 & -5 & -4 \\ 0 & -1 & 3 & 1 & -5 & 2 & 1 \\ 0 & 1 & 1 & 0 & -4 & 1 & 1 \end{bmatrix}$$

G 1169



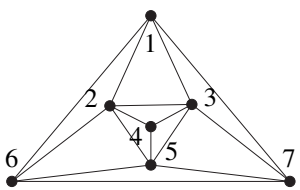
$$\begin{bmatrix} 1 & -4 & 1 & 1 & 0 & 0 & 0 \\ -4 & 21 & -2 & 0 & 1 & -3 & -1 \\ 1 & -2 & 2 & 2 & 1 & -1 & 0 \\ 1 & 0 & 2 & 6 & -1 & -3 & -2 \\ 0 & 1 & 1 & -1 & 2 & 0 & 1 \\ 0 & -3 & -1 & -3 & 0 & 2 & 1 \\ 0 & -1 & 0 & -2 & 1 & 1 & 1 \end{bmatrix}$$

G 1202



$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 1 & -3 \\ 1 & 3 & 1 & 1 & 1 & 4 & 0 \\ 1 & 1 & 3 & 3 & 1 & 0 & -4 \\ 0 & 1 & 3 & 5 & 2 & 0 & 0 \\ 0 & 1 & 1 & 2 & 1 & 1 & 1 \\ 1 & 4 & 0 & 0 & 1 & 6 & 2 \\ -3 & 0 & -4 & 0 & 1 & 2 & 14 \end{bmatrix}$$

G 1205



Acknowledgements

The authors would like to acknowledge the participants in the Early Graduate Research class 2011, led by L. Hogben, held at Iowa State University: J. Ekstrand, C. Erickson, D. Hay, R. Johnson, N. Kingsley, T. Peters, J. Roat, and A. Ross.

References

[Barioli et al. 2010] F. Barioli, W. Barrett, S. M. Fallat, H. T. Hall, L. Hogben, B. Shader, P. van den Driessche, and H. van der Holst, "Zero forcing parameters and minimum rank problems", *Linear Algebra Appl.* **433**:2 (2010), 401–411. MR 2011g:15002 Zbl 1209.05139

[Booth et al. 2008] M. Booth, P. Hackney, B. Harris, C. R. Johnson, M. Lay, L. H. Mitchell, S. K. Narayan, A. Pascoe, K. Steinmetz, B. D. Sutton, and W. Wang, "On the minimum rank among positive semidefinite matrices with a given graph", *SIAM J. Matrix Anal. Appl.* **30**:2 (2008), 731–740. MR 2009g:15003 Zbl 1226.05151

[Butler and Grout 2011] S. Butler and J. Grout, "Zq.py", 2011, https://github.com/jasongrout/minimum_rank/blob/master/Zq.py.

[Ekstrand et al. 2013] J. Ekstrand, C. Erickson, H. T. Hall, D. Hay, L. Hogben, R. Johnson, N. Kingsley, S. Osborne, T. Peters, J. Roat, A. Ross, D. D. Row, N. Warnberg, and M. Young, "Positive semidefinite zero forcing", *Linear Algebra Appl.* **439**:7 (2013), 1862–1874. MR 3090441 Zbl 1283.05165

[Hackney et al. 2009] P. Hackney, B. Harris, M. Lay, L. H. Mitchell, S. K. Narayan, and A. Pascoe, "Linearly independent vertices and minimum semidefinite rank", *Linear Algebra Appl.* **431**:8 (2009), 1105–1115. MR 2011a:15016 Zbl 1188.05085

[van der Holst 2009] H. van der Holst, "On the maximum positive semi-definite nullity and the cycle matroid of graphs", *Electron. J. Linear Algebra* **18** (2009), 192–201. MR 2010g:05216 Zbl 1173.05031

[Lovász et al. 1989] L. Lovász, M. Saks, and A. Schrijver, "Orthogonal representations and connectivity of graphs", *Linear Algebra Appl.* **114/115** (1989), 439–454. MR 90k:05095 Zbl 0681.05048

[Lovász et al. 2000] L. Lovász, M. Saks, and A. Schrijver, "A correction: "Orthogonal representations and connectivity of graphs" [Linear Algebra Appl. **114/115** (1989), 439–454; MR 90k:05095, Zbl 0681.05048]", *Linear Algebra Appl.* **313**:1-3 (2000), 101–105. MR 2001g:05070 Zbl 0954.05032

[Mitchell et al. 2010] L. H. Mitchell, S. K. Narayan, and A. M. Zimmer, "Lower bounds in minimum rank problems", *Linear Algebra Appl.* **432**:1 (2010), 430–440. MR 2010m:15004 Zbl 1220.05077

[Osborne and Warnberg 2011a] S. Osborne and N. Warnberg, "Program for calculating bounds of positive semidefinite maximum nullity of a graph using Sage", 2011, https://github.com/sosborne/psd_min_rank/blob/master/msr_program.py.

[Osborne and Warnberg 2011b] S. Osborne and N. Warnberg, "Spreadsheet of positive semidefinite maximum nullity and zero forcing number of graphs with 7 or fewer vertices", 2011, https://github.com/sosborne/psd_min_rank/blob/master/data/MpZpSpreadsheet.csv.

[Read and Wilson 1998] R. C. Read and R. J. Wilson, *An atlas of graphs*, Oxford University Press, 1998. MR 2000a:05001 Zbl 0908.05001

Received: 2011-07-19 Revised: 2011-12-13 Accepted: 2013-12-12

sosborne@iastate.edu *Department of Mathematics, Iowa State University, Ames, IA 50011, United States*

warnberg@iastate.edu *Department of Mathematics, Iowa State University, Ames, IA 50011, United States*

INVOLVE 7:5 (2014)
dx.doi.org/10.2140/involve.2014.7.611

The complement of Fermat curves in the plane

Seth Dutter, Melissa Haire and Ariel Setniker

(Communicated by Bjorn Poonen)

In this paper we will examine the affine algebraic curves defined on the complement of Fermat curves of degree five or higher in the affine plane. In particular we will bound the height of integral points over an affine curve outside of an exceptional set.

1. Introduction

Let C be a complete algebraic curve of genus g over an algebraically closed field k of characteristic 0, and $U \subset C$ be a nonempty open subset of C. The goal of this paper is to analyze morphisms

$$\phi : U \to \mathbb{P}^2_k \setminus V(z(x^n + y^n - z^n)),$$

where $V(z(x^n + y^n - z^n))$ is the zero set of the polynomial $z(x^n + y^n - z^n)$ and $[x : y : z]$ are the projective coordinates in $\mathbb{P}^2_k$. Alternatively we can think of such functions as U-points on the variety $\mathbb{P}^2_k \setminus V(z(x^n + y^n - z^n))$. We will call projective curves defined by equations of the form $x^n + y^n - z^n = 0$ Fermat curves of degree n.

A conjecture of Vojta [1987, Conjecture 3.4.3 and Proposition 4.1.2] implies that the set of integral points on the complement of a degree 4 divisor with normal crossings in $\mathbb{P}^2_k$ is not Zariski dense. Here instead of studying points over $\mathbb{Z}$ we will be looking at the split function field case of Vojta's conjecture and studying points over U. Corvaja and Zannier [2008] have proven the particular case when the divisor consists of two lines and a conic section meeting only with normal crossings. This was one of the remaining borderline cases, the other two being the union of a cubic and a line, and a quartic.

The divisor defined by $(z(x^n + y^n - z^n))$ has degree $n + 1$ and normal crossings, so we should expect that the set of U-points is not Zariski dense for $n \geq 3$. The techniques employed in this paper are able to establish results for $n \geq 5$. A counterexample is given for the case $n = 2$, leaving the cases $n = 3$ and $n = 4$ unsettled. In particular we will establish the following theorem:

MSC2010: 14G25.
Keywords: Mason's theorem, function field, Fermat curve.

Theorem 1.1. *Let C be a smooth complete algebraic curve of genus g over an algebraically closed field k of characteristic 0. Let $U \subset C$ be a nonempty open subset and $m = \#(C \setminus U)$. For $n \geq 5$ and any morphism $\phi : U \to \mathbb{P}^2_k \setminus V(z(x^n + y^n - z^n))$ either*

$$h(\phi^*(x/z), \phi^*(y/z), 1) \leq \frac{3(m + \max\{2g-2, 0\})}{n-4},$$

or $\mathrm{Im}(\phi) \subset V((x^n + y^n)(x^n - z^n)(y^n - z^n))$.

Here h is the height function over the function field of the curve C and $\phi^* : \mathcal{O}_{\mathbb{P}^2_k}(\mathbb{P}^2_k \setminus V(z(x^n + y^n - z^n))) \to \mathcal{O}_C(U)$ is the morphism of regular functions associated to ϕ. Note that the complement of the line defined by $z = 0$ can be identified with $\mathbb{A}^2_k$. Therefore we can also interpret our main result as a height bound for U-points on the complement of an affine Fermat curve in $\mathbb{A}^2_k$.

In order to establish this bound we will translate the problem into one of solving a diophantine equation over $\mathcal{O}_C(U)$. Indeed, let $\phi : U \to \mathbb{P}^2 \setminus V(z(x^n + y^n - z^n))$; then $\phi^*((x/z)^n + (y/z)^n - 1) \in \mathcal{O}_C(U)^*$ is a unit. For convenience let $X = \phi^*(x/z)$, $Y = \phi^*(y/z)$, which after substituting gives us the equation

$$X^n + Y^n - 1 = u$$

for some $u \in \mathcal{O}_C(U)^*$. Therefore we can restate our main theorem as follows:

Theorem 1.2. *Let C be a smooth complete algebraic curve of genus g over an algebraically closed field k of characteristic 0. Let $U \subset C$ be a nonempty open subset and $m = \#(C \setminus U)$. If $X, Y \in \mathcal{O}_C(U)$ and $u \in \mathcal{O}_C(U)^*$ satisfy*

$$X^n + Y^n - 1 = u$$

for some $n \geq 5$, then

$$h(X, Y, 1) \leq \frac{3(m + \max\{2g-2, 0\})}{n-4} \tag{1}$$

or $(X^n + Y^n)(X^n - 1)(Y^n - 1) = 0$.

It is this version of the theorem that we will prove. After some background material is introduced in the next section, Theorem 1.2 will be proved in Section 3.

2. Preliminaries

The main theorems we will need in order to prove (1) are Mason's theorem and its generalization by Masser and Brownawell. Before introducing these theorems, however, we will need to define some terms.

For each point p on a complete algebraic curve C there exists a discrete valuation $v_p : \mathcal{O}_{C,p} \to \mathbb{Z} \cup \{\infty\}$, which maps a function that is regular at p to its order of vanishing at p. Such valuations naturally extend to the function field K of C.

We will use these valuations to define heights which will generalize the notion of degree to a set of rational functions on an algebraic curve.

Definition 2.1. The *height* of any finite collection $u_1, u_2, \dots, u_n \in K$, not all identically 0, is defined by

$$h(u_1, u_2, \dots, u_n) = -\sum_{p \in C} \min_{1 \le j \le n} v_p(u_j)$$

and has the following properties:

(i) For any nonzero $\alpha \in K$,

$$h(\alpha u_1, \alpha u_2, \dots, \alpha u_n) = h(u_1, u_2, \dots, u_n).$$

(ii) For any j,

$$h(u_1, \dots, u_j, \dots, u_n) \ge h(u_1, \dots, u_{j-1}, u_{j+1}, \dots, u_n).$$

(iii) For any positive integer q,

$$h(u_1^q, u_2^q, \dots, u_n^q) = q h(u_1, u_2, \dots, u_n).$$

Remark 2.2. Without note we will use the fact that replacing any function in the collection with its negative does not change the height. Likewise any permutation of the collection of functions will not change their height.

When the u_i are polynomials without any common zero, their height is simply the maximum of their degrees. In this sense height generalizes the notion of degree.

Definition 2.3. For $\{u_1, u_2, \dots, u_n\} \subseteq K$ we define the *support* to be

$$\mathrm{Supp}\{u_1, u_2, \dots, u_n\} = \{p \in C : v_p(u_i) \ne 0 \text{ for some } 1 \le i \le n\},$$

that is, the set of points where at least one u_i has a zero or pole.

Mason's theorem and its generalizations give an inequality between the height of a set of linearly dependent rational functions and their support. The particular version that we will need is this:

Theorem 2.4 [Brownawell and Masser 1986, Theorem B]. *Let $u_1, u_2, \dots, u_n \in K$ be such that $u_1 + u_2 + \cdots + u_n = 0$ with no nonempty proper subset of the u_i adding to 0, and define $\gamma_s = \frac{1}{2}(s-1)(s-2)$ for $s \ge 1$ and 0 otherwise. Then*

$$h(u_1, u_2, \dots, u_n) \le \gamma_n \max\{2g-2, 0\} + \sum_{p \in C} (\gamma_n - \gamma_{r(p)}),$$

where $r(p)$ is the number of u_i not supported at p.

The specialization of this result to the case of three rational functions will be convenient to have:

Theorem 2.5 (Mason's theorem). *Let $n = 3$ and assume the hypothesis and notation of Theorem 2.4. Then*

$$h(u_1, u_2, u_3) \leq \max\{2g - 2, 0\} + \#\operatorname{Supp}\{u_1, u_2, u_3\}.$$

3. Main results

Throughout this section we will assume that C is a smooth complete curve of genus g over an algebraically closed field of characteristic 0, $U \subset C$ is a nonempty open subset, $m = \#(C \setminus U)$, and $X, Y \in \mathcal{O}_C(U)$ are regular functions on U.

As noted in Section 1 it suffices to study the solutions to the equation

$$X^n + Y^n - 1 = u,$$

where $u \in \mathcal{O}_C(U)^*$. By Theorem 2.4 we can bound the height, $h(X^n, Y^n, -1, -u)$, in terms of an expression involving the number of points in the support of each of these functions. For convenience define

$$\begin{aligned} S_1 &= \operatorname{Supp}\{X\} \cap \operatorname{Supp}\{Y\} \cap U, \\ S_2 &= (\operatorname{Supp}\{X, Y\} \setminus S_1) \cap U. \end{aligned} \tag{2}$$

Then we have the following cases:

(i) If $p \in S_1$, precisely 2 of the functions are not supported at p, so $r(p) = 2$. Hence $\gamma_4 - \gamma_{r(p)} = \gamma_4 - \gamma_2 = 3 - 0 = 3$.

(ii) If $p \in S_2$, precisely 3 of the functions are not supported at p so $\gamma_4 - \gamma_{r(p)} = 2$.

(iii) If $p \in C \setminus U$ we have $\gamma_4 - \gamma_{r(p)} \leq 3$, since $\gamma_{r(p)} \geq 0$ by definition for any p.

(iv) For all remaining points, $p \notin \operatorname{Supp}\{X, Y, u\}$, so $\gamma_4 - \gamma_{r(p)} = 0$.

Thus, provided no nonempty proper subset of $\{X^n, Y^n, -1, -u\}$ adds to 0, we have

$$h(X^n, Y^n, -1, -u) \leq 3\max\{2g - 2, 0\} + 3\#S_1 + 2\#S_2 + 3\#(C \setminus U). \tag{3}$$

By definition $\#(C \setminus U) = m$, which is fixed by the choice of U. In order to bound the height it suffices to establish bounds on $\#S_1$ and $\#S_2$. Rather than directly bounding the size of these sets we will instead bound the quantity $2\#S_1 + \#S_2$. In particular we will show that $2\#S_1 + \#S_2 \leq 2h(X, Y, 1)$. It is necessary to first establish a theorem on the addition of heights. We begin with a fact about minimums.

Lemma 3.1. *For any real numbers $x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m$,*

$$\min_{1 \leq i \leq n} \{x_i\} + \min_{1 \leq j \leq m} \{y_j\} = \min_{\substack{1 \leq i \leq n \\ 1 \leq j \leq m}} \{x_i + y_j\}.$$

Proof. Let x be the minimum of the x_is, y be the minimum of the y_js, and $x_p + y_q$ be the minimum of the $(x_i + y_j)$s. Then $x \le x_p$ and $y \le y_q$, so $x + y \le x_p + y_q$. On the other hand, $x + y \in \{x_i + y_j\}$. Therefore $x + y \ge x_p + y_q$, and equality holds. □

We are now able to come up with an alternate interpretation for the addition of heights.

Corollary 3.2. *Let $f_1, \dots, f_n$ and $g_1, \dots, g_m$ be rational functions on an algebraic curve C where at least one f_i and one g_j are not identically* 0. *Then*

$$h(f_1, f_2, \dots, f_n) + h(g_1, g_2, \dots, g_m) = -\sum_{p \in C} \min_{\substack{1 \le i \le n \\ 1 \le j \le m}} \{v_p(f_i) + v_p(g_j)\}.$$

Proof. Immediate by Lemma 3.1 and the definition of the height function. □

The utility of Corollary 3.2 comes from interpreting it as a type of distributive property. We can see this by noting $v_p(f_i) + v_p(g_j) = v_p(f_i g_j)$. The particular case that we are interested in is

$$h(X, Y, 1) + h(X, Y, 1) = h(X^2, XY, X, Y^2, Y, 1). \tag{4}$$

We can now proceed to bound the quantity $2\#S_1 + \#S_2$.

Proposition 3.3. *Suppose that neither X nor Y is identically* 0. *Then*

$$2\#S_1 + \#S_2 \le 2h(X, Y, 1).$$

Proof. As a result of Corollary 3.2 and Equation (4),

$$2h(X, Y, 1) = h(X^2, XY, X, Y^2, Y, 1).$$

By property (i) of heights (Definition 2.1) we can multiply through by $(XY)^{-1}$, giving

$$2h(X, Y, 1) = h\Big(\frac{X}{Y}, 1, \frac{1}{Y}, \frac{Y}{X}, \frac{1}{X}, \frac{1}{XY}\Big).$$

Now by the definition of height, this is equal to

$$-\sum_{p \in C} \min\Big\{v_p\Big(\frac{X}{Y}\Big), v_p\Big(\frac{1}{Y}\Big), v_p\Big(\frac{Y}{X}\Big), v_p\Big(\frac{1}{X}\Big), v_p\Big(\frac{1}{XY}\Big), 0\Big\}.$$

After distributing the negative, we get

$$\sum_{p \in C} \max\Big\{v_p\Big(\frac{Y}{X}\Big), v_p(Y), v_p\Big(\frac{X}{Y}\Big), v_p(X), v_p(XY), 0\Big\}.$$

For each $p \in S_1$ we have $v_p(XY) \ge 2$ and for each $p \in S_2$ either $v_p(X) \ge 1$ or $v_p(Y) \ge 1$. Since every term of the sum is nonnegative and $S_1 \cap S_2$ is empty, it follows that $2\#S_1 + \#S_2 \le 2h(X, Y, 1)$. □

The previous proposition required the assumption that neither X nor Y was identically 0. This is necessary as $\#\operatorname{Supp}\{0\} = \infty$. However, Theorem 1.2 still holds if either X or Y is 0. In fact a stronger bound holds.

Proposition 3.4. *Suppose that $X^n + Y^n - 1$ is a unit on U for some $n \geq 3$ and that at least one of X or Y is 0; then*

$$h(X, Y, 1) \leq \frac{m + \max\{2g-2, 0\}}{n-2}.$$

Proof. If both X and Y are 0 the inequality trivially holds. Without a loss of generality we may assume $X \neq 0$ and $Y = 0$, in which case X satisfies the equation $X^n - 1 - u = 0$ for some unit $u \in \mathcal{O}_C(U)^*$. Applying Theorem 2.5 we get the inequality

$$h(X^n, -1, -u) \leq \max\{2g-2, 0\} + \#\operatorname{Supp}\{X, 1, u\}.$$

By an argument similar to the proof of Proposition 3.3, $\#(\operatorname{Supp}\{X\} \cap U) \leq 2h(X, 1)$. Therefore we have

$$h(X^n, -1, -u) \leq \max\{2g-2, 0\} + m + 2h(X, 1).$$

By properties (ii) and (iii) of heights (Definition 2.1),

$$(n-2)h(X, -1) \leq \max\{2g-2, 0\} + m.$$

Since $h(X, 0, -1) = h(X, -1)$ and $Y = 0$,

$$h(X, Y, 1) \leq \frac{m + \max\{2g-2, 0\}}{n-2},$$

as claimed. □

Now that we have bounds established we are able to give a proof of Theorem 1.2, from which Theorem 1.1 immediately follows.

Proof of Theorem 1.2. We only need to demonstrate the case where neither X nor Y is 0, since otherwise Proposition 3.4 gives a stronger inequality. If some nonempty proper subset of $\{X^n, Y^n, -1, -u\}$ adds to 0 then $(X^n + Y^n)(X^n - 1)(Y^n - 1) = 0$. Therefore we suppose that no nonempty proper subset adds to 0 and apply Theorem 2.4 to get (3):

$$h(X^n, Y^n, -1, -u) \leq 3\max\{2g-2, 0\} + 3\#S_1 + 2\#S_2 + 3\#(C \setminus U).$$

Next we simplify this inequality by applying properties (ii) and (iii) of heights:

$$nh(X, Y, 1) \leq 3\max\{2g-2, 0\} + 3\#S_1 + 2\#S_2 + 3\#(C \setminus U).$$

Since $3\#S_1 + 2\#S_2 \leq 2(2\#S_1 + \#S_2)$ we can apply Proposition 3.3 to get

$$nh(X, Y, 1) \leq 3\max\{2g-2, 0\} + 4h(X, Y, 1) + 3m.$$

Provided $n \geq 5$ we can solve for $h(X, Y, 1)$ and get

$$h(X, Y, 1) \leq \frac{3(m + \max\{2g-2, 0\})}{n-4}. \qquad \square$$

4. Discussion

In Theorem 1.1 we were able to get a height bound provided $n \geq 5$ and that the image of the curve is not contained within a certain set. In this section we will give a geometric interpretation of this exceptional set as well as a proof that no bound can exist when $n = 2$.

Recall that a flex of an algebraic curve is a simple point where the tangent line intersects with multiplicity three or higher. Flexes can be computed by finding the zeroes of the Hessian of the defining function in the projective plane (see [Kunz 2005, Theorem 9.7] for details). In the case of Fermat curves the Hessian is

$$\begin{vmatrix} n(n-1)x^{n-2} & 0 & 0 \\ 0 & n(n-1)y^{n-2} & 0 \\ 0 & 0 & -n(n-1)z^{n-2} \end{vmatrix},$$

which is equal to $-n^3(n-1)^3x^{n-2}y^{n-2}z^{n-2}$. Therefore all of the flexes lie along the lines $x = 0$, $y = 0$, and $z = 0$. Substituting each of these into the Fermat equation gives us $y^n - z^n = 0$, $x^n - z^n = 0$, and $x^n + y^n = 0$ respectively. Each of these equations in turn determines n flexes on the Fermat curve for a total of $3n$ flexes. Additionally each of these three equations defines the union of n lines in projective space with each line being tangent to a flex. Returning to the statement of Theorem 1.1 we can see that the exceptional set $V((x^n + y^n)(x^n - z^n)(y^n - z^n))$ is just the union of the lines tangent to the $3n$ flexes of the curve.

We can also see that the exclusion of this exceptional set is necessary. For example, let $\zeta \in k$ be such that $\zeta^n = 1$. Then for each positive integer q the morphism

$$\phi_q : \mathbb{P}^1_k \setminus \{0, \infty\} \to \mathbb{P}^2_k \setminus V(z(x^n + y^n - z^n))$$

given by $[u : v] \mapsto [u^q : \zeta v^q : v^q]$ is well-defined and has its image contained within the zero set of $y^n - z^n$. Since q can be any positive integer there cannot be a bound on the height. A similar argument holds for the other components of the exceptional set.

Finally we will show that such height bounds cannot exist if $n = 2$. Let $C = \mathbb{P}^1_k$ and $U = C \setminus \{\infty, 0\}$ with coordinate ring $\mathcal{O}_C(U) = k[t, t^{-1}]$. Consider the

diophantine equation

$$X^2 + Y^2 - 1 = t^q,$$

where $X, Y \in k[t, t^{-1}]$ and q is any odd positive integer. We can rewrite this as $(X + iY)(X - iY) = 1 + t^q$. We then set

$$X + iY = 1 + t \quad \text{and} \quad X - iY = \sum_{j=0}^{q-1} (-1)^j t^j.$$

Solving for X and Y gives

$$X = \frac{1}{2}\left(1 + t + \sum_{j=0}^{q-1} (-1)^j t^j\right) \quad \text{and} \quad Y = -\frac{i}{2}\left(1 + t - \sum_{j=0}^{q-1} (-1)^j t^j\right).$$

Since q can be arbitrarily large, $h(X, Y, 1)$ is unbounded. Moreover the family of rational curves defined by (X, Y) as q varies is not contained in any proper closed subset of $\mathbb{A}^2_k$. Therefore no similar result can hold for $n = 2$.

Acknowledgements

We wish to acknowledge the support of the University of Wisconsin-Stout. We also thank Christopher Bendel for comments on an earlier version of this manuscript and the referee for pointing out a simplification to the proof of the main theorem. This research was funded in part by NSF grant 1062403.

References

[Brownawell and Masser 1986] W. D. Brownawell and D. W. Masser, "Vanishing sums in function fields", *Math. Proc. Cambridge Philos. Soc.* **100**:3 (1986), 427–434. MR 87k:11080 Zbl 0612.10010

[Corvaja and Zannier 2008] P. Corvaja and U. Zannier, "Some cases of Vojta's conjecture on integral points over function fields", *J. Algebraic Geom.* **17**:2 (2008), 295–333. MR 2008m:11124 Zbl 1221.11146

[Kunz 2005] E. Kunz, *Introduction to plane algebraic curves*, Birkhäuser, Boston, MA, 2005. MR 2006b:14001 Zbl 1078.14041

[Vojta 1987] P. Vojta, *Diophantine approximations and value distribution theory*, Lecture Notes in Mathematics **1239**, Springer, Berlin, 1987. MR 91k:11049 Zbl 0609.14011

Received: 2012-08-31 Revised: 2013-01-20 Accepted: 2013-01-21

dutters@uwstout.edu *Department of Mathematics, Statistics and Computer Science, University of Wisconsin-Stout, Menomonie, WI 54751, United States*

Melissa.Haire@gordon.edu *Department of Mathematics and Computer Science, Gordon College, Wenham, MA 01984, United States*

asetniker09@wou.edu *Mathematics Department, Western Oregon University, Monmouth, OR 97361, United States*

INVOLVE 7:5 (2014)
dx.doi.org/10.2140/involve.2014.7.619

Quadratic forms representing all primes

Justin DeBenedetto

(Communicated by Kenneth S. Berenhaut)

Building on the method used by Bhargava to prove "the fifteen theorem", we show that every integer-valued positive definite quadratic form which represents all prime numbers must also represent 205. We further this result by proving that 205 is the smallest nontrivial composite number which must be represented by all such quadratic forms.

1. Introduction and statement of results

The study of quadratic forms in various fashions dates back to the third century works of Diophantus. Diophantus worked with ways to rewrite sums of squares and found that $(a^2+b^2)(c^2+d^2)=(ac\pm bd)^2+(ad\mp bc)^2$ and that numbers of the form $4n-1$ are not able to be represented as a sum of two squares. It was not until 1625 that Albert Girard (Fermat came to the same result a few years later) wrote that a number is the sum of two squares if and only if when divided by its largest square factor, the result is a product of primes of the form $4n+1$ or twice the product of such primes [Dickson 1920].

Further exploration into representation of numbers by squares led to Lagrange proving in 1770 that every natural number is the sum of four integer squares, $n=a^2+b^2+c^2+d^2$. Ramanujan [1917] furthered this result by conjecturing[1] that there are exactly 55 sets of values for a, b, c, d such that $ax^2+by^2+cz^2+du^2$ represents all positive integers.

Willerding [1948] used an extension of Ramanujan's work to prove the following:

Theorem 1. *There are exactly* 178 *classes of universal positive definite integer matrix quaternary quadratic forms.*

Here positive definite indicates that the quadratic form represents only non-negative integers, and only represents 0 when every variable is equal to 0. A universal quadratic form represents every number in its range, thus a universal

MSC2010: primary 11E25; secondary 11E20.
Keywords: quadratic forms, number theory, prime number.

[1] Originally it was stated that there are 55 sets of values, but one was later removed when Dickson proved that exactly one of Ramanujan's forms failed to represent all positive integers.

positive definite quadratic form represents all positive integers. Integer matrix means that the coefficients on all cross terms are even.

Bhargava [2000] showed that there are actually 204 universal quaternary forms and enumerated those forms. In the same paper, Bhargava gave a proof of a theorem stated in 1993 by Conway and Schneeberger known as "the fifteen theorem".

Theorem 2. *If a positive definite quadratic form having an integer matrix represents every positive integer up to* 15 *then it represents every positive integer.*

Building on the methods used by Bhargava, we look specifically at integer-valued positive definite quadratic forms representing all prime numbers. Our goal is to determine if there are composite numbers which are represented by every such quadratic form, and if so to find the smallest such composite. Our result is a proof that there are nontrivial composites which are represented by all quadratic forms representing every prime number.

We restrict the composite numbers we are considering to composites which are not a square times a prime. This is due to the fact that all primes are represented by prime universal quadratic forms, and if n is represented then nx^2 is also represented. For this reason, we consider a square times a prime to be a trivial composite number, and search for nontrivial composite numbers which are represented by these quadratic forms.

Theorem 3. *Every integer-valued positive definite quadratic form*, Q, *representing all prime numbers*, *must represent* 205. *Furthermore*, *if* n *is a composite number less than* 205 *and* n *is represented by all quadratic forms which represent all prime numbers*, *then* n *must be a square times a prime.*

Remark. An analogous statement may be made for integer matrix positive definite quadratic forms. The same process is used, but the calculations are simpler and 66 is the smallest non-trivial composite which must be represented in that case.

2. Definitions

First, we define lattices as they pertain to quadratic forms throughout this paper. The set of all integers is denoted $\mathbb{Z}$. An n-dimensional *lattice* L is a subset $L \subseteq \mathbb{R}^n$, together with an inner product that gives a way of measuring distances and angles. Here are the properties that it must satisfy:

(i) The set L must span $\mathbb{R}^n$.

(ii) The set L must have the form $L = \left\{\sum_{i=1}^{n} a_i \vec{v}_i : a_i \in \mathbb{Z}\right\}$.

(iii) The inner product $\langle \vec{v}, \vec{w} \rangle$ is a function from $L \times L \to \mathbb{R}$.

(iv) For any $\vec{v} \in L$, $\langle \vec{v}, \vec{v} \rangle \geq 0$ with $\langle \vec{v}, \vec{v} \rangle = 0$ if and only if $\vec{v} = 0$.

(v) For any $\vec{v}, \vec{w}, \vec{x} \in L$, we have $\langle \vec{v}, \vec{w} \rangle = \langle \vec{w}, \vec{v} \rangle$ and $\langle \vec{v} + \vec{w}, \vec{x} \rangle = \langle \vec{v}, \vec{x} \rangle + \langle \vec{w}, \vec{x} \rangle$.

Note that property (iv) enforces positive definiteness.

Given a lattice L, the function

$$Q(a_1, \ldots, a_n) = \left\langle \sum_{i=1}^{n} a_i \vec{v}_i, \sum_{i=1}^{n} a_i \vec{v}_i \right\rangle$$

is a *quadratic form*. Moreover, every quadratic form arises in this way.

A quadratic form is called *integer-valued* if

$$Q(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \sum_{j=i}^{n} a_{ij} x_i x_j, \quad a_{ij} \in \mathbb{Z}.$$

Furthermore, an integer-valued quadratic form is called *integer matrix* if a_{ij} is even for all $i \neq j$.

The *Gram matrix* of a quadratic form, Q, is the matrix A such that we can write $Q = x\,{}^tAx$. The Gram matrix of an integer-valued quadratic form has entries has integer diagonal entries and half integer off diagonal entries. Similarly, the Gram matrix of an integer matrix quadratic form has integer entries.

Next, we set forth some definitions which are based upon the definitions used in [Bhargava 2000]. We first define the *prime truant* of a quadratic form to be the smallest prime not represented by the quadratic form. We also define a *prime escalation* of a lattice to be a lattice generated by the original lattice and a vector with norm equal to the prime truant of the original lattice. The dimension of a prime escalation is either equal to the dimension of the original lattice or greater by 1. A *prime escalator lattice* is a lattice which is generated by any number of prime escalations of the zero-dimensional lattice. Similarly, a quadratic form is considered to be *prime universal* if it represents all prime numbers.

Two quadratic forms, Q_1 and Q_2, are considered *equivalent* if there is an integral invertible change of variables which sends Q_1 to Q_2.

If Q is a positive definite quadratic form, let $r_Q(n)$ be the number of representations of n by Q. The *theta series* of Q is the power series

$$\Theta_Q(q) = \sum_{n=0}^{\infty} r_Q(n) q^n.$$

3. Prime escalations

We begin by giving an overview of prime escalations. Escalating the zero-dimensional lattice gives us $\begin{bmatrix} 2 \end{bmatrix}$ which leads to the form $2x^2$. This represents 2 but not 3, so our two-dimensional prime escalator lattices are

$$\begin{bmatrix} 2 & x \\ x & 3 \end{bmatrix},$$

with $x^2 \leq 6$ by the Cauchy–Schwarz inequality. Since we are looking for integer-valued quadratic forms, we allow x to be of the form $x = y/2$, $y \in \mathbb{Z}$. Thus we have the lattices with matrices

$$\begin{bmatrix} 2 & \pm\frac{1}{2} \\ \pm\frac{1}{2} & 3 \end{bmatrix}, \quad \begin{bmatrix} 2 & \pm 1 \\ \pm 1 & 3 \end{bmatrix}, \quad \begin{bmatrix} 2 & \pm\frac{3}{2} \\ \pm\frac{3}{2} & 3 \end{bmatrix}, \quad \begin{bmatrix} 2 & \pm 2 \\ \pm 2 & 3 \end{bmatrix}, \quad \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix},$$

corresponding to the quadratic forms

$$2x^2 \pm xy + 3y^2, \quad 2x^2 \pm 2xy + 3y^2, \quad 2x^2 \pm 3xy + 3y^2$$
$$2x^2 \pm 4xy + 3y^2, \quad 2x^2 + 3y^2.$$

The form $2x^2 + 3xy + 3y^2$ is equivalent to $2x^2 + xy + 2y^2$, and $2x^2 + 4xy + 3y^2$ is equivalent to $x^2 + 2y^2$. For binary quadratic forms, we may ignore the sign of the cross term since sending x to $-x$ makes the forms equivalent.

To automate this process we ran this through a math program called Magma. The Magma script began with the zero-dimensional lattice and followed the prime escalation process as above. When checking for prime truants, each form was checked for unrepresented primes beginning with 2 and checking each prime until one was unrepresented. Since prime escalator lattices of dimension two are not prime universal, every lattice was escalated at least three times. Some of these lattices were escalated a fourth time if they failed to represent a prime below 1000.

We define two sets, A and B, as follows. If a third escalation prime escalator lattice, L, represents every prime below 1000, then let $L \in A$. If the third escalation prime escalator lattice fails to represent a prime below 1000, then let a prime escalation of L be M, and let $M \in B$.

Theorem 4. *Every prime universal lattice contains a prime escalator lattice from A or B.*

Proof. Suppose L is a prime universal lattice and $L_0 \subseteq L_1 \subseteq L_2 \subseteq L_3 \subseteq \cdots \subseteq L$ is an ascending chain of escalator lattices, with L_0 zero-dimensional and L_i being a prime escalation of L_{i-1} for each i. Since every prime escalator lattice of dimension 0, 1, or 2 increases in dimension when escalated, L_3 is a 3-dimensional escalator lattice. If L_3 has a prime truant below 1000, then its escalation $L_4 \in B$. If not, then $L_3 \in A$. Thus L contains a prime escalator sublattice from A or from B. □

Forms in the set A may not be prime universal, but Theorem 5 allows us to make use of these forms.

4. Composite representation methods

Since we are searching for composite numbers which are represented by every prime universal quadratic form, we take advantage of Lagrange's four-square theorem mentioned above to form the following theorem.

Theorem 5. *Suppose $Q(\vec{y})$ is a positive definite integer-valued quadratic form and there exists a composite number m such that $Q(\vec{y})$ does not represent m. If $Q(\vec{y})$ represents every prime $p < m$, then the quadratic form*

$$R(\vec{y}, a, b, c, d, x_0, \ldots, x_{m-1})Q(\vec{y})+(m+1)(a^2+b^2+c^2+d^2)+\sum_{i=0}^{m-1}(m+1+i)x_i^2$$

is prime universal and does not represent m.

Proof. This result is due to the fact that $\sum_{i=0}^{m-1}(m+1+i)x_i^2$ represents every number between $m+1$ and $2m$, and $(m+1)(a^2+b^2+c^2+d^2)$ represents every multiple of $(m+1)$. Thus together these represent every number greater than m. The resultant quadratic form, R, does not represent m due to the fact that Q and each of the two added components are all positive definite. Since the two added components do not represent any numbers less than m and Q does not represent any negative numbers, there is no way to represent m. Finally, since we have shown that our new quadratic form represents every number greater than m and Q represents every prime less than m, R must represent all prime numbers and thus is prime universal. □

With these tools in hand we are now ready to handle the proof of our main result.

Proof of Theorem 3. In this way, if we are able to find a quadratic form which represents all primes less than a composite, but does not represent that composite number, we can construct a prime universal quadratic form which fails to represent that composite. Table 1 provides a list of quadratic forms that show that every composite number less than 205 which is not a square times a prime does not have to be represented by a prime universal quadratic form.

Next, we look at which composites are represented by prime universal quadratic forms. After each prime escalation run using Magma, we checked which composites were represented by every prime escalator lattice. In order to do this, we generated the theta series of each lattice and checked the coefficients of the composite power terms. We began this process after two prime escalations, since the zero-dimensional and one-dimensional prime escalator lattices do not represent any composites which are not a square times a prime. By comparing the represented composites for the five two-dimensional prime escalator lattices corresponding to binary quadratic forms, we find that 818 is represented by all of them. As such, we reduced our theta series to only look at which composites below 818 were represented for subsequent prime escalations. We repeated this process for the third and fourth prime escalations of the zero-dimensional lattice.

This process showed that 818 is represented by all second prime escalations, 453 is represented by all third escalations, and 205 is represented by every lattice in sets A and B.

Quadratic form	Integers not represented
$x^2+xy+2y^2+3z^2$	6, 15, 24, 33, 42, 51, 54, 60, 69, 78, 85, 87, 96, 105, 114, 123, 132, 135, 141, 150, 159, 168, 177, 186, 195, 204
$2x^2-2xz+3y^2+yz+3z^2$	1, 4, 16, 22, 38, 64, 70, 86, 88, 102, 118, 134, 152, 166, 182, 198
$2x^2-2xz+3y^2+3yz+3z^2$	26, 104
$2x^2-xz+3y^2+4z^2$	9, 81
$2x^2-xz+3y^2+3yz+4z^2$	34, 111
$2x^2+3y^2+yz+5z^2$	10, 40, 58, 74, 90, 106, 122, 136, 138, 154, 160, 170, 202
$2x^2+3y^2+3yz+5z^2$	36, 119, 144, 187
$2x^2-2xz+3y^2-yz+5z^2$	46, 178, 184
$x^2+xz+y^2+2z^2$	21, 35, 84, 91, 133, 140, 189
$2x^2+xy-xz+3y^2-yz+3z^2$	25, 30, 65, 110, 115, 155, 165, 185, 190
$x^2+xz+y^2+3z^2$	66, 77, 143
$2x^2+2xy-2xz+3y^2+yz+5z^2$	14, 56, 62, 94, 120, 126, 142, 158, 174
$2x^2+2xy+3y^2+yz+5z^2$	39, 156
$x^2+xy+xz+2y^2+3z^2+19w^2$	57
$2x^2-2xz-xw+3y^2-3yw+3z^2-2zw+14w^2$	49
$2x^2-2xz+3y^2-3yw+3z^2+15w^2$	121
$2x^2-2xz+3y^2+2yz+3z^2+13w^2$	169
$2x^2-xz+3y^2-2yz+3z^2+61w^2$	183
$2x^2-xy+2y^2-yz+3z^2+43w^2$	129
$2x^2+xz-xw+3y^2+3yz-3yw+6z^2-4zw+7w^2$	55
$2x^2+xz+2xw+3y^2+3yz+6z^2+2zw+9w^2$	95
$2x^2-2xz+2xw+3y^2-3yw+7z^2-5zw+11^2$	130
$2x^2-2xz-2xw+3y^2+2yz-3yw+7z^2-zw+9w^2$	82
$2x^2-xz+3y^2-yz+5z^2+23z^2$	161
$2x^2+2xy-xz-xw+3y^2-yw+5z^2+66w^2$	194
$2x^2+2xy-xz-xw+3y^2+2yz-3yw+5z^2+zw+26w^2$	93
$2x^2-xy+xw+2y^2-yz-yw+5z^2+70w^2$	146
$2x^2-xy+xz+2y^2-2yz+5z^2+67w^2$	201
$2x^2-xy-xw+2y^2+2yw+7z^2+8w^2$	196
$2x^2-xy+xz+xw+2y^2+yz+yw+7z^2+4zw+11z^2$	100
$x^2-xz+2y^2-yz+4z^2+29w^2$	145, 203

Table 1. The quadratic forms on the left do not represent the numbers on the right, but represent every prime less than each of those numbers. By Theorem 5, there exists a prime universal quadratic form which represents all primes and does not represent each number listed.

By Theorem 4, every prime universal lattice will contain a prime escalator lattice from sets A or B and thus will represent 205. □

5. Conclusions

There are many questions that remain regarding properties of universal quadratic forms, and specifically prime universal quadratic forms.

Question. *If we let S be a set of positive integers, T be the set of all positive definite quadratic forms that represent every number in S, and U be the set of numbers represented by everything in T, when does $T = U$ and when is U bigger than T?*

We have answered this question in the case of S being the set of all primes, and found that U is bigger than T since $205 \in U$.

Other questions regarding quadratic forms which have been answered for integers remain open when applied specifically to prime numbers. A similar examination could apply to the "290 theorem" (see [Bhargava and Hanke 2011]).

Theorem 6 (290 theorem). *If a positive definite quadratic form with integer coefficients represents the twenty-nine integers* 1, 2, 3, 5, 6, 7, 10, 13, 14, 15, 17, 19, 21, 22, 23, 26, 29, 30, 31, 34, 35, 37, 42, 58, 93, 110, 145, 203, *and* 290, *then it represents all positive integers.*

Question. *What is the smallest set of prime numbers such that all positive definite integer-valued quadratic forms which represent every prime in the set must be prime universal*? (Bhargava has answered this question in the case of positive definite integer matrices.)

Acknowledgements

We used Magma [Bosma et al. 1997] for our escalation computations. I would like to thank Jeremy Rouse for all of his guidance on this project. I would also like to thank the anonymous referee for editing and providing feedback.

References

[Bhargava 2000] M. Bhargava, "On the Conway–Schneeberger fifteen theorem", pp. 27–37 in *Quadratic forms and their applications* (Dublin, 1999), edited by E. Bayer-Fluckiger et al., Contemp. Math. **272**, Amer. Math. Soc., Providence, RI, 2000. MR 2001m:11050

[Bhargava and Hanke 2011] M. Bhargava and J. Hanke, "Universal quadratic forms and the 290-theorem", preprint, 2011; see http://wordpress.jonhanke.com/wp-content/uploads/2011/09/290-Theorem-preprint.pdf.

[Bosma et al. 1997] W. Bosma, J. Cannon, and C. Playoust, "The Magma algebra system, I: The user language", *J. Symbolic Comput.* **24**:3-4 (1997), 235–265. MR 1484478

[Dickson 1920] L. E. Dickson, *History of the theory of numbers, II: Diophantine numbers*, Carnegie Institution of Washington, 1920.

[Ramanujan 1917] S. Ramanujan, "On the expression of a number in the form $ax^2+by^2+cz^2+du^2$", *Proc. Cambridge Phil. Soc.* **19** (1917), 11–21. Reprinted as pp. 169–177 in *Collected Papers of Srinivasa Ramanujan*, edited by G. H. Hardy et al., Cambridge Univ. Press; reissued Chelsea, New York, 1962. JFM 46.0240.01

[Willerding 1948] M. F. Willerding, "Determination of all classes of positive quaternary quadratic forms which represent all (positive) integers", *Bull. Amer. Math. Soc.* **54** (1948), 334–337. MR 9,571e

Received: 2013-05-03 Revised: 2013-10-01 Accepted: 2013-12-22

debejd0@wfu.edu *Department of Mathematics, Wake Forest University, 127 Manchester Hall, Box 7388, Winston-Salem, NC 27109, United States*

INVOLVE 7:5 (2014)
dx.doi.org/10.2140/involve.2014.7.627

Counting matrices over a finite field with all eigenvalues in the field

Lisa Kaylor and David Offner

(Communicated by Kenneth S. Berenhaut)

Given a finite field $\mathbb{F}$ and a positive integer n, we give a procedure to count the $n \times n$ matrices with entries in $\mathbb{F}$ with all eigenvalues in the field. We give an exact value for any field for values of n up to 4, and prove that for fixed n, as the size of the field increases, the proportion of matrices with all eigenvalues in the field approaches $1/n!$. As a corollary, we show that for large fields almost all matrices with all eigenvalues in the field have all eigenvalues distinct. The proofs of these results rely on the fact that any matrix with all eigenvalues in $\mathbb{F}$ is similar to a matrix in Jordan canonical form, and so we proceed by enumerating the number of $n \times n$ Jordan forms, and counting how many matrices are similar to each one. A key step in the calculation is to characterize the matrices that commute with a given Jordan form and count how many of them are invertible.

1. Introduction

Let $\mathbb{F}$ be a field and let $M_n(\mathbb{F})$ denote the set of $n \times n$ matrices with entries in $\mathbb{F}$. As an example, consider $A = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right) \in M_2(\mathbb{R})$. The roots of its characteristic polynomial $\det(A - \lambda I)$ are $\lambda = \pm i$, which are the eigenvalues of A. Though the entries in A are real numbers, the eigenvalues are not. This example serves to motivate the following question: If $\mathrm{Eig}_n(\mathbb{F})$ denotes the set of elements of $M_n(\mathbb{F})$ that have all of their eigenvalues in $\mathbb{F}$, what is the cardinality of $\mathrm{Eig}_n(\mathbb{F})$? For a field like $\mathbb{R}$ that is uncountably infinite, this question is trivial, but in this paper we examine the case when $\mathbb{F}$ is a finite field with q elements. This line of research was initiated by Olšavský [2003], who determined that for any prime p,

$$|\mathrm{Eig}_2(\mathbb{Z}_p)| = \tfrac{1}{2}p^4 + p^3 - \tfrac{1}{2}p^2. \tag{1}$$

Here we present a method for determining $|\mathrm{Eig}_n(\mathbb{F})|$ for any n. We use the fact that any matrix $A \in M_n(\mathbb{F})$ with all eigenvalues in $\mathbb{F}$ is similar to a matrix J in Jordan canonical form. Thus we can determine $|\mathrm{Eig}_n(\mathbb{F})|$ using the following procedure:

MSC2010: 05A05, 15A18, 15B33.
Keywords: eigenvalues, matrices, finite fields, Jordan form.

(1) Enumerate all $n \times n$ Jordan forms.

(2) Enumerate all matrices in $M_n(\mathbb{F})$ that are similar to each Jordan form.

In Section 2, we review the definitions and notation necessary to work with Jordan forms. Then in Section 3 we explain the procedure to determine $|\mathrm{Eig}_n(\mathbb{F})|$, giving a general formula in (3). We illustrate the process for the case $n = 2$, giving a slightly shorter derivation of (1) than was given in [Olšavský 2003].

We group matrices in Jordan form by what we call their double partition type, which is defined in Section 3. In Sections 4 and 5 we find formulas for the quantities required to compute $|\mathrm{Eig}_n(\mathbb{F})|$ for any n. In Section 4 we state a formula for the number of Jordan forms of a given double partition type. Then in Section 5 we prove that the number of matrices similar to any matrix in Jordan form depends only on its double partition type, and give a formula that determines this number for any double partition type. In Section 6, we use these results to give explicit formulas for $|\mathrm{Eig}_3(\mathbb{F})|$ and $|\mathrm{Eig}_4(\mathbb{F})|$ for any finite field $\mathbb{F}$.

Olšavský [2003] also noted that the proportion of matrices in $M_2(\mathbb{Z}_p)$ with all eigenvalues in $\mathbb{Z}_p$ approaches 1/2 as p goes to infinity. In Section 7, we generalize this result to prove that the proportion of matrices in $M_n(\mathbb{F})$ with all eigenvalues in $\mathbb{F}$ approaches $1/n!$ as q approaches infinity. As a corollary, we prove that for large finite fields, if a matrix has all eigenvalues in the field, then almost surely all if its eigenvalues are distinct.

2. Jordan canonical form

Denote the set of invertible matrices in $M_n(\mathbb{F})$ by $\mathrm{GL}_n(\mathbb{F})$. We will repeatedly use the fact that any element of $M_n(\mathbb{F})$ with all eigenvalues in $\mathbb{F}$ is similar to a matrix in Jordan canonical form. Here we review the necessary definitions. For a more thorough introduction, see [Hungerford 1974, Chapter 7.4].

For $1 \le i \le n$, let A_i be a square matrix. The *direct sum* of these matrices, denoted $\bigoplus_{i=1}^n A_i$, is a block diagonal matrix such that the matrices A_i lie on the diagonal, and all other entries are zero:

$$\bigoplus_{i=1}^{n} A_i = \begin{pmatrix} A_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & A_n \end{pmatrix}.$$

A *Jordan block* of size $k \ge 1$ corresponding to some eigenvalue $\lambda \in \mathbb{F}$ is a $k \times k$ matrix with λs along the diagonal, 1s along the superdiagonal, and 0s everywhere else (see Figure 1). Let $A, J \in M_n(\mathbb{F})$. Then A is *similar* to J if there exists a matrix $P \in \mathrm{GL}_n(\mathbb{F})$ such that $A = PJP^{-1}$. It is well known (see, for example, [Hungerford 1974, Chapter 7.4, Corollary 4.7 (iii)]) that any matrix $A \in M_n(\mathbb{F})$ with all eigenvalues in $\mathbb{F}$ is similar to a matrix J which is the direct sum of Jordan blocks,

$$\begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{pmatrix} \qquad \begin{pmatrix} 3 & 1 & 0 & 0 \\ 0 & 3 & 1 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 3 \end{pmatrix}$$

Figure 1. A generic Jordan block and a Jordan block with eigenvalue 3.

$$A = \begin{pmatrix} 1 & 3 & 3 & 4 \\ 3 & 3 & 0 & 2 \\ 0 & 4 & 3 & 1 \\ 3 & 3 & 3 & 2 \end{pmatrix} = PJP^{-1} = \begin{pmatrix} 3 & 2 & 0 & 3 \\ 0 & 2 & 1 & 0 \\ 1 & 0 & 0 & 2 \\ 3 & 2 & 1 & 3 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 1 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} 0 & 1 & 4 & 4 \\ 3 & 3 & 0 & 2 \\ 4 & 0 & 0 & 1 \\ 0 & 2 & 1 & 3 \end{pmatrix},$$

$$A = \begin{pmatrix} 1 & 3 & 3 & 4 \\ 3 & 3 & 0 & 2 \\ 0 & 4 & 3 & 1 \\ 3 & 3 & 3 & 2 \end{pmatrix} = QJ'Q^{-1} = \begin{pmatrix} 3 & 3 & 2 & 0 \\ 0 & 0 & 2 & 1 \\ 1 & 2 & 0 & 0 \\ 3 & 3 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} 0 & 1 & 4 & 4 \\ 0 & 2 & 1 & 3 \\ 3 & 3 & 0 & 2 \\ 4 & 0 & 0 & 1 \end{pmatrix}.$$

Figure 2. The matrix $A \in M_4(\mathbb{Z}_5)$ is similar to J and J', which are different representations of the same Jordan form.

and this matrix J is unique up to the ordering of the blocks. The *Jordan canonical form* (or simply *Jordan form*) for A is this direct sum of Jordan blocks. We will use $J_n(\mathbb{F})$ to denote the set of Jordan forms in $M_n(\mathbb{F})$. We note that multiple matrices may correspond to a given Jordan form (e.g., $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 2 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 2 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ are representatives of the same form), so when we wish to work with a particular matrix in $J_n(\mathbb{F})$, we will specify an order for the blocks in the direct sum. For example, in Figure 2, the matrix $A \in M_4(\mathbb{Z}_5)$ has the Jordan form that is the direct sum of the Jordan blocks (1), (2), and $\left(\begin{smallmatrix} 3 & 1 \\ 0 & 3 \end{smallmatrix}\right)$. In the figure, $J \in J_4(\mathbb{Z}_5)$ and $J' \in J_4(\mathbb{Z}_5)$ are two examples of matrices that are direct sums of these blocks.

3. How to determine $|\mathrm{Eig}_n(\mathbb{F})|$

Since every matrix in $\mathrm{Eig}_n(\mathbb{F})$ is similar to a matrix in Jordan canonical form, we can determine $|\mathrm{Eig}_n(\mathbb{F})|$ by first enumerating all Jordan forms in $J_n(\mathbb{F})$ and then counting how many matrices in $M_n(\mathbb{F})$ are similar to each one. In this section we explain this process and illustrate it with the example of computing $|\mathrm{Eig}_2(\mathbb{F})|$.

For any matrix $A \in M_n(\mathbb{F})$, let $S(A) \subseteq M_n(\mathbb{F})$ be the set of matrices similar to A, and let $C(A)$ denote the subgroup of $\mathrm{GL}_n(\mathbb{F})$ of matrices P such that $PA = AP$.

Lemma 3.1. *For any $J \in M_n(\mathbb{F})$, we have $|S(J)| = |\mathrm{GL}_n(\mathbb{F})|/|C(J)|$.*

Proof. Fix $J \in M_n(\mathbb{F})$; we must find the cardinality of $S(J) = \{AJA^{-1} : A \in \mathrm{GL}_n(\mathbb{F})\}$. For any $A, B \in \mathrm{GL}_n(\mathbb{F})$, $AJA^{-1} = BJB^{-1}$ if and only if $B^{-1}AJ = JB^{-1}A$, that is, if $B^{-1}A \in C(J)$. Now, $B^{-1}A \in C(J)$ if and only if A and B are in the same coset of $C(J)$ in $\mathrm{GL}_n(\mathbb{F})$, and thus $S(J)$ has the same cardinality as the number of cosets of $C(J)$ in $\mathrm{GL}_n(\mathbb{F})$, which is equal to $|\mathrm{GL}_n(\mathbb{F})|/|C(J)|$ by Lagrange's theorem. □

It is well known (see, for example, [Stanley 2012, Proposition 1.10.1]) that if $\mathbb{F}$ has q elements, then $|\mathrm{GL}_n(\mathbb{F})| = \prod_{i=0}^{n-1}(q^n - q^i)$. Thus to find $|S(J)|$ for any J, it suffices to find $|C(J)|$, and we obtain the following formula for $|\mathrm{Eig}_n(\mathbb{F})|$:

$$|\mathrm{Eig}_n(\mathbb{F})| = \sum_{J \in J_n(\mathbb{F})} |S(J)| = \sum_{J \in J_n(\mathbb{F})} \frac{\prod_{i=0}^{n-1}(q^n - q^i)}{|C(J)|}. \tag{2}$$

For $J \in J_n(\mathbb{F})$, it turns out that $|C(J)|$ depends on what we will call the double partition type of J, which we define now.

A *partition* of a positive integer n is a collection of (not necessarily distinct) positive integers $\{n_1, n_2, \dots n_k\}$ such that $n_1 + n_2 + \cdots + n_k = n$. We define the *partition type* of a matrix $J \in J_n(\mathbb{F})$ as the partition of n given by the size of the Jordan blocks in J. For example, the partition type of the matrix $J \in J_4(\mathbb{Z}_5)$ below is $\{1, 1, 2\}$ since it has two 1×1 Jordan blocks and one 2×2 Jordan block.

$$J = \begin{pmatrix} 3 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

We define a *double partition* (or partition of a partition) of a positive integer n to be a partition where the numbers in the partition are themselves partitioned into subsets. Denote the set of double partitions of n by $\mathrm{DP}(n)$. For example, $\{\{2, 3\}, \{1, 1, 3\}, \{2, 3\}, \{4\}\} \in \mathrm{DP}(19)$. We define the *double partition type* (or simply *type*) of a matrix $J \in J_n(\mathbb{F})$ by grouping all elements of its partition type into sets where two elements of the partition are placed in the same set if their corresponding eigenvalues are the same. For example, the double partition type of the matrix J above is $\{\{1, 1\}, \{2\}\}$ since the two 1×1 blocks have the same eigenvalue, and the 2×2 block has a different eigenvalue. The study of double partitions dates back at least as far as [Cayley 1855] and [Sylvester 1851], and the values of $|\mathrm{DP}(n)|$ are collected in the *Online Encyclopedia of Integer Sequences* (oeis.org), sequence A001970.

The utility of knowing the double partition type of a matrix in Jordan form is given by the following lemma.

Lemma 3.2. *If $J_1, J_2 \in J_n(\mathbb{F})$ have the same double partition type, then $|C(J_1)| = |C(J_2)|$.*

The proof of Lemma 3.2 is one of the main results of the paper and will be deferred until Section 5. This lemma justifies the following definition: For any double partition type T define $c(T)$ and $s(T)$ so that $c(T) = |C(J)|$ and $s(T) = |S(J)|$, where J is any matrix of type T. Letting $t(T)$ denote the number of Jordan forms of type T, we can now rewrite (2) as

$$|\mathrm{Eig}_n(\mathbb{F})| = \sum_{J \in J_n(\mathbb{F})} |S(J)| = \sum_{T \in \mathrm{DP}(n)} t(T)s(T) = \sum_{T \in \mathrm{DP}(n)} t(T) \frac{\prod_{i=0}^{n-1} (q^n - q^i)}{c(T)}. \tag{3}$$

We now are prepared to illustrate our procedure for determining $|\mathrm{Eig}_n(\mathbb{F})|$ by computing $|\mathrm{Eig}_2(\mathbb{F})|$. The first step is to enumerate all the double partition types in $\mathrm{DP}(n)$. In the case $n = 2$, there are two possible partitions, $2 = 1 + 1$, and $2 = 2$. There are three double partition types: $T_{2,1} = \{\{1\}, \{1\}\}$, $T_{2,2} = \{\{1, 1\}\}$ and $T_{2,3} = \{\{2\}\}$. We give a general formula for $t(T)$ in Lemma 4.1, but for $n = 2$ we can just examine each type. The Jordan forms of type $T_{2,1}$ are represented by matrices of the form $\left(\begin{smallmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{smallmatrix}\right)$ where $\lambda_1 \neq \lambda_2$ and thus $t(T_{2,1}) = \binom{q}{2}$. Jordan forms of type $T_{2,2}$ are represented by matrices of the form $\left(\begin{smallmatrix} \lambda & 0 \\ 0 & \lambda \end{smallmatrix}\right)$, and Jordan forms of type $T_{2,3}$ are represented by matrices of the form $\left(\begin{smallmatrix} \lambda & 1 \\ 0 & \lambda \end{smallmatrix}\right)$, so $t(T_{2,2}) = t(T_{2,3}) = q$.

It remains to compute $c(T)$ for each double partition type. A general formula for $c(T)$ is given by (4), but again for $n = 2$ we can argue ad hoc. A 2×2 matrix commutes with a matrix J of type $T_{2,1}$ if and only if it is of the form $\left(\begin{smallmatrix} a & 0 \\ 0 & b \end{smallmatrix}\right)$, where $a, b \in \mathbb{F}$, and of course this is true regardless of the specific values of λ_1 and λ_2. Since an element of $C(J)$ must be invertible, there are $q - 1$ choices each for a and b, and $c(T_{2,1}) = (q - 1)^2$. Similarly, if J is of type $T_{2,2}$, $C(J) = \mathrm{GL}_2(\mathbb{F})$, and $c(T_{2,2}) = (q^2 - 1)(q^2 - q)$. Finally, if J is of type $T_{2,3}$, $C(J) = \left\{\left(\begin{smallmatrix} a & b \\ 0 & a \end{smallmatrix}\right) : a, b \in \mathbb{F}, a \neq 0\right\}$, and so $c(T_{2,3}) = q(q - 1)$.

To complete the example, we apply (3):

$$\begin{aligned} |\mathrm{Eig}_2(\mathbb{F})| &= \sum_{T \in \mathrm{DP}(2)} t(T) \frac{|\mathrm{GL}_2(\mathbb{F})|}{c(T)} = \sum_{i=1}^{3} t(T_{2,i}) \frac{(q^2 - 1)(q^2 - q)}{c(T_{2,i})} \\ &= \binom{q}{2} \frac{(q^2 - 1)(q^2 - q)}{(q - 1)^2} + q \frac{(q^2 - 1)(q^2 - q)}{(q^2 - 1)(q^2 - q)} + q \frac{(q^2 - 1)(q^2 - q)}{q(q - 1)} \\ &= \tfrac{1}{2}q^4 + q^3 - \tfrac{1}{2}q^2. \end{aligned}$$

4. The number of Jordan forms of a given type

Fix n and consider a double partition $T = \{S_1, S_2, \ldots, S_k\}$ of n where for $1 \leq i \leq k$, S_i is a set of positive integers. To determine the number of Jordan forms in $J_n(\mathbb{F})$ of type T, we count the number of ways to assign eigenvalues to the S_i. Of course, if $S_i = S_j$ then assigning eigenvalue λ_1 to S_i and λ_2 to S_j yields the same Jordan

form as assigning λ_2 to S_i and λ_1 to S_j, so we need to account for any repeated subsets in T. For example, if $n = 19$ and $T = \{\{2, 3\}, \{1, 1, 3\}, \{2, 3\}, \{4\}\}$, then the subset $\{2, 3\}$ is repeated twice, and there will be

$$\binom{q}{2}(q-2)(q-3) = \binom{q}{2}\binom{q-2}{1}\binom{q-3}{1} = \frac{q!}{2!\,1!\,1!\,(q-4)!}$$

elements of $J_n(\mathbb{F})$ of type T. The value of $t(T)$ in the general case is given by the following lemma.

Lemma 4.1. *Let $T = \{S_1, S_2, \dots, S_k\}$ be a double partition, where for $1 \le i \le k$, S_i is a set of positive integers. Let $B_1, \dots, B_l$ be equivalence classes of the sets in T so that S_i and S_j are in the same equivalence class if and only if $S_i = S_j$. Let $b_i = |B_i|$. Then the number of Jordan forms of type T in $J_n(\mathbb{F})$ is 0 if $q < k$ and otherwise is given by the formula*

$$\begin{aligned} t(T) &= \binom{q}{b_1}\binom{q-b_1}{b_2}\cdots\binom{q-b_1-b_2-\cdots-b_{l-1}}{b_l} \\ &= \frac{q!}{b_1!\,b_2!\cdots b_l!\,(q-b_1-b_2-\cdots-b_l)!}. \end{aligned}$$

Proof. There are $\binom{q}{b_1}$ ways to assign eigenvalues to the sets in B_1, $\binom{q-b_1}{b_2}$ ways to assign eigenvalues to the sets in B_2 without repeating any of the b_1 eigenvalues already assigned to sets in B_1, and so forth. □

5. Invertible matrices that commute with a Jordan form

In this section we first characterize the structure of the matrices that commute with Jordan forms, with the ultimate statement of this characterization coming in Corollary 5.3. This characterization depends only on the double partition type, and not the specific eigenvalues, and this observation is sufficient to prove Lemma 3.2. To determine $|C(J)|$, we must determine how many of the matrices of the form specified by Corollary 5.3 are invertible, and this is done starting on page 636.

For any matrix $A \in M_n(\mathbb{F})$, let $(A)_{i,j}$ denote the entry of A in the i-th row and j-th column. We say an $n \times m$ matrix A is *streaky upper triangular* if it has the following three properties (see Figure 3):

(i) $(A)_{i,1} = 0$ if $i > 1$.

(ii) $(A)_{n,j} = 0$ if $j < m$.

(iii) $(A)_{i,j} = (A)_{i+1,j+1}$ if $1 \le i \le n-1$, $1 \le j \le m-1$.

We denote the set of $n \times m$ streaky upper triangular matrices over $\mathbb{F}$ by $\mathrm{SUT}_{n,m}(\mathbb{F})$. In Lemmas 5.1 and 5.2 we examine products of the form $J_n A$ and AJ_m where A is $n \times m$, and J_n and J_m are $n \times n$ and $m \times m$ Jordan blocks, respectively (see

$$\begin{pmatrix} 2 & 1 & 0 & 3 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix} \qquad \begin{pmatrix} 0 & 0 & 3 & 3 & 1 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix} \qquad \begin{pmatrix} 2 & 5 & 1 & 0 & 5 \\ 0 & 2 & 5 & 1 & 0 \\ 0 & 0 & 2 & 5 & 1 \\ 0 & 0 & 0 & 2 & 5 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix} \qquad \begin{pmatrix} 0 & 1 & 3 & 0 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Figure 3. Some examples of streaky upper triangular matrices.

$$\begin{aligned} J_4 A &= \begin{pmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{pmatrix} \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} & a_{1,5} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} & a_{2,5} \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} & a_{3,5} \\ a_{4,1} & a_{4,2} & a_{4,3} & a_{4,4} & a_{4,5} \end{pmatrix} \\ &= \begin{pmatrix} \lambda a_{1,1} + a_{2,1} & \lambda a_{1,2} + a_{2,2} & \lambda a_{1,3} + a_{2,3} & \lambda a_{1,4} + a_{2,4} & \lambda a_{1,5} + a_{2,5} \\ \lambda a_{2,1} + a_{3,1} & \lambda a_{2,2} + a_{3,2} & \lambda a_{2,3} + a_{3,3} & \lambda a_{2,4} + a_{3,4} & \lambda a_{2,5} + a_{3,5} \\ \lambda a_{3,1} + a_{4,1} & \lambda a_{3,2} + a_{4,2} & \lambda a_{3,3} + a_{4,3} & \lambda a_{3,4} + a_{4,4} & \lambda a_{3,5} + a_{4,5} \\ \lambda a_{4,1} & \lambda a_{4,2} & \lambda a_{4,3} & \lambda a_{4,4} & \lambda a_{4,5} \end{pmatrix}, \end{aligned}$$

$$\begin{aligned} A J_5 &= \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} & a_{1,5} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} & a_{2,5} \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} & a_{3,5} \\ a_{4,1} & a_{4,2} & a_{4,3} & a_{4,4} & a_{4,5} \end{pmatrix} \begin{pmatrix} \lambda & 1 & 0 & 0 & 0 \\ 0 & \lambda & 1 & 0 & 0 \\ 0 & 0 & \lambda & 1 & 0 \\ 0 & 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & 0 & \lambda \end{pmatrix} \\ &= \begin{pmatrix} \lambda a_{1,1} & \lambda a_{1,2} + a_{1,1} & \lambda a_{1,3} + a_{1,2} & \lambda a_{1,4} + a_{1,3} & \lambda a_{1,5} + a_{1,4} \\ \lambda a_{2,1} & \lambda a_{2,2} + a_{2,1} & \lambda a_{2,3} + a_{2,2} & \lambda a_{2,4} + a_{2,3} & \lambda a_{2,5} + a_{3,4} \\ \lambda a_{3,1} & \lambda a_{3,2} + a_{3,1} & \lambda a_{3,3} + a_{3,2} & \lambda a_{3,4} + a_{3,3} & \lambda a_{3,5} + a_{3,4} \\ \lambda a_{4,1} & \lambda a_{4,2} + a_{4,1} & \lambda a_{4,3} + a_{4,2} & \lambda a_{4,4} + a_{4,3} & \lambda a_{4,5} + a_{3,4} \end{pmatrix}. \end{aligned}$$

Figure 4. Examples of $J_4 A$ and $A J_5$ where A is 4×5 and J_4 and J_5 are Jordan blocks.

Figure 4). We show $J_n A = A J_m$ if and only if J_n and J_m have the same eigenvalue and $A \in \mathrm{SUT}_{n,m}(\mathbb{F})$, or if J_n and J_m have different eigenvalues and A is the all zeros matrix. These lemmas enable us to characterize the matrices that commute with any matrix $J \in J_n(\mathbb{F})$

Lemma 5.1. *Let $J_n \in J_n(\mathbb{F})$ and $J_m \in J_m(\mathbb{F})$ be Jordan blocks with eigenvalue λ, and suppose A is an $n \times m$ matrix. Then $J_n A = A J_m$ if and only if $A \in \mathrm{SUT}_{n,m}(\mathbb{F})$.*

Proof. Suppose $J_n A = A J_m$. We will show $A \in \mathrm{SUT}_{n,m}(\mathbb{F})$.

First, we examine the individual entries in the first column. For $1 \le i \le n-1$, we have $(J_n A)_{i,1} = \lambda(A)_{i,1} + (A)_{i+1,1}$ and $(AJ_m)_{i,1} = \lambda(A)_{i,1}$. If these two quantities are equal, then $(A)_{i+1,1} = 0$ for $1 \le i \le n-1$, which implies A has property (i).

Next, we examine the individual entries in the last (n-th) row. For $2 \le j \le m$, we have $(J_n A)_{n,j} = \lambda(A)_{n,j}$ and $(AJ_m)_{n,j} = \lambda(A)_{n,j} + (A)_{n,j-1}$. If these two quantities are equal, then $(A)_{n,j-1} = 0$ for $2 \le j \le m$, which implies A has property (ii).

Finally, we examine the other entries. If $1 \le i \le n-1$ and $2 \le j \le m$, we have $(J_n A)_{i,j} = \lambda(A)_{i,j} + (A)_{i+1,j}$ and $(AJ_m)_{i,j} = \lambda(A)_{i,j} + (A)_{i,j-1}$. If these two quantities are equal, then $(A)_{i+1,j} = (A)_{i,j-1}$ for $1 \le i \le n-1$, $2 \le j \le m$, which implies A has property (iii).

Conversely, suppose $A \in \mathrm{SUT}_{n,m}(\mathbb{F})$. Then we verify $J_n A = AJ_m$ using four cases.

- For $i = n$, $j = 1$, $(J_n A)_{n,1} = \lambda(A)_{n,1} = (AJ_m)_{n,1}$.
- For $i \le n-1$, $j = 1$, $(J_n A)_{i,1} = \lambda(A)_{i,1} + (A)_{i+1,1} = \lambda(A)_{i,1} = (AJ_m)_{i,1}$.
- For $i = n$, $j \ge 2$, $(J_n A)_{n,j} = \lambda(A)_{n,j} = \lambda(A)_{n,j} + (A)_{n,j-1} = (AJ_m)_{n,j}$.
- For $i \le n-1$, $2 \le j$, $(J_n A)_{i,j} = \lambda(A)_{i,j} + (A)_{i+1,j} = \lambda(A)_{i,j} + (A)_{i,j-1} = (AJ_m)_{i,j}$. □

Lemma 5.2. *Let $J_n \in J_n(\mathbb{F})$ be a Jordan block with eigenvalue λ and $J_m \in J_m(\mathbb{F})$ be a Jordan block with eigenvalue μ, where $\lambda \ne \mu$, and suppose A is an $n \times m$ matrix. Then $J_n A = AJ_m$ if and only if $(A)_{i,j} = 0$ for all $1 \le i \le n$, $1 \le j \le m$.*

Proof. If $(A)_{i,j} = 0$ for all $1 \le i \le n$, $1 \le j \le m$, then $J_n A = AJ_m$.

Conversely, suppose $J_n A = AJ_m$. First examine the case $i = n$, $j = 1$: $(J_n A)_{n,1} = \lambda(A)_{n,1}$ and $(AJ_m)_{n,1} = \mu(A)_{n,1}$ imply $(A)_{n,1} = 0$.

Next we proceed by induction on the first column ($j = 1$). We know $(A)_{n,1} = 0$. For any $k \ge 0$, assume $(A)_{n-k,1} = 0$. Then

$$(J_n A)_{n-(k+1),1} = \lambda(A)_{n-(k+1),1} + (A)_{n-k,1} = \lambda(A)_{n-(k+1),1},$$

while $(AJ_m)_{n-(k+1),1} = \mu(A)_{n-(k+1),1}$. If these two quantities are equal, then $(A)_{n-(k+1),1} = 0$. From this we conclude that $(A)_{i,1} = 0$ for all $1 \le i \le n$.

Next we apply the same induction argument to the last row ($i = n$): For any $j \ge 1$, assume $(A)_{n,j} = 0$. Then $(J_n A)_{n,j+1} = \lambda(A)_{n,j+1}$, while

$$(AJ_m)_{n,j+1} = \mu(A)_{n,j+1} + (A)_{n,j} = \mu(A)_{n,j+1}.$$

If these two quantities are equal, then $(A)_{n,j+1} = 0$. From this we conclude that $(A)_{n,j} = 0$ for all $1 \le j \le m$.

To complete the proof, we show that for all $i < n$, $j > 1$, if $(A)_{i,j-1} = 0$ and $(A)_{i+1,j} = 0$ then $(A)_{i,j} = 0$. If $i < n$, $j > 1$, then $(J_n A)_{i,j} = \lambda(A)_{i,j} + (A)_{i+1,j}$, while $(AJ_m)_{i,j} = \mu(A)_{i,j} + (A)_{i,j-1}$. If $(A)_{i,j-1} = 0$ and $(A)_{i+1,j} = 0$ then $\lambda(A)_{i,j} =$

$\mu(A)_{i,j}$, which implies $(A)_{i,j} = 0$. We have shown if the entries below and to the left of $(A)_{i,j}$ are both zero, then $(A)_{i,j}$ is zero as well. Since we know the first column and last row of A contain only zeros, by induction all other entries must be zero as well. □

If A is a block matrix, let $A^{i,j}$ denote the block in the i-th row and j-th column.

Corollary 5.3. *Suppose $n = n_1 + n_2 + \cdots + n_k$, and $J \in J_n(\mathbb{F})$ is a Jordan form with partition type $\{n_1, n_2, \ldots, n_k\}$, that is, J is the direct sum of k Jordan blocks $J_1 \in J_{n_1}(\mathbb{F})$, $J_2 \in J_{n_2}(\mathbb{F}), \ldots, J_k \in J_{n_k}(\mathbb{F})$, with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$, some of which may be the same. We can represent J as a $k \times k$ block matrix where $J^{i,i} = J_i$ for $1 \le i \le k$, and $J^{i,j}$ is an $n_i \times n_j$ all-zeros matrix if $i \neq j$. Then $JA = AJ$ if and only if A is a $k \times k$ block matrix, where for $1 \le i, j \le k$, the block $A^{i,j}$ is an element of $\mathrm{SUT}_{n_i,n_j}(\mathbb{F})$ if $\lambda_i = \lambda_j$, and otherwise $A^{i,j}$ is an $n_i \times n_j$ matrix containing all zeros.*

Figure 5 contains an illustration of Corollary 5.3. In this example, $J \in J_{11}(\mathbb{F})$, and in the terminology of Corollary 5.3, we have $k = 6$ and

$$(n_1, n_2, n_3, n_4, n_5, n_6) = (3, 2, 2, 1, 1, 2),$$
$$(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6) = (1, 1, 1, 4, 4, 0).$$

Proof of Corollary 5.3. Decompose $A \in M_n(\mathbb{F})$ as a $k \times k$ block matrix, where block $A^{i,j}$ is $n_i \times n_j$. Then the equation $JA = AJ$ implies $J_i A^{i,j} = A^{i,j} J_j$. If $\lambda_i = \lambda_j$, then $J_i A^{i,j} = A^{i,j} J_j$ if and only if $A^{i,j} \in \mathrm{SUT}_{n_i,n_j}(\mathbb{F})$ by Lemma 5.1. If $\lambda_i \neq \lambda_j$, then $J_i A^{i,j} = A^{i,j} J_j$ if and only if $A^{i,j}$ contains all zeros, by Lemma 5.2. □

We now show that Lemma 3.2 follows from Corollary 5.3.

$$J = \left(\begin{array}{ccc|cc|cc||c|c||cc}
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
\hline\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 \\
\hline\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right), \quad
A = \left(\begin{array}{ccc|cc|cc||c|c||cc}
a & b & c & l & m & n & p & 0 & 0 & 0 & 0 \\
0 & a & b & 0 & l & 0 & n & 0 & 0 & 0 & 0 \\
0 & 0 & a & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & s & t & d & e & q & r & 0 & 0 & 0 & 0 \\
0 & 0 & s & 0 & d & 0 & q & 0 & 0 & 0 & 0 \\
\hline
0 & u & v & w & x & f & g & 0 & 0 & 0 & 0 \\
0 & 0 & u & 0 & w & 0 & f & 0 & 0 & 0 & 0 \\
\hline\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & h & y & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & z & i & 0 & 0 \\
\hline\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & j & k \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & j
\end{array}\right).$$

Figure 5. A matrix commutes with the Jordan form J if and only if it has the form of the matrix A.

Proof of Lemma 3.2. Implicit in the statement of Corollary 5.3 is the fact that the characterization of the matrices that commute with $J \in J_n(\mathbb{F})$ does not depend on the specific eigenvalues of the Jordan blocks, only on which blocks share the same eigenvalue. In other words, it depends only on the double partition type of J. The quantity $|C(J)|$ equals the number of invertible matrices that fit this characterization, and since the characterization depends only on the double partition type of J, if two Jordan forms J_1 and J_2 have the same double partition type, then $|C(J_1)| = |C(J_2)|$. □

How to determine $|C(J)|$. It remains to explain how to determine $|C(J)|$ for $J \in J_n(\mathbb{F})$. Simply characterizing the matrices that commute with J is not enough, since elements of $C(J)$ must be invertible. Thus we need to determine which of the matrices of the form given in Corollary 5.3 are invertible. Examining the example in Figure 5, we note that A can be represented as a block diagonal matrix, where one block is 7×7 and the other two blocks are 2×2. It turns out that for any $J \in J_n(\mathbb{F})$, the matrices in $C(J)$ can be represented as block diagonal matrices, and since the determinant of a block diagonal matrix is the product of the determinants of the matrices on the diagonal, we must characterize when the matrices on the diagonal are invertible. As in the example in Figure 5, the matrices on the diagonal are themselves block matrices where each block is streaky upper triangular, and Lemma 5.4 describes when such a matrix is invertible.

The proof of Lemma 5.4 proceeds by taking any block matrix where each block is streaky upper triangular and permuting the rows and columns to obtain an upper triangular block matrix, which is invertible if and only if the matrices on the diagonal are invertible. We use the fact that switching the position of two rows or columns of a matrix changes the sign of the determinant. Thus if the columns of a matrix are permuted in some way, and the rows are permuted the same way, the sign of the determinant will change an even number of times, and thus will ultimately be unchanged. To describe the diagonal blocks of the rearranged matrix, we need the following notation: If A is a $b \times b$ block matrix, where each block $A^{i,j}$ is an element of $\mathrm{SUT}_{n,n}(\mathbb{F})$, then we denote by A' the $b \times b$ matrix made up of the entries that are on the diagonal of each block $A^{i,j}$, that is, $(A')_{i,j} = (A^{i,j})_{1,1}$. For example, in Figure 6, A is a 2×2 block matrix with each block in $\mathrm{SUT}_{3,3}(\mathbb{F})$ (i.e., $b = 2$, $n = 3$) and D is a 4×4 block matrix with each block in $\mathrm{SUT}_{2,2}(\mathbb{F})$ (i.e., $b = 4$, $n = 2$). In this case, $A' = \left(\begin{smallmatrix} 3 & 0 \\ 1 & 4 \end{smallmatrix}\right)$, and D' is the 4×4 matrix shown in the figure. We write $\{b_1 n_1, \dots, b_k n_k\}$ for the collection having b_i copies of n_i, for $1 \le i \le k$.

Lemma 5.4. *Let $\{b_1 n_1, \dots, b_k n_k\}$ be a partition of n, with $n_1 > n_2 > \cdots > n_k$. Suppose $A \in M_n(\mathbb{F})$ is a $k \times k$ block matrix, where $A^{i,j}$ is $b_i n_i \times b_j n_j$, and each $A^{i,j}$ can itself be represented as a $b_i \times b_j$ block matrix, where each block is an element of* $\mathrm{SUT}_{n_i,n_j}(\mathbb{F})$. *Then A is invertible if and only if $(A^{i,i})'$ is invertible for*

$$A = \left(\begin{array}{ccc|ccc} 3 & 2 & 4 & 0 & 3 & 0 \\ 0 & 3 & 2 & 0 & 0 & 3 \\ 0 & 0 & 3 & 0 & 0 & 0 \\ \hline 1 & 0 & 2 & 4 & 1 & 4 \\ 0 & 1 & 0 & 0 & 4 & 1 \\ 0 & 0 & 1 & 0 & 0 & 4 \end{array}\right), \quad D = \left(\begin{array}{cc|cc|cc|cc} 1 & 1 & 0 & 2 & 3 & 2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 3 & 0 & 0 \\ \hline 2 & 1 & 4 & 3 & 3 & 4 & 3 & 4 \\ 0 & 2 & 0 & 4 & 0 & 3 & 0 & 3 \\ \hline 1 & 1 & 1 & 4 & 2 & 5 & 3 & 1 \\ 0 & 1 & 0 & 1 & 0 & 2 & 0 & 3 \\ \hline 1 & 0 & 0 & 1 & 1 & 2 & 2 & 2 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 2 \end{array}\right), \quad D' = \begin{pmatrix} 1 & 0 & 3 & 0 \\ 2 & 4 & 3 & 3 \\ 1 & 1 & 2 & 3 \\ 1 & 0 & 1 & 2 \end{pmatrix}.$$

Figure 6. Block matrices A and D where each block is an $n \times n$ streaky upper triangular matrix.

each i, $1 \le i \le k$. *In fact*,

$$\det A = \prod_{i=1}^{k} (\det(A^{i,i})')^{n_i}.$$

Proof. We go by induction on n_1. If $n_1 = 1$, then $A = A^{1,1} = (A^{1,1})'$ and the statement follows. Now suppose $\{b_1 n_1, \dots, b_k n_k\}$ is a partition of n, with $n_1 > n_2 > \cdots > n_k$. Let $A \in M_n(\mathbb{F})$ be a $k \times k$ block matrix, where $A^{i,j}$ is $b_i n_i \times b_j n_j$, and each $A^{i,j}$ can itself be represented as a $b_i \times b_j$ block matrix, where each block is an element of $\mathrm{SUT}_{n_i,n_j}(\mathbb{F})$ (see Figure 7).

Let $B_0 = BN_0 = 0$, and for $1 \le j \le k$, let B_j denote the sum $b_1 + b_2 + \cdots + b_j$, and BN_j denote the sum $b_1 n_1 + b_2 n_2 + \cdots + b_j n_j$. Permute the columns of A so that if $0 \le i \le B_k - 1$, and j is the smallest index so that $i < B_j$, column $BN_{j-1} + (i - B_{j-1}) n_j + 1$ becomes column $i + 1$, and all other columns become columns $B_k + 1$ to n, in the same order as they were in the original matrix A. Permute the rows the same way to obtain a 2×2 block matrix D with block structure $\left(\begin{smallmatrix} D_1 & D_2 \\ D_3 & D_4 \end{smallmatrix}\right)$ (see Figure 7). Since the rows and columns were rearranged symmetrically, $\det D = \det A$. The minor D_1 is a $k \times k$ block upper triangular matrix where for $1 \le i \le k$, $(D_1)^{i,i} = (A^{i,i})'$. Since $n_1 > n_2 > \cdots > n_k$ and each block is streaky upper triangular, the $(n - B_k) \times B_k$ matrix D_3 contains only zeros. D_4 can be represented as a $k \times k$ (if $n_k > 1$) or $k-1 \times k-1$ (if $n_k = 1$) block matrix where $(D_4)^{i,j}$ is $b_i(n_i - 1) \times b_j(n_j - 1)$, and each $(D_4)^{i,j}$ can itself be represented as a $b_i \times b_j$ block matrix, where each block is an element of $\mathrm{SUT}_{n_i-1,n_j-1}(\mathbb{F})$. Furthermore, $((D_4)^{i,i})' = (A^{i,i})'$ for each i. Thus by the inductive hypothesis, $\det D_4 = \prod_{i=1}^{k} (\det(A^{i,i})')^{n_i - 1}$. Since the determinant of an upper triangular block matrix is the product of the determinants of the blocks on the diagonal,

$$\det A = \det D_1 \det D_4 = \prod_{i=1}^{k} \det(A^{i,i})' \prod_{i=1}^{k} (\det(A^{i,i})')^{n_i - 1} = \prod_{i=1}^{k} (\det(A^{i,i})')^{n_i}. \quad \square$$

$$\left(\begin{array}{cccc||ccc|ccc||cc}
1&2&4&0&4&0&0&3&3&1&0&2\\
0&1&2&4&0&4&0&0&3&1&0&0\\
0&0&1&2&0&0&4&0&0&3&0&0\\
0&0&0&1&0&0&0&0&0&0&0&0\\
\hline\hline
0&2&1&1&1&4&2&4&2&2&3&1\\
0&0&2&1&0&1&4&0&4&2&0&1\\
0&0&0&2&0&0&1&0&0&4&0&0\\
\hline
0&1&0&3&4&0&3&0&0&2&1&1\\
0&0&1&0&0&4&0&0&0&0&0&1\\
0&0&0&1&0&0&4&0&0&0&0&0\\
\hline\hline
0&0&4&1&0&4&4&0&4&2&3&1\\
0&0&0&4&0&0&4&0&0&4&0&3
\end{array}\right)
\xrightarrow[\text{and columns}]{\text{permute rows}}
\left(\begin{array}{c|c|c|cccccccc}
1&4&3&0&2&4&0&0&0&3&1&2\\
\hline
0&1&4&3&2&1&1&4&2&2&2&1\\
\hline
0&4&0&1&1&0&3&0&3&0&2&1\\
\hline
0&0&0&3&0&4&1&4&4&4&2&1\\
\hline\hline
0&0&0&0&1&2&4&4&0&3&1&0\\
0&0&0&0&0&1&2&0&4&0&3&0\\
0&0&0&0&0&0&1&0&0&0&0&0\\
\hline
0&0&0&0&0&2&1&1&4&4&2&1\\
0&0&0&0&0&0&2&0&1&0&4&0\\
\hline
0&0&0&0&0&1&0&4&0&0&0&1\\
0&0&0&0&0&0&1&0&4&0&0&0\\
\hline\hline
0&0&0&0&0&0&4&0&4&0&4&3
\end{array}\right)$$

Figure 7. An illustration of Lemma 5.4 with a 12×12 matrix A (left). Here $n = 12$ is partitioned as $\{1 \cdot 4, 2 \cdot 3, 1 \cdot 2\}$, that is, $k = 3$, $(n_1, n_2, n_3) = (4, 3, 2)$, and $(b_1, b_2, b_3) = (1, 2, 1)$. Then $(B_1, B_2, B_3) = (1, 3, 4)$, $(BN_1, BN_2, BN_3) = (4, 10, 12)$, and in the permutation, columns 1, 5, 8, and 11 become columns 1,2,3, and 4, with the other columns becoming columns 5 through 12, keeping their original order. The rows are permuted the same way. In this example, $(A^{1,1})' = (1)$, $(A^{2,2})' = (1440)$, and $(A^{3,3})' = (3)$, and $\det A = 1^4 \cdot (\det(1440))^3 \cdot 3^2$.

The inductive argument in the proof of Lemma 5.4 implies that the columns of A can be permuted to obtain a block upper triangular matrix, where the diagonal blocks consist of n_i copies each of $(A^{i,i})'$, for $1 \le i \le k$. As a corollary, this implies that $A^{i,i}$ is invertible if and only if $(A^{i,i})'$ is invertible. For example, in Figure 6, $\det A = (\det A')^3$ and $\det D = (\det D')^2$.

Corollary 5.5. *Let $g(b_1n_1, \dots, b_kn_k)$ denote the number of invertible matrices of the type specified in Lemma 5.4. Then*

$$g(b_1n_1, \dots, b_kn_k) = \left(\prod_{i=1}^{k} q^{b_i^2(n_i-1)} \prod_{j=0}^{b_i-1} (q^{b_i} - q^j)\right)\left(\prod_{1\le i<j\le k} q^{2b_ib_jn_j}\right).$$

Proof. First we fix i and count the number of different possibilities for $A^{i,i}$. Since $(A^{i,i})'$ must be invertible, and $(A^{i,i})'$ is $b_i \times b_i$, there are $|\mathrm{GL}_{b_i}(\mathbb{F})| = \prod_{j=0}^{b_i-1}(q^{b_i} - q^j)$ different ways to choose the elements of $(A^{i,i})'$. For each of the b_i^2 blocks in $A^{i,i}$, there are $n_i - 1$ other diagonals whose entries may be nonzero, and they may be chosen arbitrarily, so there are $q^{b_i^2(n_i-1)}$ ways to choose these other entries. Thus

there are

$$f(b_i, n_i) = \left(q^{b_i^2(n_i-1)} \prod_{j=0}^{b_i-1} (q_i^n - q^j) \right)$$

ways to choose the entries in a diagonal block $A^{i,i}$.

All of the other entries in A which are not required to be zero may be chosen arbitrarily, subject to the constraint that each block is streaky upper triangular. If $i < j$, an element of $\mathrm{SUT}_{n_i,n_j}(\mathbb{F})$ has n_j diagonals that may be nonzero, and for $i \neq j$ there are $b_i b_j$ blocks in each of $\mathrm{SUT}_{n_i,n_j}(\mathbb{F})$ and $\mathrm{SUT}_{n_j,n_i}(\mathbb{F})$. Thus there are a total of

$$\prod_{1 \le i < j \le k} q^{2b_i b_j n_j}$$

ways to choose these entries. We conclude that

$$\begin{aligned} g(b_1 n_1, \dots, b_k n_k) &= \left(\prod_{i=1}^{k} f(b_i, n_i) \right) \left(\prod_{1 \le i < j \le k} q^{2b_i b_j n_j} \right) \\ &= \left(\prod_{i=1}^{k} q^{b_i^2(n_i-1)} \prod_{j=0}^{b_i-1} (q^{b_i} - q^j) \right) \left(\prod_{1 \le i < j \le k} q^{2b_i b_j n_j} \right). \qquad \square \end{aligned}$$

Now we determine $|C(J)|$ for any $J \in J_n(\mathbb{F})$. Any matrix in $C(J)$ can be represented as a block diagonal matrix, where each block on the diagonal is of the type specified in Lemma 5.4. To be precise, let $n = n_1 + n_2 + \cdots + n_l$, and $J \in J_n(\mathbb{F})$ have double partition type T given by

$$\{\{b_{1,1}n_{1,1}, \dots, b_{1,k_1}n_{1,k_1}\}, \{b_{2,1}n_{2,1}, \dots, b_{2,k_2}n_{2,k_2}\}, \dots, \{b_{l,1}n_{l,1}, \dots, b_{l,k_l}n_{l,k_l}\}\},$$

where, for $1 \le \alpha \le l$, we have $n_\alpha = \sum_{i=1}^{k_\alpha} b_{\alpha,i} n_{\alpha,i}$ and $n_{\alpha,i} > n_{\alpha,j}$ if $i < j$. Denote the eigenvalue associated with $\{b_{\alpha,1}n_{\alpha,1}, \dots, b_{\alpha,k_\alpha}n_{\alpha,k_\alpha}\}$ by λ_α. We can represent J as an $l \times l$ block diagonal matrix where $J^{\alpha,\alpha}$ contains the Jordan blocks with eigenvalue λ_α. If A commutes with J, then A can be represented as an $l \times l$ block matrix where the block $A^{\alpha,\beta}$ is $n_\alpha \times n_\beta$, and Lemma 5.2 guarantees that all off-diagonal blocks $A^{\alpha,\beta}$, $\alpha \neq \beta$ contain all zeros. So A is a block diagonal matrix, which is invertible if and only if each block on the diagonal is itself an invertible matrix. Thus to determine the number of invertible matrices $A \in C(J)$, it suffices to determine for each $1 \le \alpha \le l$ the number of invertible matrices $A^{\alpha,\alpha}$.

The matrix $A^{\alpha,\alpha}$ can be represented as a $k_\alpha \times k_\alpha$ block matrix, where $(A^{\alpha,\alpha})^{i,j}$ is $b_{\alpha,i}n_{\alpha,i} \times b_{\alpha,j}n_{\alpha,j}$, and is itself a $b_{\alpha,i} \times b_{\alpha,j}$ block matrix where Lemma 5.1 implies each block is an element of $\mathrm{SUT}_{n_{\alpha,i},n_{\alpha,j}}(\mathbb{F})$. By Lemma 5.4 it is invertible if and only if the diagonal blocks $(A^{\alpha,\alpha})^{i,i}$ are invertible. Corollary 5.5 implies the number of invertible matrices that have the required block structure is given by

$$|C(J)| = \prod_{\alpha=1}^{l} g(b_{\alpha,1}n_{\alpha,1}, \dots, b_{\alpha,k_\alpha}n_{\alpha,k_\alpha})$$
$$= \prod_{\alpha=1}^{l}\left(\left(\prod_{i=1}^{k_\alpha} q^{b_{\alpha,i}^2(n_{\alpha,i}-1)} \prod_{j=0}^{b_{\alpha,i}-1}(q^{b_{\alpha,i}} - q^j)\right) \prod_{1\le i<j\le k_\alpha} q^{2b_{\alpha,i}b_{\alpha,j}n_{\alpha,j}}\right). \tag{4}$$

Using the matrix J from Figure 5 as an example, the variables in (4) are

$$l = 3, \quad (n_1, n_2, n_3) = (7, 2, 2), \quad (k_1, k_2, k_3) = (2, 1, 1),$$
$$n_1 = 7 = 1\cdot 3 + 2\cdot 2 = b_{1,1}n_{1,1} + b_{1,2}n_{1,2},$$
$$n_2 = 2 = 2\cdot 1 = b_{2,1}n_{2,1}, \quad n_3 = 2 = 1\cdot 2 = b_{3,1}n_{3,1}.$$

Thus for the matrix A in the figure to be invertible, there are $|\mathrm{GL}_1(\mathbb{F})| = q-1$ choices each for a and j, $|\mathrm{GL}_2(\mathbb{F})| = (q^2-1)(q^2-q)$ choices each for $\{d, q, w, f\}$ and $\{h, y, z, i\}$, and q choices for each other letter. Thus

$$|C(J)| = (q-1)^2((q^2-1)(q^2-q))^2 q^{16}.$$

6. $\mathrm{Eig}_3(\mathbb{F})$ and $\mathrm{Eig}_4(\mathbb{F})$

In this section we compute $|\mathrm{Eig}_3(\mathbb{F})|$ and $|\mathrm{Eig}_4(\mathbb{F})|$ for any field $\mathbb{F}$ with q elements.

Theorem 6.1. *The number of* 3×3 *matrices with entries in* $\mathbb{F}$ *whose eigenvalues are all in* $\mathbb{F}$ *is*

$$|\mathrm{Eig}_3(\mathbb{F})| = \tfrac{1}{6}q^9 + \tfrac{5}{6}q^8 + \tfrac{2}{3}q^7 - \tfrac{1}{6}q^6 - \tfrac{5}{6}q^5 + \tfrac{1}{3}q^4.$$

Proof. There are three partitions of 3: $3 = 1+1+1 = 1+2 = 3$, and 6 double partitions:

$$T_{3,1} = \{\{1\},\{1\},\{1\}\}, \qquad T_{3,3} = \{\{1,1,1\}\}, \qquad T_{3,5} = \{\{2,1\}\},$$
$$T_{3,2} = \{\{1,1\},\{1\}\}, \qquad T_{3,4} = \{\{2\},\{1\}\}, \qquad T_{3,6} = \{\{3\}\}.$$

For $1 \le i \le 6$, Lemma 4.1 gives the value of $t(T_{3,i})$. Corollary 5.3 gives the form of $C(J)$ for $J \in T_{3,i}$, and the value of $c(T_{3,i})$ is determined by (4). This information is summarized in Table 1.

Using (3), we conclude that $|\mathrm{Eig}_3(\mathbb{F})|$ equals

$$\sum_{i=1}^{6} t(T_{3,i})\frac{(q^3-1)(q^3-q)(q^3-q^2)}{c(T_{3,i})} = \tfrac{1}{6}q^9 + \tfrac{5}{6}q^8 + \tfrac{2}{3}q^7 - \tfrac{1}{6}q^6 - \tfrac{5}{6}q^5 + \tfrac{1}{3}q^4. \quad \square$$

Theorem 6.2. *The number* $|\mathrm{Eig}_4(\mathbb{F})|$ *of* 4×4 *matrices with entries in* $\mathbb{F}$ *whose eigenvalues are all in* $\mathbb{F}$ *equals*

$$= \tfrac{1}{24}q^{16} + \tfrac{3}{8}q^{15} + \tfrac{11}{12}q^{14} + \tfrac{5}{8}q^{13} - \tfrac{1}{4}q^{12} - \tfrac{1}{8}q^{11} - \tfrac{5}{12}q^{10} + \tfrac{3}{8}q^9 + \tfrac{5}{24}q^8 - \tfrac{1}{4}q^7 - \tfrac{1}{2}q^6.$$

	$T_{3,i}$	$t(T_{3,i})$	$C(J) : J \in T_{3,i}$	$c(T_{3,i})$
$T_{3,1}$	$\left\{\begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}\right\}$	$\binom{q}{3}$	$\left\{\begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix} : a, b, c \neq 0\right\}$	$(q-1)^3$
$T_{3,2}$	$\left\{\begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_1 & 0 \\ 0 & 0 & \lambda_2 \end{bmatrix}\right\}$	$q(q-1)$	$\left\{\begin{bmatrix} a & b & 0 \\ c & d & 0 \\ 0 & 0 & e \end{bmatrix} : e, ad-bc \neq 0\right\}$	$(q^2-1)(q^2-q)(q-1)$
$T_{3,3}$	$\left\{\begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix}\right\}$	q	$\mathrm{GL}_3(\mathbb{F})$	$(q^3-1)(q^3-q)(q^3-q^2)$
$T_{3,4}$	$\left\{\begin{bmatrix} \lambda_1 & 1 & 0 \\ 0 & \lambda_1 & 0 \\ 0 & 0 & \lambda_2 \end{bmatrix}\right\}$	$q(q-1)$	$\left\{\begin{bmatrix} a & b & 0 \\ 0 & a & 0 \\ 0 & 0 & c \end{bmatrix} : a, c \neq 0\right\}$	$q(q-1)^2$
$T_{3,5}$	$\left\{\begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix}\right\}$	q	$\left\{\begin{bmatrix} a & b & c \\ 0 & a & 0 \\ 0 & d & e \end{bmatrix} : a, e \neq 0\right\}$	$q^3(q-1)^2$
$T_{3,6}$	$\left\{\begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix}\right\}$	q	$\left\{\begin{bmatrix} a & b & c \\ 0 & a & b \\ 0 & 0 & a \end{bmatrix} : a \neq 0\right\}$	$q^2(q-1)$

Table 1. Values of $t(T_{3,i})$ and $c(T_{3,i})$.

Proof. There are five partitions of 4:

$$4 = 1+1+1+1 = 1+1+2 = 2+2 = 1+3 = 4,$$

and 14 double partitions:

$$\begin{aligned}
T_{4,1} &= \{\{1\}, \{1\}, \{1\}, \{1\}\}, & T_{4,8} &= \{\{2, 1\}, \{1\}\}, \\
T_{4,2} &= \{\{1, 1\}, \{1\}, \{1\}\}, & T_{4,9} &= \{\{2, 1, 1\}\}, \\
T_{4,3} &= \{\{1, 1\}, \{1, 1\}\}, & T_{4,10} &= \{\{2\}, \{2\}\}, \\
T_{4,4} &= \{\{1, 1, 1\}, \{1\}\}, & T_{4,11} &= \{\{2, 2\}\}, \\
T_{4,5} &= \{\{1, 1, 1, 1\}\}, & T_{4,12} &= \{\{3\}, \{1\}\}, \\
T_{4,6} &= \{\{2\}, \{1\}, \{1\}\}, & T_{4,13} &= \{\{1, 3\}\}, \\
T_{4,7} &= \{\{2\}, \{1, 1\}\}, & T_{4,14} &= \{\{4\}\}.
\end{aligned}$$

The proof proceeds in the same way as the proof of Theorem 6.1, with the relevant information summarized in Table 2. □

	$T_{4,i}$	$t(T_{4,i})$	$C(J) : J \in T_{4,i}$	$c(T_{4,i})$
$T_{4,1}$	$\left\{\begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{bmatrix}\right\}$	$\binom{q}{4}$	$\left\{\begin{bmatrix} a & 0 & 0 & 0 \\ 0 & b & 0 & 0 \\ 0 & 0 & c & 0 \\ 0 & 0 & 0 & d \end{bmatrix}\right\}$	$(q-1)^4$
$T_{4,2}$	$\left\{\begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & \lambda_2 & 0 \\ 0 & 0 & 0 & \lambda_3 \end{bmatrix}\right\}$	$\binom{q}{2}(q-2)$	$\left\{\begin{bmatrix} a & b & 0 & 0 \\ c & d & 0 & 0 \\ 0 & 0 & e & 0 \\ 0 & 0 & 0 & f \end{bmatrix}\right\}$	$(q^2-q)(q^2-1)(q-1)^2$
$T_{4,3}$	$\left\{\begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & \lambda_2 & 0 \\ 0 & 0 & 0 & \lambda_2 \end{bmatrix}\right\}$	$\binom{q}{2}$	$\left\{\begin{bmatrix} a & b & 0 & 0 \\ c & d & 0 & 0 \\ 0 & 0 & e & f \\ 0 & 0 & g & h \end{bmatrix}\right\}$	$(q^2-q)^2(q^2-1)^2$
$T_{4,4}$	$\left\{\begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & \lambda_1 & 0 \\ 0 & 0 & 0 & \lambda_2 \end{bmatrix}\right\}$	$q(q-1)$	$\left\{\begin{bmatrix} a & b & c & 0 \\ d & e & f & 0 \\ g & h & i & 0 \\ 0 & 0 & 0 & j \end{bmatrix}\right\}$	$(q^3-1)(q^3-q)(q^3-q^2)(q-1)$
$T_{4,5}$	$\left\{\begin{bmatrix} \lambda & 0 & 0 & 0 \\ 0 & \lambda & 0 & 0 \\ 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & \lambda \end{bmatrix}\right\}$	q	$\mathrm{GL}_4(\mathbb{F})$	$(q^4-1)(q^4-q)(q^4-q^2)(q^4-q^3)$
$T_{4,6}$	$\left\{\begin{bmatrix} \lambda_1 & 1 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & \lambda_2 & 0 \\ 0 & 0 & 0 & \lambda_3 \end{bmatrix}\right\}$	$\binom{q}{2}(q-2)$	$\left\{\begin{bmatrix} a & b & 0 & 0 \\ 0 & a & 0 & 0 \\ 0 & 0 & c & 0 \\ 0 & 0 & 0 & d \end{bmatrix}\right\}$	$q(q-1)^3$
$T_{4,7}$	$\left\{\begin{bmatrix} \lambda_1 & 1 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & \lambda_2 & 0 \\ 0 & 0 & 0 & \lambda_2 \end{bmatrix}\right\}$	$q(q-1)$	$\left\{\begin{bmatrix} a & b & 0 & 0 \\ 0 & a & 0 & 0 \\ 0 & 0 & c & d \\ 0 & 0 & e & f \end{bmatrix}\right\}$	$q(q-1)(q^2-1)(q^2-q)$
$T_{4,8}$	$\left\{\begin{bmatrix} \lambda_1 & 1 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & \lambda_1 & 0 \\ 0 & 0 & 0 & \lambda_2 \end{bmatrix}\right\}$	$q(q-1)$	$\left\{\begin{bmatrix} a & b & c & 0 \\ 0 & a & 0 & 0 \\ 0 & d & e & 0 \\ 0 & 0 & 0 & f \end{bmatrix}\right\}$	$q^3(q-1)^3$
$T_{4,9}$	$\left\{\begin{bmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 0 & 0 \\ 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & \lambda \end{bmatrix}\right\}$	q	$\left\{\begin{bmatrix} a & b & c & d \\ 0 & a & 0 & 0 \\ 0 & e & f & g \\ 0 & h & i & j \end{bmatrix}\right\}$	$(q^2-q)(q^2-1)(q-1)q^5$

Table 2. Values of $t(T_{4,i})$ and $c(T_{4,i})$.

(*Continued on next page*)

	$T_{4,i}$	$t(T_{4,i})$	$C(J) : J \in T_{4,i}$	$c(T_{4,i})$
$T_{4,10}$	$\left\{\begin{bmatrix} \lambda_1 & 1 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & \lambda_2 & 1 \\ 0 & 0 & 0 & \lambda_2 \end{bmatrix}\right\}$	$\binom{q}{2}$	$\left\{\begin{bmatrix} a & b & 0 & 0 \\ 0 & a & 0 & 0 \\ 0 & 0 & c & d \\ 0 & 0 & 0 & c \end{bmatrix}\right\}$	$q^2(q-1)^2$
$T_{4,11}$	$\left\{\begin{bmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 0 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{bmatrix}\right\}$	q	$\left\{\begin{bmatrix} a & b & c & d \\ 0 & a & 0 & c \\ e & f & g & h \\ 0 & e & 0 & g \end{bmatrix}\right\}$	$(q^2-q)(q^2-1)q^4$
$T_{4,12}$	$\left\{\begin{bmatrix} \lambda_1 & 1 & 0 & 0 \\ 0 & \lambda_1 & 1 & 0 \\ 0 & 0 & \lambda_1 & 0 \\ 0 & 0 & 0 & \lambda_2 \end{bmatrix}\right\}$	$q(q-1)$	$\left\{\begin{bmatrix} a & b & c & 0 \\ 0 & a & b & 0 \\ 0 & 0 & a & 0 \\ 0 & 0 & 0 & d \end{bmatrix}\right\}$	$q^2(q-1)^2$
$T_{4,13}$	$\left\{\begin{bmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & \lambda \end{bmatrix}\right\}$	q	$\left\{\begin{bmatrix} a & b & c & d \\ 0 & a & b & 0 \\ 0 & 0 & a & 0 \\ 0 & 0 & e & f \end{bmatrix}\right\}$	$q^4(q-1)^2$
$T_{4,14}$	$\left\{\begin{bmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{bmatrix}\right\}$	q	$\left\{\begin{bmatrix} a & b & c & d \\ 0 & a & b & c \\ 0 & 0 & a & b \\ 0 & 0 & 0 & a \end{bmatrix}\right\}$	$q^3(q-1)$

Table 2. Values of $t(T_{4,i})$ and $c(T_{4,i})$ (*continued*).

7. The proportion of matrices with all eigenvalues in $\mathbb{F}$

Olšavský [2003] noted that as the size of the field $\mathbb{Z}_p$ increases, the proportion of matrices in $M_2(\mathbb{Z}_p)$ with all eigenvalues in $\mathbb{Z}_p$ approaches 1/2, that is,

$$\lim_{p\to\infty} \frac{|\mathrm{Eig}_2(\mathbb{Z}_p)|}{|M_2(\mathbb{Z}_p)|} = \lim_{p\to\infty} \frac{\frac{1}{2}p^4 + p^3 - \frac{1}{2}p^2.}{p^{2^2}} = \frac{1}{2}.$$

We generalize this result to show the proportion of matrices with all eigenvalues in $\mathbb{F}$ for any fixed n approaches $1/n!$ (Note that the leading coefficients for the polynomials $|\mathrm{Eig}_3(\mathbb{F})|$ and $|\mathrm{Eig}_4(\mathbb{F})|$ are $1/3!$ and $1/4!$, respectively, so the generalization is true for these cases).

Theorem 7.1. *Let $\mathbb{F}$ be a finite field with q elements. Then*

$$\lim_{q\to\infty} \frac{|\mathrm{Eig}_n(\mathbb{F})|}{|M_n(\mathbb{F})|} = \frac{1}{n!}.$$

Proof. We know $|M_n(\mathbb{F})| = q^{n^2}$ and our method for determining $|\mathrm{Eig}_n(\mathbb{F})|$ implies that it is a polynomial in the variable q: Denote the double partitions in $\mathrm{DP}(n)$ by

$T_{n,i}$ for $1 \le i \le |\mathrm{DP}(n)|$. For each i, $t(T_{n,i})$ is a polynomial, and since for each $J \in J_n(\mathbb{F})$, $C(J)$ is a subgroup of $\mathrm{GL}_n(\mathbb{F})$, the polynomial $|\mathrm{GL}_n(\mathbb{F})|$ is divisible by $|C(J)|$, and thus $s(T_{n,i})$ is also a polynomial. To evaluate the limit, we must determine the leading coefficient of $|\mathrm{Eig}_n(\mathbb{F})|$. Let $\deg f$ denote the degree (in the variable q) of the polynomial f. Then, since the degree of a sum of a fixed number of polynomials is equal to the maximum degree of the polynomials,

$$\begin{aligned}
\deg |\mathrm{Eig}_n(\mathbb{F})| &= \deg \sum_{i=1}^{|\mathrm{DP}(n)|} t(T_{n,i})s(T_{n,i}) = \max_{1\le i\le |\mathrm{DP}(n)|} \deg(t(T_{n,i})s(T_{n,i})) \\
&= \max_{1\le i\le |\mathrm{DP}(n)|} \deg\left(t(T_{n,i})\frac{|\mathrm{GL}_n(\mathbb{F})|}{c(T_{n,i})}\right) \\
&= \max_{1\le i\le |\mathrm{DP}(n)|} \deg t(T_{n,i}) + \deg |\mathrm{GL}_n(\mathbb{F})| - \deg c(T_{n,i}) \\
&= \max_{1\le i\le |\mathrm{DP}(n)|} n^2 + \deg t(T_{n,i}) - \deg c(T_{n,i}).
\end{aligned}$$

Let $T_{n,1}$ be the double partition $\{\{1\}, \{1\}, \dots, \{1\}\}$, corresponding to the type of Jordan form with all n eigenvalues distinct. For all n, we will show the maximum is attained only for $i = 1$. Indeed, since $\deg t(T_{n,i})$ is equal to the number of distinct eigenvalues in the type, $\deg t(T_{n,1}) = n$, and $\deg t(T_{n,i}) < n$, for all $i > 1$. Furthermore, by (4), $\deg c(T_{n,1}) = n$, and $\deg c(T_{n,i}) \ge n$ for all $i > 1$. Thus $\deg |\mathrm{Eig}_n(\mathbb{F})| = n^2 + \deg t(T_{n,1}) - \deg c(T_{n,1}) = n^2$, and the leading coefficient of $|\mathrm{Eig}_n(\mathbb{F})|$ is equal to the leading coefficient of $t(T_{n,1}) \cdot |\mathrm{GL}_n(\mathbb{F})|/c(T_{n,1})$.

Since $t(T_{n,1}) = \binom{q}{n}$, $|\mathrm{GL}_n(\mathbb{F})| = \prod_{i=0}^{n-1}(q^n - q^i)$, and $c(T_{n,1}) = (q-1)^n$, the leading coefficient of $t(T_{n,i}) \cdot |\mathrm{GL}_n(\mathbb{F})|/c(T_{n,i})$ is $1/n!$. □

Corollary 7.2. *Let* $\mathrm{Ediff}_n(\mathbb{F}) \subseteq \mathrm{Eig}_n(F)$ *denote the set of matrices in* $M_n(\mathbb{F})$ *with all eigenvalues in* $\mathbb{F}$ *and all eigenvalues distinct. Then*

$$\lim_{q\to\infty} \frac{|\mathrm{Ediff}_n(\mathbb{F})|}{|\mathrm{Eig}_n(\mathbb{F})|} = 1.$$

Thus, for large enough finite fields, nearly all matrices with all eigenvalues in the field have all different eigenvalues.

Proof. In the notation of the proof of Theorem 7.1,

$$|\mathrm{Ediff}_n(\mathbb{F})| = t(T_{n,1}) \cdot |\mathrm{GL}_n(\mathbb{F})|/c(T_{n,1}) = q^{n^2}/n! + o(q^{n^2}).$$

Thus $|\mathrm{Ediff}_n(\mathbb{F})|$ and $|\mathrm{Eig}_n(\mathbb{F})|$ are both polynomials in q with the same leading term. So as q increases the ratio of these cardinalities will approach 1. □

8. Conclusions

Given a finite field $\mathbb{F}$ and a positive integer n, we have given a method for computing $|\mathrm{Eig}_n(\mathbb{F})|$ using (3). In 4.1 and (4) we gave the formulas necessary for computing the pieces of (3). In Section 6 we applied our method in the cases $n = 3$ and $n = 4$, and in Section 7, we showed that as the size of the finite field increases, the proportion of matrices in $M_n(\mathbb{F})$ that have all eigenvalues in $\mathbb{F}$ approaches $1/n!$.

There are a number of interesting directions for future research. For example, Theorem 7.1 describes the asymptotic behavior of $|\mathrm{Eig}_n(\mathbb{F})|$ in the case where n is fixed and q goes to infinity. Is it possible to find an analogous statement in the case that q is fixed and n increases? It could also be interesting to find a geometric or combinatorial interpretation for these numbers. In the case $q = 2$, and $n = 1, 2, 3, 4$, $|\mathrm{Eig}_n(\mathbb{Z}_2)| = 2$, 14, 352, and 33,632, respectively, and this sequence (or even a small piece of it) is not found in the *Online Encyclopedia of Integer Sequences*. Finally the polynomials for $|\mathrm{Eig}_n(\mathbb{F})|$ could be studied further. For example, is the smallest nonzero power in $|\mathrm{Eig}_n(\mathbb{F})|$ always $2(n-1)$ and if so, why? It seems likely that the coefficients will always sum to 1, but will they always be between -1 and 1, and if so, why? Are there patterns in the magnitudes or signs of the coefficients?

Acknowledgement

The authors thank Pamela Richardson for suggesting this problem and for helpful conversations.

References

[Cayley 1855] A. Cayley, "Recherches sur les matrices dont les termes sont des fonctions linéaires d'une seule indéterminée", *J. Reine Angew. Math.* **50** (1855), 313–317. Zbl 050.1347cj

[Hungerford 1974] T. W. Hungerford, *Algebra*, Graduate Texts in Mathematics **73**, Holt, Rinehart and Winston, New York, 1974. Reprinted Springer, New York, 2003. MR 50 #6693 Zbl 0293.12001

[Olšavský 2003] G. Olšavský, "The number of 2 by 2 matrices over $\mathbb{Z}/p\mathbb{Z}$ with Eigenvalues in the same field", *Math. Mag.* **76**:4 (2003), 314–317. MR 1573703 Zbl 1057.15020

[Stanley 2012] R. P. Stanley, *Enumerative combinatorics, I*, 2nd ed., Cambridge Studies in Advanced Mathematics **49**, Cambridge University Press, 2012. MR 2868112 Zbl 1247.05003

[Sylvester 1851] J. J. Sylvester, "An enumeration of the contacts of lines and surfaces of the second order", *Phil. Mag.* **1**:2 (1851), 119–140.

Received: 2013-05-08 Revised: 2014-01-31 Accepted: 2014-02-25

kaylorlm@wclive.westminster.edu *Department of Mathematics and Computer Science, Westminster College, 319 South Market Street, New Wilmington, PA 16142, United States*

offnerde@westminster.edu *Department of Mathematics and Computer Science, Westminster College, 319 South Market Street, New Wilmington, PA 16142, United States*

INVOLVE 7:5 (2014)
dx.doi.org/10.2140/involve.2014.7.647

A not-so-simple Lie bracket expansion

Julie Beier and McCabe Olsen

(Communicated by Robert W. Robinson)

Lie algebras and quantum groups are not usually studied by an undergraduate. However, in the study of these structures, there are interesting questions that are easily accessible to an upper-level undergraduate. Here we look at the expansion of a nested set of brackets that appears in relations presented in a paper of Lum on toroidal algebras. We illuminate certain terms that must be in the expansion, providing a partial answer for the closed form.

1. Introduction

Lie algebras and quantum groups are not topics that you are apt to hear undergraduates math majors discussing in their spare time. However, there are a surprising number of nontrivial questions in this area that are undergraduate appropriate. In this paper, we will give a brief overview of the broad mathematical setting, and then discuss an accessible problem that involves expanding a nested set of brackets.

Lie algebras, their universal enveloping algebras and quantum groups are a fundamental part of representation theory that have many applications within mathematics and mathematical physics. Lie algebras and Lie groups were originally discovered by Sophus Lie in the late nineteenth century [Borel 2001]. Given a Lie algebra, we associate a unique associative algebra called the universal enveloping algebra. In 1985, Jimbo and Drinfeld discovered q-analogues of these universal enveloping algebras called "quantum groups", which have been a recent area of study (see [Lusztig 1993]).

In order to find the quantum analogue of a Lie algebra it is often desirable to understand the defining relationships of the Lie algebra inside of its universal enveloping algebra. The motivation for this project came from a paper by Lum in which he gives a nice presentation of a toroidal Lie algebra that could be useful in understanding this Lie algebra's quantum group [Lum 1998]. All of these relations utilize a nested set of brackets called $t(k)$. For simplicity, we have modified $t(k)$ by a scalar. In this paper we seek to understand the expansion of this object.

MSC2010: 17B67.
Keywords: Lie algebra, toroidal algebra.

2. The Lie bracket $t(k)$

Recall that a *Lie algebra* is defined as a vector space L over a field F that is equipped with a bilinear map $L \times L \to L$, known as a *Lie bracket*, satisfying certain conditions. The Lie bracket $(x, y) \to [x, y]$ for all $x, y \in L$ must satisfy the *alternating property*, namely

$$[x, x] = 0$$

and the *Jacobi identity*, an analog of associativity:

$$\big[x, [y, z]\big] + \big[y, [z, x]\big] + \big[z, [x, y]\big] = 0,$$

for all $x, y, z \in L$.

Example 2.1. The set $gl(n, \mathbb{R})$ of n-by-n matrices with real entries, together with the operation defined by defined by $[A, B] := AB - BA$, is a Lie algebra. To see this, consider matrices $A, B, C \in gl(n, \mathbb{R})$. It is easy to show that the bracket is bilinear. Since $[A, A] = AA - AA = 0$ alternation is satisfied. To verify that the Jacobi identity holds, note that

$$\begin{aligned}
&\big[A, [B, C]\big] + \big[B, [C, A]\big] + \big[C, [A, B]\big] \\
&\qquad = [A, BC - CB] + [B, CA - AC] + [C, AB - BA] \\
&\qquad = A(BC - CB) - (BC - CB)A + B(CA - AC) \\
&\qquad\quad - (CA - AC)B + C(AB - BA) - (AB - BA)C \\
&\qquad = ABC - ACB - BCA + CBA + BCA - BAC \\
&\qquad\quad - CAB + ACB + CAB - CBA - ABC + BAC \\
&\qquad = 0.
\end{aligned}$$

The nested set of brackets which we seek to understand is denoted $t(k)$. They are defined recursively as

$$t(1) := [x, y] = xy - yx, \quad t(k) := \underbrace{\big[\ldots[[[x, y], x], y]\ldots\big]}_{k\ xy\text{-pairs}}$$

Example 2.2. The case of $k = 3$ is given as follows:

$$\begin{aligned}
t(3) &= \big[[[[[x, y], x], y], x], y\big] \\
&= \big[[[[xy - yx, x\big], y], x], y\big] \\
&= \big[[2xyxy - 2yxyx + y^2x^2 - x^2y^2, x], y\big] \\
&= 4xyxyxy - 4yxyxyx + 2y^2xyx^2 - 2x^2yxy^2 + 2yx^2yxy - 2yxyx^2y \\
&\qquad + y^2x^3y - yx^3y^2 + yx^2y^2x - xy^2x^2y + yxy^2x^2 - x^2y^2xy + x^3y^3 - y^3x^3.
\end{aligned}$$

We choose to view the output of $t(k)$ as *words*. Unlike the combinatorial definition of words, we include the coefficient. For example, $t(1)$ consists of two words, namely xy and $-yx$. We know some additional properties of words because of how the bracket functions. No word can begin and end with an x nor can a word begin and end with y^2.

We define the *antiword* to be the associated word with reverse ordering of x's and y's, opposite sign, and same coefficient. In the example of $t(1)$, the antiword of xy is $-yx$. Similarly, $-yx$ has antiword xy. For a more interesting example, consider Example 2.2. Notice that each word is written next to its antiword. We have $4xyxyxy$ followed by $-4yxyxyx$ which is a word-antiword pair, $2y^2xyx^2$ followed by $-2x^2yxy^2$, and so on. This observation works in general and cuts the problem in half.

Theorem 2.3. *Every word in $t(k)$ has an antiword in $t(k)$.*

Proof. For the base case $k = 1$, we have $t(1) := [x, y] = xy - yx$, so the statement is clearly valid. Now assume for some integer $k \geq 1$ every word appears in $t(k)$ together with its antiword. We want to show that each word in the $t(k+1)$ has an antiword in $t(k+1)$. So consider an arbitrary word ω and its antiword $\overline{\omega}$ in the k-th iteration. By bracket expansion, we have

$$\begin{aligned}\big[[\omega+\overline{\omega}, x], y\big] &= [\omega x + \overline{\omega} x - x\omega - x\overline{\omega}, y] \\ &= \omega xy + \overline{\omega} xy - x\omega y - x\overline{\omega} y - (y\omega x + y\overline{\omega} x - yx\omega - yx\overline{\omega}) \\ &= \omega xy + \overline{\omega} xy - x\omega y - x\overline{\omega} y - y\omega x - y\overline{\omega} x + yx\omega + yx\overline{\omega}.\end{aligned}$$

Since ω and $\overline{\omega}$ are a word-antiword pair, the following are word-antiword pairs: ωxy and $yx\overline{\omega}$, $-x\omega y$ and $-y\overline{\omega} x$, $-y\omega x$ and $-x\overline{\omega} y$, and $yx\omega$ and $\overline{\omega} xy$. Each of these words will have the same coefficient as ω and $\overline{\omega}$. If two of these words in $t(k+1)$ are the same, the words in $t(k)$ that generated them have corresponding antiwords in $t(k)$. Bracketing these will necessarily give the same antiword in $t(k+1)$ causing coefficients to be preserved. Thus, while a whole pair may cancel, no word can independently disappear. □

From this result, we know that it is not possible to have any symmetric words in the output of an arbitrary $t(k)$.

3. Word patterns

Given our goal to determine the content of $t(k)$ in the universal enveloping algebra, we first look to locate patterns universal to all $t(k)$. Here we prove the existence of several such patterns of words. First we consider two fundamental lemmas.

Lemma 3.1. *If the word $x^k y^k$ exists in $t(k)$, it must be generated by the word $x^{k-1}y^{k-1}$ in $t(k-1)$. Similarly, if the word $y^k x^k$ exists in $t(k)$, it must be generated by the word $y^{k-1}x^{k-1}$ in $t(k-1)$.*

Proof. Assume the word $x^k y^k$ exists in $t(k)$. By the definition of the bracket, in order to arrive at this word, we must multiply some word in $t(k-1)$ by both an x and a y. Working backwards, we remove a y and an x in all possible ways to obtain possible root words for $x^k y^k$. Our only option is to remove a y from the end and an x from the beginning. Therefore our only root word is $x^{k-1}y^{k-1}$. Showing that $y^k x^k$ is only generated by the root word $y^{k-1}x^{k-1}$ is analogous. □

The lemma below follows in an identical fashion.

Lemma 3.2. *If the word $(xy)^k$ exists in $t(k)$, it must be generated by $(xy)^{k-1}$ or $(yx)^{k-1}$ in $t(k-1)$. Similarly, if the word $(yx)^k$ exists in $t(k)$ it must be generated by $(xy)^{k-1}$ or $(yx)^{k-1}$ in $t(k-1)$.*

The proceeding propositions use these lemmas to show some universal patterns appearing in $t(k)$ for all k.

Proposition 3.3. *The words $(-1)^{k+1}x^k y^k$ and $(-1)^k y^k x^k$ appear in $t(k)$.*

Proof. These words appear in the case of $k = 1$ since

$$[x, y] = xy - yx = (-1)^2 xy + (-1)yx.$$

Assume that for some integer $k \geq 1$, we have the words $(-1)^{k+1}x^k y^k + (-1)^k x^k y^k$. We now show that the words $(-1)^{k+2}x^{k+1}y^{k+1}$ and $(-1)^{k+1}x^{k+1}y^{k+1}$ appear in $t(k+1)$. By the definition of the bracket, we have

$$\begin{aligned}
&\big[[(-1)^{k+1}x^k y^k, x], y\big] \\
&\quad = \big[(-1)^{k+1}x^{k+1}y^k - (-1)^{k+1}x^k y^k x, y\big] \\
&\quad = \big[(-1)^{k+1}x^{k+1}y^k + (-1)^{k+2}x^k y^k x, y\big] \\
&\quad = (-1)^{k+1}yx^{k+1}y^k + (-1)^{k+2}yx^k y^k x - \big((-1)^{k+1}x^{k+1}y^{k+1} + (-1)^{k+2}x^k y^k xy\big) \\
&\quad = (-1)^{k+1}yx^{k+1}y^k + (-1)^{k+2}yx^k y^k x + (-1)^{k+2}x^{k+1}y^{k+1} + (-1)^{k+3}x^k y^k xy.
\end{aligned}$$

The word $(-1)^{k+2}x^{k+1}y^{k+1}$ appears as desired. It is an identical process to prove the existence of $(-1)^{k+1}y^{k+1}x^{k+1}$. Furthermore, we know from Lemma 3.1 that $x^{k+1}y^{k+1}$ and $y^{k+1}x^{k+1}$ cannot be generated by any other root words. Therefore the coefficient is as given. □

Using this same technique we find two more words that appear in $t(k)$.

Proposition 3.4. *The words $2^{k-1}(xy)^k$ and $-2^{k-1}(yx)^k$ appear in $t(k)$.*

4. More general recurring words

We now look to find broader patterns of words which necessarily appear in $t(k)$. Similar to before, we need a foundational lemma.

Lemma 4.1. *For $k \geq 1$ and $2 \leq j \leq k$, the word $x^j(yx)^{k-j}y^j$, if it exists in $t(k)$, can only be generated by the word $x^{j-1}(yx)^{((k-1)-(j-1))}y^{j-1}$ in $t(k-1)$ and the word $y^j(xy)^{k-j}x^j$, if it exists in $t(k)$, can only be generated by the word $y^{j-1}(xy)^{((k-1)-(j-1))}x^{j-1}$ in $t(k-1)$.*

In order to prove this lemma, we use similar techniques to that of the previous lemmas. We begin by assuming that the words appear in the k-th iteration of the bracket and we work backwards to determine possible root words. This relatively simple procedure is all that is needed to show that the lemma holds. Using this lemma, we now expand the notions of Proposition 3.3 and Proposition 3.4.

Theorem 4.2. *For $k \geq 1$ and $1 \leq j \leq k$, the word $-\tau_k^j x^j(yx)^{k-j}y^j + \tau_k^j y^j(xy)^{k-j}x^j$ appears in $t(k)$, where we have set*

$$\tau_k^j := (-1)^j 2^{k-j}.$$

To prove this more encompassing theorem, we use double induction. We know that this theorem holds for the base case $k = 1$ and $j = 1$

$$[x, y] = xy - yx = (-1)^2 2^0 x(yx)^0 y + (-1)^1 2^0 y(xy)^0 x$$

and by Proposition 3.4, we know the statement holds for arbitrary k and $j = 1$. Subsequently, we use this as a starting point for the second induction. Simply use a bracket argument similar to the one in Proposition 3.3. This argument yields all of the desired words except in the case of $j = k$. However, Proposition 3.3 already accounts for this case. Therefore, the statement is satisfied.

Returning to our running example of $k = 3$, notice that Theorem 4.2 asserts the existence of the following words: $4xyxyxy$, $-4yxyxyx$, $2y^2xyx^2$, $-2x^2yxy^2$, x^3y^3, and $-y^3x^3$. In Example 2.2, we see that all of these do indeed appear in $t(3)$.

This collection of words accounts for a share of the words in $t(k)$. Unfortunately, it does not even account for all of the words in the case of $k = 3$. However, repeated bracketing of words in Theorem 4.2 will result in more words that are always present. We leave showing the following corollary by bracket as an exercise.

Corollary 4.3. *For $k \geq 1$ and $1 \leq j \leq k$, the following sum appears in $t(k+1)$:*

$$\begin{aligned}
&-\tau_k^j x^j(yx)^{k-j}y^j xy + \tau_k^j y^j(xy)^{k-j}x^{j+1}y + \tau_k^j x^{j+1}(yx)^{k-j}y^{j+1} \\
&-\tau_k^j xy^j(xy)^{k-j}x^j y + \tau_k^j yx^j(yx)^{k-j}y^j x - \tau_k^j y^{j+1}(xy)^{k-j}x^{j+1} \\
&-\tau_k^j yx^{j+1}(yx)^{k-j}y^j + \tau_k^j yxy^j(xy)^{k-j}x^j.
\end{aligned}$$

Indeed, the words from Corollary 4.3 actually include all of the words in Theorem 4.2 as shown below.

Proposition 4.4. *All words in $t(k+1)$ of the form $-\tau_k^{j+1}x^{j+1}(yx)^{k-(j+1)}y^{j+1}$ and $\tau_k^{j+1}y^{j+1}(xy)^{k-(j+1)}x^{j+1}$ can be expressed by a form given in Corollary 4.3.*

Proof. Consider the word in $t(k+1)$ generated by Theorem 4.2 given by

$$(-1)^{(j+1)+1}2^{(k+1)-(j+1)}x^{j+1}(yx)^{(k+1)-(j+1)}y^{j+1} \;=\; \tau_k^j x^{j+1}(yx)^{k-j}y^{j+1}.$$

This word is also a word of the form given in Corollary 4.3. Furthermore, consider the other word in $t(k+1)$ generated by Theorem 4.2:

$$(-1)^{j+1}2^{(k+1)-(j+1)}y^{j+1}(xy)^{(k+1)-(j+1)}x^{j+1} \;=\; -\tau_k^j y^{j+1}(xy)^{k-j}x^{j+1}$$

which is indeed a word of the desired form. These two general words account for all words of the form given by Theorem 4.2 in $t(k+1)$ except for the case of $j=1$.

First consider $(-1)^{j+1}2^{(k+1)-j}x^j(yx)^{(k+1)-j}y^j$ generated by Theorem 4.2 evaluated at $j=1$. This yields

$$(-1)^2 2^k x(yx)^k y = 2^k(xy)^{k+1} = 2^{k-1}x(yx)^{k-1}yxy + 2^{k-1}xy(xy)^{k-1}xy$$

which are two words in Corollary 4.3 evaluated at $j=1$, namely $-\tau_k^j x^j(yx)^{k-j}y^j xy$ and $-\tau_k^j xy^j(xy)^{k-j}x^j y$. The proof that $(-1)^j 2^{(k+1)-j}y^j(xy)^{(k+1)-j}x^j$ can be expressed in a desired form when $j=1$ is identical. □

Using Proposition 4.4, we account for all of the words in $t(1)$, $t(2)$, and $t(3)$. We leave showing that Corollary 4.3 produces all of $t(3)$ as an exercise. Moreover, we believe that we can identify an even larger pattern of words.

As seen in previous cases, we first identify how the particular words can be generated.

Proposition 4.5. *If the words in the left column of the table below exist in $t(k)$, they must be generated by the corresponding root word listed on the right.*

Generated word in $t(k)$	*Root word in $t(k-1)$*
$y^m x^j(yx)^{k-(j+m)}y^j x^m$	$y^{m-1}x^j(yx)^{(k-1)-(j+(m-1))}y^j x^{m-1}$
$x^m y^j(xy)^{k-(j+m)}x^j y^m$	$x^{m-1}y^j(xy)^{k-(j+(m-1))}x^j y^{m-1}$
$x^j(yx)^{k-(j+m)}y^j(xy)^m$	$x^j(yx)^{k-(j+(m-1))}y^j(xy)^{m-1}$
$(yx)^m y^j(yx)^{k-(j+m)}x^j$	$(yx)^{m-1}y^j(yx)^{k-(j+(m-1))}x^j$
$(yx)^m x^j(yx)^{k-(j+m)}y^j$	$(yx)^{m-1}x^j(yx)^{k-(j+(m-1))}y^j$
$y^j(xy)^{k-(j+m)}x^j(xy)^m$	$y^j(xy)^{k-(j+(m-1))}x^j(xy)^{m-1}$

Despite the larger number of words in question, the proof of each follows in the same manner as our previous proofs for necessary root words, except in the more

complicated case of $m = 1$. In this instance, there are more ways to remove one x and y. However, it can be shown that some of these violate the properties of words and thus do not exist in $t(k+1)$.

Building from all of our previous work we present our largest list of necessary words.

Theorem 4.6. *Let $k \geq 1$.*

- *If $k \geq 2$, $j = 1$, and $m = 1$, then*
$$\tau_k^{j+m}(yx)^m x^j (yx)^{k-(j+m)} y^j - \tau_k^{j+m} y^j (xy)^{k-(j+m)} x^j (xy)^m$$
appears in $t(k)$.
- *If $1 \leq j \leq k$ and $m = 0$, then*
$$-\tau_k^{j+m} y^m x^j (yx)^{k-(j+m)} y^j x^m + \tau_k^{j+m} x^m y^j (xy)^{k-(j+m)} x^j y^m$$
appears in $t(k)$.
- *If $m \geq 1$, $j \geq 2$, and $j + m \leq k$, then*
$$\begin{aligned}&-\tau_k^{j+m} y^m x^j (yx)^{k-(j+m)} y^j x^m + \tau_k^{j+m} x^m y^j (xy)^{k-(j+m)} x^j y^m \\ &\quad + (-1)^m \big(-\tau_k^{j+m} x^j (yx)^{k-(j+m)} y^j (xy)^m + \tau_k^{j+m} (yx)^m y^j (yx)^{k-(j+m)} x^j \\ &\qquad - \tau_k^{j+m} (yx)^m x^j (yx)^{k-(j+m)} y^j + \tau_k^{j+m} y^j (xy)^{k-(j+m)} x^j (xy)^m\big)\end{aligned}$$
appears in $t(k)$.

Proof. Proof of the $j = 1$, $m = 1$ case follows directly from Corollary 4.3 by evaluating $-\tau_k^j yx^{j+1}(yx)^{k-j} y^j$ at $j = 1$.

Now consider the case of $m = 0$. We have
$$(-1)^{j+0+1} 2^{k-(j+0)} y^0 x^j (yx)^{k-(j+0)} y^j x^0 = -\tau_k^j x^j (yx)^{k-j} y^j.$$
This word was shown to exist in $t(k)$ by Theorem 4.2. For the same reason, we know that $\tau_k^{j+m} x^m y^j (xy)^{k-(j+m)} x^j y^m$ exists in $t(k)$ when $m = 0$.

Now we show $-\tau_k^{j+m} y^m x^j (yx)^{k-(j+m)} y^j x^m$ and $\tau_k^{j+m} x^m y^j (xy)^{k-(j+m)} x^j y^m$ appear in $t(k)$ if $j \geq 2$ and $3 \leq m + j \leq k$. We just argued the case of $m = 0$ for arbitrary $1 \leq j \leq k$ for all $t(k)$. So, we perform induction on m. In Corollary 4.3, we bracket $-\tau_k^j x^j (yx)^{k-j} y^j + (-1)^j 2^{k-j} y^j (xy)^{k-j} x^j$ in $t(k)$ with $k \geq 2$ to generate the term $(-1)^{j+2} 2^{k-j} yx^j (yx)^{k-j} y^j x - \tau_k^j xy^j (xy)^{k-j} x^j y$ which is the desired term for $m = 1$ in $t(k+1)$.

Now, assume that $k \geq 3$ and that for some $m \geq 1$ with $m + j \leq k$, the words $-\tau_k^{j+m} y^m x^j (yx)^{k-(j+m)} y^j x^m + \tau_k^{j+m} x^m y^j (xy)^{k-(j+m)} x^j y^m$ appear in $t(k)$. We want to show that
$$(-1)^{j+(m+1)+1} 2^{k-(j+(m+1))} y^{m+1} x^j (yx)^{k-(j+(m+1))} y^j x^{m+1}$$

$$+(-1)^{j+(m+1)}2^{(k+1)-(j+(m+1))}x^{m+1}y^j(xy)^{(k+1)-(j+(m+1))}x^jy^{m+1}$$

appears in $t(k+1)$. Using our bracket, we have

$$\begin{aligned}
&\big[[-\tau_k^{j+m}y^mx^j(yx)^{k-(j+m)}y^jx^m+\tau_k^{j+m}x^my^j(xy)^{k-(j+m)}x^jy^m,x],y\big]\\
&\quad=\big[-\tau_k^{j+m}y^mx^j(yx)^{k-(j+m)}y^jx^{m+1}+\tau_k^{j+m}x^my^j(xy)^{k-(j+m)}x^jy^mx\\
&\qquad+\tau_k^{j+m}xy^mx^j(yx)^{k-(j+m)}y^jx^m-\tau_k^{j+m}x^{m+1}y^j(xy)^{k-(j+m)}x^jy^m,y\big]\\
&\quad=-\tau_k^{j+m}y^mx^j(yx)^{k-(j+m)}y^jx^{m+1}y+\tau_k^{j+m}x^my^j(xy)^{k-(j+m)}x^jy^mxy\\
&\qquad+\tau_k^{j+m}xy^mx^j(yx)^{k-(j+m)}y^jx^my-\tau_k^{j+m}x^{m+1}y^j(xy)^{k-(j+m)}x^jy^{m+1}\\
&\qquad+\tau_k^{j+m}y^{m+1}x^j(yx)^{k-(j+m)}y^jx^{m+1}+\tau_k^{j+m}yx^my^j(xy)^{k-(j+m)}x^jy^mx\\
&\qquad-\tau_k^{j+m}yxy^mx^j(yx)^{k-(j+m)}y^jx^m+\tau_k^{j+m}yx^{m+1}y^j(xy)^{k-(j+m)}x^jy^m.
\end{aligned}$$

We have the resulting words (line 3 word 1 and line 2 word 2)

$$\begin{aligned}
&\tau_k^{j+m}y^{m+1}x^j(yx)^{k-(j+m)}y^jx^{m+1}\\
&\qquad=(-1)^{j+(m+1)+1}2^{(k+1)-(j+(m+1))}y^{m+1}x^j(yx)^{(k+1)-(j+(m+1))}y^jx^{m+1}
\end{aligned}$$

and

$$\begin{aligned}
&-\tau_k^{j+m}x^{m+1}y^j(xy)^{k-(j+m)}x^jy^{m+1}\\
&\qquad=(-1)^{j+(m+1)}2^{(k+1)-(j+(m+1))}x^{m+1}y^j(xy)^{(k+1)-(j+(m+1))}x^jy^{m+1}.
\end{aligned}$$

By Proposition 4.5, we know that these words cannot be generated by any other root word. The other four remaining desired words can be shown through an analogous process. □

5. Moving forward

We could consider continuing our current course of action by looking for new patterns beginning in the $k=4,5,6$ cases to try to detect another significant margin of words. One difficulty with this avenue is that an entirely new class of words appears every few cases. A second difficulty is that these become time consuming for the computer to compute. Maple 15 was unable to compute these brackets at $t(8)$ after a full day of computation for $t(7)$. It appears that every time this version of Maple encounters a noncommuting term like xyx it computes $x*y*x$. However, Sage (sagemath.org) treats xyx as an element and can compute the values much faster. Despite this, at $t(11)$ it starts to take minutes for the computation to occur, and it is expected that even using SAGE the computational time would be too high before reaching $t(20)$.

The reader may be wondering why we have taken this particular approach to the problem. The answer is quite simple. We have been unable to find a nice

k	word count	k	word count	k	word count
4	46	7	1648	10	61512
5	152	8	5506	11	206028
6	500	9	18380	12	691126

Table 1. Number of words in $t(k)$.

combinatorial method to simplify the problem. The number of terms in $t(k)$ grows rapidly; see Table 1. Our initial use of dominoes, strips, and tableaux illustrated interesting connections but did not yield useful results. Then we used the Online Encyclopedia of Integer Sequences (oeis.org) to try and find connections to other less obvious options. However, despite searching a number of related sequences, we were unable to locate any connections. It would be ideal if one could find such a connection in order to continue this problem.

This problem is thus still open, as is the question of expanding the full relations given in Lum's paper. We encourage readers to improve on our method and find connections to solve these problems. After this is done, it will be possible to give a nice presentation of the toroidal quantum group.

References

[Borel 2001] A. Borel, *Essays in the history of Lie groups and algebraic groups*, History of Mathematics **21**, American Math. Soc. and London Math. Soc., 2001. MR 2002g:01010 Zbl 1087.01011

[Lum 1998] K. H. Lum, "A presentation of toroidal algebras as homomorphic images of G.I.M. algebras", *Comm. Algebra* **26**:12 (1998), 4051–4063. MR 99j:17033

[Lusztig 1993] G. Lusztig, *Introduction to quantum groups*, Progress in Mathematics **110**, Birkhäuser, Boston, 1993. MR 94m:17016 Zbl 0788.17010

Received: 2013-05-14 Revised: 2014-02-03 Accepted: 2014-03-03

beierju@earlham.edu *Department of Mathematics, Earlham College, 801 National Road West, Drawer 138, Richmond, IN 47374, United States*

mccabe.olsen@gmail.com *Mercer University, Department of Mathematics, 1400 Coleman Avenue, Macon, GA 31207, United States*

INVOLVE 7:5 (2014)
dx.doi.org/10.2140/involve.2014.7.657

On the omega values of generators of embedding dimension-three numerical monoids generated by an interval

Scott T. Chapman, Walter Puckett and Katy Shour

(Communicated by Vadim Ponomarenko)

We offer a formula to compute the omega values of the generators of the numerical monoid $S = \langle k, k+1, k+2 \rangle$ where k is a positive integer greater than 2.

1. Introduction and the main result

The notion of a prime element is a central focus in the study of algebra and number theory. Several recent papers [Anderson and Chapman 2010; 2012; Anderson et al. 2011] have considered the following generalization of the notion of prime elements in the context of numerical monoids. This definition, which we state for a general commutative cancellative monoid, originally appeared in [Geroldinger and Hassler 2008].

Definition 1.1. Let M be a commutative, cancellative, atomic monoid with set of units $M^\times$ and set of irreducibles (or atoms) $\mathscr{A}(M)$. For $x \in M \setminus M^\times$, we define $\omega_M(x) = n$ if n is the smallest positive integer with the property that whenever $x \mid a_1 \cdots a_t$, where each $a_i \in \mathscr{A}(M)$, there is a $T \subseteq \{1, 2, \ldots, t\}$ with $|T| \leq n$ such that $x \mid \prod_{k \in T} a_k$. If no such n exists, then $\omega_M(x) = \infty$. For $x \in M^\times$, we define $\omega_M(x) = 0$.

As in [Anderson et al. 2011], when our context is clear, we will shorten $\omega_M(x)$ to $\omega(x)$. It follows easily from the definition that an element $x \in M \setminus M^\times$ is prime if and only if $\omega(x) = 1$. Hence, in some sense the omega function measures how far an element is from being prime. Some basic properties of this function can be found not only in the papers mentioned above, but also in [Geroldinger and Halter-Koch 2006]. Anderson and Chapman [2010; 2012] study the behavior of the omega function in the setting of the multiplicative monoid of a commutative ring.

MSC2010: 13A05, 13F15, 20M14.
Keywords: numerical monoid, omega function, factorizations.
The authors were supported by a 2012 Enhancement Research Grant from Sam Houston State University.

Anderson, Chapman, Kaplan and Torkornoo [Anderson et al. 2011, Section 3] offer a finite time algorithm for computing $\omega(x)$ when x is an element in a numerical monoid S. Recall that a numerical monoid is an additive submonoid of the nonnegative integers (which we denote by $\mathbb{N}_0$). Using elementary number theory, it is easy to show that such a submonoid is finitely generated and possesses a unique minimal (in terms of cardinality) generating set. If $n_1, n_2, \dots, n_t$ is the minimal generating set for a numerical monoid S, then we write

$$S = \langle n_1, \dots, n_t \rangle = \{x_1 n_1 + \cdots + x_t n_t \mid x_i \in \mathbb{N}_0 \text{ for each } i\}.$$

The value t is known as the *embedding dimension of* S. The elements $n_1, \dots, n_t$ are the irreducibles of S, and as noted in Definition 1.1, we will write $\mathscr{A}(S) = \{n_1, \dots, n_t\}$. When considering the complete class of numerical monoids, elementary isomorphism arguments allow us to reduce to the case where $\gcd(n_1, \dots, n_t) = 1$. Such a numerical monoid is called *primitive*. [Rosales and García-Sánchez 2009] is a good general reference on numerical monoids and semigroups. [Bowles et al. 2006; Chapman et al. 2006; 2009; Omidali 2012] examine factorization properties of numerical monoids which are related in various ways to the omega function.

A version of the algorithm in [Anderson et al. 2011] mentioned above has been programmed and can be found in the numerical semigroups package available for Gap (gap-system.org/Manuals/pkg/numericalsgps/doc/manual.pdf). Using data generated by this program, much of the work in [Anderson et al. 2011] is dedicated to showing that closed forms for particular values of $\omega(x)$ are highly nontrivial to determine. In [Anderson et al. 2011, Propositions 3.1 and 3.2], the authors determine formulas for this when $S = \langle n, n+1, \dots, 2n-1 \rangle$ and $S = \langle n, n+1, \dots, 2n-2 \rangle$ (where $n \geq 2$), and in [Anderson et al. 2011, Theorem 4.4] they handle the case where $S = \langle n_1, n_2 \rangle$. The paper also takes interest in computing the values $\omega(n_1)$, $\omega(n_2)$, and $\omega(n_3)$ when $S = \langle n_1, n_2, n_3 \rangle$ is of embedding dimension 3. In particular, they offer a chart [Anderson et al. 2011, p. 101] to illustrate how these omega values can differ. We include a modified form in Table 1.

There are 5 possibilities that Table 1 omits. With the programs then available, Anderson et al. [2011] were unable to find examples of these missing orderings. With some improved programming techniques, the present authors were able to compute $\omega(n_1)$, $\omega(n_2)$ and $\omega(n_3)$ for all embedding dimension-three numerical monoids with generators less than or equal to 100. This yielded two of the remaining five cases.

(i) $S = \langle 6, 7, 9 \rangle$ yields $\omega(6) = 3$, $\omega(7) = 5$, and $\omega(9) = 3$. Hence, $\omega(6) < \omega(7)$, $\omega(9) < \omega(7)$, and $\omega(6) = \omega(9)$.

(ii) $S = \langle 7, 8, 20 \rangle$ yields $\omega(7) = 6$, $\omega(8) = 4$, and $\omega(20) = 5$. Hence, $\omega(7) > \omega(8)$, $\omega(8) < \omega(20)$, and $\omega(7) > \omega(20)$.

$\langle n_1, n_2, n_3 \rangle$	$\omega(n_1)$	$\omega(n_2)$	$\omega(n_3)$	Ordering of the omega values
$\langle 6, 8, 13 \rangle$	3	4	7	$\omega(6)<\omega(8)<\omega(13)$
$\langle 5, 7, 11 \rangle$	3	5	5	$\omega(5)<\omega(7)=\omega(11)$
$\langle 4, 5, 6 \rangle$	2	4	3	$\omega(4)<\omega(5),\ \omega(5)>\omega(6),\ \omega(4)<\omega(6)$
$\langle 6, 9, 11 \rangle$	3	3	7	$\omega(6)=\omega(9)<\omega(11)$
$\langle 7, 11, 17 \rangle$	5	5	5	$\omega(7)=\omega(11)=\omega(17)$
$\langle 6, 7, 11 \rangle$	4	3	5	$\omega(6)>\omega(7),\ \omega(7)<\omega(11),\ \omega(6)<\omega(11)$
$\langle 7, 8, 12 \rangle$	5	4	4	$\omega(7)>\omega(8)=\omega(12)$
$\langle 7, 8, 13 \rangle$	5	4	5	$\omega(7)>\omega(8),\ \omega(8)<\omega(13),\ \omega(7)=\omega(13)$

Table 1. Differing values of omega (modified from [Anderson et al. 2011]).

We strongly suspect the final three orderings are not possible. Hence, we state this as a potential problem.

Problem. Let $S = \langle n_1, n_2, n_3 \rangle$ be an embedding dimension-3 numerical monoid. Show that the sequence $\omega(n_1)$, $\omega(n_2)$, and $\omega(n_3)$ does not satisfy any of the following three orderings:

- $\omega(n_1) > \omega(n_2) > \omega(n_3)$.
- $\omega(n_1) = \omega(n_2) > \omega(n_3)$.
- $\omega(n_1) < \omega(n_2),\ \omega(n_2) > \omega(n_3),\ \omega(n_3) < \omega(n_1)$.

In the course of attempting to solve this problem, numerous classes of embedding dimension-3 numerical monoids were studied. We encountered one with especially nice omega values on the generators. The remainder of this paper will consist of a proof of the following theorem.

Theorem 1.2. *Let k be a positive integer.*

(a) *If $S_1 = \langle 2k+1, 2k+2, 2k+3 \rangle$, then*

$$\omega(2k+1) = k+1 \quad \textit{and} \quad \omega(2k+2) = \omega(2k+3) = k+2.$$

(b) *If $k \geq 2$ and $S_2 = \langle 2k, 2k+1, 2k+2 \rangle$, then*

$$\omega(2k) = k,\ \omega(2k+1) = k+2 \quad \textit{and} \quad \omega(2k+2) = k+1.$$

The proof will require two results from the literature. The first allows one to reduce the definition of $\omega(x)$ from that of checking arbitrary products to checking only products of irreducibles.

Theorem 1.3 [Anderson and Chapman 2010, Theorem 2.1]. *Let M be a commutative cancellative monoid and suppose that $x \in M \backslash M^{\times}$. Then the following statements are equivalent*:

(a) $\omega(x) = m \in \mathbb{N}$.

(b) *m is the least positive integer such that if $x \mid x_1 \cdots x_n$ with each $x_i \in M$ irreducible, then $x \mid x_{i_1} \cdots x_{i_t}$ for some $\{i_1, \ldots, i_t\} \subseteq \{1, \ldots, n\}$ with $t \le m$.*

(c) *If $x \mid x_1 \cdots x_n$ with each $x_i \in M$ irreducible and $n \ge m$, then $x \mid x_{i_1} \cdots x_{i_m}$ for some $\{i_1, \ldots, i_m\} \subseteq \{1, \ldots, n\}$, and there are irreducible $x_1, \ldots, x_m \in M$ such that $x \mid x_1 \cdots x_m$, but x divides no proper subproduct of the x_i.*

For an element $x \in S$, the product $x_1 \cdots x_m$ alluded to in part (c) above will be called a *bullet* for x.

The second necessary result is an amazing characterization of the membership problem for a numerical monoid generated by an interval of integers.

Theorem 1.4 [García-Sánchez and Rosales 1999, Corollary 2]. *An element $n \in \mathbb{N}$ belongs to $S = \langle a, a+1, \ldots, a+x \rangle$ if and only if*

$$n \pmod{a} \le \left\lfloor \frac{n}{a} \right\rfloor x,$$

where $\lfloor \cdot \rfloor$ represents the greatest integer function and residues are assumed to be least.

To prove Theorem 1.2, we will verify the 6 claimed values of the omega function. To do this, we will pivot on Theorem 1.3(c) and produce a bullet for each of the six elements. The condition in Theorem 1.4 will be vital in these arguments. In the two monoids we consider, the condition will reduce to

$$n \pmod{2k+1} \le \left\lfloor \frac{n}{2k+1} \right\rfloor 2$$

for $S_1 = \langle 2k+1, 2k+2, 2k+3 \rangle$ and

$$n \pmod{2k} \le \left\lfloor \frac{n}{2k} \right\rfloor 2$$

for $S_2 = \langle 2k, 2k+1, 2k+2 \rangle$. To finish the proof, we will then verify the first part of Theorem 1.3(c); namely if the bullet is of length j, then divisibility by a sum of length greater than or equal to j yields divisibility by a subsum of length j or less.

2. Proof of Theorem 1.2 for S_1

Lemma 2.1. *In S_1 we have the following divisibility relationships*:

(a) $(2k+1) \mid \sum_{i=1}^{k+1} (2k+3)$;

(b) $(2k+2) \mid \sum_{i=1}^{k+2} (2k+1)$;

(c) $(2k+3) \mid \sum_{i=1}^{k+2} (2k+1)$.

Proof. (a) Now, $\sum_{i=1}^{k+1}(2k+3) = (k+1)(2k+3) = 2k^2+5k+3$. To prove the claim, we must show that $(2k^2+5k+3)-(2k+1) \in S_1$. Now (a) follows since $(2k^2+5k+3)-(2k+1) = 2k^2+3k+2 = k(2k+1)+(2k+2)$.

For the proof of (b) and (c), note that $\sum_{i=1}^{k+2}(2k+1) = (k+2)(2k+1) = 2k^2+5k+2$. For (b), we must show that $(2k^2+5k+2)-(2k+2) \in S_1$. Since

$$(2k^2+5k+2)-(2k+2) = 2k^2+3k = k(2k+3),$$

part (b) follows. For (c), we must show that $(2k^2+5k+2)-(2k+3) \in S_1$. Since $(2k^2+5k+2)-(2k+3) = 2k^2+3k-1 = 2k+2+(k-1)(2k+3)$, part (c) follows and the proof of the lemma is complete. □

In the next three lemmas, we show that the sums produced in Lemma 2.1 are actually bullets for $2k+1$, $2k+2$ and $2k+3$.

Lemma 2.2. *In S_1, $2k+1$ does not divide any proper subsum of* $\sum_{i=1}^{k+1}(2k+3)$.

Proof. To prove the claim, we must show that $2k+1$ does not divide $j(2k+3)$ for $1 \le j \le k$. This is equivalent to showing that

$$j(2k+3)-(2k+1) \notin S_1,$$

for each $1 \le j \le k$. Using Theorem 1.4, we must show that

$$j(2k+3)-(2k+1) \pmod{2k+1} > \left\lfloor \frac{j(2k+3)-(2k+1)}{2k+1} \right\rfloor \cdot 2, \tag{1}$$

for each $1 \le j \le k$. Now, (1) reduces to

$$2j > \left\lfloor \frac{j(2k+3)-(2k+1)}{2k+1} \right\rfloor 2, \tag{2}$$

and hence

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

Equation (3) can be rewritten as

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

Since $j \le k$, we have

$$\frac{2j}{2k+1} \le \frac{2k}{2k+1} < \frac{2k+1}{2k+1} = 1,$$

so $\lfloor 2j/(2k+1) \rfloor = 0$ and Equation (3) is true, which completes the proof. □

Lemma 2.3. *In S_1, $2k+2$ does not divide any proper subsum of $\sum_{i=1}^{k+2}(2k+1)$.*

Proof. To prove the claim, we must show that $2k+2$ does not divide $j(2k+1)$ for $1 \le j \le k+1$. This is equivalent to showing that

$$j(2k+1)-(2k+2) \notin S_1,$$

for each $1 \le j \le k+1$. Using Theorem 1.4 again, we must show that

$$j(2k+1)-(2k+2) \pmod{2k+1} > \left\lfloor \frac{j(2k+1)-(2k+2)}{2k+1} \right\rfloor 2, \tag{4}$$

for each $1 \le j \le k$. Now,

$$j(2k+1)-(2k+2) \equiv -1 \equiv 2k \pmod{2k+1},$$

and thus (4) reduces to

$$k > \left\lfloor j - \frac{2k+2}{2k+1} \right\rfloor. \tag{5}$$

Note that

$$\left\lfloor j - \frac{2k+2}{2k+1} \right\rfloor = \left\lfloor j - 1 - \frac{1}{2k+1} \right\rfloor = j - 1 + \left\lfloor -\frac{1}{2k+1} \right\rfloor = j-2.$$

Since $j \le k$, Equation (5) holds which completes the proof. □

Lemma 2.4. *In S_1, $2k+3$ does not divide any proper subsum of $\sum_{i=1}^{k+2}(2k+1)$.*

Proof. To prove the claim, we must show that $2k+3$ does not divide $j(2k+1)$ for $1 \le j \le k+1$. This is equivalent to showing that

$$j(2k+1)-(2k+3) \notin S_1,$$

for each $1 \le j \le k+1$. Using Theorem 1.4 again, we must show that

$$j(2k+1)-(2k+3) \pmod{2k+1} > \left\lfloor \frac{j(2k+1)-(2k+3)}{2k+1} \right\rfloor 2, \tag{6}$$

for each $1 \le j \le k$. Now,

$$j(2k+1)-(2k+3) \equiv -2 \equiv (2k-1) \pmod{2k+1},$$

and thus (6) reduces to

$$2k-1 > \left\lfloor j - \frac{2k+3}{2k+1} \right\rfloor 2.$$

Notice that $1 < (2k+3)/(2k+1) < 2$, and so $\lfloor j-(2k+3)/(2k+1) \rfloor = j-2$. Hence,

$$2k-1 > 2(j-2) = 2j-4,$$

and thus

$$k + \tfrac{3}{2} > j.$$

The last statement is true since $1 \leq j \leq k+1$, which completes the proof of the lemma. □

To complete the argument for S_1, we must verify that the first condition in Theorem 1.3(c) holds.

Proposition 2.5. (a) *If $(2k+1) \mid \alpha_1 + \cdots + \alpha_t$ where each α_i is irreducible in S_1 and $t \geq k+1$, then there is a proper subsum $\alpha_{i_1} + \cdots + \alpha_{i_r}$ of $\alpha_1 + \cdots + \alpha_t$ with $r \leq k+1$ such that $(2k+1) \mid \alpha_{i_1} + \cdots + \alpha_{i_r}$.*

(b) *If $(2k+2) \mid \alpha_1 + \cdots + \alpha_t$ where each α_i is irreducible in S_1 and $t \geq k+2$, then there is a proper subsum $\alpha_{i_1} + \cdots + \alpha_{i_r}$ of $\alpha_1 + \cdots + \alpha_t$ with $r \leq k+2$ such that $(2k+2) \mid \alpha_{i_1} + \cdots + \alpha_{i_r}$.*

(c) *If $(2k+3) \mid \alpha_1 + \cdots + \alpha_t$ where each α_i is irreducible in S_1 and $t \geq k+2$, then there is a proper subsum $\alpha_{i_1} + \cdots + \alpha_{i_r}$ of $\alpha_1 + \cdots + \alpha_t$ with $r \leq k+2$ such that $(2k+3) \mid \alpha_{i_1} + \cdots + \alpha_{i_r}$.*

Proof. (a) We can clearly reduce to the case where all the α_i are of the form $2k+2$ or $2k+3$. We also note that since $(2k+2)+(2k+2) = 4k+4 = (2k+1)+(2k+3)$, it follows that $(2k+1) \mid (2k+2)+(2k+2)$. Hence, if the sum $\alpha_1 + \cdots + \alpha_t$ contains two or more irreducibles of the form $2k+2$, then we are done. Assume that this is not the case. If there are no irreducibles of the form $2k+2$, then the result follows by Lemma 2.1(a). If there is exactly one copy of $2k+2$, then consider $k(2k+3)+(2k+2) = 2k^2+5k+2$. It follows that

$$(2k^2+5k+2) - (2k+1) = 2k^2+3k+1 = (k+1)(2k+1).$$

Hence, $(2k+1) \mid k(2k+3)+(2k+2)$, which completes the proof.

(b) It is only necessary to look at the case where all the α_i are of the form $2k+1$ or $2k+3$. We first note that since $(2k+1)+(2k+3) = 4k+4 = 2(2k+2)$, it follows $(2k+2) \mid (2k+1)+(2k+3)$, and if the sum $\alpha_1 + \cdots + \alpha_t$ contains at least one of each irreducible $2k+1$ and $2k+3$, then we are done. If the sum contains no copies of $2k+3$, then the result holds by Lemma 2.1(b). If the sum contains no copies of $2k+1$, then the equality

$$(k+1)(2k+3) - (2k+2) = 2k^2+3k+1 = (k+1)(2k+1)$$

completes the proof.

(c) It is only necessary to look at the case where the α_i are of the form $2k+1$ or $2k+2$. Now, $(2k+2)+(2k+2) = (2k+3)+(2k+1)$ and thus, if the sum $\alpha_1 + \cdots + \alpha_t$ contains at least 2 irreducibles of the form $2k+2$, then we are

done. If there are no irreducibles of the form $2k+2$, then this result follows by Lemma 2.1(c). If there is exactly one irreducible of the form $2k+2$, then consider $(k+1)(2k+1)+(2k+2)=2k^2+5k+3$. Now,

$$(2k^2+5k+3)-(2k+3)=2k^2+3k=k(2k+3),$$

and thus $(2k+3)\mid(k+1)(2k+1)+(2k+2)$, which completes the proof. □

3. Proof of Theorem 1.2 for S_2

Lemma 3.1. *In S_2, we have the following divisibility relationships*:

(a) $2k \,\Big|\, \sum_{i=1}^{k}(2k+2)$;

(b) $(2k+1) \,\Big|\, \sum_{i=1}^{k+2} 2k$;

(c) $(2k+2) \,\Big|\, \sum_{i=1}^{k+1} 2k$.

Proof. (a) $\sum_{i=1}^{k}(2k+2)=2k^2+2k$. Now, $(2k^2+2k)-(2k)=2k^2=2k(k)$. Thus, $2k\mid k(2k+2)$ and the result follows.

(b) $\sum_{i=1}^{k+2}(2k)=(k+2)(2k)=2k^2+4k$. Now, $(2k^2+4k)-(2k+1)=2k^2+2k-1=(k-1)(2k+2)+(2k+1)\in S_2$. Thus, $(2k+1)\mid(k+2)(2k)$ and the result follows.

(c) $\sum_{i=1}^{k+1}(2k)=(k+1)(2k)=2k^2+2k$. Now, $(2k^2+2k)-(2k+2)=(k-1)(2k+2)\in S_1$. Thus, $(2k+2)\mid 2k(k+1)$ and the result follows. □

Lemma 3.2. *In S_2, $2k$ does not divide any proper subsum of $\sum_{i=1}^{k}(2k+2)$.*

Proof. To prove this claim, we must show that $2k$ does not divide $j(2k+2)$ for $1\le j\le k-1$. This is equivalent to showing that

$$j(2k+2)-2k\notin S_2,$$

for each $1\le j\le k-1$. Using Theorem 1.4, we must show that

$$j(2k+2)-2k \pmod{2k} > 2\left\lfloor \frac{j(2k+2)-2k}{2k}\right\rfloor.$$

As in the arguments in Section 2, this reduces to

$$j>\left\lfloor \frac{jk+j-k}{k}\right\rfloor,$$

which is equivalent to

$$j>\left\lfloor j+\frac{j}{k}-1\right\rfloor.$$

Since $j \le k-1$, we have $j/k < 1$. So,

$$j - 1 = \left\lfloor j + \frac{j}{k} - 1 \right\rfloor \quad \text{and} \quad j > j - 1 = \left\lfloor j + \frac{j}{k} - 1 \right\rfloor.$$

Thus, no subsum is in S_2. □

Lemma 3.3. *In S_2, $2k+1$ does not divide any proper subsum of $\sum_{i=1}^{k+2}(2k)$.*

Proof. To prove this claim, we must show that $2k+1$ does not divide $j(2k)$ for $1 \le j \le k+1$. This is equivalent to showing that

$$j(2k) - (2k+1) \notin S_2,$$

for each $1 \le j \le k-1$. Using Theorem 1.4, we must show that

$$j(2k) - (2k+1) \pmod{2k} > 2\left\lfloor \frac{j2k - (2k+1)}{2k} \right\rfloor.$$

This is equivalent to

$$(2k-1) > 2\left\lfloor \frac{j2k - (2k+1)}{2k} \right\rfloor.$$

Since $1 < (2k+1)/2k < 2$, we know that

$$j - 2 = \lfloor j - 2 \rfloor = \left\lfloor j - \frac{2k+1}{2k} \right\rfloor = \left\lfloor \frac{j2k - (2k+1)}{2k} \right\rfloor.$$

By the limits on j, it follows that $k - \frac{1}{2} > j - 2$. Combining the last two inequalities and multiplying by 2 yields the desired result. □

Lemma 3.4. *In S_2, $2k+2$ does not divide any proper subsum of $\sum_{i=1}^{k+1} 2k$.*

Proof. To prove this claim, we must show that $2k+2$ does not divide $j(2k)$ for $1 \le j \le k$. This is equivalent to showing that

$$j(2k) - (2k+2) \notin S_2,$$

for each $1 \le j \le k$. Using Theorem 1.4, we must show that

$$j(2k) - (2k+2) \pmod{2k} > 2\left\lfloor \frac{j2k - (2k+2)}{2k} \right\rfloor.$$

This is equivalent to

$$k - 1 > \left\lfloor \frac{jk - k - 1}{k} \right\rfloor = \left\lfloor j - 1 - \frac{1}{k} \right\rfloor = j - 2,$$

from which the result follows. □

Proposition 3.5. (a) *If $2k \mid \alpha_1 + \cdots + \alpha_t$ where each α_i is irreducible in S_1 and $t \geq k$, then there is a proper subsum $\alpha_{i_1} + \cdots + \alpha_{i_r}$ of $\alpha_1 + \cdots + \alpha_t$ with $r \leq k$ such that $2k \mid \alpha_{i_1} + \cdots + \alpha_{i_r}$.*

(b) *If $(2k+1) \mid \alpha_1 + \cdots + \alpha_t$ where each α_i is irreducible in S_1 and $t \geq k+2$, then there is a proper subsum $\alpha_{i_1} + \cdots + \alpha_{i_r}$ of $\alpha_1 + \cdots + \alpha_t$ with $r \leq k+2$ such that $(2k+1) \mid \alpha_{i_1} + \cdots + \alpha_{i_r}$.*

(c) *If $(2k+2) \mid \alpha_1 + \cdots + \alpha_t$ where each α_i is irreducible in S_1 and $t \geq k+1$, then there is a proper subsum $\alpha_{i_1} + \cdots + \alpha_{i_r}$ of $\alpha_1 + \cdots + \alpha_t$ with $r \leq k+1$ such that $(2k+1) \mid \alpha_{i_1} + \cdots + \alpha_{i_r}$.*

Proof. (a) We can clearly reduce to the case where the α_i are of the form $2k+1$ and $2k+2$. Also note that we can assume that $t > k$, as the result clearly holds if $t = k$. Since

$$(2k+1) + (2k+1) = 4k+2 = (2k+2) + (2k),$$

the result holds if at least two of the α_i are of the form $2k+1$. If at least k of the α_i are of the form $2k+2$, then the result holds by Lemma 3.1(a). If not, then we have at least two of the form $2k+1$, which completes the proof of (a).

(b) We can clearly reduce to the case where the α_i are of the form $2k$ and $2k+2$. We proceed as in (a) and assume that $t > k+2$. Note that

$$(2k) + (2k+2) = (2k+1) + (2k+1).$$

Hence if at least one of the α_i is of each type, then we are done. If all the α_i are of the form $2k$, then we are done by Lemma 3.1(b). To complete the argument, note that $(k+1)(2k+2) = 2k^2 + 4k + 2$ and

$$2k^2 + 4k + 2 - (2k+1) = 2k^2 + 2k + 1 = k(2k) + (2k+1) \in S_2.$$

(c) We can clearly reduce to the case where the α_i are of the form $2k$ and $2k+1$. Assume as in (a) and (b) that $t > k+1$. As before,

$$(2k+1) + (2k+1) = 4k+2 = (2k+2) + (2k),$$

and if at least two of the α_i are of the form $2k+1$, then we are done. Otherwise, we have at least $2k+1$ copies of $2k$, and the result follows by Lemma 3.1(c). $\square$

Acknowledgement

The authors wish to thank an anonymous referee for many helpful comments and suggestions.

References

[Anderson and Chapman 2010] D. F. Anderson and S. T. Chapman, "How far is an element from being prime?", *J. Algebra Appl.* **9**:5 (2010), 779–789. MR 2012b:13003 Zbl 1203.13001

[Anderson and Chapman 2012] D. F. Anderson and S. T. Chapman, "On bounding measures of primeness in integral domains", *Internat. J. Algebra Comput.* **22**:5 (2012), 1250040, 15. MR 2949206 Zbl 1251.13002

[Anderson et al. 2011] D. F. Anderson, S. T. Chapman, N. Kaplan, and D. Torkornoo, "An algorithm to compute ω-primality in a numerical monoid", *Semigroup Forum* **82**:1 (2011), 96–108. MR 2012f:20171 Zbl 1218.20038

[Bowles et al. 2006] C. Bowles, S. T. Chapman, N. Kaplan, and D. Reiser, "On delta sets of numerical monoids", *J. Algebra Appl.* **5**:5 (2006), 695–718. MR 2007j:20092 Zbl 1115.20052

[Chapman et al. 2006] S. T. Chapman, M. T. Holden, and T. A. Moore, "Full elasticity in atomic monoids and integral domains", *Rocky Mountain J. Math.* **36**:5 (2006), 1437–1455. MR 2007j:20093 Zbl 1152.20048

[Chapman et al. 2009] S. T. Chapman, R. Hoyer, and N. Kaplan, "Delta sets of numerical monoids are eventually periodic", *Aequationes Math.* **77**:3 (2009), 273–279. MR 2010h:20141 Zbl 1204.20078

[García-Sánchez and Rosales 1999] P. A. García-Sánchez and J. C. Rosales, "Numerical semigroups generated by intervals", *Pacific J. Math.* **191**:1 (1999), 75–83. MR 2000i:20095 Zbl 1009.20069

[Geroldinger and Halter-Koch 2006] A. Geroldinger and F. Halter-Koch, *Non-unique factorizations: algebraic, combinatorial and analytic theory*, Pure and Applied Mathematics (Boca Raton) **278**, Chapman & Hall/CRC, Boca Raton, FL, 2006. MR 2006k:20001 Zbl 1113.11002

[Geroldinger and Hassler 2008] A. Geroldinger and W. Hassler, "Local tameness of v-Noetherian monoids", *J. Pure Appl. Algebra* **212**:6 (2008), 1509–1524. MR 2009b:20114 Zbl 1133.20047

[Omidali 2012] M. Omidali, "The catenary and tame degree of numerical monoids generated by generalized arithmetic sequences", *Forum Math.* **24**:3 (2012), 627–640. MR 2926638 Zbl 1252.20057

[Rosales and García-Sánchez 2009] J. C. Rosales and P. A. García-Sánchez, *Numerical semigroups*, Developments in Mathematics **20**, Springer, New York, 2009. MR 2010j:20091 Zbl 1220.20047

Received: 2013-05-14 Revised: 2013-11-21 Accepted: 2013-11-21

scott.chapman@shsu.edu *Department of Mathematics and Statistics, Lee Drain Building, Room 420, 1900 Avenue I, Sam Houston State University, Huntsville, TX 77340, United States*

wbp001@shsu.edu *Department of Mathematics and Statistics, Lee Drain Building, Room 420, 1900 Avenue I, Sam Houston State University, Huntsville, TX 77340, United States*

kns005@shsu.edu *Department of Mathematics and Statistics, Lee Drain Building, Room 420, 1900 Avenue I, Sam Houston State University, Huntsville, TX 77340, United States*

INVOLVE 7:5 (2014)
dx.doi.org/10.2140/involve.2014.7.669

Matrix coefficients of depth-zero supercuspidal representations of GL(2)

Andrew Knightly and Carl Ragsdale

(Communicated by Michael E. Zieve)

We give explicit formulas for matrix coefficients of the depth-zero supercuspidal representations of GL(2) over a nonarchimedean local field, highlighting the case where the test vector is a unit new vector. We also describe the partition of the set of such representations according to central character, and compute sums of matrix coefficients over all representations in a given class.

Introduction

Let F be a nonarchimedean local field with integer ring $\mathfrak{o}$, maximal ideal $\mathfrak{p} = \varpi\mathfrak{o}$, and residue field $k = \mathfrak{o}/\mathfrak{p}$ of cardinality q. The supercuspidal representations of $\mathrm{GL}_2(F)$ are precisely those irreducible admissible representations which do not arise as constituents of parabolic induction. They are characterized by having matrix coefficients which are compactly supported modulo the center.

In this paper we explicitly compute the matrix coefficients of depth-zero supercuspidal representations. These are the supercuspidals with the smallest possible conductor exponent, namely 2. First discovered by Mautner [1964, Section 9], they arise by compact induction from the $(q-1)$-dimensional representations of $\mathrm{GL}_2(\mathfrak{o})$ inflated from the cuspidal series of the finite group $\mathrm{GL}_2(k)$.

In the first section, we show that the matrix coefficients of any supercuspidal representation are expressible in terms of those of the finite-dimensional inducing representation. Thus, the task at hand essentially reduces to a computation of the matrix coefficients of the cuspidal representations of $\mathrm{GL}_2(k)$. The latter is achieved in Theorem 2.7 using the explicit model from [Piatetski-Shapiro 1983].

With global applications in mind, in Section 3 we single out the case where the test vector in the supercuspidal matrix coefficient is a unit new vector. The resulting function, given in (3-6) and Theorem 3.2, may be used to define an integral operator on the global automorphic spectrum of GL_2 which isolates those cuspidal

MSC2010: 22E50.
Keywords: supercuspidal, matrix coefficients, regular characters.

newforms with depth-zero supercuspidal local type (see Section 3.4). Possible applications include various trace formulas involving newforms of level p^2 which are supercuspidal (as opposed to principal series or special) at p. For a recent example involving the supercuspidal representations of conductor p^3, see [Knightly and Li 2012].

It is often desirable to organize representations according to central character. For example, there are exactly $2(q-1)$ supercuspidal representations of $\mathrm{GL}_2(F)$ of conductor p^3 with a given central character (see [Bushnell and Henniart 2014, Remark 2.2]). By contrast, there are $(q/2)(q-1)$ distinct cuspidal representations of $\mathrm{GL}_2(k)$ and $(q-1)$ possible central characters, but obviously there cannot be $q/2$ of each kind if q is odd. We sort this out in Proposition 2.3, and use it to give a formula (4-1) for the number of supercuspidals of conductor p^2 with a given central character. It depends on the parity of q and the order of the central character. We then give formulas for various sums of matrix coefficients over the set of depth-zero supercuspidal representations with a given central character. These computations rely on sum formulas for primitive characters, derived in Proposition 2.4. We close in the final section with some simple examples.

1. Matrix coefficients of supercuspidal representations

In this section let $G = \mathrm{GL}_n(F)$, or more generally, any unimodular locally profinite group [Bushnell and Henniart 2006] with center Z. Let $H \subset G$ be an open and closed subgroup containing Z with H/Z compact, and let (ρ, V) be an irreducible smooth representation of H. Consider the compact induction $\pi = \text{c-Ind}_H^G(\rho)$. It consists of the functions $\phi : G \longrightarrow V$ with compact support (mod Z) for which $\phi(hg) = \rho(h)\phi(g)$ for all $h \in H$, $g \in G$, with G acting on the space by right translation. Here we show that, as observed by Mautner [1964], the matrix coefficients of π are essentially those of ρ. These matrix coefficients are compactly supported (modulo Z), so if π is irreducible and admissible, it is supercuspidal. Conversely, it is conjectured that all supercuspidal representations arise in this way. This was proven by Bushnell and Kutzko [1993] for $G = \mathrm{GL}_n(F)$, and more recently in great (but not complete) generality in [Stevens 2008; Kim 2007].

We assume for simplicity that ρ has unitary central character, so that by the fact that H/Z is compact, ρ is unitarizable. Let $\langle v, w\rangle_V$ denote an H-equivariant inner product on V. Then the inner product on $\text{c-Ind}_H^G(\rho)$ given by

$$\langle \phi, \psi \rangle = \sum_{x \in H\backslash G} \big\langle \phi(x), \psi(x) \big\rangle_V$$

is convergent (in fact a finite sum) and well-defined. Further, for any $g \in G$,

$$\big\langle \pi(g)\phi, \pi(g)\psi \big\rangle = \sum_{x \in H\backslash G} \big\langle \phi(xg), \psi(xg) \big\rangle_V = \sum_{x \in H\backslash G} \big\langle \phi(x), \psi(x) \big\rangle_V = \langle \phi, \psi \rangle.$$

Thus π is unitary relative to this inner product.

For $v \in V$ and $y \in G$, define a function $f_{y,v} \in \text{c-Ind}_H^G(\rho)$ by

$$f_{y,v}(g) = \begin{cases} \rho(h)v & \text{if } g = hy \in Hy, \\ 0 & \text{if } g \notin Hy. \end{cases}$$

Then the set $\{f_{y,v} \mid y \in H\backslash G, v \in V\}$ spans the space $\text{c-Ind}_H^G(\rho)$. (Note that $f_{hy,v} = f_{y,\rho(h^{-1})v}$.)

Proposition 1.1. *For $y, z \in G$ and $v, w \in V$,*

$$\langle \pi(g) f_{y,v}, f_{z,w} \rangle = \begin{cases} \langle \rho(h)v, w \rangle_V & \text{if } g = z^{-1}hy \in z^{-1}Hy, \\ 0 & \text{if } g \notin z^{-1}Hy. \end{cases}$$

Proof. By definition of the inner product,

$$\begin{aligned} \langle \pi(g) f_{y,v}, f_{z,w} \rangle &= \sum_{x \in H\backslash G} \langle \pi(g) f_{y,v}(x), f_{z,w}(x) \rangle_V \\ &= \sum_{x \in H\backslash G} \langle f_{y,v}(xg), f_{z,w}(x) \rangle_V \\ &= \langle f_{y,v}(zg), w \rangle_V, \end{aligned}$$

since $f_{z,w}(x)$ vanishes unless $x \in Hz$. If $g = z^{-1}hy \in z^{-1}Hy$, then the above is equal to

$$\langle f_{y,v}(hy), w \rangle_V = \langle \rho(h)v, w \rangle_V,$$

as needed. If $g \notin z^{-1}Hy$, then $zg \notin Hy$, so $f_{y,v}(zg) = 0$ and the inner product vanishes. □

If we let $\overline{G} = G/Z$, then the formal degree d_π of π is a positive constant satisfying

$$\int_{\overline{G}} \left| \langle \pi(g) f, f \rangle \right|^2 dg = \frac{\|f\|^4}{d_\pi} \tag{1-1}$$

for all $f \in \text{c-Ind}_H^G(\rho)$. It depends on the choice of Haar measure on $\overline{G}$. (The existence of d_π is due to Godement; see, for example, [Knightly and Li 2006, Proposition 10.4].)

Proposition 1.2. *For any choice of Haar measure on $\overline{G} = G/Z$, the associated formal degree of π is given by*

$$d_\pi = \frac{\dim \rho}{\text{meas}(\overline{H})},$$

where $\overline{H}$ is the (open) image of H in $\overline{G}$.

Proof. Let $v \in V$ be a unit vector, and consider the function $f_{1,v}$. By Proposition 1.1,

$$\langle \pi(g) f_{1,v}, f_{1,v} \rangle = \begin{cases} \langle \rho(g)v, v \rangle_V & \text{if } g \in H, \\ 0 & \text{if } g \notin H. \end{cases}$$

Therefore,

$$\begin{aligned} \int_{\overline{G}} \left| \langle \pi(g) f_{1,v}, f_{1,v} \rangle \right|^2 dg &= \int_{\overline{H}} \left| \langle \rho(g)v, v \rangle_V \right|^2 dg \\ &= \frac{\|v\|^4}{\dim(\rho)} \operatorname{meas}(\overline{H}) = \frac{\operatorname{meas}(\overline{H})}{\dim(\rho)}, \end{aligned}$$

by the Schur orthogonality relations for irreducible representations of compact groups. By (1-1),

$$\frac{\operatorname{meas}(\overline{H})}{\dim(\rho)} = \frac{\|f_{1,v}\|^4}{d_\pi}.$$

Therefore, it suffices to show that $\|f_{1,v}\| = 1$. This can be done via a direct computation:

$$\begin{aligned} \|f_{1,v}\|^2 = \langle f_{1,v}, f_{1,v} \rangle &= \sum_{x \in H \backslash G} \langle f_{1,v}(x), f_{1,v}(x) \rangle_V \\ &= \langle f_{1,v}(1), f_{1,v}(1) \rangle_V = \langle v, v \rangle = 1, \end{aligned}$$

as needed. □

2. Cuspidal representations of $\mathrm{GL}_2(k)$

Let q be a prime power, let k be the finite field with q elements, let L be the unique quadratic extension of k, and let $G = \mathrm{GL}_2(k)$. Define the subgroups

$$U = \left\{ \left(\begin{smallmatrix} 1 & b \\ 0 & 1 \end{smallmatrix}\right) \in G \right\}, \quad Z = \left\{ \left(\begin{smallmatrix} a & \\ & a \end{smallmatrix}\right) \in G \right\}, \quad B = \left\{ \left(\begin{smallmatrix} a & b \\ 0 & d \end{smallmatrix}\right) \in G \right\}.$$

Note that Z is the center of G. Recall that the cuspidal representations of G are those that do not contain the trivial character of the unipotent subgroup U. These are precisely the irreducible representations that do not arise via parabolic induction. They have dimension $q-1$, and are parametrized by the Galois orbits of the primitive characters of L^*, defined below.

2.1. *Primitive characters of L^*.* For any finite abelian group H, we write $\widehat{H}$ for the dual group, consisting of the characters $\chi : H \longrightarrow \mathbb{C}^*$. We recall that

$$H \cong \widehat{H}.$$

Thus $\widehat{L^*}$ is a cyclic group of order $q^2 - 1$. A character $\nu \in \widehat{L^*}$ is *primitive* if $\nu^q \neq \nu$. Otherwise, ν is *imprimitive*. Letting $\bar{\alpha} = \alpha^q$ denote the Frobenius map, the norm

map $N : L \longrightarrow k$ is given by

$$N(\alpha) = \alpha\bar{\alpha} = \alpha^{q+1}.$$

Proposition 2.1. *Let $\nu : L^* \to \mathbb{C}^*$ be a character of L^*. Then the following are equivalent*:

(i) *ν is imprimitive*; *that is*, $\nu^q = \nu$.

(ii) *$\nu = \chi \circ N$ for some $\chi \in \widehat{k^*}$.*

(iii) *$L^1 \subset \ker(\nu)$, where L^1 is the subgroup of norm 1 elements of L^*.*

Proof. Let θ be a generator of the cyclic group L^*. Then θ^{q+1} is a generator of k^*. If $\nu^q = \nu$, then $\nu(\theta)$ is a $(q-1)$-st root of unity, so we may define $\chi \in \widehat{k^*}$ by $\chi(\theta^{q+1}) = \nu(\theta)$. Then $\nu = \chi \circ N$, so (i) implies (ii). It is clear that (ii) implies (iii). On the other hand, for any $x \in L^*$, $N(x^{q-1}) = N(x)^{q-1} = 1$, so that $x^{q-1} \in L^1$. Therefore if (iii) holds, $\nu^q(x) = \nu(x^q) = \nu(x^{q-1}x) = \nu(x^{q-1})\nu(x) = \nu(x)$. Hence (iii) implies (i). □

The imprimitive characters thus correspond bijectively with the characters of k^*, so there are $q-1$ of them. It follows that there are $(q^2-1)-(q-1) = q^2-q$ primitive characters of L^*.

Lemma 2.2. *Let ν be a primitive character of L^*. Then, for all $\alpha \in k^*$,*

$$\sum_{\substack{x \in L^* \\ N(x)=\alpha}} \nu(x) = 0.$$

Proof. By Proposition 2.1, there exists $\lambda \in L^1$ such that $\nu(\lambda) \neq 1$. Thus,

$$\sum_{N(x)=\alpha} \nu(x) = \sum_{N(x)=\alpha} \nu(\lambda x) = \nu(\lambda) \sum_{N(x)=\alpha} \nu(x).$$

It follows that $\sum_{N(x)=\alpha} \nu(x) = 0$. □

Next, we examine the partition of primitive characters into classes according to their restrictions to k^*. This will allow us to count the number of depth-zero supercuspidal representations with a given central character (see (4-1)).

Proposition 2.3. *Suppose ω is a given character of k^*. Let P_ω denote the number of primitive characters ν of L^* for which $\nu|_{k^*} = \omega$. Then*

$$P_\omega = \begin{cases} q-1 & \text{if } q \text{ is odd and } \omega^{(q-1)/2} \text{ is trivial}, \\ q+1 & \text{if } q \text{ is odd and } \omega^{(q-1)/2} \text{ is nontrivial}, \\ q & \text{if } q \text{ is even}. \end{cases}$$

Proof. Let ξ be a generator of the cyclic group $\widehat{L^*}$. Let $\nu_0 = \xi^{q-1}$. Note that for $\alpha \in k^*$,

$$\nu_0(\alpha) = \xi(\alpha^{q-1}) = \xi(1) = 1.$$

In fact, ν_0 is a generator of the order $q+1$ subgroup $\widehat{L^*/k^*}$ of $\widehat{L^*}$. Let us consider two characters of L^* to be equivalent if they have the same restriction to k^*. Then the equivalence class of a given character ν is the set

$$\{\nu, \nu\nu_0, \nu\nu_0^2, \dots, \nu\nu_0^q\}. \tag{2-1}$$

If $\omega = \nu|_{k^*}$, then $P_\omega = q+1-A_\omega$, where A_ω is the number of imprimitive elements of the above set. Write $\nu = \xi^b$. Without loss of generality (replacing ν by some $\nu\nu_0^m$), we may assume that $0 \le b < q-1$. A character ξ^a is imprimitive if and only if $\xi^{qa} = \xi^a$, or equivalently, $(q+1) \mid a$. Suppose $\nu\nu_0^s$ and $\nu\nu_0^k$ are both imprimitive for $0 \le s \le k \le q$. Then $b+(q-1)s$ and $b+(q-1)k$ are both divisible by $q+1$ and strictly less than q^2-1. Their difference $(k-s)(q-1) \ge 0$ also has these properties, and furthermore it is divisible by

$$\operatorname{lcm}(q-1, q+1) = \begin{cases} (q^2-1)/2 & \text{if } q \text{ is odd,} \\ q^2-1 & \text{if } q \text{ is even.} \end{cases}$$

It follows that $k-s=0$ if q is even, and $k-s \in \{0, (q+1)/2\}$ if q is odd. This means that $A_\omega \le 1$ if q is even, and $A_\omega \le 2$ if q is odd.

Suppose q is odd and b is even. Then there are two imprimitive elements, namely

$$b + \tfrac{1}{2}b(q-1) = \tfrac{1}{2}b(q+1), \tag{2-2}$$

giving $\nu\nu_0^{b/2} = \xi^{b/2} \circ N$, and

$$b + \frac{b+q+1}{2}(q-1) = \frac{b+q-1}{2}(q+1), \tag{2-3}$$

giving $\nu\nu_0^{(b+q+1)/2} = \xi^{(b+q-1)/2} \circ N$. Hence $A_\omega = 2$ in this case. Noting that b is even if and only if $\omega^{(q-1)/2} = 1$, we obtain the first claim of the proposition: $P_\omega = q+1-A_\omega = q-1$.

Suppose q is odd and b is odd. Then for all k, $b+k(q-1)$ is odd, and hence it cannot be divisible by the even number $q+1$. So $A_\omega = 0$, and $P_\omega = q+1$, proving the second claim of the proposition.

If q is even, then as shown above, $A_\omega \le 1$. If b is even then (2-2) is a solution, and if b is odd, (2-3) is a solution. Either way, this shows that $A_\omega \ge 1$, and hence $A_\omega = 1$, proving the final claim that $P_\omega = q+1-A_\omega = q$ when q is even. □

The character sums in the next proposition will be used in Section 4 when we sum matrix coefficients over all representations with a given central character.

Proposition 2.4. *Let ω be a character of k^*, and let $[\omega]$ denote the set of primitive characters of L^* extending ω.*

Suppose q is odd and $\omega^{(q-1)/2}$ is nontrivial. Then for $\alpha \in L^*$,

$$\sum_{\nu\in[\omega]} \nu(\alpha) = \begin{cases} (q+1)\omega(\alpha) & \text{if } \alpha \in k^*, \\ 0 & \text{if } \alpha \notin k^*. \end{cases} \tag{2-4}$$

Suppose q is odd and $\omega^{(q-1)/2}$ is trivial. Then for $\alpha \in L^*$,

$$\sum_{\nu\in[\omega]} \nu(\alpha) = \begin{cases} (q-1)\omega(\alpha) & \text{if } \alpha \in k^*, \\ -2\omega(\alpha^{(q+1)/2}) & \text{if } \alpha \notin k^*,\ \alpha^{(q^2-1)/2} = 1, \\ 0 & \text{if } \alpha^{(q^2-1)/2} = -1. \end{cases} \tag{2-5}$$

(*Note that necessarily $\alpha \notin k^*$ if $\alpha^{(q^2-1)/2} \neq 1$.*)

Suppose q is even. Then for $\alpha \in L^*$,

$$\sum_{\nu\in[\omega]} \nu(\alpha) = \begin{cases} q\omega(\alpha) & \text{if } \alpha \in k^*, \\ -\omega(N(\alpha)^{1/2}) & \text{if } \alpha \notin k^*. \end{cases} \tag{2-6}$$

Here, we note that the square root is unique in k^*, since the square function is a bijection when q is even.

Proof. If $\alpha \in k^*$, then the sum is equal to $P_\omega\omega(\alpha)$ and the assertions follow from the previous proposition. So we may assume that $\alpha \notin k^*$. We use the notation from the previous proof. Suppose q is odd and $\omega^{(q-1)/2}$ is nontrivial. By the proof of the previous proposition,

$$\sum_{\nu\in[\omega]} \nu(\alpha) = \nu(\alpha) \sum_{m=0}^{q} \nu_0^m(\alpha),$$

where on the right-hand side ν is any fixed element of $[\omega]$. Noting that

$$\sum_{m=0}^{q} \nu_0^m(\alpha) = \sum_{\chi\in\widehat{L^*/k^*}} \chi(\alpha) = \begin{cases} q+1 & \text{if } \alpha \in k^*, \\ 0 & \text{if } \alpha \notin k^*, \end{cases} \tag{2-7}$$

(2-4) follows.

Now suppose q is odd and $\omega^{(q-1)/2}$ is trivial. By the proof of the previous proposition,

$$\sum_{\nu\in[\omega]} \nu(\alpha) = \xi^b(\alpha)\Bigg(\sum_{m=0}^{q} \nu_0^m(\alpha) - \nu_0(\alpha)^{b/2} - \nu_0(\alpha)^{(b+q+1)/2}\Bigg).$$

Since $\alpha \notin k^*$, by (2-7) this is equal to

$$-\xi^b(\alpha)\big[\nu_0(\alpha)^{b/2} + \nu_0(\alpha)^{b/2}\nu_0(\alpha)^{(q+1)/2}\big].$$

Recalling that $\nu_0 = \xi^{q-1}$ and writing $\xi^b = \nu$, this is

$$= -\nu\big(\alpha^{1+((q-1)/2)}\big)\big[1+\xi\big(\alpha^{(q^2-1)/2}\big)\big] = -\nu\big(\alpha^{(q+1)/2}\big)\big[1+\xi(\alpha^{(q^2-1)/2})\big]. \quad (2\text{-}8)$$

Observe that $\alpha^{(q^2-1)/2} = \pm 1$ since its square is 1. If it is equal to $+1$, then $\alpha^{(q+1)/2} \in k^*$ since its $(q-1)$-st power is 1, and we immediately obtain the middle line of (2-5). Otherwise $\xi\big(\alpha^{(q^2-1)/2}\big) = \xi(-1) = -1$ and (2-8) vanishes.

When q is even, there exists a choice of ξ for which $\omega = \xi^b$ with b even. Then using (2-2), we find that when $\alpha \notin k^*$,

$$\sum_{\nu \in [\omega]} \nu(\alpha) = -\xi^{b/2}(N(\alpha)) = -\omega\big(N(\alpha)^{1/2}\big). \qquad \square$$

2.2. ***Model for cuspidal representations.*** There are various ways to construct the cuspidal representation ρ_ν attached to a primitive character ν. The action of L^* on the k-vector space $L \cong k^2$ by multiplication gives an identification

$$L^* \cong T \quad (2\text{-}9)$$

of L^* with a nonsplit torus $T \subset G$, with $k^* \subset L^*$ mapping onto $Z \subset T$. The characteristic polynomial of an element $g \in G$ is irreducible over k if and only if g is conjugate to an element of $T - Z$.

Fix a nontrivial character of the additive group

$$\psi : k \longrightarrow \mathbb{C}^*,$$

viewed in the obvious way as a character of U. Then one may define ρ_ν implicitly by

$$\mathrm{Ind}_{ZU}^G(\nu \otimes \psi) = \rho_\nu \oplus \mathrm{Ind}_T^G\, \nu; \quad (2\text{-}10)$$

see [Bushnell and Henniart 2006, Theorem 6.4]. Although (2-10) allows for computation of the trace of ρ_ν (see (2-20) below), it is not convenient for computing the matrix coefficients. For this purpose we shall use the explicit model for ρ_ν defined in [Piatetski-Shapiro 1983, Section 13] as follows.[1]

Given a primitive character ν of L^* and ψ as above, let

$$V = \mathbb{C}[k^*]$$

[1]There is a minus sign missing from the definition of $j(x)$ in Equation (4) of Section 13 of [Piatetski-Shapiro 1983] (otherwise his identity (6) will not hold). Likewise a minus sign is missing from (16) on page 40. The expression four lines above (16) is correct (except K should be K^*).

be the vector space of functions $f : k^* \to \mathbb{C}$. We define a representation ρ_ν of G on V as follows. For any $\left(\begin{smallmatrix} a & b \\ 0 & d \end{smallmatrix}\right) \in B$, $f \in V$, let

$$\left[\rho_\nu \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} f\right](x) = \nu(d)\psi(bd^{-1}x)f(ad^{-1}x) \quad (x \in k^*), \tag{2-11}$$

and for $g \in G - B$, define

$$(\rho_\nu(g)f)(x) = \sum_{y \in k^*} \phi(x, y; g) f(y) \quad (x \in k^*), \tag{2-12}$$

where, for $g = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in G - B$,

$$\phi(x, y; g) = -\frac{1}{q}\psi\left[\frac{ax + dy}{c}\right] \sum_{\substack{t \in L^* \\ N(t) = xy^{-1} \det g}} \psi\left(-\frac{y}{c}(t + \bar{t})\right)\nu(t). \tag{2-13}$$

Theorem 2.5. *If ν is a primitive character of L^*, (2-11) and (2-12) give a well-defined representation (ρ_ν, V) which is cuspidal. Furthermore, every cuspidal representation is isomorphic to some ρ_ν, and $\rho_\nu \cong \rho_{\nu'}$ if and only if $\nu' \in \{\nu, \nu^q\}$. In particular, there are $(q^2 - q)/2$ distinct cuspidal representations.*

Proof. See [Piatetski-Shapiro 1983, Section 13–14], where it is assumed that $q > 2$ throughout. When $q = 2$, G is isomorphic to the symmetric group S_3. The unique cuspidal representation is the character sending each permutation to its sign. It is readily checked that the above construction defines this character as well, so the theorem remains valid when $q = 2$. □

Define an inner product on V by

$$\langle f_1, f_2 \rangle = \sum_{x \in k^*} f_1(x)\overline{f_2(x)}. \tag{2-14}$$

We will work with the orthonormal basis

$$\mathcal{B} = \{f_r\}_{r \in k^*} \quad \text{for } f_r(x) = \begin{cases} 1 & \text{if } x = r, \\ 0 & \text{if } x \neq r. \end{cases}$$

Proposition 2.6. *Let ν be a primitive character of L^*, and let ρ_ν be the associated cuspidal representation of G. Then ρ_ν is unitary with respect to the inner product (2-14).*

Proof. By linearity, it suffices to prove that for all $f_r, f_s \in \mathcal{B}$ and $g \in G$,

$$\langle \rho_\nu(g) f_r, \rho_\nu(g) f_s \rangle = \langle f_r, f_s \rangle. \tag{2-15}$$

By the Bruhat decomposition $G = B \cup Bw'U$ for $w' = \left(\begin{smallmatrix} & 1 \\ -1 & \end{smallmatrix}\right)$ and the fact that ρ_ν is a homomorphism, we only need to consider $g \in B$ and $g = w'$.

Suppose first that $g = \left(\begin{smallmatrix} a & b \\ 0 & d \end{smallmatrix}\right) \in B$. Then by (2-11),

$$\left\langle \rho_\nu(g) f_r, \rho_\nu(g) f_s \right\rangle = \sum_{x \in k^*} \nu(d)\psi(bd^{-1}x) f_r(ad^{-1}x)\overline{\nu(d)\psi(bd^{-1}x) f_s(ad^{-1}x)}.$$

Using the fact that ν and ψ are unitary, and replacing x by $a^{-1}\,dx$, we see that this expression equals

$$\sum_{x \in k^*} f_r(x)\overline{f_s(x)} = \langle f_r, f_s \rangle,$$

as needed.

It remains to prove (2-15) for $g = w' = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$. By (2-12),

$$\begin{aligned}
\rho_\nu(w') f(x) &= \sum_{y \in k^*} \phi(x, y; w') f(y) = -\frac{1}{q} \sum_{y \in k^*} \sum_{\substack{t \in L^* \\ N(t) = xy^{-1}}} \psi(y(t + \bar{t}))\nu(t) f(y) \\
&= -\frac{1}{q} \sum_{y \in k^*} \sum_{\substack{u \in L^* \\ N(u) = xy}} \psi(u + \bar{u})\nu(u)\nu(y^{-1}) f(y) = \sum_{y \in k^*} \nu(y^{-1}) j(xy) f(y),
\end{aligned}$$

where

$$j(t) = -\frac{1}{q} \sum_{\substack{u \in L^* \\ N(u) = t}} \psi(u + \bar{u})\nu(u). \tag{2-16}$$

Hence

$$\rho_\nu(w') f_r(x) = \sum_{y \in k^*} \nu(y^{-1}) j(xy) f_r(y) = \nu(r^{-1}) j(rx).$$

We now see that

$$\begin{aligned}
\langle \rho_\nu(w') f_r, \rho_\nu(w') f_s \rangle &= \sum_{x \in k^*} \nu(r^{-1}) j(rx)\overline{\nu(s^{-1}) j(sx)} = \nu(sr^{-1}) \sum_{x \in k^*} j(rx)\overline{j(sx)} \\
&= \nu(sr^{-1}) \sum_{x \in k^*} j(rs^{-1}x)\overline{j(x)}.
\end{aligned}$$

Taking $r' = rs^{-1}$, it suffices to prove that

$$\sum_{x \in k^*} j(r'x)\overline{j(x)} = \begin{cases} 0 & \text{if } r' \neq 1, \\ 1 & \text{if } r' = 1. \end{cases} \tag{2-17}$$

From the definition (2-16), we have

$$\begin{aligned}
\sum_{x \in k^*} j(r'x)\overline{j(x)} &= \frac{1}{q^2} \sum_{x \in k^*} \sum_{N(\alpha) = r'x} \sum_{N(\beta) = x} \psi(\alpha + \bar{\alpha})\nu(\alpha)\psi(-\beta - \bar{\beta})\nu(\beta^{-1}) \\
&= \frac{1}{q^2} \sum_{\beta \in L^*} \sum_{N(\alpha) = r'N(\beta)} \psi(\alpha + \bar{\alpha} - \beta - \bar{\beta})\nu(\alpha\beta^{-1}).
\end{aligned}$$

Since the norm map is surjective, there exists $z \in L^*$ such that $N(z) = r'$. Then $\alpha = z\beta u$ for some $u \in L^1$. This allows us to rewrite the above sum as

$$\begin{aligned}\sum_{x\in k^*} j(r'x)\overline{j(x)} &= \frac{1}{q^2}\sum_{\beta\in L^*}\sum_{u\in L^1}\psi\big(z\beta u + \overline{z\beta u} - \beta - \bar{\beta}\big)\nu(zu)\\ &= \frac{\nu(z)}{q^2}\sum_{u\in L^1}\nu(u)\sum_{\beta\in L^*}\psi\big(\mathrm{tr}[(zu-1)\beta]\big). \end{aligned} \tag{2-18}$$

Generally, for $c \in L$, the map $R(\beta) = \psi(\mathrm{tr}[c\beta])$ is a homomorphism from L to $\mathbb{C}^*$. If $c \neq 0$, then R is nontrivial since the trace map from L to k is surjective. It follows that

$$\sum_{\beta\in L^*}\psi\big(\mathrm{tr}[c\beta]\big) = \begin{cases} -1 & \text{if } c \neq 0,\\ q^2-1 & \text{if } c = 0.\end{cases} \tag{2-19}$$

Suppose $r' \neq 1$. Then $N(zu) = N(z) = r' \neq 1$, so in particular $zu \neq 1$. Therefore (2-18) becomes

$$\sum_{x\in k^*} j(r'x)\overline{j(x)} = -\frac{\nu(z)}{q^2}\sum_{u\in L^1}\nu(u) = 0,$$

where we have used the fact (Proposition 2.1) that ν is a nontrivial character of L^1 since ν is primitive.

Now suppose $r' = 1$. Then we can take $z = 1$, so by (2-18) and (2-19),

$$\begin{aligned}\sum_{x\in k^*} j(x)\overline{j(x)} &= \frac{1}{q^2}\sum_{u\in L^1}\nu(u)\sum_{\beta\in L^*}\psi\big(\mathrm{tr}[(u-1)\beta]\big)\\ &= \frac{q^2-1}{q^2} - \frac{1}{q^2}\sum_{\substack{u\in L^1\\ u\neq 1}}\nu(u) = \left(1-\frac{1}{q^2}\right)+\frac{1}{q^2} = 1,\end{aligned}$$

since $\sum_{\substack{u\in L^1\\ u\neq 1}}\nu(u) = -1$, again because ν is nontrivial on L^1. This proves (2-17). □

2.3. ***Matrix coefficients of cuspidal representations.*** Let ν be a primitive character of L^*. Using (2-10), one finds that

$$\mathrm{tr}\,\rho_\nu(x) = \begin{cases} (q-1)\nu(x) & \text{if } x \in Z,\\ -\nu(z) & \text{if } x = zu,\ z\in Z,\ u\in U,\ u\neq 1,\\ -\nu(x)-\nu^q(x) & \text{if } x\in T,\ x\notin Z,\\ 0 & \text{if no conjugate of } x \text{ is in } T\cup ZU\end{cases} \tag{2-20}$$

(see [Bushnell and Henniart 2006, (6.4.1)]). This is a sum of matrix coefficients. For the coefficients themselves, we use the model given in the previous section to prove the following (which can also be used to derive (2-20)).

Theorem 2.7. *Let $g = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in G$ and $f_r, f_s \in \mathcal{B}$. Let ρ_ν be a cuspidal representation of G. If $g \in B$, then*

$$\langle \rho_\nu(g) f_r, f_s \rangle = \begin{cases} \nu(d)\psi(bd^{-1}s) & \text{if } r = ad^{-1}s, \\ 0 & \text{if } r \neq ad^{-1}s. \end{cases}$$

If $g \notin B$, then

$$\langle \rho_\nu(g) f_r, f_s \rangle = \phi(s, r; g),$$

where ϕ is defined in (2-13).

Proof. First suppose $g \in B$ (i.e., $c = 0$). Then

$$\begin{aligned} \langle \rho_\nu(g) f_r, f_s \rangle &= \sum_{x \in k^*} [\rho_\nu(g) f_r](x) \overline{f_s(x)} = [\rho_\nu(g) f_r](s) \\ &= \nu(d)\psi(bd^{-1}s) f_r(ad^{-1}s) = \begin{cases} \nu(d)\psi(bd^{-1}s) & \text{if } r = ad^{-1}s, \\ 0 & \text{if } r \neq ad^{-1}s. \end{cases} \end{aligned}$$

Now, suppose $g \notin B$. Then

$$[\rho_\nu(g) f_r](x) = \sum_{y \in k^*} \phi(x, y; g) f_r(y) = \phi(x, r; g).$$

Therefore

$$\langle \rho_\nu f_r, f_s \rangle = \sum_{x \in k^*} [\rho_\nu(g) f_r](x) \overline{f_s(x)} = [\rho_\nu(g) f_r](s) = \phi(s, r; g),$$

as needed. □

3. Depth-zero supercuspidal representations of $\mathrm{GL}_2(F)$

We move now to the $\mathfrak{p}$-adic setting. When no field is specified, G, Z, B, U, etc., will henceforth denote the corresponding subgroups of $\mathrm{GL}_2(F)$ rather than $\mathrm{GL}_2(k)$. Fix a primitive character ν, and let ρ_ν be the associated cuspidal representation of $\mathrm{GL}_2(k)$. We view ρ_ν as a representation of $K = \mathrm{GL}_2(\mathfrak{o})$ via reduction modulo $\mathfrak{p}$:

$$K \longrightarrow \mathrm{GL}_2(k) \longrightarrow \mathrm{GL}(V).$$

The central character of this representation is given by $z \mapsto \nu(z(1+\mathfrak{p}))$ for $z \in \mathfrak{o}^*$. Extend this character of $\mathfrak{o}^*$ to $Z \cong F^* = \bigcup_{n \in \mathbb{Z}} \varpi^n \mathfrak{o}^*$ by choosing a complex number $\nu(\varpi)$ of absolute value 1. We denote this character of F^* by ν. This allows us to view ρ_ν as a unitary representation of the group ZK. Let

$$\pi_\nu = \text{c-Ind}_{ZK}^{\mathrm{GL}_2(F)}(\rho_\nu)$$

be the representation of G compactly induced from ρ_ν. This representation is irreducible and supercuspidal (see, for example, [Bump 1997, Theorem 4.8.1]).

3.1. ***Matrix coefficients.*** Define an inner product on $\text{c-Ind}_{ZK}^{G}(\rho_\nu)$ by

$$\langle f_1, f_2\rangle = \sum_{x\in ZK\backslash G} \big\langle f_1(x), f_2(x)\big\rangle_V, \tag{3-1}$$

where $\langle\,\cdot\,,\cdot\,\rangle_V$ denotes the inner product on V defined in (2-14). As in Section 1, this inner product is G-equivariant.

The matrix coefficients of π_ν can now be computed explicitly by using Proposition 1.1 in conjunction with Theorem 2.7. Likewise, by Proposition 1.2, if we normalize so that $\operatorname{meas}(\overline{K}) = 1$, then

$$d_{\pi_\nu} = \dim \rho_\nu = q - 1. \tag{3-2}$$

Define a function $\phi_\nu : G \longrightarrow \mathbb{C}$ by

$$\phi_\nu(x) = \begin{cases} \overline{\operatorname{tr}\rho_\nu(x)} & \text{if } x \in ZK, \\ 0 & \text{otherwise.} \end{cases} \tag{3-3}$$

Then this is a *pseudocoefficient* of π_ν in the sense that for any irreducible tempered representation π of G with central character ω,

$$\operatorname{tr}\pi(\phi_\nu) = \begin{cases} 1 & \text{if } \pi \cong \pi_\nu, \\ 0 & \text{otherwise} \end{cases}$$

(see [Palm 2012, Section 9.4.1]). The function ϕ_ν may be computed explicitly using (2-20).

3.2. ***New vectors.*** For an integer $n \geq 0$, define the congruence subgroup

$$K_1(\mathfrak{p}^n) = \left\{ \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in K \mid c, (d-1) \in \mathfrak{p}^n \right\}.$$

If π is a representation of G, we let $\pi^{K_1(\mathfrak{p}^n)}$ denote the space of vectors fixed by $K_1(\mathfrak{p}^n)$. By a result of Casselman [1973], for any irreducible admissible representation π of G, there exists a unique ideal $\mathfrak{p}^n$ (the *conductor* of π) for which $\dim \pi^{K_1(\mathfrak{p}^n)} = 1$ and $\dim \pi^{K_1(\mathfrak{p}^{n-1})} = 0$. A nonzero vector fixed by $K_1(\mathfrak{p}^n)$ is called a *new vector*. The supercuspidal representations constructed above have conductor $\mathfrak{p}^2$. We shall give an elementary proof below, and exhibit a new vector. More generally, the new vectors for depth-zero supercuspidal representations of $\mathrm{GL}_n(F)$ were identified by Reeder [1991, Example (2.3)].

Proposition 3.1. *The supercuspidal representation π_ν defined above has conductor $\mathfrak{p}^2$. If we let $w \in \mathbb{C}[k^*]$ denote the constant function* 1, *that is,*

$$w = \sum_{r\in k^*} f_r, \tag{3-4}$$

then the function $f = f_{\left(\begin{smallmatrix}\varpi & \\ & 1\end{smallmatrix}\right),w}$ *supported on the coset*

$$ZK\begin{pmatrix}\varpi & 0\\ 0 & 1\end{pmatrix} = ZK\begin{pmatrix}\varpi & 0\\ 0 & 1\end{pmatrix}K_1(\mathfrak{p}^2),$$

and defined by

$$f\left(zk\begin{pmatrix}\varpi & 0\\ 0 & 1\end{pmatrix}\right) = \rho_\nu(zk)w,$$

is a new vector of π_ν.

Proof. To see that f is $K_1(\mathfrak{p}^2)$-invariant, it suffices to show that

$$f\left(\left(\begin{smallmatrix}\varpi & \\ & 1\end{smallmatrix}\right)k\right) = w \quad \text{for all } k \in K_1(\mathfrak{p}^2).$$

Writing $k = \left(\begin{smallmatrix}a & b\\ c & d\end{smallmatrix}\right)$ with $c \in \mathfrak{p}^2$ and $d \in 1+\mathfrak{p}^2$, we have

$$\begin{aligned} f\left(\begin{pmatrix}\varpi & \\ & 1\end{pmatrix}k\right) &= f\left(\begin{pmatrix}\varpi & \\ & 1\end{pmatrix}\begin{pmatrix}a & b\\ c & d\end{pmatrix}\begin{pmatrix}\varpi^{-1} & \\ & 1\end{pmatrix}\begin{pmatrix}\varpi & \\ & 1\end{pmatrix}\right)\\ &= \rho_\nu\left(\begin{pmatrix}a & \varpi b\\ \varpi^{-1}c & d\end{pmatrix}\right)w = \rho_\nu\left(\begin{pmatrix}a & 0\\ 0 & 1\end{pmatrix}\right)w\\ &= \sum_{r\in k^*}\rho_\nu\left(\begin{pmatrix}a & \\ & 1\end{pmatrix}\right)f_r = \sum_{r\in k^*} f_{a^{-1}r} = w, \end{aligned}$$

as needed. This shows that $\pi_\nu^{K_1(\mathfrak{p}^2)} \neq 0$, so the conductor divides $\mathfrak{p}^2$. There are various ways to see that the conductor is exactly $\mathfrak{p}^2$. When $n > 1$, it is straightforward to show that a continuous irreducible n-dimensional complex representation of the Weil group of F has Artin conductor of exponent at least n (see, for example, [Gross and Reeder 2010, Equation (1)]). So by the local Langlands correspondence, the conductor of any supercuspidal representation of $\mathrm{GL}_n(F)$ is divisible by $\mathfrak{p}^n$, giving the desired conclusion here when $n = 2$. For an elementary proof in the present situation, one can observe that a function $f \in \text{c-Ind}_{ZK}^G(\rho_\nu)$ supported on a coset ZKx is $K_1(\mathfrak{p})$-invariant if and only if ρ_ν is trivial on $K \cap xK_1(\mathfrak{p})x^{-1}$. Using the double coset decomposition

$$G = \bigcup_{n\geq 0} ZK\begin{pmatrix}\varpi^n & \\ & 1\end{pmatrix}K = \bigcup_{n\geq 0}\bigcup_{\delta\in\overline{K}/\overline{K_1(\mathfrak{p})}} ZK\begin{pmatrix}\varpi^n & \\ & 1\end{pmatrix}\delta K_1(\mathfrak{p})$$

(we may use the representatives $\delta \in \{1\} \cup \left\{\left(\begin{smallmatrix}y & 1\\ 1 & 0\end{smallmatrix}\right) \,\middle|\, y \in \mathfrak{o}/\mathfrak{p}\right\}$; see, for example, [Knightly and Li 2006, Lemma 13.1]), it suffices to consider $x = \left(\begin{smallmatrix}\varpi^n & \\ & 1\end{smallmatrix}\right)\delta$, and one checks that in each case ρ_ν is *not* trivial on $K \cap xK_1(\mathfrak{p})x^{-1}$, so $\pi_\nu^{K_1(\mathfrak{p})} = \{0\}$. $\square$

3.3. ***Matrix coefficient of the new vector.*** Generally, if π is a supercuspidal representation with unit new vector v and formal degree d_π, the function

$$g \mapsto d_\pi \overline{\langle \pi(g)v, v\rangle}$$

can be used to define a projection operator onto $\mathbb{C}v$.

In the present context, if f is the new vector defined in Proposition 3.1, one finds easily that $\|f\|^2 = (q-1)$, so with the standard normalization $\mathrm{meas}(\overline{K}) = 1$, by (3-2) we have

$$\Phi_\nu(g) \stackrel{\text{def}}{=} d_{\pi_\nu} \overline{\left\langle \pi_\nu(g)\frac{f}{\|f\|}, \frac{f}{\|f\|}\right\rangle} = \overline{\langle \pi_\nu(g) f, f\rangle}. \tag{3-5}$$

By Proposition 1.1,

$$\mathrm{Supp}(\Phi_\nu) = \begin{pmatrix} \varpi^{-1} & \\ & 1 \end{pmatrix} ZK \begin{pmatrix} \varpi & \\ & 1 \end{pmatrix},$$

and for $g = \left(\begin{smallmatrix} \varpi^{-1} & \\ & 1 \end{smallmatrix}\right) h \left(\begin{smallmatrix} \varpi & \\ & 1 \end{smallmatrix}\right) \in \mathrm{Supp}(\Phi_\nu)$,

$$\Phi_\nu(g) = \overline{\langle \rho_\nu(h)w, w\rangle}_V \tag{3-6}$$

for w as in (3-4). This is computed as follows.

Theorem 3.2. *Let* $h = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in G(k) = \mathrm{GL}_2(k)$, *and let* $w \in V$ *be the function defined in* (3-4). *Then*

$$\langle \rho_\nu(h)w, w\rangle_V = \begin{cases} (q-1)\nu(d) & \text{if } b = c = 0, \\ -\nu(d) & \text{if } c = 0, b \neq 0, \\ -\displaystyle\sum_{\substack{\alpha \in L^* \\ \alpha + \bar{\alpha} = \frac{aN(\alpha)}{\det h} + d}} \nu(\alpha) & \text{if } c \neq 0. \end{cases}$$

Remark. The sum may be evaluated using Proposition 3.3 below.

Proof. To ease notation, we drop the subscript V from the inner product. Suppose $h = \left(\begin{smallmatrix} a & b \\ 0 & d \end{smallmatrix}\right) \in B(k)$. Then applying Theorem 2.7,

$$\langle \rho_\nu(h)w, w\rangle = \sum_{r,s \in k^*} \langle \rho_\nu(h) f_r, f_s\rangle = \sum_{s \in k^*} \langle \rho_\nu(h) f_{ad^{-1}s}, f_s\rangle = \nu(d) \sum_{s \in k^*} \psi(bd^{-1}s).$$

This gives

$$\langle \rho_\nu(h)w, w\rangle = \begin{cases} (q-1)\nu(d) & \text{if } b = 0, \\ -\nu(d) & \text{if } b \neq 0. \end{cases}$$

Now suppose $h = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in G(k) - B(k)$. Then by Theorem 2.7,

$$\begin{aligned}\langle \rho_\nu(h)w, w\rangle &= \sum_{r,s\in k^*} \phi(s, r; h) \\ &= -\frac{1}{q} \sum_{r,s\in k^*} \psi\big(c^{-1}(sa + rd)\big) \sum_{\substack{\alpha\in L^* \\ N(\alpha)=sr^{-1}\det h}} \psi\big(-rc^{-1}(\alpha + \bar{\alpha})\big)\nu(\alpha).\end{aligned}$$

Let $l = sr^{-1}$, so $s = rl$. From the previous display we have

$$\begin{aligned}\langle \rho_\nu(h)w, w\rangle &= -\frac{1}{q} \sum_{r\in k^*}\sum_{l\in k^*} \psi\big(c^{-1}(rla + rd)\big) \sum_{N(\alpha)=l\det h} \psi\big(-rc^{-1}(\alpha + \bar{\alpha})\big)\nu(\alpha) \\ &= -\frac{1}{q} \sum_{l\in k^*} \sum_{N(\alpha)=l\det h} \nu(\alpha) \sum_{r\in k^*} \psi\big(rc^{-1}(al + d - (\alpha + \bar{\alpha}))\big) \\ &= -\frac{1}{q} \sum_{l\in k^*} \sum_{N(\alpha)=l\det h} \nu(\alpha) \sum_{r\in k^*} \psi\big(r(al + d - (\alpha + \bar{\alpha}))\big).\end{aligned}$$

There are two cases for the inner sum:

$$\sum_{r\in k^*} \psi\big(r(al + d - (\alpha + \bar{\alpha}))\big) = \begin{cases} -1 & \text{if } al + d - (\alpha + \bar{\alpha}) \neq 0, \\ q - 1 & \text{if } al + d - (\alpha + \bar{\alpha}) = 0. \end{cases}$$

Therefore,

$$\begin{aligned}\big\langle \rho_\nu(h)w, w\big\rangle &= -\frac{1}{q} \sum_{l\in k^*} \Bigg(\sum_{\substack{N(\alpha)=l\det h \\ \alpha+\bar{\alpha}\neq al+d}} -\nu(\alpha) + \sum_{\substack{N(\alpha)=l\det h \\ \alpha+\bar{\alpha}=al+d}} (q-1)\nu(\alpha) \Bigg) \\ &= -\frac{1}{q} \sum_{l\in k^*} \Bigg(- \sum_{N(\alpha)=l\det h} \nu(\alpha) + q \sum_{\substack{N(\alpha)=l\det h \\ \alpha+\bar{\alpha}=al+d}} \nu(\alpha) \Bigg).\end{aligned}$$

By Lemma 2.2, the first sum in the big parentheses vanishes. So

$$\big\langle \rho_\nu(h)w, w\big\rangle = -\frac{1}{q} \sum_{l\in k^*} q \sum_{\substack{N(\alpha)=l\det h \\ \alpha+\bar{\alpha}=al+d}} \nu(\alpha) = - \sum_{\substack{\alpha\in L^* \\ \alpha+\bar{\alpha}=\frac{aN(\alpha)}{\det h}+d}} \nu(\alpha), \tag{3-7}$$

as claimed. □

This sum can be refined as follows.

Proposition 3.3. *For $l \in k^*$ and $h = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in G(k)$ with $c \neq 0$, define*

$$p_{h,l}(X) = X^2 - \left(\frac{al}{\det h} + d\right)X + l \in k[X]. \tag{3-8}$$

Then

$$\langle \rho_\nu(h)w, w\rangle = -\sum_{\substack{l\in k^* \\ p_{h,l}(X)\text{ irred. over } k}} \sum_{\substack{\text{roots } \alpha \text{ of} \\ p_{h,l}(X) \text{ in } L^*}} \nu(\alpha) - \sum_{\substack{l\in k^* \\ p_{h,l}(X)=(X-\alpha)^2 \text{in } k[X]}} \nu(\alpha). \tag{3-9}$$

Proof. If $\alpha \in L^*$ contributes to the sum (3-7), then it is a root of

$$X^2 - \left(\frac{aN(\alpha)}{\det h} + d\right)X + N(\alpha).$$

Conversely, given $l \in k^*$, a root α_l of $p_{h,l}(X)$ contributes to (3-7) if and only if $p_{h,l}(X) = (X-\alpha_l)(X-\bar{\alpha}_l)$ in $L[X]$, which is the case if and only if either $p_{h,l}(X)$ is irreducible or $p_{h,l}(X) = (X-\alpha_l)^2$ with $\alpha_l \in k^*$. The proposition now follows. □

3.4. ***Motivation.*** Although specific global applications of the above formulas are beyond the scope of this article, perhaps a few words of motivation will be helpful. The two functions ϕ_ν and Φ_ν of (3-3) and (3-5) serve slightly different purposes. The former is simpler and hence easier to work with. Taken as a local component of a global test function, it is well suited for use in the Arthur–Selberg trace formula, for example if one is interested in detecting those automorphic representations $\pi = \bigotimes_w \pi_w$ of the adelic group $\mathrm{GL}_2(\mathbf{A}_\mathbb{Q})$ with the local condition $\pi_w \cong \pi_\nu$ at a given finite place w (e.g., to obtain a dimension formula for the associated space of classical newforms). This method was treated in detail recently by Palm [2012]. On the other hand, as mentioned in the Introduction, the matrix coefficient Φ_ν gives rise to an operator projecting onto the span of the newforms attached to the global representations π as above (see [Knightly and Li 2012, Section 2.5]). It can be used in variants of the trace formula to extract finer information, like Fourier coefficients or L-values of these newforms. Of course, the utility of Φ_ν in explicit computation is limited by the complexity of the sum in Theorem 3.2.

4. Consideration of central character

The supercuspidal representations which have a given central character ω occur naturally together as irreducible subrepresentations of the right regular representation of $G(F)$ on the space of L^2 functions $f : G(F) \to \mathbb{C}$ that transform under the center by $\bar{\omega}$. It is often the case in number theory that one can achieve a certain amount of simplification by simultaneously treating all objects in a family via averaging. Here we sum the trace functions ϕ_ν from (3-3) over all isomorphism classes of depth-zero supercuspidal π_ν with a given central character, and similarly for the new vector matrix coefficients Φ_ν of (3-5).

Let ω be a unitary character of F^*, and let S_ω denote the set of isomorphism classes of depth-zero supercuspidal representations of $\mathrm{GL}_2(F)$ with central character

ω. In order for S_ω to be nonempty, ω must be trivial on $1+\mathfrak{p}$. In fact we have

$$|S_\omega| = \begin{cases} P_\omega/2 & \text{if } \omega|_{(1+\mathfrak{p})} = 1, \\ 0 & \text{otherwise,} \end{cases} \tag{4-1}$$

for P_ω as in Proposition 2.3. (Note that ν and ν^q have the same restriction to k^* and $\rho_\nu \cong \rho_{\nu^q}$.)

Assuming $\omega|_{(1+\mathfrak{p})}$ is trivial, consider the sum of the trace functions ϕ_ν defined in (3-3). In view of the fact that $\phi_\nu = \phi_{\nu^q}$, we define

$$\phi_\omega = \frac{1}{2} \sum_{\nu \in [\omega]} \phi_\nu, \tag{4-2}$$

with notation as in Proposition 2.4. We can make it explicit with the following.

Theorem 4.1. *Suppose q is odd and $\omega^{(q-1)/2}$ is nontrivial. Then for $x \in G(k)$,*

$$\sum_{\nu \in [\omega]} \operatorname{tr} \rho_\nu(x) = \begin{cases} (q^2-1)\omega(x) & \text{if } x \in Z, \\ -(q+1)\omega(z) & \text{if } x = zu,\ z \in Z,\ u \in U,\ u \neq 1, \\ 0 & \text{if no conjugate of } x \text{ is in } ZU. \end{cases}$$

Suppose q is odd and $\omega^{(q-1)/2}$ is trivial. Then with T as in (2-9),

$$\sum_{\nu \in [\omega]} \operatorname{tr} \rho_\nu(x) = \begin{cases} (q-1)^2\omega(x) & \text{if } x \in Z, \\ -(q-1)\omega(z) & \text{if } x = zu,\ z \in Z,\ u \in U,\ u \neq 1, \\ 4\omega(x^{(q+1)/2}) & \text{if } x \in T - Z,\ x^{(q^2-1)/2} = 1, \\ 0 & \text{if } x \in T - Z,\ x^{(q^2-1)/2} = -1, \text{ or if} \\ & \text{no conjugate of } x \text{ is in } T \cup ZU. \end{cases}$$

Suppose q is even. Then

$$\sum_{\nu \in [\omega]} \operatorname{tr} \rho_\nu(x) = \begin{cases} q(q-1)\omega(x) & \text{if } x \in Z, \\ -q\omega(z) & \text{if } x = zu,\ z \in Z,\ u \in U,\ u \neq 1, \\ 2\omega(N(x)^{1/2}) & \text{if } x \in T - Z, \\ 0 & \text{if no conjugate of } x \text{ is in } T \cup ZU. \end{cases}$$

Proof. This follows immediately by examining the various cases using (2-20) and Proposition 2.4. □

Likewise, for a depth-zero supercuspidal representation $\pi = \pi_\nu$, let $\Phi_\pi = \Phi_\nu$ be the matrix coefficient defined in (3-5). Define a function Φ_ω on G by

$$\Phi_\omega(g) = \sum_{\pi \in S_\omega} \Phi_\pi(g) = \frac{1}{2} \sum_{\nu \in [\omega]} \overline{\langle \rho_\nu(h)w, w \rangle_V} \tag{4-3}$$

for $g = \left(\begin{smallmatrix} \varpi^{-1} & \\ & 1 \end{smallmatrix}\right) h \left(\begin{smallmatrix} \varpi & \\ & 1 \end{smallmatrix}\right)$ with $h \in ZK$. In principle, this function can be used to define an operator that projects the automorphic spectrum of $\mathrm{GL}_2(\mathbf{A}_\mathbb{Q})$ onto the span

of those newforms of a given weight and level p^2 that correspond to automorphic representations which are unramified away from p and are supercuspidal (as opposed to special or principal series) at p.

One can evaluate Φ_ω via the following.

Theorem 4.2. *Suppose $h = \left(\begin{smallmatrix} a & \\ & d \end{smallmatrix}\right) \in G(k)$ is diagonal. Then*

$$\sum_{\nu\in[\omega]} \langle \rho_\nu(h)w, w\rangle_V = \begin{cases} (q^2-1)\omega(d) & \text{if } q \text{ is odd and } \omega^{(q-1)/2} \text{ is nontrivial,} \\ (q-1)^2\omega(d) & \text{if } q \text{ is odd and } \omega^{(q-1)/2} \text{ is trivial,} \\ q(q-1)\omega(d) & \text{if } q \text{ is even.} \end{cases} \tag{4-4}$$

If $h = \left(\begin{smallmatrix} a & b \\ 0 & d \end{smallmatrix}\right) \in B(k)$ with $b \neq 0$, then

$$\sum_{\nu\in[\omega]} \langle \rho_\nu(h)w, w\rangle_V = \begin{cases} -(q+1)\omega(d) & \text{if } q \text{ is odd and } \omega^{(q-1)/2} \text{ is nontrivial,} \\ -(q-1)\omega(d) & \text{if } q \text{ is odd and } \omega^{(q-1)/2} \text{ is trivial,} \\ -q\omega(d) & \text{if } q \text{ is even.} \end{cases} \tag{4-5}$$

If $g \in G(k) - B(k)$, then the sum is given by (4-6) below.

Proof. Suppose $h = \left(\begin{smallmatrix} a & 0 \\ 0 & d \end{smallmatrix}\right)$ is diagonal. Then by Theorem 3.2, we can write

$$\sum_{\nu\in[\omega]} \langle \rho_\nu(h)w, w\rangle = (q-1) \sum_{\nu\in[\omega]} \nu(d).$$

Applying Proposition 2.4 now gives (4-4), using the fact that $d \in k^*$.

Similarly, if $h = \left(\begin{smallmatrix} a & b \\ 0 & d \end{smallmatrix}\right)$ with $b \neq 0$, then applying Theorem 3.2,

$$\sum_{\nu\in[\omega]} \langle \rho_\nu(h)w, w\rangle = -\sum_{\nu\in[\omega]} \nu(d).$$

Using Proposition 2.4, this gives (4-5).

Now suppose $h = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in G(k) - B(k)$. By Proposition 3.3,

$$\sum_{\nu\in[\omega]} \langle \rho_\nu(h)w, w\rangle = -\sum_{\substack{l\in k^* \\ p_{h,l}(X) \text{ irred.}}} \sum_{\substack{\text{roots } \alpha \text{ of} \\ p_{h,l}(X) \text{ in } L^*}} \sum_{\nu\in[\omega]} \nu(\alpha) - \sum_{\substack{l\in k^* \\ p_{h,l}(X)=(X-\alpha)^2}} \sum_{\nu\in[\omega]} \nu(\alpha),$$

where $p_{h,l}(X) = X^2 - \big((al/\det h) + d\big)X + l \in k[X]$. If we fix one root $\alpha_l \in L^*$ of $p_{h,l}(X)$ for each l, then the above is

$$= -2 \sum_{\substack{l\in k^* \\ p_{h,l}(X) \text{ irred.}}} \sum_{\nu\in[\omega]} \nu(\alpha_l) - P_\omega \sum_{\substack{l\in k^* \\ p_{h,l}(X)=(X-\alpha_l)^2}} \omega(\alpha_l), \tag{4-6}$$

where $P_\omega = \big|[\omega]\big|$ as in Proposition 2.3. We have used the fact that

$$\sum_{\nu\in[\omega]} \nu(\alpha_l) = \sum_{\nu\in[\omega]} \nu(\bar{\alpha}_l),$$

since ν and ν^q both belong to $[\omega]$. Once again, (4-6) can be evaluated on a case-by-case basis using Proposition 2.4.

For instance, suppose q is odd. If $\omega^{(q-1)/2}$ is nontrivial, the first term of (4-6) vanishes. If $(al + d \det h) = 0$, then l makes no contribution to the second term. In particular, the term vanishes if $a = d = 0$. Generally, at most two l contribute to the second term when q is odd since $2\alpha_l = ((al/\det h) + d)$ implies that l satisfies the quadratic equation $l = \alpha_l^2 = \frac{1}{4}\big((al/\det h) + d\big)^2$.

On the other hand, if q is even, a given $l \in k^*$ contributes to the second term of (4-6) if and only if $(al + d \det h) = 0$ since $(X - \alpha_l)^2 = X^2 - \alpha_l^2$, and as remarked earlier every l is a square when q is even. □

5. Examples

Take $q = 2$. By (4-1), there is a unique supercuspidal representation π of $\mathrm{GL}_2(\mathbb{Q}_2)$ of conductor 2^2. As mentioned before, the cuspidal representation of $\mathrm{GL}_2(\mathbb{F}_2) \cong S_3$ is the character ρ sending a matrix g to $(-1)^{|g|+1}$, where $|g|$ is the order of g in the finite group. Explicitly, ρ sends $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} 1 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ to $1, -1, -1, -1, 1, 1$ respectively. This one-dimensional representation of course coincides with its matrix coefficient:

$$\big\langle \rho(h)w, w \big\rangle = \rho(h)\langle w, w\rangle = \rho(h).$$

Indeed, one may verify that Theorem 3.2 recovers ρ when $q = 2$. For example, consider $h = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 1 \end{smallmatrix}\right)$. The polynomial (3-8) becomes $p_{h,1}(X) = X^2 + X + 1$. If $\theta \in \mathbb{F}_4^*$ is a root, then $\nu(\theta) = \exp(2\pi i/3)$ defines a primitive character. By (3-9),

$$\rho(h) = -\big(\nu(\theta) - \nu(\theta^2)\big) = 1.$$

By (3-6), the matrix coefficient Φ attached to the new vector of π has the simple expression

$$\Phi(g) = \begin{cases} \rho(h) & \text{if } g = z\left(\begin{smallmatrix} 2^{-1} & \\ & 1 \end{smallmatrix}\right) h \left(\begin{smallmatrix} 2 & \\ & 1 \end{smallmatrix}\right) \in Z\left(\begin{smallmatrix} 2^{-1} & \\ & 1 \end{smallmatrix}\right) K \left(\begin{smallmatrix} 2 & \\ & 1 \end{smallmatrix}\right), \\ 0 & \text{otherwise.} \end{cases} \tag{5-1}$$

Now consider $k = \mathbb{F}_5$ and $L = \mathbb{F}_{25}$. Then $L = k[\theta]$, where $\theta^2 = 2$. One finds that $1 + 2\theta$ generates the cyclic group L^*, so the characters of L^* are the maps

$$\nu_n(1 + 2\theta) = \zeta^n \quad (n \in \mathbb{Z}/24\mathbb{Z}),$$

where $\zeta = \exp(2\pi i/24)$. Note that ν_n is primitive if and only if $6 \nmid n$. There are exactly four characters of k^*, given by

$$\omega_n(3) = i^n \quad (n \in \mathbb{Z}/4\mathbb{Z}).$$

Noting that $\nu_n(3) = \nu_n((1+2\theta)^6) = \zeta^{6n} = i^n$, we see that $\nu_n|_{k^*} = \omega_n$. So the equivalence classes of primitive characters of L^* are as follows:

$$[\omega_0] = \{\nu_4, \nu_{20}, \nu_8, \nu_{16}\}, \qquad [\omega_1] = \{\nu_1, \nu_5, \nu_9, \nu_{21}, \nu_{13}, \nu_{17}\},$$
$$[\omega_2] = \{\nu_2, \nu_{10}, \nu_{14}, \nu_{22}\}, \qquad [\omega_3] = \{\nu_3, \nu_{15}, \nu_7, \nu_{11}, \nu_{19}, \nu_{23}\},$$

where each primitive character is listed alongside its conjugate. The above illustrates Proposition 2.3. As a simple illustration of Theorem 4.2, we now show that

$$\sum_{\nu \in [\omega_0]} \left\langle \rho_\nu\left(\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)\right) w, w \right\rangle = 0.$$

For $h = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ the polynomial (3-8) becomes $p_{h,l}(X) = X^2 + l$. The second term in (4-6) vanishes. Thus an element $l \in k^*$ contributes to (4-6) only if $-l$ is a quadratic nonresidue in k. The roots of $X^2 + 2$ are

$$\alpha_2 = (1+2\theta)^3 \quad \text{and} \quad \bar{\alpha}_2 = (1+2\theta)^{15}.$$

The roots of $X^2 + 3$ are

$$\alpha_3 = \theta = (1+2\theta)^9 \quad \text{and} \quad \bar{\theta} = (1+2\theta)^{21}.$$

For any $\nu \in [\omega_0]$, $\nu(\alpha_j)$ is a power of $\zeta^{12} = -1$ when $4|n$. Hence $\nu^5(\alpha_j) = \nu(\alpha_j)$. Thus in this case (4-6) equals

$$-2\big[2\nu_4(\alpha_2) + 2\nu_8(\alpha_2) + 2\nu_4(\alpha_3) + 2\nu_8(\alpha_3)\big] = -4[-1+1-1+1] = 0.$$

Acknowledgements

This work is based on Ragsdale's Master's thesis at the University of Maine. We thank the referee for carefully reading the manuscript and making several suggestions and corrections which have greatly improved the exposition. Ragsdale thanks the Department of Mathematics and Statistics of the University of Maine for summer support during the final stages of writing this paper. Both authors were supported by NSF grant DMS 0902145.

References

[Bump 1997] D. Bump, *Automorphic forms and representations*, Cambridge Studies in Advanced Mathematics **55**, Cambridge University Press, 1997. MR 97k:11080 Zbl 0868.11022

[Bushnell and Henniart 2006] C. J. Bushnell and G. Henniart, *The local Langlands conjecture for* GL(2), Grundlehren der Math. Wiss. **335**, Springer, Berlin, 2006. MR 2007m:22013 Zbl 1100.11041

[Bushnell and Henniart 2014] C. J. Bushnell and G. Henniart, "Langlands parameters for epipelagic representations of GL_n", *Math. Ann.* **358**:1–2 (2014), 433–463. arXiv 1302.4304

[Bushnell and Kutzko 1993] C. J. Bushnell and P. C. Kutzko, *The admissible dual of* GL(N) *via compact open subgroups*, Annals of Mathematics Studies **129**, Princeton University Press, 1993. MR 94h:22007 Zbl 0787.22016

[Casselman 1973] W. Casselman, "On some results of Atkin and Lehner", *Math. Ann.* **201** (1973), 301–314. MR 49 #2558 Zbl 0239.10015

[Gross and Reeder 2010] B. H. Gross and M. Reeder, "Arithmetic invariants of discrete Langlands parameters", *Duke Math. J.* **154**:3 (2010), 431–508. MR 2012c:11252 Zbl 1207.11111

[Kim 2007] J.-L. Kim, "Supercuspidal representations: an exhaustion theorem", *J. Amer. Math. Soc.* **20**:2 (2007), 273–320. MR 2008c:22014 Zbl 1111.22015

[Knightly and Li 2006] A. Knightly and C. Li, *Traces of Hecke operators*, Mathematical Surveys and Monographs **133**, American Mathematical Society, Providence, RI, 2006. MR 2008g:11090 Zbl 1120.11024

[Knightly and Li 2012] A. Knightly and C. Li, "Modular L-values of cubic level", *Pacific J. Math.* **260**:2 (2012), 527–563. MR 3001804 Zbl 06136442

[Mautner 1964] F. I. Mautner, "Spherical functions over $\mathfrak{b}$-adic fields, II", *Amer. J. Math.* **86** (1964), 171–200. MR 29 #3582 Zbl 0135.17204

[Palm 2012] M. Palm, *Explicit* GL(2) *trace formulas and uniform, mixed Weyl laws*, Ph.D. thesis, Göttingen, 2012. arXiv 1212.4282

[Piatetski-Shapiro 1983] I. Piatetski-Shapiro, *Complex representations of* $\mathrm{GL}(2, K)$ *for finite fields* K, Contemporary Mathematics **16**, American Mathematical Society, Providence, RI, 1983. MR 84m:20046 Zbl 0513.20026

[Reeder 1991] M. Reeder, "Old forms on GL_n", *Amer. J. Math.* **113**:5 (1991), 911–930. MR 92i:22018 Zbl 0758.11027

[Stevens 2008] S. Stevens, "The supercuspidal representations of p-adic classical groups", *Invent. Math.* **172**:2 (2008), 289–352. MR 2010e:22008 Zbl 1140.22016

Received: 2013-08-14 Revised: 2013-11-12 Accepted: 2013-11-16

knightly@math.umaine.edu *Department of Mathematics and Statistics, University of Maine, 5752 Neville Hall, Room 333, Orono, ME 04469-5752, United States*

cwragsda@syr.edu *Department of Mathematics and Statistics, University of Maine, 5752 Neville Hall, Room 333, Orono, ME 04469-5752, United States*

INVOLVE 7:5 (2014)
dx.doi.org/10.2140/involve.2014.7.691

The sock matching problem

Sarah Gilliand, Charles Johnson, Sam Rush and Deborah Wood

(Communicated by Jim Haglund)

When matching socks after doing the laundry, how many unmatched socks can appear in the process of drawing one sock at a time from the basket? By connecting the problem of sock matching to the Catalan numbers, we give the probability that k unmatched socks appear. We also show that, for each fixed k, this probability approaches 1 as the number of socks becomes large enough. The relation between the number of socks and the k for which a given probability is first reached is also discussed, but a complete answer is open.

1. Introduction

In any load of clothes to be washed by a college student, there are inevitably a variety of socks tossed in with all the other garments. By the time the clothes come out of the dryer, the socks have been thoroughly mixed in, hiding underneath shirts or in pant legs. The game of matching then begins: does the sock you just picked randomly out of the pile match any of the others you've already removed from the pile? How big is your stack of unmatched socks going to get? This creates a scenario in which there can be k unmatched socks out of n pairs. We wish to determine the likelihood of obtaining a maximum of k unmatched socks while folding a pile of laundry containing the n pairs of socks. We assume that each pair of socks is complete and unique, and that socks are drawn randomly, one at a time.

2. Background

Catalan numbers. The Catalan numbers are a sequence named after the Belgian mathematician Eugène Charles Catalan (1814–1894), who, in an 1838 paper, first defined them in their modern form [Larcombe 1999]. However, he was not the first to discover the numbers. In fact, according to J. J. Luo [1988], the Chinese mathematician Antu Ming (c. 1692–1763) discovered the numbers before anyone else [Koshy 2009; Larcombe 1999]. Leonard Euler (1707–1783) published a

MSC2010: primary 05A15, 05A16; secondary 03B48, 00A69.
Keywords: Catalan numbers, sock matching, Dyck paths.
This work supported, in part, by a QEP Mellon grant to the College of William & Mary.

recursive definition of the sequence in 1761 [Koshy 2009], almost eighty years *before* the man after whom the sequence was eventually named. He, like Catalan, discovered the sequence while investigating the problem of cutting polygons into triangles with diagonals that do not cross [Koshy 2009; Larcombe 1999]. Indeed, the Catalan numbers "have [a] delightful propensity for popping up unexpectedly, particularly in combinatorial problems" (Martin Gardner, as quoted in [Koshy 2009, p. vii]). Besides triangles within polygons, the Catalan numbers can be found in exponentiations, Pascal's triangle, binary trees, diagonals in frieze patterns, partitions, the ballot problem, folding paper, and even baseball [Conway and Guy 1996; Koshy 2009].

Definition. The Catalan numbers are defined by the sequence:

$$C_0 = 1, \quad C_{n+1} = \sum_{i=0}^{n} C_i C_{n-i}.$$

The generating function for the Catalan numbers is

$$\sum_{n=0}^{\infty} C_n x^n = \frac{2}{1+\sqrt{1-4x}},$$

from which we can determine that $C_n = \binom{2n}{n}/(n+1)$.

Examples of problems in which Catalan numbers arise. One example of the Catalan numbers is that C_n is the number of paths in an $n \times n$ grid starting from the lower left corner and ending in the upper right corner using only moves up and to the right without moving across the diagonal (called *Dyck paths*). The recursive nature of this example arises from visits to locations along the diagonal.

Another example is the number of paths from $(0, 0)$ to $(2n, 0)$ on the Cartesian plane using only moves to the northeast and southeast that do not move below the x-axis. The recursive nature of this example arises from visits to the x-axis.

Relation to our problem. Take the situation that one has n pairs of socks to match and none yet drawn and left unmatched. At this point, there are C_n ways in which socks can be drawn one at a time and set aside as unmatched until they find a match (assuming that the order in which different pairs of socks are matched is not considered). Because of this connection to the Catalan numbers, we could use information already known about the integer series in our investigation of the sock matching problem.

3. Initial observations

Every time a sock is drawn from the laundry pile, there are two possible outcomes: it could either match a sock that has already been drawn, or it is temporarily a lone

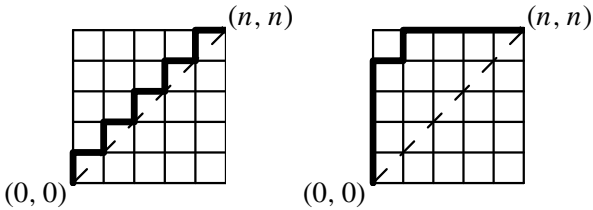


Figure 1. Left: $k = 1$; right: $k = n - 1$.

sock whose match has not yet been encountered. It is this aspect of the problem that allows us to use the grid visualization (Dyck paths) of the Catalan numbers as a model.

Every time a move is added to a path on the grid, there are also two options: it either moves one unit up or one unit to the right. Therefore, a move up can be taken to represent drawing a sock that has no match as of yet. And a move to the right represents drawing a match to some sock that has already been obtained from the laundry pile.

For example, Figure 1 (left) would result from an instance in which you continually pick a sock, and then pick its match. And Figure 1 (right) is the grid that shows the scenario in which you pick $n - 1$ socks without a single match, get a match, pick the last nonmatch, and then necessarily match the rest.

As can be seen from these grids, all of the paths that apply to this problem will begin at the origin (where no socks have yet been drawn). Because we are assuming that every sock has a match, all of the paths will also terminate at (n, n), because for every move up on the grid, there is guaranteed to be a corresponding move to the right. Whenever the path hits the line $y = x$, we can see that all socks that have been drawn thus far have a match. None of the possible paths will ever cross below the line $y = x$, because this would indicate that there have been more matches than there have been previous unmatched socks, which is impossible. Also, a grid makes it easy to examine different values of k, because whenever the path hits or crosses the line $y = x + k$, we know that at least k unmatched socks have been attained.

As opposed to using a grid, we could instead use a graph to model the paths created by drawing and matching the socks. Every time a sock without a match is pulled out, the path would move diagonally up and to the right one unit. Every time a match is obtained, it would move diagonally down and to the right. So, every return to the x-axis indicates an instance in which all socks that have been drawn have also been matched, and any time it hits or passes above the line $y = k$ indicates an instance in which at least k unmatched socks have been reached. Any path would still begin at the origin, but must terminate at the point $(2n, 0)$.

Expected value for maximum k. Drawing randomly from n pairs of unmatched socks, how many socks may one expect to find drawn but unmatched at any point in

the pairing process? Mathematically, this question asks for the average maximum value k may reach, and in terms of Dyck paths of order n, this is the average distance from the diagonal to the path.

Here, the expected value equals the number of ways in which n socks can be matched weighted by the maximum k reached on that path divided by C_n. That is,

$$E(n) = \frac{\sum_{i=1}^{n}(\text{\# of ways to match socks such that at most } i \text{ are unmatched at any time} \times i)}{C_n}.$$

The numerator is the sum of heights of all Dyck paths of order n, sequence A136439 in the Online Encyclopedia of Integer Sequences (OEIS) [Finch 2008], and it must only be divided by C_n, the number of all those paths, in order to find the average. From OEIS, as well as Bruijn, Knuth and Rice [de Bruijn et al. 1972], we have as an equation $E(n)$ for the expected maximum number of socks to be left unmatched while matching n pairs of socks:

$$E(n)=(n+1)\left[\sum_{j=1}^{n+1}\left(\sum_{i\mid j} i^0\right)\frac{n!}{(n+a+j)!(j-a)!}-2\sum_{j=1}^{n}\left(\sum_{i\mid j} i^0\right)\frac{n!}{(n+a+j)!(j-a)!}\right.$$
$$\left.+\sum_{j=1}^{n-1}\left(\sum_{i\mid j} i^0\right)\frac{n!}{(n+a+j)!(j-a)!}\right]-1.$$

Recurrence formula. In this section, we focus upon the grid model for our problem. We define $B_{n,k}$ as the number of ways to get from $(0, 0)$ to (n, n) without crossing the diagonal but reaching the line $y = x + k$. This can be thought of as the total number of ways to get at least k unmatched socks at least once during the matching process.

To do this, we must hit at least one point on the diagonal $y = x$ after $(0, 0)$, since at the very least we must hit (n, n). Let us consider the point (i, i), which is the first point on the line $y = x$ that the path visits after $(0, 0)$. For analysis, we consider three possibilities: the line hits $y = x + k$ before (i, i), the line hits $y = x + k$ after (i, i), and the line hits $y = x + k$ both before and after (i, i) (which is counted twice so we want to subtract case 3 from the other two).

Case 1: The number of ways to hit $y = x + k$ between $(0, 0)$ and (i, i) is just $B_{i,k}$. The number of ways to get from (i, i) to (n, n) without hitting $y = x$ is the same as the number of ways to pass from $(i, i + 1)$ to $(n - 1, n)$ without crossing $y = x + 1$, which is C_{n-i-1}. Therefore the number of paths for this case is $B_{i,k}C_{n-i-1}$.

Case 2: The number of ways to get from $(0, 0)$ to (i, i) without crossing $y = x$ is C_i. Then, the number of ways to get from (i, i) to (n, n) hitting $y = x + k$ but not $y = x$ is the same as the number of ways to get from $(i, i + 1)$ to $(n - 1, n)$ hitting

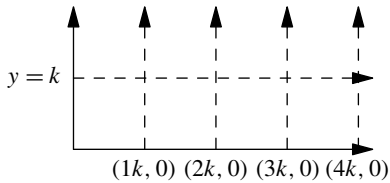


Figure 2. $\lim_{n\to\infty} P_{n,k}$.

$y = x + k$ but not crossing $y = x + 1$, which is $B_{n-i-1,k-1}$. Therefore the number of paths in this case is $C_i B_{n-i-1,k-1}$.

Case 3: In this case, we want to count trajectories that hit $y = x + k$ before and after. This is a combination of our previous cases, $B_{i,k} B_{n-i-1,k-1}$. Now, we just need to add up the total paths for all of our total cases, so our final recurrence is:

$$B_{n,k} = \sum_{i=1}^{n} \big(B_{i,k} C_{n-i-1} + C_i B_{n-i-1,k-1} - B_{i,k} B_{n-i-1,k-1} \big). \tag{1}$$

We refer to the above as the *Sock Matching Theorem*.

4. Asymptotic behavior

The question that arises from the recurrence formula is whether or not the basic patterns we see for small values of n and k hold true for all values of n and k.

Large n. In particular, we first ask if the probability of reaching a given, fixed k approaches 1 as n approaches infinity. In this section we show that it does.

Let us define $P_{n,k}$ as the probability that we reach k unmatched socks at least once in a draw of n pairs. Notice that $P_{n,k} = B_{n,k}/C_n$. We must prove that $\lim_{n\to\infty} P_{n,k} = 1$. To do this, we return to the graph model.

First, split up the graph into sections of length k as shown in Figure 2. Call the probability that the path reaches $y = k$ in the first section p_1. Since moving up at every step reaches $y = k$ in k steps, p_1 is positive. The probability that the path reaches $y = k$ in any subsequent section is dependent on where the path terminated in the prior section. However, the probability of reaching $y = k$ in any section is at least p_1 *no matter what happened in prior sections*. That is, $p_i \geq p_1$ for section i, or $1 - p_i \leq 1 - p_1$. This allows us to say that the probability we never reach $y = k$, which is $1 - P_{n,k}$, is at most $\prod_{i=1}^{2n/k} (1 - p_1)$. Therefore,

$$\lim_{n\to\infty} 1 - P_{n,k} = \lim_{n\to\infty} \prod_{i=1}^{2n/k} (1 - p_1) = \lim_{n\to\infty} (1 - p_1)^{2n/k} = 0.$$

Therefore, $\lim_{n\to\infty} P_{n,k} = 1 - \lim_{n\to\infty}(1 - P_{n,k}) = 1 - 0 = 1$.

k	$P_{n,k} \geq 0.99$	$P_{n,k} \geq 0.999$	$P_{n,k} \geq 0.9999$
1	1	1	1
2	6	8	10
3	12	16	20
4	20	27	33
5	30	39	49
6	41	54	67
7	55	72	88
8	70	91	>93
9	86	>93	>93

Table 1. First value of n at which $P_{n,k}$ has reached a certain threshold, for various values of k.

A quadratic relationship? In the previous section, we considered the asymptotic behavior of the model as n approaches infinity given a fixed k. For our next step, we instead fixed the probability, $P_{n,k}$ in order to discover the behavior of k as n again approaches infinity. We started by setting the probability at 0.99, and from the tables of data generated by a computer program we acquired the necessary data to speculate.

This investigation proved intriguing but unsatisfying. For $1 < k < 6$ when $P_{n,k} = 0.99$, the relationship between k and the first n for which the probability of reaching k is greater than or equal to $P_{n,k}$ can be described by the quadratic equation $n = k^2 + k$. This, however, fails for all other values of k and all other probabilities. When the constant probability is 0.999, n increases more rapidly as k increases, and for the constant probability 0.9999, the rate of increase for n rises even more. Our data, moreover, end at $k = 8$ for 0.999 and at $k = 7$ for 0.9999. Although the patterns in the Table 1 suggest a quadratic relationship exists in this context, a specific equation is not sustained by high values of k or the given probability.

References

[de Bruijn et al. 1972] N. G. de Bruijn, D. E. Knuth, and S. O. Rice, "The average height of planted plane trees", pp. 15–22 in *Graph theory and computing*, edited by R. C. Read, Academic Press, New York, 1972. MR 58 #21737 Zbl 0247.05106

[Conway and Guy 1996] J. H. Conway and R. K. Guy, *The book of numbers*, Springer, New York, 1996. MR 98g:00004 Zbl 0866.00001

[Finch 2008] S. Finch, "Sum of heights of all 1-watermelons with wall of length $2n$", in *The online encyclopedia for integer sequences*, 2008.

[Koshy 2009] T. Koshy, *Catalan numbers with applications*, Oxford University Press, Oxford, 2009. MR 2010g:05008 Zbl 1159.05001

[Larcombe 1999] P. J. Larcombe, "The 18th century Chinese discovery of the Catalan numbers", *Mathematical Spectrum* (1999), 5–7.

[Luo 1988] J. J. Luo, "Antu Ming, the first inventor of Catalan numbers in the world", *Neimenggu Daxue Xuebao* **19** (1988), 239–245. In Chinese.

Received: 2013-08-21 Accepted: 2013-11-25

scgilliand@email.wm.edu *Department of Biology, The College of William & Mary, College Station Unit 3011, P.O. Box 8793, Williamsburg, VA 23187, United States*

crjohn@wm.edu *Department of Mathematics, The College of William & Mary, P.O. Box 8795, Williamsburg, VA 23187, United States*

samuel.j.rush@gmail.com *Department of Computer Science, California Institute of Technology, 1200 East California Boulevard, MS 305-16, Pasadena, CA 91125, United States*

dwood@email.wm.edu *Department of Mathematics, The College of William & Mary, College Station Unit 4085, P.O. Box 8793, Williamsburg, VA 23187, United States*

INVOLVE 7:5 (2014)
dx.doi.org/10.2140/involve.2014.7.699

Superlinear convergence via mixed generalized quasilinearization method and generalized monotone method

Vinchencia Anderson, Courtney Bettis, Shala Brown, Jacqkis Davis, Naeem Tull-Walker, Vinodh Chellamuthu and Aghalaya S. Vatsala

(Communicated by Johnny Henderson)

The method of upper and lower solutions guarantees the interval of existence of nonlinear differential equations with initial conditions. To compute the solution on this interval, we need coupled lower and upper solutions on the interval of existence. We provide both theoretical as well as numerical methods to compute coupled lower and upper solutions by using a superlinear convergence method. Further, we develop monotone sequences which converge uniformly and monotonically, and with superlinear convergence, to the unique solution of the nonlinear problem on this interval. We accelerate the superlinear convergence by means of the Gauss–Seidel method. Numerical examples are developed for the logistic equation. Our method is applicable to more general nonlinear differential equations, including Riccati type differential equations.

1. Introduction

Qualitative study such as existence, uniqueness of nonlinear differential equations with initial and boundary conditions play an important role in modeling science and engineering problems. Explicit solutions of such nonlinear problems are rarely possible [Adams et al. 2012; Cronin 1994; Holt and Pickering 1985; Lakshmikantham et al. 1989; Jin et al. 2004]. Approximate methods such as Picard's method provide only local existence. The generalized monotone method combined with coupled lower and upper solutions provides a method to compute coupled

MSC2010: primary 34A12; secondary 34A34.
Keywords: coupled lower and upper solutions, superlinear convergence.
This work was done under the auspices of the 2013 "Smooth Transition for Advancement to Graduate Education" (STAGE) for Underrepresented Minorities in Mathematical Sciences, an undergraduate summer research and professional development experience organized by the University of Louisiana at Lafayette. STAGE is a pilot project that is supported by National Science Foundation grant DMS-1043223.

minimal and maximal solutions [Bhaskar and McRae 2002; Sokol and Vatsala 2001; Stutson and Vatsala 2011; West and Vatsala 2004]. Noel, Sheila, Zenia, Dayonna, Jasmine, Vatsala, and Sowmya [Noel et al. 2012] have developed both theoretical and numerical approaches to compute coupled minimal and maximal solutions using the idea of generalized monotone method. See [Muniswamy and Vatsala 2013; Noel et al. 2012] for details. However, the order of convergence of the sequences generated is linear. We note that the generalized monotone method is useful when the nonlinear function is the sum of increasing and decreasing functions.

In this paper, we develop a method when the nonlinear function is the sum of a convex function and a decreasing function. The sequences constructed yield quadratic convergence when the decreasing function is not present and yields linear convergence when the convex term is not present. We develop a methodology to compute coupled lower and upper solutions whose convergence rate is superlinear on any desired interval. Using the computed coupled upper and lower solutions, we can develop sequences which converge uniformly and monotonically to the unique solution of the nonlinear problem. The rate of convergence of the sequences is superlinear. In addition, the superlinear convergence can be accelerated by using the Gauss–Seidel approach. We have presented some numerical examples of the population model of single species, namely the logistic equation. Our method is applicable to more general nonlinear problems such as Ricatti type differential equation.

2. Preliminary results

In this section, we recall known definitions and results which we need to develop our main results. For that purpose, consider the first-order differential equation of the form

$$u' = f(t,u) + g(t,u), \quad u(0) = u_0 \quad \text{on } [0,T] = J, \tag{2-1}$$

where f, g lie in $C(J \times \mathbb{R}, \mathbb{R})$, the space of continuous functions from $J \times \mathbb{R}$ to $\mathbb{R}$.

Definition 2.1. The functions $v_0, w_0 \in C^1(J, \mathbb{R})$ are called *natural lower and upper solutions* of (2-1) if

$$\begin{aligned} v_0' &\le f(t, v_0) + g(t, v_0), \qquad v_0(0) \le u_0, \\ w_0' &\ge f(t, w_0) + g(t, w_0), \quad w_0(0) \ge u_0. \end{aligned}$$

Definition 2.2. The functions $v_0, w_0 \in C^1(J, \mathbb{R})$ are called *coupled lower and upper solutions* of (2-1) of type I if

$$\begin{aligned} v_0' &\le f(t, v_0) + g(t, w_0), \qquad v_0(0) \le u_0, \\ w_0' &\ge f(t, w_0) + g(t, v_0), \quad w_0(0) \ge u_0. \end{aligned}$$

Next we recall a comparison theorem which will be useful in establishing the uniqueness of the solution of (2-1). For that purpose we assume

$$f(t,u)+g(t,u)=F(t,u).$$

Theorem 2.3. *Let $v, w \in C^1(J,\mathbb{R})$ be lower and upper solutions of (2-1) respectively. Suppose that $F(t,x)-F(t,y) \le L(x-y)$ whenever $x \ge y$, and $L>0$ is a constant, then $v(0) \le w(0)$ implies that $v(t) \le w(t)$, $t \in J$.*

Corollary 2.4. *Let $p(t) \in C(J,\mathbb{R})$ be a function such that $p'(t) \le L(t)p(t)$, where $L(t) \in C(J,\mathbb{R})$. Then $p(0) \le 0$ implies $p(t) \le 0$.*

Remark. If in Corollary 2.4 all the inequalities are reversed, the conclusion holds with reversed inequality.

We define the following sector Ω for convenience. That is,

$$\Omega = \big\{(t,u) \mid v(t) \le u(t) \le w(t),\, t \in J\big\}.$$

Theorem 2.5. *Suppose $v, w \in C^1(J,\mathbb{R})$ are natural upper and lower solutions of (2-1) such that $v(t) \le w(t)$ on J and $F \in C(\Omega,\mathbb{R})$. Then there exists a solution $u(t)$ of (2-1) such that $v(t) \le u(t) \le w(t)$ on J, provided $v(0) \le u(0) \le w(0)$.*

Proof. See [Ladde et al. 1985] for details. □

Remark. If $g(t,u)$ in (2-1) is nonincreasing in u, then the existence of coupled lower and upper solutions of (2-1) on J implies that they are also natural lower and upper solutions. From Theorem 2.5 it follows that there exists a solution of (2-1) such that $v(t) \le u(t) \le w(t)$ on J, provided $v(0) \le u(0) \le w(0)$.

The next result is to prove the existence of coupled lower and upper solutions by the generalized monotone method.

Theorem 2.6. *Let $v_0, w_0 \in C^1(J,\mathbb{R})$ be coupled upper and lower solutions of type I such that $v_0(t) \le w_0(t)$ on J, and assume that f, g are elements of $C(J\times\mathbb{R},\mathbb{R})$ such that $f(t,u)$ is nondecreasing in u and $g(t,u)$ is nonincreasing in u on J.*

There exist monotone sequences $\{v_n(t)\}$ and $\{w_n(t)\}$ on J such that

$$v_n(t) \to v(t) \quad \text{and} \quad w_n(t) \to w(t)$$

uniformly and monotonically, and (v,w) are coupled minimal and maximal solutions, respectively, to (2-1). That is, (v,w) satisfy on J the equations

$$v' = f(t,v)+g(t,w), \qquad v(0)=u_0, \tag{2-2}$$

$$w' = f(t,w)+g(t,v), \qquad w(0)=u_0. \tag{2-3}$$

Here the iterative scheme is given on J by

$$v'_{n+1} = f(t, v_n) + g(t, w_n), \quad v_{n+1}(0) = u_0, \tag{2-4}$$

$$w'_{n+1} = f(t, w_n) + g(t, v_n), \quad w_{n+1}(0) = u_0. \tag{2-5}$$

Proof. See [Noel et al. 2012; Sokol and Vatsala 2001; West and Vatsala 2004] for details of the proof. □

We now give an existence result based on the generalized monotone method using natural lower and upper solutions.

Theorem 2.7. *Let $v_0, w_0 \in C^1(J, \mathbb{R})$ be natural lower and upper solutions with $v_0 \le w_0$ on J, and assume that f, g are elements of $C(J \times \mathbb{R}, \mathbb{R})$ such that $f(t, u)$ is nondecreasing in u and $g(t, u)$ is nonincreasing in u on J.*

There exist monotone sequences $\{v_n(t)\}$ and $\{w_n(t)\}$ on J such that

$$v_n(t) \to v(t) \quad \text{and} \quad w_n(t) \to w(t)$$

uniformly and monotonically, and (v, w) are coupled minimal and maximal solutions, respectively, to (2-1). That is, (v, w) satisfy

$$v' = f(t, v) + g(t, w), \quad v(0) = u_0,$$
$$w' = f(t, w) + g(t, v), \quad w(0) = u_0,$$

on J, provided also that $v_0 \le v_1$ and $w_1 \le w_0$ on J.

Proof. See [Noel et al. 2012; West and Vatsala 2004] for details of the proof. □

Noel et al. [2012] have developed computational methods to compute coupled lower and upper solutions to any desired interval by applying Theorem 2.7, by redefining the sequences on the interval J. However, the rate of convergence of these sequences is linear.

Before we recall the next result, which provides quadratic convergence, we need the Gronwall lemma which will be used to compute the rate of convergence.

Lemma 2.8 (Gronwall lemma). *Let $v \in C^1(J, \mathbb{R}^N)$ and $v' \le Av + \sigma$, where $A = (a_{ij})$ is an $N \times N$ constant matrix satisfying $a_{ij} \ge 0$, $i \ne j$, and $\sigma \in C(J, \mathbb{R}^N)$. Then we have*

$$v' \le v(0)e^{At} + \int_0^t e^{A(t-s)}\sigma(s)\,ds, \quad t \in J. \tag{2-6}$$

Theorem 2.9. *Assume that*

(i) *$v_0, w_0 \in C^1(J, \mathbb{R})$, $v_0(t) \le w_0(t)$ on J, with $v_0(t)$ and $w_0(t)$ coupled lower and upper solutions of type I for (2-1), such that $v_0(t) \le w_0(t)$ on J;*

(ii) *$f, g \in C(\Omega, \mathbb{R})$, f_u, g_u, f_{uu}, g_{uu} exist, are continuous and satisfy $f_{uu}(t, u) \ge 0$, $g_{uu}(t, u) \le 0$ for $(t, u) \in \Omega = \{t \in J \mid v_0(t) \le u \le w_0(t)\}$;*

(iii) *$g_u(t, u) \le 0$ on Ω.*

Then there exist monotone sequences $\{v_n(t)\}$, $\{w_n(t)\}$ *that converge uniformly to the unique solution of* (2-1). *The convergence is quadratic.*

Proof. See [Lakshmikantham and Vatsala 1998] for details of the proof. □

In Theorem 2.9 the iterations are as follows:

$$\begin{cases} v_n' = f(t, v_{n-1}) + f_u(t, v_{n-1})(v_n - v_{n-1}) \\ \qquad\qquad + g(t, w_{n-1}) + g_u(t, v_{n-1})(w_n - w_{n-1}), \\ v_n(0) = u_0, \end{cases} \tag{2-7}$$

$$\begin{cases} w_n' = f(t, w_{n-1}) + f_u(t, v_{n-1})(w_n - w_{n-1}) \\ \qquad\qquad + g(t, v_{n-1}) + g_u(t, v_{n-1})(v_n - v_{n-1}), \\ w_n(0) = u_0. \end{cases} \tag{2-8}$$

In this theorem the sequences $\{v_n\}$ and $\{w_n\}$ are solutions of the two linear systems of coupled equations with variable coefficients, which are not easy to compute. In the next result under a slightly weaker assumption, we obtain superlinear convergence.

Theorem 2.10. *Assume that*

(i) $v_0, w_0 \in C^1(J, \mathbb{R})$, $v_0(t) \le w_0(t)$ *on* J, *with* $v_0(t)$ *and* $w_0(t)$ *coupled lower and upper solutions of type I for* (2-1), *such that* $v_0(t) \le w_0(t)$ *on* J;

(ii) $f, g \in C(\Omega, \mathbb{R})$, f_u, g_u, f_{uu} *exist, are continuous and satisfy* $f_{uu}(t, u) \ge 0$, *for* $(t, u) \in \Omega = \{t \in J \mid v_0(t) \le u \le w_0(t)\}$;

(iii) $g_u(t, u) \le 0$ *on* Ω.

Then there exist monotone sequences $\{v_n(t)\}$, $\{w_n(t)\}$ *that converge uniformly to the unique solution of* (2-1), *and the convergence is superlinear.*

Proof. See [Muniswamy and Vatsala 2013] for details of proof. □

3. Main results

In this section we will provide a method to compute coupled lower and upper solutions of (2-1) to any desired interval when we have natural lower and upper solutions. Natural lower and upper solutions are relatively easy to compute. For example, equilibrium solutions are natural lower and upper solutions for all time. This means the solution of the nonlinear problem exists for all time by upper and lower solution method. In order to develop this method, we modify Theorem 2.10 using natural lower and upper solutions.

Theorem 3.1. *Let*

(i) $v_0, w_0 \in C^1(J, R)$, $v_0(t) \le w_0(t)$ *on* J, *with* $v_0(t)$ *and* $w_0(t)$ *natural lower and upper solutions of* (2-1), *such that* $v_0(t) \le w_0(t)$ *on* J;

(ii) *$f, g \in C(\Omega, \mathbb{R})$, f_u, g_u, f_{uu} exist, are continuous and satisfy $f_{uu}(t,u) \geq 0$ for $(t,u) \in \Omega = \{t \in J \mid v_0(t) \leq u \leq w_0(t)\}$;*

(iii) *$g_u(t,u) \leq 0$ on Ω;*

(iv) *$v_0 \leq v_1$ and $w_1 \leq w_0$ on J.*

Then there exist monotone sequences $\{v_n(t)\}$, $\{w_n(t)\}$ that converge uniformly to the unique solution of (2-1). The convergence is superlinear.

Here and in Theorem 2.10 the iterations are computed as follows:

$$v_n' = f(t, v_{n-1}) + f_u(t, w_{n-1})(v_n - v_{n-1}) + g(t, w_{n-1}), \qquad v_n(0) = u_0, \tag{3-1}$$

$$w_n' = f(t, w_{n-1}) + f_u(t, v_{n-1})(w_n - w_{n-1}) + g(t, v_{n-1}), \qquad w_n(0) = u_0. \tag{3-2}$$

Proof. The proof follows on the same lines as in Theorem 2.10. Here we briefly prove the superlinear convergence part. In order to prove superlinear convergence, we let $p_n(t) = u(t) - v_n(t)$ and $q_n(t) = w_n(t) - u(t)$ on J, where u, v_n, and w_n are solutions of (2-1), (3-1), and (3-2) respectively. It is easy to see that $p_n(0) = 0 = q_n(0)$. Using Gronwall lemma and the estimate on $|f_{uu}(+,\cdot)|$ and $|g_u(+,\cdot)|$ on J, we can prove that

$$\max_J\big(|p_n + q_n|\big) \leq L_1 \max_J\big(|p_{n-1} + q_{n-1}|^2\big) + L_2 \max_J\big(|p_{n-1} + q_{n-1}|\big),$$

where L_1 and L_2 depends on bounds of $|f_{uu}(+,\cdot)|$ and $|g_u(+,\cdot)|$ on J. If $g \equiv 0$, then $L_2 \equiv 0$, we have quadratic convergence. If $f \equiv 0$, then $L_1 \equiv 0$, which means we have linear convergence. □

Consider the following example

$$u' = u - u^2, \quad u(0) = \tfrac{1}{2}, \quad t \in [0, T], \quad T \geq 1.$$

It is easy to observe that $v_0(t) = 0$ and $w_0(t) = 1$ are natural lower and upper solutions. Starting with $v_0 = 0$ and $w_0 = 1$, which are natural lower and upper solutions, and using the iterations as in Theorem 3.1, we get

$$v_1 = 1 - \tfrac{1}{2}e^t \quad \text{and} \quad w_1 = \tfrac{1}{2}e^t.$$

We can see that

$$\begin{aligned} v_0 \leq v_1, &\qquad 0 \leq 1 - \tfrac{1}{2}e^t \quad \text{on } [0, 0.69], \\ w_1 \leq w_0, &\quad \tfrac{1}{2}e^t \leq 1 \qquad\qquad \text{on } [0, 0.69]. \end{aligned}$$

This means $v_0 \leq v_1$ and $w_1 \leq w_0$ on $[0, 0.69]$. Here $0.69 < T$. This is the motivation for our next main result: developing a method to compute coupled lower and upper solutions to any desired interval.

Theorem 3.2. *Let all the hypothesis of Theorem 3.1 hold.*

Then there exist monotone sequences $\{v_n(t)\}$ *and* $\{w_n(t)\}$ *on* J *such that*

$$v_n(t) \to v(t) \quad \text{and} \quad w_n(t) \to w(t)$$

uniformly and monotonically and (v, w) *are coupled minimal and maximal solutions, respectively to* (2-1).

The sequences $\{v_n\}$ *and* $\{w_n\}$ *are computed using the following iterative scheme*:

$$\begin{aligned} v_n' &= f(t, v_{n-1}) + f_u(t, v_{n-1})(v_n - v_{n-1}) + g(t, w_{n-1}), \qquad v_n(0) = u_0, \\ w_n' &= f(t, w_{n-1}) + f_u(t, v_{n-1})(w_n - w_{n-1}) + g(t, v_{n-1}), \quad w_n(0) = u_0. \end{aligned}$$

Proof. We compute the iterations using v_0 and w_0 in the following form:

$$\begin{aligned} v_1' &= f(t, v_0) + f_u(t, v_0)(v_1 - v_0) + g(t, w_0), \qquad v_1(0) = u_0, \\ w_1' &= f(t, w_0) + f_u(t, v_0)(w_1 - w_0) + g(t, v_0), \quad w_1(0) = u_0. \end{aligned}$$

After computing v_1 and w_1, if $v_0 \le v_1$ and $w_1 \le w_0$ on $[0, T]$, then there is nothing to prove. If not, then $v_1(t_1) = v_0(t_1)$ and $w_1(\bar{t}_1) = w_0(\bar{t}_1)$. It is obvious that t_1 and $\bar{t}_1$ are less than T. We relabel $v_1(t)$ and $w_1(t)$ as

$$\begin{aligned} v_1(t) &= v_1(t) \text{ on } [0, t_1], \qquad w_1(t) = w_1(t) \text{ on } [0, \bar{t}_1], \\ v_1(t) &= v_0(t) \text{ on } [t_1, T], \qquad w_1(t) = w_0(t) \text{ on } [\bar{t}_1, T]. \end{aligned}$$

It is easy to see that $v_0 \le v_1$ and $w_1 \le w_0$ on $[0, T]$. Continuing this process, we can compute $v_n(t)$ and $w_n(t)$ as

$$v_n' = f(t, v_{n-1}) + f_u(t, v_{n-1})(v_n - v_{n-1}) + g(t, w_{n-1}), \qquad v_n(0) = u_0, \tag{3-3}$$

$$w_n' = f(t, w_{n-1}) + f_u(t, v_{n-1})(w_n - w_{n-1}) + g(t, v_{n-1}), \quad w_n(0) = u_0, \tag{3-4}$$

on $[0, t_n]$ and $[0, \bar{t}_n]$, respectively. Again relabeling $v_n(t)$ and $w_n(t)$ on $[0, T]$ as before, we can prove

$$v_0 \le v_1 \le \cdots \le v_n \le u \le w_n \le \cdots \le w_1 \le w_0$$

on $[0, T]$. Note that this is the redefined sequences $\{v_n(t)\}$ and $\{w_n(t)\}$ on $[0, T]$. We can show that the redefined sequences $\{v_n(t)\}$ and $\{w_n(t)\}$ are equicontinuous and uniformly bounded on J. We will show that the sequence $\{v_n(t)\}$ is uniformly bounded. Since $v_0 \le v_n \le w_0$, and v_0, w_0 are continuous functions on a closed bounded set, it follows that $0 \le |v_n(t) - v_0(t)| \le |w_0(t) - v_0(t)| \le K_1$ on $[0, T]$. From this, and using the triangle inequality, we can show that

$$|v_n| = |v_n - v_0 + v_0| \le |v_n - v_0| + |v_0| \le K_1 + K_2 = K,$$

on $[0, T]$ where K is independent of n and t. This proves that $\{v_n(t)\}$ is uniformly

bounded. Similarly, we can prove that $\{w_n(t)\}$ is uniformly bounded. The equicontinuity of these sequences follows from the integral representation of v_n and w_n. This is achieved using the fact that the functions $f(t,u)$, $g(t,u)$ are continuous on Ω, and the uniform boundedness of $v_n(t)$ and $w_n(t)$. Hence, by the Arzelà–Ascoli theorem, a subsequence converges uniformly and monotonically. Since the sequences are monotone, the entire sequence converges uniformly and monotonically to v and w respectively. Further, we can prove the rate of convergence for these sequences are superlinear. □

We note that the elements of the sequences $\{v_n(t)\}$ and $\{w_n(t)\}$ are also coupled lower and upper solutions of (2-1) on $[0,T]$. We demonstrate this here.

Since $v_{n-1} \le v_n$ on $[0,T]$ and $w_n \le w_{n-1}$ on $[0,T]$, we get

$$f(t, v_{n-1}) + f_u(t, v_{n-1})(v_n - v_{n-1}) \le f(t, v_n) \quad \text{on } [0,T]$$

and

$$g(t, w_{n-1}) \le g(t, w_n),$$

using the assumptions on f and g from the hypothesis. This proves

$$v_n' \le f(t, v_n) + g(t, w_n),\ v_n(0) = u_0 \quad \text{on } [0,T].$$

Using the nature of $f(t,u)$ and $g(t,u)$, we can prove that

$$f(t, w_{n-1}) + f_u(t, v_{n-1})(w_n - w_{n-1}) \ge f(t, w_n) \quad \text{and} \quad g(t, w_{n-1}) \ge g(t, v_n).$$

Now from the iterates w_n, we can show that

$$w_n' \ge f(t, w_n) + g(t, v_n),\ w_n(0) = u_0 \quad \text{on } [0,T].$$

This proves that v_n and w_n are coupled lower and upper solutions of type I on the interval $[0,T]$.

Remark. Note that we can accelerate the rate of convergence of the sequences in Theorem 3.2 by using the Gauss–Seidel method.

We will apply the Gauss–Seidel method to Theorem 2.10.

Theorem 3.3. *We assume that*

(i) $v_0, w_0 \in C^1(J,R)$, $v_0(t) \le w_0(t)$ *on* J, *with* $v_0(t)$ *and* $w_0(t)$ *coupled lower and upper solutions of type I for* (2-1), *such that* $v_0(t) \le w_0(t)$ *on* J;

(ii) $f, g \in C(\Omega, \mathbb{R})$, f_u, g_u, f_{uu} *exist, are continuous and satisfy* $f_{uu}(t,u) \ge 0$ *for* $(t,u) \in \Omega = \{t \in J \mid v_0(t) \le u \le w_0(t)\}$;

(iii) $g_u(t,u) \le 0$ *on* Ω.

Then there exist monotone sequences $\{v_n^*(t)\}$, $\{w_n^*(t)\}$ *that converge uniformly to the unique solution of* (2-1), *and the convergence is faster than superlinear.*

The iterative scheme is given by:

$$\left(v_n^*\right)' = f(t, v_{n-1}^*) + f_u(t, v_{n-1}^*)(v_n^* - v_{n-1}) + g(t, w_{n-1}^*), \qquad v_n^*(0) = u_0, \tag{3-5}$$

$$\left(w_n^*\right)' = f(t, w_{n-1}^*) + f_u(t, v_n^*)(w_n^* - w_{n-1}^*) + g(t, v_n^*), \qquad w_n^*(0) = u_0, \tag{3-6}$$

starting with $v_0^* = v_1$ *on* J.

Remark. Here $v_1(t)$ is computed using Theorem 2.10.

Proof. Initially, compute $v_1(t)$ using

$$v_1' = f(t, v_0) + f_u(t, v_0)(v_1 - v_0) + g(t, w_0), \qquad v_1(0) = u_0.$$

Relabel $v_1(t)$ as $v_0^*(t)$. Now compute $w_1(t)$ using $w_0(t)$ and $v_0^*(t)$. That is, $w_1(t)$ is the solution of

$$w_1' = f(t, w_0) + f_u(t, v_0^*)(w_1 - w_0) + g(t, v_0^*), \qquad w_1(0) = u_0.$$

Relabel $w_1(t)$ as $w_0^*(t)$ and continue the process.

It is obvious that $v_0(t) \le v_1(t) = v_0^*(t)$ and $w_1(t) \le w_0(t)$ on J. Therefore $g(v_0) \ge g(v_0^*)$ and $f_u(t, v_0) \le f_u(t, v_0^*)$ from the hypothesis. Let $p(t) = w_1(t) - w_0^*(t)$. Then $p(0) = 0$. Also,

$$\begin{aligned}
p'(t) &= (w_1)'(t) - (w_0^*)'(t) \\
&= f_u(t, v_0)(w_1 - w_0) + g(t, v_0) - f_u(t, v_0^*)(w_0^* - w_0) - g(t, v_0^*) \\
&\ge f_u(t, v_0)(w_1 - w_0) - f_u(t, v_0^*)(w_0^* - w_0) \\
&\ge f_u(t, v_0^*)(w_1 - w_0) - f_u(t, v_0^*)(w_0^* - w_0) \\
&= f_u(t, v_0^*)p(t).
\end{aligned}$$

It follows that $p'(t) \ge f_u(t, v_0^*)p(t)$. Using Corollary 2.4, we know $p(t) \ge 0$, that is, $w_1 \ge w_0^*$ on J. Continuing the process, we will be able to show that the sequences $\{v_n^*\}$ and $\{w_n^*\}$ converges faster than the sequences $\{v_n\}$ and $\{w_n\}$ computed using Theorem 2.10. □

4. Numerical results

Here we develop numerical results as an application of the theoretical main results in Section 3. All the numerical simulations are done using Euler's method, implemented in Matlab.

To begin with, consider the simple logistic equations

$$u' = u - u^2, \qquad u_0(0) = \tfrac{1}{2}, \tag{4-1}$$

$$u' = 2u - 3u^2, \quad u_0(0) = \tfrac{1}{2}. \tag{4-2}$$

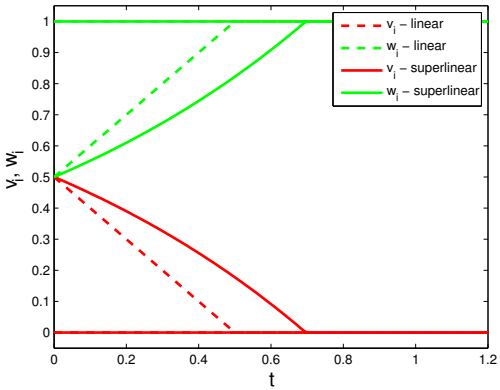


Figure 1. Comparison between linear and superlinear convergence.

It is easy to observe that $v_0(t) = 0$ and $w_0(t) = 1$ are the equilibrium solutions. In addition, they are also natural lower and upper solutions. Using the existence of solution by upper and lower solution method, the solution of (4-1) exists for all time. In Figure 1, we have computed coupled lower and upper solution using our superlinear convergence method as well as the linear convergence method as in [Noel et al. 2012]. In Figure 1, $v_0 \le v_1$ and $w_1 \le w_0$ on $[0, 0.5]$ by the generalized monotone method, whereas $v_0 \le v_1$ and $w_1 \le w_0$ on $[0, 0.7]$ by the superlinear convergence method.

Using Theorem 3.2, we computed v_i and w_i for (4-1), for $i = 1, 2, 3$. In Figure 2, we can see that in three iterations, that is, v_3 and w_3 are coupled lower and upper solutions of (4-1) on $[0, 1]$.

Using v_3 and w_3 from Figure 2 as v_0 and w_0 in Theorem 2.10, we compute the unique solution of (4-1), in Figure 3. This is achieved in four iterations. In Figure 4, we use superlinear convergence and the Gauss–Seidel method to compute coupled lower and upper solutions of (4-1) on $[0, 1]$. We achieved this in two iterations.

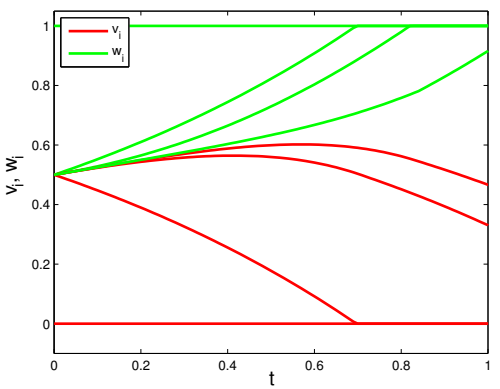


Figure 2. Coupled lower and upper solutions of (4-1) using Theorem 3.2.

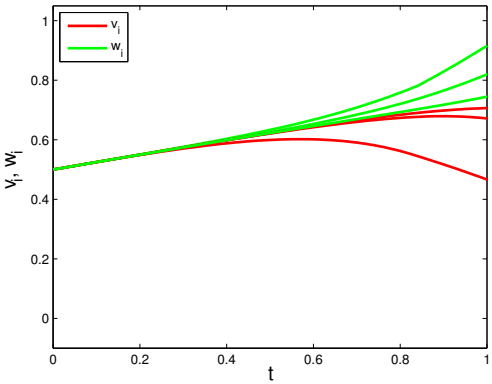


Figure 3. Four iterations of (4-1) using Theorem 2.10.

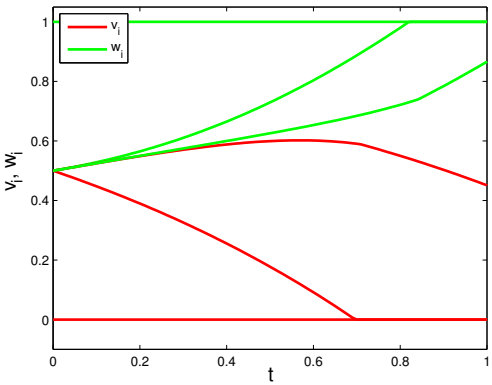


Figure 4. Coupled lower and upper solutions of (4-1) using Theorem 3.3.

Using the coupled lower and upper solution of Figure 4, we have computed the unique solution of (4-1) on [0, 1] using Theorem 3.3; see Figure 5. This combines

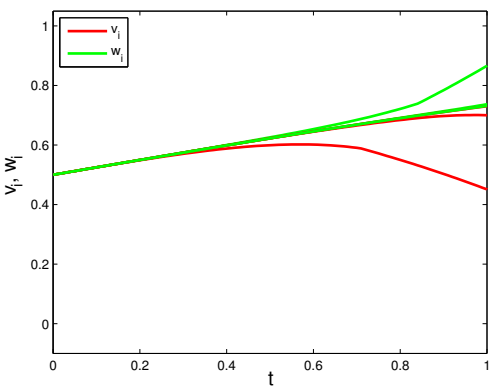


Figure 5. Three iterations of (4-1) using Theorem 3.3.

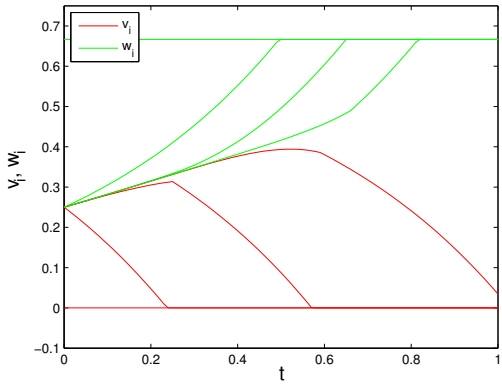


Figure 6. Three iterations of (4-2) using Theorem 3.2.

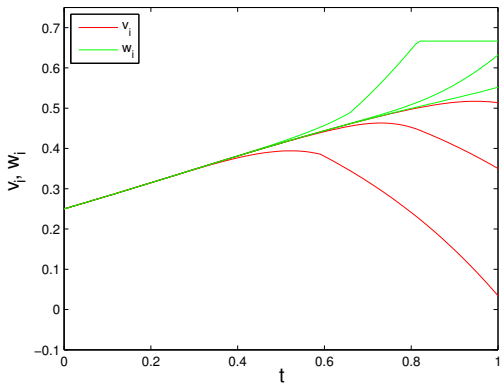


Figure 7. Three iterations of (4-2) using Theorem 2.10.

superlinear convergence and the Gauss–Seidel method. We achieved this in three iterations.

We have used the superlinear convergence method to compute coupled lower and upper solutions of (4-2) on [0, 1]. We achieved this in Figure 6 in three iterations.

Using the coupled lower and upper solutions of Figure 6, we have computed the unique solution of (4-2) on [0, 1] using Theorem 2.10. In Figure 7, this is achieved in three iterations.

5. Conclusion

We have developed a method to compute coupled upper and lower solutions for a nonlinear differential equation with initial conditions to any desired interval or to the interval of existence. We note that the natural lower and upper solutions guarantee the interval of existence of the solution. However, to compute the solutions by the generalized monotone method or the generalized quasilinearization method, we need coupled lower and upper solutions of type I on that interval. The method we have

developed requires the construction of sequences or iterates that are solutions of the linear equation. The rate of convergence of these sequences is superlinear. Further, the rate of convergence can be accelerated using the Gauss–Seidel acceleration method. Linear convergence methods are developed in [Noel et al. 2012]. Although we have applied our theoretical method to the logistic equation in our numerical results, our method is applicable to a variety of nonlinear problems, including Ricatti type differential equations. We plan to extend our method to two or more systems of differential equations. We anticipate being able to apply it to two species biological models (the Lotka–Volterra equation, for example), which can be cooperative, competitive or predator–prey models.

References

[Adams et al. 2012] B. M. Adams, N. Davis, P. Epps, F. Miller, D. Mullen, and A. S. Vatsala, "Existence of solution for impulsive hybrid differential equation", *Math. Sci.* (*Springer*) **6** (2012), Art. 19, 9. MR 3030367

[Bhaskar and McRae 2002] T. G. Bhaskar and F. A. McRae, "Monotone iterative techniques for nonlinear problems involving the difference of two monotone functions", *Appl. Math. Comput.* **133**:1 (2002), 187–192. MR 2003f:34005 Zbl 1035.34002

[Cronin 1994] J. Cronin, *Differential equations: introduction and qualitative theory*, 2nd ed., Monographs and Textbooks in Pure and Applied Mathematics **180**, Marcel Dekker, New York, 1994. MR 95b:34001 Zbl 0798.34001

[Holt and Pickering 1985] R. D. Holt and J. Pickering, "Infectious disease and species coexistence: a model of Lotka–Voltera Form", *The American Naturalist* **126**:2 (1985), 196–211.

[Jin et al. 2004] Z. Jin, M. Zhien, and H. Maoan, "The existence of periodic solutions of the n-species Lotka–Volterra competition systems with impulsive", *Chaos Solitons Fractals* **22**:1 (2004), 181–188. MR 2005d:34088 Zbl 1058.92046

[Ladde et al. 1985] G. S. Ladde, V. Lakshmikantham, and A. S. Vatsala, *Monotone iterative techniques for non-linear differential equations*, Pitman, 1985.

[Lakshmikantham and Vatsala 1998] V. Lakshmikantham and A. S. Vatsala, *Generalized quasi-linearization for nonlinear problems*, Mathematics and its Applications **440**, Kluwer Academic Publishers, Dordrecht, 1998. MR 99k:34013 Zbl 0997.34501

[Lakshmikantham et al. 1989] V. Lakshmikantham, D. D. Baĭnov, and P. S. Simeonov, *Theory of impulsive differential equations*, Series in Modern Applied Mathematics **6**, World Scientific Publishing Co., Teaneck, NJ, 1989. MR 91m:34013 Zbl 0719.34002

[Muniswamy and Vatsala 2013] S. Muniswamy and A. S. Vatsala, "Superlinear convergence for Caputo fractional differential equations with applications", *Dynam. Systems Appl.* **22**:2-3 (2013), 479–492. MR 3100218

[Noel et al. 2012] C. Noel, H. Sheila, N. Zenia, P. Dayonna, W. Jasmine, A. S. Vatsala, and M. Sowmya, "Numerical application of generalized monotone method for population models", *Neural Parallel Sci. Comput.* **20**:3-4 (2012), 359–372. MR 3057736 Zbl 1278.34033

[Sokol and Vatsala 2001] M. Sokol and A. S. Vatsala, "A unified exhaustive study of monotone iterative method for initial value problems", *Nonlinear Stud.* **8**:4 (2001), 429–438. MR 2002i:34009 Zbl 1094.34503

[Stutson and Vatsala 2011] D. Stutson and A. S. Vatsala, "Generalized monotone method for Caputo fractional differential systems via coupled lower and upper solutions", *Dynam. Systems Appl.* **20**:4 (2011), 495–503. MR 2012j:34012 Zbl 1236.34020

[West and Vatsala 2004] I. H. West and A. S. Vatsala, "Generalized monotone iterative method for initial value problems", *Appl. Math. Lett.* **17**:11 (2004), 1231–1237. MR 2099322 Zbl 1112.34304

Received: 2013-09-15 Revised: 2013-11-22 Accepted: 2013-11-24

vinchencia.anderson@gmail.com *Medgar Evers College, 1650 Bedford Avenue, New York, NY 11225, United States*

topaz5276@yahoo.com *Xavier University of Louisiana, 1 Drexel Drive, New Orleans, LA 70125, United States*

sbrown77@scmail.spelman.edu *Spelman College, 350 Spelman Lane SW, Atlanta, GA 30314, United States*

jxd5345@louisiana.edu *Department of Mathematics, University of Louisiana at Lafayette, 104 East University Avenue, Lafayette, LA 70504, United States*

ntullwal@live.com *Grambling State University, 403 Main Street, Grambling, LA 71245, United States*

vxc1794@louisiana.edu *Department of Mathematics, University of Louisiana at Lafayette, 104 East University Avenue, Lafayette, LA 70504-1010, United States*

vatsala@louisiana.edu *Department of Mathematics, University of Louisiana at Lafayette, 104 East University Avenue, Lafayette, LA 70504-1010, United States*

Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the Submission page at the Involve website.

Originality. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles in *Involve* are usually in English, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

Format. Authors are encouraged to use LaTeX but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@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

2014 vol. 7 no. 5

Infinite cardinalities in the Hausdorff metric geometry 585
ALEXANDER ZUPAN

Computing positive semidefinite minimum rank for small graphs 595
STEVEN OSBORNE AND NATHAN WARNBERG

The complement of Fermat curves in the plane 611
SETH DUTTER, MELISSA HAIRE AND ARIEL SETNIKER

Quadratic forms representing all primes 619
JUSTIN DEBENEDETTO

Counting matrices over a finite field with all eigenvalues in the field 627
LISA KAYLOR AND DAVID OFFNER

A not-so-simple Lie bracket expansion 647
JULIE BEIER AND MCCABE OLSEN

On the omega values of generators of embedding dimension-three numerical monoids generated by an interval 657
SCOTT T. CHAPMAN, WALTER PUCKETT AND KATY SHOUR

Matrix coefficients of depth-zero supercuspidal representations of GL(2) 669
ANDREW KNIGHTLY AND CARL RAGSDALE

The sock matching problem 691
SARAH GILLIAND, CHARLES JOHNSON, SAM RUSH AND DEBORAH WOOD

Superlinear convergence via mixed generalized quasilinearization method and generalized monotone method 699
VINCHENCIA ANDERSON, COURTNEY BETTIS, SHALA BROWN, JACQKIS DAVIS, NAEEM TULL-WALKER, VINODH CHELLAMUTHU AND AGHALAYA S. VATSALA