involve

a journal of mathematics

msp

2017 vol. 10, no. 4

involve

msp.org/involve

INVOLVE YOUR STUDENTS IN RESEARCH

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

PRODUCTION

Silvio Levy, Scientific Editor

Cover: Alex Scorpan

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2017 is US $175/year for the electronic version, and $235/year (+$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP.

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

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

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

INVOLVE 10:4 (2017)
dx.doi.org/10.2140/involve.2017.10.541

New algorithms for modular inversion and representation by the form $x^2+3xy+y^2$

Christina Doran, Shen Lu and Barry R. Smith

(Communicated by Filip Saidak)

We observe structure in the sequences of quotients and remainders of the Euclidean algorithm with two families of inputs. Analyzing the remainders, we obtain new algorithms for computing modular inverses and representing prime numbers by the binary quadratic form $x^2+3xy+y^2$. The Euclidean algorithm is commenced with inputs from one of the families, and the first remainder less than a predetermined size produces the modular inverse or representation.

1. The algorithms

Intuitively, the iterative nature of the Euclidean algorithm makes the sequences of quotients and remainders "sensitive to initial conditions". A small perturbation to the inputs can induce a chain reaction of increasingly large perturbations in the sequence of quotients and remainders, leading to considerable alterations to both the lengths of the sequences and their entries. Later entries are especially prone to change because of cumulative effects.

Our first result, Theorem 8, provides a surprising example of regularity under perturbation. When v is a solution of the congruence $v^2+v-1\equiv 0 \pmod{u}$, we show that the Euclidean algorithm with u and $v-1$ always takes one step fewer than the Euclidean algorithm with u and v. The sequences of quotients in both cases are almost identical, differing only in their middle one or two entries. (They are also symmetric outside of those middle entries.) We also obtain explicit formulas for the remainders of the Euclidean algorithm with u and $v-1$ in terms of the remainders produced by u and v.

From these formulas we obtain a new algorithm for representing prime numbers by the indefinite quadratic form $x^2+3xy+y^2$. When such a representation exists,

MSC2010: 11A05.

Keywords: number theory, continued fraction, binary quadratic form, algorithm.

This work was supported by a grant from The Edward H. Arnold and Jeanne Donlevy Arnold Program for Experiential Education, which supports research experiences for undergraduates at Lebanon Valley College.

the algorithm produces one with $x > y > 0$. Lemma 3 at the end of this section shows this representation is unique.

Algorithm 1. *Let p be a prime number congruent to* 1 *or* 4 *modulo* 5. *To compute the unique representation $p = b^2 + 3bc + c^2$ with $b > c > 0$, first compute a solution v to the congruence $v^2 + v - 1 \equiv 0 \pmod{p}$, then perform the Euclidean algorithm with p and v. The first remainder less than $\sqrt{p/5}$ is c, and the remainder just preceding is either b or $b + c$.*

This algorithm is similar to earlier algorithms that use the Euclidean algorithm to produce representations by binary quadratic forms [Brillhart 1972; Cornacchia 1908; Hardy et al. 1990a; 1990b; Matthews 2002; Wilker 1980]. Of these, [Matthews 2002] is the only one to produce representations by forms with positive discriminant, namely, the forms $x^2 - wy^2$ with $w = 2, 3, 5, 6$, or 7. The algorithm we present is a new contribution to this body of work.

We study a second family of inputs to the Euclidean algorithm, pairs $u > v$ for which $(v \pm 1)^2 \equiv 0 \pmod{u}$. This condition implies that there must exist a, b, and c with $u = ab^2$ and $v = abc \pm 1$. Theorems 9 and 10 give an explicit description of the quotients and remainders of the Euclidean algorithm with u and v in terms of the quotients and remainders of the Euclidean algorithm with b and c.

The relationship between the quotients of the Euclidean algorithm with b and c and with ab^2 and $abc \pm 1$ is essentially the "folding lemma" for continued fractions, first explicated independently in [Mendès France 1973; Shallit 1979]. This lemma has inspired a significant body of work concerning the quotients of continued fractions. These works give attention only to continued fractions — the remainders in the Euclidean algorithm are never explicitly considered. The description of the entire Euclidean algorithm with ab^2 and $abc \pm 1$ in Theorems 9 and 10 is new. They are unified by Theorem 11, which arithmetically characterizes the quotient pattern that will appear in the Euclidean algorithm with u and v when $(v \pm 1)^2 \equiv 0 \pmod{u}$.

Analysis of the remainders leads to another new algorithm, this time for modular inversion.

Algorithm 2. *If m and n are relatively prime positive integers, then the multiplicative inverse of m modulo n is the first remainder less than n when the Euclidean algorithm is performed with n^2 and $mn + 1$.*

A similar algorithm was obtained by Seysen [2005]. In his algorithm, an integer f is arbitrarily chosen with $f > 2n$, and the Euclidean algorithm is run with fn and $fm + 1$. The algorithm is stopped at the first remainder r less than $f + n$, and the modular inverse of m modulo n is then $r - f$ (which can be negative). If f were allowed to equal n, then this would be similar indeed to the algorithm above. However, Seysen's algorithm does not work generally in this case. For instance, with $n = 12$ and $m = 5$, Seysen's algorithm with $f = 12$ would say to run the

Euclidean algorithm with 144 and 61, stopping at the first remainder less than 24. This remainder is 22, and Seysen's algorithm would output 10, which is not an inverse for 5 modulo 12. Our algorithm above instead produces the inverse 5.

The inputs to Algorithm 2 are less than half the size of the inputs to Seysen's. But Seysen's algorithm has the flexibility arising from choosing the factor f. It would be interesting to see if both algorithms can fit in a common framework.

Our results are a new contribution to the literature on algorithmic number theory, but we believe the modular inversion algorithm also has pedagogical value. Students are less prone to mistakes working by hand with the new algorithm rather than the extended Euclidean algorithm or Blankinship's matrix algorithm [1963]. The new algorithm might seem nonintuitive, but our proof is elementary and is an amalgam of topics encountered by a student learning formal reasoning: the Euclidean algorithm, congruences, and mathematical induction.

We conclude this section with the result guaranteeing the uniqueness of the representation produced by Algorithm 1.

Lemma 3. *If p is a prime number congruent to* 1 *or* 4 *modulo* 5, *then there is a unique pair of positive integers $b > c$ satisfying*

$$p = b^2 + 3bc + c^2.$$

Proof. We work in the field $\mathbb{Q}(\sqrt{5})$. The algebraic integers in this field are

$$\mathcal{O} = \left\{ \tfrac{1}{2}m + \tfrac{1}{2}n\sqrt{5} : m, n, \in \mathbb{Z}, m \equiv n \bmod 2 \right\}.$$

Denote by τ the nontrivial automorphism of $\mathbb{Q}(\sqrt{5})$ and by N the norm map $N\gamma = \gamma\gamma^\tau$. The unit $\varepsilon = \frac{3}{2} + \frac{1}{2}\sqrt{5}$ generates the group of units of norm 1 in $\mathbb{Z}[\sqrt{5}]$. The map

$$(b, c) \mapsto \left(b + \tfrac{3}{2}c\right) + \left(\tfrac{1}{2}c\right)\sqrt{5}$$

gives a bijection between all pairs of integers (b, c) with $b^2 + 3bc + c^2 = p$ and all elements of $\mathcal{O}$ of norm p. The condition $b > c > 0$ for a pair with $b^2 + 3bc + c^2 = p$ is equivalent to the corresponding element $\frac{1}{2}x + \frac{1}{2}y\sqrt{5}$ of $\mathcal{O}$ satisfying $x > 5y > 0$.

By quadratic reciprocity, p splits in $\mathbb{Q}(\sqrt{5})$. The ring $\mathcal{O}$ is a principal ideal domain, so we may pick a generator γ of one of the prime ideals dividing p. Multiplying γ by $\frac{1}{2} + \frac{1}{2}\sqrt{5}$ if necessary, we may assume γ has norm p.

There is therefore at least one algebraic integer with norm p of the form $\frac{1}{2}x + \frac{1}{2}y\sqrt{5}$. Among all such elements, let α be one for which x is positive and is as small as possible (i.e., α has minimal positive trace). Replacing α by α^τ if necessary, we may assume also that y is positive. The lemma will be proved by showing that α is the unique element $\frac{1}{2}x + \frac{1}{2}y\sqrt{5}$ in $\mathcal{O}$ with norm p and $x > 5y > 0$.

Define a_n, b_n as the integers for which

$$\alpha\varepsilon^n = \tfrac{1}{2}a_n + \tfrac{1}{2}b_n\sqrt{5}.$$

Then

$$(\alpha\varepsilon^{-1})^\tau = \tfrac{1}{4}(3a_0 - 5b_0) + \tfrac{1}{4}(a_0 - 3b_0)\sqrt{5}.$$

If we suppose $a_0 - 3b_0 < 0$, then $\frac{1}{4}(5b_0 - 3a_0) > -\frac{1}{3}a_0$. If $5b_0 - 3a_0$ were negative, then $(\alpha\varepsilon^{-1})^\tau$ would have norm p and smaller positive trace than α, a contradiction. Thus, again by our choice of α, we have $\frac{1}{2}(5b_0 - 3a_0) \geq a_0$; hence $a_0 \leq b_0$. But then

$$N\alpha = \tfrac{1}{4}(a_0^2 - 5b_0^2) \leq -b_0^2 < 0,$$

which contradicts the assumption that α has norm p.

It must be then that $a_0 - 3b_0 > 0$, and thus, $3a_0 - 5b_0 > 0$. Again using our assumption on α, we have $\frac{1}{2}(3a_0 - 5b_0) \geq a_0$. It follows that $a_0 \geq 5b_0 > 0$ (and, in fact, $a_0 > 5b_0$ since $p \neq 5$).

It remains to show that α is the unique algebraic integer $\frac{1}{2}x + \frac{1}{2}y\sqrt{5}$ with norm p satisfying $x > 5y > 0$. Suppose x and y are integers and set $\frac{1}{2}w + \frac{1}{2}z\sqrt{5} = \big(\frac{1}{2}x + \frac{1}{2}y\sqrt{5}\big)\varepsilon$. It is readily checked that if $x > 0$ and $y > 0$, then $w > 0$ and $z > 0$ and $w < 5z$. It follows that all for all $n \geq 0$, we have $a_n > 0$ and $b_n > 0$, but $a_n > 5b_n$ only when $n = 0$. Recall that $\alpha\varepsilon^{-1} = \frac{1}{2}a_{-1} + \frac{1}{2}b_{-1}\sqrt{5}$. From the above two paragraphs, we have $a_{-1} > 0$ and $b_{-1} < 0$. If we set $\frac{1}{2}w' + \frac{1}{2}z'\sqrt{5} = \big(\frac{1}{2}x + \frac{1}{2}y\sqrt{5}\big)\varepsilon^{-1}$ and if $x > 0$ and $y < 0$, then $w' > 0$ and $y' < 0$. Thus, $a_n > 0$ and $b_n < 0$ for all $n \leq -1$.

The numbers in $\mathcal{O}$ of norm p are exactly $\pm\frac{1}{2}a_n \pm \frac{1}{2}b_n\sqrt{5}$ for n in $\mathbb{Z}$. It follows that the only possible element $\frac{1}{2}x + \frac{1}{2}y\sqrt{5}$ with norm p and $x > 5y > 0$ other than α is $\frac{1}{2}a_{-1} - \frac{1}{2}b_{-1}\sqrt{5} = \frac{1}{4}(3a_0 - 5b_0) + \frac{1}{4}(a_0 - 3b_0)\sqrt{5}$. But $3a_0 - 5b_0 > 5(a_0 - 3b_0)$ implies that $a_0 < 5b_0$, which we know is not true. The uniqueness is proved. □

2. Euclidean algorithm background

For positive integers $u > v$, the sequence of equations of the Euclidean algorithm when commenced by dividing v into u has the form

$$\begin{aligned} u &= q_1 v + r_1, \\ v &= q_2 r_1 + r_2, \\ r_1 &= q_3 r_2 + r_3, \\ &\vdots \\ r_{s-3} &= q_{s-1} r_{s-2} + r_{s-1}, \\ r_{s-2} &= q_s r_{s-1} + r_s, \end{aligned} \tag{1}$$

with $r_{s-1} = \gcd(u, v)$ and $r_s = 0$. We define

$$r_{-1} = u \quad \text{and} \quad r_0 = v.$$

Because $r_{s-1} < r_{s-2}$, it follows that $q_s \geq 2$.

Our study of the Euclidean algorithm is streamlined by allowing it to unfold in two different ways. These parallel the two continued fraction expansions of a rational number. The expansion of u/v with final quotient ≥ 2 is the sequence of quotients of the Euclidean algorithm with u and v. We will modify the Euclidean algorithm to make it produce the other expansion. If the Euclidean algorithm with u and v is written as (1), we replace the final equation by the two equations

$$r_{s-2} = (q_{s-1} - 1)r_{s-1} + r_{s-1}, \quad r_{s-1} = 1 \cdot r_{s-1} + 0. \tag{2}$$

This modification changes the length parities of the sequences of quotients and remainders.

Definition. If u and v are positive integers and $\delta = 0$ or 1, we denote by $\mathrm{EA}(u, v, \delta)$ the sequence of equations of the Euclidean algorithm when commenced with u and v. When $\delta = 0$, we use whichever of the standard or modified Euclidean algorithms takes an even number of steps, and when $\delta = 1$, whichever takes an odd number. When considering only the standard algorithm, we write simply $\mathrm{EA}(u, v)$. We denote the i-th equation by $\mathrm{EA}^i(u, v, \delta)$ or $\mathrm{EA}^i(u, v)$ and call the associated sequences (q_i) and (r_i) the *sequence of quotients* and *sequence of remainders*.

Reasoning about the Euclidean algorithm is facilitated by continuants. Properties of continuants can be found in Section 6.7 of the book by Graham, Knuth, and Patashnik [Graham et al. 1989].

Definition. Associated with a sequence $(q_1, \dots, q_s)$ of integers, we define a doubly indexed sequence of *continuants*

$$\mathfrak{q}_{i,j} = q_i \mathfrak{q}_{i+1,j} + \mathfrak{q}_{i+2,j} \qquad \text{and} \qquad \mathfrak{q}_{i+1,i} = 1, \quad \mathfrak{q}_{i+2,i} = 0 \tag{3}$$

for $1 \leq i \leq j + 2 \leq s + 2$. When a more explicit description of the $\mathfrak{q}_i$ is required, we will use alternate notation (for $i \leq j$):

$$[q_i, \dots, q_j] := \mathfrak{q}_{i,j}.$$

The properties of continuants that we will need are the recursion (3) and the surprising symmetry

$$[q_i, \dots, q_j] = [q_j, \dots, q_i],$$

which can be proved by induction. An illuminating combinatorial proof is in [Benjamin et al. 2000]. From the symmetry of continuants and recurrence (3) we obtain the alternate recurrence

$$\mathfrak{q}_{i,j} = q_j \mathfrak{q}_{i,j-1} + \mathfrak{q}_{i,j-2}. \tag{4}$$

Lemma 4. *Let u and v be relatively prime integers. If $(q_i)_{i=1}^{s}$ and $(r_i)_{i=-1}^{s}$ are the sequences of quotients and remainders of* $\mathrm{EA}(u, v, \delta)$ *and $\mathfrak{q}_{i,j}$ are the continuants corresponding to the sequence of quotients, then*

$$r_i = \mathfrak{q}_{i+2,s}$$

for $i = -1, \dots, s$. In particular, $u = \mathfrak{q}_{1,s}$ and $v = \mathfrak{q}_{2,s}$.

Proof. Because u and v are relatively prime, we have $r_{s-1} = 1 = \mathfrak{q}_{s+1,s}$ and $r_s = 0 = \mathfrak{q}_{s+2,s}$. The formula $r_i = \mathfrak{q}_{i+2,s}$ follows from the observation that the recurrence (3) with $j = s$ is the same recurrence satisfied by the remainders. □

The continuants $\mathfrak{q}_{1,i}$ have a prominent role in studying the Euclidean algorithm. They are the numerators of the convergents of the simple continued fraction expansion of u/v, and they are the absolute values of coefficients commonly computed as part of the extended Euclidean algorithm. We therefore make the following definition.

Definition. Let $q_1, q_2, \dots, q_s$ be the sequence of quotients of $\mathrm{EA}(u, v, \delta)$ with associated continuants $\mathfrak{q}_{i,j}$. We define the *Bézout coefficients* of u and v by

$$\beta_i = \mathfrak{q}_{1,i}$$

for $-1 \le i \le s$.

The following lemmas reveal a close connection between the sequence of remainders of $\mathrm{EA}(u, v, \delta)$ and the corresponding Bézout coefficients. Each makes a fine exercise in mathematical induction.

Lemma 5. *If $(q_i)_{i=1}^{s}$ and $(r_i)_{i=-1}^{s}$ are the sequences of quotients and remainders of $\mathrm{EA}(u, v, \delta)$ and $(\beta_i)_{i=-1}^{s}$ are the Bézout coefficients, then*

$$v\beta_i \equiv (-1)^i r_i \pmod{u} \quad \text{for } -1 \le i \le s.$$

Proof. The cases $i = -1$ and $i = 0$ simply say that $0 \equiv -u \pmod{u}$ and $v \equiv v \pmod{u}$. Further, if the congruence holds for $i - 1$ and i with $0 \le i \le s - 1$, then

$$\begin{aligned} v\beta_{i+1} &= vq_{i+1}\beta_i + v\beta_{i-1} \\ &\equiv (-1)^i q_{i+1} r_i + (-1)^{i-1} r_{i-1} \pmod{u} \\ &= (-1)^{i+1} r_{i+1}. \end{aligned}$$

The lemma follows by induction. □

Lemma 6. *If $(q_i)_{i=1}^{s}$ and $(r_i)_{i=-1}^{s}$ are the sequences of quotients and remainders of $\mathrm{EA}(u, v, \delta)$ and $(\beta_i)_{i=-1}^{s}$ are the Bézout coefficients, then $u = \beta_i r_{i-1} + \beta_{i-1} r_i$ for $0 \le i \le s$.*

Proof. For $i = 0$, the equation is just $u = u$. Assume that $u = \beta_i r_{i-1} + \beta_{i-1} r_i$ for some i with $0 \le i \le s - 1$. Then using (4),

$$u = \beta_i (q_{i+1} r_i + r_{i+1}) + (\beta_{i+1} - q_{i+1}\beta_i) r_i = \beta_{i+1} r_i + \beta_i r_{i+1}.$$

The lemma follows by induction. □

We now discuss background for studying structure in the Euclidean algorithm quotients. Fix a positive integer k. In recent work [Smith 2015], it was proved that if v with $0 < v < u$ satisfies the congruence

$$v^2 + kv \pm 1 \equiv 0 \pmod{u},$$

then the sequence of quotients of $\mathrm{EA}(u, v, \delta)$ (with $\delta = 0$ if the plus sign is used in the above congruence and $\delta = 1$ otherwise) fits one of a finite list of "end-symmetric" patterns. The list of patterns depends only on k. We will use this result only when $k = 1$, 2, or 3.

Lemma 7. *The sequence of quotients of* $\mathrm{EA}(u, v, 1)$ *when* $v^2 + v - 1 \equiv 0 \pmod{u}$ *has the form*

$$q_1,\ \dots,\ q_{s-1},\ q_s + (-1)^{s+1},\ 1,\ q_s,\ q_{s-1},\ \dots,\ q_1$$

for some positive integers $q_1, \dots, q_s$.

When $v^2 + 3v + 1 \equiv 0 \pmod{u}$, *then* $\mathrm{EA}(u, v, 0)$ *has quotient sequence of the form*

$$q_1,\ \dots,\ q_{s-1},\ q_s + (-1)^{s+1} \cdot 3,\ q_s,\ q_{s-1},\ \dots,\ q_1$$

for some positive integers $q_1, \dots, q_s$.

When $v^2 + (-1)^\delta 2v + 1 \equiv 0 \pmod{u}$, *that is, when*

$$(v + (-1)^\delta)^2 \equiv 0 \pmod{u}, \tag{5}$$

then $\mathrm{EA}(u, v, 0)$ *has quotient sequence fitting one of the patterns*

$$\begin{gathered} q_1,\ \dots,\ q_{s-1},\ q_s - (-1)^{s+\delta},\ q_s + (-1)^{s+\delta},\ q_{s-1},\ \dots,\ q_1, \\ q_1,\ \dots,\ q_{s-1},\ q_s + 1,\ x,\ 1,\ q_s,\ q_{s-1},\ \dots,\ q_1, \\ q_1,\ \dots,\ q_{s-1},\ q_s - 1,\ 1,\ x,\ q_s,\ q_{s-1},\ \dots,\ q_1 \end{gathered} \tag{6}$$

for some positive integers $q_1, \dots, q_s$ *and* x.

The patterns (6) are well known, being related to paper-folding sequences and folded continued fractions [Shallit 1979; van der Poorten 2002]. What seems to be new is their appearance in the quotients of the Euclidean algorithm with u and v when v satisfies (5). Theorem 11 gives an arithmetical criteria for deciding which of the patterns (6) describes the simple continued fraction expansion of u/v.

3. Explicating the Euclidean algorithm

Suppose u and v are positive integers with $u > v$ and $v^2 + v - 1 \equiv 0 \pmod{u}$. Then $v - 1$ satisfies the congruence $v^2 + 3v + 1 \equiv 0 \pmod{u}$. According to Lemma 7, $\mathrm{EA}(u, v, 1)$ has sequence of quotients of the form $q_1, \dots, q_s + \delta_1, 1, q_s + \delta_0, \dots, q_1$, while $\mathrm{EA}(u, v - 1, 0)$ has sequence of quotients of the form $\tilde{q}_1, \dots, \tilde{q}_s + \delta_1 \cdot 3$, $\tilde{q}_s + \delta_0 \cdot 3, \dots, \tilde{q}_1$. In both cases, $\delta_1 = 1$ if s is odd and 0 if s is even, while $\delta_0 = 1$ if s is even and 0 if s is odd. There is no a priori reason for the sequence of q_i to equal the sequence of $\tilde{q}_i$. Nevertheless, that is the conclusion of the following theorem, which also gives explicit formulas for the remainders of $\mathrm{EA}(u, v - 1, 0)$ in terms of the remainders of $\mathrm{EA}(u, v, 1)$.

Theorem 8. *Let u and v be positive integers $u > v$, with $v^2 + v - 1 \equiv 0 \pmod{u}$. Write the sequence of quotients of* $\mathrm{EA}(u, v, 1)$ *as*

$$q_1,\ \ldots,\ q_s + \delta_1,\ 1,\ q_s + \delta_0,\ \ldots,\ q_1.$$

Let $(r_i)_{i=-1}^{2s+1}$ be the sequence of remainders, and for $i = -1, \ldots, s-1$, set

$$t_i = r_i + (-1)^{i+1} r_{2s-i}.$$

Then $\mathrm{EA}(u, v-1, 0)$ *is the sequence of $2s$ equations*

$$\begin{aligned} t_{i-2} &= \quad q_i \qquad\quad \cdot t_{i-1} + t_i && \text{for } 1 \le i \le s-1,\\ t_{s-2} &= (q_s + \delta_1 \cdot 3) \cdot t_{s-1} + r_{s+1}, &&\\ t_{s-1} &= (q_s + \delta_0 \cdot 3) \cdot r_{s+1} + r_{s+2}, &&\\ r_{i-1} &= \quad q_{2s+1-i} \quad\ \cdot r_i \quad + r_{i+1} && \text{for } s+2 \le i \le 2s. \end{aligned}$$

Proof. A quick check verifies that $t_{-1} = u$ and $t_0 = v-1$, which begin the remainder sequence of $\mathrm{EA}(u, v-1, 0)$. Because the sequence $(r_i)_{i=1}^{2s+1}$ is decreasing, it is clear that the purported quotients and remainders are all positive. We check that the purported remainders form a strictly decreasing sequence (except that the final two may be equal when $\mathrm{EA}(u, v-1, 0)$ is computed using the modification (2) of the Euclidean algorithm.) This is apparent for $r_{s+1}, \ldots, r_{2s+1}$. Also, $t_{s-1} \ge r_{s-1} - r_{s+1} = r_s \ge r_{s+1}$. (The equality is because the middle quotient of $\mathrm{EA}(u, v, 1)$ is 1. The final equality is strict unless $u = 5$ and $v = 2$.)

We must show $t_i > t_{i+1}$ for $1 \le i \le s-2$. From the division algorithm, we have $r_i \ge r_{i+1} + r_{i+2}$ for $-1 \le i \le 2s-1$. Thus, for $-1 \le i \le s-3$, we have

$$r_i - r_{i+1} \ge r_{i+2} \ge r_{i+3} + r_{i+4} > r_{2s-i} + r_{2s-i-1}.$$

It follows that $t_i > t_{i+1}$ for $1 \le i \le s-3$. The above chain of inequalities also holds with the final inequality replaced by an equality when $i = s-2$. The second inequality is strict when $i = s-2$ unless $q_s + \delta_0 = 1$, which only happens if s is odd. But in that case, $t_{s-2} = r_{s-2} + r_{s+2} > r_{s-1} - r_{s+1} = t_{s-1}$ holds anyway.

To ensure the equations in the theorem are the steps of $\mathrm{EA}(u, v-1, 0)$, it remains to check the algebraic validity of each step. The theorem will then follow from the uniqueness of the quotients and remainders.

The equation $t_{i-2} = q_i \cdot t_{i-1} + t_i$ is equivalent to

$$(-1)^{i+1}(r_{i-2} - q_i r_{i-1} - r_i) = r_{2s-i} - q_i r_{2s+1-i} - r_{2s+2-i}.$$

The expression on the left is 0. Also, examining the pattern of the sequence of quotients of $\mathrm{EA}(u, v, 1)$, we see that $q_{2s+2-i} = q_i$ for $i = 1, \ldots, s-1$. Thus, the

$(2s-i+1)$-th step of EA$(u, v, 1)$ is

$$r_{2s-i} = q_i r_{2s+1-i} + r_{2s+2-i}, \tag{7}$$

and the right side is also 0. Substituting $2s+1-i$ for i in (7), we find as well that $r_{i-1} = q_{2s+1-i} r_i + r_{i+1}$ for $s+2 \le i \le 2s$, which verifies steps $i = s+2$ through $i = 2s$ in the theorem.

We now check the middle pair of equations. We know that the s-th through $(s+2)$-th equations of EA$(u, v, 1)$ are

$$\begin{aligned} r_{s-2} &= (q_s + \delta_1) r_{s-1} + r_s, \\ r_{s-1} &= \qquad\qquad r_s \quad + r_{s+1}, \\ r_s &= (q_s + \delta_0) r_{s+1} + r_{s+2}. \end{aligned} \tag{8}$$

Assume first that s is odd so that $\delta_1 = 1$ and $\delta_0 = 0$. The equation

$$t_{s-2} = (q_s + \delta_1 \cdot 3) t_{s-1} + r_{s+1}$$

is equivalent to

$$r_{s-2} = (q_s + 3)(r_{s-1} - r_{s+1}) + r_{s+1} - r_{s+2}.$$

Substituting in turn $r_{s+2} = r_s - q_s r_{s+1}$ and $r_{s+1} = r_{s-1} - r_s$ from (8), this is equivalent to

$$\begin{aligned} r_{s-2} &= (q_s + 3)(r_{s-1} - r_{s+1}) + r_{s+1} - r_s + q_s r_{s+1} \\ &= (q_s + 3) r_s + r_{s-1} - 2r_s + q_s r_{s-1} - q_s r_s \\ &= (q_s + 1) r_{s-1} + r_s, \end{aligned}$$

which is the first of equations (8).

If, instead, s is even, so $\delta_1 = 0$ and $\delta_0 = 1$, then $t_{s-2} = (q_s + \delta_1 \cdot 3) t_{s-1} + r_{s+1}$ is equivalent to

$$r_{s-2} = q_s (r_{s-1} + r_{s+1}) + r_{s+1} + r_{s+2}.$$

Substituting in turn $r_{s+2} = r_s - q_s r_{s+1} - r_{s+1}$ and $r_{s+1} = r_{s-1} - r_s$, this is equivalent to

$$\begin{aligned} r_{s-2} &= q_s (r_{s-1} + r_{s+1}) + r_s - q_s r_{s+1} \\ &= q_s (2r_{s-1} - r_s) + r_s - q_s r_{s-1} + q_s r_s \\ &= q_s r_{s-1} + r_s, \end{aligned}$$

which is the first of equations (8).

The verification that $t_{s-1} = (q_s + \delta_0 \cdot 3) \cdot r_{s+1} + r_{s+2}$ is entirely similar, using the latter two equations of (8). □

Proof of Algorithm 1. Let the quotients and remainders of EA$(u, v, 1)$ be written as in Theorem 8. Suppose first that s is odd. Applying Lemma 6 with $i = s$ to

EA(u, v, 1), we have $u = [q_1, \dots, q_{s-1}, q_s + 1]r_{s-1} + [q_1, \dots, q_{s-1}]r_s$. By the symmetry of continuants and recurrence (4), it follows that

$$\begin{aligned} u &= [q_s + 1, q_{s-1}, \dots, q_1]r_{s-1} + [q_{s-1}, \dots, q_1]r_s \\ &= [q_{s-1}, \dots, q_1](r_{s-1} + r_s) + [q_s, \dots, q_1]r_{s-1}. \end{aligned}$$

Now use the "end-symmetric" form of the quotient sequence of EA(u, v, 1) and Lemma 4 to obtain

$$u = r_{s+1}(r_{s-1} + r_s) + r_s r_{s-1}.$$

Substituting out r_{s-1} using the middle of equations (8) gives

$$u = r_s^2 + 3r_s r_{s+1} + r_{s+1}^2.$$

Suppose now that s is even. Applying Lemma 6 with $i = s$ to EA(u, v, 1) in this case gives $u = [q_1, \dots, q_s]r_{s-1} + [q_1, \dots, q_{s-1}]r_s$. Again using the recurrence (4), it follows that

$$u = [q_s + 1, q_{s-1}, \dots, q_1]r_{s-1} + [q_{s-1}, \dots, q_1](r_s - r_{s-1}),$$

and Lemma 4 shows

$$u = r_s r_{s-1} + r_{s+1}(r_s - r_{s-1}).$$

Substituting with (8) once more gives

$$u = (r_s - r_{s+1})^2 + 3(r_s - r_{s+1})r_{s+1} + r_{s+1}^2.$$

Thus, in either case, $r_{s+1} = c$ in the unique representation $p = b^2 + 3bc + c^2$ with $b > c > 0$. If s is odd, then $r_s = b$, and if s is even, then $r_s = b + c$. The inequalities $5b^2 > b^2 + 3bc + c^2 > 5c^2$ show that

$$b + c > b > \sqrt{\frac{p}{5}} > c.$$

Therefore regardless of whether s is odd or even, c is the first remainder smaller than $\sqrt{p/5}$. □

Fix anew positive integers b and c with $\gcd(b, c) = 1$. We next give an explicit description of the quotients and remainders of EA(b^2, $bc \pm 1$) in terms of the quotients, remainders, and Bézout coefficients of EA(b, c). The algorithm for computing inverses in modular arithmetic falls out of this description.

Theorem 9. *Let $b > c > 1$ be integers with* $\gcd(b, c) = 1$. *Let $(q_i)_{i=1}^s$ and $(r_i)_{i=-1}^s$ be the sequences of quotients and remainders of the standard (i.e., unmodified) Euclidean algorithm with b and c, let $(\beta_i)_{i=-1}^s$ be the corresponding Bézout coefficients, and set $t_i = r_i b \pm (-1)^i \beta_i$ for $-1 \le i \le s - 1$. Then* EA(b^2, $bc \pm 1$, 0) *is the* *sequence of* $2s$ *equations*

$$\begin{aligned}
t_{i-2} &= q_i && \cdot t_{i-1} && + t_i && \textit{for } 1 \le i \le s-1,\\
t_{s-2} &= (q_s \pm (-1)^s) && \cdot t_{s-1} && + \beta_{s-1}, &&\\
t_{s-1} &= (q_s \pm (-1)^{s-1}) && \cdot \beta_{s-1} && + \beta_{s-2}, &&\\
\beta_{2s+1-i} &= q_{2s+1-i} && \cdot \beta_{2s-i} && + \beta_{2s-1-i} && \textit{for } s+2 \le i \le 2s.
\end{aligned}$$

Proof. The proof can be conducted in an analogous manner to the proof of Theorem 8. One readily checks that the first two remainders are $t_{-1} = b^2$ and $t_0 = bc \pm 1$. The observation $q_s \ge 2$ was made in the first paragraph of Section 2, so the purported quotients are all positive. So are the remainders since $b \ge \beta_i$ for $-1 \le i \le s-1$.

For $s+2 \le i \le 2s$, the equation $\beta_{2s+1-i} = q_{2s+1-i} \cdot \beta_{2s-i} + \beta_{2s-1-i}$ follows from (4). For $1 \le i \le s-1$, the equality $t_{i-2} = q_i t_{i-1} + t_i$ can be deduced from the equation $\mathrm{EA}^i(b, c)$ and (4). To verify the middle two equations, we first note that because b and c are relatively prime, we have $r_{s-1} = 1$, $t_{s-1} = b \pm (-1)^{s-1}\beta_{s-1}$, and $q_s = r_{s-2}$. The equations can then be verified using Lemma 6 with $u = b$, $v = c$, and $i = s-1$:

$$\begin{aligned}
(q_s \pm (-1)^s)t_{s-1} + \beta_{s-1} &= (r_{s-2} \pm (-1)^s)b \pm (-1)^{s-1} r_{s-2}\beta_{s-1}\\
&= r_{s-2}\, b \pm (-1)^{s-2}\beta_{s-2}\\
&= t_{s-2}
\end{aligned}$$

and

$$\begin{aligned}
(q_s \pm (-1)^{s-1})\beta_{s-1} + \beta_{s-2} &= r_{s-2}\beta_{s-1} \pm (-1)^{s-1}\beta_{s-1} + \beta_{s-2}\\
&= b \pm (-1)^{s-1}\beta_{s-1}\\
&= t_{s-1}.
\end{aligned}$$

Finally, the remainders form a decreasing sequence. For $-1 < i < s-1$, the inequality $(r_i - r_{i+1})b > \beta_i + \beta_{i+1}$ follows from Lemma 6 and implies $t_i > t_{i+1}$. The inequality $\beta_{s-1} < t_{s-1}$ follows from the equation $t_{s-1} = (q_s \pm (-1)^{s-1})\beta_{s-1} + \beta_{s-2}$ verified in the last paragraph. And $\beta_{i-1} < \beta_i$ for $0 \le i \le s$ follows from the recurrence (4). □

Proof of Algorithm 2. When $m = 1$, the algorithm is easily validated. If $m > n$, then the third step of $\mathrm{EA}(n^2, mn+1)$ will be division of $rn + 1$ into n^2, where r is the remainder when m is divided by n. Thus, it suffices to assume $n > m > 1$, so also $s > 1$.

Theorem 9 implies the first remainder less than n in $\mathrm{EA}(n^2, mn+1)$ is β_{s-1} when s is odd and t_{s-1} when s is even. We apply Lemma 5 to $\mathrm{EA}(n, m)$ to find $m\beta_{s-1} \equiv (-1)^{s-1} \pmod n$. Thus when s is odd, the product of m and the first remainder less than n is

$$m\beta_{s-1} \equiv 1 \pmod{n}.$$

When s is even, the product is

$$mt_{s-1} = mn - m\beta_{s-1} \equiv 1 \pmod{n}. \qquad \square$$

We now give a complete description of $\mathrm{EA}(ab^2, abc \pm 1)$ for positive integers $a \geq 2$, b, and c and $\gcd(b, c) = 1$.

Theorem 10. *Let a, b, c, and k be integers with $b > c > 1$, $\gcd(b, c) = 1$, and $a \geq 2$. Let $(q_i)_{i=1}^{s}$ and $(r_i)_{i=-1}^{s}$ be the sequences of quotients and remainders in* $\mathrm{EA}(b, c)$, *let $(\beta_i)_{i=-1}^{s}$ be the corresponding Bézout coefficients, and set $t_i = abr_i + (-1)^{i+k}\beta_i$ for $-1 \leq i \leq s-1$. If $(-1)^{s+k} = -1$, then* $\mathrm{EA}(ab^2, abc + (-1)^k, 0)$ *is the sequence of $2s+2$ equations*

$$\begin{array}{rllll}
t_{i-2} = & q_i & \cdot\, t_{i-1} & + t_i & \textit{for } 1 \leq i \leq s-1, \\
t_{s-2} = & (q_s - 1) & \cdot\, t_{s-1} & + (t_{s-1} - b), & \\
t_{s-1} = & 1 & \cdot\, (t_{s-1} - b) & + b, & \\
t_{s-1} - b = & (a-1) & \cdot\, b & + \beta_{s-1}, & \\
b = & q_s & \cdot\, \beta_{s-1} & + \beta_{s-2}, & \\
\beta_{2s+3-i} = & q_{2s+3-i} & \cdot\, \beta_{2s+2-i} & + \beta_{2s+1-i} & \textit{for } s+4 \leq i \leq 2s+2.
\end{array}$$

When $(-1)^{s+k} = 1$, steps s through $s+3$ change to

$$\begin{array}{rlll}
t_{s-2} = & q_s & \cdot\, t_{s-1} & + b, \\
t_{s-1} = & (a-1) & \cdot\, b & + (b - \beta_{s-1}), \\
b = & 1 & \cdot\, (b - \beta_{s-1}) & + \beta_{s-1}, \\
b - \beta_{s-1} = & (q_s - 1) & \cdot\, \beta_{s-1} & + \beta_{s-2}.
\end{array}$$

Proof. It follows as in the proof of Theorem 9 that the purported quotients and remainders are positive (excluding the final remainder). The equations $\beta_{2s+3-i} = q_{2s+3-i} \cdot \beta_{2s+2-i} + \beta_{2s+1-i}$ and $t_{i-2} = q_i t_{i-1} + t_i$ can be deduced as in the proof of Theorem 9. The equations $t_{s-1} = 1 \cdot (t_{s-1} - b) + b$ and $b = 1 \cdot (b - \beta_{s-1}) + \beta_{s-1}$ are clearly true. Lemma 4 shows that $\beta_s = b$. Thus, the equations $b = q_s \cdot \beta_{s-1} + \beta_{s-2}$ and $b - \beta_{s-1} = (q_s - 1)\beta_{s-1} + \beta_{s-2}$ are consequences of (4).

Since $\gcd(b, c) = 1$, we have $r_{s-1} = 1$, $t_{s-1} = ab - (-1)^{s+k}\beta_{s-1}$, and $q_s = r_{s-2}$. From this, we obtain the equations $t_{s-1} - b = (a-1)b + \beta_{s-1}$ when $(-1)^{s+k} = -1$ and $t_{s-1} = (a-1)b + (b - \beta_{s-1})$ when $(-1)^{s+k} = 1$.

When $(-1)^{s+k} = -1$, the s-th equation is valid since

$$\begin{aligned}(q_s - 1)t_{s-1} + (t_{s-1} - b) &= q_s(ab + \beta_{s-1}) - \beta_s \\ &= abr_{s-2} + (\beta_s - \beta_{s-2}) - \beta_s, \\ &= t_{s-2}.\end{aligned}$$

Similarly, when $(-1)^{s+k} = 1$,

$$\begin{aligned}q_s t_{s-1} + b &= q_s(ab - \beta_{s-1}) + b \\ &= abr_{s-2} - (\beta_s - \beta_{s-2}) + \beta_s \\ &= t_{s-2}.\end{aligned}$$

When $(-1)^{s+k} = -1$, the inequality $t_{s-1} - b < t_{s-1}$ is clear and the inequality $b < t_{s-1} - b$ follows from the assumption that $a \geq 2$. When $(-1)^{s+k} = 1$, the inequality $b < t_{s-1}$ follows from the assumption that $a \geq 2$ and from $b = \beta_s > \beta_{s-1}$. The inequality $b - \beta_{s-1} < b$ is clear, and the inequality $\beta_{s-1} < b - \beta_{s-1}$ follows from $b = q_s\beta_{s-1} + \beta_{s-2}$ and $q_s \geq 2$. That $t_i < t_{i-1}$ and $\beta_i > \beta_{i-1}$ for $1 \leq i \leq s-1$ follows as in the proof of Theorem 9. □

To conclude, we provide an arithmetical characterization of which quotient pattern will appear when performing the Euclidean algorithm with u and v with $(v \pm 1)^2 \equiv 0 \pmod{u}$.

Theorem 11. *Let u be a positive integer and write $u = ab^2$, where a is the square-free part of u. Assume v with $0 < v < u$ satisfies $(v + (-1)^\delta)^2 \equiv 0 \pmod{u}$. Then there is an integer c such that*

$$v = abc + (-1)^{\delta+1}.$$

Let $q_1, \ldots, q_s$ be the quotient sequence of the simple continued fraction expansion of b/c.

The continued fraction expansion of u/v with even length has quotient sequence fitting the first of the patterns (6) *if and only if* $\gcd(b, c) = a = 1$. *Otherwise, it fits one of the other patterns with $x = \gcd(b, c)^2 \cdot a - 1$. The second pattern appears if $s + \delta$ is odd, and the third if $s + \delta$ is even.*

Proof. By assumption, there exists some integer w such that $(v + (-1)^\delta)^2 = uw$. Consideration of prime factorizations shows that a is also the square-free part of w, say $w = ac^2$. Then $v = abc + (-1)^{\delta+1}$.

If $\gcd(b, c) = d$ and we set $\tilde{a} = ad^2$, $\tilde{b} = b/d$, and $\tilde{c} = c/d$, then

$$u = \tilde{a}\tilde{b}^2, \quad v = \tilde{a}\tilde{b}\tilde{c} + (-1)^{\delta+1}, \quad \text{and} \quad \gcd(\tilde{b}, \tilde{c}) = 1.$$

Theorem 11 now follows from Theorem 9 and Theorem 10. □

References

[Benjamin et al. 2000] A. T. Benjamin, F. E. Su, and J. J. Quinn, “Counting on continued fractions”, *Math. Mag.* **73**:2 (2000), 98–104. MR

[Blankinship 1963] W. A. Blankinship, “Classroom notes: a new version of the Euclidean algorithm”, *Amer. Math. Monthly* **70**:7 (1963), 742–745. MR Zbl

[Brillhart 1972] J. Brillhart, “Note on representing a prime as a sum of two squares”, *Math. Comp.* **26** (1972), 1011–1013. MR Zbl

[Cornacchia 1908] G. Cornacchia, “Su di un metodo per la risoluzione in numeri interi dell’equazione $\sum_{h=0}^{n} C_h x^{n-h} y^h = P$”, *Giorn. Mat. Battaglini* **46** (1908), 33–90. JFM

[Graham et al. 1989] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete mathematics: a foundation for computer science*, Addison-Wesley, Reading, MA, 1989. MR Zbl

[Hardy et al. 1990a] K. Hardy, J. B. Muskat, and K. S. Williams, “A deterministic algorithm for solving $n = fu^2 + gv^2$ in coprime integers u and v”, *Math. Comp.* **55**:191 (1990), 327–343. MR Zbl

[Hardy et al. 1990b] K. Hardy, J. B. Muskat, and K. S. Williams, “Solving $n = au^2 + buv + cv^2$ using the Euclidean algorithm”, *Utilitas Math.* **38** (1990), 225–236. MR Zbl

[Matthews 2002] K. Matthews, “Thue’s theorem and the Diophantine equation $x^2 - Dy^2 = \pm N$”, *Math. Comp.* **71**:239 (2002), 1281–1286. MR Zbl

[Mendès France 1973] M. Mendès France, “Sur les fractions continues limitées”, *Acta Arith.* **23**:2 (1973), 207–215. MR Zbl

[van der Poorten 2002] A. J. van der Poorten, “Symmetry and folding of continued fractions”, *J. Théor. Nombres Bordeaux* **14**:2 (2002), 603–611. MR Zbl

[Seysen 2005] M. Seysen, “Using an RSA accelerator for modular inversion”, pp. 226–236 in *Cryptographic hardware and embedded systems – CHES 2005* (Edinburgh, 2005), edited by J. R. Rao and B. Sunar, Lecture Notes in Computer Science **3659**, Springer, Berlin, 2005.

[Shallit 1979] J. Shallit, “Simple continued fractions for some irrational numbers”, *J. Number Theory* **11**:2 (1979), 209–217. MR Zbl

[Smith 2015] B. R. Smith, “End-symmetric continued fractions and quadratic congruences”, *Acta Arith.* **167**:2 (2015), 173–187. MR Zbl

[Wilker 1980] P. Wilker, “An efficient algorithmic solution of the Diophantine equation $u^2+5v^2=m$”, *Math. Comp.* **35**:152 (1980), 1347–1352. MR Zbl

Received: 2013-09-10 Revised: 2015-05-06 Accepted: 2016-07-11

christina.doran@chubb.com *Lebanon Valley College, 101 College Ave.,*
Annville, PA 17003, United States

shlu6807@colorado.edu *Lebanon Valley College, 101 College Ave.,*
Annville, PA 17003, United States

barsmith@lvc.edu *Lebanon Valley College, 101 College Ave.,*
Annville, PA 17003, United States

INVOLVE 10:4 (2017)
dx.doi.org/10.2140/involve.2017.10.555

New approximations for the area of the Mandelbrot set

Daniel Bittner, Long Cheong, Dante Gates and Hieu D. Nguyen

(Communicated by Kenneth S. Berenhaut)

Due to its fractal nature, much about the area of the Mandelbrot set M remains to be understood. While a series formula has been derived by Ewing and Schober (1992) to calculate the area of M by considering its complement inside the Riemann sphere, to date the exact value of this area remains unknown. This paper presents new improved upper bounds for the area based on a parallel computing algorithm and for the 2-adic valuation of the series coefficients in terms of the sum-of-digits function.

1. Introduction

The Mandelbrot set (hereafter M) is defined as the set of complex numbers $c \in \mathbb{C}$ such that the sequence $\{z_n\}$ defined by the recursion

$$z_n = z_{n-1}^2 + c, \tag{1}$$

with initial value $z_0 = 0$, remains bounded for all $n \geq 0$. Douady and Hubbard [1982] proved that M is connected and Shishikura [1998] proved that M has a fractal boundary of Hausdorff dimension 2. However, it is unknown whether the boundary has positive Lebesgue measure.

Ewing and Schober [1992] derived a series formula for the area of M by considering its complement, $\tilde{M}$, inside the Riemann sphere $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$, i.e., $\tilde{M} = \overline{\mathbb{C}} - M$. It is known that $\tilde{M}$ is simply connected with mapping radius 1 [Douady and Hubbard 1982]. In other words, there exists an analytic homeomorphism

$$\psi(z) = z + \sum_{m=0}^{\infty} b_m z^{-m} \tag{2}$$

that maps the domain $\Delta = \{z : 1 < |z| \leq \infty\} \subset \overline{\mathbb{C}}$ onto $\tilde{M}$. It follows from the classic result of Gronwall [1914/15] that the area of the Mandelbrot set $M = \overline{\mathbb{C}} - \tilde{M}$

MSC2010: 37F45.
Keywords: Mandelbrot set, sum of digits.

is given by

$$A=\pi\left[1-\sum_{m=1}^{\infty} m|b_m|^2\right]. \tag{3}$$

The arithmetic properties of the coefficients b_m have been studied in depth, first by Jungreis [1985], then independently by Levin [1989; 2014], Bielefeld, Fisher, and Haeseler [Bielefeld et al. 1993], Ewing and Schober [1990; 1992], and more recently by Shimauchi [2015]. We note that the results of Levin [1989; 2014] and Shimauchi [2015] hold for Multibrot sets defined by generalizing (1) to higher-order recurrences.

There are three approaches to calculating the coefficients b_m. The first approach involves expressing b_m as a contour integral, found independently by Levin [1989] and by Ewing and Schober [1990]:

$$b_m=-\frac{1}{2\pi m i}\int_{|z|=R} p_n(z)^{m/2^n}\,\mathrm{d}z, \tag{4}$$

where $1\le m\le 2^{n+1}-3$ and R is chosen sufficiently large. The polynomials $p_n(w)$ in (4) are defined recursively by

$$\begin{aligned} p_0(w)&=w,\\ p_n(w)&=p_{n-1}^2(w)+w. \end{aligned} \tag{5}$$

Ewing and Schober [1990] proved that the polynomials $p_n(w)$ are Faber polynomials of degree 2^n for M, i.e., $p_n(\psi(z))=z^{2^n}+o(1)$ as $z\to\infty$, a fact that they used to prove (4). Jungreis [1985] proved earlier that $b_{2^{n+1}}=0$ for $n\ge 1$ (see also [Bielefeld et al. 1993; Ewing and Schober 1990; Levin 1989]). Bielefeld, Fisher, and Haeseler [Bielefeld et al. 1993] proved that no constants ϵ and K exist so that $|b_m|<K/m^{1+\epsilon}$ for all m.

The second approach to calculating b_m, due to [Bielefeld et al. 1993], involves substituting (2) into (5) to obtain

$$p_n(\psi(z))=p_{n-1}^2(\psi(z))+\psi(z)=z^{2^n}+o(1),$$

and then equating coefficients to recursively solve for b_m. In this paper, we follow a variation of this approach, due to Ewing and Schober [1992], by expanding $p_n(\psi(z))$ in the form

$$p_n(\psi(z))=\sum_{m=0}^{\infty}\beta_{n,m}z^{2^n-m}, \tag{6}$$

where $b_m = \beta_{0,m+1}$. It follows that $\beta_{n,m} = 0$ for $n \geq 1$ and $1 \leq m \leq 2^n$.Moreover, this range of zero values can be extended to $1 \leq m \leq 2^{n+1} - 2$ because of the recursion

$$\beta_{n,m} = 2\beta_{n-1,m} + \sum_{k=1}^{m-1} \beta_{n-1,k}\beta_{m-1,m-k}, \tag{7}$$

which can be derived by substituting (6) into (5) and equating coefficients. Formula (7) can then be manipulated to obtain the backward recursion formula [Ewing and Schober 1992]

$$\beta_{n,m} = \frac{1}{2}\left[\beta_{n+1,m} - \sum_{k=2^{n+1}-1}^{m-2^{n+1}+1} \beta_{n,k}\beta_{n,m-k} - \beta_{0,m-2^{n+1}+1}\right], \tag{8}$$

where $\beta_{n,0} = 1$ and $\beta_{0,m} = b_{m-1}$ for $m \geq 1$.

No explicit formula is known for $\beta_{n,m}$ (nor for b_m), except those at certain positions. However, it is clear from (8) that $\beta_{n,m}$ is rational and that its denominator equals a power of 2 when expressed in lowest terms. In their paper, Ewing and Schober [1992] established the following upper bound on its 2-adic valuation.

Theorem 1 [Ewing and Schober 1992]. *Let $n \in \mathbb{N}$ and m be a positive integer. Then $2^{2m+3-2^{n+2}}\beta_{n,m}$ is an integer, i.e.,*

$$-\nu(\beta_{n,m}) \leq 2m + 3 - 2^{n+2} \tag{9}$$

for nonzero $\beta_{n,m}$.

Here, the 2-adic valuation $\nu(x)$ of a positive integer x is defined to be the greatest integer for which $2^{\nu(x)}$ divides x, and if x/y is a fraction in lowest terms, then we define $\nu(x/y) = \nu(x) - \nu(y)$. If $x = 0$, then we set $\nu(x) = \infty$. Observe that in the special case $b_m = \beta_{0,m+1}$, (9) reduces to

$$-\nu(b_m) \leq 2m - 1. \tag{10}$$

Zagier [Bielefeld et al. 1993] observed earlier that

$$-\nu(b_m) \leq \nu((2m+2)!)$$

for $0 \leq m \leq 1000$. Moreover, he observed that equality holds when m is odd (or zero). These results were later proven by Levin [1989] and Shimauchi [2015].

Theorem 2 [Levin 1989]. *If m is a positive odd integer, then*

$$-\nu(b_m) = \nu((2m+2)!). \tag{11}$$

Theorem 3 [Shimauchi 2015]. *Let m be a nonnegative integer. Then*

$$-\nu(b_m) \leq \nu((2m+2)!). \tag{12}$$

Moreover, equality holds precisely when m is odd.

Ewing and Schober [1992] used (8) to compute the first 240,000 coefficients for b_n by computer. Since

$$A \leq A_N \equiv \pi \left[1 - \sum_{m=1}^{N} m|b_m|^2 \right], \tag{13}$$

their calculation of $A_{240,000} \approx 1.7274$ yielded an upper bound for the area of M. They were able to slightly improve their result to 1.72 by extending their computations to the first 500,000 coefficients as reported by Ewing [1995]. They also calculated a crude lower bound of $7\pi/16 \approx 1.3744$ by estimating the size of the main cardioid ($3\pi/8$) and the main bulb ($\pi/16$). However, they reported a discrepancy with their approximation of 1.52 obtained by pixel counting. More recent calculations by Förstemann [2012] provide an estimate of 1.50659 based on a resolution of almost 88 trillion pixels. In addition, Andreadis and Karakasidis [2015] obtained an estimate of 1.5052 based on the boundary scanning method. Thus, as noted by Ewing and Schober, either the series (3) converges so slowly that the approximation $A_{500,000} \approx 1.72$ is poor or else the pixel counting method fails to account for the boundary of *M*. Recently, Buff and Chéritat [2012] found Julia sets with positive area. Therefore, coupled with Shishikura's result that the boundary of M has Hausdorff dimension 2, it is not far-fetched to suspect that the boundary of M may have positive area.

In this paper, we report on progress in obtaining new upper bounds for A and new results involving the two-dimensional sequence $\beta_{n,m}$. In particular, we were able to compute the first five million coefficients for b_n by developing a parallel processing implementation of (8). This extends the calculation of the first one million coefficients by Chen, Kawahira, Li, and Yuan [Chen et al. 2011] by five-fold, where they reported the upper bound $A_{1,000,000} = 1.703927$. As a result of our calculations, we obtained the new upper bound

$$A_{5,000,000} \approx 1.68288. \tag{14}$$

Moreover, we were able to improve on (9) by establishing the tighter bound (Theorem 9)

$$-\nu(\beta_{n,m}) \leq 2m - 2^{n+2} + 4 - s(n, m) \tag{15}$$

for nonzero $\beta_{n,m}$, where $s(n, m)$ is the base-2 sum-of-digits function of degree n (Definition 4). In the special case $b_m = \beta_{0,m+1}$, we obtain as a corollary

$$-\nu(b_m) \leq 2(m + 1) - s(0, m + 1). \tag{16}$$

This is equivalent to Shimauchi's result (12) because $\nu(k!) = k - s(0, k)$ for any positive integer k. Observe that the equality in (16) holds for all odd values of m, which follows from Shimauchi's result (Theorem 3), whereas (10) holds only when $m+1$ equals a power of 2.

Our new upper bound (15) is significant on two levels. First, from a computational perspective, it allows the values of $\beta_{n,m}$ to be calculated by integer arithmetic (as discussed by Ewing and Schober [1992]) using less memory than (9). Such an approach would increase the accuracy in which upper bounds for the area of the M are calculated over floating-point arithmetic where the values of $\beta_{n,m}$ are stored as truncated decimals. Secondly, (16) confirms Levin's work that the sum-of-digits function is a crucial ingredient in determining the exact area of M by using the series formula (3).

2. Two-adic valuation of $\beta_{n,m}$

In this section we consider the 2-adic valuation of $\beta_{n,m}$ and prove the bound (15), which is a refinement of (9). We begin by defining the sum-of-digits function and present a series of lemmas on properties of this function that will be utilized in the proof. Throughout this paper, $\mathbb{N}$ denotes the set of nonnegative integers.

Definition 4. Let $m \in \mathbb{N}$ with base-2 expansion $m = d_L 2^L + d_{L-1} 2^{L-1} + \cdots + d_0 2^0$, where $d_L = 1$ and $d_i \in \{0, 1\}$ for $i < L$. We define the *base-*2 *sum-of-digits function* $s(n, m)$ of *degree n* by

$$s(n, m) = \sum_{i=n}^{L} d_i.$$

Lemma 5. *Let $m, n \in \mathbb{N}$. Then $s(n,m)$ is subadditive, i.e.,*

$$s(n, l+m) \le s(n, l) + s(n, m)$$

for all $l \in \mathbb{N}$.

Proof. We follow the proof in [Rivoal 2008]. Let $l = c_K 2^K + c_{K-1} 2^{K-1} + \cdots + c_0 2^0$ and $m = d_L 2^L + d_{L-1} 2^{L-1} + \cdots + d_0 2^0$. Since $s(n, m+2^i) \le s(n, m)$ for $i < n$ and $s(n, m+2^i) \le s(n, m) + 1$ for $i \ge n$, it follows that

$$\begin{aligned} s(n, l+m) = s\Bigl(n, m + \sum_{i=0}^{K} c_i 2^i\Bigr) \le s\Bigl(n, m + \sum_{i=n}^{K} c_i 2^i\Bigr) \le s(n, m) + \sum_{i=n}^{K} c_i \\ \le s(n, m) + s(n, l). \end{aligned}$$ □

Lemma 6. *For all $m, n \in \mathbb{N}$, we have*

(a) $0 \le s(n, m) - s(n+1, m) \le 1$,

(b) $s(n, 2^{n+1} - 1) = 1$,

(c) $s(0, m) \le 2s\bigl(0, \tfrac{1}{2}m\bigr) - 1$

for positive even integers m.

Proof. (a) We express m as in Definition 4. It follows that

$$\begin{aligned} s(n,m)-s(n+1,m) &= \sum_{i=n}^{L} d_i - \sum_{i=n+1}^{L} d_i \\ &= d_n + \sum_{i=n+1}^{L} d_i - \sum_{i=n+1}^{L} d_i \\ &= d_n, \end{aligned}$$

where d_n must equal either 0 or 1. This completes the proof for (a).

(b) The result follows immediately from the fact that $2^{n+1}-1=2^0+\cdots+2^{n-1}+2^n$.

(c) Assume m is even. Then m can be expressed as

$$m=\sum_{i=r}^{L} d_i 2^i$$

for some integers r, L, where $r \geq 1$ by assumption. It follows that

$$\tfrac{1}{2}m=\sum_{i=r-1}^{L-1} d_{i+1}2^i.$$

Therefore,

$$\begin{aligned} s(0,m)=s\bigl(0,\tfrac{1}{2}m\bigr) &= 2s\bigl(0,\tfrac{1}{2}m\bigr)-s\bigl(0,\tfrac{1}{2}m\bigr) \\ &\leq 2s\bigl(0,\tfrac{1}{2}m\bigr)-1, \end{aligned}$$

since $s\bigl(0,\tfrac{1}{2}m\bigr) \geq 1$. □

Next, we present a lemma regarding the convolution described in (8).

Lemma 7. *Let $m \in \mathbb{N}$ with $m \geq 2^{n+2}-2$:*

(a) *For even m, we have*

$$\sum_{k=2^{n+1}-1}^{m-2^{n+1}+1} \beta_{n,k}\beta_{n,m-k} = 2\Biggl[\sum_{k=2^{n+1}-1}^{m/2-1} \beta_{n,k}\beta_{n,m-k}\Biggr] + (\beta_{n,m/2})^2. \tag{17}$$

(b) *For odd m, we have*

$$\sum_{k=2^{n+1}-1}^{m-2^{n+1}+1} \beta_{n,k}\beta_{n,m-k} = 2\Biggl[\sum_{k=2^{n+1}-1}^{(m-1)/2} \beta_{n,k}\beta_{n,m-k}\Biggr]. \tag{18}$$

Proof. When m is even, we have

$$\sum_{k=2^{n+1}-1}^{m-2^{n+1}+1} \beta_{n,k}\beta_{n,m-k} = \sum_{k=2^{n+1}-1}^{m/2-1} \beta_{n,k}\beta_{n,m-k}+\sum_{m/2}^{m/2} \beta_{n,k}\beta_{n,m-k}+\sum_{m/2+1}^{m-2^{n+1}+1} \beta_{n,k}\beta_{n,m-k}.$$

Letting $h = m - k$, we obtain

$$\begin{aligned}\sum_{k=2^{n+1}-1}^{m-2^{n+1}+1} \beta_{n,k}\beta_{n,m-k} &= \sum_{k=2^{n+1}-1}^{m/2-1} \beta_{n,k}\beta_{n,m-k}+(\beta_{n,m/2})(\beta_{n,m/2})+\sum_{h=m/2-1}^{2^{n+1}-1} \beta_{n,m-h}\beta_{n,h}\\ &= 2\Bigg[\sum_{k=2^{n+1}-1}^{m/2-1} \beta_{n,k}\beta_{n,m-k}\Bigg]+(\beta_{n,m/2})^2.\end{aligned}$$

This proves (a).

On the other hand, when m is odd,

$$\sum_{k=2^{n+1}-1}^{m-2^{n+1}+1} \beta_{n,k}\beta_{n,m-k} = \sum_{k=2^{n+1}-1}^{(m-1)/2} \beta_{n,k}\beta_{n,m-k} + \sum_{k=(m+2)/2}^{m-2^{n+1}+1} \beta_{n,k}\beta_{n,m-k}.$$

Letting $l = m - k$, we have

$$\begin{aligned}\sum_{k=2^{n+1}-1}^{m-2^{n+1}+1} \beta_{n,k}\beta_{n,m-k} &= \sum_{k=2^{n+1}-1}^{(m-1)/2} \beta_{n,k}\beta_{n,m-k} + \sum_{l=(m-1)/2}^{2^{n+1}-1} \beta_{n,m-l}\beta_{n,l}\\ &= 2\Bigg[\sum_{k=2^{n+1}-1}^{(m-1)/2} \beta_{n,k}\beta_{n,m-k}\Bigg].\end{aligned}$$

This justifies (b). □

We now present one final lemma involving the right-hand side of (15).

Lemma 8. *Let $m, n \in \mathbb{N}$ and define*

$$p(n, m) = 2m - 2^{n+2} + 4 - s(n, m). \tag{19}$$

Then the following inequalities hold:

(a) $p(n, m) - 1 \geq p(n + 1, m)$.

(b) $p(n, m) \geq p(n, k) + p(n, m - k)$ *for* $0 \leq k \leq m$.

(c) $p(0, m) - 1 \geq 2p(0, m/2)$ *for m is even.*

(d) $p(n, m) - 1 \geq p(0, m - 2^{n+1} + 1)$.

Proof. (a) Since $-1 \le s(n,m) - s(n+1,m) \le 0$ because of Lemma 6(a), we have

$$\begin{aligned} p(n,m) - 1 - p(n+1,m) &= 2^{n+3} - 2^{n+2} - 1 - s(n,m) + s(n+1,m) \\ &= 2^{n+2} - 1 + s(n,m) - s(n+1,m) \\ &\ge 2^{n+2} - 2 \ge 0. \end{aligned}$$

(b) Using subadditivity of $s(n,m)$ (Lemma 5) and the fact that $2^{n+2} - 4 \ge 0$ for $n \in \mathbb{N}$, we have

$$\begin{aligned} p(n,m) - p(n,k) - p(n,m-k) &\ge s(n,m-k) + s(n,k) - s(n,m) + 2^{n+2} - 4 \\ &\ge s(n,m-k) + s(n,k) - s(n,m) \ge 0. \end{aligned}$$

(c) We have

$$p(0,m) - 1 - 2p\big(0, \tfrac{1}{2}m\big) \ge 2s\big(0, \tfrac{1}{2}m\big) - 1 - s(0,m) \ge 0,$$

where the last inequality above follows from Lemma 6(c).

(d) We have

$$p(n,m) - 1 - p(0, m - 2^{n+1} + 1) \ge s(0, m - 2^{n+1} + 1) + 1 - s(n,m) \ge 0,$$

where last inequality above follows from Lemmas 5 and 6(b), namely

$$\begin{aligned} s(n,m) \le s(n, m - 2^{n+1} + 1 + 2^{n+1} - 1) &\le s(n, m - 2^{n+1} + 1) + s(n, 2^{n+1} - 1) \\ &\le s(0, m - 2^{n+1} + 1) + 1, \qquad \square \end{aligned}$$

We now have presented all lemmas needed to prove the following theorem.

Theorem 9. *Let $m, n \in \mathbb{N}$ and assume $m \ge 2^{n+1} - 1$. Then $2^{p(n,m)}\beta_{n,m}$ is an integer, i.e.,*

$$-\nu(\beta_{n,m}) \le p(n,m). \tag{20}$$

Proof. From (8) we have

$$2^{p(n,m)}\beta_{n,m} = 2^{p(n,m)-1}\Bigg[\beta_{n+1,m} - \sum_{k=2^{n+1}-1}^{m-2^{n+1}+1} \beta_{n,k}\beta_{n,m-k} - \beta_{0,m-2^{n+1}+1}\Bigg]$$

whose right-hand side can be rewritten as

$$2^{p(n,m)-1}\beta_{n+1,m} \sum_{k=2^{n+1}-1}^{m-2^{n+1}+1} 2^{p(n,m)-1}\beta_{n,k}\beta_{n,m-k} - 2^{p(n,m)-1}\beta_{0,m-2^{n+1}+1}. \tag{21}$$

It suffices to show that each term in (21) is an integer by induction on m, which we will do so using properties of $p(n,m)$ established in Lemma 8. Assume that the values of $\beta_{n,m}$ are arranged in a two-dimensional array where the rows are indexed by n and the columns indexed by m. Since $\beta_{n,m} = 0$ for $n \geq 1$ and $1 \leq m \leq 2^{n+1} - 2$, we shall call the values in this range trivial and those outside this range, i.e., $m \geq 2^{n+1} - 1$, nontrivial. It follows that each column has at most a finite number of nontrivial entries.

Therefore, we shall apply induction by moving upwards along the nontrivial values in each column from left to right following Ewing and Schober [1992]. We first establish the base case. Assume $n = 0$ and $m = 1$. Since $\beta_{0,1} = -\frac{1}{2}$ and $p(0, 1) = 1$, it is clear that $2^{p(0,1)}\beta_{0,1} = -1$ is an integer.

Next, to prove that the result holds for $\beta_{n,m}$, we assume inductively that $2^{p(j,k)}\beta_{j,k}$ is an integer for all columns to the left of $\beta_{n,m}$ and all entries below it, i.e., for $1 \leq k \leq m-1$, we have $0 \leq j \leq \log_2(k+1) - 1$ and for $k = m$, we have $n+1 \leq j \leq \log_2(m+1) - 1$, respectively. Let us consider the first term $2^{p(n,m)-1}\beta_{n+1,m}$ in (21). Since $p(n, m) - 1 \geq p(n+1, m)$ (due to Lemma 8(a)) and $2^{p(n+1,m)}\beta_{n+1,m}$ is an integer by the assumption, it follows that $2^{p(n,m)-1}\beta_{n+1,m}$ is an integer.

Next, we rewrite the summation in (21) according to whether m is even or odd by using Lemma 7. If m is odd, then

$$\sum_{k=2^{n+1}-1}^{m-2^{n+1}+1} 2^{p(n,m)-1}\beta_{n,k}\beta_{n,m-k} = \sum_{k=2^{n+1}-1}^{(m-1)/2} 2^{p(n,m)}\beta_{n,k}\beta_{n,m-k}.$$

Since $p(n, m) \geq p(n, k) + p(n, m-k)$ for $0 \leq k \leq m$ from Lemma 8(b) and

$$(2^{p(n,k)}\beta_{n,k})(2^{p(n,m-k)}\beta_{n,m-k})$$

is an integer by the assumption, it follows that each term $2^{p(n,m)-1}\beta_{n,k}\beta_{n,m-k}$ in the summation must be an integer. On the other hand, if m is even, then

$$\sum_{k=2^{n+1}-1}^{m-2^{n+1}+1} 2^{p(n,m)-1}\beta_{n,k}\beta_{n,m-k} = \sum_{k=2^{n+1}-1}^{m/2-1} 2^{p(n,m)}\beta_{n,k}\beta_{n,m-k} + 2^{p(n,m)-1}(\beta_{n,m/2})^2.$$

By the same argument as before, we have that $2^{p(n,m)}\beta_{n,k}\beta_{n,m-k}$ is an integer. Moreover, since $p(n, m) - 1 \geq 2p\big(n, \frac{1}{2}m\big)$ (due to Lemma 8(c)) and $2^{p(n,m/2)}\beta_{n,m/2}$ is an integer by the assumption, it follows that $2^{p(n,m)-1}(\beta_{n,m/2})^2$ must also be an integer. Thus, each term $2^{p(n,m)-1}\beta_{n,k}\beta_{n,m-k}$ in the summation must also be an integer.

As for the last term $2^{p(n,m)-1}\beta_{0,m-2^{n+1}+1}$ in (21), we know from Lemma 8(d) that $p(n, m) - 1 \geq p(0, m - 2^{n+1} + 1)$. Since $2^{p(0,m-2^{n+1}-1)}\beta_{0,m-2^{n+1}+1}$ is an integer by the assumption, it follows by the same reasoning that $2^{p(n,m)-1}\beta_{0,m-2^{n+1}+1}$ must be an integer. □

3. Special values of $\beta_{n,m}$

In this section we derive recurrences for special values of $\beta_{n,m}$ where m is restricted to a certain interval. Recall that $\beta_{n,m} = 0$ for $1 \leq m \leq 2^{n+1} - 2$. We therefore begin with an unpublished result by Malik Ahmed and one of the authors regarding $\beta_{n,m}$ in the interval $2^{n+1}-1 \le m \le 2^{n+2}-3$.

Theorem 10 (Ahmed–Nguyen). *Let $n \in \mathbb{N}$ and m be a positive integer satisfying $2^{n+1}-1 \le m \le 2^{n+2}-3$. Then for all $p \in \mathbb{N}$, we have*

$$\beta_{n,m} = \beta_{n+p,m+2^{n+1}(2^p-1)} = -\tfrac{1}{2}\beta_{0,m-2^{n+1}+1}. \tag{22}$$

Proof. It follows from (8) that

$$\beta_{n,m} = -\tfrac{1}{2}\beta_{0,m-2^{n+1}+1}. \tag{23}$$

Next, set

$$n' = n+p, \quad m' = m+2^{n+1}(2^p-1).$$

Then

$$m' - 2^{n'+1} + 1 = m - 2^{n+1} + 1,$$

which proves

$$\beta_{n,m} = \beta_{n',m'}, \tag{24}$$

as desired. □

As a corollary of Theorem 10, we establish a special case of (9).

Corollary 11. *Let n be a positive integer and m a positive integer satisfying $2^{n+1} \le m \le 2^{n+2}-3$. Then $2^{2m+2-2^{n+2}}\beta_{n,m}$ is an integer.*

Proof. We know from (9) that

$$2^{2(m-2^{n+1}+1)+3-2^2}\beta_{0,m-2^{n+1}+1} = 2^{2m+1-2^{n+2}}\beta_{0,m-2^{n+1}+1}$$

is an integer. It follows from Theorem 10 that

$$2^{2m+2-2^{n+2}}\beta_{n,m} = 2^{2m+2-2^{n+2}}\big(-\tfrac{1}{2}\beta_{0,m-2^{n+1}+1}\big) = -2^{2m+1-2^{n+2}}\beta_{0,m-2^{n+1}+1} \tag{25}$$

must also be an integer. □

Observe that the corollary above fails for $m = 2^{n+1}-1$. By Theorem 10 we have $\beta_{n,2^{n+1}-1} = -\frac{1}{2}\beta_{0,0}$, but (9) doesn't apply to $\beta_{0,0} = 1$.

We next focus on deriving recurrences for special values of $\beta_{n,m}$ located at certain positions for m between $2^{n+2}-2$ and $2^{n+2}+6$.

Lemma 12. *Let $n \in \mathbb{N}$. Then*

$$\beta_{n,2^{n+2}-2} = -\tfrac{1}{2}\big(\beta_{0,2^{n+1}-1} + \tfrac{1}{4}\big), \tag{26}$$

$$\beta_{n,2^{n+2}-1} = -\tfrac{1}{2}\big(\beta_{0,2^{n+1}} + \tfrac{1}{4}\big). \tag{27}$$

Proof. Recall that $\beta_{n,m} = -\frac{1}{2}\beta_{0,m-2^{n+1}+1}$ for $n \geq 0$ and $2^{n+1}-1 \leq m \leq 2^{n+2}-3$. We have

$$\begin{aligned}
\beta_{n,2^{n+2}-2} &= \frac{1}{2}\Bigg[\beta_{n+1,2^{n+2}-2} - \sum_{k=2^{n+1}-1}^{2^{n+1}-1} \beta_{n,k}\beta_{n,2^{n+2}-2-k} - \beta_{0,2^{n+1}-1}\Bigg] \\
&= \frac{1}{2}\big[0 - \beta^2_{n,2^{n+1}-1} - \beta_{0,2^{n+1}-1}\big] \\
&= \tfrac{1}{2}\big[-\tfrac{1}{4}\beta^2_{0,0} - \beta_{0,2^{n+1}-1}\big] \\
&= -\tfrac{1}{2}\big[\beta_{0,2^{n+1}-1} + \tfrac{1}{4}\big],
\end{aligned}$$

and

$$\begin{aligned}
\beta_{n,2^{n+2}-1} &= \tfrac{1}{2}\Bigg[\beta_{n+1,2^{n+2}-1} - \sum_{k=2^{n+1}-1}^{2^{n+1}} \beta_{n,k}\beta_{n,m-k} - \beta_{0,2^{n+1}}\Bigg] \\
&= \tfrac{1}{2}\big[\beta_{n+1,2^{n+2}-1} - 2(\beta_{n,2^{n+1}-1}\beta_{n,2^{n+1}}) - \beta_{0,2^{n+1}}\big] \\
&= \tfrac{1}{2}\big[\big(-\tfrac{1}{2}\beta_{0,0}\big) - 2\big(\big(-\tfrac{1}{2}\beta_{0,0}\big)\big(-\tfrac{1}{2}\beta_{0,1}\big)\big) - \beta_{0,2^{n+1}}\big] \\
&= -\tfrac{1}{2}\big[\beta_{0,2^{n+1}} + \tfrac{1}{4}\big]. \qquad \square
\end{aligned}$$

In the case where $m = 2^{n+2}$, we find that $\beta_{n,m}$ is constant.

Lemma 13. *Let n be a positive integer. Then $\beta_{n,2^{n+2}} = \frac{1}{16}$.*

Proof. Recall from Theorem 10 that $\beta_{n,m} = -\frac{1}{2}\beta_{0,m-2^{n+1}+1}$ for $n \geq 0$ and $2^{n+1}-1 \leq m \leq 2^{n+2}-3$. Moreover, recall that $b_{2^{n+1}} = 0$ for $n \geq 1$ [Jungreis 1985]. Using these results, we have

$$\begin{aligned}
\beta_{n,2^{n+2}} &= \frac{1}{2}\Bigg[\beta_{n+1,2^{n+2}} - \sum_{k=2^{n+1}-1}^{2^{n+1}+1} \beta_{n,k}\beta_{n,2^{n+2}-k} - \beta_{0,2^{n+1}+1}\Bigg] \\
&= \tfrac{1}{2}\big[-\tfrac{1}{2}\beta_{0,1} - 2\beta_{n,2^{n+1}-1}\beta_{n,2^{n+1}+1} - \beta^2_{n,2^{n+1}} - b_{0,2^{n+1}}\big] \\
&= \tfrac{1}{2}\big[-\tfrac{1}{2}\beta_{0,1} - \tfrac{1}{2}\beta_{0,0}\beta_{0,2} - \tfrac{1}{4}\beta^2_{0,1} - 0\big] \\
&= \tfrac{1}{2}\big[-\tfrac{1}{2}\big(-\tfrac{1}{2}\big) - \tfrac{1}{2}(1)\big(\tfrac{1}{8}\big) - \tfrac{1}{4}\big(-\tfrac{1}{2}\big)^2\big] = \tfrac{1}{16}. \qquad \square
\end{aligned}$$

We end this section by considering three other special cases.

Lemma 14. *Let $n \in \mathbb{N}$. Then:*

(a) $\beta_{n,2^{n+2}+2} = -\frac{1}{2}\beta_{0,2^{n+1}+3}$ *for* $n \geq 2$.

(b) $\beta_{n,2^{n+2}+4} = -\frac{1}{2}\beta_{0,2^{n+1}+5}$ *for* $n \geq 2$.

(c) $\beta_{n,2^{n+2}+6} = -\frac{1}{2}\beta_{0,2^{n+1}+7}$ *for* $n \geq 3$.

Proof. We have

$$\begin{aligned}
\beta_{n,2^{n+2}+2} &= \frac{1}{2}\Bigg[\beta_{n+1,2^{n+2}+2} - \sum_{k=2^{n+1}-1}^{2^{n+1}+3} \beta_{n,k}\beta_{n,2^{n+2}+2-k} - \beta_{0,2^{n+1}+3}\Bigg] \\
&= \frac{1}{2}\Bigg[\beta_{n+1,2^{n+2}+2} - 2\sum_{k=2^{n+1}-1}^{2^{n+1}} \beta_{n,k}\beta_{n,2^{n+2}+2-k} - \beta_{n,2^{n+1}+1}^2 - \beta_{0,2^{n+1}+3}\Bigg] \\
&= \frac{1}{2}\Bigg[-\tfrac{1}{2}\beta_{0,3} - \tfrac{1}{2}\sum_{j=0}^{1} \beta_{0,j}\beta_{0,4-j} - \tfrac{1}{4}\beta_{0,2}^2 - \beta_{0,2^{n+1}+3}\Bigg] \\
&= \tfrac{1}{2}\big[-\tfrac{1}{2}\big(-\tfrac{1}{4}\big) - \tfrac{1}{2}(\beta_{0,0}\beta_{0,4} + \beta_{0,1}\beta_{0,3}) - \tfrac{1}{4}\big(\tfrac{1}{8}\big)^2 - \beta_{0,2^{n+1}+3}\big] \\
&= \tfrac{1}{2}\big[-\tfrac{1}{2}\big(-\tfrac{1}{4}\big) - \tfrac{1}{2}\big[(1)\big(\tfrac{15}{128}\big) + \big(-\tfrac{1}{2}\big)\big(-\tfrac{1}{4}\big)\big] - \tfrac{1}{4}\big(\tfrac{1}{8}\big)^2 - \beta_{0,2^{n+1}+3}\big] \\
&= -\tfrac{1}{2}\beta_{0,2^{n+1}+3}.
\end{aligned}$$

This proves (a), and (b) and (c) can be proven in a similar manner. □

4. New area approximations

In this section, we describe a parallel processing algorithm to compute the values of $\beta_{n,m}$ and present new upper bounds for the area of M that were calculated using this algorithm. Assume as before that the values of $\beta_{n,m}$ are arranged in a two-dimensional array with the rows indexed by n and columns indexed by m. We recall Ewing and Schober's backwards algorithm for computing the nontrivial values of $\beta_{n,m}$ recursively one at a time by moving upwards along each column from left to right as described in our induction proof of Theorem 9. Thus, the order of computation would be $\beta_{0,1}, \beta_{0,2}, \beta_{1,3}, \beta_{0,3}, \beta_{1,4}, \beta_{0,4}, \ldots$

Our new method is as follows: We calculate values of $\beta_{n,m}$ across multiple columns simultaneously in a parallel fashion while moving up along them as before until we reach a critical row near the top, where from this point on, all remaining column values must be computed one at a time. This is then repeated for the next set of columns, etc.

To illustrate our method, consider for example the calculation of $\beta_{1,7}$ and $\beta_{1,8}$ in row $n = 1$ using the backward recursion formula (8):

$$\beta_{1,7} = \frac{1}{2}\Bigg[\beta_{2,7} - \sum_{k=3}^{4} \beta_{1,k}\beta_{1,7-k} - \beta_{0,4}\Bigg] = \tfrac{1}{2}[\beta_{2,7} - 2\beta_{1,3}\beta_{1,4} - \beta_{0,4}],$$

$$\beta_{1,8} = \frac{1}{2}\Bigg[\beta_{1,8} - \sum_{k=3}^{4} \beta_{1,k}\beta_{1,8-k} - \beta_{0,5}\Bigg] = \tfrac{1}{2}[\beta_{2,8} - 2\beta_{1,3}\beta_{1,5} - \beta_{1,4}^2 - \beta_{0,4}].$$

These two values do not depend on each other and can be computed independently in parallel. However, this is not the case for $\beta_{0,7}$ and $\beta_{0,8}$ in the top row ($n = 0$), where the latter depends on the former:

$$\begin{aligned}\beta_{0,7} &= \frac{1}{2}\left[\beta_{1,7} - \sum_{k=1}^{6} \beta_{0,k}\beta_{0,7-k} - \beta_{0,6}\right]\\ &= \tfrac{1}{2}[\beta_{1,7} - 2\beta_{0,1}\beta_{0,6} - 2\beta_{0,2}\beta_{0,5} - 2\beta_{0,3}\beta_{0,4} - \beta_{0,4}],\\ \beta_{0,8} &= \frac{1}{2}\left[\beta_{1,8} - \sum_{k=1}^{7} \beta_{0,k}\beta_{0,8-k} - \beta_{0,7}\right]\\ &= \tfrac{1}{2}[\beta_{1,8} - 2\beta_{0,1}\beta_{0,7} - 2\beta_{0,2}\beta_{0,6} - 2\beta_{0,3}\beta_{0,5} - \beta_{0,4}^2 - \beta_{0,7}].\end{aligned}$$

In general, the values $\beta_{n,m}$ and $\beta_{n,m+1}$ and $\beta_{n,m+2}$ in three consecutive columns can be calculated in parallel as long as $n \geq 1$. This is because $\beta_{n,m+1}$ depends only on the values $\beta_{n,k}$ in row n, where $k = 3, 4, \ldots, m-2$, which are prior to $\beta_{n,m}$. Similarly, $\beta_{n,m+2}$ depends only on $\beta_{n,k}$, where $k = 3, 4, \ldots, m-1$. Since the number of nonzero values in each column increases with m, this parallel algorithm becomes more effective and asymptotically three times as fast than if calculating the $\beta_{n,m}$ one at a time. This approach can be extended to calculate the values $\beta_{n,m}$, $\beta_{n,m+1}, \ldots, \beta_{n,m+6}$ in seven consecutive columns simultaneously as long as $n \geq 2$. More generally, if $n \geq N$, then up to $2^{N+1} - 1$ columns can be computed in parallel.

We were able to use this parallel algorithm to calculate the first five million terms of b_m and obtain a new upper bound of $A_{5,000,000} \approx 1.68288$ for the area of the Mandelbrot set. This algorithm was implemented using the programming language C++ and message passing interface Open MPI. In particular, we calculated the values of $\beta_{n,m}$ across four columns in parallel for $n \geq 2$, beginning with the first group of columns $\beta_{n,8}$, $\beta_{n,9}$, $\beta_{n,10}$, $\beta_{n,11}$ (we initialized columns $\beta_{n,0}, \ldots, \beta_{n,7}$ with their known values). Our code was executed on a Linux cluster with 32 GB of available RAM and required four processors (1.05 Ghz AMD Opteron 2352 quad-core processors) to execute it since each column was computed using a different processor. After computing its column of values, each processor would pass these values to the other three processors before calculating to its next designated column. Thus, each processor was required to store all values of $\beta_{m,n}$ (generated from all four processors) separately in its own RAM in order to compute its next column. This parallel approach improved the performance of our implementation significantly; asymptotically, the run-time was decreased by a factor of four in comparison to using a single processor, but at the cost of quadrupling our RAM memory requirements. It is possible to reduce this cost using shared memory; however, we did not implement this approach since we had sufficient RAM available. The only computational cost to our algorithm involves having each processor pass its values to the other three processors. Since the number of nonzero values for $\beta_{n,m}$ in each column grows on

N (millions)	A_N	N (millions)	A_N
0.5	1.72 [Ewing and Schober 1992]	3	1.68895
1	1.70393 [Chen et al. 2011]	3.5	1.6874
1.5	1.69702	4	1.68633
2	1.69388	4.5	1.68447
2.5	1.69096	5	1.68288

Table 1. New upper bounds for the area of the Mandelbrot set.

the order of $\log_2 m$, the computational cost in passing these values is insignificant in comparison to the cost of computing $\beta_{n,m}$ itself using (8), whose summation term grows on the order of m since $2^{n+1}-1 \le m \le 2^{n+2}-2$.

Table 1 gives values for the approximations A_N, where N ranges from 500,000 to 5 million in increments of 500,000, based on our computed values of $\beta_{n,m}$, and thus $b_m = \beta_{0,m+1}$. These values were computed in batches over a five-month period between August and December of 2014, although the actual total run-time was approximately 3 months. Table 2 gives a sense of the run-time required to compute b_m in batches of 500,000 starting at $m = 2{,}500{,}000$.

To estimate the error in our upper bounds, we use Ewing and Schober's [1992] analysis of their calculation of $\beta_{n,m}$ using (8) and double-precision floating-point arithmetic. First, they considered propagation error due to errors in computing previous coefficients. They argued probabilistically that the propagation error is on the same order of magnitude as machine error, so the computations for $\beta_{n,m}$ are stable. That is, write

$$\tilde{\beta}_{n,m} = \beta_{n,m} + \epsilon_{n,m}, \tag{28}$$

where $\beta_{n,m}$ is the true value, $\tilde{\beta}_{n,m}$ the calculated value, and $\epsilon_{n,m}$ the corresponding error. Substituting $\epsilon_{n,m} = \beta_{n,m} - \tilde{\beta}_{n,m}$ into (8) gives for the propagation error

$$\begin{aligned}\epsilon_{n,m} &\\ =&\frac{1}{2}\biggl(\beta_{n+1,m}-\tilde{\beta}_{n+1,m}-\sum_{k=2^{n+1}-1}^{m-2^{n+1}+1}(\beta_{n,k}\beta_{n,m-k}-\tilde{\beta}_{n,k}\tilde{\beta}_{n,m-k})-\beta_{0,m-2^{n+1}+1}+\tilde{\beta}_{0,m-2^{n+1}+1}\biggr)\\ \approx&\tfrac{1}{2}(-\epsilon_{n+1,m}+\epsilon_{0,m-2^{n+1}+1})+\sum_{k=2^{n+1}-1}^{m-2^{n+1}+1}\beta_{n,k}\epsilon_{n,m-k},\end{aligned} \tag{29}$$

range of m (millions)	2.5–3	3–3.5	3.5–4	4–4.5	4.5–5
runtime to compute b_m (days)	9	10.8	12.5	14.4	16.2

Table 2. Runtimes for calculating b_m in batches of 500,000.

where the quadratic error terms are ignored. Next, assume that $\epsilon_{n,m}$ is uniformly distributed with a small probability of exceeding some threshold value ϵ. Moreover, assume that the sum

$$E_{n,m} = \sum_{k=2^{n+1}-1}^{m-2^{n+1}+1} |\beta_{n,k}|$$

is bounded, which we verified in computing $A_{5{,}000{,}000}$. In particular, we found $E_{n,m}$ to be approximately bounded by

$$E_{n,m} \leq 13.2254$$

for $m = 5{,}000{,}000$ and $0 \leq n \leq 21$ with equality holding when $n = 9$. It follows from the law of large numbers that the error term in (29),

$$\sum_{k=2^{n+1}-1}^{m-2^{n+1}+1} \beta_{n,k}\epsilon_{n,m-k}, \tag{30}$$

which we view as a weighted sum of independent and identically distributed random variables, approaches zero as $m \to \infty$. Thus, (30) is negligible in contributing towards the propagation error in (29). Hence, if all previous errors are bounded by ϵ, then so is the propagation error.

To check the accuracy of our calculations, we compared our calculated values of b_m (in double-precision floating point format) with exact values that are given by closed formulas at certain positions. For example, Levin [1989] and Ewing and Schober [1992] proved independently that $b_m = 0$ for all $m = (2k+1)2^\nu$, where $k, \nu \in \mathbb{N}$ satisfy $k + 3 \leq 2^\nu$. We confirmed this for our calculated values. Moreover, Ewing and Schober [1992] proved that

$$b_m = \begin{cases} \dfrac{-1}{2^{\nu+3}(2^\nu-1)}\dbinom{2^\nu-\frac{5}{2}}{2^\nu-2}, & m = (2^\nu-1)2^\nu,\ \nu \geq 1, \\ \dfrac{3(2^\nu-6)}{2^{\nu+5}(2^\nu+1)(2^\nu-5)}\dbinom{2^\nu-\frac{3}{2}}{2^\nu-1}, & m = (2^\nu+1)2^\nu,\ \nu \geq 2, \\ \dfrac{-(214{\cdot}2^{3\nu}-767{\cdot}2^{2\nu}+146{\cdot}2^\nu+452)}{2^{\nu+8}(2^{\nu+1}-7)(2^{2\nu}-1)(2^\nu+2)}\dbinom{2^\nu-\frac{5}{2}}{2^\nu-2}, & m = (2^\nu+3)2^\nu,\ \nu \geq 2. \end{cases}$$

A comparison of these exact values of b_m with our calculated values yielded a maximum error of $5.00034 \cdot 10^{-16}$. Thus, as summarized in [Ewing and Schober 1992], the computations suggest that the error in calculating b_m for $m \leq 5{,}000{,}000$ is at most $6 \cdot 10^{-16}$ and it is expected that the error in our upper bound for $A_{5{,}000{,}000}$ is at most $3 \cdot 10^{-9}$. We note that our calculation of $A_{1{,}000{,}000} \approx 1.70393$ agrees with that reported in [Chen et al. 2011]. In Table 3, we give values for b_m at certain

m	b_m	m	b_m
500,000	$5.5221313 \cdot 10^{-8}$	3,000,000	$8.150385 \cdot 10^{-9}$
1,000,000	$-4.7138830 \cdot 10^{-8}$	3,500,000	$-3.911993 \cdot 10^{-9}$
1,500,000	$8.4477641 \cdot 10^{-8}$	4,000,000	$2.315128 \cdot 10^{-9}$
2,000,000	$-6.4378660 \cdot 10^{-9}$	4,500,000	$-8.87746 \cdot 10^{-9}$
2,500,000	$1.6594295 \cdot 10^{-8}$	5,000,000	$8.0532 \cdot 10^{-11}$

Table 3. Some calculated values of b_m at positions where no closed formula is known.

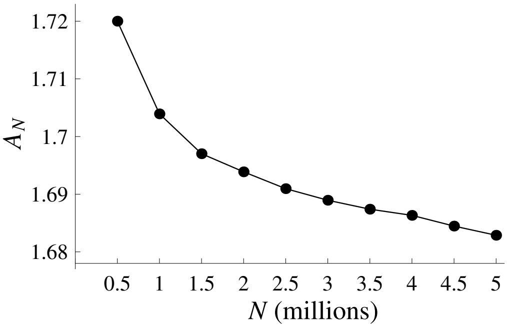


Figure 1. Plot of A_N.

positions where no closed formula is known so that the reader may verify our calculations.

Figure 1 shows a plot of Table 1 that clearly reveals the slow convergence of A_N. If the exact value of A lies closer to 1.50659 as computed by pixel counting, then certainly using A_N to closely approximate A is impractical due to the extremely large number of terms required. On the other hand, if the exact value lies closer to 1.68, then this would indicate that the boundary of the Mandelbrot set may have positive area.

5. Conclusions

In this paper we presented new results which improve on known upper bounds for the area of the Mandelbrot set and 2-adic valuations of the series coefficients $\beta_{n,m}$ given by Ewing and Schober [1992]. Of course, our calculations of the first five million terms of b_m were performed using more powerful computers that those available to Ewing and Schober two decades ago. Therefore, it would be interesting to find out in the next two decades what improvements can be made to our results by using computers that will be even more powerful, unless we are fortunate enough to see the exact area calculated before then.

Acknowledgments

The authors wish to thank the reviewer for providing us with feedback that greatly improved this paper and bringing to our attention important references that allowed us to give a more accurate survey of previous results. We also wish to thank Mark Sedlock and Jason Faulcon for their technical support of the computing cluster at Rowan University on which our computations were performed.

References

[Andreadis and Karakasidis 2015] I. Andreadis and T. E. Karakasidis, "On a numerical approximation of the boundary structure and of the area of the Mandelbrot set", *Nonlinear Dynam.* **80**:1–2 (2015), 929–935. MR

[Bielefeld et al. 1993] B. Bielefeld, Y. Fisher, and F. v. Haeseler, "Computing the Laurent series of the map $\Psi : \mathbf{C} - \bar{D} \to \mathbf{C} - M$", *Adv. Appl. Math.* **14**:1 (1993), 25–38. MR Zbl

[Buff and Chéritat 2012] X. Buff and A. Chéritat, "Quadratic Julia sets with positive area", *Ann. of Math.* **176**:2 (2012), 673–746. MR Zbl

[Chen et al. 2011] Y.-C. Chen, T. Kawahira, H.-L. Li, and J.-M. Yuan, "Family of invariant Cantor sets as orbits of differential equations, II: Julia sets", *Int. J. Bifur. Chaos Appl. Sci. Eng.* **21**:1 (2011), 77–99. MR Zbl

[Douady and Hubbard 1982] A. Douady and J. H. Hubbard, "Itération des polynômes quadratiques complexes", *C. R. Acad. Sci. Paris* **294**:3 (1982), 123–126. MR Zbl

[Ewing 1995] J. Ewing, "Can we see the Mandelbrot set?", *College Math. J.* **26**:2 (1995), 90–99.

[Ewing and Schober 1990] J. H. Ewing and G. Schober, "On the coefficients of the mapping to the exterior of the Mandelbrot set", *Michigan Math. J.* **37**:2 (1990), 315–320. MR Zbl

[Ewing and Schober 1992] J. H. Ewing and G. Schober, "The area of the Mandelbrot set", *Numer. Math.* **61**:1 (1992), 59–72. MR Zbl

[Förstemann 2012] T. Förstemann, "Numerical estimation of the area of the Mandelbrot set", 2012, available at http://tinyurl.com/zpodkk3.

[Gronwall 1914/15] T. H. Gronwall, "Some remarks on conformal representation", *Ann. of Math.* **16**:1–4 (1914/15), 72–76. MR Zbl

[Jungreis 1985] I. Jungreis, "The uniformization of the complement of the Mandelbrot set", *Duke Math. J.* **52**:4 (1985), 935–938. MR Zbl

[Levin 1989] G. M. Levin, "On the theory of iterations of polynomial families in the complex plane", *Teor. Funktsiĭ Funktsional. Anal. i Prilozhen.* 51 (1989), 94–106. In Russian; translated in *J. Soviet Math.* **52**:6 (1990), 3512–3522. MR Zbl

[Levin 2014] G. Levin, "On the Laurent coefficients of the Riemann map for the complement of the Mandelbrot set", 2014. arXiv

[Rivoal 2008] T. Rivoal, "On the bits counting function of real numbers", *J. Aust. Math. Soc.* **85**:1 (2008), 95–111. MR Zbl

[Shimauchi 2015] H. Shimauchi, "A remark on Zagier's observation of the Mandelbrot set", *Osaka J. Math.* **52**:3 (2015), 737–746. MR Zbl

[Shishikura 1998] M. Shishikura, "The Hausdorff dimension of the boundary of the Mandelbrot set and Julia sets", *Ann. of Math.* **147**:2 (1998), 225–267. MR Zbl

Received: 2014-10-05 Revised: 2016-09-19 Accepted: 2016-10-17

bittne12@students.rowan.edu *Department of Mathematics, Rowan University, Robinson Hall,*
201 Mullica Hill Road, Glassboro, NJ 08028, United States

cheong94@students.rowan.edu *Department of Mathematics, Rowan University, Robinson Hall,*
201 Mullica Hill Road, Glassboro, NJ 08028, United States

gatesd78@students.rowan.edu *Department of Mathematics, Rowan University, Robinson Hall,*
201 Mullica Hill Road, Glassboro, NJ 08028, United States

nguyen@rowan.edu *Department of Mathematics, Rowan University, Robinson Hall,*
201 Mullica Hill Road, Glassboro, NJ 08028, United States

mathematical sciences publishers msp

INVOLVE 10:4 (2017)
dx.doi.org/10.2140/involve.2017.10.573

Bases for the global Weyl modules of $\mathfrak{sl}_n$ of highest weight $m\omega_1$

Samuel Chamberlin and Amanda Croan

(Communicated by Jim Haglund)

We utilize a theorem of B. Feigin and S. Loktev to give explicit bases for the global Weyl modules for the map algebras of the form $\mathfrak{sl}_n \otimes A$ of highest weight $m\omega_1$. These bases are given in terms of specific elements of the universal enveloping algebra, $\boldsymbol{U}(\mathfrak{sl}_n \otimes A)$, acting on the highest weight vector.

1. Introduction

Let $\mathfrak{g}$ be a simple finite-dimensional complex Lie algebra. For the loop algebras $\mathfrak{g}\otimes\mathbb{C}[t, t^{-1}]$, the global Weyl modules were introduced by Chari and Pressley [2001]. Feigin and Loktev [2004] extended these global Weyl modules to the case where the Laurent polynomials above were replaced by the coordinate ring of a complex affine variety. Chari, Fourier and Khandai [Chari et al. 2010] then generalized this definition to the map algebras, $\mathfrak{g} \otimes A$, where A is a commutative, associative, complex unital algebra. Feigin and Loktev [2004] also gave an isomorphism which explicitly determines the structure of the global Weyl modules for the map algebras of $\mathfrak{sl}_n$ of highest weight $m\omega_1$.

The goal of this work is to use the structure isomorphism given by Feigin and Loktev to give nice bases for the global Weyl modules for the map algebras $\mathfrak{sl}_n \otimes A$ of $\mathfrak{sl}_n$ of highest weight $m\omega_1$. These bases will be given in terms of specific elements of $\boldsymbol{U}(\mathfrak{sl}_n \otimes A)$ acting on the highest weight vector. This was done in [Chamberlin 2011] in the case $n = 2$, but the case $n > 2$ has not previously appeared in the literature.

2. Preliminaries

2.1. ***The structure of $\mathfrak{sl}_n$.*** Recall that $\mathfrak{sl}_n$ is the Lie algebra of all complex traceless matrices. The Lie bracket is the commutator bracket given by $[A, B] = AB - BA$.

MSC2010: 17B10.
Keywords: Lie algebra, module, representation, Weyl.

Given any matrix $[b_{i,j}]$, define $\varepsilon_k([b_{i,j}]) := b_{k,k}$. For $i \in \{1, \dots, n-1\}$, define $\alpha_i := \varepsilon_i - \varepsilon_{i+1}$. Define

$$R^{\pm} := \{\pm(\alpha_i + \cdots + \alpha_j) \mid 1 \le i < j \le n-1\}$$

to be the positive and negative roots respectively, and define $R = R^+ \cup R^-$ to be the set of roots. Let $e_{i,j}$ be the $n \times n$ matrix with a one in the i-th row and j-th column and zeros in every other position. Define $h_i := h_{\alpha_i} = e_{i,i} - e_{i+1,i+1}$, for $i \in \{1, \dots, n-1\}$. Then $\mathfrak{h} := \operatorname{span}\{h_i \mid 1 \le i \le n\}$ is a Cartan subalgebra of $\mathfrak{sl}_n$. Given $\alpha = \alpha_i + \cdots + \alpha_j \in R^+$, define $x_\alpha := e_{i,j}$ and $x_{-\alpha} := e_{j,i}$. Then $\{h_i,\ x_{\pm\alpha} \mid 1 \le i \le n-1,\ \alpha \in R\}$ is a Chevalley basis for $\mathfrak{sl}_n$. Given $i \in \{1, \dots, n-1\}$, define $x_i := x_{\alpha_i} = e_{i,i+1}$ and $x_{-i} := x_{-\alpha_i} = e_{i+1,i}$. Note that, for all $1 \le i \le n-1$, $\operatorname{span}\{x_{-i}, h_i, x_i\} \cong \mathfrak{sl}_2$.

Define nilpotent subsuperalgebras $\mathfrak{n}^{\pm} := \operatorname{span}\{x_\alpha | \alpha \in R^{\pm}\}$ and note that $\mathfrak{sl}_n = \mathfrak{n}^- \oplus \mathfrak{h} \oplus \mathfrak{n}^+$. Define the set of fundamental weights $\{\omega_1, \dots, \omega_{n-1}\} \subset \mathfrak{h}^*$ by $\omega_i(h_j) = \delta_{i,j}$ for all $i, j \in \{1, \dots, n-1\}$. Define $P^+ := \operatorname{span}_{\mathbb{Z}_{\ge 0}}\{\omega_1, \dots, \omega_{n-1}\}$ to be the set of dominant integral weights.

2.2. ***Map algebras and Weyl modules.*** For the remainder of this work fix a commutative, associative, complex unital algebra A. Define the map algebra of $\mathfrak{sl}_n$ to be $\mathfrak{sl}_n \otimes A$ with Lie bracket given by linearly extending the bracket

$$[z \otimes a, w \otimes b] = [z, w] \otimes ab$$

for all $z, w \in \mathfrak{sl}_n$ and $a, b \in A$.

Define $\boldsymbol{U}(\mathfrak{sl}_n \otimes A)$ to be the universal enveloping algebra of $\mathfrak{sl}_n \otimes A$.

As in [Chari et al. 2010] we define the global Weyl model for $\mathfrak{sl}_n \otimes A$ of highest weight $\lambda \in P^+$ to be the module generated by a vector w_λ, called the highest weight vector, with relations

$$(x \otimes a)w_\lambda = 0, \quad (h \otimes 1)w_\lambda = \lambda(h)w_\lambda, \quad (x_{-i} \otimes 1)^{\lambda(h_i)+1}w_\lambda = 0,$$

for all $a \in A,\ x \in \mathfrak{n}^+,\ h \in \mathfrak{h}$, and $1 \le i \le n-1$.

2.3. ***Multisets.*** Given any set S, define a multiset of elements of S to be a multiplicity function $\chi : S \to \mathbb{Z}_{\ge 0}$. Define $\mathcal{F}(S) := \{\chi : S \to \mathbb{Z}_{\ge 0} : |\operatorname{supp} \chi| < \infty\}$. For $\chi \in \mathcal{F}(S)$, define $|\chi| := \sum_{s \in S} \chi(s)$. Notice that $\mathcal{F}(S)$ is an abelian monoid under function addition. For $\psi, \chi \in \mathcal{F}(S)$, we write $\psi \subseteq \chi$ if $\psi(s) \le \chi(s)$ for all $s \in S$. Define $\mathcal{F}(\chi)(S) := \{\psi \in \mathcal{F}(S) \mid \psi \subseteq \chi\}$. In the case $S = A$, the S will be omitted from the notation, so that $\mathcal{F} := \mathcal{F}(A)$ and $\mathcal{F}(\chi) = \mathcal{F}(\chi)(A)$.

If $\psi, \chi \in \mathcal{F}$ with $\psi \in \mathcal{F}(\chi)$ we define $\chi - \psi$ by standard function subtraction.

Also define $\pi : \mathcal{F} - \{0\} \to A$ by

$$\pi(\psi) := \prod_{a\in A} a^{\psi(a)},$$

and extend π to $\mathcal{F}$ by setting $\pi(0) = 1$. Define $\mathcal{M} : \mathcal{F} \to \mathbb{Z}$ by

$$\mathcal{M}(\psi) := \frac{|\psi|!}{\prod_{a\in A} \psi(a)!}.$$

$\mathcal{M}(\psi) \in \mathbb{Z}$ for all $\psi \in \mathcal{F}$ because if $\operatorname{supp}\psi = \{a_1, \dots, a_k\}$ then $\mathcal{M}(\psi)$ is the multinomial coefficient

$$\binom{|\psi|}{\psi(a_1), \dots, \psi(a_k)}.$$

For $s \in S$, define χ_s to be the characteristic function of the set $\{s\}$. Then, for all $\chi \in \mathcal{F}(S)$,

$$\chi = \sum_{s\in S} \chi(s)\chi_s.$$

2.4. ***The symmetric tensor space.*** Given any vector space W, there is an action of the symmetric group S_k on $W^{\otimes k} = W \otimes W \otimes \cdots \otimes W$ (k times) given by

$$\sigma(w_1\otimes w_2\otimes\cdots\otimes w_k) = w_{\sigma^{-1}(1)}\otimes w_{\sigma^{-1}(2)}\otimes\cdots\otimes w_{\sigma^{-1}(k)}, \quad \text{where } w_1, \dots, w_k \in W.$$

For any vector space W, define its k-th symmetric tensor space

$$S^k(W) = \operatorname{span}\left\{ \sum_{\sigma\in S_k} \sigma(w_1 \otimes \cdots \otimes w_k) \;\middle|\; w_1, \dots, w_k \in W \right\}.$$

Define $V \cong \mathbb{C}^n$ to be an $\mathfrak{sl}_n$-module via left matrix multiplication and write the basis as $v_1 := (1, 0, \dots, 0)$, and for $i \in \{1, \dots, n+m-1\}$, set $v_{i+1} := x_{-i}v_i$. Then $V \otimes A$ is an $\mathfrak{sl}_n \otimes A$-module under the action $(z \otimes a)(w \otimes b) = zw \otimes ab$.

Given $\varphi_1, \dots, \varphi_n \in \mathcal{F}$ with $k := \sum_{i=1}^n |\varphi_i|$, define

$$w(\varphi_1, \dots, \varphi_n) := \bigotimes_{a_1\in\operatorname{supp}\varphi_1} (v_1\otimes a_1)^{\otimes\varphi_1(a_1)} \otimes\cdots\otimes \bigotimes_{a_n\in\operatorname{supp}\varphi_n} (v_n\otimes a_n)^{\otimes\varphi_n(a_n)} \in (V\otimes A)^{\otimes k}$$

and

$$v(\varphi_1, \dots, \varphi_n) := \sum_{\sigma\in S_k} \sigma\big(w(\varphi_1, \dots, \varphi_n)\big) \in S^k(V \otimes A).$$

We will need the following theorem:

Theorem 2.4.1 [Feigin and Loktev 2004, Theorem 6]. *For all $m \in \mathbb{N}$, $W_A(m\omega_1) \cong S^m(V \otimes A)$ via the map given by*

$$w_{m\omega_1} \mapsto (v_1 \otimes 1)^{\otimes m}.$$

We will also need the following lemma:

Lemma 2.4.2. *Let $\boldsymbol{B}$ be a basis for A. Then the set*

$$\mathfrak{B} := \left\{ v(\varphi_1, \dots, \varphi_n) \;\middle|\; \varphi_1, \dots, \varphi_n \in \mathcal{F}(\boldsymbol{B}),\ \sum_{i=1}^{n} |\varphi_i| = m \right\}$$

is a basis for $S^m(V \otimes A)$.

Proof. $\mathfrak{B}$ spans $S^m(V \otimes A)$ because $\boldsymbol{B}$ spans A and $v_1, \dots, v_n$ span V. $\mathfrak{B}$ is linearly independent because the set

$$\big\{ (v_{j_1} \otimes b_1) \otimes \cdots \otimes (v_{j_m} \otimes b_m) \;\big|\; j_1, \dots, j_m \in \{1, \dots, n\},\ b_1, \dots, b_m \in \boldsymbol{B} \big\}$$

is a basis for $(V \otimes A)^{\otimes m}$ and hence is linearly independent. □

Given $k \in \mathbb{N}$, define $\Delta^{k-1} : \boldsymbol{U}(\mathfrak{sl}_n \otimes A) \to \boldsymbol{U}(\mathfrak{sl}_n \otimes A)^{\otimes k}$ by extending the map $\mathfrak{sl}_n \otimes A \to \boldsymbol{U}(\mathfrak{sl}_n \otimes A)^{\otimes k}$ given by

$$\Delta^{k-1}(z \otimes a) = \sum_{j=0}^{k-1} 1^{\otimes j} \otimes (z \otimes a) \otimes 1^{\otimes k-1-j}.$$

Note that $\Delta^{k-1}(1) = 1^{\otimes k}$, not $k1^{\otimes k}$.

Since $V \otimes A$ is a $\boldsymbol{U}(\mathfrak{sl}_n \otimes A)$-module, $(V \otimes A)^{\otimes m}$ is a left $\boldsymbol{U}(\mathfrak{sl}_n \otimes A)$-module with u acting as $\Delta^{m-1}(u)$ followed by coordinatewise module actions. Moreover $S^m(V \otimes A)$ is a submodule under this action. Thus $S^m(V \otimes A)$ is a left $\boldsymbol{U}(\mathfrak{sl}_n \otimes A)$-module under this Δ^{m-1} action.

2.5. For all $i = 1, \dots, n-1$ and $\chi, \varphi \in \mathcal{F}$, recursively define $q_i(\varphi, \chi) \in \boldsymbol{U}(\mathfrak{sl}_n \otimes A)$ as follows:

$$\begin{aligned}
q_i(0,0) &:= 1,\\
q_i(0,\chi) &:= -\frac{1}{|\chi|} \sum_{0 \neq \psi \in \mathcal{F}(\chi)} \mathcal{M}(\psi)\big(h_i \otimes \pi(\psi)\big) q_i(0, \chi - \psi),\\
q_i(\varphi,\chi) &:= -\frac{1}{|\varphi|} \sum_{\psi \in \mathcal{F}(\chi)} \sum_{d \in \operatorname{supp} \varphi} \mathcal{M}(\psi)\big(x_{-i} \otimes d\pi(\psi)\big) q_i(\varphi - \chi_d, \chi - \psi).
\end{aligned}$$

Given $\varphi_n, \dots, \varphi_n \in \mathcal{F}$, define

$$\begin{aligned}
q(\varphi_1, \dots, \varphi_n) := q_{n-1}(\varphi_n, \varphi_{n-1}) q_{n-2}\big((|\varphi_n| + |\varphi_{n-1}|)\chi_1, \varphi_{n-2}\big)&\\
\times \cdots \times q_2\bigg(\Big(\sum_{j=3}^{n} |\varphi_j|\Big)\chi_1, \varphi_2\bigg) q_1\bigg(\Big(\sum_{k=2}^{n} |\varphi_j|\Big)\chi_1, \varphi_1\bigg)&.
\end{aligned}$$

Remark. The $q_i(0, \chi)$ coincide with the $p_i(\chi)$ defined in [Bagci and Chamberlin 2014].

3. Main theorem

The main result of this work is the theorem stated below.

Theorem 3.0.1. *Given a basis $\boldsymbol{B}$ for A and $m \in \mathbb{Z}_{>0}$, the set*

$$\left\{ q(\varphi_1, \dots, \varphi_n) w_{m\omega_1} \;\middle|\; \varphi_1, \dots, \varphi_n \in \mathcal{F}(\boldsymbol{B}),\ \sum_{i=1}^{n} |\varphi_i| = m \right\}$$

is a basis for $W_A(m\omega_1)$.

The proof of this theorem will be given after several lemmas and propositions.

3.1. *Necessary lemmas and propositions.*

Proposition 3.1.1. *For all $k \in \mathbb{N}$ $\Delta^k = (1^{\otimes k-1} \otimes \Delta^1) \circ \Delta^{k-1}$.*

Proof. The case $k = 1$ is trivial. For $k \geq 2$ and $u \in \boldsymbol{U}(\mathfrak{sl}_n \otimes A)$ we have

$$\begin{aligned}
(1^{\otimes k-1} \otimes \Delta^1)\big(\Delta^{k-1}(u)\big) &= (1^{\otimes k-1} \otimes \Delta^1)\bigg(\sum_{j=0}^{k-1} 1^{\otimes j} \otimes u \otimes 1^{\otimes k-1-j} \bigg) \\
&= (1^{\otimes k-1} \otimes \Delta^1)\bigg(\sum_{j=0}^{k-2} 1^{\otimes j} \otimes u \otimes 1^{\otimes k-1-j} + 1^{\otimes k-1} \otimes u \bigg) \\
&= \sum_{j=0}^{k-2} 1^{\otimes j} \otimes u \otimes 1^{\otimes k-2-j} \otimes \Delta^1(1) + 1^{\otimes k-1} \otimes \Delta^1(u) \\
&= \sum_{j=0}^{k-2} 1^{\otimes j} \otimes u \otimes 1^{\otimes k-2-j} \otimes 1 \otimes 1 + 1^{\otimes k-1} \otimes (u \otimes 1 + 1 \otimes u) \\
&= \sum_{j=0}^{k-2} 1^{\otimes j} \otimes u \otimes 1^{\otimes k-j} + 1^{\otimes k-1} \otimes u \otimes 1 + 1^{\otimes k-1} \otimes 1 \otimes u \\
&= \sum_{j=0}^{k} 1^{\otimes j} \otimes u \otimes 1^{\otimes k-j} = \Delta^k(u).
\end{aligned}$$

□

Given $\chi \in \mathcal{F}$ and $k \in \mathbb{N}$, define

$$\circ_k(\chi) = \left\{ \psi : \{1, \dots, k\} \to \mathcal{F}(\chi) \;\middle|\; \sum_{j=1}^{k} \psi(j) = \chi \right\}.$$

Lemma 3.1.2. *For all $i \in \{1, \dots, n-1\}$,*

$$\Delta^{k-1}\big(q_i(\varphi, \chi)\big) = \sum_{\substack{\psi \in \circ_k(\chi) \\ \phi \in \circ_k(\varphi)}} q_i\big(\phi(1), \psi(1)\big) \otimes \cdots \otimes q_i\big(\phi(k), \psi(k)\big)$$

Proof. This can be proven by induction on k. The case $k=1$ is trivial. In the case $k=2$ the lemma becomes

$$\Delta^1\big(q_i(\varphi,\chi)\big)=\sum_{\substack{\psi\in\mathcal{F}(\chi)\\ \phi\in\mathcal{F}(\varphi)}} q_i(\phi,\psi)\otimes q_i(\varphi-\phi,\chi-\psi).$$

This can be proven by induction on $|\varphi|$. For $k>2$ use Proposition 3.1.1. The details in the $\mathfrak{sl}_2$ case can be found in [Chamberlin 2011]. This can be extended to the $\mathfrak{sl}_n$ case via the injection $\Omega_i:\mathfrak{sl}_2\otimes A\to\mathfrak{sl}_n\otimes A$ given by

$$\Omega_i(x^-\otimes a)=x_{-i}\otimes a,\quad \Omega_i(h\otimes a)=h_i\otimes a,\quad \Omega_i(x^+\otimes a)=x_i\otimes a,$$

for all $i\in\{1,\dots,n-1\}$ and $a\in A$. □

Lemma 3.1.3. *For all $\varphi,\chi\in\mathcal{F}$ with $|\varphi|+|\chi|>1$ and all $i\in\{1,\dots,n-1\}$, $q_i(\varphi,\chi)(v_i\otimes 1)=0$.*

Proof. Assume that $\varphi=0$. This case will proceed by induction on $|\chi|>1$. If $|\chi|=2$ (so that $\chi=\{a,b\}$ for some $a,b\in A$) we have

$$\begin{aligned}q_i(0,\{a,b\})(v_i\otimes 1)&=[(h_i\otimes a)\otimes(h_i\otimes b)-(h_i\otimes ab)](v_i\otimes 1)\\&=(h_i\otimes a)\otimes(v_i\otimes b)-(v_i\otimes ab)\\&=(v_i\otimes ab)-(v_i\otimes ab)\\&=0.\end{aligned}$$

For the next case assume that $|\chi|>2$ then

$$q_i(0,\chi)(v_i\otimes 1)=-\frac{1}{|\chi|}\sum_{\varnothing\neq\psi\in\mathcal{F}(\chi)}\mathcal{M}(\psi)\big(h_i\otimes\pi(\psi)\big)q_i(\chi-\psi)(v_i\otimes 1)=0$$

by induction. Now assume that $|\varphi|=1$ (or $\varphi=\chi_b$ for some $b\in A$). Then

$$\begin{aligned}&q_i(\chi_b,\chi)(v_i\otimes 1)\\&\quad=-\sum_{\psi\in\mathcal{F}(\chi)}\mathcal{M}(\psi)\big(x_{-i}\otimes b\pi(\psi)\big)q_i(0,\chi-\psi)(v_i\otimes 1)\\&\quad=-\mathcal{M}(\chi)\big(x_{-i}\otimes b\pi(\chi)\big)(v_i\otimes 1)\\&\qquad\qquad-\sum_{a\in\operatorname{supp}\chi}\mathcal{M}(\chi-\chi_a)\big(x_{-i}\otimes b\pi(\chi-\chi_a)\big)q_i(0,\chi_a)(v_i\otimes 1)\\&\quad=-\mathcal{M}(\chi)\big(v_{i+1}\otimes b\pi(\chi)\big)\\&\qquad\qquad-\sum_{a\in\operatorname{supp}\chi}\mathcal{M}(\chi-\chi_a)\big(x_{-i}\otimes b\pi(\chi-\chi_a)\big)(-h_i\otimes a)(v_i\otimes 1)\\&\quad=-\mathcal{M}(\chi)\big(v_{i+1}\otimes b\pi(\chi)\big)+\sum_{a\in\operatorname{supp}\chi}\mathcal{M}(\chi-\chi_a)\big(x_{-i}\otimes b\pi(\chi-\chi_a)\big)(v_i\otimes a)\end{aligned}$$

$$
\begin{aligned}
&= -\mathcal{M}(\chi)\big(v_{i+1}\otimes b\pi(\chi)\big) + \sum_{a\in\operatorname{supp}\chi} \mathcal{M}(\chi-\chi_a)\big(v_{i+1}\otimes b\pi(\chi)\big)\\
&= -\mathcal{M}(\chi)\big(v_{i+1}\otimes b\pi(\chi)\big) + \sum_{a\in\operatorname{supp}\chi} \frac{(|\chi|-1)!}{\prod\limits_{c\in\operatorname{supp}(\chi-\chi_a)}(\chi-\chi_a)(c)!}\big(v_{i+1}\otimes b\pi(\chi)\big)\\
&= -\mathcal{M}(\chi)\big(v_{i+1}\otimes b\pi(\chi)\big) + \sum_{a\in\operatorname{supp}\chi} \frac{(|\chi|-1)!}{\prod\limits_{\substack{c\in\operatorname{supp}\chi\\ c\neq a}}\chi(c)!\big(\chi(a)-1\big)!}\big(v_{i+1}\otimes b\pi(\chi)\big)\\
&= -\mathcal{M}(\chi)\big(v_{i+1}\otimes b\pi(\chi)\big) + \sum_{a\in\operatorname{supp}\chi} \frac{\chi(a)}{|\chi|}\mathcal{M}(\chi)\big(v_{i+1}\otimes b\pi(\chi)\big)\\
&= -\mathcal{M}(\chi)\big(v_{i+1}\otimes b\pi(\chi)\big) + \mathcal{M}(\chi)\big(v_{i+1}\otimes b\pi(\chi)\big)\\
&= 0.
\end{aligned}
$$

Finally assume that $|\varphi| > 1$. Then

$$
\begin{aligned}
&q_i(\varphi,\chi)(v_i\otimes 1)\\
&\quad = -\frac{1}{|\varphi|}\sum_{\psi\in\mathcal{F}(\chi)}\sum_{d\in\operatorname{supp}\varphi} \mathcal{M}(\psi)\big(x_{-i}\otimes d\pi(\psi)\big)q_i(\varphi-\chi_d,\chi-\psi)(v_i\otimes 1)\\
&\quad = -\frac{1}{|\varphi|}\sum_{\psi\in\mathcal{F}(\chi)}\sum_{d\in\operatorname{supp}\varphi} \mathcal{M}(\psi)\\
&\qquad \Bigg(-\frac{1}{|\varphi|-1}\sum_{\psi_1\in\mathcal{F}(\chi-\psi)}\sum_{d_1\in\operatorname{supp}(\varphi-\chi_d)} \mathcal{M}(\psi_1)\\
&\qquad\quad \big(x_{-i}\otimes d\pi(\psi)\big)\big(x_{-i}\otimes d_1\pi(\psi_1)\big)q_i(\varphi-\chi_d-\chi_{d_1},\chi-\psi-\psi_1)\Bigg)(v_i\otimes 1)\\
&\quad = 0,
\end{aligned}
$$

because at least two x_{-i} terms act on a single v_i as 0. □

Lemma 3.1.4. *For all $i\in\{1,\dots,n-1\}$ and $\varphi,\chi\in\mathcal{F}$ with $|\varphi|+|\chi| = k$ we have*

$$
q_i(\varphi,\chi)(v_i\otimes 1)^{\otimes k} = (-1)^k v(0,\dots,0,\chi,\varphi,0,\dots,0),
$$

where χ is in the i-th position and φ in the $(i+1)$-st.

Proof. We have $q_i(\varphi,\chi)(v_i\otimes 1)^{\otimes k}i = \Delta^{k-1}\big(q_i(\varphi,\chi)\big)(v_i\otimes 1)^{\otimes k}$. By Lemma 3.1.2, this equals

$$
\Bigg(\sum_{\substack{\psi\in\circ_k(\chi)\\ \phi\in\circ_k(\varphi)}} q_i\big(\phi(1),\psi(1)\big)\otimes\cdots\otimes q_i\big(\phi(k),\psi(k)\big)\Bigg)(v_i\otimes 1)^{\otimes k},
$$

which can be rewritten as

$$\sum_{\substack{\psi\in\circ_k(\chi)\\ \phi\in\circ_k(\varphi)}} \big(q_i(\phi(1),\psi(1))(v_i\otimes 1)\big)\otimes\cdots\otimes\big(q_i(\phi(k),\psi(k))(v_i\otimes 1)\big).$$

By Lemma 3.1.3 we see that the only potentially nonzero terms in the sum are those for which $|\phi(j)|+|\psi(j)|\le 1$ for all $j\in\{1,\dots,k\}$. Since $|\varphi|+|\chi|=k$ if we have $|\psi(j)|+|\phi(j)|=0$ for some $j\in\{1,\dots,n-1\}$, then there is a $r\in\{1,\dots,n-1\}$ such that $|\psi(r)|+|\phi(r)|>1$. So the only potentially nonzero terms in the sum are those for which $|\phi(j)|+|\psi(j)|=1$ for all $j\in\{1,\dots,k\}$. Suppose that $\phi(j)=\chi_a$ and $\psi(j)=0$ for some $j\in\{1,\dots k\}$ and some $a\in A$. Then

$$q_i(\chi_a,0)(v_i\otimes 1)=-(x_{-i}\otimes a)(v_i\otimes 1)=-(v_{i+1}\otimes a).$$

Suppose that $\phi(j)=0$ and $\psi(j)=\chi_a$ for some $j\in\{1,\dots k\}$ and some $a\in A$. Then

$$q_i(0,\chi_a)(v_i\otimes 1)=-(h_i\otimes a)(v_i\otimes 1)=-(v_i\otimes a).$$

So $-(v_{i+1}\otimes a)$ and $-(v_i\otimes a)$ are the only possibilities for factors in the tensor product above. Since we are summing over all possible submultisets of φ and χ, we have the result. □

Lemma 3.1.5. *For all $m\in\mathbb{N}$ and all $\varphi_1,\dots,\varphi_n\in\mathcal{F}$ with $\sum_{i=1}^n|\varphi_i|=m$,*

$$q(\varphi_1,\dots,\varphi_n)(v_1\otimes 1)^{\otimes m}=(-1)^{\sum_{j=1}^n j|\varphi_j|}v(\varphi_1,\dots,\varphi_n).$$

Proof. Since for all $j\in\{1,\dots,n-1\}$ and $k\in\{1,\dots,n\}$,

$$x_{-j}v_k=\delta_{j,k}v_{j+1},\qquad h_jv_k=\delta_{j,k}v_j-\delta_{j+1,k}v_{j+1},$$

so by Lemma 3.1.4 we have

$$\begin{aligned}
&q(\varphi_1,\dots,\varphi_n)(v_1\otimes 1)^{\otimes m}\\
&\quad=q_{n-1}(\varphi_n,\varphi_{n-1})q_{n-2}\big((|\varphi_n|+|\varphi_{n-1}|)\chi_1,\varphi_{n-2}\big)\\
&\qquad\qquad\times\cdots\times q_1\Big(\Big(\sum_{j=2}^n|\varphi_j|\Big)\chi_1,\varphi_1\Big)(v_1\otimes 1)^{\otimes m}\\
&\quad=(-1)^m q_{n-1}(\varphi_n,\varphi_{n-1})\dots q_2\Big(\Big(\sum_{j=3}^n|\varphi_j|\Big)\chi_1,\varphi_2\Big)v\Big(\varphi_1,\Big(\sum_{j=2}^n|\varphi_j|\Big)\chi_1,0,\dots,0\Big)\\
&\quad=(-1)^{|\varphi_1|+2\sum_{j=2}^n|\varphi_j|}q_{n-1}(\varphi_n,\varphi_{n-1})\\
&\qquad\qquad\times\cdots\times q_3\Big(\Big(\sum_{j=4}^n|\varphi_j|\Big)\chi_1,\varphi_3\Big)vi\Big(\varphi_1,\varphi_2,\Big(\sum_{j=3}^n|\varphi_j|\Big)\chi_1,0,\dots,0\Big)
\end{aligned}$$

$$= (-1)^{\sum_{j=1}^{n-2} j|\varphi_j|} q_{n-1}(\varphi_n, \varphi_{n-1}) v\big(\varphi_1, \dots, \varphi_{n-2}, (|\varphi_{n-1}| + |\varphi_n|)\chi_1, 0\big)$$
$$= (-1)^{\sum_{j=1}^{n} j|\varphi_j|} v(\varphi_1, \dots, \varphi_n). \qquad \square$$

3.2. *The proof of Theorem 3.0.1.*

Proof. By Lemmas 3.1.5 and 2.4.2

$$\left\{ q(\varphi_1, \dots, \varphi_n)(v_1 \otimes 1)^{\otimes m} \;\middle|\; \varphi_1, \dots, \varphi_n \in \mathcal{F}(\boldsymbol{B}),\ \sum_{i=1}^{n} |\varphi_i| = m \right\}$$

is a basis for $S^m(V \otimes A)$. Therefore by Theorem 2.4.1

$$\left\{ q(\varphi_1, \dots, \varphi_n) w_{m\omega_1} \;\middle|\; \varphi_1, \dots, \varphi_n \in \mathcal{F}(\boldsymbol{B}),\ \sum_{i=1}^{n} |\varphi_i| = m \right\}$$

is a basis for $W_A(m\omega_1)$. □

References

[Bagci and Chamberlin 2014] I. Bagci and S. Chamberlin, "Integral bases for the universal enveloping algebras of map superalgebras", *J. Pure Appl. Algebra* **218**:8 (2014), 1563–1576. MR Zbl

[Chamberlin 2011] S. H. Chamberlin, *Integral bases for the universal enveloping algebras of map algebras*, Ph.D. thesis, University of California, Riverside, 2011, http://search.proquest.com/docview/883977727. MR

[Chari and Pressley 2001] V. Chari and A. Pressley, "Weyl modules for classical and quantum affine algebras", *Represent. Theory* **5** (2001), 191–223. MR Zbl

[Chari et al. 2010] V. Chari, G. Fourier, and T. Khandai, "A categorical approach to Weyl modules", *Transform. Groups* **15**:3 (2010), 517–549. MR Zbl

[Feigin and Loktev 2004] B. Feigin and S. Loktev, "Multi-dimensional Weyl modules and symmetric functions", *Comm. Math. Phys.* **251**:3 (2004), 427–445. MR Zbl

Received: 2015-06-29 Accepted: 2015-08-25

samuel.chamberlin@park.edu
Department of Information Systems, Computer Science and Mathematics, Park University, 8700 NW River Park Drive #30, Parkville, 64152, United States

amanda.croan@park.edu
Department of Information Systems, Computer Science and Mathematics, Park University, 8700 NW River Park Drive #30, Parkville, 64152, United States

INVOLVE 10:4 (2017)
dx.doi.org/10.2140/involve.2017.10.583

Leverage centrality of knight's graphs and Cartesian products of regular graphs and path powers

Roger Vargas, Jr., Abigail Waldron, Anika Sharma,
Rigoberto Flórez and Darren A. Narayan

(Communicated by Kenneth S. Berenhaut)

In 2010, Joyce et al. defined the leverage centrality of a graph as a means to analyze connections within the brain. In this paper we investigate this property from a mathematical perspective and determine the leverage centrality for knight's graphs, path powers, and Cartesian products.

1. Introduction

We recall that the degree of a vertex v is the number of edges incident to v and is denoted $\deg v$. Joyce, Laurienti, Burdette, and Hayasaka [Joyce et al. 2010] defined the property of leverage centrality based on vertex degrees.

Definition 1. Leverage centrality is a measure of the relationship between the degree of a given node v and the degree of each of its neighbors v_i, averaged over all neighbors of v, denoted N_v, and is defined as

$$l(v) = \frac{1}{\deg v} \sum_{v_i \in N_v} \frac{\deg v - \deg v_i}{\deg v + \deg v_i}.$$

This property was used by Joyce et al. [2010] in the analysis of functional magnetic resonance imaging (fMRI) data and has been used to analyze real-world networks including airline connections, electrical power grids, and coauthorship collaborations [Li et al. 2015]. The leverage centralities of complete multipartite graphs and the Cartesian product of paths were investigated by Sharma, Vargas, Waldron, Flórez, and Narayan [Sharma et al. 2017]. Bounds on leverage centrality were determined by Li, Li, Van Mieghem, Stanley, and Wang [Li et al. 2015]. We restate one of their results as our first theorem.

MSC2010: 05C07.
Keywords: leverage centrality, knight's graphs.

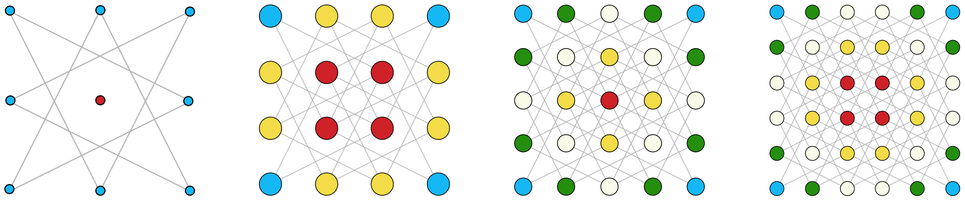


Figure 1. The 3×3, 4×4, 5×5 and 6×6 knight's graphs.

Theorem 2. *For any vertex v, we have $|l(v)| \le 1-\frac{2}{n}$. Furthermore, these bounds are tight in the cases of stars and complete graphs.*

We note that the bounds are also tight for regular graphs with degree $r > 1$.

In this paper we investigate leverage centrality for various families of graphs including the knight's graphs, path powers, and the Cartesian products of graphs.

2. Leverage centrality of a knight's graph

We define an $n \times n$ *knight's graph* to be the graph with n^2 vertices in which every vertex represents a square in an $n \times n$ chessboard. The vertices on the $n \times n$ chessboard can be placed in an $n \times n$ table where two vertices v_i and v_j are adjacent if they are exactly four entries apart (including the entries of v_i and v_j) and they form an "L" shape. We give examples of knight's graphs of small order in Figure 1, where in each graph all of the vertices of same degree are the same color.

We next state the leverage centrality of each vertex in the $n \times n$ knight's graph. We use t_j to denote the j-th triangular number $\frac{1}{2}j(j+1)$.

Theorem 3. *Let G_n be the $n \times n$* knight's graph.

(1) *The leverage centrality of every vertex of G_3 is zero.*

(2) *If $n = 4, 6,$ or 8, then G_n has exactly $t_{n/2}$ distinct leverage centralities.*

(3) *If $n = 5$ or 7, then G_n has exactly $t_{(n+1)/2} - 1$ distinct leverage centralities.*

(4) *If $n \ge 9$, then G_n has exactly 15 distinct leverage centralities.*

Proof. We first find the degree of each vertex in the knight's graph on an $n \times n$ chessboard, where $n \ge 3$. To describe the degree of each vertex in the graph G_n, we will arrange the vertices of G_n in an $n\times n$ table. The vertices corresponding to entries $(1, 1)$, $(1, n)$, $(n, 1)$, and (n, n) have degree 2. Those corresponding to entries $(1, 2)$, $(1, n-1)$, $(2, 1)$, $(2, n)$, $(n-1, 1)$, $(n, 2)$, $(n-1, n)$, and $(n, n-1)$ have degree 3. Those corresponding to entries $(2, 2)$, $(2, n-1)$, $(n-1, 2)$, $(n-1, n-1)$ and $(1, i)$, $(i, 1)$, (n, i), and (i, n), where $i = 3, 4, \dots, n-2$, have degree 4. Those corresponding to entries $(2, i)$, $(i, 2)$, $(n-1, i)$, and $(i, n-1)$, where $i = 3, 4, \dots, n-2$, have degree 6. Vertices corresponding to entries (i, j), where $i = 3, 4, \dots, n-3$ and $j = 3, 4, \dots, n-2$, have degree 8; see, for example, Figure 2 (left).

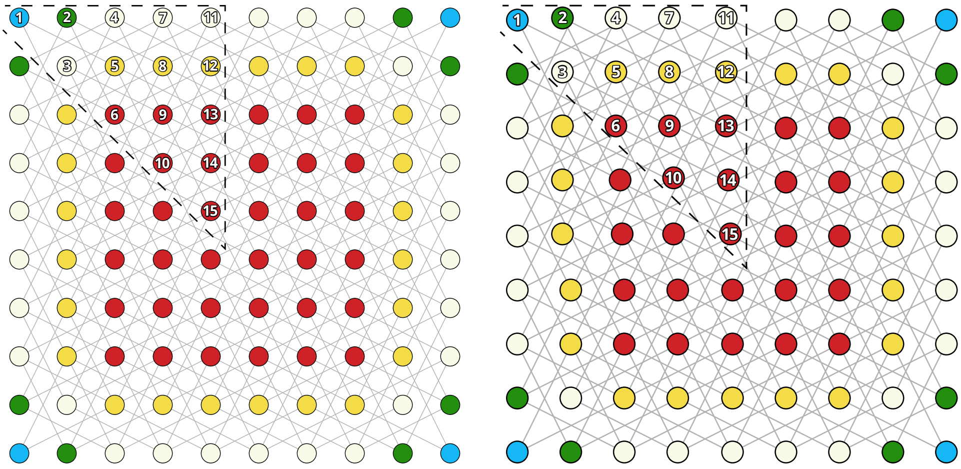


Figure 2. The 10 × 10 knight's graph (left) and the 9 × 9 knight's graph (right).

If n is even we subdivide the knight's graph's vertical and horizontal axes and the two diagonals to obtain eight regions. Each region forms a right triangle where the legs have $\frac{1}{2}n$ vertices; see, for example, Figure 2 (left). Using symmetry we can calculate the leverage centrality of all vertices by only analyzing a single triangle.

If n is odd, as in Figure 2 (right), we do the same subdivision; however, in this case two adjacent triangles will overlap — the legs of the right triangle will have $\frac{1}{2}(n+1)$ vertices.

We choose the triangle with vertices

$$\begin{array}{lllll} v_1 = (1,1), & v_2 = (1,2), & v_3 = (2,2), & v_4 = (1,3), & v_5 = (2,3), \\ v_6 = (3,3), & v_7 = (1,4), & v_8 = (2,4), & v_9 = (3,4), & v_{10} = (4,4), \\ v_{11} = (1,5), & v_{12} = (2,5), & v_{13} = (3,5), & v_{14} = (4,5), & v_{15} = (5,5). \end{array} \tag{2-1}$$

Note that if $n < 10$, we take triangles with vertices v_i for $i = 1, 2, \ldots, k$, where $k = \frac{1}{2}n$ if $n = 2(k)$ or $k = \frac{1}{2}(n+1)$ if $n = 2(k) - 1$.

Proof of (1). Since G_3 is regular, the leverage centrality of all of its vertices is 0.

Proof of (2). For case $n = 4$, it is easy to see that $l(v_1) = -\frac{1}{3}$, $l(v_2) = -\frac{1}{21}$, and $l(v_3) = \frac{5}{21}$.

Now consider the cases $n = 6$ and $n = 8$. From the above analysis we only need to calculate the leverage centrality for a triangle with legs that have $\frac{1}{2}n$ vertices (see Figure 2 (left) for an example of those triangles). Thus, to calculate the leverage centrality of these special cases, we consider the triangle with vertices $v_1, \ldots, v_{t_i}$, where t_i is the i-th triangular number where $i = 1, 2, \ldots, \frac{1}{2}n$, and then use Tables 1 and 3, respectively.

vertex v_i	deg v_i	AD(v_i)	$l(v_i)$
v_1	2	6, 6	−1/2
v_2	3	4, 6, 8	−215/693
v_3	4	4, 4, 8, 8	−1/6
v_4	4	3, 6, 4, 8	−41/420
v_5	6	2, 3, 4, 6, 6, 8	187/1260
v_6	8	3, 3, 4, 4, 4, 4, 6, 6	73/231

Table 1. Leverage centrality when $n = 6$. Here AD(v_i) denotes the degrees of vertices adjacent to v_i.

vertex v_i	deg v_i	AD(v_i)	$l(v_i)$
v_1	2	6, 6	−1/2
v_2	3	4, 6, 8	−215/693
v_3	4	4, 4, 8, 8	−1/6
v_4	4	3, 6, 6, 8	−31/210
v_5	6	2, 4, 4, 6, 8, 8	43/420
v_6	8	3, 3, 4, 4, 6, 6, 8, 8	215/924
v_7	4	4, 4, 8, 8	−1/6
v_8	6	3, 3, 6, 6, 8, 8	4/63
v_9	8	4, 4, 4, 4, 6, 6, 8, 8	17/84
v_{10}	8	6, 6, 6, 6, 6, 6, 6, 6	1/7

Table 2. Leverage centrality when $n = 7$. Here AD(v_i) denotes the degrees of vertices adjacent to v_i.

Proof of (3). First consider the case $n = 5$. From the above analysis we only need to calculate the leverage centrality for a triangle with legs that have three vertices. It is easy to see that $l(v_1) = -\frac{1}{2}$, $l(v_2) = -\frac{19}{77}$, $l(v_3) = -\frac{1}{35}$, $l(v_4) = -\frac{1}{35}$, $l(v_5) = \frac{3}{10}$, and $l(v_6) = \frac{5}{11}$. This shows that there are only five distinct leverage centralities in G_5.

Now consider the case $n = 7$. From the above analysis we only need to calculate the leverage centrality for a triangle with legs that have four vertices. From Table 2 we can see that G_7 has only nine distinct leverage centralities.

Proof of (4). For $n = 8$ and $n = 10$, the proof is similar to those of parts (2) and (3).

We now suppose $n > 10$. Consider the 15 vertices in the triangle given in (2-1) and their relevant data, given in Table 3.

The analysis for the remaining vertices in the triangle is as follows. From the definition of the knight's graph we know that if two vertices v_i and v_j are adjacent, then they are four entries apart (including the entries of v_i and v_j) and they form an "L" shape. This implies that if $n \geq 11$ then the leverage centrality of every vertex located

n	vertex v_i	deg v_i	AD(v_i)	$l(v_i)$
$8, 9, 10$	v_1	2	$6, 6$	$-1/2$
$8, 9, 10$	v_2	3	$4, 6, 8$	$-215/693$
$8, 9, 10$	v_3	4	$4, 4, 8, 8$	$-1/6$
$8, 9, 10$	v_4	4	$3, 6, 6, 8$	$-31/210$
$8, 9, 10$	v_5	6	$2, 4, 4, 6, 8, 8$	$43/420$
$8, 9, 10$	v_6	8	$3, 3, 4, 4, 6, 6, 8, 8$	$215/924$
$8, 9, 10$	v_7	4	$4, 6, 8, 8$	$-13/60$
$8, 9, 10$	v_8	6	$3, 4, 6, 8, 8, 8$	$11/630$
$8, 9, 10$	v_9	8	$4, 4, 4, 6, 6, 8, 8, 8$	$9/56$
$8, 9, 10$	v_{10}	8	$6, 6, 6, 6, 8, 8, 8, 8$	$1/14$
$9, 10$	v_{11}	4	$6, 6, 8, 8$	$-4/45$
$9, 10$	v_{12}	6	$4, 4, 8, 8, 8, 8$	$-1/35$
$9, 10$	v_{13}	8	$4, 4, 6, 6, 8, 8, 8, 8$	$5/42$
$9, 10$	v_{14}	8	$6, 6, 8, 8, 8, 8, 8, 8$	$1/28$
$9, 10$	v_{15}	8	$8, 8, 8, 8, 8, 8, 8, 8$	0

Table 3. Leverage centrality with $n = 8, 9, 10$. Here AD(v_i) denotes the degrees of vertices adjacent to v_i.

in entries (r, t) is zero for $r = 5, 6, \dots, k$ and $t = 5, 6, \dots, k$, where $k = \frac{1}{2}n$ if $n = 2k$ or $k = \frac{1}{2}(n+1)$ if $n = 2k-1$. Moreover, every vertex located in position (i, j) will have the same leverage centrality as the vertices located in entries $(i, 6)$, where $i = 1, 2, \dots, k$ and $j = 7, \dots, k$, where $k = \frac{1}{2}n$ if $n = 2k$ or $k = \frac{1}{2}(n+1)$ if $n = 2k-1$. □

3. Leverage centralities of P_n^k

Let P_n^k be the graph with vertices $v_1, v_2, \dots, v_n$ and edges (v_i, v_j) when $1 \le |i-j| \le k \le n-1$. In this paper we will assume $n > 1$. This family contains both paths (when $k = 1$) and complete graphs (when $k = n-1$). Note that $\deg v_i = \min\{i+k-1, 2k\}$. The neighbors of v_i are $v_{i-1}, v_{i-2}, \dots, v_{i-s}$ and $v_{i+1}, v_{i+2}, \dots, v_{i+t}$, where $s = \min\{k, i-1\}$ and $t = \min\{k, n-i\}$. The above conditions can be combined in the next lemma to give the leverage centrality of any vertex in P_n^k.

Lemma 4. *Suppose the vertex $v_i \in V(P_n^k)$ has neighbors $v_{i-1}, v_{i-2}, \dots, v_{i-s}$ and $v_{i+1}, v_{i+2}, \dots, v_{i+t}$, where $s = \min\{k, i-1\}$ and $t = \min\{k, n-i\}$. Then*

$$l(v_i) = \frac{1}{\delta_i} \sum_{i-s \le j \le i+t} \frac{\delta_i - \delta_j}{\delta_i + \delta_j},$$

where $\delta_x = \min\{x+k-1, 2k\}$ for $x = i, j$.

We begin by determining the leverage centrality of vertices in a path P_n, where $n \ge 2$. We note that by symmetry $l(v_i) = l(v_{n+1-i})$ for all $1 \le i \le n$. We start with small values of n. When $n = 2$, both vertices have a leverage centrality of zero. When $n = 3$, the two vertices of degree 1 have leverage centrality $\frac{1-2}{1+2} = -\frac{1}{3}$ and the vertex of degree 2 has leverage centrality $\frac{2-1}{1+2} = \frac{1}{3}$. When $n = 4$, the two vertices of degree 1 have leverage centrality $\frac{1-2}{1+2} = -\frac{1}{3}$ and the two vertices of degree 2 have leverage centrality $\frac{1}{2}\left(\frac{2-1}{1+2} + \frac{2-2}{2+2}\right) = \frac{1}{6}$.

Next, we use the operation of edge subdivision to handle cases where $n \geq 5$. Recall that in an edge subdivision an edge $u - v$ is replaced by a path on three vertices $u - w - v$. We note that if we extend the length of a path by subdividing the edge between vertices c and d, the new vertices will have a leverage centrality of zero. Further subdivision of an edge connecting two vertices with degree 2 will include a new vertex with leverage centrality zero. Hence, there will be exactly three distinct leverage centralities in any path with five or more vertices. The general result follows.

Theorem 5. *Let P_n be a path where $n \geq 5$. Then $l(v_1) = l(v_n) = -\frac{1}{3}$, $l(v_2) = l(v_{n-1}) = \frac{1}{6}$, and for all $3 \leq i \leq n-2$, we have $l(v_i) = 0$.*

3.1. *Leverage centralities of P_n^2.*

We now calculate the leverage centralities for paths P_n^2. Again by symmetry, we have $l(v_i) = l(v_{n+1-i})$ for all $1 \leq i \leq n$.

- $n = 3$: For $1 \leq i \leq 3$, we have $l(v_i) = 0$.
- $n = 4$: $l(v_1) = -\frac{4}{15}$ and $l(v_2) = \frac{2}{15}$.
- $n = 5$: $l(v_1) = -\frac{4}{15}$, $l(v_2) = \frac{2}{105}$ and $l(v_3) = \frac{5}{21}$.
- $n = 6$: $l(v_1) = -\frac{4}{15}$, $l(v_2) = -\frac{1}{35}$ and $l(v_3) = \frac{13}{84}$.
- $n = 7$: $l(v_1) = -\frac{4}{15}$, $l(v_2) = -\frac{1}{35}$ $l(v_3) = \frac{5}{42}$ and $l(v_4) = \frac{3}{28}$.
- $n = 8$: $l(v_1) = -\frac{4}{15}$, $l(v_2) = -\frac{1}{35}$, $l(v_3) = \frac{5}{42}$ and $l(v_4) = \frac{1}{28}$.
- $n \geq 9$: $l(v_1) = -\frac{4}{15}$, $l(v_2) = -\frac{1}{35}$, $l(v_3) = \frac{5}{42}$ $l(v_4) = \frac{1}{28}$ and for all $5 \leq i \leq n-4$, $l(v_i) = 0$.

It is clear that to calculate the leverage centralities of all vertices in P_n^k for all k in this manner would require lengthy computation. However by noticing that the leverage centralities become fixed when n becomes large enough ($n \geq 4k+1$), we can compute the leverage centralities in a more formal manner.

First we give an elementary result with the leverage centralities for the first vertex in any path power.

Proposition 6. *If $v_1 \in V(P_n^k)$, then*

$$l(v_1) = \sum_{i=1}^{k} \frac{-i}{2k+i}.$$

Proof. The vertex v_1 has k neighbors, with degrees $k+1, k+2, \ldots, 2k$. Then

$$l(v_1) = \frac{1}{k}\left(k \sum_{i=1}^{k} \frac{-i}{2k+i}\right) = \sum_{i=1}^{k} \frac{-i}{2k+i}. \qquad \square$$

We continue with three lemmas which will help us determine the relationships between the leverage centralities of different vertices in P_n^k.

Lemma 7. *If i is an integer and $1 < \frac{1}{2}a \le i < a$ then we have*

$$\frac{1}{a}\left(\frac{a-i}{a+i}\right) > \frac{1}{a-1}\left(\frac{a-1-i}{a-1+i}\right).$$

Proof. Let $\frac{1}{2}a \le i < a$. This implies

$$\begin{aligned} 2ia - a^2 + (a-i) > 0 &\Rightarrow -a^2 + a + 2ia - 1 - i > 0 \\ &\Rightarrow a^3 - 2a^2 + (2+i)a - 1 - i > a^3 - a^2 + (1-i)a \\ &\Rightarrow (a-i)(a-1)(a-1+i) > a(a+i)(a-1-i) \\ &\Rightarrow \frac{1}{a}\left(\frac{a-i}{a+i}\right) > \frac{1}{a-1}\left(\frac{a-1-i}{a-1+i}\right). \qquad \square \end{aligned}$$

Lemma 8. *For all $1 \le a \le 2k$, we have*

$$\frac{1}{a}\left(\frac{-1}{2a+1}\right) > \frac{-1}{(a-1)a}.$$

Proof. We first note that when $a = 2k$,

$$\frac{1}{a}\left(\frac{a-(a+1)}{a+a+1}\right) > \frac{1}{a-1}\left(\frac{a-1-(a+1)}{a-1+a+1}\right)$$

is clear since the left side is positive and the right side is negative.

Let $1 \le a$. Then

$$\begin{aligned} 2a^2 + a > a^2 + 1 &\Rightarrow \frac{-1}{2a^2+a} > \frac{-1}{a^2+1} \\ &\Rightarrow \frac{1}{a}\left(\frac{-1}{2a+1}\right) > \frac{1}{a+1}\left(\frac{-2}{2a}\right) \\ &\Rightarrow \frac{1}{a}\left(\frac{a-(a+1)}{a+a+1}\right) > \frac{1}{a-1}\left(\frac{a-1-(a+1)}{a-1+a+1}\right) \\ &\Rightarrow \frac{1}{a}\left(\frac{-1}{a+a+1}\right) > \frac{1}{a-1}\left(\frac{-2}{a-1+a+1}\right). \qquad \square \end{aligned}$$

Lemma 9. *Let $2 \le i \le k-1$. Then*

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

Proof. Note that

$$0 > \frac{1-i}{2k+i+1} > \frac{-i}{2k+i+1} > \frac{-i}{2k+i}.$$

Hence

$$0 > \frac{k+1-(k+i)}{k+1+k+i} > \frac{k-(k+i)}{k+k+i}.$$

Since $\frac{1}{k+1} < \frac{1}{k}$, we have

$$0 > \frac{1}{k+1}\left(\frac{1-i}{k+1+k+i}\right) > \frac{1}{k}\left(\frac{-i}{k+k+i}\right). \qquad \square$$

Proposition 6 and Lemmas 7, 8, and 9 can be combined as follows.

Proposition 10. *Let $G = P_n^k$, where $n \geq 4k+1$. Then*:

(i) $l(v_i) = l(v_{n+1-i})$.

(ii) *For all* $0 \leq j \leq k-1$,

$$l(v_{k+j+1}) = \frac{1}{2k}\left(\sum_{i=k+j}^{2k-1} \frac{2k-i}{2k+i}\right).$$

(iii) *For all* $0 \leq j \leq k-1$,

$$l(v_{k-j}) = \frac{1}{2k-j-1}\sum_{i=k}^{2k-1}\frac{2k-j-i}{2k-j+i} + \frac{k-j}{2k-1-j}\left(\frac{2k-j-1-2k}{2k-j-1+2k}\right).$$

(iv) *For all* $2k+1 \leq j \leq n-2k$, *we have* $l(v_j) = 0$.

This leads to the following theorem.

Theorem 11. *Let $G = P_n^k$, where $n \geq 4k+1$. Then the vertex with the largest leverage centrality in G is v_{k+1}, and furthermore $l(v_{k+1}) > l(v_k) > \cdots > l(v_1)$ and $l(v_{k+1}) > l(v_{k+2}) > \cdots > l(v_{2k+1})$.*

Proof. For the first part, we recall that

$$l(v_{k+1}) = \frac{1}{2k}\sum_{i=k}^{2k-1}\left(\frac{2k-i}{2k+i}\right) + \frac{k}{2k}\left(\frac{2k-2k}{2k+2k}\right),$$

and for $0 \leq j \leq k-1$,

$$l(v_{k-j}) = \frac{1}{2k-j-1}\sum_{i=k}^{2k-1}\frac{2k-j-i}{2k-j+i} + \frac{k-j}{2k-1-j}\left(\frac{2k-j-1-2k}{2k-j-1+2k}\right).$$

We seek to show that $l(v_{k+1}) > l(v_k) > \cdots > l(v_1)$. When comparing terms from $l(v_r)$ with $l(v_{r-1})$ for a fixed i, five cases are needed to show that the i-th term of $l(v_r)$ is larger than the i-th term of $l(v_{r-1})$.

Case (i): $2k - j - i > 1$. Use Lemma 7.

Case (ii): $2k - j - i = 1$. In the i-th term, the numerator is positive for the j-th term and the numerator is zero for $(j+1)$-th term.

Case (iii): $2k - j - i = 0$. In the i-th term, the numerator is zero for the j-th term and the numerator is negative for $(j+1)$-th term.

Case (iv): $2k - j - i < 0$. Use Lemma 8.

Case (v):

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

Use Lemma 8.

The combination of these five cases yields $l(v_{k+1}) > l(v_k) > \cdots > l(v_1)$. For the second part we note that for $0 \le r \le k-1$, we have $l(v_{k+r}) > l(v_{k+r+1})$ as terms with positive value are replaced by zeros in each successive case. Hence, $l(v_{k+1}) > l(v_{k+2}) > \cdots > l(v_{2k+1})$. We note that we have not obtained a linear ordering, but two separate linear orderings both starting with the largest leverage centrality $l(v_{k+1})$. □

4. Cartesian product of graphs

In this next section we give some general results about the leverage centrality of the Cartesian product of graphs. These build upon results by Sharma et al. [2017].

Definition 12. Given a graph F with vertex set $V(F)$ and edge set $E(F)$, and a graph H with vertex set $V(H)$ and edge set $E(H)$, we let G define the Cartesian product of F and H to be the graph $G = F \times H$, which is defined as

$$\begin{aligned} V(G) &= \{(u, v) \mid u \in V(F) \text{ and } v \in V(H)\}, \\ E(G) &= \bigl\{(u_1, v_1), (u_2, v_2) \bigm| u_1 = u_2 \text{ and } (v_1, v_2) \in E(H), \\ &\qquad\qquad\qquad\qquad\qquad \text{or } v_1 = v_2 \text{ and } (u_1, u_2) \in E(F)\bigr\}. \end{aligned}$$

We next present an elementary result from graph theory.

Lemma 13. *If $G = F \times H$, then the degree of a vertex (u, v) in G is the sum of the degrees of vertices u and v, where $u \in V(F)$ and $v \in V(H)$.*

Theorem 14. *Let G be a graph and let RG_r be a regular graph where each vertex has degree r. Let $u \in V(RG_r)$ and let v_i and v_j be vertices in G with degrees k_i and k_j respectively. For each vertex $(u, v_i) \in V(RG_r \times G)$ we have*

$$l(u, v_i) = \frac{1}{r + k_i} \sum_{j \ne i} \frac{k_i - k_j}{2r + k_i + k_j}.$$

Proof. Consider a vertex $(u, v_i) \in V(RG_r \times G)$. We note that

$$\deg(u, v_i) = \deg u + \deg v_i = r + k_i.$$

Then

$$l(u, v_i) = \frac{1}{r+k_i} \sum_{j \neq i} \frac{k_i - k_j}{2r + k_i + k_j}. \qquad \square$$

We conclude by posing the following problem where the graphs may not be regular.

Problem 15. Given graphs F and H where the leverage centralities are known for all vertices in F and H, determine the leverage centralities for all vertices in $F \times H$.

Acknowledgements

We would like to thank an anonymous referee for helpful comments. Research was supported by National Science Foundation Research Experience for Undergraduates Grant #1358583.

References

[Joyce et al. 2010] K. E. Joyce, P. J. Laurienti, J. H. Burdette, and S. Hayasaka, “A new measure of centrality for brain networks”, *PLOS ONE* **5**:8 (2010), art. id. e12200.

[Li et al. 2015] C. Li, Q. Li, P. Van Mieghem, H. E. Stanley, and H. Wang, “Correlation between centrality metrics and their application to the opinion model”, *Eur. Phys. J. B* **88**:3 (2015), art. id. 65. MR

[Sharma et al. 2017] A. Sharma, R. Vargas, A. Waldron, R. Florez, and D. A. Narayan, “Leverage centrality of graphs”, preprint, 2017. arXiv

Received: 2015-09-09 Revised: 2016-01-04 Accepted: 2016-03-25

roger_vargas@g.harvard.edu
*Mathematics and Statistics, Williams College,
Williamstown, MA 01267, United States*

awaldron@presby.edu
*Mathematics Department, Presbyterian College,
Clinton, SC 29325, United States*

anikasha@buffalo.edu
Department of Computer Science and Department of Mathematics, University of Buffalo, Buffalo, NY 14260, United States

rigoflorez@gmail.com
*Department of Mathematics and Computer Science,
The Citadel, Charleston, SC 29409, United States*

dansma@rit.edu
*School of Mathematical Sciences, Rochester Institute of
Technology, Rochester, NY 14623, United States*

INVOLVE 10:4 (2017)
dx.doi.org/10.2140/involve.2017.10.593

Equivalence classes of $\mathrm{GL}(p,\mathbb{C})\times\mathrm{GL}(q,\mathbb{C})$ orbits in the flag variety of $\mathfrak{gl}(p+q,\mathbb{C})$

Leticia Barchini and Nina Williams

(Communicated by Ken Ono)

We consider the pair of complex Lie groups

$$(G,K)=\bigl(\mathrm{GL}(p+q,\mathbb{C}),\mathrm{GL}(p,\mathbb{C})\times\mathrm{GL}(q,\mathbb{C})\bigr)$$

and the finite set $\{\mathcal{Q}:K\text{-orbits on the flag variety }\mathfrak{B}\}$. The moment map μ of the G-action on the cotangent bundle $T^*\mathfrak{B}$ maps each conormal bundle closure $\overline{T^*_{\mathcal{Q}}\mathfrak{B}}$ onto the closure of a single nilpotent K-orbit, $\mathcal{O}_K$. We use combinatorial techniques to describe $\mu^{-1}(\mathcal{O}_K)=\{\mathcal{Q}\in\mathfrak{B}:\mu(T^*_{\mathcal{Q}}\mathfrak{B})=\mathcal{O}_K\}$.

Introduction

We consider the pair (G,K) of complex groups equal to

$$\bigl(\mathrm{GL}(p+q,\mathbb{C}),\mathrm{GL}(p,\mathbb{C})\times\mathrm{GL}(q,\mathbb{C})\bigr).$$

Such a pair comes from the real Lie group $U(p,q)$, and K is the complexification of the maximal compact subgroup $K_{\mathbb{R}}=U(p)\times U(q)$. We denote by $\mathfrak{g}$ the Lie algebra of G. The group K acts with finitely many orbits both on $\mathcal{N}$, the nilpotent cone of $\mathfrak{g}$, and on $\mathfrak{B}$, the flag variety of $\mathfrak{g}$. The points in the cotangent bundle $T^*\mathfrak{B}$ can be thought of as pairs $(\mathfrak{b},\xi)$ consisting of a Borel subalgebra $\mathfrak{b}=\mathfrak{h}\oplus\mathfrak{n}$ and a covector $\xi\in\mathfrak{n}^*$. The projection $\mu:(\mathfrak{b},\xi)\to\xi$ from the cotangent bundle $T^*\mathfrak{B}$ to $\mathcal{N}$ is the moment map for the G-action on $T^*\mathfrak{B}$. If $\mathcal{Q}$ is a K-orbit on $\mathfrak{B}$, the image $\mu(\overline{T^*_{\mathcal{Q}}\mathfrak{B}})$ lies in $\mathcal{N}$ and it is the closure of a nilpotent K-orbit. We write $\mathcal{O}_K$ for the nilpotent K-orbit. We give a combinatorial algorithmic description, amenable to computer computations, of the set

$$\mu^{-1}(\mathcal{O}_K)=\{\mathcal{Q}\in\mathfrak{B}:\mu(T^*_{\mathcal{Q}}\mathfrak{B})=\mathcal{O}_K\}. \tag{0.1}$$

This is the content of Theorem 4.3. Our approach relies heavily on work by Devra Garfinkle [1993], and on work by Peter Trapa [1999]. Our goal is to keep the presentation accessible to an advanced undergraduate student. Some of our

MSC2010: primary 22E47; secondary 22E46.
Keywords: nilpotent orbits, flag variety, Young tableaux.

arguments can be simplified by using advanced results in representation theory, but we choose instead a combinatorial approach.

We use the combinatorial notion of a *clan* to parametrize K-orbits in $\mathfrak{B}$, as in [Matsuki and Oshima 1990]. For each nilpotent orbit, $\mathcal{O}_K$, we identify a *distinguished clan* $\boldsymbol{c}_{\text{dis}} \in \mu^{-1}(\mathcal{O}_K)$. All other clans in $\mu^{-1}(\mathcal{O}_K)$ are obtained from the distinguished clan in a combinatorial manner. Following [Garfinkle 1993], we attach to each clan $\boldsymbol{c}$ a pair of equally shaped tableaux, one signed and the other numbered. It is known, see [Trapa 1999], that the signed tableau determines $\mu(T^*_{\mathcal{Q}_c}\mathfrak{B}) = \mathcal{O}_K$, where $\mathcal{Q}_{\boldsymbol{c}}$ is the K-orbit parametrized by $\boldsymbol{c}$. The resulting map

$$E : \{\text{clans}\} \to \{(T_\pm, ST_{\boldsymbol{c}})\}$$

is a bijection. Thus, if we fix $\mathcal{O}_K$ and we let $T^{\text{dis}}_\pm$ be the signed tableau that corresponds to $\boldsymbol{c}_{\text{dis}} \in \mu^{-1}(\mathcal{O}_K)$ under E, we have

$$\mu^{-1}(\mathcal{O}_K) = \{\mathcal{Q}_{\boldsymbol{c}} \text{ clans} : E(\boldsymbol{c}) = (T^{\text{dis}}_\pm, ST_{\boldsymbol{c}})\}.$$

That is, $\mu^{-1}(\mathcal{O}_K)$ is the set of K-obits on $\mathfrak{B}$ parametrized by clans $\boldsymbol{c}$ having $T^{\text{dis}}_\pm$ as the signed tableau in $E(\boldsymbol{c})$. In order to explicitly describe the set $\mu^{-1}(\mathcal{O}_K)$, we use combinatorially defined operators $T_{i,j}$ acting both on clans and on numbered tableaux. The bijection E is compatible with the action of such operators. We conclude that if $\boldsymbol{c} \in \mu^{-1}(\mathcal{O}_K)$, then so is $T_{i,j}\,\boldsymbol{c}$. We argue that any clan in $\mu^{-1}(\mathcal{O}_K)$ can be obtained from the distinguished clan by applying an appropriate sequence of operators $T_{i,j}$. This is the content of Theorem 4.3. If $n = p+q$, and the shape of the tableau is fixed, then the action of operators $T_{i,j}$ on numbered tableaux of that given shape determines $\mu^{-1}(O_{\mathrm{GL}(r,\mathbb{C})\times\mathrm{GL}(s,\mathbb{C})})$ for any (r, s) with $r+s = n$. This implies that the algorithm is in a sense independent of the real form; see Theorem 4.5. When nilpotent K-orbits are parametrized by two-column signed tableaux, we give explicit effective sequences of operators $T_{i,j}$ to generate $\mu^{-1}(\mathcal{O}_K)$. We use this result to describe the clans in $\mu^{-1}(\mathcal{O}_K)$ in special cases. The two column case is discussed in Section 5.

The problem of describing $\mu^{-1}(\mathcal{O}_K)$ when $K = \mathrm{GL}(p, \mathbb{C})\times\mathrm{GL}(q, \mathbb{C})$, considered in this paper, is a particular instance (and an easy one) of a more general question posted by David Vogan.

The paper is organized as follows. We fix notation, and we introduce combinatorial parametrizations of nilpotent orbits and K-orbits in $\mathfrak{B}$ in Section 1. In Section 2, we summarize Garfinkle's algorithm, we describe some of its properties, and we introduce the notion of distinguished clan. We include in Section 3 the definition of operators $T_{i,j}$ at both the tableau and clan level, and we explain some of their properties. We obtain an algorithmic description of $\mu^{-1}(\mathcal{O}_K)$ and prove our main theorem in Section 4. In Section 5, we restrict our attention to nilpotent K-orbits parametrized by two-column signed tableaux and give a detailed description of $\mu^{-1}(\mathcal{O}_K)$ in special cases.

1. Preliminaries

The real form $U(p,q)$. In this section we carefully define the real form of interest. Assume p and q are positive integers with $p \geq q$. Write $n = p+q$, and let

$$I_{p,q} = \begin{pmatrix} I_{p\times p} & 0 \\ 0 & -I_{q\times q} \end{pmatrix},$$

where $I_{p\times p}$, $I_{q\times q}$ are identity matrices. Define

$$G_{\mathbb{R}} = U(p,q) = \{g \in \mathrm{GL}(n,\mathbb{C}) : \bar{g}^T I_{p,q} g = I_{p,q}\}.$$

The map Θ given by

$$\begin{aligned} \Theta : \mathrm{GL}(n,\mathbb{C}) &\to \mathrm{GL}(n,\mathbb{C}), \\ A &\mapsto I_{p,q} A I_{p,q}, \end{aligned}$$

is an involution. We call Θ the Cartan involution. Then,

$$\begin{aligned} \mathrm{GL}(n,\mathbb{C})^{\Theta} &= \left\{A \in \mathrm{GL}(n,\mathbb{C}) : \Theta(A) = A\right\} = K \\ &= \left\{\begin{pmatrix} Z_1 & 0 \\ 0 & Z_2 \end{pmatrix} : Z_1 \in \mathrm{GL}(p,\mathbb{C}),\, Z_2 \in \mathrm{GL}(q,\mathbb{C})\right\}. \end{aligned}$$

Similarly, we have

$$U(p,q)^{\Theta} = U(p) \times U(q) = K_{\mathbb{R}}.$$

The differential of Θ, denoted by θ, is an involution at the Lie-algebra level. That is $\theta : \mathfrak{gl}(n,\mathbb{C}) \to \mathfrak{gl}(n,\mathbb{C})$ has $\theta^2 = 1$. The ±-eigenspace decomposition of $\mathfrak{gl}(n,\mathbb{C})$ is

$$\mathfrak{g} = \mathfrak{gl}(n,\mathbb{C}) = \mathfrak{k} \oplus \mathfrak{p},$$

where

$$\mathfrak{k} = \left\{\begin{pmatrix} z_1 & 0 \\ 0 & z_2 \end{pmatrix} : z_1 \in \mathfrak{gl}(p,\mathbb{C}),\, z_2 \in \mathfrak{gl}(q,\mathbb{C})\right\},$$

$$\mathfrak{p} = \left\{\begin{pmatrix} 0 & A \\ B & 0 \end{pmatrix} : A \in M(p\times q),\, B \in M(q \times p)\right\}.$$

Define $\mathfrak{h} \subset \mathfrak{k}$ as the Cartan subalgebra consisting of diagonal matrices of the form $\mathrm{diag}(t_1, t_2, \ldots, t_{p+q})$. This is a maximally abelian subalgebra of $\mathfrak{g}$. The matrices $E_{i,j}$ with all entries zero but for a 1 in the intersection of the i-th row, j-th column satisfy

$$[\mathrm{diag}(t_1, t_2, \ldots, t_{p+q}),\, E_{i,j}] = (t_i - t_j)\, E_{i,j}.$$

In other words, the $E_{i,j}$ are common eigenvectors of the matrices in $\mathfrak{h}$. They are called root vectors. Their eigenvalues $\epsilon_i - \epsilon_j$, given by

$$(\epsilon_i - \epsilon_j)(\operatorname{diag}(t_1, t_2, \dots, t_{p+q})) = t_i - t_j,$$

are called roots. A root $\epsilon_i - \epsilon_j$ is said to be positive if $i < j$. We set

$$\mathfrak{n} = \bigoplus_{i<j} \mathbb{C}E_{i,j}, \qquad \mathfrak{b} = \mathfrak{h} \oplus \mathfrak{n}, \quad \text{upper triangular matrices.} \tag{1.1}$$

The subalgebra $\mathfrak{b} \subset \mathfrak{g}$ is a Borel subalgebra.

***K*-orbits on the flag variety of *G*.** The flag variety of G is the variety of Borel subalgebras of $\mathfrak{g}$. We describe this variety geometrically as follows.

Definition 1.2. A flag of G is a sequence of $n+1$ complex vector spaces, $\mathcal{F} = (V_0, V_1, \dots, V_n)$, satisfying the conditions

(1) $\dim V_i = i$;

(2) $\{0\} = V_0 \subset V_1 \subset V_2 \subset \cdots \subset V_n = \mathbb{C}^n$.

We define $\mathfrak{B} = \{\text{flags in } \mathbb{C}^n\}$.

The group G acts on $\mathfrak{B}$ via

$$g \cdot \mathcal{F} = (g \cdot V_0, g \cdot V_1, \dots, g \cdot V_n).$$

Let $\{e_1, \dots, e_n\}$ denote the standard basis of $\mathbb{C}^n$, and for each integer $1 \le i \le n$, set $V_i^0 = \langle e_1, \dots, e_i \rangle$. Define $\mathcal{F}_0 = (\{0\}, V_1^0, \dots, V_n^0)$. It is not difficult to see that for any flag, $\mathcal{F}$, there exists a $g \in G$ so that $\mathcal{F} = g \cdot \mathcal{F}_0$. This implies that the action of G on $\mathfrak{B}$ is transitive.

Theorem 1.3. *G acts transitively on $\mathfrak{B}$.*

If $\mathcal{F}_0 = \big(\{0\}, \langle e_1 \rangle, \langle e_1, e_2 \rangle, \dots, \langle e_1, \dots, e_{n-1} \rangle, \mathbb{C}^n\big)$, then $G \cdot \mathcal{F}_0 \cong \mathfrak{B} \cong G/B$, where

$$B = \operatorname{Stab}_G(\mathcal{F}_0) = \begin{pmatrix} e_{11} & e_{12} & \cdots & \cdots & e_{1n} \\ 0 & e_{22} & & & \vdots \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & 0 & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & e_{nn} \end{pmatrix}.$$

The following known theorem will play an important role in our work.

Theorem 1.4. *K acts on $\mathfrak{B}$ with finitely many orbits.*

Clan parametrization of K-orbits on the flag variety of G. It will be useful to parametrize K-orbits in $\mathfrak{B}$ in a combinatorial manner. To this end, we use the notion of clans. Clans have been introduced in [Matsuki and Oshima 1990]. We follow the presentation in [Yamamoto 1997].

Definition 1.5. An n-indication is a sequence of symbols $(c_1 \cdots c_n)$ so that

(1) c_i is $+$, $-$, or a natural number;

(2) if $c_i = a \in \mathbb{N}$, then there exists a unique c_j with $c_i = c_j = a$;

(3) $\#\{i : c_i = +\} + \#\{\text{pairs of equal numbers}\} = p$.

We define an equivalence relation between two indications. Two indications $(c_1 \cdots c_n)$ and $(c'_1 \cdots c'_n)$ are equivalent if and only if there exists a permutation σ so that

$$c_i = \begin{cases} \sigma(c'_i) & \text{if } c'_i \in \mathbb{N}, \\ + & \text{if } c'_i = +, \\ - & \text{if } c'_i = -. \end{cases}$$

A *clan* is an equivalence class of indications with respect to the equivalence relation.

Define $V_+ = \langle e_1, \dots, e_p \rangle$ and $V_- = \langle e_{p+1}, \dots, e_{p+q} \rangle$.

Proposition 1.6 [Yamamoto 1997, Proposition 2.2.7]. *Let $p + q = n$. Given a flag $\mathcal{F} = (V_0, V_1, \dots, V_n)$ there exists a clan $\mathbf{c} = (c_1 \cdots c_n)$ so that*

(1) $\dim V_i \cap V_+ = \#\{l : c_l = + \textit{ for } l \leq i\} + \#\{a \in \mathbb{N} : c_s = c_t = a \textit{ for } s < t \leq i\}$;

(2) $\dim V_i \cap V_- = \#\{l : c_l = - \textit{ for } l \leq i\} + \#\{a \in \mathbb{N} : c_s = c_t = a \textit{ for } s < t \leq i\}$;

(3) $\dim V_i - \dim V_i \cap V_+ - \dim V_i \cap V_- = \#\{a \in \mathbb{N} : c_s = c_t = a \textit{ for } s \leq i < t\}$;

(4) $\dim V_j + \pi_+(V_i) = j + \#\{a \in \mathbb{N} : c_s = c_t = a \textit{ for } s \leq i < j < t\}$.

Moreover, the set of flags that corresponds to a given clan $\mathbf{c}$, *constitutes a K-orbit in $\mathfrak{B}$.*

The converse of the proposition also holds. Hence, we have the following theorem.

Theorem 1.7 [Yamamoto 1997]. *Clans parametrize K-orbits in $\mathfrak{B}$.*

Example. Assume $G_{\mathbb{R}} = U(2, 2)$.

- The clan $(+ + - -)$ corresponds to the flag

$$\mathcal{F}_0 = \big(\{0\} \subset \langle e_1 \rangle \subset \langle e_1, e_2 \rangle \subset \langle e_1, e_2, e_3 \rangle \subset \mathbb{C}^4\big).$$

- The clan $(1\,2\,2\,1)$ corresponds to the flag

$$\mathcal{F} = \big(\{0\} \subset \langle e_1 \rangle \subset \langle e_1, e_2 + e_3 \rangle \subset \langle e_1, e_2, e_3 \rangle \subset \mathbb{C}^4\big).$$

Example. Assume $G_{\mathbb{R}} = U(4,4)$. We attach a flag $\mathcal{F}_{\boldsymbol{c}}$, satisfying (1) through (4) of Proposition 1.6, to the clan $\boldsymbol{c} = (1\,2+3\,1-2\,3) = (c_1\,c_2\,c_3\,c_4\,c_5\,c_6\,c_7\,c_8)$. Write $\mathcal{F} = (V_0{=}\{0\}, V_1, V_2, \dots, \mathbb{C}^8)$. As $c_1 = c_5 = 1$, we set $V_1 = \langle e_1 + e_5\rangle$. Note that

$$\begin{cases} \dim V_1 \cap V_+ = 0, \\ \dim V_1 \cap V_- = 0. \end{cases}$$

Similarly, we note that $c_2 = c_7 = 2$ and define $V_2 = \langle e_1 + e_5, e_2 + e_7\rangle$. Next, as $c_3 = +$, we set $V_3 = \langle e_1 + e_5, e_2 + e_7, e_3\rangle$. It is easy to check, as $c_1 = c_5$ and $c_2 = c_7$, that $\dim V_3 \cap V_+ = 1$, $\dim V_3 \cap V_- = 0$, and $\dim V_3 - \dim V_3 \cap V_+ - \dim V_3 \cap V_- = 2$.

Continuing in similar manner we get

$$\mathcal{F}_{\boldsymbol{c}} = \big(\langle e_1 + e_5\rangle \subset \langle e_1 + e_5, e_2 + e_7\rangle \subset \langle e_1 + e_5, e_2 + e_7, e_3\rangle \\ \subset \langle V_3, e_4 + e_8\rangle \subset \langle V_4, e_1 - e_5\rangle \subset \langle V_5, e_6\rangle \subset \langle V_6, e_2 - e_7\rangle \subset \mathbb{C}^8\big).$$

Example. Assume $G_{\mathbb{R}} = U(3,2)$. The flag

$$\big(\{0\} \subset \langle e_1\rangle \subset \langle e_1, e_2 + e_4\rangle \subset \langle e_1, e_2 + e_4, e_3\rangle \subset \langle e_1, e_3, e_4, e_5\rangle \subset \mathbb{C}^5\big)$$

is parametrized by $(+\,1+1\,-)$.

Young diagrams. We introduce some combinatorial tools used in our work.

Definition 1.8. A partition of n is a tuple $[d_1, d_2, \dots, d_k]$ of positive integers with

(1) $d_1 \geq d_2 \geq \cdots \geq d_k > 0$, and

(2) $\sum d_k = n$.

Given a partition $[d_1, d_2, \dots, d_k]$, we form a left-justified array of n rows of empty boxes so that the i-th row has length d_i. This is called a Young diagram.

Definition 1.9. A signed tableau is a labeled Young diagram in which boxes are labeled by $+$ and $-$ signs in such a way that the signs alternate along rows. Two signed tableaux are regarded as equal if and only if one can be obtained from the other by interchanging rows of equal length.

Definition 1.10. The signature of a signed tableau is a pair of numbers (i, j), where $i = \#\{+ \text{ signs in the tableau}\}$ and $j = \#\{- \text{ signs in the tableau}\}$.

Definition 1.11. A standard tableau is a labeled Young diagram in which boxes are labeled by numbers that monotonically increase along rows (from left to right) and increase strictly along columns (from top to bottom). We write $b_{i,j}$ for the box in the intersection of the i-th row and j-th column.

Nilpotent G and K-orbits. We think of a nilpotent matrix $X_{n\times n}$ as a linear transformation

$$T_X : \mathbb{C}^n \to \mathbb{C}^n \text{ such that } T^k = 0 \text{ for some } k.$$

Linear algebra tells us that we can write

$$\mathbb{C}^n = V_{p_1} \oplus V_{p_2} \oplus \cdots \oplus V_{p_r}$$

as a sum of vector subspaces with the following properties:

- $T_X : V_{p_i} \to V_{p_i}$.
- Each V_{p_i} admits a basis such that

$$e^i_{p_i} \xrightarrow{T_X} e^i_{p_i-1} \xrightarrow{T_X} \cdots \xrightarrow{T_X} e^i_1 \xrightarrow{T_X} 0.$$

In this basis T_X is represented by its Jordan form J. Moreover, if $Y = g^{-1}Xg$ for some $g \in G$, then the matrix of T_Y with respect to the basis $\{g^{-1}e^i\}$ is also J. We conclude that G acts on the set of nilpotent matrices by conjugation and that this action yields a finite number of orbits.

The Jordan decomposition theorem implies that we can attach to each nilpotent G-orbit, $G \cdot X$, a Young diagram which is completely determined by the Jordan form of X. Indeed, the lengths of the rows of the corresponding Young diagram are given by the size of the Jordan blocks. The following known proposition states that the map from nilpotent G-orbits to Young diagrams is a bijection.

Proposition 1.12 [Collingwood and McGovern 1993]. *There is a one-to-one correspondence between the set of nilpotent orbits and the set of partitions of n. The correspondence sends a nilpotent element X to the partition determined by the block-size of its Jordan form. The orbit* 0 *corresponds to the partition* $[1, 1, \dots, 1]$.

The group K acts by conjugation of the set $\mathcal{N} \cap \mathfrak{p}$ of nilpotent matrices of the form

$$X = \begin{pmatrix} 0 & A_{p\times q} \\ B_{q\times p} & 0 \end{pmatrix}.$$

If we write

$$\mathbb{C}^n = V^+ \oplus V^-, \quad \text{where } V^+ = \langle e_1, \dots, e_p\rangle,\ V^- = \langle e_{p+1}, \dots, e_{p+q}\rangle,$$

then

$$\begin{aligned} X &: V^+ \to V^-, \\ X &: V^- \to V^+. \end{aligned} \tag{1.13}$$

A generalized version of the Jordan decomposition theorem, combined with (1.13), yields a parametrization of K-orbits on $\mathcal{N} \cap \mathfrak{p}$ via Young diagrams with boxes labeled by alternating signs, $+$ and $-$. Our next proposition is well-known and follows from the above discussion.

Proposition 1.14. *There is a one-to-one correspondence between K-orbits in $\mathcal{N} \cap \mathfrak{p}$ and signed tableaux.*

We fix $p \geq q$ with $p+q=n$ and a partition $\lambda = [r_1, r_2, \dots, r_\ell]$ of n. Such a partition determines a Young diagram of size n. Let $[p_1, p_2, \dots, p_r]$ be the length of the columns of the Young diagram determined by λ.

Proposition 1.15. *Fix $p \geq q$ with $p+q=n$, and fix $[p_1, p_2, \dots, p_r]$ integers with $\sum p_i = n$. There is a bijection*

$$\left\{ \begin{matrix} \text{nilpotent } K\text{-orbits } \mathcal{O}_K \text{ parametrized by} \\ \text{tableaux of column lengths } [p_1, \dots, p_r] \end{matrix} \right\} \longleftrightarrow \left\{ \begin{matrix} (t_1, \dots, t_s) \text{ integers, } s \leq p_1, \\ t_1 < t_2 < \cdots < t_s \end{matrix} \right\}.$$

Proof. Assume $\mathcal{O}_K$ is a nilpotent K-orbit parametrized by a signed tableau of shape λ. Note that such a signed tableau is completely determined by its shape and the position of the $-$ signs on the first column of the tableau. The proposition follows by letting $t_1 < t_2 < \cdots < t_s$ denote the positions of the $-$ signs in the first column of the parametrizing tableau. □

2. Garfinkle's algorithm

In this section we describe the algorithm defined in [Garfinkle 1993]. The algorithm assigns to each clan a pair of equally shaped tableaux; one signed, the other numbered. The resulting map has significant representational theoretical meaning. The relevance of the algorithm in our work is explained in the introduction.

Garfinkle's algorithm. Starting with a clan $\boldsymbol{c} = (c_1, c_2, \dots, c_n)$ form a sequence of pairs

$$(i, \epsilon_i) \quad \text{if } c_i = \epsilon_i,$$
$$(i, j) \quad \text{if } c_i = c_j.$$

Arrange the pairs in order by the largest entry, with the convention that a sign has numerical size 0. Write $\pi_1, \dots, \pi_r$ for the resulting ordered sequence. Suppose that a smaller, equally shaped pair of tableaux $(T_\pm, ST)$ has been constructed from $\pi_1, \dots, \pi_{j-1}$. If $\pi_j = (k, \epsilon_k)$, then first add the sign ϵ_k to the topmost row of (a signed tableau in the equivalence class of) $T_\pm$ so that the resulting tableau has signs alternating across rows. Then add the integer k to ST in the unique position so that the two new tableaux have the same shape. If $\pi_j = (k, \ell)$, first add k to ST using the Robinson–Schensted bumping algorithm to get a new tableau ST', and then add a sign ϵ (either $+$ or $-$ as needed) to $T_\pm$ so that the result is a signed tableau $T'_\pm$ of the same shape as ST'. Then add $(\ell, -\epsilon)$ (by the same recipe as the first case) to the first row strictly below the row to which ϵ was added.

Example. Assume $G_\mathbb{R} = U(2, 2)$, and consider the clan $(1 - + 1)$. Attach to $(1 - + 1)$ the sequence $(2, -)(3, +)(1, 4)$.

We associate to $(2,-)(3,+)$ a pair of tableaux, one a signed tableau, the other a standard tableau:

$$\begin{array}{|c|c|}\hline - & + \\ \hline\end{array} \qquad\qquad \begin{array}{|c|c|}\hline 2 & 3 \\ \hline\end{array}$$

Next, we add $(1,4)$ to obtain

$$\begin{array}{|c|c|}\hline - & + \\ \hline + & \multicolumn{1}{c}{} \\ \cline{1-1} - & \multicolumn{1}{c}{} \\ \cline{1-1}\end{array} \qquad\qquad \begin{array}{|c|c|}\hline 1 & 3 \\ \hline 2 & \multicolumn{1}{c}{} \\ \cline{1-1} 4 & \multicolumn{1}{c}{} \\ \cline{1-1}\end{array}$$

The algorithm assigns to $(1-+1)$ the signed tableau

$$\begin{array}{|c|c|}\hline - & + \\ \hline + & \multicolumn{1}{c}{} \\ \cline{1-1} - & \multicolumn{1}{c}{} \\ \cline{1-1}\end{array}$$

Example. Assume $G_{\mathbb{R}} = U(5,4)$, and consider the K-orbit parametrized by the clan $(+\,1+2\,3\,3\,2-1)$. Attach to $(+\,1+2\,3\,3\,2-1)$ the sequence

$$(1,+)(3,+)(5,6)(4,7)(8,-)(2,9).$$

We associate to $(1,+)(3,+)(5,6)$ a pair of tableaux, one a signed tableau, the other a standard tableau:

$$\begin{array}{|c|c|}\hline + & - \\ \hline + & \multicolumn{1}{c}{} \\ \cline{1-1} + & \multicolumn{1}{c}{} \\ \cline{1-1}\end{array} \qquad\qquad \begin{array}{|c|c|}\hline 1 & 5 \\ \hline 3 & \multicolumn{1}{c}{} \\ \cline{1-1} 6 & \multicolumn{1}{c}{} \\ \cline{1-1}\end{array}$$

Next we add $(4,7)$ to obtain

$$\begin{array}{|c|c|}\hline + & - \\ \hline + & - \\ \hline + & \multicolumn{1}{c}{} \\ \cline{1-1} + & \multicolumn{1}{c}{} \\ \cline{1-1}\end{array} \qquad\qquad \begin{array}{|c|c|}\hline 1 & 4 \\ \hline 3 & 5 \\ \hline 6 & \multicolumn{1}{c}{} \\ \cline{1-1} 7 & \multicolumn{1}{c}{} \\ \cline{1-1}\end{array}$$

Our next goal is to include the pair $(8,-)$. This gives

$$\begin{array}{|c|c|}\hline + & - \\ \hline + & - \\ \hline + & - \\ \hline + & \multicolumn{1}{c}{} \\ \cline{1-1}\end{array} \qquad\qquad \begin{array}{|c|c|}\hline 1 & 4 \\ \hline 3 & 5 \\ \hline 6 & 8 \\ \hline 7 & \multicolumn{1}{c}{} \\ \cline{1-1}\end{array}$$

The next step is a little different. When we add the pair (2, 9), we get

$$\begin{array}{|c|c|}\hline + & - \\ \hline + & - \\ \hline + & - \\ \hline \multicolumn{1}{|c|}{+} & \\ \cline{1-1} \multicolumn{1}{|c|}{+} & \\ \cline{1-1} \multicolumn{1}{|c|}{-} & \\ \cline{1-1} \end{array} \qquad \begin{array}{|c|c|}\hline 1 & 2 \\ \hline 3 & 4 \\ \hline 5 & 8 \\ \hline \multicolumn{1}{|c|}{6} & \\ \cline{1-1} \multicolumn{1}{|c|}{7} & \\ \cline{1-1} \multicolumn{1}{|c|}{9} & \\ \cline{1-1} \end{array}$$

Theorem 2.1 [Trapa 2005; 1999, Theorem 5.6]. (1) *Garfinkle's algorithm defines a bijection between $\{\mathcal{Q} \in K/\mathfrak{B}\}$ and the set of pairs $\{(T_\pm, ST)\}$ consisting of a signed Young tableau and a standard Young tableau of the same shape.*

(2) *If $T_{\pm,\mathcal{Q}}$ is the signed tableau attached via Garfinkle's algorithm to $\mathcal{Q}$, then $T_{\pm,\mathcal{Q}}$ parametrizes $\mu(T^*_{\mathcal{Q}}(\mathfrak{B}))$.*

A distinguished set of K-orbits in $\mathfrak{B}$ that parametrizes nilpotent K-orbits.

Definition 2.2. Fix $p \geq q$ with $p+q=n$, and fix $[p_1, p_2, \ldots, p_r]$ integers with $\sum p_i = n$. Define $\mathcal{S}_{\text{dis}}$ to be the set of clans of length n satisfying the following conditions:

(1) The first p_1 components of the clan (from left to right) are of the form

$$(1 \cdots a_1\ \epsilon_1 \cdots \epsilon_1\ a_1 \cdots 1),$$

where ϵ_1 is either $+$ or $-$.

(2) Components $(c_{\sum_1^{i-1} p_k+1} \cdots c_{\sum_1^{i} p_k})$ are of the form

$$\left(\sum_1^{i-1} a_k + 1 \cdots \sum_1^{i-1} a_k + a_i\ \ \epsilon_i \cdots \epsilon_i\ \ \sum_1^{i-1} a_k + a_i \cdots \sum_1^{i-1} a_k + 1\right),$$

where ϵ_i is either $+$ or $-$.

(3) $a_1 \geq a_2 \geq \cdots \geq a_r$.

(4) $q = \sum a_i + \sum \delta_{\epsilon_j,-}$ with $\delta_{\epsilon_i,-} = 1$ if $\epsilon_i = -$ and $\delta_{\epsilon_i,-} = 0$ if $\epsilon_i = +$.

An element of $\mathcal{S}_{\text{dis}}$ is called a *distinguished clan.*

Example. The clan $(1\,2 + + + 2\,1\,3\,4 - 4\,3\,5\,5)$ is a distinguished clan. Observe that $p_1 = 7$, $p_2 = 5$, $p_3 = 2$; $a_1 = a_2 = 2$, $a_3 = 1$, and $q = 6$. The clan $(1\,2\,3\,4\,4\,3\,2\,1)$ is distinguished.

Proposition 2.3. *Fix $p \geq q > 0$ integers so that $p+q=n$. Let $[p_1 \cdots p_r]$ be a sequence of positive integers with $\sum_i p_i = n$. Denote by $\mathcal{O}^{[p_1 \cdots p_r]}$ the nilpotent*

G-orbit parametrized by a tableau with column lengths $p_1, \dots, p_r$. There is a bijection

$$\left\{ \begin{matrix} \textit{nilpotent } K\textit{-orbits } \mathcal{O}_K \textit{ such that} \\ G \cdot \mathcal{O}_K = \mathcal{O}^{[p_1 \cdots p_r]} \end{matrix} \right\} \longleftrightarrow \mathcal{S}_{\text{dis}}^{[p_1 \cdots p_r]}.$$

Proof. Let $\mathcal{O}_K$ be a nilpotent K-orbit. Assume the signed tableau that parametrizes $\mathcal{O}_K$ has columns of lengths $p_1, p_2, \dots, p_r$. By Proposition 1.15, $\mathcal{O}_K$ is completely determined by the position of $-$ signs in the first column of its corresponding signed tableau $T_\pm$. Counting the numbers of the boxes that contain a $-$ sign from top to bottom, list the position of the $-$ signs in the first column as $(t_1, t_2, \dots, t_s)$. Define

$$\begin{aligned} \ell_1 &= \#\{- \text{ signs in the first column of } T_\pm\}, \\ \ell_2 &= \#\{t_i : t_i \le p_2\}, \\ &\ \vdots \\ \ell_r &= \#\{t_i : t_i \le p_r\}. \end{aligned}$$

We assign to the nilpotent K-orbit, $\mathcal{O}_K$, a distinguished K-orbit $\mathcal{Q} \subset \mathfrak{B}$. We describe the clan $\boldsymbol{c}_\mathcal{Q}$ that identifies $\mathcal{Q}$ as follows. Write

$$\boldsymbol{c}_\mathcal{Q} = \big(c_1 \cdots c_{p_1} c_{p_1+1} \cdots c_{p_1+p_2} c_{p_1+p_2+1} \cdots c_{\sum p_i}\big).$$

The first p_1 entries of $\boldsymbol{c}_\mathcal{Q}$ are given by

$$(c_1 \cdots c_{p_1}) = \begin{cases} (1 \cdots \ell_1 + \cdots + \ell_1 \cdots 1) & \text{if } p_1 \ge 2\ell_1, \\ (1 \cdots (p_1 - \ell_1) - \cdots - (p_1 - \ell_1) \cdots 1) & \text{if } p_1 < 2\ell_1. \end{cases}$$

Note that $\ell_1 = \frac{1}{2}\#\{c_i \in \mathbb{N}\} + \#\{c_i = -\}$.

The next p_2 entries are

$$\big(c_{p_1+1} \cdots c_{p_1+p_2}\big) = \begin{cases} (a_1 \cdots a_{\ell_2} - \cdots - a_{\ell_2} \cdots a_1) & \text{if } p_2 \ge 2\ell_2, \\ (a_1 \cdots a_{p_2-\ell_2} + \cdots + a_{p_2-\ell_2} \cdots a_1) & \text{if } p_2 < 2\ell_2, \end{cases} \tag{2.4}$$

where the integers a_i are consecutive and

$$a_1 = \begin{cases} \ell_1 + 1 & \text{if } p_1 \ge 2\ell_1, \\ p_1 - \ell_1 + 1 & \text{if } p_1 < 2\ell_1. \end{cases}$$

Note that $\ell_2 = \frac{1}{2}\#\{c_i \in \mathbb{N} : p_1+1 \le i \le p_1+p_2\} + \#\{c_i = + : p_1+1 \le i \le p_1+p_2\}$. Continuing inductively we define the remaining entries in $\boldsymbol{c}_\mathcal{Q}$.

The above construction assigns to $\mathcal{O}_K$ a unique distinguished $\mathbf{c}_\mathcal{Q}$. It is easy to check that Garfinkle's algorithm attaches to $\boldsymbol{c}_\mathcal{Q}$ a pair of tableaux with the signed tableau parametrizing $\mathcal{O}_K$. By Theorem 2.1 , the orbit $\mathcal{Q}$ is such that $\mu(T^*_\mathcal{Q}\mathfrak{B}) = \mathcal{O}_K$. The definition of distinguished clan guarantees that the map from nilpotent orbits to distinguished clans is onto. □

Example. Consider the nilpotent orbit $\mathcal{O}_K$ corresponding to

$$\begin{array}{|c|c|c|c|c|c|c|}
\hline
+ & - & + & - & + & - & + \\
\hline
- & + & - & + \\
\cline{1-4}
+ & - & + \\
\cline{1-3}
+ & - \\
\cline{1-2}
\end{array}$$

We have $p_1 = p_2 = 4$, $p_3 = 3$, $p_4 = 2$, $p_5 = p_6 = p_7 = 1$ and $\ell_i = 1$ for all $1 \leq i \leq 7$. The construction described in the proof of Proposition 2.3 gives $\boldsymbol{c}_\Omega = \big(1 + + 1\, 2 - - 2\, 3 + 3\, 4\, 4 + - +\big)$. In particular the K-orbit Ω parametrized by clan $\boldsymbol{c}_\Omega$ belongs to $\mu^{-1}(\mathcal{O}_K)$.

3. The operators $T_{\alpha,\beta}$

We now describe some combinatorial tools that will play an important role in our work. Indeed, given a nilpotent K-orbit $\mathcal{O}_K$, we have defined a distinguished clan $\boldsymbol{c}_{\text{dis}}$ so that $\boldsymbol{c}_{\text{dis}} \in \mu^{-1}(\mathcal{O}_K)$. We will show in Section 4 that each $\boldsymbol{c} \in \mu^{-1}(\mathcal{O}_K)$ can be obtained from $\boldsymbol{c}_{\text{dis}}$ by applying an appropriate sequence of operators $T_{\cdot,\cdot}$. These operators are defined both at the level of standard tableaux and at the level of clans.

$T_{\alpha,\beta}$ on standard tableaux. We follow [Garfinkle 1993, Chapter 3] and we let T be a standard tableau.

Definition 3.1. We say that a root $\alpha_i = \epsilon_i - \epsilon_{i+1}$ is in the τ-invariant of T if the box in T labeled i lies on a row above that containing the box labeled $i + 1$.

Example. The τ-invariant of

$$T = \begin{array}{|c|c|}
\hline
1 & 5 \\
\hline
2 & 6 \\
\hline
3 & 7 \\
\hline
4 & 8 \\
\hline
9 & 11 \\
\hline
10 \\
\cline{1-1}
\end{array}$$

is $\tau(T) = \{\alpha_1, \alpha_2, \alpha_3, \alpha_5, \alpha_6, \alpha_7, \alpha_8, \alpha_9\}$.

Definition 3.2. Given $\alpha = \epsilon_i - \epsilon_{i+1}$ and $\beta = \epsilon_{i+1} - \epsilon_{i+2}$, we say that T is in $D_{\alpha,\beta}$, the domain of $T_{\alpha,\beta}$, if $\alpha \notin \tau(T)$ and $\beta \in \tau(T)$. This is the case when either (a) the row containing label $i + 2$ is below the row containing label i, which in turn is equal to or below the row that contains $i + 1$ or (b) the row containing label $i + 1$ is above the row containing label i, which in turn is equal to the row that contains $i + 2$. We define

$$\begin{aligned} T_{\alpha,\beta} : D_{\alpha,\beta} &\to D_{\beta,\alpha}, \\ T &\mapsto T_{\alpha,\beta}(T), \end{aligned}$$

by switching the labels $i+1$ and $i+2$ in case (a) and by switching the labels i and $i+1$ in case (b).

Remark 3.3. The above definition is extended to the case $\beta = \alpha_{i-1} = \epsilon_{i-1} - \epsilon_i$ in the obvious manner. We often use the abbreviated notation $T_{i,j}$ for T_{α_i,α_j}.

Example. The operator $T_{4,5}$ maps the tableau

$$T = \begin{array}{|c|c|}\hline 1 & 5 \\ \hline 2 & 6 \\ \hline 3 & 7 \\ \hline 4 & 8 \\ \hline 9 & 11 \\ \hline 10 \\ \cline{1-1}\end{array}$$

to the tableau

$$\begin{array}{|c|c|}\hline 1 & 4 \\ \hline 2 & 6 \\ \hline 3 & 7 \\ \hline 5 & 8 \\ \hline 9 & 11 \\ \hline 10 \\ \cline{1-1}\end{array}$$

Theorem 3.4 [Vogan 1979]. *Fix λ a partition of n and denote by S_λ the set of standard tableaux of a fixed shape λ. The operators $T_{\alpha,\beta}$ act transitively on S_λ.*

$T_{\alpha,\beta}$ on clans. In this subsection we introduce the notion of τ-invariant on clans and define operations $T_{\alpha,\beta}$ on clans. These notions are not new. The work of Borho, Jantzen and Duflo established the important invariant of an irreducible representation, its τ-invariant. This is a subset of simple roots defined in terms of wall-crossing. As part of an important study of wall-crossing, [Speh and Vogan 1980] and [Vogan 1979] give formulas for the τ-invariant of a representation and related $T_{\alpha,\beta}$ in terms of $\mathbb{Z}_2$-data (in type A, $\mathbb{Z}_2$-data can be interpreted as clan-data). Our combinatorial description of τ-invariant and $T_{\alpha,\beta}$-operations on clans agrees with the work in [Speh and Vogan 1980].

Definition 3.5. Let $\boldsymbol{c} = (c_1, c_2, \dots, c_n)$ be a clan. We define the τ-invariant of $\boldsymbol{c}$ as

$$\begin{aligned}\bigl\{\epsilon_i - \epsilon_{i+1} : \;& (c_i, c_{i+1}) \text{ is a pair of equal signs,}\\ & (c_i, c_{i+1}) \text{ is a pair of equal numbers,}\\ & (c_i, c_{i+1}) = (\pm, a) \text{ so that there is } j < i \text{ with } c_j = a \in \mathbb{N},\\ & (c_i, c_{i+1}) = (a, \pm) \text{ so that there is } j > i+1 \text{ with } c_j = a \in \mathbb{N},\\ & (c_i, c_{i+1}) = (a, b) \text{ so that there are } j < k \text{ with } c_j = b, c_k = a \in \mathbb{N}\bigr\}.\end{aligned}$$

Remark 3.6. At the Lie-algebra level, each clan determines a Borel subalgebra

$$\mathfrak{b}_{\boldsymbol{c}} = \mathfrak{h}_{\boldsymbol{c}} \oplus \mathfrak{n}_{\boldsymbol{c}} \subset \mathfrak{g}.$$

The parametrization of K-orbits in G/B via clans is arranged to have the following property: there is a unique automorphism of $\mathfrak{g}$ carrying $\mathfrak{b}_{\boldsymbol{c}}$ to the Borel $\mathfrak{b} = \mathfrak{h} \oplus \mathfrak{n}$ of equation (1.1). Using such an automorphism, one can keep track of the action of θ on $\Delta(\mathfrak{n}_{\boldsymbol{c}})$. In particular if $\alpha \in \Delta(\mathfrak{h}_{\boldsymbol{c}}, \mathfrak{n}_{\boldsymbol{c}})$ corresponds to $\epsilon_i - \epsilon_{i+1}$ via the mentioned automorphism, then $\theta(\alpha)$ corresponds to

$$\begin{cases} \epsilon_i - \epsilon_k & \text{if } c_i \text{ is a sign and } c_{i+1} = c_k \in \mathbb{N}, \\ \epsilon_k - \epsilon_{i+1} & \text{if } c_{i+1} \text{ is a sign and } c_i = c_k \in \mathbb{N}, \\ \epsilon_k - \epsilon_\ell & \text{if } c_i = c_k \in \mathbb{N} \text{ and } c_{i+1} = c_\ell \in \mathbb{N}, \\ \epsilon_i - \epsilon_j & \text{if } c_i, c_j \text{ are signs.} \end{cases}$$

We say that $\alpha \in \Delta(\mathfrak{n}_{\boldsymbol{c}})$ corresponding to $\epsilon_i - \epsilon_{i+1}$ is

$$\begin{cases} \text{imaginary compact} & \text{if } (c_i, c_{i+1}) \text{ is a pair of equal signs,} \\ \text{imaginary noncompact} & \text{if } (c_i, c_{i+1}) \text{ is a pair of distinct signs,} \\ \text{real} & \text{if } (c_i, c_{i+1}) \text{ is a pair of equal numbers,} \\ \text{complex} & \text{otherwise.} \end{cases}$$

We write $\boldsymbol{i}_n$ for imaginary noncompact roots, $\boldsymbol{i}_c$ for imaginary compact roots, and $\boldsymbol{r}$ for real roots. For α, a positive complex root with $\theta(\alpha) > 0$, we write $\mathbb{C}^+$. For α, a positive complex root with $\theta(\alpha) < 0$, we write $\mathbb{C}^-$.

Hence, the τ-invariant of clan $\boldsymbol{c}$ is

$$\tau(\boldsymbol{c}) = \{\text{simple roots } \alpha \in \Delta(\mathfrak{n}_{\boldsymbol{c}}) : \alpha \text{ is } \boldsymbol{i}_c \text{ or } \boldsymbol{r} \text{ or } \mathbb{C}^-\}.$$

In order to define the combinatorial $T_{\alpha,\beta}$-action on clans we introduce a technical definition.

Definition 3.7. Let $\boldsymbol{c}$ be a clan, and write $\mathfrak{b}_{\boldsymbol{c}} = \mathfrak{h}_{\boldsymbol{c}} \oplus \mathfrak{n}_{\boldsymbol{c}}$ for the corresponding Borel subalgebra. Write ϵ for a sign (could be $+$ or $-$). Let $\alpha_i \in \Delta(\mathfrak{n}_{\boldsymbol{c}})$, where α_i corresponds to $\epsilon_i - \epsilon_{i+1}$.

(1) If α_i is imaginary noncompact ($\boldsymbol{i}_n$), we define the Cayley map

$$\mathrm{Cay}_i(c_1 \cdots c_i{=}\epsilon \;\; c_{i+1}{=}-\epsilon \cdots c_n) = (c_1 \cdots c_i{=}1 \;\; c_{i+1}{=}1 \cdots c_n).$$

(2) If α_i is real ($\boldsymbol{r}$), we define the inverse Cayley map

$$\begin{aligned} &\mathrm{Cay}_i^{-1}(c_1 \cdots c_i{=}1 \;\; c_{i+1}{=}1 \cdots c_n) \\ &\qquad = \big\{(c_1 \cdots c_i{=}+ \;\; c_{i+1}{=}- \cdots c_n); (c_1 \cdots c_i{=}- \;\; c_{i+1}{=}+ \cdots c_n)\big\}. \end{aligned}$$

(3) If α_i is complex ($\mathbb{C}^+$), the $\theta(\alpha_i)$ corresponds to $\epsilon_j - \epsilon_k$ with $j<k$. We define the cross-action $s_i \times \boldsymbol{c}$ as

$$s_i \times (c_1 \cdots c_i{=}\epsilon \;\; c_{i+1}{=}a \cdots a \cdots c_n) = (c_1 \cdots c_i{=}a \;\; c_{i+1}{=}\epsilon \cdots a \cdots c_n),$$
$$s_i \times (c_1 \cdots a \cdots c_i{=}a \;\; c_{i+1}{=}\epsilon \cdots c_n) = (c_1 \cdots a \cdots c_i{=}\epsilon \;\; c_{i+1}{=}a \cdots c_n),$$
$$s_i \times (c_1 \cdots c_i{=}a \;\; c_{i+1}{=}b \cdots c_n) = (c_1 \cdots c_i{=}b \;\; c_{i+1}{=}a \cdots c_n)$$

for any clan $\boldsymbol{c}$ with the companion of a to the left of the companion of b.

(4) If α_i is complex ($\mathbb{C}^-$), the $\theta(\alpha_i)$ corresponds to $\epsilon_j - \epsilon_k$ with $j>k$. We define the cross-action

$$s_i \times (c_1 \cdots a \cdots c_i{=}\epsilon \;\; c_{i+1}{=}a \cdots c_n) = (c_1 \cdots a \cdots c_i{=}a \;\; c_{i+1}{=}\epsilon \cdots c_n),$$
$$s_i \times (c_1 \cdots c_i{=}a \;\; c_{i+1}{=}\epsilon \cdots a \cdots c_n) = (c_1 \cdots c_i{=}\epsilon \;\; c_{i+1}{=}a \cdots c_n),$$
$$s_i \times (c_1 \cdots c_i{=}a \;\; c_{i+1}{=}b \cdots c_n) = (c_1 \cdots c_i{=}b \;\; c_{i+1}{=}a \cdots c_n)$$

for any clan with the companion of a to the right of the companion of b.

Definition 3.8. Given $\boldsymbol{c}$, a clan, we define $D^{\boldsymbol{c}}_{\alpha,\beta} = \{\text{clans} : \alpha \notin \tau(\boldsymbol{c}) \text{ and } \beta \in \tau(\boldsymbol{c})\}$, and we define $T_{\alpha,\beta} : D^{\boldsymbol{c}}_{\alpha,\beta} \to D^{\boldsymbol{c}}_{\beta,\alpha}$ as

$$T_{\alpha,\beta}(\boldsymbol{c}) = \begin{cases} s_\alpha \times \boldsymbol{c} & \text{if } \alpha \in \mathbb{C}^+,\ \beta \in \mathbb{C}^- \text{ and } \alpha+\beta \in \{\mathbb{C}^+, \boldsymbol{i}_n\}, \\ s_\alpha \times \boldsymbol{c} & \text{if } \alpha \in \mathbb{C}^+,\ \beta \in \boldsymbol{i}_c \text{ and } \alpha+\beta \in \mathbb{C}^+, \\ s_\alpha \times \boldsymbol{c} & \text{if } \alpha \in \mathbb{C}^+,\ \beta \in \boldsymbol{r} \text{ and } \theta(\alpha+\beta) \neq \alpha, \\ s_\beta \times \boldsymbol{c} & \text{if } \alpha \in \mathbb{C}^+,\ \beta \in \mathbb{C}^- \text{ and } \alpha+\beta \in \{\mathbb{C}^-, \boldsymbol{i}_c, \boldsymbol{r}\}, \\ s_\beta \times \boldsymbol{c} & \text{if } \alpha \in \boldsymbol{i}_n,\ \beta \in \mathbb{C}^-, \\ \mathrm{Cay}_\alpha\, \boldsymbol{c} & \text{if } \alpha \in \boldsymbol{i}_n,\ \beta \in \boldsymbol{i}_c, \\ \mathrm{Cay}_\beta^{-1}\, \boldsymbol{c} \cap D_{\beta,\alpha} & \text{if } \alpha \in \mathbb{C}^+,\ \beta \in \boldsymbol{r} \text{ and } \theta(\alpha+\beta) = \alpha. \end{cases}$$

Remark 3.9. We verify that $T_{\alpha,\beta}$ in Definition 3.8 is well-defined, i.e., $T_{\alpha,\beta}(\boldsymbol{c}) \in D^{\boldsymbol{c}}_{\beta,\alpha}$, by using the formulas given in Definition 3.7 and the definition of τ-invariant of a clan.

Compatibility of $T_{\alpha,\beta}$-actions. We have defined operators $T_{\alpha,\beta}$ both at the level of clans and of standard tableaux. In representation theoretic language these actions correspond to actions on $\mathbb{Z}_2$-data and on primitive ideals. Crucial to our work is the following theorem.

Theorem 3.10 [Garfinkle 1993, Section 4.2]. *Assume $p > q$. Let*

$$E : \{\textit{clans of signature } (p,q)\} \equiv \{\mathfrak{Q} \in K/\mathfrak{B}\} \to \{(T_\pm, ST)\},$$
$$\boldsymbol{c} \mapsto (T^{\boldsymbol{c}}_\pm, ST_{\boldsymbol{c}}),$$

be the bijection between $\{\mathfrak{Q} : K\text{-orbits on } \mathfrak{B}\}$ and pairs of equally shaped tableaux (the first one signed and the second one standard) induced by Garfinkle's algorithm.

Then if $\alpha, \beta \in D_{\alpha,\beta}(\text{clan } \boldsymbol{c})$, then $\alpha, \beta \in D_{\alpha,\beta}(ST_{\boldsymbol{c}})$. Moreover,

$$E(T_{\alpha,\beta}\boldsymbol{c}) = (T_{\pm}^{\boldsymbol{c}}, T_{\alpha,\beta}(ST_{\boldsymbol{c}})).$$

Remark 3.11. Each clan $\boldsymbol{c}$ determines an orbit $\mathcal{Q} \in \mathfrak{B}$. Via the Beilinson–Bernstein classification, such a $\mathcal{Q}$ determines an irreducible Harish-Chandra module with trivial infinitesimal character, $X(\boldsymbol{c}) = X(\mathcal{Q})$. By [Trapa 2005, Theorem 5.6], $T_{\pm}^{\boldsymbol{c}}$ parametrizes the associated variety of $X(\mathcal{Q})$ (which, under our assumptions, agrees with $\mu(T_{\boldsymbol{c}}^{*}\mathfrak{B})$. A result by Vogan guarantees that $T_{\alpha,\beta}$ preserves associated variety. Hence it preserves signed tableaux.

4. Characterization of $\mu^{-1}(\mathcal{O}_K)$

In this section we identify K-orbits on $\mathfrak{B}$ with their clan parametrization. Then, we freely write "τ-invariant of $\mathcal{Q}$" meaning the τ-invariant of the associated clan, as given in Section 3. Similarly we write "$T_{\alpha,\beta}$ of an orbit", meaning the corresponding action on clans. Theorem 4.3 gives a combinatorial description of the set $\mu^{-1}(\mathcal{O}_K)$. Theorem 4.5 implies that the combinatoric in Theorem 4.3 is independent of the real form.

Definition 4.1. Given $\boldsymbol{c}, \boldsymbol{c}'$ two clans parametrizing K-orbits $\mathcal{Q}, \mathcal{Q}' \in \mathfrak{B}$, we write $\mathcal{Q} \mapsto \mathcal{Q}'$ if there exist simple adjacent roots α, β with $\alpha \notin \tau(\boldsymbol{c})$, $\beta \in \tau(\boldsymbol{c})$ so that $T_{\alpha,\beta}\boldsymbol{c} = \boldsymbol{c}'$. We say that $\mathcal{Q}$ and $\mathcal{Q}'$ are *τ-linked* if there exists a sequence $(\mathcal{Q}_0, \mathcal{Q}_1, \dots, \mathcal{Q}_r)$ of K-orbits on $\mathfrak{B}$ so that $\mathcal{Q}_0 = \mathcal{Q}$, $\mathcal{Q}_r = \mathcal{Q}'$ and $\mathcal{Q}_0 \mapsto \mathcal{Q}_1 \mapsto \cdots \mapsto \mathcal{Q}_r$.

Lemma 4.2. *The τ-linked relation on the set $K/\mathfrak{B}$ is an equivalence relation.*

Proof. The lemma holds since in type A the operators $T_{\alpha,\beta}$ are injective. □

Theorem 4.3. *Let $\mathcal{O}_K$ be a nilpotent K-orbit. Then, $\mathcal{Q}, \mathcal{Q}' \in \mu^{-1}(\mathcal{O}_K)$ if and only if $\mathcal{Q}$ and $\mathcal{Q}'$ are τ-linked.*

Proof. By Theorem 2.1, two orbits $\mathcal{Q}, \mathcal{Q}'$ belong to $\mu^{-1}(\mathcal{O}_K)$ if and only if $E(T_{\mathcal{Q}}^{*}\mathfrak{B}) = (T_{\pm}^{\mathcal{Q}}, ST_{\mathcal{Q}})$ and $E(T_{\mathcal{Q}'}^{*}\mathfrak{B}) = (T_{\pm}^{\mathcal{Q}'}, ST_{\mathcal{Q}'})$ have $T_{\pm}^{\mathcal{Q}} = T_{\pm}^{\mathcal{Q}'}$. On the other hand, by Theorem 3.4 there exists a sequence $\{T_{\alpha_i,\beta_i}\}$ so that $ST_{\mathcal{Q}'} = T_{\alpha_r,\beta_r} \circ \cdots \circ T_{\alpha_1,\beta_1} ST_{\mathcal{Q}}$. Now the theorem follows from Theorem 3.10. □

Definition 4.4. Fix a partition $[r_1, r_2, \dots, r_k]$ of $n = p + q$. Define a τ-graph of standard tableaux of shape $[r_1, r_2, \dots, r_k]$ as follows. The vertices of the graph are the standard tableaux of shape $[r_1, r_2, \dots, r_k]$. Two standard tableaux (T_1, T_2) are linked if there is a pair of adjacent simple roots with (α, β) with $\alpha \notin \tau(T_1)$ $\beta \in \tau(T_1)$ and $T_2 = T_{\alpha,\beta}T_1$.

Theorem 4.5. *Fix a partition $[r_1, r_2, \dots, r_k]$ of n. Let (r, t) be any pair of integers so that $r+t = n$. Let $\mathcal{O}_K$ be a nilpotent* $\mathrm{GL}(r, \mathbb{C}) \times \mathrm{GL}(s, \mathbb{C})$*-orbit with parametrizing tableau of shape $[r_1, r_2, \dots, r_k]$. Let $\boldsymbol{c}$ be the distinguished clan associated to $\mathcal{O}_K$ as* *in Proposition 2.3. Then,* $\mu^{-1}(\mathcal{O}_K)$ *is completely determined by* $\boldsymbol{c}$ *and the* τ*-graph of standard tableaux of shape* $[r_1, r_2, \ldots, r_k]$.

Proof. The distinguished clan $\boldsymbol{c}$ parametrizes an orbit $\mathcal{Q}_0 \in \mu^{-1}(\mathcal{O}_K)$. Garfinkle's algorithm attaches to $\mathcal{Q}_0$ a pair $(T^{\boldsymbol{c}}_{\pm}, ST_{\boldsymbol{c}})$ of shape $[r_1, r_2, \ldots, r_k]$. By Theorem 4.3, $\mathcal{Q} \in \mu^{-1}(\mathcal{O}_K)$ if and only if $\mathcal{Q}$ is τ-linked to $\mathcal{Q}_0$. Since Garfinkle's map commutes with the action of operators $T_{\alpha,\beta}$, we conclude that $\mathcal{Q} \in \mu^{-1}(\mathcal{O}_K)$ if and only if the standard tableau associated to $\mathcal{Q}$ via Garfinkle's map belongs to the τ-graph of $ST_{\boldsymbol{c}}$. □

Remark 4.6. The previous theorems imply that the equivalence relation $\mathcal{Q} \simeq \mathcal{Q}'$ if and only if $\mu(T^*_{\mathcal{Q}}\mathfrak{B}) = \mu(T^*_{\mathcal{Q}'}\mathfrak{B})$ is independent of the real form $U(r, t)$ of GL($n = r + t, \mathbb{C}$).

Remark 4.7. It is important to note that the sequence of operators $\{T_{\alpha_i,\beta_i}\}$ that link two standard tableaux of the same shape is not unique. Our next example illustrates Theorem 4.5. The example concerns tableaux of shape [2, 2, 2, 1, 1]. We show that each standard tableau T of shape [2, 2, 2, 1, 1] can be obtained from

$$\begin{array}{|c|c|}\hline 1 & 6 \\ \hline 2 & 7 \\ \hline 3 & 8 \\ \hline \multicolumn{1}{|c|}{4} \\ \cline{1-1} \multicolumn{1}{|c|}{5} \\ \cline{1-1}\end{array}$$

by a sequence of $T_{i,j}$. This sequence is not unique. In Section 5, in the setting of two-column standard tableaux, we give explicit effective sequences of operators $T_{i,j}$ to generate $\mu^{-1}(\mathcal{O}_K)$.

Example. We illustrate Theorem 4.5 in an example. First we draw the τ-graph of tableaux of shape [2, 2, 2, 1, 1]. This is a connected graph. In order to fit the diagram, we have divided the graph into halves, shown in Figures 1 and 2. The tableaux on the first row of Figure 2 are indeed obtained by applying $T_{7,6}$ to appropriate tableaux listed in Figure 1.

Next we consider two different real forms, $U(5, 3)$ and $U(4, 4)$. We set

$$T_1 = \begin{array}{|c|c|}\hline + & - \\ \hline + & - \\ \hline + & - \\ \hline \multicolumn{1}{|c|}{+} \\ \cline{1-1} \multicolumn{1}{|c|}{+} \\ \cline{1-1}\end{array}, \quad T_2 = \begin{array}{|c|c|}\hline - & + \\ \hline - & + \\ \hline + & - \\ \hline \multicolumn{1}{|c|}{+} \\ \cline{1-1} \multicolumn{1}{|c|}{+} \\ \cline{1-1}\end{array}, \quad \text{and} \quad T_3 = \begin{array}{|c|c|}\hline - & + \\ \hline + & - \\ \hline + & - \\ \hline \multicolumn{1}{|c|}{+} \\ \cline{1-1} \multicolumn{1}{|c|}{-} \\ \cline{1-1}\end{array}.$$

We describe $\mu^{-1}(T_1)$, $\mu^{-1}(T_2)$ and $\mu^{-1}(T_3)$.

We start with the standard tableau

$$ST = \begin{array}{|c|c|}\hline 1&6\\\hline 2&7\\\hline 3&8\\\hline 4\\\cline{1-1} 5\\\cline{1-1}\end{array}$$

and we choose a sequence of operators $T_{\cdot,\cdot}$ that generates all standard tableaux of shape $[2, 2, 2, 1, 1]$. Next, we determine ${\boldsymbol{c}^i}_{\mathrm{dis}} \in \mu^{-1}(T_i)$ for $i = 1, 2, 3$. It is useful to observe that $E({\boldsymbol{c}^i}_{\mathrm{dis}}) = (T_i, ST)$. We show that the chosen sequence of operators $T_{\cdot,\cdot}$ allows us to describe $\mu^{-1}(T_1)$, $\mu^{-1}(T_2)$ and $\mu^{-1}(T_3)$ simultaneously when applied to $\boldsymbol{c}^i_{\mathrm{dis}}$. The example illustrates Theorem 4.5.

$$\begin{tikzcd}
\begin{array}{|c|c|}\hline 1&6\\\hline 2&7\\\hline 3&8\\\hline 4\\\cline{1-1} 5\\\cline{1-1}\end{array} \arrow[r, "T_{5,4}"] & \begin{array}{|c|c|}\hline 1&5\\\hline 2&7\\\hline 3&8\\\hline 4\\\cline{1-1} 6\\\cline{1-1}\end{array} \arrow[r, "T_{4,3}"] & \begin{array}{|c|c|}\hline 1&4\\\hline 2&7\\\hline 3&8\\\hline 5\\\cline{1-1} 6\\\cline{1-1}\end{array} \arrow[r, "T_{3,2}"] \arrow[d, "T_{6,5}"'] & \begin{array}{|c|c|}\hline 1&3\\\hline 2&7\\\hline 4&8\\\hline 5\\\cline{1-1} 6\\\cline{1-1}\end{array} \arrow[r, "T_{2,1}"] \arrow[d, "T_{6,5}"'] & \begin{array}{|c|c|}\hline 1&2\\\hline 3&7\\\hline 4&8\\\hline 5\\\cline{1-1} 6\\\cline{1-1}\end{array} \arrow[d, "T_{6,5}"'] & \\
 & \begin{array}{|c|c|}\hline 1&5\\\hline 2&6\\\hline 3&8\\\hline 4\\\cline{1-1} 7\\\cline{1-1}\end{array} & \begin{array}{|c|c|}\hline 1&4\\\hline 2&6\\\hline 3&8\\\hline 5\\\cline{1-1} 7\\\cline{1-1}\end{array} \arrow[l, "T_{4,3}"'] & \begin{array}{|c|c|}\hline 1&3\\\hline 2&6\\\hline 4&8\\\hline 5\\\cline{1-1} 7\\\cline{1-1}\end{array} \arrow[d, "T_{5,4}"'] & \begin{array}{|c|c|}\hline 1&2\\\hline 3&6\\\hline 4&8\\\hline 5\\\cline{1-1} 7\\\cline{1-1}\end{array} \arrow[d, "T_{5,4}"'] & \\
 & & \begin{array}{|c|c|}\hline 1&4\\\hline 2&5\\\hline 3&8\\\hline 6\\\cline{1-1} 7\\\cline{1-1}\end{array} & \begin{array}{|c|c|}\hline 1&3\\\hline 2&5\\\hline 4&8\\\hline 6\\\cline{1-1} 7\\\cline{1-1}\end{array} \arrow[l, "T_{4,3}"'] & \begin{array}{|c|c|}\hline 1&2\\\hline 3&5\\\hline 4&8\\\hline 6\\\cline{1-1} 7\\\cline{1-1}\end{array} \arrow[d, "T_{4,3}"'] & \\
 & & & & \begin{array}{|c|c|}\hline 1&2\\\hline 3&4\\\hline 5&8\\\hline 6\\\cline{1-1} 7\\\cline{1-1}\end{array} \arrow[r, "T_{3,2}"] & \begin{array}{|c|c|}\hline 1&3\\\hline 2&4\\\hline 5&8\\\hline 6\\\cline{1-1} 7\\\cline{1-1}\end{array}
\end{tikzcd}$$

Figure 1

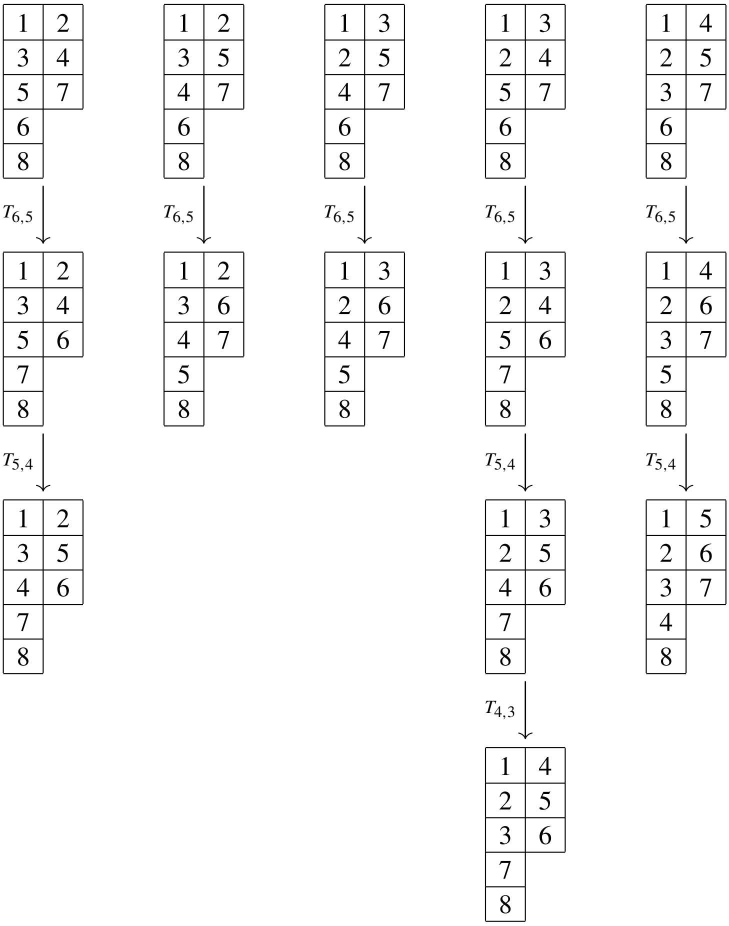


Figure 2

The GL(5, $\mathbb{C}$) × GL(3, $\mathbb{C}$)-orbits in $\mathfrak{B}$ that belong to $\mu^{-1}(T_1)$ are parametrized by the clans

$$\begin{tikzcd}
{+{+}{+}{+}{+}{-}{-}{-}} \arrow[r, "T_{5,4}"] & {+{+}{+}{+}11{-}{-}} \arrow[r, "T_{4,3}"] & {+{+}{+}1{+}1{-}{-}} \arrow[r, "T_{3,2}"] \arrow[d, "T_{6,5}"'] & {+{+}1{+}{+}1{-}{-}} \arrow[r, "T_{2,1}"] \arrow[d, "T_{6,5}"'] & {+1{+}{+}{+}1{-}{-}} \arrow[d, "T_{6,5}"'] \\
 & {+{+}{+}{+}1{-}1{-}} \arrow[r, dotted, leftrightarrow, "T_{4,3}"'] & {+{+}{+}1{+}{-}1{-}} & {+{+}1{+}{+}{-}1{-}} \arrow[d, "T_{5,4}"'] & {+1{+}{+}{+}{-}1{-}} \arrow[d, "T_{5,4}"'] \\
 & & {+{+}{+}1221{-}} & {+{+}1{+}221{-}} \arrow[l, "T_{4,3}"] & {+1{+}{+}221{-}} \arrow[d, "T_{4,3}"'] \\
 & & & & {+1{+}2{+}21{-}} \arrow[d, "T_{3,2}"'] \\
 & & & & {+{+}12{+}21{-}}
\end{tikzcd}$$

$$\begin{tikzcd}
+1+2+2-1 \arrow[d, "T_{6,5}"'] & +1++22-1 \arrow[d, "T_{6,5}"'] & ++1+22-1 \arrow[d, "T_{6,5}"'] & ++12+2-1 \arrow[d, "T_{6,5}"'] & +++122-1 \arrow[d, "T_{6,5}"'] \\
+1+2+-21 \arrow[d, "T_{5,4}"'] & +1+++--1 & ++1++--1 & ++12+-21 \arrow[d, "T_{5,4}"'] & +++1+--1 \arrow[d, "T_{5,4}"'] \\
+1++2-21 & & & ++1+2-21 \arrow[d, "T_{4,3}"'] & ++++1--1 \\
 & & & +++12-21 &
\end{tikzcd}$$

The $\mathrm{GL}(5,\mathbb{C}) \times \mathrm{GL}(3,\mathbb{C})$-orbits in $\mathfrak{B}$ that belong to $\mu^{-1}(T_2)$ are parametrized by the clans

$$\begin{tikzcd}
12+213+3 \arrow[r, "T_{5,4}"] & 12+231+3 \arrow[r, "T_{4,3}"] & 12+321+3 \arrow[r, "T_{3,2}"] \arrow[d, "T_{6,5}"'] & 123+21+3 \arrow[r, "T_{2,1}"] \arrow[d, "T_{6,5}"'] & 132+21+3 \arrow[d, "T_{6,5}"'] \\
 & 12+23+13 \arrow[r, dotted, leftrightarrow, "T_{4,3}"'] & 12+32+13 \arrow[d, "T_{5,4}"'] & 123+2+13 \arrow[d, "T_{5,4}"'] & 132+2+13 \arrow[d, "T_{5,4}"'] \\
 & & 1223++13 & 1232++13 \arrow[l, "T_{4,3}"] & 1322++13 \arrow[d, "T_{4,3}"'] \\
 & & & & 13-+++13 \arrow[d, "T_{3,2}"] \\
 & & & & 1-3+++13
\end{tikzcd}$$

$$\begin{tikzcd}
13-++1+3 \arrow[d, "T_{6,5}"'] & 1322+1+3 \arrow[d, "T_{6,5}"'] & 1232+1+3 \arrow[d, "T_{6,5}"'] & 1-3++1+3 \arrow[d, "T_{6,5}"'] & 1223+1+3 \arrow[d, "T_{6,5}"'] \\
13-+1++3 \arrow[d, "T_{5,4}"'] & 13221++3 & 12321++3 & 1-3+1++3 \arrow[d, "T_{5,4}"'] & 12231++3 \arrow[d, "T_{5,4}"'] \\
13-1+++3 & & & 1-31+++3 \arrow[d, "T_{4,3}"'] & 12213++3 \\
 & & & 1-13+++3 &
\end{tikzcd}$$

The $\mathrm{GL}(4,\mathbb{C}) \times \mathrm{GL}(4,\mathbb{C})$-orbits in $\mathfrak{B}$ that belong to $\mu^{-1}(T_3)$ are parametrized by the clans

$$\begin{tikzcd}
12+213-3 \arrow[r, "T_{5,4}"] & 12+231-3 \arrow[r, "T_{4,3}"] & 12+321-3 \arrow[r, "T_{3,2}"] \arrow[d, "T_{6,5}"'] & 123+21-3 \arrow[r, "T_{2,1}"] \arrow[d, "T_{6,5}"'] & 132+21-3 \arrow[d, "T_{6,5}"'] \\
 & 12+23-13 \arrow[r, "T_{4,3}"'] & 12+32-13 \arrow[d, "T_{5,4}"'] & 123+2-13 \arrow[d, "T_{5,4}"'] & 132+2-13 \arrow[d, "T_{5,4}"'] \\
 & & 12+3-213 & 123+-213 \arrow[l, "T_{4,3}"] & 132+-213 \arrow[d, "T_{4,3}"'] \\
 & & & & 13+2-213 \arrow[d, "T_{3,2}"] \\
 & & & & 1+32-213
\end{tikzcd}$$

$$\begin{tikzcd}
13{+}2{-}123 \arrow[d, "T_{6,5}"'] & 132{+}{-}123 \arrow[d, "T_{6,5}"'] & 123{+}{-}123 \arrow[d, "T_{6,5}"'] & 1{+}32{-}123 \arrow[d, "T_{6,5}"'] & 12{+}3{-}231 \arrow[d, "T_{6,5}"'] \\
13{+}21{-}23 \arrow[d, "T_{5,4}"'] & 132{+}1{-}23 & 123{+}1{-}23 & 1{+}321{-}23 \arrow[d, "T_{5,4}"'] & 12{+}32{-}31 \arrow[d, "T_{5,4}"'] \\
13{+}12{-}23 & & & 1{-}312{-}23 \arrow[d, "T_{4,3}"'] & 12{+}23{-}31 \\
 & & & 1{+}132{-}23 &
\end{tikzcd}$$

5. The two-column case

Explicit computations of the action of $T_{\alpha,\beta}$-operators on two-column standard tableaux.

Proposition 5.1. *Assume T is a standard tableau of shape $[2^t, 1^{r-t}]$. Further assume that T has its $b_{r,1}$ box labeled $r+\ell$ with $\ell \le t$, and has its $b_{1,2}$ box labeled j. Then, there exists a tableau $\widetilde{T}$ with $\tilde{b}_{r,1}$ labeled $r+\ell-1$ so that one of the following holds*:

(1) $\ell = 1$ *and* $T = T_{r,r-1}(\widetilde{T})$.

(2) $\ell > 1$ *and* $T = T_{r+\ell-2,r+\ell-3} \circ T_{r+\ell-1,r+\ell-2}(\widetilde{T})$.

(3) $\ell > 1$ *and* $T = T_{r+\ell-1,r+\ell}(\widetilde{T})$.

(4) *T has box $b_{\ell,2}$ labeled by an integer $k \ge j+\ell-1$, the box with label $k-1$ is on the first column, and $T = T_{k,k-1} \circ \cdots \circ T_{r+\ell-2,r+\ell-3} \circ T_{r+\ell-1,r+\ell-2}(\widetilde{T})$.*

(5) *T has box $b_{\ell,2}$ labeled by an integer $k \ge j+\ell-1$, the box with label $k-1$ is on the second column, and there is a label s with $j-1 \le s \le k-1$ so that $T = T_{s,s-1} \circ \cdots \circ T_{k-1,k-2} \circ T_{k,k-1} \circ \cdots \circ T_{r+\ell-2,r+\ell-3} \circ T_{r+\ell-1,r+\ell-2}(\widetilde{T})$.*

The proposition is proved by induction on the label of the box $b_{r,1}$ in the intersection of the last row and first column of T. As the standard tableau T has shape $[2^t, 1^{r-t}]$, the box $b_{r,1}$ is labeled by an integer of the form $r+\ell$ for some $\ell \ge 0$. For expository purposes we first prove the proposition when $\ell = 1$ and $\ell = 2$. Lemma 5.2 concerns the case $\ell = 1$. Lemma 5.3 treats the case $\ell = 2$.

Let T_o be the standard tableau of shape $[2^t, 1^{r-t}]$ with box $b_{r,1}$ labeled r and box $b_{t,2}$ labeled $r+t$.

Lemma 5.2. *Assume T is a standard tableau of shape $[2^t, 1^{r-t}]$. Further assume that T has its $b_{r,1}$ box labeled $r+1$. Then, there exists a tableau $\widetilde{T}$ with $\tilde{b}_{r,1}$ labeled r such that either*

(1) $T = T_{r,r-1}(\widetilde{T})$, *or*

(2) $T = T_{j,j-1} \circ \cdots \circ T_{r-1,r-2} \circ T_{r,r-1}(\widetilde{T})$ *for some integer $j < r$.*

Proof. T has $b_{r,1}$ labeled $r+1$. Then $b_{r-1,1}$ is either labeled $r-1$ or is labeled r. There is exactly one such tableau with $b_{r-1,1}$ labeled $r-1$. This is $T_{r,r-1}(T_o)$. Thus $\widetilde{T} = T_o$ and $T = T_{r,r-1}(T_o)$. If the label of $b_{r-1,1}$ is r, then T is of the form

$$T = \begin{array}{|c|c|}\hline 1 & j \\ \hline \cdot & r{+}2 \\ \hline \cdot & \cdot \\ \hline \cdot & \cdot \\ \hline \cdot & r{+}t \\ \hline \cdot \\ \cline{1-1} \cdot \\ \cline{1-1} r \\ \cline{1-1} r{+}1 \\ \cline{1-1} \end{array}.$$

In this case, $T = T_{j,j-1} \circ \cdots \circ T_{r-1,r-2} \circ T_{r,r-1}(T_o)$. □

Lemma 5.3. *Assume T is a standard tableau of shape $[2^t, 1^{r-t}]$. Further assume that T has its $b_{r,1}$ box labeled $r+2$.*

(1) *If $b_{1,2}$ has label r and $r+1$ is the label of $b_{2,2}$, then there exists a tableau $\widetilde{T}$ with $\tilde{b}_{r,1}$ labeled $r+1$ such that $T = T_{r,r-1} \circ T_{r+1,r}(\widetilde{T})$.*

(2) *If $b_{1,2}$ has label $j < r$ and $r+1$ is the label of $b_{2,2}$, then there exists a tableau $\widetilde{T}$ with $\tilde{b}_{r,1}$ labeled $r+1$ such that $T = T_{r+1,r}(\widetilde{T})$.*

(3) *If the label of $b_{r-1,1}$ is $r+1$, then there exists a tableau $\widetilde{T}$ with $\tilde{b}_{r,1}$ labeled $r+1$ such that*

$$T = T_{i,i-1} \circ T_{i+1,i} \circ \cdots \circ T_{r,r-1} \circ T_{r+1,r}(\widetilde{T})$$

for some integer $i < r$.

Proof. Assume first that $r+1$ is the label of $b_{2,2}$. Then $b_{1,2}$ has label j with $j \le r$. When $j \ne r$, we have

$$T = T_{r+1,r}\left(\begin{array}{|c|c|}\hline 1 & j \\ \hline \cdot & r{+}2 \\ \hline \cdot & \cdot \\ \hline \cdot & \cdot \\ \hline \cdot & r{+}t \\ \hline \cdot \\ \cline{1-1} \cdot \\ \cline{1-1} r \\ \cline{1-1} r{+}1 \\ \cline{1-1} \end{array}\right),$$

When $j = r$, we have $T = T_{r,r-1} \circ T_{r+1,r}(\widetilde{T})$, where $\tilde{b}_{1,2} = r-1$.

We next consider the tableaux T with $b_{r-1,1}$ labeled $r+1$. Observe that T is of the form

$$\begin{array}{|c|c|}
\hline
1 & j \\ \hline
\cdot & k \\ \hline
\cdot & r{+}3 \\ \hline
\cdot & r{+}4 \\ \hline
\cdot & \cdot \\ \hline
\cdot & \cdot \\ \hline
\cdot & r{+}t \\ \hline
\cdot \\ \cline{1-1}
\cdot \\ \cline{1-1}
r{+}1 \\ \cline{1-1}
r{+}2 \\ \cline{1-1}
\end{array}$$

where $k \leq j+1$.

When $k = j+1$, the tableau $T_{r-1,r} \circ \cdots \circ T_{j,j+1} \circ T_{j-1,j}(T)$ has box $b_{r,1}$ labeled $r+2$ and $b_{2,2}$ labeled $r+1$. We have $T_{r-1,r} \circ \cdots \circ T_{j,j+1} \circ T_{j-1,j}(T) = T_{r+1,r}(\widetilde{T})$, with $\widetilde{T}$ a tableau of shape $[2^t, 1^{r-t}]$ having $\tilde{b}_{r,1}$ labeled $r+1$. As the operators $T_{z,z-1}$ are injective (with inverses $T_{z-1,z}$), we have

$$T = T_{j,j-1} \circ T_{j+1,j} \circ \cdots \circ T_{r,r-1} \circ T_{r+1,r}(\widetilde{T}).$$

When $k \neq j+1$, some box in the first column of T has label $k-1$. Then, T is of the form

$$T = \begin{array}{|c|c|}
\hline
1 & j \\ \hline
\cdot & k \\ \hline
\cdot & r{+}3 \\ \hline
\cdot & r{+}4 \\ \hline
\cdot & \cdot \\ \hline
\cdot & \cdot \\ \hline
k{-}1 & \cdot \\ \hline
k{+}1 & \cdot \\ \hline
\cdot & r{+}t \\ \hline
\cdot \\ \cline{1-1}
\cdot \\ \cline{1-1}
\cdot \\ \cline{1-1}
\cdot \\ \cline{1-1}
r{+}1 \\ \cline{1-1}
r{+}2 \\ \cline{1-1}
\end{array}\,.$$

Hence, $T_{r-1,r} \circ \cdots \circ T_{k,k+1} \circ T_{k-1,k}(T)$ is a tableau with box $b_{2,2}$ labeled $r+1$. By part (2) of this lemma, we have $T_{r-1,r} \circ \cdots \circ T_{k,k+1} \circ T_{k-1,k}(T) = T_{r+1,r}(\widetilde{T})$, where $\widetilde{T}$ is a tableau of shape $[2^t,\, 1^{r-t}]$ having $\tilde{b}_{r,1}$ labeled $r+1$. We conclude that $T = T_{k,k-1} \circ T_{k+1,k} \circ \cdots \circ T_{r,r-1} \circ T_{r+1,r}(\widetilde{T})$.

Note that our argument above is independent of r and t. □

Proof of Proposition 5.1. The proof is by induction on the label of the box in the intersection of the last row first column of T. Assume T is a standard tableau of shape $[2^t,\, 1^{r-t}]$. By Lemmas 5.2 and 5.3, the proposition holds when $\ell = 1, 2$. Assume the statement of the proposition holds for any tableau of shape $[2^n,\, 1^{r-n}]$ with box $b_{r,1}$ labeled $r+m$ with $m < \ell$. We prove that the result holds for a tableau of shape $[2^t,\, 1^{r-t}]$ with box $b_{r,1}$ labeled $\ell + r$. We have two cases. Either $r+\ell-1$ occurs as a label of a box in the second column of T or $r+\ell-1$ is the label of $b_{r-1,1}$.

Assume that $r+\ell-1$ occurs as label of a box in the second column of T. Such a T is of the form

$$T = \begin{array}{|c|c|}
\hline
1 & j \\ \hline
\cdot & \cdot \\ \hline
\cdot & \cdot \\ \hline
\cdot & r{+}\ell{-}1 \\ \hline
\cdot & r{+}\ell{+}1 \\ \hline
\cdot & \cdot \\ \hline
k{-}1 & \cdot \\ \hline
k{+}1 & \cdot \\ \hline
\cdot & r{+}t \\ \hline
\multicolumn{1}{|c|}{\cdot} & \multicolumn{1}{c}{} \\ \cline{1-1}
\multicolumn{1}{|c|}{\cdot} & \multicolumn{1}{c}{} \\ \cline{1-1}
\multicolumn{1}{|c|}{\cdot} & \multicolumn{1}{c}{} \\ \cline{1-1}
\multicolumn{1}{|c|}{r{+}\ell} & \multicolumn{1}{c}{} \\ \cline{1-1}
\end{array}\,.$$

Observe that $T_{r+\ell,r+\ell-1}(T) = \widetilde{T}$ is a tableau with $\tilde{b}_{r,1}$ labeled $r+\ell-1$. Since the $T_{\cdot,\cdot}$ are injective, we conclude that $T = T_{r+\ell-1,r+\ell}(\widetilde{T})$.

If $r+\ell-1$ is the label of $b_{r-1,1}$, then T is of the form

$$T = \begin{array}{|c|c|}
\hline
1 & j \\ \hline
\cdot & \cdot \\ \hline
\cdot & \cdot \\ \hline
\cdot & k \\ \hline
\cdot & r{+}\ell{+}1 \\ \hline
\cdot & \cdot \\ \hline
k{+}1 & \cdot \\ \hline
k{+}2 & \cdot \\ \hline
\cdot & r{+}t \\ \hline
\multicolumn{1}{|c|}{\cdot} & \multicolumn{1}{c}{} \\ \cline{1-1}
\multicolumn{1}{|c|}{\cdot} & \multicolumn{1}{c}{} \\ \cline{1-1}
\multicolumn{1}{|c|}{r{+}\ell{-}1} & \multicolumn{1}{c}{} \\ \cline{1-1}
\multicolumn{1}{|c|}{r{+}\ell} & \multicolumn{1}{c}{} \\ \cline{1-1}
\end{array}$$

with $k \geq \ell - 1 + j$. Note that $k-1$ can be either in the first or in the second column.

We consider the smaller tableau

$$\widehat{T} = \begin{array}{|c|c|}\hline 1 & j \\ \hline \cdot & \cdot \\ \hline \cdot & \cdot \\ \hline \cdot & k \\ \hline \cdot \\ \cline{1-1} \cdot \\ \cline{1-1} r+\ell-1 \\ \cline{1-1}\end{array}\,.$$

By induction hypothesis there exists $\widetilde{\widehat{T}}$, with the box in the intersection of the last row and first column labeled $r+\ell-2$, so that $\widehat{T}$ is either

- $\widehat{T} = T_{k,k-1} \circ \cdots \circ T_{r+\ell-2,r+\ell-3}\big(\widetilde{\widehat{T}}\big) = \mathcal{S}_1\big(\widetilde{\widehat{T}}\big)$,
- $\widehat{T} = T_{s,s-1} \circ \cdots \circ T_{k,k-1} \circ \cdots \circ T_{r+\ell-2,r+\ell-3}\big(\widetilde{\widehat{T}}\big) = \mathcal{S}_2\big(\widetilde{\widehat{T}}\big)$ with $j-1 \leq s$, or
- $\widehat{T} = T_{r+\ell-3,r+\ell-4} \circ T_{r+\ell-2,r+\ell-3}\big(\widetilde{\widehat{T}}\big) = \mathcal{S}_3\big(\widetilde{\widehat{T}}\big)$.

In each case, $\widetilde{\widehat{T}}$ has $r+\ell-2$ occurring in the first column. Enlarge $\widetilde{\widehat{T}}$ to a tableau of shape $[2^t, 1^{r-t}]$ by adding a box with label $r+\ell$ to the first column and $t-\ell$ boxes to the end of the second column with consecutive labels $r+\ell+1$ to $r+t$. Call this new tableau $\widetilde{T}$. It is useful to note that $\widetilde{T}$ has box $\tilde{b}_{r-1,1}$ labeled $r+\ell-2$ and box $\tilde{b}_{\ell,2}$ labeled $r+\ell-1$. It follows that

$$T = \mathcal{S}_i\big(\widetilde{T}\big) \quad \text{with } i \in \{1, 2, 3\}. \tag{5.4}$$

On the other hand, as $\widetilde{T}$ has box $\tilde{b}_{r-1,1}$ labeled $r+\ell-2$ and box $\tilde{b}_{\ell,2}$ labeled $r+\ell-1$,

$$T_{r+\ell-2,r+\ell-1}\big(\widetilde{T}\big) = \widetilde{\widetilde{T}} \quad \text{with } \tilde{\tilde{b}}_{\ r,1} \text{ labeled } r+\ell-1. \tag{5.5}$$

Combining equations (5.4) and (5.5) we have that T can be obtained from $\widetilde{\widetilde{T}}$ with $\tilde{\tilde{b}}_{\ r,1}$ labeled $r+\ell-1$ by a sequence of operators $T_{\cdot,\cdot}$ as prescribed by the proposition. □

Example. Consider the standard tableau

$$T = \begin{array}{|c|c|}\hline 1 & 5 \\ \hline 2 & 6 \\ \hline 3 & 7 \\ \hline 4 & 8 \\ \hline 9 & 11 \\ \hline 10 \\ \cline{1-1}\end{array}\,.$$

We have $r=6$, $\ell=4$, and $k=8$. Observe that $k-1=7$, $k-2=6$, and $k-3=5$ are labels of boxes in the second column of T. Take $s=5$. Then

$$T = T_{5,4}\circ T_{6,5}\circ T_{7,6}\circ T_{8,7}\circ T_{9,8}\left(\begin{array}{|c|c|}\hline 1&4\\\hline 2&5\\\hline 3&6\\\hline 7&10\\\hline 8&11\\\hline 9&\\\cline{1-1}\end{array}\right).$$

The equivalence class of $+++\cdots+--\cdots-$.

Proposition 5.6. *Let $\mathcal{O}$ be the nilpotent K-orbit parametrized by a two-column tableau with length-sizes (p,q) having all boxes in the first column labeled by $+$. Assume that $\boldsymbol{c}$ is a clan that parametrizes a K-orbit in $\mu^{-1}(\mathcal{O})$. Then*:

(1) $c_1=+$.

(2) *The first p-entries of $\boldsymbol{c}$ are either $+$ signs or natural numbers.*

(3) *The last q-entries of $\boldsymbol{c}$ are either $-$ signs or natural numbers.*

(4) *If c_k is the last integer entry in $\boldsymbol{c}$, then for all $t>k$ $c_t=-$.*

(5) *If $j\le p$ and $c_j\in\mathbb{N}$, then there is exactly one $i\ge p+1$ so that $c_j=c_i$.*

(6) *If $i<j$ and (c_i,c_{p+t}) and (c_j,c_{p+s}) are pairs of equal numbers, then $s<t$.*

(7) *If $j<p$ and $c_j\in\mathbb{N}$, then $\#\{c_t\in\mathbb{N}$ with $t\le j\}\le\#\{c_t=+$ with $t<j\}$.*

Proof. We first observe that if $\boldsymbol{c}\in\mu^{-1}(\mathcal{O}_K)$, then $c_1=+$. This is an easy consequence of Garfinkle's algorithm, as otherwise the algorithm would produce a signed tableau having both a $+$ sign and a $-$ sign in the first column. Call c_j the first entry in $\boldsymbol{c}$ (counting from left to right) such that $c_j=a\in\mathbb{N}$. Let c_i be the unique entry of $\boldsymbol{c}$ with $i\neq j$ and $c_i=c_j$. Then we know that each entry $c_t\in\boldsymbol{c}$ with $t<j$ is a $+$ as otherwise the algorithm would not produce a two-column tableau. Similar considerations allow us to conclude that $i\ge p+1$ and that all entries in $\boldsymbol{c}$ with indices larger than i are $-$ signs. Hence, we can write $c_i=c_{p+\ell}$ with $\ell\ge 1$.

Our proof is by induction on ℓ. We first prove that all clans in $\mu^{-1}(\mathcal{O}_K)$ for which the last integer entry (counting from left to right) is c_{p+1} satisfy the proposition. Let $\boldsymbol{c}$ be one such clan. As $q=\#\{-$ signs in $\boldsymbol{c}\}+\#\{$pairs of equal numbers$\}$, we have

$$\boldsymbol{c}=(+\cdots+1+\cdots+1-\cdots-),\quad \text{with } c_j=c_{p+1}=1.$$

Hence, $\boldsymbol{c}$ satisfies the proposition.

Assume next that clans with last numerical entry in position $p+\ell-1$ satisfy the proposition. We prove that it is so for those clans with last numerical entry in position $p+\ell$. Let $\boldsymbol{c}_\ell$ be a clan that parametrizes an orbit $\mathcal{Q}_{\mathbf{c}_\ell}\in\mu^{-1}(\mathcal{O}_K)$

such that the last numerical entry in $\mathbf{c}_\ell$ is in position $p+\ell$. By Theorem 4.3 and Proposition 5.1, there exists an orbit $\mathcal{Q}_{\boldsymbol{c}_{\ell-1}} \in \mu^{-1}(\mathcal{O}_K)$ which is τ-linked to $\mathcal{Q}_{\boldsymbol{c}_\ell}$. In particular, $\boldsymbol{c}_\ell$ can be obtained from a clan $\boldsymbol{c}_{\ell-1}$, having its last numerical entry in position $p+\ell-1$, by an appropriate sequence of operators $T_{\cdot,\cdot}$ as prescribed by Proposition 5.1. By our induction hypothesis, clan $\boldsymbol{c}_{\ell-1}$ satisfies the proposition; that is:

(a) Each of the first p entries is either a $+$ sign or a natural number with $c_1=+$.

(b) If (c_i,c_j) is a pair of equal numbers, then $i\le p$ and $j\ge p+1$.

(c) After the last numerical entry, the clan consists of $-$ signs.

(d) For each $c_j\in\mathbb{N}$ with $j\le p$, $\#\{c_t\in\mathbb{N} \text{ with } t\le j\}\le\#\{c_t=+ \text{ with } t<j\}$.

In order to show that $\boldsymbol{c}_\ell$ also satisfies the proposition, we study the effect of the sequence of operators $T_{\cdot,\cdot}$ on $\boldsymbol{c}_{\ell-1}$. The sequence of relevant operators $T_{\cdot,\cdot}$ is that of Proposition 5.1. The first operator in the sequence is $T_{p+\ell-1,p+\ell-2}$. Since $\boldsymbol{c}_{\ell-1}\in D_{p+\ell-1,p+\ell-2}$ and it satisfies the proposition, its entries $c_{p+\ell-2}$, $c_{p+\ell-1}$, $c_{p+\ell}$ are of the form $(\cdots a\cdots b\cdots\,|\cdots b\,a\,-)$ or $(\cdots a\cdots+|\,a\,-)$. Thus, $T_{p+\ell-1,p+\ell-2}(\boldsymbol{c}_{\ell-1})$ gives $(\cdots a\cdots b\cdots\,|\cdots b-a)$ or $(\cdots a\cdots+\,|\,-a)$. All such new clans satisfy the proposition. The action of $T_{p+\ell-2,p+\ell-3}$ on one such new clan depends on its $c_{p+\ell-3}$ entry. We have the following possibilities:

$$(\cdots a\cdots b\cdots|\cdots-b-a),\qquad(\cdots a\cdots++\,|-a),\quad(\cdots a\cdots b\cdots+\,|\,b-a),$$
$$(\cdots a\cdots b\cdots c\cdots|\cdots c\,b-a),\quad(\cdots a\cdots b\,\,|\,\,b{-}a),\qquad(\cdots a\cdots b+\,|{-}b\cdots a).$$

Thus, $T_{p+\ell-2,p+\ell-3}$ applied to the clans above gives

$$(\cdots a\cdots b\cdots|\cdots b--a),\qquad(\cdots a\cdots+b\,|\,b\,a),\quad(\cdots a\cdots b\cdots+|-b\,a),$$
$$(\cdots a\cdots b\cdots c\cdots|\cdots c-b\,a),\quad(\cdots a\cdots+|--a),\quad(\cdots a\cdots+b\,|-b\cdots a).$$

The clans so produced clearly satisfy the proposition. When studying the consecutive action of $T_{\cdot,\cdot}$, as prescribed by Proposition 5.1, we need to also consider clans containing the patterns

$$(\cdots++a\cdots|\cdots a),\quad(\cdots+a+b\cdots|\cdots b-a),\quad(\cdots+a+c\,|\,c\cdots\ a),$$
$$(\cdots a\,b+c\,|\,c-b\cdots a),\quad(\cdots++a\,|\cdots a).$$

In these cases, $T_{i,i+1}$ maps the above clans to new clans containing the patterns

$$(\cdots+a+\cdots|\cdots a),\quad(\cdots++a\,b\cdots|\cdots b-a),\quad(\cdots++a\,c\,|\,c-a),$$
$$(\cdots a+b\,c\,|\,c-b\cdots a),\quad(\cdots+a+|\cdots a).$$

Conditions (1) through (6) of the proposition are clearly satisfied by these new clans. The only nonobvious conclusion is that the clans

$$\boldsymbol{c}' = T_{\cdot,\cdot}\big(\boldsymbol{c}{=}(\cdots + + a \cdots \mid \cdots a \cdots)\big) = (\cdots + a + \cdots \mid \cdots a \cdots)$$

and

$$\boldsymbol{c}' = T_{\cdot,\cdot}\big(\boldsymbol{c}{=}\cdots + + a \mid \cdots a \cdots)\big) = (\cdots + a + \mid \cdots a \cdots)$$

satisfy condition (7). Let $A = \#\{+ \text{ signs in } \boldsymbol{c} \text{ that occur to the left of } a\}$, and let $B = \#\{c_t \in \boldsymbol{c} : \text{integer entry to the left or at the position of } a\}$. By the induction hypothesis we have $B \leq A$. If $B < A$, then $\boldsymbol{c}'$ satisfies (7). We assume that $A = B$ and derive a contradiction. Write the first p-entries of $\boldsymbol{c}$ as $[+\,\gamma\ +\ +a\cdots]$. Let A_γ denote the number of $+$ signs in γ and let B_γ denote the number of integers in γ. We have $A = A_\gamma + 3 = B = B_\gamma + 1$. Hence,

$$B_\gamma = A_\gamma + 2. \tag{5.7}$$

If the last numerical entry in γ is c_{t_γ} then, as $\mathbf{c}$ satisfies (7) by the induction hypothesis,

$$B_\gamma \leq \#\{+ \text{ signs to the left of } c_{t_\gamma}\}. \tag{5.8}$$

On the other hand,

$$\begin{aligned} A_\gamma = \#\{&+ \text{ signs in } \boldsymbol{c} \text{ to the left of } c_{t_\gamma}\} \\ &+ \#\{+ \text{ signs in } \gamma \text{ occurring to the right of } c_{t_\gamma}\} - 1. \end{aligned} \tag{5.9}$$

Combining the identities in (5.7) and (5.9) with the inequality (5.8), we obtain

$$\begin{aligned} \#\{+ \text{ signs to the left of } c_{t_\gamma}\} + \#\{+ \text{ signs in } \gamma \text{ occurring to the right of } c_{t_\gamma}\} + 1 \\ \leq \#\{+ \text{ signs to the left of } c_{t_\gamma}\}. \end{aligned} \tag{5.10}$$

As inequality (5.10) cannot hold, we conclude that $A < B$. □

Corollary 5.11. *Let $\mathcal{O}_K$ be the nilpotent K-orbit parametrized by a two-column tableau with length-sizes (p, q) having all boxes in the first column labeled by $+$. Assume that $\boldsymbol{c}$ is a clan that parametrizes a K-orbit in $\mu^{-1}(\mathcal{O}_K)$. Then,*

$$0 \leq \#\{\textit{ pairs of equal numbers in } \boldsymbol{c}\} \leq \min\big\{\big[\tfrac{1}{2}p\big], q\big\}.$$

Proof. Garfinkle's algorithm assigns to $\boldsymbol{c}$ a signed tableau and a standard tableau. The algorithm is such that each pair of equal numbers in $\boldsymbol{c}$ produces a $-$ sign in the corresponding signed tableau. Hence, under our assumptions

$$\#\{\text{pairs of equal numbers in } \boldsymbol{c}\} \leq q.$$

On the other hand, part (7) of Proposition 5.6 implies

$$\#\{\text{pairs of equal numbers in } \boldsymbol{c}\} \leq \big[\tfrac{1}{2}p\big].$$

The corollary follows. □

On $\mu^{-1}(\mathcal{O}_K)$ for orbits $\mathcal{O}_K$ parametrized by a two-column signed tableau. A bijection between the set of nilpotent K-orbits and a set consisting of distinguished clans is exhibited in Proposition 2.3. In this subsection we give the explicit parametrization of nilpotent K-orbits in terms of clans in the two-column case. We introduce some notation. We consider two-column tableaux with column lengths (r,t) with $r+t=p+q=n$. Set

$$L_1 = \#\{- \text{ signs in the first column}\}, \tag{5.12}$$

$$L_2 = \#\{+ \text{ signs in the second column}\}. \tag{5.13}$$

Proposition 5.14. *Let $\mathcal{O}_K$ be a nilpotent K-orbit. Assume that the signed tableau parametrizing $\mathcal{O}$ has two columns. Then $\mu^{-1}(\mathcal{O}_K)$ contains the K-orbit $Q_{\boldsymbol{c}}$ in $\mathcal{B}$ for exactly one of the following*:

(1) $\boldsymbol{c}=\big(12\cdots r-L_1-\cdots-r-L_1\cdots 1r+1\cdots r+t-L_2+\cdots+r+t-L_2\cdots r+1\big)$, *with* $L_1\geq\left[\frac{r}{2}\right]$, $L_2\geq\left[\frac{t}{2}\right]$.

(2) $\boldsymbol{c}=\big(1\,2\cdots r-L_1-\cdots-r-L_1\cdots 1\,r+1\cdots r+L_2-\cdots-r+L_2\cdots r+1\big)$, *with* $L_1\geq\left[\frac{r}{2}\right]$, $L_2\leq\left[\frac{t}{2}\right]$.

(3) $\boldsymbol{c}=\big(1\,2\cdots L_1+\cdots+L_1\cdots 1\,r+1\cdots r+t-L_2+\cdots+r+t-L_2\cdots r+1\big)$, *with* $L_1\leq\left[\frac{r}{2}\right]$, $L_2\geq\left[\frac{t}{2}\right]$.

(4) $\boldsymbol{c}=\big(1\,2\cdots L_1+\cdots+L_1\cdots 1\,r+1\cdots r+L_2-\cdots-r+L_2\cdots r+1\big)$, *with* $L_1\leq\left[\frac{r}{2}\right]$, $L_2\leq\left[\frac{t}{2}\right]$.

Proof. The proposition follows from Proposition 2.3 and Garfinkle's algorithm. □

Proposition 5.15. *Keep the notation just introduced. Assume $\boldsymbol{c}\in\mu^{-1}(\mathcal{O}_K)$, and let $N_{\boldsymbol{c}}=\#\{$pairs of equal numbers in $\boldsymbol{c}\}$. Then one has the following*:

(1) *If* $L_1\geq\left[\frac{r}{2}\right]$, $L_2\geq\left[\frac{t}{2}\right]$, *and*

$$M=\min\left\{\left[\tfrac{1}{2}\max\{2L_1-r,2L_2-t\}\right],\min\{2L_1-r,2L_2-t\}\right\},$$

then for each integer k with

$$n-(L_1+L_2)\leq k\leq n-(L_1+L_2)+M,$$

there exists a clan $\boldsymbol{c}_k\in\mu^{-1}(\mathcal{O}_K)$ so that $N_{\boldsymbol{c}_k}=k$.

(2) *If* $L_1\leq\left[\frac{r}{2}\right]$, $L_2\leq\left[\frac{t}{2}\right]$, *and*

$$M=\min\left\{\left[\tfrac{1}{2}\max\{r-2L_1,t-2L_2\}\right],\min\{r-2L_1,t-2L_2\}\right\},$$

then for each integer k with

$$L_1+L_2\leq k\leq(L_1+L_2)+M,$$

there exists a clan $\boldsymbol{c}_k\in\mu^{-1}(\mathcal{O}_K)$ so that $N_{\boldsymbol{c}_k}=k$.

(3) *If* $L_1 \leq \left[\frac{r}{2}\right]$ *and* $L_2 \geq \left[\frac{t}{2}\right]$, *then for each integer* k *with*

$$t - L_2 \leq k \leq t - L_2 + L_1,$$

there exists a clan $\boldsymbol{c}_k \in \mu^{-1}(\mathcal{O}_K)$ *so that* $N_{\boldsymbol{c}_k} = k$.

(4) *If* $L_1 \geq \left[\frac{r}{2}\right]$ *and* $L_2 \leq \left[\frac{t}{2}\right]$, *then for each integer* k *with*

$$L_2 \leq k \leq r - L_1 + L_2,$$

there exists a clan $\boldsymbol{c}_k \in \mu^{-1}(\mathcal{O}_K)$ *so that* $N_{\boldsymbol{c}_k} = k$.

Proof. We prove that (2) holds. Statements (1), (3), and (4) can be proved using similar arguments. By Proposition 5.6 it is enough to show that clans of the form

$$\big(a_1\, b_1\, b_2 \cdots b_{L_2}\, a_2 \cdots a_{L_1} + \cdots + - - \cdots - a_{L_1} \cdots a_1\, b_{L_2} \cdots b_1\big) \tag{5.16}$$

are in $\mu^{-1}(\mathcal{O}_K)$.

We start by observing that Proposition 5.14 guarantees that $\mu^{-1}(\mathcal{O}_K)$ contains the clan

$$\boldsymbol{c} = \big(a_1\, a_2 \cdots a_{L_1} + \cdots + a_{L_1} \cdots a_1\, b_1 \cdots b_{L_2} - \cdots - b_{L_2} \cdots b_1\big).$$

By Theorem 4.3, the proposition is settled once an appropriate sequence of operators $T_{\cdot,\cdot}$, when applied to $\boldsymbol{c}$, produces clans of the desired shape.

Clan $\boldsymbol{c}$ is in the domain of $T_{r,r-1}$. Hence, by Theorem 4.3, $T_{r,r-1}\boldsymbol{c} \in \mu^{-1}(\mathcal{O}_K)$. Similarly, we argue that $T_{2,1} \circ T_{3,2} \circ \cdots \circ T_{r,r-1}(\boldsymbol{c}) \in \mu^{-1}(\mathcal{O}_K)$. That is,

$$\begin{aligned} \boldsymbol{c}' &= \big(a_1 b_1 a_2 \cdots a_{L_1} + \cdots + a_{L_1} \cdots a_1 b_2 \cdots b_{L_2} - \cdots - b_{L_2} \cdots b_1\big), \\ \boldsymbol{c}'' &= \big(a_1\, b_1\, b_2 \cdots b_{L_2}\, a_2 \cdots a_{L_1} + \cdots + a_{L_1} \cdots a_1 - \cdots - b_{L_2} \cdots b_1\big) \end{aligned}$$

are clans in $\mu^{-1}(\mathcal{O}_K)$. The next operator in the sequence is $T_{r+L_2,r+L_2+1}$, which when applied to $\boldsymbol{c}''$ gives

$$\boldsymbol{c}''' = \big(a_1\, b_1\, b_2 \cdots b_{L_2}\, a_2 \cdots a_{L_1} + \cdots + - a_{L_1} \cdots a_2 - a_1 - \cdots - b_{L_2} \cdots b_1\big).$$

Next, we compute $T_{r-L_1+L_2,r-L_1+L_2-1} \circ \cdots \circ T_{r+L_2,r+L_2+1}(\boldsymbol{c}'')$ to obtain

$$\boldsymbol{c}^{iv} = \big(a_1\, b_1\, b_2 \cdots b_{L_2}\, a_2 \cdots a_{L_1} + \cdots + - a_{L_1} \cdots a_2 - \cdots - a_1\, b_{L_2} \cdots b_1\big).$$

Note that now, at “the center” of the clan we have the $+ + \cdots + -$ pattern. Further applications of similar operators yield the clan in (5.16). □

References

[Collingwood and McGovern 1993] D. H. Collingwood and W. M. McGovern, *Nilpotent orbits in semisimple Lie algebras*, Van Nostrand Reinhold, New York, 1993. MR Zbl

[Garfinkle 1993] D. Garfinkle, “The annihilators of irreducible Harish-Chandra modules for $\mathrm{SU}(p, q)$ and other type A_{n-1} groups”, *Amer. J. Math.* **115**:2 (1993), 305–369. MR Zbl

[Matsuki and Oshima 1990] T. Matsuki and T. Oshima, "Embeddings of discrete series into principal series", pp. 147–175 in *The orbit method in representation theory* (Copenhagen, 1988), edited by M. Duflo et al., Progr. Math. **82**, Birkhäuser, Boston, 1990. MR Zbl

[Speh and Vogan 1980] B. Speh and D. A. Vogan, Jr., "Reducibility of generalized principal series representations", *Acta Math.* **145**:3–4 (1980), 227–299. MR Zbl

[Trapa 1999] P. E. Trapa, "Generalized Robinson–Schensted algorithms for real groups", *Internat. Math. Res. Notices* **1999**:15 (1999), 803–834. MR Zbl

[Trapa 2005] P. E. Trapa, "Richardson orbits for real classical groups", *J. Algebra* **286**:2 (2005), 361–385. MR Zbl

[Vogan 1979] D. A. Vogan, Jr., "A generalized τ-invariant for the primitive spectrum of a semisimple Lie algebra", *Math. Ann.* **242**:3 (1979), 209–224. MR Zbl

[Yamamoto 1997] A. Yamamoto, "Orbits in the flag variety and images of the moment map for classical groups, I", *Represent. Theory* **1** (1997), 329–404. MR Zbl

Received: 2015-09-27 Revised: 2016-03-01 Accepted: 2016-07-11

leticia@math.okstate.edu *Department of Mathematics, Oklahoma State University,*
Stillwater, OK 74078, United States

nina.l.williams@okstate.edu *Department of Mathematics, Oklahoma State University,*
Stillwater, OK 74078, United States

INVOLVE 10:4 (2017)
dx.doi.org/10.2140/involve.2017.10.625

Global sensitivity analysis in a mathematical model of the renal insterstitium

Mariel Bedell, Claire Yilin Lin,
Emmie Román-Meléndez and Ioannis Sgouralis

(Communicated by Suzanne Lenhart)

The pressure in the renal interstitium is an important factor for normal kidney function. Here we develop a computational model of the rat kidney and use it to investigate the relationship between arterial blood pressure and interstitial fluid pressure. In addition, we investigate how tissue flexibility influences this relationship. Due to the complexity of the model, the large number of parameters, and the inherent uncertainty of the experimental data, we utilize Monte Carlo sampling to study the model's behavior under a wide range of parameter values and to compute first- and total-order sensitivity indices. Characteristically, at elevated arterial blood pressure, the model predicts cases with increased or reduced interstitial pressure. The transition between the two cases is controlled mostly by the compliance of the blood vessels located before the afferent arterioles.

1. Introduction

Kidneys are the core organs in the urinary system. Their principal functions are to remove metabolic waste from the blood and to regulate blood salt and water levels [Eaton et al. 2009]. Through the regulation of salt and water, kidneys also play an important role in the regulation of arterial blood pressure [Cowley 1997; Wolgast et al. 1981]. To perform these functions, each kidney adjusts the composition of the urine it produces.

Each kidney has an outer layer, called the *cortex*, and an inner layer, known as the *medulla* [Kriz and Bankir 1988]. Much of the space in these regions is filled by the functional units of the kidney, which are termed *nephrons*. Depending on the organism, each kidney contains thousands to millions of nephrons. Nephrons are responsible for the production of urine.

Kidneys contain two types of nephrons, cortical (short) and juxtamedullary (long) nephrons, each of which is surrounded by a net of capillaries. Cortical

MSC2010: primary 49Q12, 65C05, 92C30, 93A30; secondary 92C50, 92C42.
Keywords: mathematical model, sensitivity analysis, Monte Carlo, kidney, interstitium.

nephrons remain almost entirely in the cortex, while juxtamedullary nephrons extend deep into the medulla. Each nephron consists of a *glomerulus* and a *renal tubule*. Furthermore, each renal tubule consists of various permeable or impermeable segments [Eaton et al. 2009; Kriz and Bankir 1988]. Additionally, each nephron has access to a collecting duct for removal of the produced urine.

Kidneys are connected with the rest of the body by two blood vessels, the renal artery, which carries blood into the kidney, and the renal vein, which carries blood out of the kidney to recirculate into the body. In addition, urine is excreted from the body through the ureter. Blood coming from the renal artery is delivered to the afferent arterioles. A steady flow of blood coming from the afferent arteriole of a nephron is filtered in the glomerulus and flows into the renal tubule. The blood flow is maintained constant in each glomerulus by the constriction or relaxation of its afferent arteriole [Holstein-Rathlou and Marsh 1994; Sgouralis and Layton 2015]. Nearly all of the fluid that passes through the renal tubules is reabsorbed and only a minor fraction results in urine. Fluid is reabsorbed from the renal tubules in two stages: first by the renal interstitium and then by the surrounding capillaries. The processes underlying reabsorption are driven by the pressures in the interstitial spaces [Cowley 1997; Wolgast et al. 1981].

The pressures in the renal interstitium are important determinants of kidney function. There is a lack of investigations that look at the factors affecting them. We develop a computational model of the rat kidney, for which several experimental data exist, and use it to study the relationship between arterial blood pressure and interstitial fluid pressure. In addition, we study how tissue flexibility affects this relationship and how the model predictions are affected by the uncertainty of key model parameters. We model the uncertain parameters as random variables and quantify their impact using Monte Carlo sampling and global sensitivity analysis.

2. Methods

2.1. ***Model description.*** The model consists of a collection of compartments that follow the characteristic anatomy of the kidneys of mammals [Kriz and Bankir 1988; Moffat and Fourman 1963]. The compartments fall in three categories:

(i) *regions* that model the cortical and medullary interstitial spaces,

(ii) *pipes* that model the blood vessels and renal tubules, and

(iii) *spheres* that model the glomeruli.

A schematic diagram depicting the arrangement of the compartments (1–35) is shown in Figure 1 and a summary is given in Table 1. To facilitate the description of the model equations below, we use a set of nodes (c1–c32) to mark the connections of the compartments; these nodes are also included in Figure 1 and Table 1.

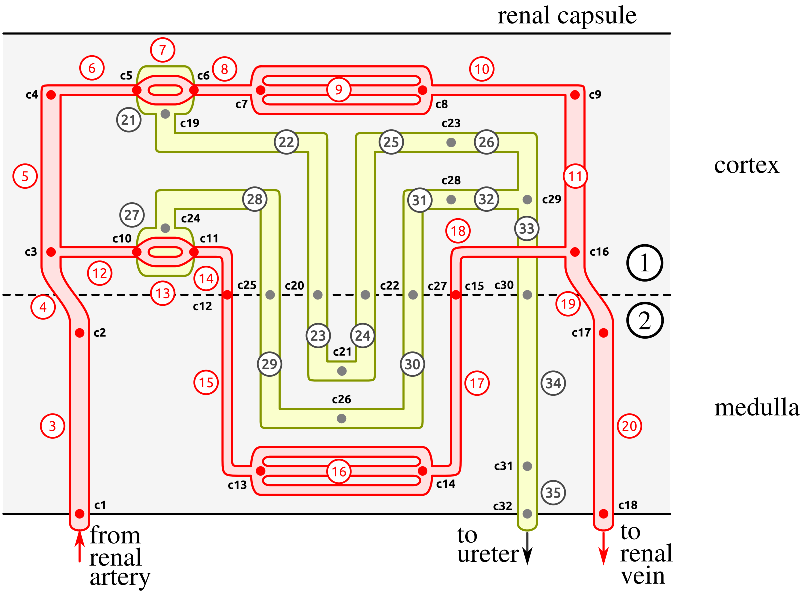


Figure 1. Schematic diagram of the model kidney. It shows the arrangement of blood vessels (red) and nephrons (yellow) within the interstitial spaces (gray). With the exceptions of the capillaries, the schematic displays only one of each of the different compartments contained in the full model. Nodes c1–c32 mark the connections of the compartments.

Briefly described, blood enters through the renal artery (node c1) and splits into a number of large arteries (compartments 3–5) that drain to the afferent arterioles (compartments 6 and 12). Each afferent arteriole supplies one glomerulus (compartments 21 and 27). In the glomeruli, blood is divided between the efferent arterioles (compartments 8 and 14) and the renal tubules (compartments 22–26 and 28–32). Leaving the efferent arterioles, blood passes through the cortical microcirculation (compartments 9 and 10) or the medullary microcirculation (compartments 15–18), before it rejoins in large veins (compartments 11, 19, and 20) and leaves through the renal vein (node c18).

The model represents short (compartments 21–26) and long nephrons (compartments 27–32) that both drain in the same collecting duct (compartments 33–35), which, in turn, drains to the ureter (node c32). The model accounts for the spacial as well as the anatomical differences between the two nephrons that are developed in the mammalian kidney [Kriz and Bankir 1988; Moffat and Fourman 1963]. For example, the model accounts for differences in the location within the cortex or medulla, in the pre- and postglomerular vascular supply, dimensions, reabsorptive capacity, etc.

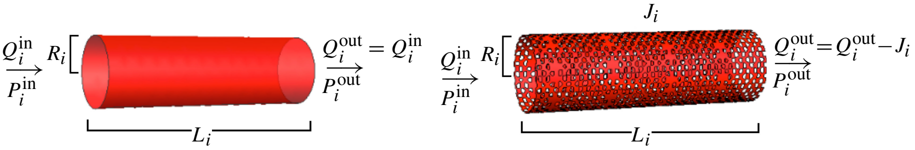


Figure 2. Model pipes: impermeable pipe, left, and permeable pipe, right.

2.1.1. *Model pipes and spheres.* Blood vessels and renal tubules are modeled as distensible pipes, while glomeruli are modeled as distensible spheres. Fluid flows through a compartment i at a volumetric rate of Q_i (Figure 2). Following the physiology, some of the pipes are considered permeable while others are impermeable [Eaton et al. 2009]. For simplicity, we assume that the only pipes modeling blood vessels that are permeable are those that model capillaries.

The flow that passes through the walls of a permeable pipe is denoted by J_i. According to the common convention, $J_i > 0$ denotes fluid leaving the pipe and $J_i < 0$ denotes fluid entering the pipe. Due to conservation of mass, the flow that leaves from an impermeable pipe Q_i^{out} is the same as the flow that enters Q_i^{in}, thus

$$Q_i^{\text{out}} = Q_i^{\text{in}}, \tag{1}$$

while the flow that leaves a permeable pipe is given by

$$Q_i^{\text{out}} = Q_i^{\text{in}} - J_i. \tag{2}$$

We assume that the flow crossing through the walls of renal tubules and glomerular capillaries is a constant fraction of the corresponding inflow

$$J_i = f_i Q_i^{\text{in}}, \tag{3}$$

where f_i is the fraction of fluid that crosses through the pipe's wall. For the coefficients f_i we use the values listed in Table 1, which are chosen such that the model predicts flows similar to the antidiuretic rat model in [Moss and Layton 2014].

Flows through the walls of the cortical and medullary capillaries are computed by the Starling equation [Wolgast et al. 1981]:

$$J_9 = K_f^9(P_9 - P_1 + \pi_9 - \pi_1), \tag{4}$$

$$J_{16} = K_f^{16}(P_{16} - P_2 + \pi_{16} - \pi_2), \tag{5}$$

where $K_f^9 = 1.59\,\mu\text{m}^3/\text{mmHg/min}$ and $K_f^{16} = 2.28\,\mu\text{m}^3/\text{mmHg/min}$ are the filtration coefficients of the cortical and medullary capillaries and π_1, π_2, π_9, and π_{16} are the oncotic pressures and P_1, P_2, P_9, and P_{15} are the hydrostatic pressures in the

i	compartment (type)	number	P_i^{int}	P_i^{ext}	nodes	frac. coeff.
1	Cortical interstitium (region)	1	P_1	-	-	-
2	Medullary interstitium (region)	1	P_2	-	-	-
3	Medullary artery (pipe)	8	P_3	P_2	c1–c2	0
4	Arcuate artery (pipe)	24	P_4	$\frac{1}{2}(P_1+P_2)$	c2–c3	0
5	Cortical radial artery (pipe)	864	P_5	P_1	c3–c4	0
6	Afferent arteriole$^{\text{sn}}$ (pipe)	20736	P_6	P_1	c4–c5	0
7	Glomerular capillary$^{\text{sn}}$ (pipe)	5598720	P_7	P_{c19}	c5–c6	$^{3}/_{28}$
8	Efferent arteriole$^{\text{sn}}$ (pipe)	20736	P_8	P_1	c6–c7	0
9	Cortical capillary (pipe)	1658880	P_9	P_1	c7–c8	see (4)
10	Venule$^{\text{sn}}$ (pipe)	20736	P_{10}	P_1	c8–c9	0
11	Cortical radial vein (pipe)	864	P_{10}	P_1	c9–c16	0
12	Afferent arteriole$^{\text{ln}}$ (pipe)	10368	P_{12}	P_1	c3–c10	0
13	Glomerular capillary$^{\text{ln}}$ (pipe)	4302720	P_{13}	P_{c24}	c10–c11	$^{3}/_{28}$
14	Efferent arteriole$^{\text{ln}}$ (pipe)	10368	P_{14}	P_1	c11–c12	0
15	Descending vas rectum (pipe)	207360	P_{15}	P_2	c12–c13	0
16	Medullary capillary (pipe)	10368000	P_{16}	P_2	c13–c14	see (5)
17	Ascending vas rectum (pipe)	414720	P_{17}	P_2	c14–c15	0
18	Venule$^{\text{sn}}$ (pipe)	10368	P_{18}	P_1	c15–c16	0
19	Arcuate vein (pipe)	24	P_{19}	$\frac{1}{2}(P_1+P_2)$	c16–c17	0
20	Medullary vein (pipe)	8	P_{20}	P_2	c17–c18	0
21	Glomerulus$^{\text{sn}}$ (sphere)	20736	P_{c19}	P_1	c19	–
22	Proximal tubule$^{\text{sn}}$ (pipe)	20736	P_{22}	P_1	c19–c20	$^{2}/_{3}$
23	Descending limb$^{\text{sn}}$ (pipe)	20736	P_{23}	P_2	c20–c21	$^{3}/_{10}$
24	Medullary ascending limb$^{\text{sn}}$ (pipe)	20736	P_{24}	P_2	c21–c22	0
25	Cortical ascending limb$^{\text{sn}}$ (pipe)	20736	P_{c24}	P_1	c22–c23	0
26	Distal tubule$^{\text{sn}}$ (pipe)	20736	P_{26}	P_1	c23–c29	$^{13}/_{84}$
27	Glomerulus$^{\text{ln}}$ (sphere)	10368	P_{24}	P_1	c24	–
28	Proximal tubule$^{\text{ln}}$ (pipe)	10368	P_{28}	P_1	c24–c25	$^{2}/_{3}$
29	Descending limb$^{\text{ln}}$ (pipe)	10368	P_{29}	P_2	c25–c26	$^{5}/_{12}$
30	Medullary ascending limb$^{\text{ln}}$ (pipe)	10368	P_{30}	P_2	c26–c27	0
31	Cortical ascending limb$^{\text{ln}}$ (pipe)	10368	P_{31}	P_1	c27–c28	0
32	Distal tubule$^{\text{ln}}$ (pipe)	10368	P_{32}	P_1	c28–c29	0
33	Cortical collecting duct (pipe)	144	P_{33}	P_1	c29–c30	$^{13}/_{84}$
34	Medullary collecting duct (pipe)	144	P_{34}	P_2	c30–c31	$^{12}/_{13}$
35	Papillary collecting duct (pipe)	8	P_{35}	P_2	c31–c32	0

Table 1. Summary of the compartments contained in the kidney model. Superscripts sn and ln denote short and long nephrons, respectively. Number refers to the total number of compartments contained in the full model.

associated compartments. The oncotic pressures are obtained by an approximation of the Landis–Pappenheimer relation

$$\pi_i = \alpha C_i + \beta C_i^2, \tag{6}$$

where $\alpha = 1.63\,\text{mmHg}\cdot\text{dl/gr}$ and $\beta = 0.29\,\text{mmHg}\cdot\text{dl}^2/\text{gr}^2$ as used in [Deen et al. 1972]. In (6), C_i denotes the concentration of protein in the compartment i. We assume a fixed protein concentration of the blood entering through the renal artery of $C_a = 5.5\,\text{gr/dl}$ and compute concentrations throughout the blood vessels (compartments 3–9 and 12–16) by taking into consideration conservation of mass:

$$C_i^{\text{out}} = \frac{Q_i^{\text{in}}}{Q_i^{\text{in}} - J_i} C_i^{\text{in}}, \tag{7}$$

where C_i^{in} and C_i^{out} denote the inflow and outflow concentrations of the compartment i. The oncotic pressures π_9 and π_{16} in (4) and (5) are computed based on the averages

$$C_9 = \tfrac{1}{2}(C_9^{\text{in}} + C_9^{\text{out}}), \tag{8}$$

$$C_{16} = \tfrac{1}{2}(C_{19}^{\text{in}} + C_{19}^{\text{out}}). \tag{9}$$

In each pipe and glomerulus, the internal pressure is denoted P_i^{int} and the external P_i^{ext}. For pipes, P_i^{int} is computed by the average of the pressures at the associated inflow and outflow nodes (Figure 1). For the glomeruli, internal pressure equals to the pressure of the associated node (Figure 1 and Table 1). For all pipes and glomerulus compartments, the external pressures equal the internal pressure of the surrounding compartment, which, in the case of the cortical and medullary regions, are denoted by P_1 and P_2, respectively. Exceptions to this are the arcuate arteries and veins (compartments 4 and 19, respectively), which anatomically are located between the cortex and the medulla [Kriz and Bankir 1988], so we compute P_i^{ext} for these compartments by the average of P_1 and P_2.

The volumes of the compartments, besides the regions and the afferent arterioles (compartments 1, 2, 6, and 12), depend *passively* on the pressure difference that is developed across their walls:

$$V_i = V_i^{\text{ref}} + s_i(P_i^{\text{int}} - P_i^{\text{ext}} + \Delta P_i^{\text{ref}}), \tag{10}$$

where V_i^{ref}, ΔP_i^{ref}, and s_i are constants. In particular, V_i^{ref} and ΔP_i^{ref} denote a reference volume and the pressure difference across the walls of the compartment when V_i equals V_i^{ref}, respectively. The parameters s_i are a measure of the distensibility of the compartments. A large s_i value indicates a compartment that is very distensible, while a low value s_i indicates a more rigid compartment. In the model, we use $s_i \geq 0$ such that an increase in P_i^{int} or a decrease in P_i^{ext} leads to an expansion of the volume V_i, and vice versa.

For a model pipe, let P_i^{in} and P_i^{out} denote the pressures at its inflow and outflow nodes, respectively. These pressures are related by a modified form of the Poiseuille law:

$$P_i^{\text{in}} - P_i^{\text{out}} = \frac{8\mu_i L_i}{\pi R_i^4}\left(Q_i^{\text{in}} - \tfrac{2}{3} J_i\right), \tag{11}$$

where μ_i is the viscosity of the flowing fluid, L_i is the length of the pipe, and R_i is its radius. In the model, we assume μ_i and L_i to be constants, while we compute R_i based on the compartment's volume (i.e., $V_i = \pi R_i^2 L_i$). Equation (11) reduces to the common Poiseuille equation for the impermeable pipes [Sgouralis and Layton 2015], while for the permeable pipes, it is assumed that J_i is linearly distributed along the length of the pipe with a value of zero at the end of the pipe.

Pressure at node c1 equals the arterial blood pressure P_a, which in our model is a free variable. Pressures at nodes c18 and c32 are kept constant at 4 mmHg and 2 mmHg, respectively, in agreement with the values of venous and ureter pressures used in previous modeling studies [Moss and Thomas 2014; Layton et al. 2012].

2.1.2. *Model afferent arterioles.* The afferent arterioles are unique vessels in the sense that they *actively* adjust radii such that blood flows through them at a fixed rate [Holstein-Rathlou and Marsh 1994; Sgouralis and Layton 2015]. In the model, we assume that blood flows in the afferent arterioles that feed the short and long nephrons (i.e., Q_6 and Q_{12}, respectively) are fixed at 280 nl/min and 336 nl/min, respectively, as in previous modeling studies of renal hemodynamics (see, for example, [Moss and Layton 2014; Fry et al. 2014; Sgouralis and Layton 2014]).

We compute the radii of the afferent arterioles by using the Poiseuille equation [Sgouralis and Layton 2015], which yields

$$R_6 = \left(\frac{8\mu_6 L_6}{\pi}\frac{Q_6}{P_{c4} - P_{c5}}\right)^{1/4}, \tag{12}$$

$$R_{12} = \left(\frac{8\mu_{12} L_{12}}{\pi}\frac{Q_{12}}{P_{c3} - P_{c10}}\right)^{1/4}. \tag{13}$$

Note that (12) and (13) imply that whenever the pressure difference along the afferent arterioles $P_{c4} - P_{c5}$ and $P_{c3} - P_{c10}$ increases, the radii R_6 and R_{12} decrease. This, in turn, implies that whenever the arterial blood pressure P_a increases, the afferent arterioles constrict, and thus the total volumes occupied by them, $V_6 = \pi R_6^2 L_6$ and $V_{12} = \pi R_{12}^2 L_{12}$, are reduced.

2.1.3. *Model interstitial regions.* The cortical and medullary interstitial spaces, i.e., compartments 1 and 2, lie outside of the compartments 3–35 and therefore must be calculated separately using a different set of equations. We obtain the first of such relationships by assuming that the net accumulation of interstitial fluid within the cortex and medulla is zero. That is,

$$J_9 + \tfrac{1}{80} J_{22} + \tfrac{1}{80} J_{26} + \tfrac{1}{160} J_{28} + \tfrac{1}{160} J_{32} = 0, \tag{14}$$

$$J_{16} + \tfrac{1}{500} J_{23} + \tfrac{1}{1000} J_{29} + \tfrac{1}{72000} J_{34} = 0, \tag{15}$$

where the flows J_i are weighted based on the total number of the compartments contained in the full model (Table 1).

Equations (4) and (5) require the oncotic pressures π_1 and π_2, which in turn require the cortical and medullary protein concentrations C_1 and C_2 for (6). Protein concentrations in the cortical and medullary regions are computed assuming that the total mass of protein contained in each region, M_1 and M_2, respectively, remains constant. Thus,

$$C_1 = M_1 / V_1, \tag{16}$$

$$C_2 = M_2 / V_2. \tag{17}$$

We use the values $M_1 = 1.93\,\text{mgr}$ and $M_2 = 1.25\,\text{mgr}$, which are computed such that the resulting model predicts reference pressures in the renal cortex and medulla of $\sim 6\,\text{mmHg}$, similar to those estimated experimentally [Cowley 1997].

Cortical and medullary interstitial volumes V_1 and V_2 are assumed to change proportionally; thus,

$$V_1 / V_2 = \kappa, \tag{18}$$

where κ is the proportionality constant. The combined volume of the interstitial regions $V_1 + V_2$ is calculated based on the total volume of the kidney V_0 according to

$$V_1 + V_2 = V_0 - V_{\text{cortex}} - V_{\text{medulla}}, \tag{19}$$

where V_{cortex} and V_{medulla} are found by summing the total volumes of the pipe and glomerulus compartments contained within each region. Finally, the total volume of the kidney V_0 is calculated by

$$V_0 = V_0^{\text{ref}} + s_0 \, (P_1 - P_0^{\text{ext}} + \Delta P_0^{\text{ref}}), \tag{20}$$

where in this case P_0^{ext} refers to the pressure external to the kidney, which is set to 0 mmHg. Equation (20) assumes that the total volume of the kidney is determined by the distensibility of the renal capsule s_0, which is stretched by the difference of the pressures developed across it, i.e., $P_1 - P_0^{\text{ext}}$.

2.2. ***Model parameters.*** Values for the model parameters are given in Table 2. These values are chosen such that at a reference arterial blood pressure $P_a^{\text{ref}} = 100\,\text{mmHg}$, the model predicts pressures and volumes that are in good agreement

i	L_i μm	μ_i	P_i^{ref} mmHg	ΔP_i^{ref} mmHg	R_i^{ref} μm	V_i^{ref} μm^3	$\tilde{\sigma}_i$	ci	$P_{\text{ci}}^{\text{ref}}$ mmHg
1	-	-	6	-	-	$7.62 \cdot 10^{10}$	-	c1	100
2	-	-	6	-	-	$4.92 \cdot 10^{10}$	-	c2	97.51
3	$7 \cdot 10^3$	μ_L	98.75	-92.75	270	$1.60 \cdot 10^9$	$\tilde{\sigma}_{G4}$	c3	95.02
4	$2 \cdot 10^3$	μ_L	96.26	-90.26	150	$1.41 \cdot 10^8$	$\tilde{\sigma}_{G4}$	c4	93.97
5	$3 \cdot 10^3$	μ_L	94.50	-88.50	75	$5.30 \cdot 10^7$	$\tilde{\sigma}_{G4}$	c5	51.17
6	300	μ_A	72.57	-66.57	10	$9.42 \cdot 10^4$	-	c6	48.08
7	80	μ_C	49.62	-37.27	4.2	$4.43 \cdot 10^3$	$\tilde{\sigma}_{G5}$	c7	14.38
8	310	μ_E	31.23	25.23	11	$1.17 \cdot 10^5$	$\tilde{\sigma}_{G5}$	c8	8.92
9	40	μ_C	11.65	-5.65	4.2	$2.21 \cdot 10^3$	$\tilde{\sigma}_{G5}$	c9	5.44
10	50	μ_L	7.17	-1.18	12	$2.26 \cdot 10^4$	$\tilde{\sigma}_{G5}$	c10	50.52
11	$3 \cdot 10^3$	μ_L	5.40	0.60	150	$2.12 \cdot 10^8$	$\tilde{\sigma}_{G5}$	c11	47.51
12	260	μ_A	72.77	-66.77	10	$8.16 \cdot 10^4$	-	c12	12.94
13	100	μ_C	49.02	-35.35	4.2	$5.54 \cdot 10^3$	$\tilde{\sigma}_{G5}$	c13	9.88
14	265	μ_E	30.22	-24.22	11	$1.00 \cdot 10^5$	$\tilde{\sigma}_{G5}$	c14	9.12
15	210	μ_E	11.41	-5.41	9	$5.34 \cdot 10^4$	$\tilde{\sigma}_{G5}$	c15	7.78
16	60	μ_C	9.50	-3.50	4.2	$3.32 \cdot 10^3$	$\tilde{\sigma}_{G5}$	c16	5.37
17	210	μ_A	8.45	-2.45	9	$5.34 \cdot 10^4$	$\tilde{\sigma}_{G5}$	c17	4.41
18	30	μ_A	6.58	-0.58	12	$1.35 \cdot 10^4$	$\tilde{\sigma}_{G5}$	c18	4
19	$2 \cdot 10^3$	μ_L	4.89	1.11	190	$2.26 \cdot 10^8$	$\tilde{\sigma}_{G5}$	c19	12.36
20	$7 \cdot 10^3$	μ_L	4.20	1.79	425	$3.97 \cdot 10^9$	$\tilde{\sigma}_{G5}$	c20	11.73
21	-	-	12.36	-6.36	80	$2.14 \cdot 10^6$	$\tilde{\sigma}_{G2}$	c21	11.30
22	$14 \cdot 10^3$	μ_N	12.04	-6.04	15	$9.89 \cdot 10^6$	$\tilde{\sigma}_{G3}$	c22	10.93
23	$2 \cdot 10^3$	μ_N	11.51	-5.51	8.5	$4.53 \cdot 10^5$	$\tilde{\sigma}_{G3}$	c23	10.79
24	$2 \cdot 10^3$	μ_N	11.12	5.11	8.5	$4.53 \cdot 10^5$	$\tilde{\sigma}_{G3}$	c24	13.66
25	$3 \cdot 10^3$	μ_N	10.86	-4.86	12	$1.35 \cdot 10^6$	$\tilde{\sigma}_{G3}$	c25	12.90
26	$5 \cdot 10^3$	μ_N	10.73	-4.73	13.5	$2.86 \cdot 10^6$	$\tilde{\sigma}_{G3}$	c26	11.76
27	-	-	13.66	-7.66	100	$4.18 \cdot 10^6$	$\tilde{\sigma}_{G2}$	c27	10.84
28	$14 \cdot 10^3$	μ_N	13.28	-7.28	55	$9.89 \cdot 10^6$	$\tilde{\sigma}_{G3}$	c28	10.79
29	$5 \cdot 10^3$	μ_N	12.33	-6.33	8.5	$1.13 \cdot 10^6$	$\tilde{\sigma}_{G3}$	c29	10.66
30	$5 \cdot 10^3$	μ_N	11.30	-5.30	8.5	$1.13 \cdot 10^6$	$\tilde{\sigma}_{G3}$	c30	6.64
31	$1 \cdot 10^3$	μ_N	10.82	-4.82	12	$4.52 \cdot 10^5$	$\tilde{\sigma}_{G3}$	c31	2.00
32	$5 \cdot 10^3$	μ_N	10.73	-4.73	13.5	$2.86 \cdot 10^6$	$\tilde{\sigma}_{G3}$	c32	2
33	$1.5 \cdot 10^3$	μ_N	8.65	-2.65	16	$1.20 \cdot 10^6$	$\tilde{\sigma}_{G3}$		
34	$4.5 \cdot 10^3$	μ_N	4.32	1.68	16	$3.61 \cdot 10^6$	$\tilde{\sigma}_{G3}$		
35	$2.5 \cdot 10^3$	μ_N	2.00	4.00	2.3	$4.15 \cdot 10^{10}$	$\tilde{\sigma}_{G1}$		

Table 2. Parameter and reference values for the model compartments (indexed by i) and nodes (indexed by ci). Values for viscosity and flexibility are given in Table 3.

viscosity values (min·mmHg)	flexibility values (mmHg^{-1})
$\mu_L = 6.4 \cdot 10^{-7}$	$\tilde{\sigma}_{G1} = 0.002$
$\mu_A = 2 \cdot 10^{-6}$	$\tilde{\sigma}_{G2} = 0.005$
$\mu_E = 2.5 \cdot 10^{-6}$	$\tilde{\sigma}_{G3} = 0.045$
$\mu_C = 4.9 \cdot 10^{-6}$	$\tilde{\sigma}_{G4} = 0.004$
$\mu_N = 5.4 \cdot 10^{-8}$	$\tilde{\sigma}_{G5} = 0.065$

Table 3. Viscosity and flexibility values.

with either direct experimental measurements [Nyengaard 1993; Nordsletten et al. 2006; Jensen and Steven 1977; Heilmann et al. 2012; Cortes et al. 1996] or previous modeling studies [Moss and Layton 2014; Moss and Thomas 2014; Edwards and Layton 2011; Sgouralis and Layton 2012; 2013; 2016; Oien and Aukland 1991; Sgouralis et al. 2015; Chen et al. 2011].

The pressure-volume relationships used in the model, (10) and (20), require values for the parameters s_i. We assume that

(i) s_i scale proportionally to the reference volumes

$$s_i = \sigma_i V_i^{\text{ref}}, \tag{21}$$

and

(ii) the coefficients σ_i depend only on the histology of the associated compartment.

That is, we group the compartments as follows:

*Group G*1: renal capsule (s_0) and papillary collecting duct (s_{35}),

*Group G*2: glomeruli (s_{21} and s_{27}),

*Group G*3: renal tubules (s_{22}–s_{26}) and proximal collecting ducts (s_{28}–s_{34}),

*Group G*4: preafferent arteriole blood vessels (s_3–s_5),

*Group G*5: postafferent arteriole blood vessels (s_7–s_{11} and s_{13}–s_{20}).

Then we assign the same flexibility value σ_i to all members of each group (Table 2). With this formulation, the model compartments in each histological group experience the same fractional change in volume whenever they are challenged by the same pressure gradient $P_i^{\text{int}} - P_i^{\text{ext}}$.

The available experimental data do not permit an accurate estimate of the values of the flexibility parameters. For this reason, we treat the flexibilities of the five groups σ_g as *independent random variables*. To facilitate the comparison among the different groups, we set

$$\sigma_g = \tilde{\sigma}_g \Lambda_g, \tag{22}$$

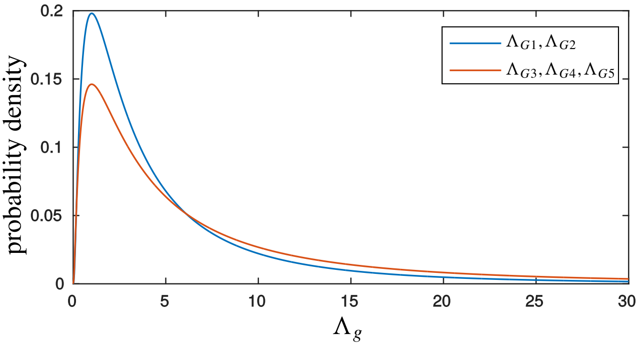


Figure 3. Probability densities of the flexibility parameters Λ_g of the histological groups G1–G5 used in this study.

where $\tilde{\sigma}_g$ are constants, and Λ_g are random variables configured to have mode 1. We estimate the values of $\tilde{\sigma}_g$ empirically based on ex vivo measurements reported in [Hebert et al. 1975; Zhu et al. 1992; Cortes et al. 1996; Cortell et al. 1973; Yamamoto et al. 1983] (Table 2).

For each simulation, Λ_g are drawn from the log-normal distribution (Figure 3), which is chosen such that

(i) s_g attain nonnegative values,

(ii) arbitrarily large values of s_g are allowed, and

(iii) low s_g values are more frequent than large ones.

We choose the latter condition assuming that the experimental procedures (anesthesia, renal decapsulation, tissue isolation, etc.) utilized in [Hebert et al. 1975; Cortes et al. 1996; Cortell et al. 1973; Yamamoto et al. 1983] likely increase rather than decrease tissue flexibility, thus our computed $\tilde{\sigma}_g$ likely overestimate rather than underestimate σ_g.

Finally, we configure the log-normal distributions such that Λ_{G1} and Λ_{G2} have a log-standard deviation of 1.1, and Λ_{G3}, Λ_{G4}, and Λ_{G5} have a log-standard deviation of 1.25 (Figure 3). According to our experience, such configuration reflects the degree of the uncertainty in our estimated values of $\tilde{\sigma}_g$, for which we consider $\tilde{\sigma}_{G3}$, $\tilde{\sigma}_{G4}$, and $\tilde{\sigma}_{G5}$ less accurately estimated than $\tilde{\sigma}_{G1}$ and $\tilde{\sigma}_{G2}$.

2.3. *Sensitivity analysis.*

2.3.1. *Formulation.* For the sensitivity analysis of the model described in the previous sections, we adopt a *variance-based method* which is best suited for nonlinear models [Saltelli et al. 2000; Sobol' 2001]. Let

$$y = f(x_1, x_2, \dots, x_k) \tag{23}$$

denote a generic model, where y is an output value and $x_1, x_2, \ldots, x_k$ are some random inputs (in our case those represent the uncertain parameters). For a factor x_g, the first- and total-order sensitivity indices are given by

$$S_g = \frac{\mathbb{V}(\mathbb{E}(y \mid x_g))}{\mathbb{V}(y)}, \tag{24}$$

$$T_g = 1 - \frac{\mathbb{V}(\mathbb{E}(y \mid x_{-g}))}{\mathbb{V}(y)}, \tag{25}$$

respectively [Saltelli et al. 2000; Saltelli 2002; Sobol' 2001]. In the equations above, $\mathbb{E}$ and $\mathbb{V}$ denote mean value and variance, respectively. In (24), first the mean of y is computed by fixing the factor x_g to some value $\tilde{x}_g$, and then the variance of the mean values is computed over all possible $\tilde{x}_g$. In (25), first the mean value is computed by fixing all factors except x_g (which is denoted by x_{-g}), and then the variance of the mean values is computed over all possible x_{-g}.

According to the above definitions, the first-order index S_g indicates the fraction by which the variance of y will be reduced if only the value of the factor x_g is certainly specified [Saltelli et al. 2000]. Similarly, the total-order index T_g indicates the fraction of the variance of y that will be left if all factors besides x_g are certainly specified [Saltelli et al. 2000]. We compute both indices, because generally for a nonlinear model the factors are expected to interact in a nonadditive way, and therefore T_g is expected to be larger than S_g. The difference $T_g - S_g$ characterizes the extent of the interactions with the other factors that x_g is involved with.

2.3.2. *Evaluation of sensitivity indices.* To better characterize the contribution of the individual factors Λ_g of (22), in the variance of P_1 and P_2, we calculate their first- and total-order sensitivity indices given in (24) and (25). We compute the indices according to the method proposed by Saltelli [2002], which is computationally less demanding than a straightforward application of the formulas (24) and (25).

Briefly, according to the Saltelli method we form two input matrices:

$$M_A = \begin{bmatrix} \Lambda_{G1}^{1,A} & \Lambda_{G2}^{1,A} & \Lambda_{G3}^{1,A} & \Lambda_{G4}^{1,A} & \Lambda_{G5}^{1,A} & \Lambda_{G6}^{1,A} \\ \Lambda_{G1}^{2,A} & \Lambda_{G2}^{2,A} & \Lambda_{G3}^{2,A} & \Lambda_{G4}^{2,A} & \Lambda_{G5}^{2,A} & \Lambda_{G6}^{2,A} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \Lambda_{G1}^{N,A} & \Lambda_{G2}^{N,A} & \Lambda_{G3}^{N,A} & \Lambda_{G4}^{N,A} & \Lambda_{G5}^{N,A} & \Lambda_{G6}^{N,A} \end{bmatrix}, \tag{26}$$

$$M_B = \begin{bmatrix} \Lambda_{G1}^{1,B} & \Lambda_{G2}^{1,B} & \Lambda_{G3}^{1,B} & \Lambda_{G4}^{1,B} & \Lambda_{G5}^{1,B} & \Lambda_{G6}^{1,B} \\ \Lambda_{G1}^{2,B} & \Lambda_{G2}^{2,B} & \Lambda_{G3}^{2,B} & \Lambda_{G4}^{2,B} & \Lambda_{G5}^{2,B} & \Lambda_{G6}^{2,B} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \Lambda_{G1}^{N,B} & \Lambda_{G2}^{N,B} & \Lambda_{G3}^{N,B} & \Lambda_{G4}^{N,B} & \Lambda_{G5}^{N,B} & \Lambda_{G6}^{N,B} \end{bmatrix} \tag{27}$$

by generating Monte Carlo samples $\Lambda_g^{j,A}$ and $\Lambda_g^{j,B}$ for the factors Λ_g. Subsequently, for each factor, we form a matrix M_g. Each M_g is formed by the columns of M_A, except the column that corresponds to the factor Λ_g, which is taken from M_B. For instance, M_{G2} is given by

$$M_{G2} = \begin{bmatrix} \Lambda_{G1}^{1,A} & \boldsymbol{\Lambda_{G2}^{1,B}} & \Lambda_{G3}^{1,A} & \Lambda_{G4}^{1,A} & \Lambda_{G5}^{1,A} & \Lambda_{G6}^{1,A} \\ \Lambda_{G1}^{2,A} & \boldsymbol{\Lambda_{G2}^{2,B}} & \Lambda_{G3}^{2,A} & \Lambda_{G4}^{2,A} & \Lambda_{G5}^{2,A} & \Lambda_{G6}^{2,A} \\ \vdots & \boldsymbol{\vdots} & \vdots & \vdots & \vdots & \vdots \\ \Lambda_{G1}^{N,A} & \boldsymbol{\Lambda_{G2}^{N,B}} & \Lambda_{G3}^{N,A} & \Lambda_{G4}^{N,A} & \Lambda_{G5}^{N,A} & \Lambda_{G6}^{N,A} \end{bmatrix}. \tag{28}$$

We use each row of the matrices M_A, M_B, and M_g to solve the model equations at $P_a = 180\,\text{mmHg}$ and combine the solutions in the vectors

$$m_A^k = \begin{bmatrix} P_k^{1,A} \\ P_k^{2,A} \\ \vdots \\ P_k^{N,A} \end{bmatrix}, \qquad m_B^k = \begin{bmatrix} P_k^{1,B} \\ P_k^{2,B} \\ \vdots \\ P_k^{N,B} \end{bmatrix}, \qquad m_g^k = \begin{bmatrix} P_k^{1,g} \\ P_k^{2,g} \\ \vdots \\ P_k^{N,g} \end{bmatrix}, \tag{29}$$

where $k = 1$ corresponds to the pressure in the cortical region P_1, and $k = 2$ to the pressure in the medullary region P_2. The first- and total-order sensitivity indices are then computed by

$$S_g^k = \frac{1/(N-1)\sum_{j=1}^N (P_k^{j,A} P_k^{j,g}) - 1/N \sum_{j=1}^N (P_k^{j,A} P_k^{j,B})}{\mathbb{V}(m_A^k)}, \tag{30}$$

$$T_g^k = 1 - \frac{1/(N-1)\sum_{j=1}^N (P_k^{j,B} P_k^{j,g}) - \left(1/N \sum_{j=1}^N P_k^{j,B}\right)^2}{\mathbb{V}(m_B^k)}, \tag{31}$$

respectively. In (30) and (31), $\mathbb{V}$ denotes the sample variance. For further details on the method, see [Saltelli 2002].

2.4. *Numerical methods.* For the numerical solution, we combine the model equations (1)–(20) into a system of 69 coupled nonlinear equations. Given a value for the arterial blood pressure P_a and a choice for the flexibility parameters Λ_g, the resulting system is solved to yield the values for the pressures at the interstitial regions P_1 and P_2, the pressures at the model nodes P_{c1}–P_{c32}, and the volumes of the compartments V_1–V_{35}.

To obtain solutions, we implement the system in MATLAB and use the standard root-finding function (fsolve). This function computes solutions to the model equations iteratively by starting from a given initial approximation. For the initial approximation we use the reference values from literature (Table 2). Note that by

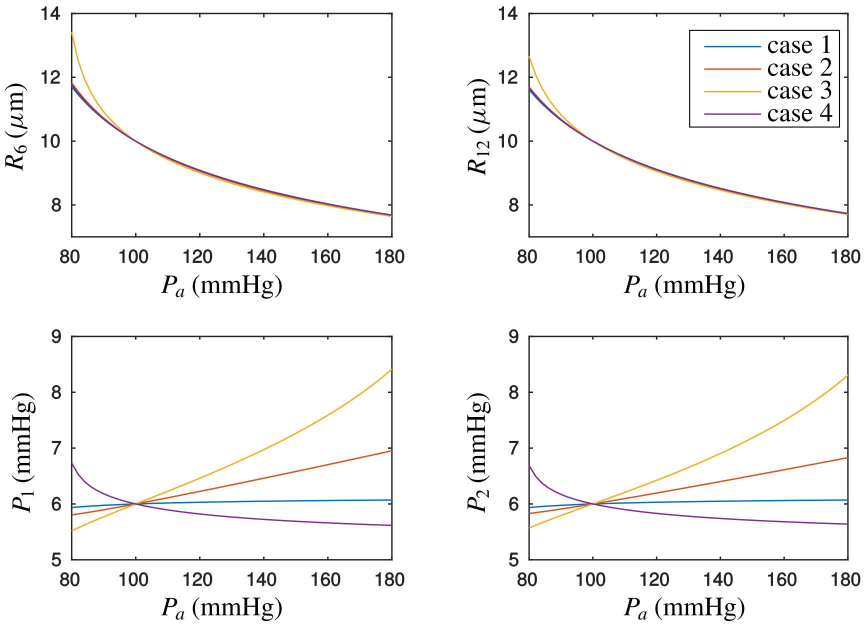


Figure 4. Model predictions for selected parameter choices. Upper panels: radii of the afferent arterioles. Lower panels: pressures in the interstitial regions.

the construction of the model, the solution at reference can be obtained trivially, and thus no root-finding is necessary for this step.

3. Results

3.1. ***Selected case studies.*** In the first set of simulations, we investigate how the pressures in the interstitial regions P_1 and P_2 are affected by the arterial blood pressure P_a for selected choices of the flexibility parameters when P_a varies in the range 80–180 mmHg. In particular, we make the following choices for the flexibility parameters:

Case 1: $\Lambda_{G1} = \Lambda_{G2} = \Lambda_{G3} = \Lambda_{G4} = \Lambda_{G5} = 0,$

Case 2: $\Lambda_{G1} = \Lambda_{G2} = \Lambda_{G3} = \Lambda_{G4} = \Lambda_{G5} = 1,$

Case 3: $\Lambda_{G1} = \Lambda_{G2} = \Lambda_{G3} = \Lambda_{G4} = \Lambda_{G5} = 4,$

Case 4: $\Lambda_{G3} = 0$ and $\Lambda_{G4} = 0.27$ and $\Lambda_{G5} = 0.2.$

Figure 4 shows key solution values.

Case 1 corresponds to a kidney with rigid compartments. In this case, pressure does not affect the volume of the compartments except for the two afferent arterioles V_6 and V_{12}. For example, at elevated P_a, the pressure differences along the afferent arterioles $P_{c4} - P_{c5}$ and $P_{c3} - P_{c10}$ increase. As a result, the arterioles constrict in order to maintain constant blood flow; see (12) and (13). Given that total kidney volume V_0, given by (20), does not change, the reduction in afferent arteriole volume increases the volume of the interstitial regions V_1 and V_2 given by (18). In turn, increases in interstitial volumes reduce the protein concentrations C_1 and C_2 by (16) and (17) and the oncotic pressures π_1 and π_2 that promote uptake J_9 and J_{16} of interstitial fluid by (4) and (5). However, due to tubular reabsorption J_{22}–J_{34}, the flow of fluid into the interstitial spaces is kept constant; see (14) and (15). Thus, in order to maintain a constant uptake and avoid accumulation of interstitial fluid, P_1 and P_2 increase. Vice versa, a decrease in P_a has the opposite effect and results in a decrease of P_1 and P_2. Because the total volume of the afferent arterioles is only a minor fraction of the volume of the interstitial regions ($\sim 2\%$, see Table 2), even large changes in R_6 and R_{12} induce small changes in π_1 and π_2. Therefore, the total change in P_1 and P_2, across the full range of P_a variation, is in the order of 0.1 mmHg (see blue curves in Figure 4).

Case 2 corresponds to a kidney with distensible compartments. This case is similar to Case 3; however, the changes in P_1 induced by the constriction of the afferent arterioles is followed by an expansion of the renal capsule (20), which increases whole kidney volume V_0. In this case, the cortical and medullary interstitial volumes V_1 and V_2 increase to a larger extent compared with Case 1 in order to accommodate the expansion of V_0. As a result, interstitial protein concentrations C_1 and C_2, and oncotic pressures π_1 and π_2 drop by larger amounts than in Case 1. Consequently, significant drops in P_1 and P_2 follow (see orange curves in Figure 4).

Case 3 corresponds to a kidney with very flexible compartments and renal capsule. Through the same effects as in Cases 1 and 2, changes in arterial pressure P_a lead to similar changes in P_1 and P_2. Because in this case the expansion of whole kidney volume V_0 is greater than in Case 2, due to the increased flexibility of the renal capsule s_0, the interstitial pressures are affected to a greater extent (see yellow curves in Figure 4).

Case 4 shows a different behavior that corresponds to a kidney with a flexible capsule but relatively rigid compartments. As in all cases, P_a affects severely the pressures in the preafferent arteriole vascular compartments P_3, P_4, and P_5 — see (11) — which are not regulated by the active constriction/dilation of the afferent arterioles. As a result, whenever P_a increases, P_3, P_4, and P_5 also increase, leading to an increase of the associated preafferent arteriole vascular volumes V_3, V_4, and V_5. Note that the increase of V_3, V_4, and V_5 opposes the reduction of V_6 and V_{12} caused by constriction of the afferent arterioles. In this particular case, opposite to what happens in Cases 1–3, the increase of the total volume of the preafferent arteriole compartments V_3, V_4, and V_5 exceeds the reduction of the total volume of the afferent arterioles V_6 and V_{12}. As a result, the interstitial regions are compressed, which in turn leads to increases of the protein concentrations C_1 and C_2 and oncotic pressures π_1 and π_2. Because the uptake of interstitial fluid is maintained constant, this leads to reductions of P_1 and P_2. Finally, the reductions of P_1 and P_2 are further amplified by constriction of the renal capsule that follows the reduction of P_1.

3.2. ***Sensitivity analysis.*** From the previous section, it is apparent that the predictions of the model depend on the choice of the flexibility parameters Λ_g, which are not well-characterized (Section 2.2). To assess the degree to which different choices affect the pressures in the interstitial regions P_1 and P_2, we sample the parameter space. For each sample point, we evaluate the model solution at an elevated arterial blood pressure P_a. For all simulations, we keep P_a constant at 180 mmHg.

3.2.1. *Summary statistics.* The model utilizes five factors that correspond to the flexibility parameters associated with the histological groups of Section 2.2. We use a sample size of $N = 41 \cdot 10^3$ and perform sampling with the Monte Carlo method. The resulting probability densities and cumulative distributions of P_1 and P_2 are shown in Figure 5.

As can be seen in Figure 5, the model predicts mostly increased P_1 and P_2 at elevated P_a. However, the uncertainty in the flexibility parameters Λ_g induces a significant degree of variability for both pressures. The mean values of P_1 and P_2 are 9.1 and 8.6 mmHg, and the standard deviations are 4.1 and 3.7 mmHg, respectively. Both pressure distributions are heavily skewed towards large values.

Interestingly, the model also predicts low or even negative pressures. Negative pressure values indicate that the pressures in the interstitial regions fall below the pressure in the space surrounding the kidney P_0^{ext}, which in this study is set to 0 mmHg. In summary, 84% of P_1 and 77% of P_2 values at $P_a = 180$ mmHg are above the corresponding values at $P_a = 100$ mmHg, and 16% of P_1 and 11% of P_2 values lie below 0 mmHg or above 15 mmHg.

Scatter plots between the input factors Λ_g and the computed pressures P_1 and P_2 are shown in Figure 6. Only Λ_{G4} shows a clear influence on P_1 and P_2, with high values of Λ_{G4} being associated generally with higher interstitial pressures. No apparent trend can be identified for the rest of the factors. Linear regressions between the computed pressures and the input factors (shown by the dashed lines in Figure 6) yield low R^2. Precisely, values of R^2 for Λ_{G4} equal 0.25 for P_1 and 0.16 for P_2. The rest of the factors yield R^2 for 0.02 or less. Such low R^2 indicate strong nonlinear dependencies of the interstitial pressures on the input factors, a behavior that most likely stems from the inverse-forth-power in the Poiseuille law given by (11).

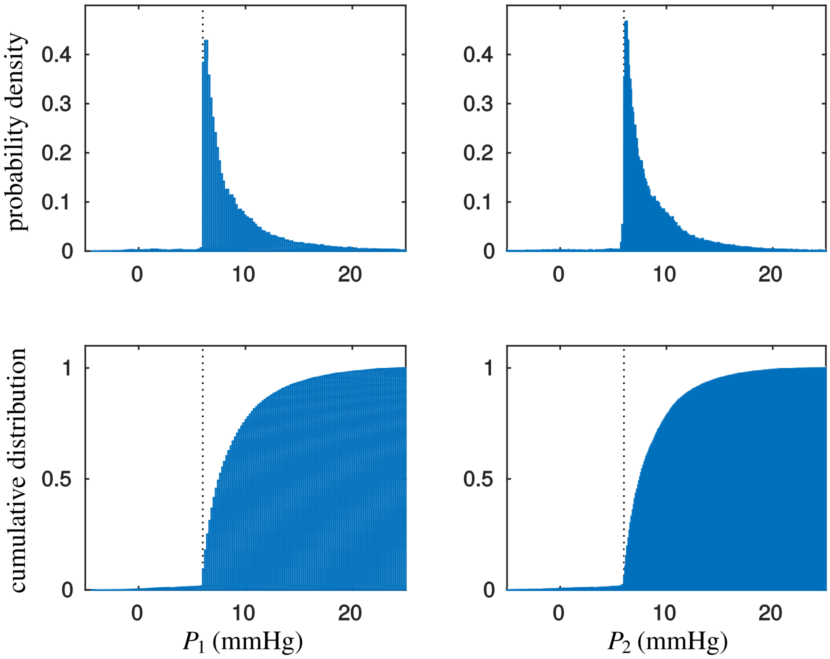


Figure 5. Probability densities of P_1, left panels, and P_2, right panels, at elevated arterial blood pressure ($P_a = 180\,\text{mmHg}$) as estimated by model simulations. Vertical lines indicate the values at the reference arterial blood pressure ($P_a = 100\,\text{mmHg}$).

Correlation coefficients computed between the input factors Λ_g and the computed pressures P_1 and P_2 are shown in Figure 7, left panels. As is suggested by Figure 6, Λ_{G4} is positively correlated (weakly) with P_1 and P_2. From the rest of the factors, Λ_{G1}, Λ_{G3}, and Λ_{G5} are negatively correlated with P_1 and P_2 to an even weaker extent than for Λ_{G4}, and Λ_{G2} shows no correlation with either P_1 or P_2.

In contrast to the apparent lack of any trend between the computed pressures P_1 and P_2 and the input factors Λ_{G4}, the model predicts a high degree of correlation between P_1 and P_2. The associated correlation coefficient reaches as high as 0.95 (Figure 7 right panels), which indicates that P_1 and P_2 are predicted to change *in tandem* in a seemingly linear way.

3.2.2. *Sensitivity indices.* To better characterize the contribution of the individual factors Λ_g in the variance of P_1 and P_2, we calculate their first- and total-order sensitivity indices shown in (24) and (25). Details on the adopted computational methods can be found in Section 2.3.

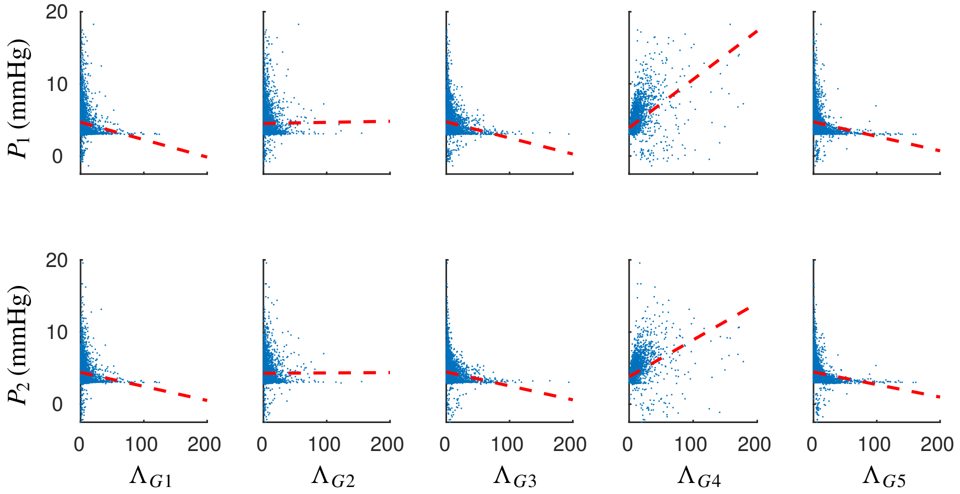


Figure 6. Interstitial pressures P_1, upper panels, and P_2, lower panels, with respect to the sampled input factors Λ_g. Dashed lines indicate the linear regression estimates. For clarity, only $\frac{1}{5}$ of the computed points are shown.

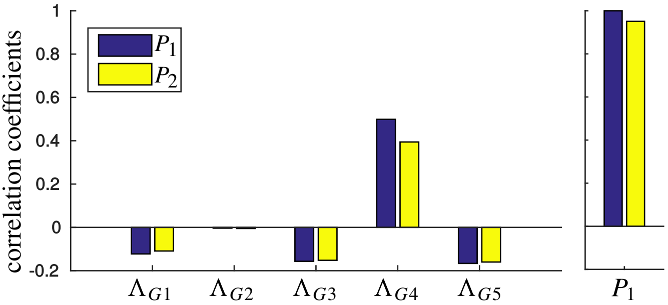


Figure 7. Correlation coefficients between the input factors Λ_g and the computed pressures in the cortical and medullary interstitial spaces P_1 and P_2, respectively.

Figure 8 shows the computed indices. Evidently, the flexibility of the preafferent arteriole vascular segments (group G4) accounts for most of the variation in P_1 or P_2 with respect to either the first- or total-order indices. The postafferent arteriole vasculature (group G5) has the second-most significant contribution. Groups G1–G3 have only minor contributions according to the first-order sensitivity indices. However, this is not the case with the total-order indices, which indicate that G1 and G3 are involved to a significant degree in interactions. On the contrary, the glomeruli (group G2) have only a minor involvement in interactions.

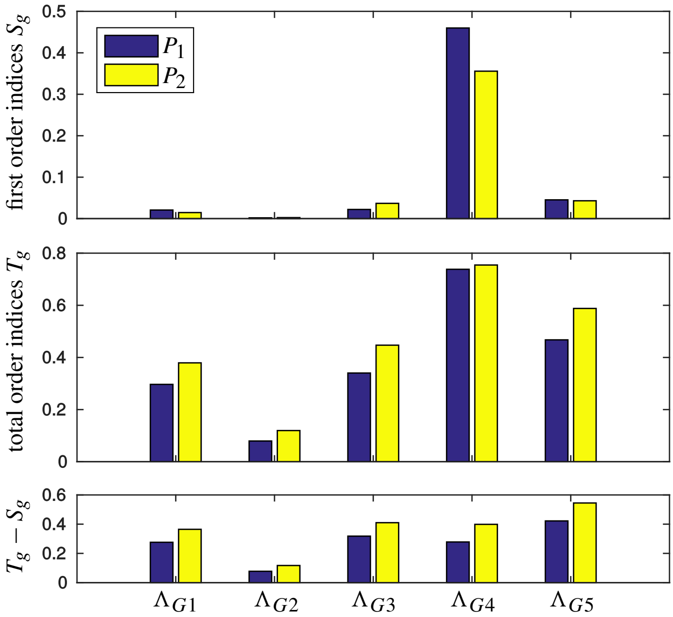


Figure 8. First-order, upper panel, and total-order, middle panel, sensitivity indices of P_1 and P_2 at elevated arterial blood pressure ($P_a = 180$ mmHg). Lower panel shows the difference between the first- and total-order sensitivity indices.

For all groups, it is observed that $T_g^1 < T_g^2$ and $T_g^1 - S_g^1 < T_g^2 - S_g^2$, which indicate that the medullary pressure P_2 is more susceptible to interactions than cortical pressure P_1. This behavior is expected, given that the afferent arterioles (compartments 6 and 12), which initiate the changes in P_1 and P_2, are located exclusively in the cortex, while the medulla is susceptible mostly to secondary interactions initiated by the expansion/constriction of the renal capsule.

4. Conclusions

We developed a multicompartmental computational model of the rat kidney. The model is constructed using conservation laws (2) and (7), fluid dynamics (11), simplified pressure-volume relationships (10) and (20), and constitutive equations specific to the physiology of the kidney (3) and (14) and (15).

We assigned values to the model parameters (Tables 1 and 2) using experimental measurements when such measurements were available and previous modeling studies when direct measurements were not available. However, the data required for the flexibility parameters σ_i are sparse and do not suffice for an accurate estimation of their values. To that end, we chose to model these parameters as random variables with probability distributions that permit values spanning multiple orders of magnitude (Section 2.2 and Figure 3).

To determine the probability distributions of the random variables, we defined five histological groups within the model kidney. Group $G1$ models thick and relatively inflexible structures, for which we used pressure-mass data obtained from whole kidneys in dogs [Hebert et al. 1975; Zhu et al. 1992]. Group $G2$ models the glomeruli, for which we used pressure-volume data measured in rats [Cortes et al. 1996]. Group $G3$ models the various segments of the nephrons and the proximal parts of the collecting duct, for which we used pressure-radius measurements of the rat proximal tubule [Cortell et al. 1973]. Groups $G4$ and $G5$ model the blood vessels, for which we used pressure-volume measurements of the systemic circulation measured in rats [Yamamoto et al. 1983]. We combined the postafferent arteriole vasculature in one group (group $G5$), despite that it consists of segments of the arterial and venous vascular trees [Kriz and Bankir 1988]. We were motivated to do so by the fact that these vascular segments have considerably thinner walls and therefore should be considerably more flexible than the preafferent arteriole segments [Rhodin 1980].

Output from the model leads to a range of predictions depending on the choices of the flexibility values. Generally, increased arterial blood pressure is predicted to increase the pressure in both interstitial spaces (Figure 5). As arterial blood pressure increases from 100 mmHg to 180 mmHg, interstitial pressures are predicted to increase on average by ~ 3 mmHg. Changes of similar magnitude have been observed in the kidneys of rats [Garcia-Estan and Roman 1989; Khraibi et al. 2001; Skarlatos et al. 1994; Khraibi 2000] and dogs [Majid et al. 2001; Granger and Scott 1988]. Upon a limited number of flexibility choices, however, the model predicts decreased interstitial pressures as a result. Furthermore, the model predicts a tight correlation between the cortical and the medullary pressures (Figure 7, right panels) which is also in agreement with the experimental observations reported in [Garcia-Estan and Roman 1989]. Concerning the four case studies of Section 3.1, Cases 2 and 3 are in best agreement with the experimental observations in [Garcia-Estan and Roman 1989; Khraibi et al. 2001; Skarlatos et al. 1994; Khraibi 2000; Majid et al. 2001; Granger and Scott 1988]. In contrast, Case 4 deviates from the experimental observations.

As arterial blood pressure P_a increases, mainly two distinct pathways that lead to interstitial pressure P_1 and P_2 changes can be identified (Figure 9). The first pathway (denoted with red) leads to an *increase* of interstitial pressure upon constriction of the afferent arterioles. The second pathway (denoted with blue) leads to a *decrease* of interstitial pressure upon dilation of the preafferent arteriole blood

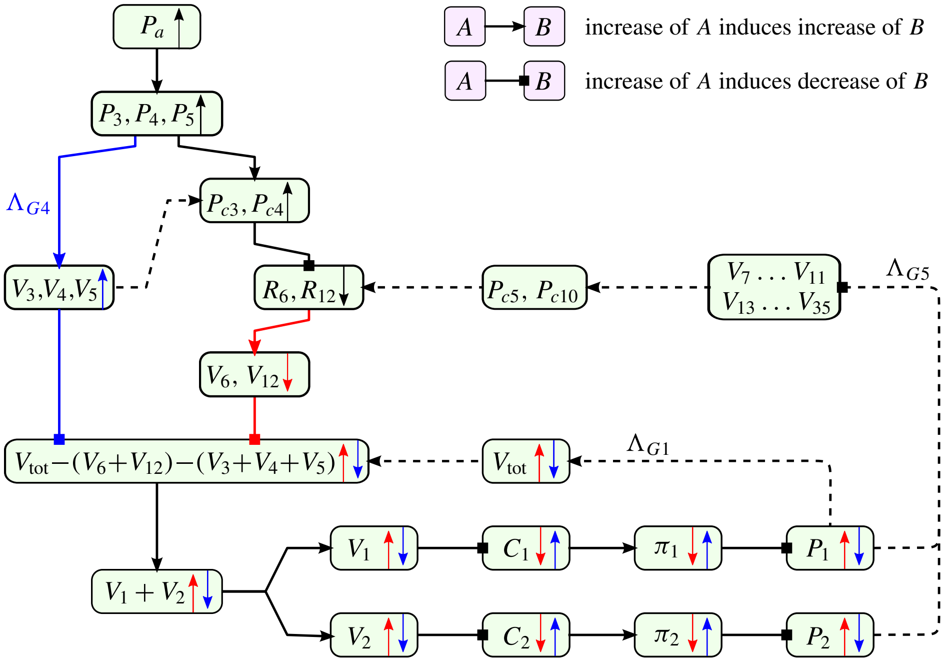


Figure 9. A summary of the mechanism relating arterial blood pressure P_a and interstitial pressures P_1 and P_2. Changes in P_a are transmitted to P_1 and P_2 primarily by two pathways: one is mediated by afferent arteriole volumes (V_6, V_{12}) which is marked with red arrows, the other is mediated by preafferent arteriole volumes (V_3, V_4, V_5) and is marked with blue arrows. The two pathways have competing effects. Secondary interactions are denoted with dashed lines. For simplicity, some of the secondary interactions are omitted.

vessels. Primarily, both pathways lead to changes in interstitial volumes V_1 and V_2, which are subsequently transmitted to protein concentrations C_1 and C_2, oncotic pressures π_1 and π_2, and finally to P_1 and P_2. The two pathways have competing effects; the first leads to changes of P_1 and P_2 towards the same direction as P_a, while the second leads to changes of P_1 and P_2 towards the opposite direction of P_a. It is important to note that, in general, both pathways are active. However, the model results (Figure 5) indicate that under most circumstances the first pathway dominates over the second.

The model predictions appear particularly sensitive to the flexibility of the preafferent arteriole blood vessels (histological group G4) (Figure 8). Such behavior is attributed mostly to the fact that blood pressure is only regulated by the afferent arterioles, which are located after these vessels [Sgouralis and Layton 2015].

The lack of pressure regulation, in the preafferent arteriole compartments, leads to larger internal pressure P_i^{int} changes upon increases in arterial pressure P_a than in the rest of the compartments. For example, as P_a increases from 100 mmHg to 180 mmHg, assuming an increase in the interstitial pressures of ~ 5 mmHg, we see that the compartments of group G4 are stretched by a pressure difference of ~ 70–75 mmHg, while the walls of the rest of the compartments are stretched by a pressure difference of ~ 5 mmHg. Thus, in view of the pressure-volume relations given by (10), the resulting change in total kidney volume V_0, which mediates the changes in interstitial pressures, is mostly affected by s_{G4} rather than s_{G1}, s_{G2}, s_{G3}, or s_{G5}.

The model developed in this study uses several simplifications. For example, the current model assumes perfect autoregulation of blood flow for equations (12) and (13), which limits its applicability to cases with arterial blood pressures between 80 mmHg and 180 mmHg [Sgouralis and Layton 2015]. The model does not account for the differences in tubular reabsorption, e.g., coefficients f_i in (3), occurring between diuretic and antidiuretic animals or for pressure-diuretic responses [Cowley 1997; Moss and Thomas 2014]. Furthermore, the model assumes linear pressure-volume relationships for (10) and (20). Lifting those limitations requires a more detailed model, the development of which will be the focus of future studies. Despite these limitations, the present model could be a useful component in comprehensive models of renal physiology.

Acknowledgments

The authors thank Dr. Vasileios Maroulas for assistance with the statistical analysis in this study and for other helpful discussions. This work is conducted as a part of the 2015 Summer Research Experience for Undergraduates and Teachers at the National Institute for Mathematical and Biological Synthesis (NIMBioS), sponsored by the National Science Foundation through NSF Award #DBI-1300426, with additional support from The University of Tennessee, Knoxville.

References

[Chen et al. 2011] J. Chen, I. Sgouralis, L. C. Moore, H. E. Layton, and A. T. Layton, "A mathematical model of the myogenic response to systolic pressure in the afferent arteriole", *Am. J. Physiol. Renal Physiol.* **300**:3 (2011), F669–F681.

[Cortell et al. 1973] S. Cortell, F. J. Gennari, M. Davidman, W. H. Bossert, and W. B. Schwartz, "A definition of proximal and distal tubular compliance: practical and theoretical implications", *J. Clin. Invest.* **52**:9 (1973), 2330–2339.

[Cortes et al. 1996] P. Cortes, X. Zhao, B. L. Riser, and R. G. Narins, "Regulation of glomerular volume in normal and partially nephrectomized rats", *Am. J. Physiol. Renal Physiol.* **270**:2 (1996), F356–F370.

[Cowley 1997] A. W. Cowley, "Role of the renal medulla in volume and arterial pressure regulation", *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **273**:1 (1997), R1–R15.

[Deen et al. 1972] W. M. Deen, C. R. Robertson, and B. M. Brenner, "A model of glomerular ultrafiltration in the rat", *Am. J. Physiol. Legacy* **223**:5 (1972), 1178–1183.

[Eaton et al. 2009] D. C. Eaton, J. Pooler, and A. J. Vander, *Vander's renal physiology*, 7th ed., McGraw-Hill, New York, 2009.

[Edwards and Layton 2011] A. Edwards and A. T. Layton, "Modulation of outer medullary NaCl transport and oxygenation by nitric oxide and superoxide", *Am. J. Physiol. Renal Physiol.* **301**:5 (2011), F979–F996.

[Fry et al. 2014] B. C. Fry, A. Edwards, I. Sgouralis, and A. T. Layton, "Impact of renal medullary three-dimensional architecture on oxygen transport", *Am. J. Physiol. Renal Physiol.* **307**:3 (2014), F263–F272.

[Garcia-Estan and Roman 1989] J. Garcia-Estan and R. J. Roman, "Role of renal interstitial hydrostatic pressure in the pressure diuresis response", *Am. J. Physiol. Renal Physiol.* **256**:1 (1989), F63–F70.

[Granger and Scott 1988] J. P. Granger and J. W. Scott, "Effects of renal artery pressure on interstitial pressure and Na excretion during renal vasodilation", *Am. J. Physiol. Renal Physiol.* **255**:5 (1988), F828–F833.

[Hebert et al. 1975] L. A. Hebert, K. A. Stuart, and J. A. Stemper, "Whole kidney volume/pressure relationships", *Kidney Int.* **7**:1 (1975), 45–54.

[Heilmann et al. 2012] M. Heilmann, S. Neudecker, I. Wolf, L. Gubhaju, C. Sticht, D. Schock-Kusch, W. Kriz, J. F. Bertram, L. R. Schad, and N. Gretz, "Quantification of glomerular number and size distribution in normal rat kidneys using magnetic resonance imaging", *Nephrol. Dial. Transplant.* **27**:1 (2012), 100–107.

[Holstein-Rathlou and Marsh 1994] N. H. Holstein-Rathlou and D. J. Marsh, "Renal blood flow regulation and arterial pressure fluctuations: a case study in nonlinear dynamics", *Physiol. Reviews* **74**:3 (1994), 637–681.

[Jensen and Steven 1977] P. K. Jensen and K. Steven, "Angiotensin II induced reduction of peritubular capillary diameter in the rat kidney", *Pflügers Archiv* **371**:3 (1977), 245–250.

[Khraibi 2000] A. A. Khraibi, "Renal interstitial hydrostatic pressure and pressure natriuresis in pregnant rats", *Am. J. Physiol. Renal Physiol.* **279**:2 (2000), F353–F357.

[Khraibi et al. 2001] A. A. Khraibi, M. Liang, and T. J. Berndt, "Role of gender on renal interstitial hydrostatic pressure and sodium excretion in rats", *Am. J. Hypertens.* **14**:9 (2001), 893–896.

[Kriz and Bankir 1988] W. Kriz and L. Bankir, "A standard nomenclature for structures of the kidney", *Kidney Int.* **33**:1 (1988), 1–7.

[Layton et al. 2012] A. T. Layton, P. Pham, and H. Ryu, "Signal transduction in a compliant short loop of Henle", *Int. J. Numer. Methods Biomed. Eng.* **28**:3 (2012), 369–383. MR Zbl

[Majid et al. 2001] D. S. A. Majid, K. E. Said, S. A. Omoro, and L. G. Navar, "Nitric oxide dependency of arterial pressure-induced changes in renal interstitial hydrostatic pressure in dogs", *Circul. Res.* **88**:3 (2001), 347–351.

[Moffat and Fourman 1963] D. B. Moffat and J. Fourman, "A vascular pattern of the rat kidney", *J. Anat. Lond.* **97** (1963), 543–553.

[Moss and Layton 2014] R. Moss and A. T. Layton, "Dominant factors that govern pressure natriuresis in diuresis and antidiuresis: a mathematical model", *Am. J. Physiol. Renal Physiol.* **306**:9 (2014), F952–F969.

[Moss and Thomas 2014] R. Moss and S. R. Thomas, "Hormonal regulation of salt and water excretion: a mathematical model of whole kidney function and pressure natriuresis", *Am. J. Physiol. Renal Physiol.* **306**:2 (2014), F224–F248.

[Nordsletten et al. 2006] D. A. Nordsletten, S. Blackett, M. D. Bentley, E. L. Ritman, and N. P. Smith, “Structural morphology of renal vasculature”, *Am. J. Physiol. Heart Circul. Physiol.* **291**:1 (2006), H296–H309.

[Nyengaard 1993] J. R. Nyengaard, “Number and dimensions of rat glomerular capillaries in normal development and after nephrectomy”, *Kidney Internat.* **43**:5 (1993), 1049–1057.

[Oien and Aukland 1991] A. H. Oien and K. Aukland, “A multinephron model of renal blood flow autoregulation by tubuloglomerular feedback and myogenic response”, *Acta Physiol. Scand.* **143**:1 (1991), 71–92.

[Rhodin 1980] J. A. G. Rhodin, “Architecture of the vessel wall”, pp. 1–31 in *Handbook of physiology, supplement 7: The cardiovascular system, vascular smooth muscle*, American Physiological Society, Bethesda, MD, 1980.

[Saltelli 2002] A. Saltelli, “Making best use of model evaluations to compute sensitivity indices”, *Comput. Phys. Commun.* **145**:2 (2002), 280–297. Zbl

[Saltelli et al. 2000] A. Saltelli, S. Tarantola, and F. Campolongo, “Sensitivity analysis as an ingredient of modeling”, *Statist. Sci.* **15**:4 (2000), 377–395. MR

[Sgouralis and Layton 2012] I. Sgouralis and A. T. Layton, “Autoregulation and conduction of vasomotor responses in a mathematical model of the rat afferent arteriole”, *Am. J. Physiol. Renal Physiol.* **303**:2 (2012), F229–F239.

[Sgouralis and Layton 2013] I. Sgouralis and A. T. Layton, “Control and modulation of fluid flow in the rat kidney”, *Bull. Math. Biol.* **75**:12 (2013), 2551–2574. MR Zbl

[Sgouralis and Layton 2014] I. Sgouralis and A. T. Layton, “Theoretical assessment of renal autoregulatory mechanisms”, *Am. J. Physiol. Renal Physiol.* **306**:11 (2014), F1357–F1371.

[Sgouralis and Layton 2015] I. Sgouralis and A. T. Layton, “Mathematical modeling of renal hemodynamics in physiology and pathophysiology”, *Math. Biosci.* **264** (2015), 8–20. MR Zbl

[Sgouralis and Layton 2016] I. Sgouralis and A. T. Layton, “Conduction of feedback-mediated signal in a computational model of coupled nephrons”, *Math. Med. Biol.* **33**:1 (2016), 87–106. MR Zbl

[Sgouralis et al. 2015] I. Sgouralis, R. G. Evans, B. S. Gardiner, J. A. Smith, B. C. Fry, and A. T. Layton, “Renal hemodynamics, function, and oxygenation during cardiac surgery performed on cardiopulmonary bypass: a modeling study”, *Physiol. Rep.* **3**:1 (2015).

[Skarlatos et al. 1994] S. Skarlatos, P. H. Brand, P. J. Metting, and S. L. Britton, “Spontaneous changes in arterial blood pressure and renal interstitial hydrostatic pressure in conscious rats”, *J. Physiol.* **481**:3 (1994), 743–752.

[Sobol’ 2001] I. M. Sobol’, “Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates”, *Math. Comput. Simulation* **55**:1-3 (2001), 271–280. MR Zbl

[Wolgast et al. 1981] M. Wolgast, M. Larson, and K. Nygren, “Functional characteristics of the renal interstitium”, *Am. J. Physiol. Renal Physiol.* **241**:2 (1981), F105–F111.

[Yamamoto et al. 1983] J. Yamamoto, Y. Goto, M. Nakai, K. Ogino, and M. Ikeda, “Circulatory pressure-volume relationship and cardiac output in DOCA-salt rats”, *Hypertension* **5**:4 (1983), 507–513.

[Zhu et al. 1992] M. Q. Zhu, W. Vaneerdeweg, N. Buyssens, and M. E. De Broe, “Quantitative relationships between body weight, kidney weight and nephron size in mongrel dogs”, *Nephron* **62**:2 (1992), 187–191.

Received: 2015-09-29 Revised: 2016-04-26 Accepted: 2016-07-11

mbedell@andrew.cmu.edu
Carnegie Mellon University, 5000 Forbes Ave.,
Pittsburgh, PA 15213, United States

claire.lin@emory.edu
Emory University, 201 Dowman Dr., Atlanta, GA 30322,
United States

roman.melendez2104@gmail.com
University of Puerto Rico Mayaguez, 259 Blvd. Alfonso
Valdez Cabian, Mayagüez 00680, Puerto Rico

sgouralis@nimbios.org
National Institute for Mathematical and Biological Synthesis (NIMBioS), University of Tennessee, 1122 Volunteer
Blvd., Suite 106, Knoxville, TN 37996, United States

mathematical sciences publishers

msp

INVOLVE 10:4 (2017)
dx.doi.org/10.2140/involve.2017.10.651

Sums of squares in quaternion rings

Anna Cooke, Spencer Hamblen and Sam Whitfield

(Communicated by Michael E. Zieve)

Lagrange's four squares theorem states that any positive integer can be expressed as the sum of four integer squares. We investigate the analogous question for quaternion rings, focusing on squares of elements of quaternion rings with integer coefficients. We determine the minimum necessary number of squares for infinitely many quaternion rings, and give global upper and lower bounds.

1. Introduction and definitions

Waring's problem.

Theorem 1.1 (Waring's problem/Hilbert–Waring theorem). *For every integer $k \geq 2$ there exists a positive integer $g(k)$ such that every positive integer is the sum of at most $g(k)$ k-th powers of integers.*

Generalizations of Waring's problem have been studied in a variety of settings (for example, number fields [Siegel 1921] and polynomial rings over finite fields [Car 1973]). Additionally, calculation of the exact values of $g(k)$ for all $k \geq 2$ was completed only relatively recently. For an excellent and thorough exposition of the research on Waring's problem and its generalizations, see Vaughan and Wooley [2002]. We will examine a generalization of Waring's problem to quaternion rings.

Definition 1.2. Let $Q_{a,b}$ denote the quaternion ring

$$\{\alpha_0 + \alpha_1 \boldsymbol{i} + \alpha_2 \boldsymbol{j} + \alpha_3 \boldsymbol{k} \mid \alpha_n, a, b \in \mathbb{Z}, \boldsymbol{i}^2 = -a, \boldsymbol{j}^2 = -b, \boldsymbol{i}\boldsymbol{j} = -\boldsymbol{j}\boldsymbol{i} = \boldsymbol{k}\}.$$

Let $Q^n_{a,b}$ denote the additive group generated by all n-th powers in $Q_{a,b}$.

Note here that $\boldsymbol{k}^2 = -ab$, and if $a = b = 1$, we have what are called the *Lipschitz quaternions*. We then have the following analogue of Waring's Problem.

Conjecture 1.3. For every integer $k \geq 2$ and all positive integers a, b there exists a positive integer $g_{a,b}(k)$ such that every element of $Q^k_{a,b}$ can be written as the sum of at most $g_{a,b}(k)$ k-th powers of elements of $Q_{a,b}$.

MSC2010: primary 11E25, 11P05; secondary 11E20.
Keywords: Waring's problem, quaternions, sums of squares.
Research supported by the McDaniel College Student-Faculty Summer Research Fund.

Main results. We will examine sums of squares in quaternion rings, that is, when $k = 2$. We are therefore looking to generalize Lagrange's four squares theorem, the inspiration for Waring's initial conjecture.

Theorem 1.4 (Lagrange's four squares theorem). *Any positive integer can be written as the sum of four integer squares.*

We prove the following general result giving the upper and lower bounds for $g_{a,b}(2)$ for any positive integers a and b.

Theorem 1.5. *For all positive integers a, b, we have*

$$3 \leq g_{a,b}(2) \leq 5.$$

Additionally, each possible value of $g_{a,b}(2)$ (i.e., 3, 4, and 5) occurs infinitely often.

We prove the general upper and lower bounds in Section 2; more specific results, including the proof of the latter half of Theorem 1.5, are given in Section 3. Note that for any positive integers a and b, the quaternion rings $Q_{a,b}$ and $Q_{b,a}$ are naturally isomorphic; we therefore generally assume $a \leq b$.

2. Squares of quaternions — upper and lower bounds

In this section we prove the upper and lower bounds of Theorem 1.5. We will use the following classical result on sums of squares extensively; for this result and a more general look at sums of squares of integers, see [Grosswald 1985].

Theorem 2.1 (Legendre's three squares theorem). *A positive integer N can be written as the sum of three integer squares if and only if N is not of the form $4^m(8\ell + 7)$ with ℓ, m nonnegative integers.*

To study $g_{a,b}(2)$, we first need to establish the general form of squares of quaternions, and to characterize elements of $Q^2_{a,b}$.

Let $\alpha = \alpha_0 + \alpha_1 \boldsymbol{i} + \alpha_2 \boldsymbol{j} + \alpha_3 \boldsymbol{k} \in Q_{a,b}$. We call α_0 the *real* part of α and $\alpha_1 \boldsymbol{i} + \alpha_2 \boldsymbol{j} + \alpha_3 \boldsymbol{k}$ the *pure* part of α, with $\alpha_1, \alpha_2, \alpha_3$ the *pure coefficients*. Then note that

$$\alpha^2 = \alpha_0^2 - a\alpha_1^2 - b\alpha_2^2 - ab\alpha_3^2 + 2\alpha_0\alpha_1 \boldsymbol{i} + 2\alpha_0\alpha_2 \boldsymbol{j} + 2\alpha_0\alpha_3 \boldsymbol{k}. \tag{1}$$

We therefore have that all the pure coefficients of squares of quaternions, and therefore the pure coefficients of all elements of $Q^2_{a,b}$, are even. Additionally, any set of even pure coefficients can be achieved (for example, set $\alpha_0 = 1$ in (1)), as can any negative real coefficient (since we are assuming $a, b \geq 1$). We therefore have

$$Q^2_{a,b} = \{\alpha_0 + 2\alpha_1 \boldsymbol{i} + 2\alpha_2 \boldsymbol{j} + 2\alpha_3 \boldsymbol{k} \mid \alpha_n \in \mathbb{Z}\}. \tag{2}$$

In 1946, Niven computed $g_{1,1}(2)$ and studied extensions of Waring's problem in other various settings, including complex numbers.

Theorem 2.2 [Niven 1946]. *Every element in $Q_{1,1}^2$ can be written as the sum of at most three squares in $Q_{1,1}$. Additionally, $6+2\boldsymbol{i}$ is not expressible as the sum of two squares in $Q_{1,1}$, so $g_{1,1}(2)=3$.*

We extend this result to $Q_{a,b}$ for all positive integers a, b. The proofs for the lower bounds are similar to Niven's work (i.e., finding examples); the proofs for the upper bounds take more work.

Lemma 2.3. *Suppose a and b are positive integers.*

1. *If $a \equiv 1$ or $2 \bmod 4$, then $2+2\boldsymbol{i}$ is not expressible as the sum of two squares in $Q_{a,b}$.*
2. *If $a \equiv 0$ or $3 \bmod 4$, then $4+2\boldsymbol{i}$ is not expressible as the sum of two squares in $Q_{a,b}$.*

Proof. Let $x=x_0+x_1\boldsymbol{i}+x_2\boldsymbol{j}+x_3\boldsymbol{k}$, and $y=y_0+y_1\boldsymbol{i}+y_2\boldsymbol{j}+y_3\boldsymbol{k}$, with $x_m, y_n \in \mathbb{Z}$ for $m, n \in \{0, 1, 2, 3\}$. Then if $x^2+y^2=\alpha$ with $\alpha=\alpha_0+2\alpha_1\boldsymbol{i}+2\alpha_2\boldsymbol{j}+2\alpha_3\boldsymbol{k} \in Q_{a,b}^2$, we have

$$\alpha_0 = x_0^2+y_0^2-a(x_1^2+y_1^2)-b(x_2^2+y_2^2)-ab(x_3^2+y_3^2), \tag{3}$$

$$\alpha_1 = x_0x_1+y_0y_1, \tag{4}$$

$$\alpha_2 = x_0x_2+y_0y_2, \tag{5}$$

$$\alpha_3 = x_0x_3+y_0y_3. \tag{6}$$

Case 1 ($a \equiv 1, 2 \bmod 4$): Suppose $a \equiv 1, 2 \bmod 4$, and let $\alpha = 2+2\boldsymbol{i}$, so $\alpha_0=2$, $\alpha_1=1$, and $\alpha_2=\alpha_3=0$. Since $\alpha_1=1$, (4) and Bézout's identity then imply that x_0 and y_0 must be relatively prime, since they have a linear combination equal to 1. Then, by (5), we must have $x_0 \mid y_2$ and $y_0 \mid x_2$. However, since $b \geq 1$, if $x_2, y_2 \neq 0$, (3) then implies $\alpha_0 \leq 0$. As $\alpha_0 = 2$, we must have $x_2 = y_2 = 0$. A similar argument using (6) implies $x_3 = y_3 = 0$.

By (4), since $\alpha_1 = 1$, we have that exactly one of the products x_0x_1 and y_0y_1 must be odd; we therefore assume y_0 and y_1 are odd. The following table shows that (3) has no solutions mod 4 if $a \equiv 1, 2 \bmod 4$:

x_0	x_1	equation (3) mod 4
even	odd	$\alpha_0 = 2 \equiv 1-2a$
even	even	$\alpha_0 = 2 \equiv 1-a$
odd	even	$\alpha_0 = 2 \equiv 2-a$

Therefore $2+2\boldsymbol{i}$ cannot be written as the sum of two squares in $Q_{a,b}$.

Case 2 ($a \equiv 0, 3 \bmod 4$): Suppose $a \equiv 0, 3 \bmod 4$. Then let $\alpha = 4+2\boldsymbol{i}$. By the same argument as above, we get three possibilities for (3) mod 4, none of which have solutions. Therefore $4+2\boldsymbol{i}$ cannot be written as the sum of two squares in $Q_{a,b}$. □

As both $2{+}2\boldsymbol{i}$ and $4{+}2\boldsymbol{i}$ are in $Q^2_{a,b}$, this gives us the lower bound in Theorem 1.5. We then turn to the upper bound; we establish an algorithm for expressing every element as a sum of squares.

Lemma 2.4. *Every element in $Q^2_{a,b}$ can be written as a sum of at most five squares in $Q_{a,b}$.*

Proof. Let $\alpha = \alpha_0 + 2\alpha_1\boldsymbol{i} + 2\alpha_2\boldsymbol{j} + 2\alpha_3\boldsymbol{k} \in Q^2_{a,b}$; we want to show that we can represent α as a sum of squares of no more than five quaternions.

Let $v = 1 + U\boldsymbol{i} + \alpha_2\boldsymbol{j} + \alpha_3\boldsymbol{k}$ for some $U \in \mathbb{Z}$, and note that

$$\alpha - v^2 = \alpha_0 - 1 + aU^2 + b\alpha_2^2 + ab\alpha_3^2 + 2(\alpha_1 - U)\boldsymbol{i}.$$

If we also let $A = \alpha_0 - 1 + a\alpha_1^2 + b\alpha_2^2 + ab\alpha_3^2$, we have

$$\alpha - v^2 = A + a(U^2 - \alpha_1^2) + 2(\alpha_1 - U)\boldsymbol{i}. \tag{7}$$

We then have three cases:

(1) $A \geq 0$,

(2) $A < 0$ and A cannot be written as $4^m(8\ell + 7)$ for any nonnegative integer m and $\ell \in \mathbb{Z}$, and

(3) $A < 0$ and $A = 4^m(8\ell + 7)$ for some nonnegative integer m and $\ell \in \mathbb{Z}$.

Case 1 ($A \geq 0$): If $A \geq 0$, then by Theorem 1.4, there exists $w, x, y, z \in \mathbb{Z}$ such that $A = w^2 + x^2 + y^2 + z^2$. Letting $U = \alpha_1$, (7) becomes

$$\alpha - v^2 = A = w^2 + x^2 + y^2 + z^2,$$

so we can represent α as the sum of five squares.

Case 2 ($A < 0$ and $A \neq 4^m(8\ell + 7)$): Here we again let $U = \alpha_1$, so $\alpha - v^2 = A$. Then let e_1 be the greatest exponent of 4 such that 4^{e_1} divides A, and let e_2 be the least exponent of 4 such that $4^{2e_2} + A \geq 0$. We then let $e = \max\{e_1 + 1, e_2\}$, and let $w = 4^e\boldsymbol{i}$.

We then have $\alpha - v^2 - w^2 = A + a4^{2e} \geq 0$. Additionally, since $2e \geq 2e_1 + 2$, if A cannot be written in the form $4^m(8\ell + 7)$, then neither can $A + 4^{2e}$. Therefore by Theorem 2.1, there exist $x, y, z \in \mathbb{Z}$ such that $A + 4^{2e} = x^2 + y^2 + z^2$. So

$$\alpha - v^2 - w^2 = A + 4^{2e} = x^2 + y^2 + z^2,$$

so we can represent α as the sum of five squares.

Case 3 ($A < 0$ and $A = 4^m(8\ell + 7)$): We first treat the case when $m > 0$. Here we let

$$w = 2^{m-1} + \left(\frac{\alpha_1 - U}{2^{m-1}}\right)\boldsymbol{i}$$

and choose $U = \alpha_1 + 2^{m-1}U_1$, where U_1 satisfies the following three conditions:

(a) $4^{m+1} \mid U_1$,

(b) $U_1 > -(2^m\alpha_1)/(4^{m-1}+1)$, and

(c) $U_1 > (A-4^{m-1})/a$.

Note that it is always possible to meet these conditions; for example,

$$U_1 = 4^{m+1}|A| \cdot \max\{1, |\alpha_1|\}$$

satisfies all three. We then have

$$\begin{aligned}\alpha - v^2 - w^2 &= (A + a(U^2-\alpha_1^2) + 2(\alpha_1 - U)\boldsymbol{i}) \\ &\qquad\qquad - \left(4^{m-1} + 2(\alpha_1-U)\boldsymbol{i} - a\left(\frac{\alpha_1-U}{2^{m-1}}\right)^2\right) \\ &= A + a(\alpha_1^2 + 2^m\alpha_1U_1 + 4^{m-1}U_1^2 - \alpha_1^2) - 4^{m-1} + aU_1^2 \\ &= A - 4^{m-1} + aU_1(2^m\alpha_1 + (4^{m-1}+1)U_1).\end{aligned}$$

Note that condition (b) on U_1 ensures the quantity in parentheses must be positive, and condition (c) ensures that $\alpha - v^2 - w^2$ is positive. Letting $A = 4^m(8\ell+7)$ and since 4^{m+1} divides U_1, the remainder of the equation equals $4^{m+1}\ell_1$ for some $\ell_1 \in \mathbb{Z}$, we have

$$\begin{aligned}\alpha - v^2 - w^2 &= 4^m(8\ell+7) - 4^{m-1} + 4^{m+1}\ell_1 \\ &= 4^{m-1}(4(8\ell+7) - 1 + 16\ell_1) \\ &= 4^{m-1}(8(4\ell+3+2\ell_1)+3).\end{aligned}$$

Since this is not of the form excluded by Legendre's Three Squares Theorem, there exist $x, y, z \in \mathbb{Z}$ such that $\alpha - v^2 - w^2 = x^2 + y^2 + z^2$, so we can represent α as the sum of five squares.

Lastly, we treat the case when $A = 8\ell+7$ for some negative integer ℓ. Here we let $U = \alpha_1 + U_1$ and $w = 1 + U_1\boldsymbol{i}$, choosing U_1 such that $8 \mid U_1$ and $U_1 > \max\{|A|, |\alpha_1|\}$. Then

$$\begin{aligned}\alpha - v^2 - w^2 &= (A + a(U^2 - \alpha_1^2) + 2(\alpha_1 - U)\boldsymbol{i}) - (1 - U_1\boldsymbol{i})^2 \\ &= A + a(2\alpha_1U_1 + U_1^2) - 1 + aU_1^2 \\ &= 8\ell + 6 + 8\ell_1,\end{aligned}$$

where we have $\ell_1 + \ell \geq 0$ by the conditions on U_1. Since this is a positive number that is 6 mod 8, it is expressible as the sum of three integer squares by Legendre's three squares theorem. So we can represent α as the sum of five squares here and in all cases. □

Lemmas 2.3 and 2.4 combined give the bounds for $g_{a,b}(2)$ in Theorem 1.5.

3. Values of $g_{a,b}(2)$

In this section, we establish exact values for $g_{a,b}(2)$ for several infinite families of quaternion rings, and for each of the possible values of $g_{a,b}(2)$. We note that the methods for showing each are different: for example, to show $g_{a,b}(2)=3$, all we need is an algorithm to express every element in $Q^2_{a,b}$ as a sum of three squares, and to show $g_{a,b}(2)=5$, all we need is to find an element that cannot be expressed as the sum of four squares.

Quaternion rings with $g_{a,b}(2)=3$. We examine $Q_{1,b}$, where $b\in\mathbb{N}$. We can view $Q_{1,b}$ as an extension of the Gaussian integers $\mathbb{Z}[\sqrt{-1}]=\{x+y\sqrt{-1}\mid x,\,y\in\mathbb{Z}\}$ by adjoining $\boldsymbol{j}$ and $\boldsymbol{k}$. The following lemma then provides a shortcut for representing elements of $Q_{1,b}$ as sums of squares.

Lemma 3.1 [Niven 1940, Theorem 2]. *The equation $\alpha_0+2\alpha_1\boldsymbol{i}=x^2+y^2$ is solvable in $\mathbb{Z}[\sqrt{-1}]$ if $\frac{1}{2}\alpha_0$ and α_1 are not both odd integers.*

Note that this lemma also implies that $g_{\mathbb{Z}[\sqrt{-1}]}(2)=3$.

Theorem 3.2. *For all $b\in\mathbb{N}$, every element in $Q^2_{1,b}$ can be written as the sum of at most three squares in $Q_{1,b}$. Therefore $g_{1,b}(2)=3$ for all $b\in\mathbb{N}$.*

Proof. Let $\alpha=\alpha_0+2\alpha_1\boldsymbol{i}+2\alpha_2\boldsymbol{j}+2\alpha_3\boldsymbol{k}\in Q^2_{1,b}$; we wish to find $x,\,y,\,z\in Q_{1,b}$ such that $\alpha=x^2+y^2+z^2$. Since $\mathbb{Z}[\sqrt{-1}]\subset Q_{1,b}$, Lemma 3.1 implies that it is sufficient to find $z\in Q_{1,b}$ such that $\alpha-z^2\in\mathbb{Z}[\sqrt{-1}]$ and satisfies the hypotheses of Lemma 3.1.

Therefore, let $z=1+U\boldsymbol{i}+\alpha_2\boldsymbol{j}+\alpha_3\boldsymbol{k}$, where $U=0$ if α_1 is even and $U=1$ if α_1 is odd. We then examine $\alpha-z^2$:

$$\begin{aligned}\alpha-z^2&=\alpha_0+2\alpha_1\boldsymbol{i}+2\alpha_2\boldsymbol{j}+2\alpha_3\boldsymbol{k}-1+U^2+b\alpha_2^2+b\alpha_3^2-2U\boldsymbol{i}-2\alpha_2\boldsymbol{j}-2\alpha_3\boldsymbol{k}\\&=\alpha_0-1+U^2+b\alpha_2^2+b\alpha_3^2+2(\alpha_1-U)\boldsymbol{i}.\end{aligned}$$

Note that if α_1 is even, then $U=0$, so α_1-U is even; conversely, if α_1 is odd, then $U=1$, so α_1-U is again even. We can therefore apply Lemma 3.1 to find $x,\,y\in\mathbb{Z}[\sqrt{-1}]\subset Q_{1,b}$ such that $\alpha-z^2=x^2+y^2$. □

We note that the proof relies on the fact that squares in the Gaussian integers can be easily characterized. This is not generally true of imaginary quadratic fields (see [Eljoseph 1954; Niven 1940, Theorem 3]).

Quaternion rings with $g_{a,b}(2)=4$. We combine a standard lower bound proof and a constructive upper bound proof to find a family of quaternion rings with $g_{a,b}(2)=4$.

Lemma 3.3. *There exist elements in $Q^2_{4m,4n+3}$ that are not the sum of three squares.*

Proof. Suppose that there exist $x, y, z \in Q_{4m,4n+3}$ such that $x^2+y^2+z^2=9+2\boldsymbol{j}$. Letting

$$\begin{aligned} x &= x_0+x_1\boldsymbol{i}+x_2\boldsymbol{j}+x_3\boldsymbol{k},\\ y &= y_0+y_1\boldsymbol{i}+y_2\boldsymbol{j}+y_3\boldsymbol{k},\\ z &= z_0+z_1\boldsymbol{i}+z_2\boldsymbol{j}+z_3\boldsymbol{k},\end{aligned}$$

the resulting equations for the real and $\boldsymbol{j}$ coefficients of $9+2\boldsymbol{j}$ are, respectively,

$$x_0^2+y_0^2+z_0^2-4m(x_1^2+y_1^2+z_1^2)-(4n+3)(x_2^2+y_2^2+z_2^2)$$
$$-(4m)(4n+3)(x_3^2+y_3^2+z_3^2)=9, \tag{8}$$
$$x_0x_2+y_0y_2+z_0z_2=1. \tag{9}$$

Examining (8) mod 4, we have

$$x_0^2+y_0^2+z_0^2+x_2^2+y_2^2+z_2^2 \equiv 1 \bmod 4. \tag{10}$$

Recall then that for all integers ℓ, we have $\ell^2 \equiv 0 \bmod 4$ (if ℓ is even) or $\ell^2 \equiv 1 \bmod 4$ (if ℓ is odd). From this we have two possibilities that satisfy (10): we must have either 1 or 5 of $x_0, y_0, z_0, x_2, y_2, z_2$ odd in order for the left side of (10) to sum to 1 mod 4.

If only one of the terms is odd, then the left side of (9) will be even since the lone odd term must be multiplied by an even term, and therefore cannot equal 1. Likewise, if there are five odd terms, the left side of (9) will be the sum of two odd terms and one even term, which cannot sum to 1.

Since (8) and (9) cannot simultaneously be satisfied, $9+2\boldsymbol{j}$ cannot be expressed as the sum of three squares in $Q^2_{4m,4n+3}$. □

When a is a sum of two integer squares. When a is a sum of integer squares, we can construct an algorithm to express elements of $Q^2_{a,b}$ as the sum of four squares. This gives us a general result when combined with the lower bound results of Lemma 3.3.

Lemma 3.4. *Every element of $Q^2_{a,b}$ is the sum of at most four squares in $Q_{a,b}$ in the following two cases*:

- $a=n_1^2+n_2^2$ *with* $\gcd(n_1,n_2)=1$; *or*
- $a=n_1^2+n_2^2$ *with* $\gcd(n_1,n_2)=2$ *and* $n_1\equiv 0 \bmod 4$, *and* $b \not\equiv 0 \bmod 4$.

Note that we allow $n_1=0$ only if $n_2=1$ or 2; in the latter case we get $a=4$, which will be useful in light of Lemma 3.3.

Proof. Let $\alpha=\alpha_0+2\alpha_1\boldsymbol{i}+2\alpha_2\boldsymbol{j}+2\alpha_3\boldsymbol{k}$. If we let $z=1+\alpha_1\boldsymbol{i}+\alpha_2\boldsymbol{j}+\alpha_3\boldsymbol{k}\in Q_{a,b}$, then $\alpha-z^2\in\mathbb{Z}$. We claim that every integer can be represented as the sum of three squares in $Q_{a,b}$; we could then represent α as the sum of four squares.

Let $x = n_1\ell + r$, $y = n_2\ell + s$, and $w = \ell\boldsymbol{i} + \delta\boldsymbol{j}$ for some $\ell, r, s, \delta \in \mathbb{Z}$. We then have

$$\begin{aligned} x^2 + y^2 + w^2 &= (n_1\ell + r)^2 + (n_2\ell + s)^2 + (\ell\boldsymbol{i} + \delta\boldsymbol{j})^2 \\ &= 2(rn_1 + sn_2)\ell + r^2 + s^2 - b\delta^2. \end{aligned} \tag{11}$$

Our method will be to choose r and s to determine a "modulus" $(rn_1 + sn_2)$ and residue class $(r^2 + s^2 - b\delta^2)$. Since ℓ is independent of r and s, we will therefore be able to represent every integer in that residue class (we will only use δ in one particularly troublesome case).

Recall that by Bézout's identity there exist $r_0, s_0 \in \mathbb{Z}$ such that $r_0n_1 + s_0n_2 = \gcd(n_1, n_2) \in \{1, 2\}$; these will inform our choices of r and s. We then have three cases (relabeling if necessary) which we address separately:

(a) n_1 odd, n_2 even, and $\gcd(n_1, n_2) = 1$;

(b) n_1, n_2 odd, and $\gcd(n_1, n_2) = 1$; and

(c) $\frac{1}{2}n_1$ even, $\frac{1}{2}n_2$ odd, and $\gcd(n_1, n_2) = 2$.

Case (a): Our modulus here will be 2. Note that if $r = r_0$, $s = s_0$, and $\delta = 0$, we have from (11)

$$x^2 + y^2 + w^2 = 2\ell + r_0^2 + s_0^2.$$

Next, if $r = r_0 - n_2$, $s = s_0 + n_1$, and $\delta = 0$, (11) yields

$$x^2 + y^2 + w^2 = 2\ell + (r_0 - n_2)^2 + (s_0 + n_1)^2.$$

Recalling that n_1 is assumed to be odd and n_2 is assumed to be even, we necessarily have that $r_0^2 + s_0^2$ and $(r_0 - n_2)^2 + (s_0 + n_1)^2$ cover all residue classes mod 2 with the two equations above. With a proper choice of ℓ, we can therefore directly find $x, y, w \in Q_{a,b}$ such that $\alpha - z^2 = x^2 + y^2 + w^2$, and so we can write α as a sum of four squares in $Q_{a,b}$.

Case (b): Our modulus here will be 4. Since n_1 and n_2 are here both odd, we may assume without loss of generality that r_0 is odd and s_0 is even.

We then use three choices of r and s to represent all possible residue classes mod 4; we let $\delta = 0$ for all subcases. First, let $r = r_0$ and $s = s_0$. Equation (11) is then

$$x^2 + y^2 + w^2 = 2\ell + r_0^2 + s_0^2,$$

which represents all odd integers, since r_0 is odd and s_0 is even.

If we then let $r = 2r_0$ and $s = 2s_0$, (11) then yields

$$x^2 + y^2 + w^2 = 4\ell + 4(r_0^2 + s_0^2).$$

This allows us to represent all multiples of 4.

If, instead, we let $r = 2r_0 - n_2$ and $s = 2s_0 + n_1$, (11) then yields

$$x^2 + y^2 + w^2 = 4\ell + (2r_0 - n_2)^2 + (2s_0 + n_1)^2.$$

As $2r_0 - n_2$ and $2s_0 + n_1$ are necessarily both odd, this allows us to represent all integers that are 2 mod 4. Combined with the above two choices, this covers all residue classes mod 4, and so, similarly to Case (a), we are done.

Case (c): Our modulus here will be 8. We will need four choices of r and s, along with letting $\delta = 1$ if $\alpha - z^2 \equiv 3 \bmod 4$. Note that we are assuming $n_2 \equiv 2 \bmod 4$, so we know that $\frac{1}{2}n_2$ is odd. Additionally, we may assume that s_0 is odd and r_0 is even.

First, let $r = r_0$ and $s = s_0$. Equation (11) is then

$$x^2 + y^2 + w^2 = 4\ell + r_0^2 + s_0^2. \tag{12}$$

If we let $r = r_0 - \frac{1}{2}n_2$ and $s = s_0 + \frac{1}{2}n_1$, (11) yields

$$x^2 + y^2 + w^2 = 4\ell + (r_0 - \tfrac{1}{2}n_2)^2 + (s_0 + \tfrac{1}{2}n_1)^2. \tag{13}$$

Since s_0 and $\frac{1}{2}n_2$ are both odd, while r_0 and $\frac{1}{2}n_1$ are even, (12) represents all integers that are 1 mod 4, while (13) represents all integers that are 2 mod 4.

Next, let $r = 2r_0$ and $2s = s_0$. Equation (11) is then

$$x^2 + y^2 + w^2 = 8\ell + 4(r_0^2 + s_0^2).$$

As r_0 is even and s_0 is odd, this represents all integers that are 4 mod 8.

If we let $r = 2r_0 - n_2$ and $s = 2s_0 + n_1$, (11) yields

$$x^2 + y^2 + w^2 = 8\ell + (2r_0 - n_2)^2 + (2s_0 + n_1)^2.$$

Since $2r_0 \equiv n_1 \equiv 0 \bmod 4$ and $2s_0 \equiv n_2 \equiv 2 \bmod 4$, this represents all integers that are 0 mod 8, and we therefore have all integers that are 0 mod 4.

We still need to represent integers that are 3 mod 4; this is where δ comes in. If we let $\delta = 1$, (11) becomes

$$x^2 + y^2 + w^2 = 2(rn_1 + sn_2)\ell + r^2 + s^2 - b.$$

If $b \not\equiv 0 \bmod 4$ and $\alpha - z^2 \equiv 3 \bmod 4$, this allows us to represent $\alpha - z^2 + b$ via one of the choices of r and s above. Therefore we can always represent α as the sum of four squares in $Q_{a,b}$ in Case (c). □

If $a = n_1^2 + n_2^2$ with $\gcd(n_1, n_2) = 2$, then necessarily $a \equiv 0 \bmod 4$; we can then combine Lemmas 3.3 and 3.4 to get the following theorem.

Theorem 3.5. *Suppose $a = n_1^2 + n_2^2$, where $n_1, n_2 \in \mathbb{N}$ are such that $\gcd(n_1, n_2) = 2$, and $m \in \mathbb{N}$. Then $g_{a,4m+3} = 4$.*

Specifically, if $n_1 = 0$ and $n_2 = 2$, we get that $g_{4,4m+3} = 4$ for all $m \in \mathbb{N}$.

Quaternion rings with $g_{a,b}(2) = 5$. In this section, we find $a, b \in \mathbb{N}$ such that there exists elements of $Q_{a,b}$ that require five squares, which by Lemma 2.4 gives us that $g_{a,b}(2) = 5$.

Theorem 3.6. *For all $m, n \in \mathbb{N}$, there are elements of $Q^2_{4m,4n}$ that are not the sum of four squares in $Q_{4m,4n}$. Therefore $g_{4m,4n}(2) = 5$ for all $m, n \in \mathbb{N}$.*

Proof. Suppose that there exist $w, x, y, z \in Q_{4m,4n}$ such that $w^2 + x^2 + y^2 + z^2 = 8 + 2\boldsymbol{k}$. Letting

$$\begin{aligned} w &= w_0 + w_1\boldsymbol{i} + w_2\boldsymbol{j} + w_3\boldsymbol{k}, \\ x &= x_0 + x_1\boldsymbol{i} + x_2\boldsymbol{j} + x_3\boldsymbol{k}, \\ y &= y_0 + y_1\boldsymbol{i} + y_2\boldsymbol{j} + y_3\boldsymbol{k}, \\ z &= z_0 + z_1\boldsymbol{i} + z_2\boldsymbol{j} + z_3\boldsymbol{k}, \end{aligned}$$

the resulting equations for the real, $\boldsymbol{i}$, $\boldsymbol{j}$, and $\boldsymbol{k}$ coefficients are, respectively,

$$\begin{aligned} w_0^2 + x_0^2 + y_0^2 + z_0^2 - 4m(w_1^2 + x_1^2 + y_1^2 + z_1^2)& \\ - 4n(w_2^2 + x_2^2 + y_2^2 + z_2^2) - 16mn(w_3^2 + x_3^2 + y_3^2 + z_3^2) &= 8, \quad (14) \\ w_0w_1 + x_0x_1 + y_0y_1 + z_0z_1 &= 0, \quad (15) \\ w_0w_2 + x_0x_2 + y_0y_1 + z_0z_1 &= 0, \quad (16) \\ w_0w_3 + x_0x_3 + y_0y_3 + z_0z_3 &= 1. \quad (17) \end{aligned}$$

We start by examining (17) mod 2, and note that at least one of w_0, x_0, y_0, z_0 must be odd, as otherwise the sum of the terms would be even. Since at least one of these terms must be odd, we assume without loss of generality that $w_0 \equiv 1 \bmod 2$. With that in mind, (14) mod 8 is

$$1 + x_0^2 + y_0^2 + z_0^2 - 4m(w_1^2 + x_1^2 + y_1^2 + z_1^2) - 4n(w_2^2 + x_2^2 + y_2^2 + z_2^2) \equiv 0 \bmod 8. \quad (18)$$

Recall then that for all odd ℓ, we have $\ell^2 \equiv 1 \bmod 8$, and for all even ℓ, $\ell^2 \equiv 0$ or 4 mod 8. Since the left side of (18) is 1 added to three squares followed by multiples of 4; in order for it to sum to 0 mod 8, the values of x_0^2, y_0^2, z_0^2 must all be 1 mod 8. So $w_0^2, x_0^2, y_0^2, z_0^2$ are odd.

Then $w_0^2 + x_0^2 + y_0^2 + z_0^2 \equiv 4 \bmod 8$, so an odd number of $w_1^2, x_1^2, y_1^2, z_1^2$ or $w_2^2, x_2^2, y_2^2, z_2^2$ must be odd to contribute an additional 4 mod 8. But this forces an odd number of odd terms on the left side of either (15) or (16), which contradicts their even sums.

Since the equations required for $8 + 2\boldsymbol{k}$ to be a sum of four squares in $Q_{4m,4n}$ cannot hold, $8 + 2\boldsymbol{k}$ cannot be expressed as a sum of four squares in $Q_{4m,4n}$. □

4. Other individual cases

We were able to find $g_{a,b}(2)$ in several other cases for specific values of a and b. We include these here for completeness but also to demonstrate the methods used, which vary significantly from those used in Section 3.

Theorem 4.1. $$g_{2,2}(2) = g_{2,3}(2) = 3.$$

These proofs rely of the theory of quadratic forms — specifically, representations of integers via ternary diagonal quadratic forms. A ternary diagonal quadratic form is a function $f(x, y, z) = rx^2 + sy^2 + tz^2$; for our purposes, we have $r, s, t \in \mathbb{N}$. We say a ternary diagonal quadratic form *represents* $n \in \mathbb{N}$ if there exists an integer solution to $f(x, y, z) = n$. Lastly, we say that a ternary diagonal quadratic form is *regular* if the only positive integers it does not represent coincide with certain arithmetic progressions. The most common example of this is Legendre's three squares theorem: every positive integer not of the form $4^m(8\ell+7)$ can be represented in the form $x^2+y^2+z^2$ with $x, y, z \in \mathbb{Z}$. For more information on representation of integers via quadratic forms, see [Jones and Pall 1939] or (more recently) [Hanke 2004].

Noting that
$$(x\boldsymbol{i} + y\boldsymbol{j} + z\boldsymbol{k})^2 = -(ax^2 + by^2 + abz^2),$$
for our theorem, we will examine the expressions $2x^2+2y^2+4z^2$ and $2x^2+3y^2+6z^2$. Dickson has a complete list of regular diagonal ternary quadratic forms, from whence we get the following lemma.

Lemma 4.2 [Dickson 1939, Table 5]. (1) *Let* $f_{2,2}(x, y, z) = 2x^2 + 2y^2 + 4z^2$. *Then* $f_{2,2}$ *represents all even integers not of the form* $2 \cdot 4^n(16\ell + 14)$.

(2) *Let* $f_{2,3}(x, y, z) = 2x^2 + 3y^2 + 6z^2$. *Then* $f_{2,3}$ *represents all positive integers not of the form* $4^n(8\ell + 7)$ *or* $3m + 1$.

Proof of Theorem 4.1. Let $\alpha = \alpha_0 + 2\alpha_1\boldsymbol{i} + 2\alpha_2\boldsymbol{j} + 2\alpha_3\boldsymbol{k} \in Q_{a,b}^2$. Then, letting $x = 1 + \alpha_1\boldsymbol{i} + \alpha_2\boldsymbol{j} + \alpha_3\boldsymbol{k}$, we have
$$\alpha - x^2 = \alpha_0 - 1 + a\alpha_1^2 + b\alpha_2^2 + ab\alpha_3^2 := A \in \mathbb{Z}. \tag{19}$$
It then suffices to find elements $y, z \in Q_{a,b}$ with $y = y_0 \in \mathbb{Z}$ and $z = z_1\boldsymbol{i} + z_2\boldsymbol{j} + z_3\boldsymbol{k}$ such that
$$A = y^2 + z^2 = y_0^2 - az_1^2 - bz_2^2 - abz_3^2, \tag{20}$$
as we would then have $\alpha = x^2 + y^2 + z^2$.

Case 1 ($a = b = 2$): In light of Lemma 4.2 and the regularity of the associated quadratic form, we know that if we can represent the residue class of A mod 32, then we can find y_0, z_0, z_1, z_2 that satisfy (20).

We let $S_{a,b;m}$ be the set of residue classes mod m that are completely represented by

$$f_{a,b}(z_0, z_1, z_2) = az_0^2 + bz_1^2 + abz_2^2.$$

For example, $2 \in S_{2,2;32}$ since $f_{2,2}(1, 0, 0) = 2$, $2 \not\equiv 2 \cdot 4^n(16\ell + 14) \bmod 32$ for any $n, \ell \in \mathbb{N}$, and by Lemma 4.2 $f_{2,2}$ represents all even integers not of the form $2 \cdot 4^n(16\ell + 14)$. But $16 \notin S_{2,2;32}$ since $16 \equiv 2 \cdot 4^1(16\ell + 14) \bmod 32$.

When $a = b = 2$ and $m = 32$, we have

$$S_{2,2;32} = \{2, 4, 6, 8, 10, 12, 14, 18, 20, 22, 24, 26, 30\};$$

our goal then is to show that for any $A \in \mathbb{Z}$, we can find $y_0 \in \mathbb{Z}$ and $s \in S_{2,2;32}$ such that $A \equiv y_0^2 - s \bmod 32$. By Lemma 4.2, there would then exist $z = z_1\boldsymbol{i} + z_2\boldsymbol{j} + z_3\boldsymbol{k} \in Q_{2,2}$ such that $-s \equiv z^2 \bmod 32$ and $A = y_0^2 + z^2$.

We can then break this search for y_0 and s into cases:

- if $A \not\equiv 0, 1, 4, 5, 16$, or $17 \bmod 32$, then A is congruent to either $-s$ or $1 - s$ for some $s \in S_{2,2;32}$;
- if $A \equiv 0, 16 \bmod 32$, then $A \equiv 4 - s \bmod 32$ for $s = 4, 20 \in S_{2,2;32}$;
- if $A \equiv 1, 5, 17 \bmod 32$, then $A \equiv 9 - s \bmod 32$ for $s = 8, 4, 24 \in S_{2,2;32}$; and
- if $A \equiv 4 \bmod 32$, then $A \equiv 16 - s \bmod 32$ for $s = 12 \in S_{2,2;32}$.

Therefore we can represent A as a sum of two squares from $Q_{2,2}$, and so we can always express α as a sum of three squares from $Q_{2,2}$.

Case 2 ($a = 2$, $b = 3$): We again use the set $S_{a,b;m}$, letting $m = 24$; this yields

$$S_{2,3;24} = \{2, 3, 5, 6, 9, 11, 14, 17, 18, 21\}.$$

Similarly to Case 1, we search for $y_0 \in \mathbb{Z}$ and $s \in S_{2,3;24}$ such that $A \equiv y_0^2 - s \bmod 24$:

- if $A \not\equiv 0, 1, 2, 5, 9, 12$, or $17 \bmod 24$, then A is congruent to either $-s$ or $1 - s$ for some $s \in S_{2,3;24}$;
- if $A \equiv 1, 2, 17 \bmod 24$, then $A \equiv 4 - s \bmod 24$ for $s = 3, 2, 11 \in S_{2,3;24}$;
- if $A \equiv 0, 12 \bmod 24$, then $A \equiv 9 - s \bmod 24$ for $s = 9, 21 \in S_{2,3;24}$;
- if $A \equiv 5 \bmod 24$, then $A \equiv 16 - s \bmod 24$ for $s = 11 \in S_{2,3;24}$; and
- if $A \equiv 9 \bmod 24$, then $A \equiv 36 - s \bmod 24$ for $s = 3 \in S_{2,3;24}$.

Therefore as above we can always express α as a sum of three squares from $Q_{2,3}$. Given the lower bound for $g_{a,b}(2)$ in Lemma 2.3, we therefore have $g_{a,b}(2) = 3$ in both cases. □

The proof of Theorem 4.1 relies entirely on the regularity of the associated ternary quadratic forms given in Lemma 4.2. There are, unfortunately, only finitely many regular diagonal ternary quadratic forms ([Dickson 1939, Table 5] is a complete list), so this exact method has limited general use. Nonetheless, there does seem to be a close relationship between these quaternion rings and ternary quadratic forms, and one might be able to relax the regularity condition slightly and be able to represent "enough" integers to use a similar method as in Theorem 4.1.

5. Open questions

There are many questions left to explore here. It seems like it should be possible to find $g_{a,b}(2)$ for all a and b positive; at the very least, we'd like to know the proportion of such quaternion rings that have each of the possible values of $g_{a,b}(2)$. We have also been using as our analog of the integers the Lipschitz quaternions; the Hurwitz Quaternions would be an equally good choice, especially since we would get unique factorization. Lastly, we have been focusing on the cases when $\boldsymbol{i}^2$ and $\boldsymbol{j}^2$ are negative; one could easily investigate the cases when one or both are positive.

References

[Car 1973] M. Car, "Le problème de Waring pour l'anneau des polynômes sur un corps fini", pp. 13 in *Séminaire de théorie des nombres* (Talence, France), Lab. Théorie des Nombres, Centre Nat. Recherche Sci., Talence, France, 1973. MR Zbl

[Dickson 1939] L. Dickson, *Modern elementary theory of numbers*, University of Chicago Press, 1939. Zbl

[Eljoseph 1954] N. Eljoseph, "On the representation of a number as a sum of squares", *Riveon Lematematika* **7** (1954), 38–43. MR

[Grosswald 1985] E. Grosswald, *Representations of integers as sums of squares*, Springer, New York, 1985. MR Zbl

[Hanke 2004] J. Hanke, "Some recent results about (ternary) quadratic forms", pp. 147–164 in *Number theory*, edited by H. Kisilevsky and E. Z. Goren, CRM Proc. Lecture Notes **36**, Amer. Math. Soc., Providence, RI, 2004. MR Zbl

[Jones and Pall 1939] B. W. Jones and G. Pall, "Regular and semi-regular positive ternary quadratic forms", *Acta Math.* **70**:1 (1939), 165–191. MR Zbl

[Niven 1940] I. Niven, "Integers of quadratic fields as sums of squares", *Trans. Amer. Math. Soc.* **48** (1940), 405–417. MR Zbl

[Niven 1946] I. Niven, "A note on the number theory of quaternions", *Duke Math. J.* **13** (1946), 397–400. MR Zbl

[Siegel 1921] C. Siegel, "Darstellung total positiver Zahlen durch Quadrate", *Math. Z.* **11**:3 (1921), 246–275. MR

[Vaughan and Wooley 2002] R. C. Vaughan and T. D. Wooley, "Waring's problem: a survey", pp. 301–340 in *Number theory for the millennium, III* (Urbana, IL, 2000), edited by M. A. Bennett et al., A K Peters, Natick, MA, 2002. MR Zbl

Received: 2015-11-14 Revised: 2016-04-17 Accepted: 2016-05-02

anc003@mcdaniel.edu *McDaniel College, 2 College Hill, Westminster, MD 21157,*
United States

shamblen@mcdaniel.edu *McDaniel College, 2 College Hill, Westminster, MD 21157,*
United States

slw001@mcdaniel.edu *McDaniel College, 2 College Hill, Westminster, MD 21157,*
United States

INVOLVE 10:4 (2017)
dx.doi.org/10.2140/involve.2017.10.665

On the structure of symmetric spaces of semidihedral groups

Jennifer Schaefer and Kathryn Schlechtweg

(Communicated by Scott T. Chapman)

We investigate the symmetric spaces associated to the family of semidihedral groups of order 2^n. We begin this study by analyzing the structure of the automorphism group and by determining which automorphims are involutions. We then determine the symmetric spaces corresponding to each involution and the orbits of the fixed-point groups on these spaces.

1. Introduction

Real symmetric spaces were first introduced by Élie Cartan [1926; 1927] as a special class of homogeneous Riemannian manifolds. They were later generalized by Berger [1957] who gave classifications of the irreducible semisimple symmetric spaces. Since then the theory of symmetric spaces, a theory that plays a key role in many areas of active research, including Lie theory, differential geometry, harmonic analysis, and physics, has developed into an extensive field. The theory of symmetric spaces also has numerous generalizations. Symmetric varieties, symmetric k-varieties, Vinberg's theta-groups, spherical varieties, Gelfand pairs, Bruhat–Tits buildings, Kac–Moody symmetric spaces, and generalized symmetric spaces are among these generalizations which have found importance in many areas of mathematics and physics such as number theory, algebraic geometry, and representation theory.

The majority of these generalizations can be studied in the context of generalized symmetry spaces. Generalized symmetric spaces are defined as the homogeneous spaces G/H with G an arbitrary group and $H = G^\theta = \{g \in G \mid \theta(g) = g\}$ the fixed-point group of an order-n automorphism θ. Of special interest are automorphisms of order 2, also called *involutions*. If G is an algebraic group defined over a field k and θ an involution defined over k, then these spaces are also called symmetric k-varieties, first introduced in [Helminck 1994].

For involutions there is a natural embedding of the homogeneous spaces G/H into the group G as follows. Let $\tau : G \to G$ be a morphism of G given by

MSC2010: 20D15, 53C35.
Keywords: semidihedral group, quasidihedral group, symmetric spaces, automorphisms, involutions.

$\tau(g) = g\theta(g)^{-1}$ for $g \in G$, where θ is an involution of G. The map τ induces an isomorphism of the coset space G/H onto $\tau(G) = \{g\theta(g)^{-1} \mid g \in G\}$. We will take the image $Q = \{g\theta(g)^{-1} \mid g \in G\}$ as our definition of the *generalized symmetric space determined by* (G, θ). In addition, we define the *extended symmetric space* determined by (G, θ) as $R = \{g \in G \mid \theta(g) = g^{-1}\}$. Extended symmetric spaces play an important role in generalizing the Cartan decomposition for real reductive groups to reductive algebraic groups defined over an arbitrary field. While for real groups it suffices to use Q for the Cartan decomposition, in the general case one needs the extended symmetric space R. Symmetric spaces and symmetric k-varieties are well known for their role in many areas of mathematics. They are probably best known for their fundamental role in representation theory. The generalized symmetric spaces as defined above are of importance in a number of areas as well, including group theory, number theory, and representation theory.

In this paper, we investigate the symmetric spaces associated to one particular family of finite groups, namely the semidihedral groups of order 2^n. Semidihedral groups, also known as quasidihedral groups, appear as Sylow-2 subgroups of certain finite simple groups (see [Alperin et al. 1970]). In Section 2, we analyze the family of semidihedral groups of order 2^n, SD_{2^n}, for $n \geqslant 4$. In Section 3, we classify the automorphisms of SD_{2^n} and determine which automorphisms are involutions. In Section 4, we describe the fixed-point group H, the generalized symmetric space Q, and the extended symmetric space R associated with each involution of SD_{2^n}. In Section 5, we study the orbit decomposition of Q by H and SD_{2^n}. Finally in the Appendix, we provide the H, Q, and R associated to each involution of SD_{16}.

The symmetric spaces associated to the more general family of semidihedral groups of order $8k$, SD_{8k}, where $k \geqslant 1$ are considered in [Raza and Imran 2014]. Their result, Lemma 6, regarding the automorphism group of SD_{8k} is incorrect and as a consequence their results about H, Q, and R associated with each involution of SD_{8k} are not completely accurate. The techniques used in our paper and based on the undergraduate honors thesis of the second author under the supervision of the first author could be utilized to consider this more general family of semidihedral groups and the associated symmetric spaces.

2. Preliminaries

Throughout this paper, we consider the semidihedral group SD_{2^n}, which can be described using the following presentation from [Gorenstein 1968]:

$$\mathrm{SD}_{2^n} = \langle r, s \mid r^{2^{n-1}} = s^2 = 1,\ sr = r^{2^{n-2}-1}s \rangle,$$

where $n \geq 4$ is an integer. This particular presentation is convenient for describing the automorphism group of SD_{2^n}.

We begin by providing some basic facts relating to the structure and properties of the elements of SD_{2^n} that will be useful. It is clear from the group presentation given above that SD_{2^n} is a non-Abelian group. The first result we state provides a commutation relation which we will use to simplify the structure of the group's elements.

Lemma 1. *For any integer $k \geq 1$, we have $sr^k = r^{(2^{n-2}-1)k}s$.*

Using the relation $r^{2^{n-1}} = s^2 = 1$ and the outcome of Lemma 1 repeatedly, we have the following results.

Theorem 2. *Every element of SD_{2^n} has a unique presentation as $r^i s^j$, where i and j are integers with $0 \leq i < 2^{n-1}$ and $j \in \{0, 1\}$.*

We call the presentation given in Theorem 2 the *normal form* of an element of SD_{2^n} and by writing all elements of the group in their normal form, we have the subsequent corollary.

Corollary 3. *The non-Abelian group SD_{2^n} has order 2^n and consists of the elements $1, r, r^2, \dots, r^{(2^{n-1}-1)}, s, rs, \dots, r^{(2^{n-1}-1)}s$.*

When determining the automorphism group and the future symmetric spaces, it will be necessary to know the order of each group element and its inverse. The next two results provide this information.

Theorem 4. *For any integer i with $0 \leq i < 2^{n-1}$, we have*

$$|r^i| = \frac{2^{n-1}}{\gcd(i, 2^{n-1})},$$

$|r^i s| = 2$ when i is even, and $|r^i s| = 4$ when i is odd.

Proof. Because $|\,\mathrm{SD}_{2^n}\,| = 2^n$, we know that the order of every element of SD_{2^n} is a power of 2. By basic properties of cyclic groups, $|r^i| = 2^{n-1}/\gcd(i, 2^{n-1})$. Consider $r^i s$ where $i = 2l$ for some $l \in \mathbb{Z}$. Then by Lemma 1 and the relation $r^{2^{n-1}} = s^2 = 1$,

$$r^i s r^i s = r^{i+i(2^{n-2}-1)} s^2 = r^{2^{n-2}(2l)} = r^{2^{n-1}(l)} = 1.$$

Consider $r^i s$ where $i = 2k + 1$ for some $k \in \mathbb{Z}$. Then

$$(r^i s)^2 = (r^i s)(r^i s) = r^{2^{n-2} i} = r^{2^{n-2}(2k+1)} = r^{2^{n-2}} \neq 1.$$

However, it follows that $(r^i s)^4 = (r^{2^{n-2}})^2 = r^{2^{n-1}} = 1$. □

Theorem 5. *For any integer i with $0 \leq i < 2^{n-1}$, we have $(r^i)^{-1} = r^{2^{n-1}-i}$. When i is even, $(r^i s)^{-1} = r^i s$ and when i is odd, $(r^i s)^{-1} = r^{i+2^{n-2}} s$.*

Proof. Using the relation $r^{2^{n-1}} = 1$, it follows that $(r^i)^{-1} = r^{2^{n-1}-i}$ and by Theorem 4, we know that $(r^i s)^{-1} = r^i s$ when i is even. Consider $r^i s$ where $i = 2k+1$ for some $k \in \mathbb{Z}$. Then again by Lemma 1 and the relation $r^{2^{n-1}} = s^2 = 1$, we have

$$\begin{aligned} r^i s r^{i+2^{n-2}} s = r^i r^{(i+2^{n-2})(2^{n-2}-1)} s^2 &= r^{(2^{n-2})i+(2^{n-2})(2^{n-2}-1)} \\ &= r^{(2^{n-2})[(2k+1)+(2^{n-2}-1)]} = r^{2^{n-1}(k+2^{n-3})} = 1. \end{aligned}$$

Thus the result follows. □

3. Automorphisms and involutions of SD_{2^n}

In this section, we investigate the automorphism group of SD_{2^n}, which we denote by $\mathrm{Aut}(\mathrm{SD}_{2^n})$. We begin by analyzing the structure of each automorphism and then move to proving some properties of the automorphism group as a whole. We conclude this section by determining which elements of $\mathrm{Aut}(\mathrm{SD}_{2^n})$ are involutions.

Theorem 6. *A homomorphism $\phi : \mathrm{SD}_{2^n} \to \mathrm{SD}_{2^n}$ is an automorphism if and only if $\phi(r) = r^a$ and $\phi(s) = r^b s$, where a is odd and b is even.*

Proof. Let $\phi \in \mathrm{Aut}(\mathrm{SD}_{2^n})$. Then by properties of automorphisms, r must map to an element of order 2^{n-1} and s must map to an element of order 2 under ϕ. Thus by Theorem 4, $\phi(r) = r^a$, where a is odd, and $\phi(s) = r^b s$ or $r^{2^{n-2}}$, where b is even. However, ϕ would not be onto if s mapped to $r^{2^{n-2}}$. Therefore, if ϕ is an automorphism, $\phi(r) = r^a$ and $\phi(s) = r^b s$, where a is odd and b is even. The converse of this statement can easily be shown. □

Based on the results of Theorem 6, we can represent each automorphism uniquely as ϕ_{ab} where $\phi_{ab}(r) = r^a$ and $\phi_{ab}(s) = r^b s$, where a is odd and b is even. Using this notation, we see that ϕ_{ab} maps an arbitrary element $r^i s^j$ to $r^{ai+bj} s^j$ and ϕ_{10} denotes the identity automorphism.

Corollary 7. *The automorphism group*, $\mathrm{Aut}(\mathrm{SD}_{2^n})$, *has order* 2^{2n-4}.

Proof. Since there are 2^{n-2} elements r^a where a is odd and 2^{n-2} elements $r^b s$ where b is even, $|\,\mathrm{Aut}(\mathrm{SD}_{2^n})| = 2^{n-2} \cdot 2^{n-2} = 2^{2n-4}$. □

As one of the most important examples of an automorphism of a group G is provided by conjugation by a fixed element in G, it is interesting to determine which elements of $\mathrm{Aut}(\mathrm{SD}_{2^n})$ are inner automorphisms. Given an arbitrary group G and an element $g \in G$, we will let $\psi_g \in \mathrm{Aut}(G)$ denote conjugation by g and $\mathrm{Inn}(G)$ denote the collection of inner automorphisms of G.

Theorem 8. *The inner automorphisms of* SD_{2^n} *are ϕ_{1b} and $\phi_{(2^{n-2}-1)b}$ where $b \in \mathbb{Z}_{2^{n-1}}$ is even.*

Proof. Consider ψ_g for some $g \in \mathrm{SD}_{2^n}$. Suppose $g = r^i$. Then

$$\begin{aligned}\psi_{r^i}(r) &= r^i r r^{2^{n-1}-i} = r^{2^{n-1}+1} = r,\\ \psi_{r^i}(s) &= r^i s r^{2^{n-1}-i} = r^i r^{(2^{n-2}-1)(2^{n-1}-i)} s = r^{2i-2^{n-2}i} s = r^{2(i-2^{n-3}i)} s.\end{aligned}$$

Next, consider $g = r^i s$ where $i \in \mathbb{Z}_{2^{n-1}}$ is even. Then

$$\psi_{r^i s}(r) = r^i s r r^i s = r^i r^{(1+i)(2^{n-2}-1)} s^2 = r^{2^{n-2}-1}$$

and $\psi_{r^i s}(s) = r^i s s r^i s = r^{2i} s$. Finally, consider the case when $g = r^i s$ where $i \in \mathbb{Z}_{2^{n-1}}$ is odd. Then

$$\begin{aligned}\psi_{r^i s}(r) &= (r^i s) r (r^{i+2^{n-2}} s) = r^i r^{(2^{n-2}-1)(1+i+2^{n-2})} s^2 = r^{2^{n-2}-1},\\ \psi_{r^i s}(s) &= r^i s s r^{i+2^{n-2}} s = r^{2i+2^{n-2}} s = r^{2(i+2^{n-3})} s.\end{aligned}$$

Conversely, consider $\phi_{1b} \in \mathrm{Aut}(\mathrm{SD}_{2^n})$. Note that conjugation by $r^{(b/2)(1-2^{n-3})^{-1}}$ gives

$$r^{(b/2)(1-2^{n-3})^{-1}} r r^{-(b/2)(1-2^{n-3})^{-1}} = r$$

and

$$r^{(b/2)(1-2^{n-3})^{-1}} s r^{-(b/2)(1-2^{n-3})^{-1}} = r^b s.$$

Thus, $\phi_{1b} \in \mathrm{Inn}(\mathrm{SD}_{2^n})$. Similarly, consider $\phi_{(2^{n-2}-1)b} \in \mathrm{Aut}(\mathrm{SD}_{2^n})$. If $b/2$ is even, then conjugation by $r^{b/2}s$ gives

$$r^{b/2} s r r^{b/2} s = r^{2^{n-2}-1}$$

and

$$r^{b/2} s s r^{b/2} s = r^b s.$$

If $b/2$ is odd, then conjugation by $r^{b/2-2^{n-3}} s$ gives

$$r^{b/2-2^{n-3}} s r r^{b/2-2^{n-3}+2^{n-2}} s = r^{2^{n-2}-1}$$

and

$$r^{b/2-2^{n-3}} s s r^{b/2-2^{n-3}+2^{n-2}} s = r^b s.$$

Thus, $\phi_{(2^{n-2}-1)b} \in \mathrm{Inn}(\mathrm{SD}_{2^n})$. Therefore, ϕ_{ab} is an inner automorphism of SD_{2^n} if and only if a is 1 or $2^{n-2}-1$ and $b \in \mathbb{Z}_{2^{n-1}}$ is even. □

It follows from this result that 2^{n-1} of the 2^{2n-4} automorphisms in $\mathrm{Aut}(\mathrm{SD}_{2^n})$ are inner automorphisms, which one knew would be the case as $\mathrm{Inn}(\mathrm{SD}_{2^n}) \cong \mathrm{SD}_{2^n}/Z(\mathrm{SD}_{2^n})$ and $|Z(\mathrm{SD}_{2^n})| = 2$ (see [Gorenstein 1968]). In Section 4, we will find it useful to understand the structure of the involutions arising from inner automorphisms because it will allow us to simplify the presentation of the fixed-point groups, the generalized symmetric spaces, and the extended symmetric spaces in these cases.

Before we characterize the automorphisms of finite order, and in particular the involutions, we provide the following lemma.

Lemma 9. *For any* $\phi_{ab}, \phi_{cd} \in \mathrm{Aut}(\mathrm{SD}_{2^n})$, *we have*

$$\phi_{ab} \circ \phi_{cd} = \phi_{[ac \bmod 2^{n-1}][ad+b \bmod 2^{n-1}]}.$$

Proof. Let $r^i s^j \in \mathrm{SD}_{2^n}$, where $i, j \in \mathbb{Z}$ such that $0 \le i \le 2^{n-1} - 1$ and $0 \le j \le 1$. Then

$$\begin{aligned}\phi_{ab} \circ \phi_{cd}(r^i s^j) = \phi_{ab}(r^{ci+dj} s^j) = r^{a(ci+dj)+bj} s^j = r^{(ac)i+(ad+b)j} s^j \\ = \phi_{[ac \bmod 2^{n-1}][ad+b \bmod 2^{n-1}]}(r^i s^j).\end{aligned}$$ □

This result concerning composition of automorphisms of SD_{2^n} is quite useful. It allows to us to answer our question regarding automorphisms of finite order via a straightforward modulo 2^{n-1} calculation.

Theorem 10. *Let* $\phi_{ab} \in \mathrm{Aut}(\mathrm{SD}_{2^n})$. *Then* $(\phi_{ab})^d = \phi_{10}$ *if and only if* $a^d \equiv 1 \bmod 2^{n-1}$ *and* $b(1 + a + a^2 + \cdots + a^{d-1}) \equiv 0 \bmod 2^{n-1}$.

Proof. Consider $\phi_{ab} \in \mathrm{Aut}(\mathrm{SD}_{2^n})$. By repeated use of Lemma 9, we find that $(\phi_{ab})^d(r) = r^{a^d}$ and $(\phi_{ab})^d(s) = r^{b(1+a+a^2+\cdots+a^{d-1})}s$. Since $r^{a^d} = r$ when $a^d \equiv 1 \bmod (2^{n-1})$ and $r^{b(1+a+a^2+\cdots+a^{d-1})}s = s$ when $b(1 + a + a^2 + \cdots + a^{d-1}) \equiv 0 \bmod 2^{n-1}$, the result follows. □

We are now able to determine which automorphisms of SD_{2^n} are involutions and the number of involutions in $\mathrm{Aut}(\mathrm{SD}_{2^n})$ for any n.

Corollary 11. *Let* $\phi_{ab} \in \mathrm{Aut}(\mathrm{SD}_{2^n})$. *Then* $(\phi_{ab})^2 = \phi_{10}$ *if and only if* $a^2 \equiv 1 \bmod 2^{n-1}$ *and* $b(1 + a) \equiv 0 \bmod 2^{n-1}$.

Corollary 12. *For integers* $n \geqslant 4$, $\mathrm{Aut}(\mathrm{SD}_{2^n})$ *contains* $2^{n-1} + 3$ *involutions.*

Proof. By Corollary 11, for any odd integer a in $\mathbb{Z}_{2^{n-1}}$ such that $a^2 \equiv 1 \bmod 2^{n-1}$, we have $\gcd(a+1, 2^{n-1})$ even elements b in $\mathbb{Z}_{2^{n-1}}$ such that $b(1+a) \equiv 0 \bmod 2^{n-1}$. There are four elements a in $\mathbb{Z}_{2^{n-1}}$ with $a^2 \equiv 1 \bmod 2^{n-1}$ by [Burton 2011], namely 1, -1, $1 + 2^{n-2}$, and $-1 + 2^{n-2}$. Thus we have $2 + 2^{n-2} + 2 + 2^{n-2} = 2^{n-1} + 4$ elements $\phi_{ab} \in \mathrm{Aut}(\mathrm{SD}_{2^n})$ with $(\phi_{ab})^2 = \phi_{10}$. Because ϕ_{10} has order 1, it follows that there are $2^{n-1} + 3$ involutions in $\mathrm{Aut}(\mathrm{SD}_{2^n})$. □

Example. Consider SD_{16}. Then by Corollary 12 there are 11 involutions in $\mathrm{Aut}(\mathrm{SD}_{16})$, namely ϕ_{14}, ϕ_{30}, ϕ_{32}, ϕ_{34}, ϕ_{36}, ϕ_{50}, ϕ_{54}, ϕ_{70}, ϕ_{72}, ϕ_{74}, ϕ_{76}.

As stated earlier, it is useful to know which of these involutions arise from inner automorphisms. Using Theorem 8 and Corollary 11, it is clear that when $a = 1$, b must have order 2^{n-2} to satisfy the equation $b(1 + a) \equiv 0 \bmod 2^{n-1}$. However, in the case that $a = 2^{n-2} - 1$, it is not as restrictive, for the equation $b(1 + a) =$

$b(2^{n-2}) \equiv 0 \bmod 2^{n-1}$ is satisfied by any even in $\mathbb{Z}_{2^{n-1}}$. Thus, we have the following result that characterizes which inner automorphisms are also involutions.

Theorem 13. *The involutions of* SD_{2^n} *which arise from inner automorphisms are* $\phi_{12^{n-2}}$ *and* $\phi_{(2^{n-2}-1)b}$, *where* $b \in \mathbb{Z}_{2^{n-1}}$ *is even.*

Example. Consider SD_{16}. It follows from Theorem 13 that the involutions in $\mathrm{Aut}(\mathrm{SD}_{16})$ that arise from inner automorphisms are ϕ_{14}, ϕ_{30}, ϕ_{32}, ϕ_{34}, and ϕ_{36}.

We complete this section by determining which elements of $\mathrm{Aut}(\mathrm{SD}_{2^n})$ are equivalent, for equivalent involutions produce the same generalized symmetric spaces.

Definition 14. Let G be a group and ϕ, $\sigma \in \mathrm{Aut}(G)$. Then ϕ and σ are said to be isomorphic, written $\phi \sim \sigma$, if and only if there exists $\rho \in \mathrm{Aut}(G)$ such that $\rho\phi\rho^{-1} = \sigma$, i.e., ϕ and σ are conjugate to each other. Two isomorphic automorphisms are said to be in the same equivalence class.

Theorem 15. *For any* $\phi_{ab}, \phi_{cd} \in \mathrm{SD}_{2^n}$, *we have* $\phi_{ab}^{-1} = \phi_{cd}$ *if and only if* $c = a^{-1}$ *and* $d \equiv a^{-1}(-b) \bmod 2^{n-1}$.

Proof. Consider $\phi_{ab}, \phi_{cd} \in \mathrm{SD}_{2^n}$. It follows by Lemma 9 that

$$\phi_{ab} \circ \phi_{cd} = \phi_{[ac \bmod 2^{n-1}][(ad+b) \bmod 2^{n-1}]} = \phi_{10}$$

if and only if $ac \equiv 1 \bmod 2^{n-1}$ and $ad + b \equiv 0 \bmod 2^{n-1}$. Now c must equal a^{-1} to satisfy $ac \equiv 1 \bmod 2^{n-1}$. Next, $ad + b \equiv 0 \bmod 2^{n-1}$ becomes $ad \equiv -b \bmod 2^{n-1}$. Then, by multiplying both sides by a^{-1}, we get $d \equiv a^{-1}(-b) \bmod 2^{n-1}$. □

Theorem 16. *For any* $\phi_{ab}, \phi_{cd} \in \mathrm{SD}_{2^n}$, *we have*

$$\phi_{ab} \circ \phi_{cd} \circ \phi_{ab}^{-1} = \phi_{[c \bmod 2^{n-1}][(-bc+ad+b) \bmod 2^{n-1}]}.$$

Proof. Consider $\phi_{ab}, \phi_{cd} \in \mathrm{SD}_{2^n}$. Then

$$\begin{aligned}\phi_{ab} \circ \phi_{cd} \circ \phi_{ab}^{-1} &= \phi_{ab} \circ \phi_{cd} \circ \phi_{[a^{-1}][a^{-1}(-b) \bmod 2^{n-1}]} \\ &= \phi_{ab} \circ \phi_{[a^{-1}c \bmod 2^{n-1}][(c(a^{-1}(-b))+d) \bmod 2^{n-1}]} \\ &= \phi_{[aa^{-1}c \bmod 2^{n-1}][(a(-ca^{-1}b+d)+b) \bmod 2^{n-1}]} \\ &= \phi_{[c \bmod 2^{n-1}][(-bc+ad+b) \bmod 2^{n-1}]}.\end{aligned}$$ □

Theorem 17. *Two elements* $\phi_{ab}, \phi_{cd} \in \mathrm{Aut}(\mathrm{SD}_{2^n})$ *are equivalent if there exists an* $\phi_{ef} \in \mathrm{Aut}(\mathrm{SD}_{2^n})$ *such that* $a = c$ *and* $d \equiv (f(1-a) + be) \bmod 2^{n-1}$.

Proof. Let $\phi_{ab}, \phi_{cd} \in \mathrm{Aut}(\mathrm{SD}_{2^n})$. These elements are conjugate if there exists an $\phi_{ef} \in \mathrm{Aut}(\mathrm{SD}_{2^n})$ such that $\phi_{ef} \circ \phi_{ab} \circ \phi_{ef}^{-1} = \phi_{cd}$. Thus, using the results of the previous theorem, $\phi_{cd} = \phi_{[a \bmod 2^{n-1}][-af+be+f \bmod 2^{n-1}]}$. This is true if and only if $a = c$ and $d \equiv (f(1-a) + be) \bmod 2^{n-1}$. □

Example. Consider SD_{16} and the 11 involutions in $\mathrm{Aut}(\mathrm{SD}_{16})$, namely ϕ_{14}, ϕ_{30}, ϕ_{32}, ϕ_{34}, ϕ_{36}, ϕ_{50}, ϕ_{54}, ϕ_{70}, ϕ_{72}, ϕ_{74}, ϕ_{76}. Then by the previous theorem, the equivalence classes of involutions in $\mathrm{Aut}(\mathrm{SD}_{16})$ are $\{\phi_{14}\}$, $\{\phi_{30}, \phi_{32}, \phi_{34}, \phi_{36}\}$, $\{\phi_{50}, \phi_{54}\}$, and $\{\phi_{70}, \phi_{72}, \phi_{74}, \phi_{76}\}$.

4. Fixed-point groups and symmetric spaces of SD_{2^n}

Recall again from the Introduction that we are interested in determining the fixed-point group H, the generalized symmetric space Q, and the extended symmetric space R for each involution of SD_{2^n} found in Corollary 11. It is important to note that for the remainder of this paper we will let $a \equiv b$ represent $a \equiv b \bmod 2^{n-1}$.

Theorem 18. *For an involution $\phi_{ab} \in \mathrm{Aut}(\mathrm{SD}_{2^n})$, the fixed-point group is*

$$H_{\phi_{ab}} = \{r^i s^j \in \mathrm{SD}_{2^n} \mid i(a-1) + jb \equiv 0\},$$

where $i \in \mathbb{Z}_{2^{n-1}}$ and $j \in \mathbb{Z}_2$.

Proof. Let $\phi_{ab} \in \mathrm{Aut}(\mathrm{SD}_{2^n})$. Then $H_{\phi_{ab}} = \{r^i s^j \in \mathrm{SD}_{2^n} \mid \phi_{ab}(r^i s^j) = r^i s^j\}$, where $i \in \mathbb{Z}_{2^{n-1}}$ and $j \in \mathbb{Z}_2$.

Case 1. Let $j = 0$. Then $\phi_{ab}(r^i) = r^{ai} = r^i$ if and only if $ia \equiv i$ or $i(a-1) \equiv 0$.

Case 2. Let $j = 1$. Then $\phi_{ab}(r^i s) = r^{ai+b} s = r^i s$ if and only if $ai + b \equiv i$ or $i(a-1) + b \equiv 0$. □

Example. Consider SD_{16} and four of its involutions: ϕ_{14}, ϕ_{36}, ϕ_{54}, and ϕ_{70}. Using the results of Theorem 18, we have $H_{\phi_{14}} = \{1, r, \dots, r^7\}$, $H_{\phi_{36}} = \{1, r^4, rs, r^5 s\}$, $H_{\phi_{54}} = \{1, r^2, r^4, r^6, rs, r^3 s, r^5 s, r^7 s\}$, and $H_{\phi_{70}} = \{1, r^4, s, r^4 s\}$.

Theorem 19. *For an involution $\phi_{ab} \in \mathrm{Aut}(\mathrm{SD}_{2^n})$, the generalized symmetric space is*

$$Q_{\phi_{ab}} = \{r^{i(1-a)-jb} \mid i \in \mathbb{Z}_{2^{n-1}} \text{ and } j \in \mathbb{Z}_2\}.$$

Proof. Let ϕ_{ab} be an involution of SD_{2^n}. Then $Q_{\phi_{ab}} = \{(r^i s)\phi_{ab}(r^i s)^{-1} \mid r^i s^j \in \mathrm{SD}_{2^n}\}$, where $i \in \mathbb{Z}_{2^{n-1}}$ and $j \in \mathbb{Z}_2$.

Case 1. Let $j = 0$. Then $(r^i)\phi_{ab}(r^i)^{-1} = r^i (r^{ai})^{-1} = r^i r^{2^{n-1}-ai} = r^{i(1-a)}$.

Case 2. Let $j = 1$. Then $(r^i s)\phi_{ab}(r^i s)^{-1} = (r^i s)(r^{ai+b} s)^{-1}$. Notice that $ai + b$ can be even or odd depending on the value of i since a is odd and b is even.

(i) Suppose i is even. It follows that $ai + b$ is even. Then

$$(r^i s)(r^{ai+b} s)^{-1} = r^i s r^{ai+b} s = r^i r^{(2^{n-2}-1)(ai+b)} s^2 = r^{i-(ai+b)} = r^{i(1-a)-b}.$$

(ii) Suppose i is odd. It follows that $ai + b$ is odd. Then

$$\begin{aligned}(r^i s)(r^{ai+b} s)^{-1} &= (r^i s)(r^{(ai+b)+2^{n-2}} s)\\ &= r^i r^{(2^{n-2}-1)((ai+b)+2^{n-2})} s^2 = r^{i-ai-b+(ai-1)2^{n-2}} = r^{i(1-a)-b}\end{aligned}$$

since $ai - 1$ is even. □

Theorem 20. *For an involution $\phi_{ab} \in \mathrm{Aut}(\mathrm{SD}_{2^n})$, the extended symmetric space is*

$$\begin{aligned} R_{\phi_{ab}} = \{r^i \in \mathrm{SD}_{2^n} \mid i(a+1) \equiv 0\} \\ \cup \{r^i s \in \mathrm{SD}_{2^n} \mid i(a-1)+b \equiv 0 \bmod 2^{n-1} \textit{ and } i \textit{ is even}\} \\ \cup \{r^i s \in \mathrm{SD}_{2^n} \mid i(a-1)+b \equiv 2^{n-2} \bmod 2^{n-1} \textit{ and } i \textit{ is odd}\}. \end{aligned}$$

Proof. Let ϕ_{ab} be an involution of SD_{2^n}. Then

$$R_{\phi_{ab}} = \{r^i s^j \in \mathrm{SD}_{2^n} \mid \phi_{ab}(r^i s^j) = (r^i s^j)^{-1}\}.$$

Case 1. Let $j = 0$. Then $\phi_{ab}(r^i) = r^{ai} = r^{-i} = r^{2^{n-1}-i}$ if and only if $ai \equiv 2^{n-1} - i$. In other words, $i(a+1) \equiv 0$.

Case 2. Let $j = 1$ and i be even. Then $\phi_{ab}(r^i s) = r^{ai+b}s = (r^i s)^{-1} = r^i s$ if and only if $ai + b \equiv i$. In other words, $i(a-1)+b \equiv 0$.

Case 3. Let $j = 1$ and i be odd. Then $\phi_{ab}(r^i s) = r^{ai+b}s = (r^i s)^{-1} = r^{i+2^{n-2}}s$ if and only if $ai + b \equiv i + 2^{n-2}$. In other words, $i(a-1)+b \equiv 2^{n-2}$. □

Example. Consider SD_{16} and four of its involutions: ϕ_{14}, ϕ_{36}, ϕ_{54}, and ϕ_{70}. Using the results of Theorem 19, we have that $Q_{\phi_{14}} = \{1, r^4\}$, $Q_{\phi_{36}} = \{1, r^2, r^4, r^6\}$, $Q_{\phi_{54}} = \{1, r^4\}$, and $Q_{\phi_{70}} = \{1, r^2, r^4, r^6\}$. However, by Theorem 20, we have that $R_{\phi_{14}} = \{1, r^4, rs, r^3s, r^5s, r^7s\}$, $R_{\phi_{36}} = \{1, r^2, r^4, r^6, r^3s, r^7s\}$, $R_{\phi_{54}} = \{1, r^4\}$, and $R_{\phi_{70}} = \{1, r, \dots, r^7, s, r^4s\}$. We see that $Q_{ab} \subseteq R_{ab}$ in all instances, which should be, as $Q \subseteq R$ for all arbitrary groups and all of their respective involutions. However, it is usually the case that $Q \neq R$. Thus the fact that $Q_{\phi_{54}} = R_{\phi_{54}}$ for SD_{16} is noteworthy. We provide the fixed-point group, the generalized symmetric space, and the extended symmetric space for each involution of SD_{16} in the Appendix.

The descriptions of H, Q, and R are more specific when ϕ_{ab} is an inner automorphism. Recall that from Theorem 13, an involution arising from an inner automorphism is of the form $\phi_{12^{n-2}}$ or $\phi_{(2^{n-2}-1)b}$, where $b \in \mathbb{Z}_{2^{n-1}}$ is even.

Theorem 21. *Let ϕ_{ab} be an involution of $\mathrm{SD}_{2^{n-1}}$ which arises from an inner automorphism.*

1. *If $a = 1$ and $b = 2^{n-2}$, then $H_{\phi_{ab}} = \{1, r, r^2, \dots, r^{2^{n-2}}\}$, $Q_{\phi_{ab}} = \{1, r^{2^{n-2}}\}$, and $R_{\phi_{ab}} = \{1, r^{2^{n-2}}, rs, r^3s, \dots, r^{2^{n-1}-1}s\}$.*
2. *If $a = 2^{n-2} - 1$ and b is even, then $H_{\phi_{ab}} = \{1, r^{2^{n-2}}\} \cup \{r^i s \mid i(2^{n-2}-2)+b \equiv 0\}$, $Q_{\phi_{ab}} = \{1, r^2, r^4, \dots, r^{2^{n-1}-2}\}$, and*

$$\begin{aligned} R_{\phi_{ab}} = \{r^i \in \mathrm{SD}_{2^n} \mid i \textit{ is even}\} \\ \cup \{r^i s \in \mathrm{SD}_{2^n} \mid i(2^{n-2}-2)+b \equiv 0 \bmod 2^{n-1} \textit{ and } i \textit{ is even}\} \\ \cup \{r^i s \in \mathrm{SD}_{2^n} \mid i(2^{n-2}-2)+b \equiv 2^{n-2} \bmod 2^{n-1} \textit{ and } i \textit{ is odd}\}. \end{aligned}$$

5. Orbits

By Theorem 18, we can view $H_{\phi_{ab}}$ as the disjoint union of $\{r^i \in \mathrm{SD}_{2^n} \mid i(a-1) \equiv 0\}$ and $\{r^i s \in \mathrm{SD}_{2^n} \mid i(a-1)+b \equiv 0\}$. The first set will contain at least the identity and $r^{2^{n-2}}$. However, the second set may be empty if there is no solution, i, to the equation $i(a-1)+b \equiv 0$ for fixed a and b. The question of the existence of such a solution produces two possible outcomes for the $H_{\phi_{ab}}$-orbits on $Q_{\phi_{ab}}$.

Theorem 22. *Let ϕ_{ab} be an involution of* SD_{2^n}*.*

(1) *If there is no solution, i, to the equation $i(a-1)+b \equiv 0$ for fixed a and b, then the $H_{\phi_{ab}}$-orbits on $Q_{\phi_{ab}}$ are*

$$H_{\phi_{ab}} \backslash Q_{\phi_{ab}} = \bigl\{\{r^k\} \bigm| k = i(1-a) - jb \text{ where } i \in \mathbb{Z}_{2^{n-1}} \text{ and } j \in \mathbb{Z}_2\bigr\}.$$

(2) *If there is a solution, i, to the equation $i(a-1)+b \equiv 0$ for fixed a and b, then the $H_{\phi_{ab}}$-orbits on $Q_{\phi_{ab}}$ are*

$$H_{\phi_{ab}} \backslash Q_{\phi_{ab}} = \bigl\{\{r^k, r^{-k}\} \bigm| k = i(1-a) - jb \text{ where } i \in \mathbb{Z}_{2^{n-1}} \text{ and } j \in \mathbb{Z}_2\bigr\}.$$

Proof. In general, a group G acts on its extended symmetric space R, and thus its generalized symmetric space Q, via θ-twisted conjugation defined as $g.r = gr\theta(g)^{-1}$ for $g \in G$ and $r \in R$, where θ is an involution of G. Given that $H_{\phi_{ab}}$ is the fixed-point group of ϕ_{ab}, the action of $H_{\phi_{ab}}$ on $Q_{\phi_{ab}}$ reduces to conjugation. In addition, we found in Theorem 19 that $Q_{\phi_{ab}} \subset \langle r^2 \rangle \subset \mathrm{SD}_{2^n}$. Thus to determine the orbits of $H_{\phi_{ab}}$ on $Q_{\phi_{ab}}$, it is sufficient to evaluate the action of $H_{\phi_{ab}}$ on a general element r^k, keeping in mind that k is even. Let $r^i \in H_{\phi_{ab}}$ such that $i(a-1) \equiv 0$. Then $r^i r^k (r^i)^{-1} = r^k$ and it follows that elements of the form $r^i \in H_{\phi_{ab}}$ fix $Q_{\phi_{ab}}$ pointwise. Now suppose $r^i s \in H_{\phi_{ab}}$ such that $i(a-1)+b \equiv 0$. Consider the case when i is even. Then

$$\begin{aligned}(r^i s)(r^k)(r^i s)^{-1} &= (r^i s)(r^k)(r^i s)\\ &= (r^i s)(r^{k+i} s) = r^i r^{(2^{n-2}-1)(k+i)} s^2 = r^{(2^{n-2}-1)k} = r^{-k}\end{aligned}$$

since k is even. Finally, suppose i is odd. Then

$$\begin{aligned}(r^i s)(r^k)(r^i s)^{-1} &= (r^i s)(r^k)(r^{i+2^{n-2}} s)\\ &= (r^i s)(r^{k+i+2^{n-2}} s) = r^i r^{(2^{n-2}-1)(k+i+2^{n-2})} s^2 = r^{2^{n-2}(i-1)-k} = r^{-k}\end{aligned}$$

since k and $i-1$ are both even. □

Theorem 23. *Let ϕ_{ab} be an involution of* SD_{2^n}*. There is one* SD_{2^n}*-orbit on $Q_{\phi_{ab}}$, i.e.,* $\mathrm{SD}_{2^n} \backslash Q_{\phi_{ab}} = \{Q_{\phi_{ab}}\}$.

Proof. We proceed by proving that every element of SD_{2^n} is in the SD_{2^n}-orbit of the identity, 1, in $Q_{\phi_{ab}}$. By Theorem 19, every element of $Q_{\phi_{ab}}$ can be written in the form $r^{i(1-a)}$ or $r^{i(1-a)-b}$ for some $i \in \mathbb{Z}_{2^{n-1}}$. We know $r^i \in \mathrm{SD}_{2^n}$ for $i \in \mathbb{Z}_{2^{n-1}}$

and $r^i.1 = r^i\phi_{ab}(r^i)^{-1} = r^i(r^{ai})^{-1} = r^ir^{-ai} = r^{i(1-a)}$. We also know $r^is \in \mathrm{SD}_{2^n}$ for $i \in \mathbb{Z}_{2^{n-1}}$. In the case that i is even,

$$\begin{aligned} r^is\,.\,1 &= r^is\phi_{ab}(r^is)^{-1} \\ &= r^is(r^{ai+b}s)^{-1} = r^is(r^{ai+b}s) = r^ir^{(2^{n-2}-1)(ai+b)}s^2 = r^{i(1-a)-b} \end{aligned}$$

by $ai+b$ even. Likewise when i is odd,

$$\begin{aligned} r^is\,.\,1 &= r^is\phi_{ab}(r^is)^{-1} \\ &= r^is(r^{ai+b}s)^{-1} = r^is(r^{ai+b+2^{n-2}}s) = r^ir^{(2^{n-2}-1)(ai+b+2^{n-2})}s^2 = r^{i(1-a)-b} \end{aligned}$$

since $ai-1$ is even. □

Example. Again, consider the involutions ϕ_{14}, ϕ_{36}, ϕ_{54}, and ϕ_{70} of SD_{16} and their respective fixed-point groups and generalized symmetric spaces from Section 4. By applying Theorem 22, we find that because $i(0)+4 \equiv 0$ has no solution for i and $Q = \{1, r^4\}$,

$$H_{\phi_{14}}\backslash Q_{\phi_{14}} = \big\{\{1\}, \{r^4\}\big\} \quad \text{for } \phi_{14};$$

because $i(2)+6 \equiv 0$ has $i = 1$ as a solution and $Q = \{1, r^2, r^4, r^6\}$,

$$H_{\phi_{36}}\backslash Q_{\phi_{36}} = \big\{\{1\}, \{r^4\}, \{r^2, r^6\}\big\} \quad \text{for } \phi_{36};$$

because $i(4)+4 \equiv 0$ has $i = 1$ as a solution and $Q = \{1, r^4\}$,

$$H_{\phi_{54}}\backslash Q_{\phi_{54}} = \big\{\{1\}, \{r^4\}\big\} \quad \text{for } \phi_{54};$$

because $i(6)+0 \equiv 0$ has $i = 4$ as a solution and $Q = \{1, r^2, r^4, r^6\}$,

$$H_{\phi_{70}}\backslash Q_{\phi_{70}} = \big\{\{1\}, \{r^4\}, \{r^2, r^6\}\big\} \quad \text{for } \phi_{70}.$$

Appendix: Symmetric spaces and fixed-point groups for SD_{16}

involution	H	Q	R
ϕ_{14}	$\{1, r, \dots, r^7\}$	$\{1, r^4\}$	$\{1, r^4, rs, r^3s, r^5s, r^7s\}$
ϕ_{30}	$\{1, r^4, s, r^4s\}$	$\{1, r^2, r^4, r^6\}$	$\{1, r^2, r^4, r^6, s, r^4s\}$
ϕ_{32}	$\{1, r^4, r^3s, r^7s\}$	$\{1, r^2, r^4, r^6\}$	$\{1, r^2, r^4, r^6, rs, r^5s\}$
ϕ_{34}	$\{1, r^4, r^2s, r^6s\}$	$\{1, r^2, r^4, r^6\}$	$\{1, r^2, r^4, r^6, r^2s, r^6s\}$
ϕ_{36}	$\{1, r^4, rs, r^5s\}$	$\{1, r^2, r^4, r^6\}$	$\{1, r^2, r^4, r^6, r^3s, r^7s\}$
ϕ_{50}	$\{1, r^2, r^4, r^6, s, r^2s, r^4s, r^6s\}$	$\{1, r^4\}$	$\{1, r^4, s, rs, \dots, r^7s\}$
ϕ_{54}	$\{1, r^2, r^4, r^6, rs, r^3s, r^5s, r^7s\}$	$\{1, r^4\}$	$\{1, r^4\}$
ϕ_{70}	$\{1, r^4, s, r^4s\}$	$\{1, r^2, r^4, r^6\}$	$\{1, r, \dots, r^7, s, r^4s\}$
ϕ_{72}	$\{1, r^4, rs, r^5s\}$	$\{1, r^2, r^4, r^6\}$	$\{1, r, \dots, r^7, r^3s, r^7s\}$
ϕ_{74}	$\{1, r^4, r^2s, r^6s\}$	$\{1, r^2, r^4, r^6\}$	$\{1, r, \dots, r^7, r^2s, r^6s\}$
ϕ_{76}	$\{1, r^4, r^3s, r^7s\}$	$\{1, r^2, r^4, r^6\}$	$\{1, r, \dots, r^7, rs, r^5s\}$

Acknowledgements

This paper is based on the undergraduate honors thesis of Schlechtweg, which Schaefer supervised. Schaefer would like to thank the Research Experiences for Undergraduate Faculty (REUF) program, a joint program of the American Institute of Mathematics and the Institute for Computational and Experimental Research in Mathematics, and Aloysius G. Helminck, in particular, for introducing her to the deep and rich theory of generalized symmetric spaces.

References

[Alperin et al. 1970] J. L. Alperin, R. Brauer, and D. Gorenstein, "Finite groups with quasi-dihedral and wreathed Sylow 2-subgroups", *Trans. Amer. Math. Soc.* **151** (1970), 1–261. MR Zbl

[Berger 1957] M. Berger, "Les espaces symétriques noncompacts", *Ann. Sci. École Norm. Sup.* (3) **74** (1957), 85–177. MR Zbl

[Burton 2011] D. M. Burton, *Elementary number theory*, 7th ed., McGraw-Hill, Boston, 2011.

[Cartan 1926] E. Cartan, "Sur une classe remarquable d'espaces de Riemann", *Bull. Soc. Math. France* **54** (1926), 214–264. MR JFM

[Cartan 1927] E. Cartan, "Sur une classe remarquable d'espaces de Riemann, II", *Bull. Soc. Math. France* **55** (1927), 114–134. MR JFM

[Gorenstein 1968] D. Gorenstein, *Finite groups*, Harper & Row, New York, 1968. MR Zbl

[Helminck 1994] A. G. Helminck, "Symmetric k-varieties", pp. 233–279 in *Algebraic groups and their generalizations: classical methods* (University Park, PA, 1991), edited by W. J. Haboush and B. J. Parshall, Proc. Sympos. Pure Math. **56**, Part 1, American Mathematical Society, Providence, RI, 1994. MR Zbl

[Raza and Imran 2014] Z. Raza and Imran, "On the structure of involutions and symmetric spaces of quasi dihedral group", preprint, 2014. arXiv

Received: 2016-02-03 Revised: 2016-04-20 Accepted: 2016-05-12

schaefje@dickinson.edu *Department of Mathematics and Computer Science,*
Dickinson College, Carlisle, PA 17013, United States

kms2278@cumc.columbia.edu *Columbia University, New York, NY 10027, United States*

INVOLVE 10:4 (2017)
dx.doi.org/10.2140/involve.2017.10.677

Spectrum of the Laplacian on graphs of radial functions

Rodrigo Matos and Fabio Montenegro

(Communicated by Martin J. Bohner)

We prove that if M is a complete, noncompact hypersurface in $\mathbb{R}^{n+1}$, which is the graph of a real radial function, then the spectrum of the Laplace operator on M is the interval $[0, \infty)$.

1. Introduction

Let M be a simply connected Riemannian manifold. The Laplace operator $\Delta : C_0^\infty(M) \to C_0^\infty(M)$, defined as $\Delta = \operatorname{div} \circ \operatorname{grad}$ and acting on $C_0^\infty(M)$ (the space of smooth functions with compact support), is a second-order elliptic operator and, provided M is complete, it has a unique extension Δ to an unbounded self-adjoint operator on $L^2(M)$ whose domain is $\operatorname{Dom}(\Delta) = \{f \in L^2(M) : \Delta f \in L^2(M)\}$; see [Grigor'yan 2009, Theorem 11.5]. Since $-\Delta$ is positive and symmetric, its spectrum is the set of $\lambda \geq 0$ such that $\Delta + \lambda I$ does not have a bounded inverse. Sometimes we say "spectrum of M" rather than "spectrum of $-\Delta$", and we denote it by $\sigma(M)$. One defines the *essential spectrum*, $\sigma_{\text{ess}}(M)$, to be those λ in the spectrum which are either accumulation points of the spectrum or eigenvalues of infinite multiplicity. The *discrete spectrum* is the set $\sigma_d = \sigma(M) \setminus \sigma_{\text{ess}}(M)$ of all eigenvalues of finite multiplicity which are isolated points of the spectrum.

There is a vast literature on the spectrum of the Laplace operator on complete noncompact manifolds. The first result we mention was published by Tayoshi [1971]. He showed the absence of eigenvalues of $-\Delta$ for a class of surfaces of revolution, determined by nonnegative radial growth.

Donnelly [1981] showed

$$\sigma_{\text{ess}}(M) = \big[(n-1)^2 \tfrac{1}{4} c^2, \infty\big),$$

provided M is a Hadamard manifold whose sectional curvature approaches $-c^2$ at infinity. Karp [1984] gave sufficient conditions for a class of manifolds to have

MSC2010: primary 58J50; secondary 58C40.

Keywords: Complete surface, Laplace operator, spectrum.

purely continuous spectrum ($\sigma_d(M) = \varnothing$) under some curvature conditions. Eight years later, Donnelly and Garofalo [1992] obtained results in a similar direction, using the hypothesis of nonnegative radial sectional curvature, without restrictions on the metric.

Cheng and Zhiqin Lu [1992] proved $\sigma_{\text{ess}}(M) = [0, \infty)$ when M has nonnegative radial sectional curvature and Li [1994] proved $\sigma_{\text{ess}}(M) = [0, \infty)$, provided M has nonnegative Ricci curvatures and a pole. Zhou [1994] proved $\sigma_{\text{ess}}(M) = [0, \infty)$ when M has nonnegative sectional curvatures, generalizing the work of Escobar and Freire [1992].

Kumura [1997] found a result which generalized [Donnelly 1981]. He showed $\sigma_{\text{ess}}(M) = \left[\frac{1}{4}c^2, \infty\right)$ whenever

$$\lim_{n\to\infty} \sup_{t>n} |\Delta t - c| = 0,$$

where t denotes the distance function on M.

Wang [1997] showed that the spectrum of a complete, noncompact Riemannian manifold with asymptotically nonnegative Ricci curvature is equal to $[0, \infty)$.

Zhiqin Lu and Detang Zhou [2011] proved that the L^p essential spectrum of M is equal to $[0, \infty)$ when

$$\liminf_{x\to\infty} \operatorname{Ric}_M(x) = 0$$

and M is noncompact and complete. We should mention here that almost all the above works were strongly motivated by the decomposition principle [Donnelly and Li 1979], which states that the essential spectrum of a Riemannian manifold is invariant under compact perturbations of the metric, thus it is a function of the geometry of the ends. In [Monte and Montenegro 2015], it was proved that $\sigma_{\text{ess}}(M) \supset \left[(n-1)^2\frac{1}{4}c^2, \infty\right)$ for a class of Riemannian manifolds, not necessarily complete, whose metric is given by

$$g_M = dr^2 + \psi^2(rw)g_{\mathbb{S}^{n-1}},$$

using curvature conditions only in a neighborhood of a ray.

See also [Bessa et al. 2010; 2012; 2015; Donnelly and Li 1979; Kleine 1988; 1989; Tayoshi 1971] for geometric conditions implying the discreteness of the spectrum, $\sigma_{\text{ess}}(M) = \varnothing$.

In this work we consider complete hypersurfaces which are graphs of radial functions. Our main result is the following theorem.

Theorem 1. *Let M be a complete hypersurface in $\mathbb{R}^{n+1}$, which is the graph of a real radial function. Then, the spectrum of the Laplace operator on M is* $[0, \infty)$.

Without loss of generality, we may assume the domain Dom f to be connected and symmetric with respect to $0 \in \mathbb{R}^n$. From the completeness of M we further deduce Dom f is an open ball or annulus. The theorem above allows us to construct a bounded hypersurface with the same spectrum of $\mathbb{R}^{n+1}$ by taking M to be the graph of the real function $f(x) = \cos\bigl(\tan\bigl(\frac{1}{2}\pi\,|x|\bigr)\bigr)$ defined on the unit open ball.

Throughout the following discussion, for simplicity, we deal with the case where $f : D \to \mathbb{R}$ is defined in an open ball. Let $X : [0, R) \times \Omega \to D$ be defined by $X(r, x_1, \dots, x_{n-1}) = rw(x_1, \dots, x_{n-1})$, where $0 < R \le +\infty$ and w is a coordinate system on S^{n-1} defined on an open set Ω of $\mathbb{R}^n$. Note that M has a natural coordinate system $Y : [0, R) \times \Omega \to M$, given by $Y(r, x_1, \dots, x_{n-1}) = (rw(x_1, \dots, x_{n-1}), f(r))$, but we are interested in the spherical coordinate system for M on $p = (0, f(0))$. Consider $t : [0, R) \to [0, \infty)$, given by

$$t(r) = \int_0^r \bigl(1 + f'(\tau)^2\bigr)^{1/2}\,d\tau.$$

We claim that t is a diffeomorphism. Observe that t is increasing and

$$\lim_{r\to R} t(r) = +\infty.$$

We denote by $r : [0, \infty) \to [0, R)$ the inverse diffeomorphism. By the inverse function theorem,

$$0 < r'(t) = \bigl(1 + f'(r)^2\bigr)^{-1/2} \le 1. \tag{1}$$

Finally, the system of spherical coordinates on M, denoted $Z : [0, \infty) \times \Omega \to M$, is defined by

$$Z(t, x_1, \dots, x_{n-1}) = \bigl(r(t)w(x_1, \dots, x_{n-1}),\ f \circ r(t)\bigr).$$

The metric of M on such a system is given by

$$g_M = dt^2 + r(t)^2 g_{\mathbb{S}^{n-1}}.$$

Because of this observation, Theorem 1 is a simple consequence of the theorem below.

Theorem 2. *Let $I \subset \mathbb{R}$ be an unbounded interval and $M = I \times \mathbb{S}^{n-1}$ with metric given by $g_M = dt^2 + r^2(t)g_{\mathbb{S}^{n-1}}$, where $0 < r'(t) \le c$ for all t. Then, the spectrum of the Laplace operator on M is* $[0, \infty)$.

Remark. (1) If M has a pole at $p \in M$, then $\exp_p : T_pM \to M$ is a diffeomorphism so that M isometric to T_pM with the pullback metric. Therefore, Theorem 2 implies that if M has a pole p and $g_M = dt^2 + r^2(t)g_{\mathbb{S}^{n-1}}$ with respect to p and $0 < r'(t) < c$, then M has spectrum equal to $[0, \infty)$.

(2) To the best of our knowledge, this natural result has only been verified in less general settings. For instance, since $r'(t) > 0$, then $r(t)$ is increasing and there are only two possibilities:

(a) $\lim_{t\to\infty} r(t) = \infty$, or

(b) $\lim_{t\to\infty} r(t) = R$.

In the first case, since $r'(t)$ is bounded, we have

$$\lim_{t\to\infty} \Delta t = \lim_{t\to\infty} \frac{r'(t)}{r(t)} = 0.$$

By [Kumura 1997, Theorem 1.2], it follows that the spectrum of M is purely continuous and equal to $[0, \infty)$. In the second case, if $r' \to 0$ we still have $r'(t)/r(t) \to 0$. Therefore, the main contribution of this paper is the proof of the case where $r'(t)$ does not converge to zero and $\lim_{t\to\infty} r(t) = R < +\infty$. This is the scenario for the graph of the function $f(x) = \cos\bigl(\tan\bigl(\frac{1}{2}\pi\,|x|\bigr)\bigr)$ presented above.

In the next section we prove Theorem 2. The Appendix is devoted to the Sturm–Liouville theory used in this note.

2. Proof of Theorem 2

We concentrate our efforts for the case where $\lim_{t\to\infty} r(t) = R$. Our approach is variational, based on the following lemma.

Lemma 3 [Davies 1995, Lemma 4.1.2]. *A number $\lambda \in \mathbb{R}$ lies in the spectrum of a self-adjoint operator H if and only if there exists a sequence of functions $f_n \in \operatorname{Dom} H$ with $\|f_n\| = 1$ such that*

$$\lim_{n\to\infty} \|Hf_n - \lambda f_n\| = 0.$$

To deduce Theorem 2 from Lemma 3 we will construct, for each $\lambda > 0$, a sequence of radial smooth functions $f_p : M \to \mathbb{R}$ with compact support such that

$$\|\Delta f_p + \lambda f_p\|_{L^2(M)} \le \frac{c}{p}\, \|f_p\|_{L^2(M)} \tag{2}$$

for any natural p, where c is a constant which does not depend on p. It will follow that $g_p = f_p/\|f_p\|$ has norm one and

$$\lim_{p\to\infty} \|\Delta g_p + \lambda g_p\|_{L^2(M)} = 0.$$

Therefore, by Lemma 3, λ belongs to the spectrum. To construct the function f_p, we fix $t_0 > 0$ and prove that there are $t_1(\lambda) > t_0$ and a radial function $u = u(t)$ solution of the problem

$$\begin{cases} \Delta u + \lambda u = 0 & \text{in } [t_0, t_1], \\ u(t_0) = u(t_1) = 0, & \\ u > 0 & \text{in } (t_0, t_1). \end{cases} \tag{3}$$

Using Sturm–Liouville theory, we showed that u can be extended to the whole interval $[t_0, \infty)$ and it has infinite zeros $t_0 < t_1 < \cdots < t_p < \cdots$. The next step is to consider (for each p) a smooth bump function h_p whose support is the interval $[t_0, t_{3p}]$. We then define $f_p = uh_p$ and show that each f_p in this sequence satisfies (2). The function $t \mapsto r^{n-1}(t)$ has a geometric meaning and plays an important role in the proof, thus deserving a special notation. In the sequence of the paper, we let $v(t) = r^{n-1}(t)$.

We observe that the first equation in (3) is equivalent to

$$(v(t)u'(t))' + \lambda v(t)u(t) = 0 \tag{4}$$

if $u = u(t)$ is a radial function. By Theorem 9 in the Appendix, given positive t_0 and λ, (4) has a solution defined on $[t_0, \infty)$ and satisfying $u(t_0) = 0$.

Moreover, Corollary 8 allows us to consider a sequence of zeros $t_0 < t_1 < \cdots$ of u.

For $p \in \mathbb{N}$, we choose a smooth bump function $h = h_p \colon \mathbb{R} \mapsto \mathbb{R}$ with $0 \le h \le 1$ satisfying

$$\begin{cases} h(t) = 0, & t \in (-\infty, t_0] \cup [t_{3p}, \infty), \\ h(t) = 1, & t \in [t_p, t_{2p}]. \end{cases}$$

Such a function can be defined in the following way: let $\varphi \in C_0^\infty(\mathbb{R})$ be nonnegative with $\operatorname{supp} \varphi = [0, 1]$ and $\int \varphi = 1$. Let

$$h_p(t) = \int_{-\infty}^{t} \varphi_p(s)\, ds,$$

where

$$\varphi_p(t) = \frac{1}{t_p - t_0}\varphi\left(\frac{t - t_0}{t_p - t_0}\right) - \frac{1}{t_{3p} - t_{2p}}\varphi\left(\frac{t - t_{2p}}{t_{3p} - t_{2p}}\right).$$

This construction is useful since it leads to the following estimates:

$$\begin{aligned} \|h_p'\|_\infty &\le \max\left\{\frac{\|\varphi\|_\infty}{t_p - t_0}, \frac{\|\varphi\|_\infty}{t_{3p} - t_{2p}}\right\} \le \frac{C}{p}, \\ \|h_p''\|_\infty &\le \max\left\{\frac{\|\varphi'\|_\infty}{(t_p - t_0)^2}, \frac{\|\varphi'\|_\infty}{(t_{3p} - t_{2p})^2}\right\} \le \frac{C}{p^2}. \end{aligned} \tag{5}$$

Here, we have made use of Corollary 11 in the Appendix.

Consider $f = f_p = uh_p$. We are going to prove that such a function satisfies the inequality in (2). Computing $\Delta f + \lambda f$, we obtain

$$\Delta f + \lambda f = 2u'h' + uh'' + (n-1)\frac{r'}{r}h'u.$$

Using the inequalities in (5), together with the fact that r is increasing and r' is bounded, we have

$$|\Delta f+\lambda f|\le \frac{c}{p}(|u'|+|u|)\chi_{[t_0,t_{3p}]}.$$

Then,

$$|\Delta f+\lambda f|^2\le \frac{c}{p^2}(|u'|^2+|u|^2)\chi_{[t_0,t_{3p}]},$$

$$\int_M |\Delta f+\lambda f|^2\,dM\le \frac{c}{p^2}\biggl(\int_{t_0}^{t_{3p}}|u'|^2 v\,dt+\int_{t_0}^{t_{3p}}|u|^2 v\,dt\biggr).$$

Multiplying (4) by u and using integration by parts we find

$$\int_{t_0}^{t_{3p}}|u'|^2 v(t)\,dt=\lambda\int_{t_0}^{t_{3p}}|u|^2 v(t)\,dt,$$

$$\|\Delta f_p+\lambda f_p\|_{L^2(M)}\le \frac{c}{p}\|u\cdot\chi_{[t_0,t_{3p}]}\|_{L^2(M)}\le \frac{c}{p}\|u\cdot\chi_{[t_p,t_{2p}]}\|_{L^2(M)}\le \frac{c}{p}\|f_p\|_{L^2(M)},$$

where the second inequality comes from Lemma 4 below.

Lemma 4. *There is a positive constant C independent on p such that*

$$\int_{t_0}^{t_{3p}}u^2 v\,dt\le C\int_{t_p}^{t_{2p}}u^2 v\,dt,$$

where u is solution of (4) *and $t_0<t_1<\cdots$ are zeros of u.*

This result is a manifestation of the oscillatory behavior of u. Before justifying its veracity, we state a useful way of estimating u between two zeros.

Lemma 5. *Let u be a solution of* (4), *and choose t_k, t_{k+1} to be consecutive zeros for u. Define*

$$\alpha_k(t)=a_k\sin\biggl(\lambda^{1/2}R^{n-1}\int_{t_k}^{t}v^{-1}(s)\,ds\biggr)$$

and

$$\beta_k(t)=b_k\sin\biggl(\lambda^{1/2}v(t_k)\int_{t_k}^{t}v^{-1}(s)\,ds\biggr),$$

where $a_k=v(t_k)b_k/(R^{n-1}\lambda^{1/2})$ and $b_k=u'(t_k)/\lambda^{1/2}$. Then $|\alpha_k|\le|u|$ on $(t_k,\tilde{t}_k)$ and $|u|\le|\beta_k|$ on (t_k,t_{k+1}), where $\tilde{t}_k$ is the next zero of α_k after t_k.

To make the exposition more fluid, we postpone the proof until the Appendix.

Proof of Lemma 4. Observe that multiplying (4) by $v(t)u'$ we get

$$(v(t)u')'v(t)u'+\lambda v^2uu'=0,$$

and so,

$$\bigl((v(t)u')^2\bigr)'+\lambda v^2(u^2)'=0.$$

Integrating from t_0 to t_k, we have

$$v(t_k)^2u'(t_k)^2 - v(t_0)^2u'(t_0)^2 = -\lambda \int_{t_0}^{t_k} v^2(s)(u^2(s))'\,ds.$$

Integrating the right hand side by parts, we find

$$v(t_k)^2u'(t_k)^2 - v(t_0)^2u'(t_0)^2 = 2\lambda \int_{t_0}^{t_k} vv'u^2\,ds. \tag{6}$$

Since $r, r' > 0$, we have $v, v' > 0$. Also, $r(t) < R$ and as a consequence,

$$u'(t_k)^2 > \frac{v(t_0)^2u'(t_0)^2}{R^{2(n-1)}} \tag{7}$$

for $k \geq 1$.

To obtain an estimate in the other direction, we observe that the function $\beta = \beta_0(t)$ in Lemma 5 satisfies $\beta'(t_0) = u'(t_0) > 0$ and

$$(v(t)\beta'(t))' + \frac{\lambda v(t_0)^2}{v(t)}\beta(t) = 0. \tag{8}$$

Multiplying by $v(t)\beta'$ we get, as in the preceding computations,

$$(v(t)^2(\beta')^2)' + \lambda v(t_0)^2(\beta^2)' = 0. \tag{9}$$

Now, if $\bar{t_1}$ is the next root of β after t_0, integrating the last equation we find

$$\begin{aligned} v(\bar{t_1})^2\beta'(\bar{t_1})^2 &= v(t_0)^2\beta'(t_0)^2 \\ &= v(t_0)^2u'(t_0)^2. \end{aligned} \tag{10}$$

We take $k = 1$ and estimate the right side of (6) as follows:

$$\begin{aligned} \lambda \int_{t_0}^{t_1} (v^2)'u^2\,dt &\leq \lambda \int_{t_0}^{t_1} (v^2)'\beta^2\,dt \\ &\leq \lambda \int_{t_0}^{\bar{t_1}} (v^2)'\beta^2\,dt \\ &= -\lambda \int_{t_0}^{\bar{t_1}} v^2(\beta^2)'\,dt \\ &= -\frac{1}{v(t_0)^2} \int_{t_0}^{\bar{t_1}} v^2(\lambda v(t_0)^2\beta^2)'\,dt. \end{aligned} \tag{11}$$

By (9) we infer

$$\begin{aligned} -\frac{1}{v(t_0)^2}\int_{t_0}^{\bar{t_1}} v^2(\lambda v(t_0)^2\beta^2)'\,dt &= \frac{1}{v(t_0)^2}\int_{t_0}^{\bar{t_1}} v^2(v^2(\beta')^2)'\,dt \\ &= \frac{1}{v(t_0)^2}\int_{t_0}^{\bar{t_1}} (v^4(\beta')^2)' - (v^2)'v^2(\beta')^2\,dt \\ &< \frac{v^4(\bar{t_1})(\beta')^2(\bar{t_1}) - v^4(t_0)(\beta')^2(t_0)}{v(t_0)^2}. \end{aligned} \tag{12}$$

Now, using (10) and that $\beta'(t_0) = u'(t_0)$, we find

$$\lambda\int_{t_0}^{t_1} (v^2)'u^2 \le (v(\bar{t_1})^2 - v(t_0)^2)u'(t_0)^2\,dt.$$

Then, by (6),

$$v(t_1)^2u'(t_1)^2 - v(t_0)^2u'(t_0)^2 \le \big(v(\bar{t_1})^2 - v(t_0)^2\big)u'(t_0)^2.$$

Since $v(t)$ is increasing, it follows that

$$\begin{aligned} v(t_1)^2u'(t_1)^2 &\le v(\bar{t_1})^2u'(t_0)^2 \\ &\le v(t_2)^2u'(t_0)^2. \end{aligned} \tag{13}$$

Then,

$$u'(t_1)^2 \le \frac{v(t_2)^2}{v(t_0)^2}u'(t_0)^2.$$

Using the same argument, one shows by induction that

$$u'(t_k)^2 \le \frac{v(t_{k+1})^2v(t_k)^2}{v(t_1)^2v(t_0)^2}u'(t_0)^2.$$

Since $r(t) < R$, we find that

$$u'(t_k)^2 \le \frac{R^{4(n-1)}}{v(t_0)^2v(t_1)^2}u'(t_0)^2. \tag{14}$$

Now, using Lemma 5, it's easy to check that

$$\begin{aligned} \int_{t_0}^{t_{3p}} u^2v\,dt &= \sum_{k=0}^{3p-1}\int_{t_k}^{t_{k+1}} u^2v(t)\,dt \\ &\le \frac{1}{\lambda}\sum_{k=0}^{3p-1} u'(t_k)^2\int_{t_k}^{t_{k+1}} \sin^2\left(\lambda^{1/2}v(t_k)\int_{t_k}^{t}\frac{ds}{v(s)}\right)v(t)\,dt. \end{aligned} \tag{15}$$

Letting

$$\tau = \lambda^{1/2}v(t_k)\int_{t_k}^{t}\frac{ds}{v(s)},$$

the change of variables formula shows that

$$\begin{aligned}\frac{1}{\lambda}\sum_{k=0}^{3p-1} u'(t_k)^2 \int_{t_k}^{t_{k+1}} \sin^2\Bigl(\lambda^{1/2}v(t_k)\int_{t_k}^{t}\frac{ds}{v(s)}\Bigr)v(t)\,dt \\ &= \frac{1}{\lambda^{3/2}}\sum_{k=0}^{3p-1}\frac{u'(t_k)^2}{v(t_k)}\int_0^{\pi}\sin^2(\tau)v^2(\tau(t))\,d\tau \\ &\le \frac{\pi R^{2(n-1)}}{2\lambda^{3/2}r^{n-1}(t_0)}\sum_{k=0}^{3p-1}u'(t_k)^2 \\ &= C\sum_{k=0}^{3p-1}u'(t_k)^2.\end{aligned} \tag{16}$$

By (7) and (14), the following inequalities hold:

$$\begin{aligned}\sum_{k=0}^{3p-1}u'(t_k)^2 &\le 3Cpu'(t_0)^2 \\ &\le C\sum_{k=p}^{2p-1}u'(t_k)^2.\end{aligned} \tag{17}$$

We have

$$\int_{t_0}^{t_{3p}} u^2 v\,dt \le C\sum_{k=p}^{2p-1}u'(t_k)^2. \tag{18}$$

Here, the last inequality comes from (7), for some suitable constant $C > 0$. Again by the change of variables formula (this time applied to each α_k) and by Lemma 5, one sees that if $\tilde{t}_k$ is the next zero of α_k after t_k we have

$$\begin{aligned}\int_{t_p}^{t_{2p}} u^2v(t)\,dt &= \sum_{k=p}^{2p-1}\int_{t_k}^{t_{k+1}}u^2v(t)\,dt \\ &\ge \sum_{k=p}^{2p-1}\int_{t_k}^{\tilde{t}_{k+1}}\alpha_k^2v(t)\,dt \\ &\ge C\sum_{k=p}^{2p-1}u'(t_k)^2.\end{aligned} \tag{19}$$

From (18) we conclude that

$$\int_{t_0}^{t_{3p}} u^2r^{n-1}\,dt \le C\int_{t_p}^{t_{2p}}u^2r^{n-1}\,dt$$

for every $p \in \mathbb{N}$ and for a constant $C = C(\lambda, R)$, independent of p.

Appendix: Elements of Sturm–Liouville theory

For the convenience of the reader, we present some facts about Sturm–Liouville problems used in the previous section. Our motivation relies on the study of

$$(v(t)u')' + \lambda v(t)u = 0 \quad t \geq t_0 > 0, \tag{20}$$

where $v(t) = r^{n-1}(t)$ for fixed $n \in \mathbb{N}$. In the following we assume the function $r(t)$ to be positive; moreover:

(I) $0 < r'(t) \leq c$.

(II) $\lim\limits_{t \to \infty} r(t) = R < +\infty$.

We start with a classical terminology.

Definition 6. Equation (20) is said to be oscillatory if any of its solutions has arbitrarily large zeros.

The following theorem is a practical criterion for oscillation.

Theorem 7. *Let $v(t)$ be a positive continuous function on $[t_0, \infty)$ and $\lambda > 0$. Then, the equation*

$$(v(t)u')' + \lambda v(t)u = 0$$

for $t \geq t_0$ is oscillatory, provided $\int_{t_0}^{\infty} v(t)\,dt = +\infty$ and $\int_{t_0}^{t} v(t)\,dt \leq Ct^a$, for some positive constants C and a.

The proof is discussed in [do Carmo and Zhou 1999, Theorem 2.1]. Since $\lim_{t \to \infty} r(t) = R$, we easily have the following.

Corollary 8. *Equation* (20) *is oscillatory.*

Theorem 9. *For positive v, any solution u of* (20) *on a interval $[t_0, t_0 + \delta]$ with initial values $u(t_0) = x_0$ and $u'(t_0) = x_1$ can be extended to $[t_0, \infty)$.*

Again, the proof is presented in [do Carmo and Zhou 1999, Theorem 2.2].

The next propositions appear in the literature as Sturm comparison theorems; see [Hartman 1982, Theorem 3.1]. These are standard results, but for the sake of self-containment we decided to present their proofs. They emerge as useful ways to compare solutions for ordinary differential equations, as we did in Section 2.

Proposition 10. *Let x, y be nontrivial solutions for*

$$\begin{cases} (p(t)x')' + q(t)x = 0, \\ (p_1(t)y')' + q_1(t)y = 0, \end{cases}$$

where $p(t) \geq p_1(t) > 0$ and $q_1(t) \geq q(t)$ for every $t \in I$. If $t_1 < t_2$ are consecutive zeros of x, then either y has a zero on $J = (t_1, t_2)$ or there is a $d \in \mathbb{R}$ for which $y = dx$ on J.

Proof. As a starting point, note that if $y(t_i) = 0$, then by uniqueness we have $y = dx$ for $d = y'(t_i)/x'(t_i)$. Uniqueness also implies that the set of zeroes of x does not have a cluster point, so the interval J is well-defined. Therefore, it is enough to consider the case where x and y are linearly independent. Observe that if y does not have a zero on J, then

$$\left(x\frac{(p(t)x'y - p_1(t)xy')}{y}\right)' = (q_1 - q)x^2 + (p - p_1)(x')^2 + \frac{p_1(x'y - xy')^2}{y^2}.$$

Integrating from t_1 to t_2, we have

$$\int_{t_1}^{t_2}(q_1 - q)x^2\,dt + \int_{t_1}^{t_2}(p - p_1)(x')^2\,dt + \int_{t_1}^{t_2} p_1\frac{(x'y - xy')^2}{y^2}\,dt = 0.$$

Then, if y is not multiple of x, the Wronskian $(xy' - x'y)$ is nonzero on J and we get a contradiction with the last equation. □

As a consequence, we obtain a universal estimate from below to the distance between two consecutive zeros of a solution of (20).

Corollary 11. *Let $\{t_p\}_{p=1}^{\infty}$ be an increasing sequence of zeros of u. There is a universal constant $C > 0$ such that $t_{p+1} - t_p > C$ for any $p \in \mathbb{N}$.*

Proof. Given $p \in \mathbb{N}$, define $\varphi(t) = \sin(2^{(n-1)/2}\lambda^{1/2}(t - t_p))$. Then, φ has a zero at $t = t_p$ and

$$\left(\tfrac{1}{2}R\right)^{n-1}\varphi'' + \lambda R^{n-1}\varphi = 0.$$

Now, $\left(\frac{1}{2}R\right)^{n-1} < v(t) < R^{n-1}$ for t sufficiently large, lets say for $t > c_0$. As a consequence, if p is sufficiently large, we can apply Proposition 10 for u and φ to conclude that the next zero of φ is on (t_p, t_{p+1}).

Since the next zero of φ after t_p is on $t = t_p + \pi/(2^{(n-1)/2}\lambda)$, we have

$$t_{p+1} - t_p > \frac{\pi}{2^{(n-1)/2}\lambda}$$

for $t_p > c_0$, from which the corollary follows. □

Proposition 12. *Let x, y be nontrivial solutions for*

$$\begin{cases} (p(t)x')' + q(t)x = 0, \\ (p_1(t)y')' + q_1(t)y = 0, \end{cases}$$

on an interval $[a, b]$, where $p \geq p_1 > 0$, $q_1 > q$ and $x(a) = 0$. Suppose that $c \in (a, b]$ is such that $x(c) \neq 0$, $y(c) \neq 0$ and x has the same number of zeros as y

on (a, c). *Then*

$$\frac{p(c)x'(c)}{x(c)} \geq \frac{p_1(c)y'(c)}{y(c)}.$$

Proof. We only deal with the case where y is different from dx, otherwise there is nothing to prove. Let $a = a_0, \dots, a_n$ be the zeros of x on $[a, c)$ and $b_0, \dots, b_{n-1}$ be the zeros of y on (a, c). By Proposition 10, we have

$$a_i < b_i < a_{i+1}$$

for $i = 0, \dots, n-1$. Consequently, y has no zero on (a_n, c). Now, we can use the same idea from the proof of Proposition 10 to conclude that

$$\Big((px'y - p_1xy')\frac{x}{y}\Big)' \geq 0$$

on (a_n, c). Integrating both sides from a_n to c and using that $x(a_n) = 0$, we get

$$(px'y - p_1xy')(c)\frac{x(c)}{y(c)} \geq 0,$$

and since we can always assume that $x(c)y(c) > 0$, we find

$$\frac{p(c)x'(c)}{x(c)} \geq \frac{p_1y'(c)}{y(c)}. \qquad \square$$

Proof of Lemma 5. Observe that $\alpha_k(t_k) = 0$, $\alpha_k'(t_k) = u_k'(t_k)$ and

$$(v(t)\alpha_k')' + \lambda\frac{R^{2(n-1)}}{v(t)}\alpha_k = 0.$$

Since

$$\frac{R^{2(n-1)}}{v(t)} \geq R^{n-1} \geq v(t)$$

for all $t \geq t_k$, we can apply Proposition 12 to u and α_k and establish that

$$\frac{u'(t)}{u(t)} \geq \frac{\alpha_k'(t)}{\alpha_k(t)}, \quad t \in (t_k, \tilde{t}_k).$$

So, taking $\epsilon > 0$ and integrating the inequality above from $t_k + \epsilon$ to t, we get

$$\log\bigg(\frac{|u(t)|}{|u(t_k+\epsilon)|}\bigg) \geq \log\bigg(\frac{|\alpha_k(t)|}{|\alpha_k(t_k+\epsilon)|}\bigg),$$
$$\frac{|u(t)|}{|\alpha_k(t)|} \geq \frac{|u(t_k+\epsilon)|}{|\alpha_k(t_k+\epsilon)|}.$$

Sending $\epsilon \to 0$ and using that $u'(t_k) = \alpha_k'(t_k) \neq 0$, we find $|\alpha_k| \leq |u|$.

The proof of the other inequality follows the same ideas and is omitted.

Acknowledgements

This work is part of the first author's senior thesis at Universidade Federal do Ceará, Brazil, supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). The second author is partially supported by CNPq. We are grateful to the anonymous referee for pointing out the importance of the reference [Hartman 1982] as well as observing that our method supports more intrinsic results.

References

[Bessa et al. 2010] G. P. Bessa, L. P. Jorge, and J. F. Montenegro, "The spectrum of the Martin–Morales–Nadirashvili minimal surfaces is discrete", *J. Geom. Anal.* **20**:1 (2010), 63–71. MR Zbl

[Bessa et al. 2012] G. P. Bessa, J. F. Montenegro, and P. Piccione, "Riemannian submersions with discrete spectrum", *J. Geom. Anal.* **22**:2 (2012), 603–620. MR Zbl

[Bessa et al. 2015] G. P. Bessa, L. P. Jorge, and L. Mari, "On the spectrum of bounded immersions", *J. Differential Geom.* **99**:2 (2015), 215–253. MR Zbl

[do Carmo and Zhou 1999] M. P. do Carmo and D. Zhou, "Eigenvalue estimate on complete noncompact Riemannian manifolds and applications", *Trans. Amer. Math. Soc.* **351**:4 (1999), 1391–1401. MR Zbl

[Chen and Lu 1992] Z. H. Chen and Z. Q. Lu, "Essential spectrum of complete Riemannian manifolds", *Sci. China Ser. A* **35**:3 (1992), 276–282. MR Zbl

[Davies 1995] E. B. Davies, *Spectral theory and differential operators*, Cambridge Studies in Advanced Mathematics **42**, Cambridge University Press, 1995. MR Zbl

[Donnelly 1981] H. Donnelly, "On the essential spectrum of a complete Riemannian manifold", *Topology* **20**:1 (1981), 1–14. MR Zbl

[Donnelly and Garofalo 1992] H. Donnelly and N. Garofalo, "Riemannian manifolds whose Laplacians have purely continuous spectrum", *Math. Ann.* **293**:1 (1992), 143–161. MR Zbl

[Donnelly and Li 1979] H. Donnelly and P. Li, "Pure point spectrum and negative curvature for noncompact manifolds", *Duke Math. J.* **46**:3 (1979), 497–503. MR Zbl

[Escobar and Freire 1992] J. F. Escobar and A. Freire, "The spectrum of the Laplacian of manifolds of positive curvature", *Duke Math. J.* **65**:1 (1992), 1–21. MR Zbl

[Grigor'yan 2009] A. Grigor'yan, *Heat kernel and analysis on manifolds*, AMS/IP Studies in Advanced Mathematics **47**, American Mathematical Society, Providence, RI, 2009. MR Zbl

[Hartman 1982] P. Hartman, *Ordinary differential equations*, revised 2nd ed., Birkhäuser, Boston, 1982. MR Zbl

[Karp 1984] L. Karp, "Noncompact Riemannian manifolds with purely continuous spectrum", *Michigan Math. J.* **31**:3 (1984), 339–347. MR Zbl

[Kleine 1988] R. Kleine, "Discreteness conditions for the Laplacian on complete, noncompact Riemannian manifolds", *Math. Z.* **198**:1 (1988), 127–141. MR Zbl

[Kleine 1989] R. Kleine, "Warped products with discrete spectra", *Results Math.* **15**:1–2 (1989), 81–103. MR Zbl

[Kumura 1997] H. Kumura, "On the essential spectrum of the Laplacian on complete manifolds", *J. Math. Soc. Japan* **49**:1 (1997), 1–14. MR Zbl

[Li 1994] J. Y. Li, "Spectrum of the Laplacian on a complete Riemannian manifold with nonnegative Ricci curvature which possess a pole", *J. Math. Soc. Japan* **46**:2 (1994), 213–216. MR Zbl

[Lu and Zhou 2011] Z. Lu and D. Zhou, "On the essential spectrum of complete non-compact manifolds", *J. Funct. Anal.* **260**:11 (2011), 3283–3298. MR Zbl

[Monte and Montenegro 2015] L. A. C. Monte and J. F. B. Montenegro, "Essential spectrum of a class of Riemannian manifolds", *J. Geom. Anal.* **25**:4 (2015), 2241–2261. MR Zbl

[Tayoshi 1971] T. Tayoshi, "On the spectrum of the Laplace–Beltrami operator on a non-compact surface", *Proc. Japan Acad.* **47** (1971), 187–189. MR Zbl

[Wang 1997] J. Wang, "The spectrum of the Laplacian on a manifold of nonnegative Ricci curvature", *Math. Res. Lett.* **4**:4 (1997), 473–479. MR Zbl

[Zhou 1994] D. Zhou, "Essential spectrum of the Laplacian on manifolds of nonnegative curvature", *Int. Math. Res. Not.* 5 (1994), 209–214. MR Zbl

Received: 2016-02-22 Revised: 2016-06-26 Accepted: 2016-07-11

matosrod@msu.edu *Department of Mathematics, Michigan State University,*
East Lansing, MI 48824, United States

fabio@mat.ufc.br *Departamento de Matemática, Universidade Federal*
do Ceará (UFC), Av. Humberto Monte, s/n,
Campus do Pici, Bloco 914, 60455-760 Fortaleza-, Brazil

INVOLVE 10:4 (2017)
dx.doi.org/10.2140/involve.2017.10.691

A generalization of Eulerian numbers via rook placements

Esther Banaian, Steve Butler, Christopher Cox,
Jeffrey Davis, Jacob Landgraf and Scarlitte Ponce

(Communicated by Jim Haglund)

We consider a generalization of Eulerian numbers which count the number of placements of cn rooks on an $n \times n$ chessboard where there are exactly c rooks in each row and each column, and exactly k rooks below the main diagonal. The standard Eulerian numbers correspond to the case $c = 1$. We show that for any c the resulting numbers are symmetric and give generating functions of these numbers for small values of k.

1. Introduction

Rook placements on boards have a wonderful and rich history in combinatorics (see, e.g., [Butler et al. 2011]). Traditionally the rooks are placed in a nonattacking fashion (i.e., at most one rook in each row and column) and the combinatorial aspects come from considering variations on the board shapes.

Instead of varying the board, we could also change the restrictions on how many rooks are allowed in each row and each column. If we have a square board and the number of rooks in each column and row is fixed, then this corresponds to counting nonnegative matrices with fixed row and column sums; see A000681, A001500, A257493, etc., in the On-Line Encyclopedia of Integer Sequences (OEIS).

In this paper, we will look at this latter case of placing multiple rooks in each row and column more closely. We begin in Section 2 by exploring the connections between these rook placements and juggling patterns. In Section 3, we look at Eulerian numbers (which correspond to the number of nonattacking rook placements on an $n \times n$ board with a fixed number of rooks below the main diagonal) and in Section 4 generalize to the case in which c rooks are placed in each row and each column. In Section 5, we provide generating functions for special cases of these generalized Eulerian numbers. We end with concluding remarks and open problems in Section 6.

MSC2010: 05A15.
Keywords: Eulerian numbers, juggling, recursion, multiplex.

2. Minimal juggling patterns and rook placements

Juggling patterns can be described by a *siteswap* sequence listing the throws that the pattern requires, i.e., $t_1t_2\cdots t_n$ where at time $s \equiv i \pmod n$ we throw the ball so that it will land t_i beats in the future. A sequence of throws can be juggled if and only if there are no collisions, i.e., two balls landing at the same time, which is equivalent to $1+t_1, 2+t_2, \ldots, n+t_n$ being distinct modulo n. One well-known property of siteswap sequences is that the average of the throws is the number of balls needed to juggle the pattern (see [Buhler et al. 1994; Polster 2003]).

A *minimal juggling pattern* is a valid juggling pattern $t_1t_2\cdots t_n$ with $0 \le t_i \le n-1$. These form the basic building blocks of juggling patterns since all juggling patterns of period n arise by starting from some minimal juggling pattern and adding multiples of n to the various throws (such additions do not affect modular conditions). More about this approach was found by Buhler, Eisenbud, Graham and Wright [1994].

This naturally leads to the problem of enumerating minimal juggling patterns. This is done by relating such patterns to rook placements on a square board. In particular we will consider the $n \times n$ board $\mathcal{B}_n$, with labels on each cell (i, j) given by the rule

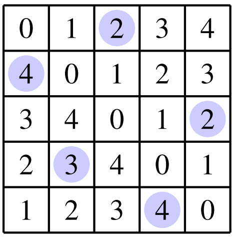
$$\begin{cases} j-i & \text{if } j \ge i, \\ n+j-i & \text{if } j < i. \end{cases}$$

We can interpret the rows of $\mathcal{B}_n$ as the *throwing* times (modulo n) and the columns of $\mathcal{B}_n$ as the *landing* times (modulo n). The label of the cell (i, j) is then the smallest possible throw required to throw at time i and land at time j.

Given a minimal juggling pattern $t_1t_2\cdots t_n$, we form a rook placement by placing a rook in row i on the cell labeled t_i for $1 \le i \le n$ (note that this forces the rook to be placed in the column corresponding to the landing time modulo n). Since landing times are unique modulo n, no two rooks will be in the same column, so this forms a nonattacking rook placement with n rooks. Conversely, given a nonattacking rook placement with n rooks we can form a minimal juggling pattern by reading off the cell labels of the covered square starting at the first row and reading down. This establishes the bijective relationship between minimal juggling patterns and nonattacking rook patterns on $\mathcal{B}_n$. An example of this is shown in Figure 1 for the minimal juggling pattern 24234.

We can extract information about the minimal juggling pattern by properties of the rook placements, including, for example, the number of balls.

Proposition 1. *The number of rooks below the main diagonal in a nonattacking rook placement on $\mathcal{B}_n$ is the same as the number of balls necessary to juggle the corresponding minimal juggling pattern.*



0	1	2	3	4
4	0	1	2	3
3	4	0	1	2
2	3	4	0	1
1	2	3	4	0

Figure 1. A nonattacking rook placement on $\mathcal{B}_5$ corresponding to the minimal juggling pattern 24234.

Proof. Suppose there are k rooks below the main diagonal in a placement of n nonattacking rooks on $\mathcal{B}_n$. Then when we sum the labels of all the cells covered by a rook, i.e., we sum the throw heights for the juggling sequence, we have

$$\sum_{\ell=1}^{n} t_\ell = kn + \sum_{j=1}^{n} j - \sum_{i=1}^{n} i = kn.$$

Since the average of the throws is the number of balls needed for the sequence, the claim follows. □

Note that in Figure 1 there are three rooks below the main diagonal and that the juggling pattern 24234 requires three balls to juggle.

2.1. ***Multiplex juggling and c-rook placements.*** A natural variation in juggling is to allow multiple balls to be caught and thrown at a time. This is known as *multiplex juggling*, and we will see that many of the basic ideas generalize well to this setting.

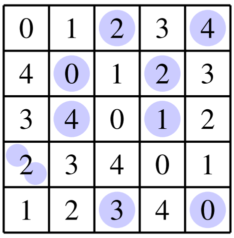
We will let c denote a hand capacity; i.e., at each beat we make c throws (allowing some of the throws to be 0, which happens when the number of actual balls thrown is less than c). Siteswap sequences of period n now correspond to a sequence of n sets, $T_1 T_2 \cdots T_n$, where each T_i is a (multi)set of the form $\{t_{i,1}, t_{i,2}, \ldots, t_{i,c}\}$, denoted in shorthand notation as $[t_{i,1} t_{i,2} \cdots t_{i,c}]$. A multiplex juggling sequence is valid if and only if the *juggling modular condition* is satisfied. Namely, every $1 \le \ell \le n$ appears exactly c times in the multiset

$$\{t_{i,j} + i \pmod{n}\}_{\substack{1\le i\le n \\ 1\le j\le c}}.$$

In other words, no more than c balls land at each time.[1] As in standard juggling patterns, the number of balls b needed to juggle the pattern relates to an average. In particular,

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c} t_{i,j} = b.$$

[1] A 0 throw indicates a ball is not landing.



0	1	2	3	4
4	0	1	2	3
3	4	0	1	2
2	3	4	0	1
1	2	3	4	0

Figure 2. A 2-rook placement on $\mathcal{B}_5$ corresponding to the minimal multiplex juggling pattern [24][02][14][22][03].

We say a multiplex juggling sequence is a *minimal multiplex juggling sequence* if and only if $0 \le t_{i,j} \le n-1$ for all throws $t_{i,j}$.

There is a relationship between period n, hand capacity c, multiplex juggling sequences and placements of "rooks" on $\mathcal{B}_n$. This is done by generalizing from nonattacking rook placements to *c-rook placements*, placements of cn rooks with exactly c rooks in every row and column, where multiple rooks are allowed in cells.

There is a bijection between minimal multiplex juggling patterns of period n with hand capacity c and c-rook placements on $\mathcal{B}_n$. In particular, for each i we place c rooks in the i-th row corresponding to $t_{i,1}, \dots, t_{i,c}$. Conversely, given a c-rook placement we can form a minimal multiplex juggling sequence by letting T_i denote the cells covered by the rooks in row i (with appropriate multiplicity). An example of this is shown in Figure 2 for the minimal multiplex juggling pattern [24][02][14][22][03].

By the same argument used for Proposition 1 we have the following.

Proposition 2. *The number of rooks below the main diagonal in a c-rook placement on $\mathcal{B}_n$ is the same as the number of balls necessary to juggle the corresponding minimal multiplex juggling pattern.*

For example, the multiplex juggling pattern in Figure 2 requires four balls to juggle.

3. Eulerian numbers

The Eulerian numbers, denoted $\left\langle {n \atop k} \right\rangle$, are usually defined as the number of permutations of $[n]$, $\pi = \pi_1\pi_2\cdots\pi_n$, with k ascents ($\pi_i < \pi_{i+1}$), or equivalently the number of permutations with k descents ($\pi_i > \pi_{i+1}$). There is a bijection between permutations of $[n]$ with k descents and permutations with k drops ($i > \pi_i$), so that $\left\langle {n \atop k} \right\rangle$ also counts permutations of $[n]$ with k drops (see [Buhler et al. 1994]). Given an $n \times n$ board, with rows and columns labeled $1, 2, \dots, n$, we can use our permutation to form a nonattacking rook placement by placing rooks at positions (i, π_i). A drop

$n \downarrow k \rightarrow$	0	1	2	3	4	5	6
1	1						
2	1	1					
3	1	4	1				
4	1	11	11	1			
5	1	26	66	26	1		
6	1	57	302	302	57	1	
7	1	120	1191	2416	1191	120	1

Table 1. The Eulerian numbers $\left\langle {n \atop k} \right\rangle$ for $1 \le n \le 7$.

in the permutation corresponds to a rook below the main diagonal, so we will call any rook below the main diagonal a *drop*.

By Proposition 1, the number of drops in a nonattacking rook placement equals the number of balls necessary for the corresponding juggling pattern. Therefore, $\left\langle {n \atop k} \right\rangle$ also counts the number minimal juggling patterns of period n using k balls.

The Eulerian numbers have many nice properties, some of which can be seen in Table 1. For example, they are symmetric, i.e., $\left\langle {n \atop k} \right\rangle = \left\langle {n \atop n-k-1} \right\rangle$. This can be shown by noting if we start with a permutation with k ascents and reverse the permutation, we now have $n-1-k$ ascents (i.e., ascents go to descents and vice-versa; and there are $n-1$ consecutive pairs). We will give a different proof of this symmetry in the next section using rook placements.

Another well-known property of the Eulerian numbers is a recurrence relation.

Proposition 3. *The Eulerian numbers satisfy* $\left\langle {n \atop k} \right\rangle = (n-k)\left\langle {n-1 \atop k-1} \right\rangle + (k+1)\left\langle {n-1 \atop k} \right\rangle$.

This recurrence is again proven using permutations and ascents. Here, we provide an alternate proof using rook placements and drops.

Proof. Start by considering a nonattacking rook placement on an $(n-1) \times (n-1)$ board with $k-1$ drops. Add an n-th row and n-th column, and place a rook in position (n, n). The newly added rook is not below the diagonal and so we have not created any new drops. We can now create one additional drop by taking any rook (other than the one just added) which is *on or above* the main diagonal, say in position (i, j), move that rook to position (n, j) and move the rook in position (n, n) to position (i, n). This moves the rook in the j-th column below the main diagonal creating a new drop. Since no other rook moves, we now have precisely k drops and a nonattacking rook placement. Note that there are $(n-1)-(k-1) = n-k$ ways we could have chosen which rook to move, so that in total this gives $(n-k)\left\langle {n-1 \atop k-1} \right\rangle$ boards of size $n \times n$ with k drops.

Now, consider a nonattacking rook placement on an $(n-1) \times (n-1)$ board with k drops. Add an n-th row and n-th column, and place a rook in position (n, n).

As before we switch, but now only switch with a rook which is *below* the main diagonal (i.e., a drop). This will not change the number of drops, so the result is a nonattacking rook placement on an $n \times n$ board with k drops. There are k rooks we can choose to switch with, or alternatively, we can leave the n-th rook in position (n, n); thus, there are $k+1$ ways to build the desired rook placement, so that in total this gives $(k+1)\genfrac{\langle}{\rangle}{0pt}{}{n-1}{k}$ boards of size $n \times n$ with k drops.

Finally, we note that each $n \times n$ board with k drops is formed uniquely from one of these operations. This can be seen by taking such a board and then noting the location of the rook(s) in the last row and in the last column. Suppose these are in positions (i, n) and (n, j), respectively. We then move these rooks to positions (i, j) and (n, n). This can at most decrease the number of drops by one (i.e., moving the rook in the last column does not affect the number of drops). Now removing the last row and column gives an $(n-1) \times (n-1)$ board having a nonattacking rook placement with either k or $k-1$ drops. $\square$

4. Generalized Eulerian numbers

The *generalized Eulerian numbers*, denoted $\genfrac{\langle}{\rangle}{0pt}{}{n}{k}_c$, are the number of c-rook placements on the $n \times n$ board with k drops. Just as the Eulerian numbers count the number of minimal juggling patterns of period n with k balls, the generalized Eulerian numbers count the number of minimal multiplex juggling patterns of period n with k balls and hand capacity c. Notice that the generalized Eulerian numbers reduce to the Eulerian numbers when $c = 1$. In Table 2 we give some of the generalized Eulerian numbers for $c = 2$ and 3.

$n \downarrow k \rightarrow$	0	1	2	3	4	5	6	7	8
1	1								
2	1	1	1						
3	1	4	11	4	1				
4	1	11	72	114	72	11	1		
5	1	26	367	1492	2438	1492	367	26	1

$n \downarrow k \rightarrow$	0	1	2	3	4	5	6	7	8	9
1	1									
2	1	1	1	1						
3	1	4	11	23	11	4	1			
4	1	11	72	325	595	595	325	72	11	1

Table 2. Small values of the generalized Eulerian $\genfrac{\langle}{\rangle}{0pt}{}{n}{k}_c$ numbers for $c = 2$ (top) and 3 (bottom).

These numbers appear to satisfy a symmetry property similar to Eulerian numbers. We will give two proofs of this symmetry, one in terms of rook placements and the other using minimal multiplex juggling patterns.

Theorem 4. *Let n, k and c be nonnegative integers. Then $\left\langle {n \atop k} \right\rangle_c = \left\langle {n \atop c(n-1)-k} \right\rangle_c$.*

Proof. We construct a bijection between the rook placements with k rooks below the main diagonal and those with $c(n-1)-k$ rooks below the diagonal. Consider a rook placement with c rooks in every row and column, and k rooks below the diagonal. Now, shift every rook one space to the right cyclically. Let us consider the number of rooks which are *strictly* above the main diagonal.

- All c rooks in the last column were shifted to the first column. So, none of these rooks are above the main diagonal.
- All of the k rooks that were initially below the main diagonal are now either on or still below the main diagonal.
- All other rooks will be above the diagonal.

Since there are cn rooks on the board total, there are $cn-c-k = c(n-1)-k$ rooks above the diagonal after this shift. Finally, we switch the rows and columns of the board. This flips the rook placement across the main diagonal. After this transformation, there are now $c(n-1)-k$ rooks *below* the main diagonal. This composition of transformations is invertible by switching rows and columns then shifting every rook left one space. Thus, the transformation gives a bijection, completing the proof. □

Before we can give the second proof, we must first establish some basic properties of (multiplex) juggling sequences.

Lemma 5. *If the sequence $T_1T_2\cdots T_n$ satisfies the juggling modular conditions with hand capacity c, and $\alpha \in \mathbb{Z}_n$ with $\gcd(\alpha, n) = 1$, then $(\alpha T_{1\alpha^{-1}})(\alpha T_{2\alpha^{-1}})\cdots(\alpha T_{n\alpha^{-1}})$, where*

$$\alpha T_i := \{\alpha t_{i,1}, \dots, \alpha t_{i,c}\},$$

and the subscripts are taken modulo n, also satisfies the juggling modular conditions.

Proof. We have

$$A = \{\alpha t_{i\alpha^{-1},j} + i\}_{\substack{1\le i\le n\\ 1\le j\le c}} = \{\alpha(t_{i\alpha^{-1},j} + i\alpha^{-1})\}_{\substack{1\le i\le n\\ 1\le j\le c}} = \{\alpha(t_{i',j} + i')\}_{\substack{1\le i'\le n,\\ 1\le j\le c}}$$

where we use that $\gcd(\alpha, n) = 1$ so that α is invertible modulo n and as i ranges between 1 and n, then so does $i' := i\alpha^{-1}$. Since $\{t_{i,j} + i\}_{1\le i\le n, 1\le j\le c}$ has c occurrences each of 1 through n, then scaling by α and taking terms modulo n we also have that A will have c occurrences each of 1 through n. □

Lemma 6. *If the sequence $T_1T_2\cdots T_n$ satisfies the juggling modular conditions of hand capacity c, and $\beta\in\mathbb{Z}$, then $(T_1+\beta)(T_2+\beta)\cdots(T_n+\beta)$, where*

$$T_i+\beta := [t_{i,1}+\beta, t_{i,2}+\beta, \dots, t_{i,c}+\beta],$$

still satisfies the juggling modular conditions.

Proof. The multiset

$$A=\{(t_{i,j}+\beta)+i\}_{\substack{1\le i\le n\\ 1\le j\le c}}$$

is found by taking $\{t_{i,j}+i\}_{1\le i\le n, 1\le j\le c}$ and shifting each element by β. Since $T_1T_2\cdots T_n$ satisfy the juggling modular conditions then so also must A. □

Juggling proof of Theorem 4. We show there is a bijection between the minimal multiplex juggling sequences using k balls and those using $c(n-1)-k$ balls for a fixed length n and hand capacity c. So let $T_1T_2\cdots T_n$ be a valid minimal multiplex juggling sequence with k balls and hand capacity c. By Lemma 5 and Lemma 6, if we scale each T_i by -1 (reversing the indexing) and add $n-1$ then the resulting sets still satisfy the modular juggling conditions. In particular we have that the following satisfies the modular juggling condition:

$$(n-1-T_n)(n-1-T_{n-1})\cdots(n-1-T_1).$$

We also note the resulting throws all lie between 0 and $n-1$ so that this is indeed a minimal juggling pattern.

The number of balls in the new juggling sequence is

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}((n-1)-t_{i,j})=\frac{1}{n}\left(cn(n-1)-\sum_{i=1}^{n}\sum_{j=1}^{c}t_{i,j}\right)=c(n-1)-k.$$

Finally, we note that this operation is its own inverse, and thus gives the desired bijection. □

5. Generalized Eulerian numbers for small k

We now look at determining the values of the generalized Eulerian numbers $\left\langle {n \atop k} \right\rangle_c$ for small k. This depends of course on both n and c. However, for a fixed k there are only finitely many c that need to be considered. This is a consequence of the following lemma.

Lemma 7. *For $c\ge k$ we have $\left\langle {n \atop k} \right\rangle_c=\left\langle {n \atop k} \right\rangle_k$.*

Proof. It will suffice to establish the following claim.

Claim. Every c-rook placement with k drops has at least $c-k$ rooks in every entry on the main diagonal.

We proceed to establish this by using induction on $k+c$. For $k+c=1$, the only possible case is $k=0$ and $c=1$ for which there is only one placement, namely one rook in each cell on the main diagonal.

Now assume that we have established the claim for all k, c with $k+c<\ell$, and let $k+c=\ell$. Let S be a c-rook placement with k drops. We can interpret the rook placement as an incidence relationship of a regular bipartite graph. By Hall's marriage theorem, we know we can find a perfect matching in this bipartite graph which corresponds to T, a 1-rook placement contained in S. Suppose there are i drops in T. Then, $S-T$ is a $(c-1)$-rook placement with $k-i$ drops. Since $(c-1)+(k-i)<c+k=\ell$, by our induction hypothesis, there are at least $c-k+i-1$ rooks on each entry on the main diagonal in $S-T$, and hence also in S. If $i\geq 1$, we are done. If $i=0$, then T is again the unique 1-rook placement where every rook is on the main diagonal, so S still has at least $c-k$ rooks on each entry on the main diagonal. □

This can also be established in terms of minimal multiplex juggling patterns.

Juggling proof of Lemma 7. If there are k balls, then at each step we can throw at most k balls, i.e., each T_i has at least $c-k$ entries of 0. It follows that in the corresponding c-rook placement each row has at least $c-k$ rooks on the diagonal. □

We will be looking at the generalized Eulerian numbers $\left\langle {n \atop k} \right\rangle_c$ for $k=1,2,3$. By Lemma 7 this reduces down to only six cases to consider, namely, $\left\langle {n \atop 1} \right\rangle_1$, $\left\langle {n \atop 2} \right\rangle_1$, $\left\langle {n \atop 2} \right\rangle_2$, $\left\langle {n \atop 3} \right\rangle_1$, $\left\langle {n \atop 3} \right\rangle_2$ and $\left\langle {n \atop 3} \right\rangle_3$. Since $\left\langle {n \atop k} \right\rangle_1=\left\langle {n \atop k} \right\rangle$, the cases $\left\langle {n \atop 1} \right\rangle_1$, $\left\langle {n \atop 2} \right\rangle_1$ and $\left\langle {n \atop 3} \right\rangle_1$ have been previously determined (see A000295, A000460 and A000498, respectively, in the OEIS). So that leaves $\left\langle {n \atop 2} \right\rangle_2$, $\left\langle {n \atop 3} \right\rangle_2$ and $\left\langle {n \atop 3} \right\rangle_3$ and in Table 3 we give the generating function for these three sequences. In the remainder of this section we will demonstrate the techniques used to determine the generating functions by working through the case for $\left\langle {n \atop 2} \right\rangle_2$.

5.1. ***Placing rooks in a generic rook placement.*** We break the problem of counting c-rook placements into several subproblems according to the way the rooks below the main diagonal are placed relative to one another (i.e., relative placements instead of absolute placements). Given some generic placement of the k rooks below the main diagonal we can determine the number of ways to place the remaining rooks on or above the main diagonal. We then combine the results over all possible generic placements.

We will carefully work through the rook placement shown in Figure 3, which consists of two rooks below the main diagonal and where both rooks are in the same column and different rows. Here a, b, c and d are the number of rows between the various transition points (a transition point to passing a rook, or rooks, in a row or a column as we move along the main diagonal).

We place the remaining rooks one row at a time starting from the bottom and going to the top. For each new row, the way we place rooks will depend on all of

$$\begin{aligned}\sum_{n\geq 0}\left\langle {n \atop 2} \right\rangle_2 x^n &= x^2+11x^3+72x^4+367x^5+1630x^6+6680x^7+26082x^8+\cdots\\ &= \frac{x^2-x^3-x^4-3x^5+5x^6}{(1-x)^3(1-2x)^2(1-5x+5x^2)}\end{aligned}$$

$$\begin{aligned}\sum_{n\geq 0}\left\langle {n \atop 3} \right\rangle_2 x^n &= 4x^3+114x^4+1492x^5+13992x^6+109538x^7+769632x^8+\cdots\\ &= \frac{\begin{array}{r}4x^3+2x^4-300x^5+1748x^6-4676x^7+7058x^8\\ -6648x^9+4397x^{10}-2206x^{11}+625x^{12}\end{array}}{(1-x)^4(1-2x)^3(1-5x+5x^2)^2(1-8x+13x^2)}\end{aligned}$$

$$\begin{aligned}\sum_{n\geq 0}\left\langle {n \atop 3} \right\rangle_3 x^n &= x^2+23x^3+325x^4+3368x^5+28819x^6+218788x^7+\cdots\\ &= \frac{\begin{array}{r}x^2-7x^3+39x^4-336x^5+1844x^6-5545x^7+9697x^8\\ -10404x^9+7532x^{10}-4558x^{11}+2435x^{12}-700x^{13}\end{array}}{(1-x)^4(1-2x)^3(1-5x+5x^2)^2(1-10x+27x^2-20x^3)}\end{aligned}$$

Table 3. Generating functions for some of the generalized Eulerian numbers.

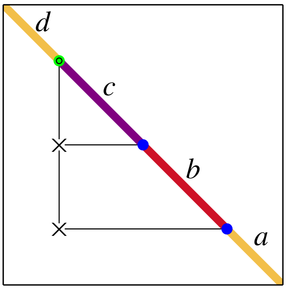


Figure 3. A 2-rook placement with two rooks below the main diagonal where both rooks are in the same column and different rows.

the choices we have made previously. However, it suffices to know only what is happening locally. In particular, we only need to know how many columns can have rooks placed into them, as well as the respective numbers that can go into those columns. We can represent these by a partition of what we will call the *excess* (the total number of rooks that can still be placed in the columns *after* the row has had its rooks placed). As we move one row up the board we will gain a new column (from the diagonal) and the excess will change in one of several ways.

- There are no rooks below or to the left of the new diagonal cell. Initially we now have a new column that can take up to c rooks, and we place c rooks in the row. The excess remains unchanged.
- There are τ rooks below the new diagonal cell. Initially we have the new column, but that can only take up to $c-\tau$ rooks (i.e., τ rooks have already

gone into the column), and we still have to place c rooks in the row. The excess decreases by τ.

- There are σ rooks to the left of the new diagonal cell. Initially we have the new column that can take up to c rooks, and we place $c-\sigma$ rooks in the row (i.e., σ rooks have already gone into the row). The excess increases by σ.

We note that it is possible for the last two situations to occur simultaneously.

In going from row to row we will transition from partitions of the old excess to partitions of the new excess. We illustrate this with an example in which case the excesses are both 2. We indicate a column which can still have r rooks placed into it by $\boxed{r}$, then underneath look at all possible ways we can place 2 rooks into those columns, and finally note the resulting set of columns contributing to the new excess:

$$\begin{array}{cccc} \boxed{2} & \boxed{2} & & \\ 2 & 0 & \to & \boxed{2} \\ 0 & 2 & \to & \boxed{2} \\ 1 & 1 & \to & \boxed{1}\boxed{1} \end{array} \qquad\qquad \begin{array}{ccccc} \boxed{2} & \boxed{1} & \boxed{1} & & \\ 2 & 0 & 0 & \to & \boxed{1}\boxed{1} \\ 1 & 1 & 0 & \to & \boxed{1}\boxed{1} \\ 1 & 0 & 1 & \to & \boxed{1}\boxed{1} \\ 0 & 1 & 1 & \to & \boxed{2} \end{array}$$

This can be modeled by a transition matrix where the *columns* of the transition matrix correspond to the excess of the original row and the *rows* of the transition matrix correspond to the partitions of the excess of the new row:

$$\begin{array}{c|cc} & \boxed{2} & \boxed{1}\boxed{1} \\ \hline \boxed{2} & 2 & 1 \\ \boxed{1}\boxed{1} & 1 & 3 \end{array}.$$

Repeating this for all possible situations that might arise for transitioning between excesses 0, 1, or 2, we get the transition matrices in the following table:

transition from

transition to	∅	$\boxed{1}$	$\boxed{2}$	$\boxed{1}\boxed{1}$
∅	(1)	(1)	(1	1)
$\boxed{1}$	(1)	(2)	(2	3)
$\boxed{2}$	(1	(1	(2	1)
$\boxed{1}\boxed{1}$	0)	1)	1	3)

We now start below the bottom row (in one possible way) and we move up from row to row and multiply on the *left* by the transition that we perform between the two rows. At any point we stop, the resulting vector will denote the number of ways to fill up the board to that row with a particular excess. In particular, if we carry this procedure all the way to the top we will get a 1×1 matrix whose entry is the number of ways to fill in the rooks on and above the main diagonal.

For Figure 3, where we have of runs of a, b, c and d rows as well as three other transitions to make, the resulting product that gives our count is

$$(1)^d(1\ 1)\begin{pmatrix}2 & 1\\ 1 & 3\end{pmatrix}^c\begin{pmatrix}1\\1\end{pmatrix}(2)^b(1)(1)^a.$$

Finally, for this generic rook placement we sum over all possible choices of a, b, c and d that gives an $n \times n$ board, i.e.,

$$\sum_{a+b+c+d=n-3}(1)^d(1\ 1)\begin{pmatrix}2 & 1\\ 1 & 3\end{pmatrix}^c\begin{pmatrix}1\\1\end{pmatrix}(2)^b(1)(1)^a.$$

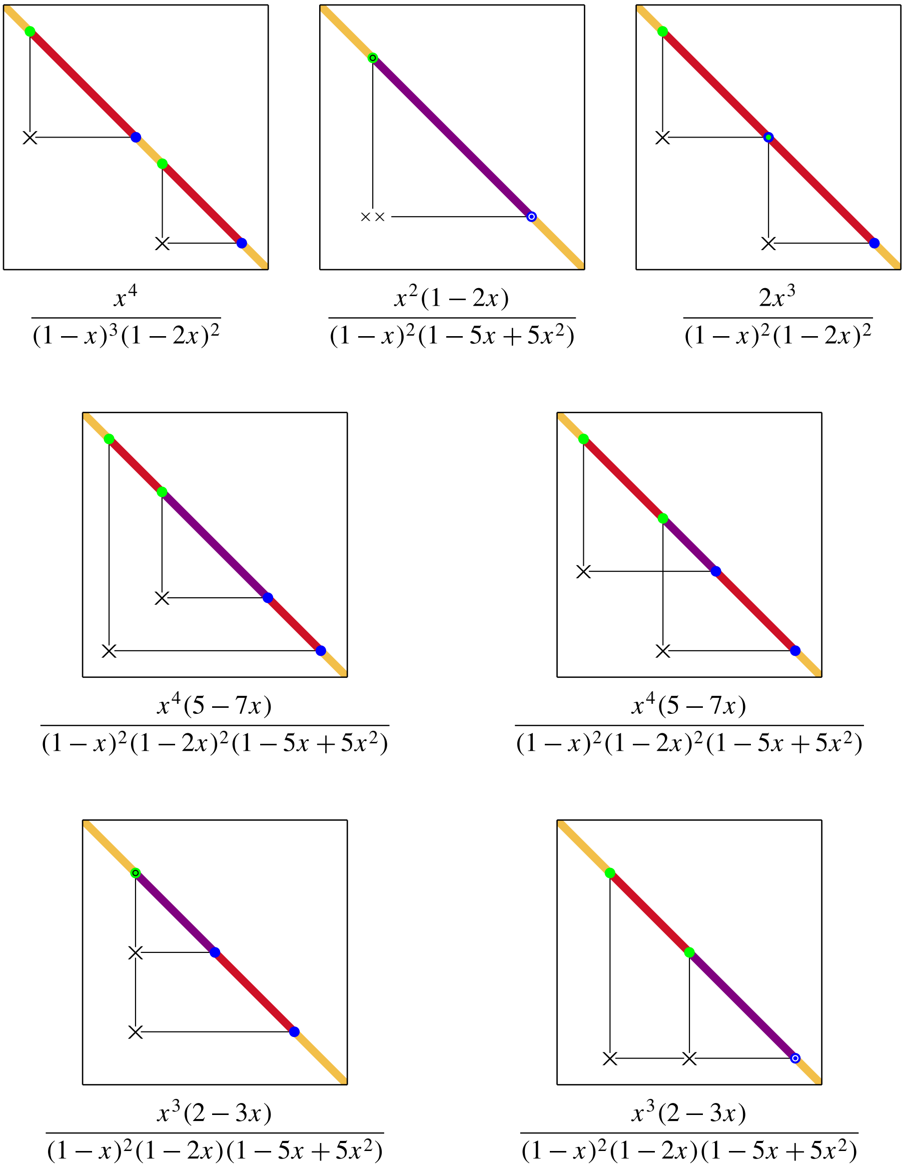
In order to help evaluate this sum, we will add in an extra parameter x that keeps track of how many of each transition we made, or viewed another way, the power of x corresponds to the number of rows we have. Therefore when counting the number of placements on an $n \times n$ board, we are interested in the coefficient of x^n of the expression

$$\sum_{a,b,c,d\geq 0}(x)^d x(1\ 1)\begin{pmatrix}2x & x\\ x & 3x\end{pmatrix}^c x\begin{pmatrix}1\\1\end{pmatrix}(2x)^b x(x)^a.$$

This sum can be decomposed as a combination of geometric sums giving

$$\begin{aligned}\sum_{a,b,c,d\geq 0}&(x)^d x\,(1\ 1)\begin{pmatrix}2x & x\\ x & 3x\end{pmatrix}^c x\begin{pmatrix}1\\1\end{pmatrix}(2x)^b x(x)^a\\ &= x^3\Big(\sum_{d\geq 0}x^d\Big)(1\ 1)\Big(\sum_{c\geq 0}\begin{pmatrix}2x & x\\ x & 3x\end{pmatrix}^c\Big)\begin{pmatrix}1\\1\end{pmatrix}\Big(\sum_{b\geq 0}(2x)^b\Big)\Big(\sum_{a\geq 0}x^a\Big)\\ &= x^3\cdot\frac{1}{1-x}\cdot(1\ 1)\Big(I-\begin{pmatrix}2x & x\\ x & 3x\end{pmatrix}\Big)^{-1}\begin{pmatrix}1\\1\end{pmatrix}\cdot\frac{1}{1-2x}\cdot\frac{1}{1-x}\\ &= \frac{x^3}{(1-x)^2(1-2x)}\cdot(1\ 1)\Big(\frac{1}{1-5x+5x^2}\begin{pmatrix}1-3x & x\\ x & 1-2x\end{pmatrix}\Big)\begin{pmatrix}1\\1\end{pmatrix}\\ &= \frac{x^3(2-3x)}{(1-x)^2(1-2x)(1-5x+5x^2)}.\end{aligned}$$

This is the generating function for one of the generic ways to place rooks. We can now repeat this procedure for every way in which we can place rooks below the main diagonal and add the individual generating functions together. All the seven generic cases, with their corresponding generating functions, are shown in Figure 4. Adding the individual generating functions together then gives us the overall generating function that was given in Table 3.



$$\frac{x^4}{(1-x)^3(1-2x)^2}$$

$$\frac{x^2(1-2x)}{(1-x)^2(1-5x+5x^2)}$$

$$\frac{2x^3}{(1-x)^2(1-2x)^2}$$

$$\frac{x^4(5-7x)}{(1-x)^2(1-2x)^2(1-5x+5x^2)}$$

$$\frac{x^4(5-7x)}{(1-x)^2(1-2x)^2(1-5x+5x^2)}$$

$$\frac{x^3(2-3x)}{(1-x)^2(1-2x)(1-5x+5x^2)}$$

$$\frac{x^3(2-3x)}{(1-x)^2(1-2x)(1-5x+5x^2)}$$

Figure 4. All generic 2-rook placements and corresponding generating functions.

This same process works for determining the generating function of $\left\langle {n \atop k} \right\rangle_c$ for any fixed k and c. The main challenge lies in that the number of generic cases that have to be considered grows drastically as we increase c and k. This is demonstrated in Table 4. It is possible to automate this process, which was used for determining the generating functions for $k=3$ given in Table 3.

$k \downarrow c \rightarrow$	1	2	3	4	5	6	7
1	1						
2	4	7					
3	26	68	75				
4	236	940	1090	1105			
5	2752	16645	20360	20790	20821		
6	39208	360081	464111	477242	478376	478439	
7	660032	9202170	12492277	12933423	12974826	12977688	12977815

Table 4. The number of generic c-rook placements with k rooks below the main diagonal.

6. Conclusion

The generalized Eulerian numbers are a natural extension of the Eulerian numbers, at least in regards to the interpretation coming from rook placements. We have also seen that these numbers exhibit a symmetry similar to that of the Eulerian numbers. It would be interesting to know which other properties and relationships involving Eulerian numbers generalize. Some natural candidates to try and generalize include the following:

- Is there a generalization of the recurrence in Proposition 3 for Eulerian numbers to generalized Eulerian numbers? Related to this, is there a simple generating function for the generalized Eulerian numbers?
- Is there a generalization of Worpitzky's identity, $x^n = \sum_k \left\langle {n \atop k} \right\rangle \binom{x+k}{n}$, to generalized Eulerian numbers? Worpitzky's identity is used in counting the number of juggling patterns (see [Buhler et al. 1994]), so a generalization might be useful in counting multiplex juggling patterns.
- Is there a generalization of the identity of Chung, Graham and Knuth [2010],
$$\sum_k \binom{a+b}{k} \left\langle {k \atop a-1} \right\rangle = \sum_k \binom{a+b}{k} \left\langle {k \atop b-1} \right\rangle ?$$
Note that this uses the convention $\left\langle {0 \atop 0} \right\rangle = 0$.

More about the Eulerian numbers and various identities and relationships that could be considered are given by Graham, Knuth and Patashnik [1994, §6.2].

We also note the original motivation for investigating these numbers was looking into the mathematics of multiplex juggling. There is a close connection between the mathematics of juggling and the mathematics of rook placements. We hope to see this relationship strengthened in future work.

Acknowledgment

The research was conducted at the 2015 REU program held at Iowa State University, which was supported by NSF DMS 1457443.

References

[Buhler et al. 1994] J. Buhler, D. Eisenbud, R. Graham, and C. Wright, "Juggling drops and descents", *Amer. Math. Monthly* **101**:6 (1994), 507–519. MR Zbl

[Butler et al. 2011] F. Butler, M. Can, J. Haglund, and J. Remmel, "Rook theory notes", 2011, available at http://www.math.ucsd.edu/~remmel/files/Book.pdf.

[Chung et al. 2010] F. Chung, R. Graham, and D. Knuth, "A symmetrical Eulerian identity", *J. Comb.* **1**:1 (2010), 29–38. MR Zbl

[Graham et al. 1994] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete mathematics: a foundation for computer science*, 2nd ed., Addison-Wesley, Reading, MA, 1994. MR Zbl

[Polster 2003] B. Polster, *The mathematics of juggling*, Springer, 2003. MR Zbl

Received: 2016-04-26 Accepted: 2016-07-11

embanaian@csbsju.edu *College of St. Benedict, Collegeville, MN 56321, United States*

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

cocox@andrew.cmu.edu *Department of Mathematical Sciences, Carnegie Mellon*
University, Pittsburgh, PA 15213, United States

jeffrey.laylon.davis@emory.edu *Department of Mathematics and Computer Science,*
Emory University, Atlanta, GA 30322, United States

jlandgr1@nd.edu *Department of Mathematics, University of Notre Dame,*
Notre Dame, IN 46556, United States

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

mathematical sciences publishers msp

INVOLVE 10:4 (2017)
dx.doi.org/10.2140/involve.2017.10.707

The H-linked degree-sum parameter for special graph families

Lydia East Kenney and Jeffrey Scott Powell

(Communicated by Jerrold Griggs)

For a fixed graph H, a graph G is H-linked if any injection $f : V(H) \to V(G)$ can be extended to an H-subdivision in G. The concept of H-linked generalizes several well-known graph theory concepts such as k-connected, k-linked, and k-ordered. In 2012, Ferrara et al. proved a sharp σ_2 (or degree-sum) bound for a graph to be H-linked. In particular, they proved that any graph G with $n > 20|E(H)|$ vertices and $\sigma_2(G) \geq n + a(H) - 2$ is H-linked, where $a(H)$ is a parameter maximized over certain partitions of $V(H)$. However, they do not discuss the calculation of $a(H)$ in their work. In this paper, we prove the exact value of $a(H)$ in the cases when H is a path, a cycle, a union of stars, a complete graph, and a complete bipartite graph. Several of these results lead to new degree-sum conditions for particular graph classes while others provide alternate proofs of previously known degree-sum conditions.

1. Introduction

We only consider finite, undirected graphs. Let G and H be graphs with vertex sets $V(G)$ and $V(H)$ and edge sets $E(G)$ and $E(H)$, respectively. Let $\mathcal{P}(G)$ denote the set of paths in G. An *H-subdivision* in G is a pair of mappings $f_1 : V(H) \to V(G)$ and $f_2 : E(H) \to \mathcal{P}(G)$ such that:

(i) f_1 is injective.

(ii) For every edge $xy \in E(H)$, the image $f_2(xy)$ in a path in G from $f_1(x)$ to $f_1(y)$ and distinct edges of H map to internally disjoint paths in G.

Note that the existence of an H-subdivision in G means that H is a topological minor of G and, as a result, H is also a minor of G. See Figure 1 for an illustration of an H-subdivision.

A graph G is *H-linked* if any injection $f : V(G) \to V(H)$ can be extended to an H-subdivision. The concept of H-linked was introduced in [Jung 1970], and

MSC2010: primary 05C35, 05C38; secondary 05C83.
Keywords: H-linked, path, cycle, degree-sum, Ore condition.

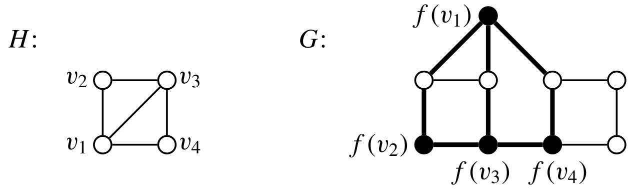


Figure 1. An H-subdivision: the vertices $v_1, v_2, v_3, \dots, v_4$ of H are mapped via an injection f to vertices in G. The subgraph in G induced by the thick edges and the vertices incident with these edges is an H-subdivision in G.

for appropriate choices of H with $|V(H)| = k$, H-linked generalizes several graph properties including k-connected, k-linked, and k-ordered.

Several recent publications have proven degree conditions for a graph to be H-linked. In [Ferrara et al. 2006; Gould et al. 2006; Kostochka and Yu 2005], sharp minimum degree conditions were proved. Degree-sum conditions were proved in [Kostochka and Yu 2008; Ferrara et al. 2012], and as this paper examines a parameter related to these conditions, we will examine them in further detail. Let $\sigma_2(G)$ denote the minimum degree sum of nonadjacent vertices in G. The minimum degree sum required to guarantee the existence of a property is known as a degree-sum condition or a σ_2 condition. Kostochka and Yu [2008] proved a sharp σ_2 condition for G to be H-linked for every graph H with minimum degree at least two.

Theorem 1.1 [Kostochka and Yu 2008]. *Let G be a graph of order n and let H be a simple graph with k edges and minimum degree at least two. If*

$$\sigma_2(G) \geq \begin{cases} \lceil n + \frac{1}{2}(3k-9) \rceil, & n > 2.5k - 5.5, \\ \lceil n + \frac{1}{2}(3k-8) \rceil, & 2k \leq n \leq 2.5k - 5.5, \\ 2n-3, & k \leq 2.5k - 1, \end{cases}$$

then G is H-linked.

Note that Theorem 1.1 provides an upper bound on the minimum degree-sum required for any possible H with minimum degree at least two, but it does not supply the optimal bound for every choice of H. A sharp σ_2 bound for this latter case was proved by Ferrara et al. [2012]. Their bound is a function of a parameter of H, called $a(H)$, that is maximized over certain partitions of $V(H)$ into two nonempty sets A and B. We use (A, B) to denote a specific partition of $V(H)$ into these two sets. Let $e(A, B)$ denote the number of edges with one vertex in A and one vertex in B. We will say that these edges "cross the partition". For a vertex v, we let $d_B(v)$ denote the number of neighbors of v in B. For the partition of H given by (A, B), let $\Delta_B(A)$ equal the maximum value of $d_B(v)$ for all $v \in A$.

We are now ready to define $a(H)$. Let

$$a(H) = \max_{\substack{A \cup B = V(H) \\ e(A,B) \geq 1}} \bigl(e(A, B) + |B| - \Delta_B(A)\bigr).$$

Using $a(H)$, one can find a sharp $\sigma_2(G)$ condition for G to be H-linked:

Theorem 1.2 [Ferrara et al. 2012]. *Let H be a simple graph and G be a graph on n vertices with $n > 20|E(H)|$. If*

$$\sigma_2(G) \geq n + a(H) - 2,$$

then G is H-linked. This result is sharp.

The same paper also gave a sharp $\sigma_2(G)$ bound for when H is a multigraph. However, in this paper, we restrict our attention to the case when H is a graph. Ferrara et al. [2012] assert that, for particular choices of H, Theorem 1.2 has (as corollaries) the previously proven σ_2 conditions for k-linked and k-ordered. However, no formal proof for these assertions is included and no further examination of the parameter $a(H)$ is presented for any particular H.

In this paper, we prove the value of $a(H)$ when H is a path, cycle, union of stars, complete graph, or complete bipartite graph. Some of these proofs specify new σ_2 conditions while others provide alternate proofs of well-known conditions. One of our aims is to supply some initial results for $a(H)$, as Theorem 1.2 could potentially be a useful tool when routing specific paths between arbitrarily chosen vertices. Additionally, we hope that these initial results for $a(H)$ encourage further study of this unusual parameter. To that end, two examples are given in the conclusion to illustrate some surprising properties of $a(H)$.

To continue, we need some further notation. For a given graph H, let $\mathbb{P}(H)$ be the set of all possible partitions of $V(H)$ into two nonempty sets with at least one edge of H that crosses the partition. For a partition $(A, B) \in \mathbb{P}(H)$, let $a(A, B) = e(A, B) + |B| - \Delta_B(A)$. Thus,

$$a(H) = \max_{(A,B) \in \mathbb{P}(H)} a(A, B).$$

For a partition (A, B), we say that F is an *induced subpartition* of H if F is an induced subgraph of H and the vertices of F are partitioned in the exact same manner in which they were partitioned in H. Note that it is possible for an induced subpartition not to have any edges that cross the partition. See Figure 2 for an illustration of these terms.

Additionally, note that for a partition (A, B), we will often speak of "moving" a vertex from A to B or from B to A. In that language, the labels A and B refer to the two sides of the partition in addition to the sets themselves. For terms and notation not defined here, see [West 1996].

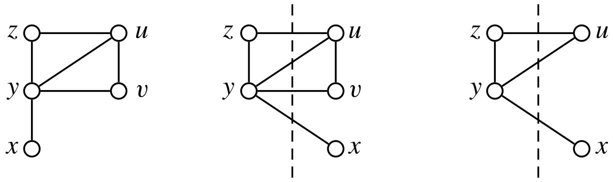


Figure 2. Suppose the graph shown on the left is *H*. The partition $(A, B) \in \mathbb{P}(H)$ with $A = \{y, z\}$ and $B = \{u, v, x\}$ is illustrated in the center. The vertical line is a visual aid to distinguish between the sets A and B. Note that in this case, $a(A, B) = 4$. The graph on the right is an induced subpartition of the partition (A, B).

2. Lemmas

To start, we prove two lemmas regarding the structure of optimal partitions of *H*, i.e., partitions $(A, B) \in \mathbb{P}(H)$ for which $a(A, B) = a(H)$. The first lemma notes that certain subpartitions cannot be induced subpartitions of an optimal partition of *H*.

Let H_1 be the induced subpartition consisting of an induced path of length two with all three vertices in A. Let H_2 be the induced subpartition consisting of an induced path of length three with one edge in A, one edge that crosses the partition, and one edge in B. See Figure 3 for H_1 and H_2.

This first lemma proves that H_1 and H_2 cannot be induced subpartitions in any optimal partition of the graph *H*.

Lemma 2.1. *Let H be any graph. Suppose* $(A, B) \in \mathbb{P}(H)$ *with* $a(A, B) = a(H)$. *Then,* H_1 *and* H_2 *are not induced subpartitions in* (A, B).

Proof. Suppose for the sake of contradiction that H_1 is an induced subpartition of (A, B). Let $x, y, z \in A$ be the vertices of H_1 with $d(y) = 2$. Also, let

$$\xi = \begin{cases} 1 & \text{if } \Delta_B(A) = d_B(x) \text{ or } \Delta_B(A) = d_B(z), \\ 0 & \text{otherwise.} \end{cases}$$

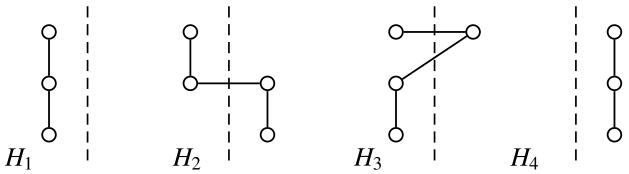


Figure 3. The induced subpartitions H_1, H_2, H_3, and H_4 referenced in Lemmas 2.1, 4.1, and 4.2. The vertical dashed line provides a visual reference to the partition of the vertices into the sets A and B for $(A, B) \in \mathbb{P}(H)$. The vertices to the left of the line in each graph are in A and the vertices on the right are in B.

Consider the partition (A', B') identical to (A, B) except that the vertex y is moved from A to B. Then,

$$\begin{aligned} a(A', B') &= e(A, B) + 2 + |B| + 1 - \Delta_B(A) - \xi \\ &= a(A, B) + 3 - \xi \\ &> a(A, B). \end{aligned}$$

This contradicts our choice of the optimal partition (A, B).

For the sake of contradiction, suppose that H_2 is an induced subpartition of (A, B). Let x, y, z, w be the vertices of H_2 so that $x, y \in A$ and $z, w \in B$, and the edge yz crosses the partition. As H_2 is an induced path of length three, note that $d_G(y) = d_G(z) = 2$. Also, let

$$\xi = \begin{cases} 1 & \text{if } \Delta_B(A) = d_B(x) \text{ or } \Delta_B(A) = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Consider the partition (A', B') identical to (A, B) except that the vertex y is moved from A to B and the vertex z is moved from B to A. Then,

$$\begin{aligned} a(A', B') &= e(A, B) + 2 + |B| - \Delta_B(A) - \xi \\ &= a(A, B) + 2 - \xi \\ &> a(A, B). \end{aligned}$$

Once again, this contradicts our choice of the optimal partition (A, B). □

The next lemma is useful for dealing with vertices of degree one in H.

Lemma 2.2. *For a graph H, there exists a partition $(A, B) \in \mathbb{P}(H)$ with $a(A, B) = a(H)$ and the edges incident with vertices of degree one cross the partition.*

Proof. Consider all $(A, B) \in \mathbb{P}(H)$ with $a(A, B) = a(H)$. Among these, choose the partition which has the maximum number of edges incident with degree one vertices which cross the partition. For the sake of contradiction, suppose there is at least one edge incident to a degree one vertex that does not cross the partition. Let x be this degree one vertex and let y be the neighbor of x. Now, let

$$\xi = \begin{cases} 1 & \text{if } d_B(y) \geq \Delta_B(A), \\ 0 & \text{otherwise.} \end{cases}$$

Suppose first that $x \in A$ and the edge xy does not cross the partition. Consider the partition $(A', B') \in \mathbb{P}(H)$, which is identical to (A, B) except that x is moved from A to B. Then, $a(A', B') = a(A, B) + 2 - \xi > a(A, B)$, which contradicts our choice of the optimal partition (A, B).

Suppose now that $x \in B$ and the edge xy does not cross the partition. Consider the partition $(A', B') \in \mathbb{P}(H)$, which is identical to (A, B) except that x is moved from B to A. Then, $a(A', B') = a(A, B)$, which contradicts our choice of the optimal partition $a(A, B)$ which maximizes the number of edges incident with degree one vertices that cross the partition. □

Lemma 2.2 can be used to provide an alternate proof of the σ_2 condition for a graph to be k-linked. A graph G is *k-linked* if, for every list of $2k$ vertices $\{s_1, \dots, s_k, t_1, \dots, t_k\}$, there exist internally disjoint paths $P_1, \dots, P_k$ such that each P_i is a path joining s_i and t_i. If H is the union of k independent edges (i.e., k copies of the complete graph K_2), then a graph being H-linked is equivalent to the graph being k-linked. As each vertex in H has degree one, Lemma 2.2 states that there exists an optimal partition of H where all of the edges cross the partition. Thus, $a(H) = 2k - 1$ and Theorem 1.2 gives the σ_2 condition proved previously (and independently) in [Kawarabayashi et al. 2006] and [Gould and Whalen 2006]. Note that the bound on the number of vertices in G given by Theorem 1.2 is higher than the bounds in those references.

The next result follows directly from the first case in the proof of Lemma 2.2. The result differs from Lemma 2.2 in that it applies to every optimal partition of H, whereas Lemma 2.2 applies to only a subset of optimal partitions of H.

Corollary 2.3. *If $(A, B) \in \mathbb{P}(H)$ with $a(A, B) = a(H)$, then the vertices of degree one in A must be incident to edges that cross the partition.*

3. Stars

In this section, we determine the value of $a(H)$ when H is a star or a union of stars. Let $K_{1,k}$ denote a star with one vertex of degree k and k vertices of degree one. By Lemma 2.2, an optimal partition of H exists where all degree one vertices cross the partition. Thus, we have the following:

Corollary 3.1. *If $H = K_{1,k}$ for $k \geq 1$, then $a(H) = k$.*

When $H = K_{1,k}$, G being H-linked is equivalent to G being k-connected. This follows from a theorem by Dirac [1960]. With this fact, Theorem 1.2, and Corollary 3.1, we get the well-known σ_2 condition for a graph G to be k-connected (i.e., $\sigma_2(G) \geq n + k - 2$).

We now determine the value of $a(H)$ when H is a union of stars. For $H = K_{1,k_1} \cup K_{1,k_2} \cup \dots \cup K_{1,k_m}$, we call the vertex of maximum degree in each star the *hub vertex* or *hub* of that star. Note that, for $K_{1,1}$, either vertex can be considered a hub vertex.

Theorem 3.2. *If $H = K_{1,k_1} \cup K_{1,k_2} \cup \dots \cup K_{1,k_m}$ with $k_i \geq 1$ for $1 \leq i \leq m$, then*

$$a(H) = 2\sum_{j=1}^{m} k_j - \max\{k_1, k_2, \dots, k_m\}.$$

Proof. Assume without loss of generality that $k_m \geq k_i$ for all $1 \leq i \leq m-1$. By Lemma 2.2, there exists a partition $(A, B) \in \mathbb{P}(H)$ with $a(A, B) = a(H)$ where all edges incident with vertices of degree one cross the partition. Among all optimal partitions that satisfy that property, choose the partition with the maximum number of hub vertices in A. We will now show that, under the assumptions above, all of the hub vertices are in A.

Claim 3.3. The hub of the star K_{1,k_m} must be in A.

Proof. Let x be the hub of K_{1,k_m} and suppose that $x \in B$. Note that $d_G(x) = k_m$. Consider the partition (A', B') obtained by moving x from B to A and moving its leaves from A to B. Then, noting that $\Delta_B(A) \geq 1$,

$$\begin{aligned} a(A', B') &= a(A, B) + k_m - 1 - (k_m - \Delta_B(A)) \\ &= a(A, B) + \Delta_B(A) - 1 \\ &\geq a(A, B). \end{aligned}$$

However, this contradicts our assumption that (A, B) is an optimal partition of H, which has the maximum number of hubs in A. So, the hub of maximum degree must be in A. □

Assume without loss of generality that the hubs of $K_{1,k_1}, K_{1,k_2}, \ldots, K_{1,k_i}$ are in B (where $i \geq 0$) and the remaining hubs are in A. By the above claim, $i < m$. Now, we have

$$\begin{aligned} a(A, B) &= \sum_{j=1}^{m} k_j + \Bigg(\sum_{t=i+1}^{m} k_t \Bigg) + i - k_m = \sum_{j=1}^{m} k_j + \Bigg(\sum_{t=i+1}^{m-1} k_t \Bigg) + i \\ &\leq \sum_{j=1}^{m} k_j + \sum_{t=1}^{m-1} k_t = 2\Bigg(\sum_{j=1}^{m} k_i \Bigg) - k_m. \end{aligned}$$

So, this gives us an upper bound on $a(A, B)$ for all possible locations of the hubs. For the lower bound, note that the partition $(A', B') \in \mathbb{P}(H)$ where all of the hubs of H are in A' has

$$a(A', B') = 2\Bigg(\sum_{j=1}^{m} k_j \Bigg) - k_m.$$

Therefore, $a(H) = 2\big(\sum_{j=1}^{m} k_j\big) - k_m$, where $k_m = \max\{k_1, k_2, \ldots, k_m\}$. □

Note that Theorem 3.2 can also be used to show that $a(H) = 2k - 1$ when H is the union of k independent edges (which was discussed in the previous section).

4. Cycles and paths

We now move our attention to paths and cycles. Let C_k (for $k \geq 3$) denote a cycle on k vertices and P_k (for $k \geq 2$) denote a path on k vertices.

The following lemmas prove that H_3 (shown in Figure 3) cannot appear as an induced subpartition in any optimal partition of H.

Lemma 4.1. *Let $k \geq 4$. For $H \in \{C_k, P_k\}$, the graph H_3 cannot be an induced subpartition of any partition $(A, B) \in \mathbb{P}(H)$ with $a(A, B) = a(H)$.*

Proof. Suppose for the sake of contradiction that H_3 is an induced subpartition of some partition (A, B) with $a(A, B) = a(H)$. Assume the vertices of H_3 are x, y, z, and w with $x, z, w \in A$ and $y \in B$, and the edges are xy, yz, and zw.

By Corollary 2.3, $d_G(w) \neq 1$ since the edge incident to w does not cross the partition. Let t be a neighbor of w in H. By Lemma 2.1, $t \in B$.

Now, either $d_G(x) = 1$, x has a neighbor in A, or x has a second neighbor in B. If $d_G(x) = 1$ or if x has a neighbor in A, then the partition (A', B') formed from (A, B) by moving x and z to from A to B and y from B to A has $a(A', B') > a(A, B)$. As this contradicts our choice of the optimal partition (A, B), x must have a second neighbor in B.

Let v be the other neighbor of x in B. As a result, $\Delta_B(A) = 2$. Consider the partition (A', B'), which modifies the partition (A, B) by moving w from A to B. Then,

$$\begin{aligned} a(A', B') &= e(A, B) + |B| + 1 - \Delta_B(A) - 0 \\ &= a(A, B) + 1. \end{aligned}$$

Thus, the partition (A', B') has $a(A', B') > a(A, B)$. However, this contradicts the assumption that the partition (A, B) has $a(A, B) = a(H)$.

As all possibilities are exhausted and lead to contradictions, we conclude that H_3 is not an induced subpartition of any partition (A, B) with $a(A, B) = a(H)$. □

This final lemma proves that there exists an optimal partition of H which does not contain H_4 (shown in Figure 3) as an induced subpartition.

Lemma 4.2. *If $H \in \{C_k, P_k\}$ with $k \geq 3$, then there is a partition $(A, B) \in \mathbb{P}(H)$ with $a(A, B) = a(H)$ which does not have H_4 as a subpartition.*

Proof. For the sake of contradiction, assume all partitions (A, B) with $a(A, B) = a(H)$ have H_4 as a subpartition. Consider one such partition (A, B) which contains H_4. Let the vertices of H_4 (all of which are in B) be x, y, and z with the two edges being xy and yz. Consider the partition (A', B') which is identical to (A, B)

except the vertex y is moved from B to A. Then,

$$\begin{aligned} a(A', B') &\geq e(A, B) + 2 + |B| - 1 - \Delta_B(A) - 1 \\ &= a(A, B) + 1 - 1 \\ &= a(A, B). \end{aligned}$$

Note that equality occurs in the first line above only when $\Delta_B(A) = 1$ as the partition (A', B') has $\Delta_{B'}(A') = 2$. Otherwise, $a(A', B') > a(A, B)$. In either case, as $a(A', B') \geq a(A, B)$ and (A', B') does not contain H_4 as a subpartition, we have a contradiction. □

With these lemmas, we are now able to prove the value for $a(H)$ when H is a cycle or path with three or more vertices. Note that by Lemma 2.2, for the single edge P_2, we have $a(P_2) = 1$.

Theorem 4.3. *For $k \geq 3$, we have $a(C_k) = \left\lceil \frac{1}{2}(3k - 5) \right\rceil$ and $a(P_k) = \left\lceil \frac{1}{2}(3k - 6) \right\rceil$.*

Proof. Let $H \in \{P_k, C_k\}$ and assume that $V(H) = \{1, 2, 3, \dots, k\}$ with the vertices numbered based on an arbitrary orientation of H. If $k = 3$, then it is straightforward to show that $a(C_3) = a(P_3) = 2 = \left\lceil \frac{1}{2}(3(3) - 5) \right\rceil = \left\lceil \frac{1}{2}(3(3) - 6) \right\rceil$. If $k = 4$, then it is also straightforward to show that $a(C_4) = 4 = \left\lceil \frac{1}{2}(3(4) - 5) \right\rceil$ and $a(P_4) = 3 = \left\lceil \frac{1}{2}(3(4) - 6) \right\rceil$. So, assume $k \geq 5$. Consider a partition $(A, B) \in \mathbb{P}(H)$ with $a(A, B) = a(H)$. By Lemma 4.2, we may assume H_4 is not an induced subpartition of (A, B). It follows from Lemma 2.1, Lemma 4.1, Corollary 2.3, and the fact that $k \geq 5$ that the partition (A, B) cannot have any edge with both endpoints in A. Consequently, $\Delta_B(A) = 2$.

Assume for the sake of contradiction that the partition (A, B) has at least two edges with both endpoints in B. Among the edges with both endpoints in B, choose the two edges with the fewest edges of H between them based on the orientation of H. Let $(i, i+1)$ and $(j, j+1)$ with $j > i$ be two edges with both endpoints in B. Note that $i + 1 \neq j$ since H_4 is not an induced subpartition. In particular, vertex $i + 2$ must be in A and by Lemma 2.1, $i + 3$ must be in B as otherwise H_2 would be an induced subpartition. Lemma 2.1, Lemma 4.2, and our choice of j imply that $j = i + t$ for some positive odd integer t and the vertices $i + 1, i + 3, \dots, i + t$ are in B while the vertices $i + 2, i + 4, \dots, i + t - 1$ are in A.

Consider the partition (A', B') formed by starting with (A, B) and moving vertices $i + 1, i + 3, \dots i + t$ from B to A and moving $i + 2, i + 4, \dots, i + t - 1$ from A to B. Then,

$$\begin{aligned} a(A', B') &= e(A, B) + 2 + |B| - 1 - \Delta_A(B) \\ &= a(A, B) + 1 \\ &> a(A, B). \end{aligned}$$

However, this contradicts our choice of the partition (A, B). Thus, as no edge of the partition can have both endpoints in A, all of the edges of (A, B) must cross the partition with the possible exception of exactly one edge which must have both endpoints in B.

If $H = C_k$ with k even, then (A, B) can have no edge with both endpoints in B as one edge in B would force the existence of another edge with either both endpoints in B or both endpoints in A. Thus, (A, B) must either be the partition with $B = \{1, 3, \dots k-1\}$ and $A = \{2, 4, \dots, k\}$ or the same partition with the vertices in A and B swapped. Consequently, $a(A, B) = a(C_k) = k + \frac{1}{2}k - 2 = \frac{1}{2}(3k-4)$.

If $H = C_k$ with k odd, then (A, B) must have exactly one edge with both endpoints in B as all edges cannot cross the partition. Thus, (A, B) must be the partition with $A = \{1, 3, \dots, k-2\}$ and $B = \{2, 4, \dots, k-1, k\}$ or a vertex relabeling of this partition. Consequently, $a(A, B) = a(C_k) = k + \lceil \frac{1}{2}k \rceil - 2 = \lceil \frac{1}{2}(3k-5) \rceil$.

If $H = P_k$ with k odd, then (A, B) must have all edges crossing the partition. Thus, (A, B) must be the partition with $A = \{2, 4, \dots, k-1\}$ and $B = \{1, 3, \dots, k\}$. If $H = P_k$ with k even, then (A, B) either has all edges crossing the partition or exactly one edge with both endpoints in B (and all other edges crossing the partition). Thus, in either case, $a(P_k) = k - 1 + \lceil \frac{1}{2}k \rceil - 2 = \lceil \frac{1}{2}(3k-6) \rceil$. □

When $H = C_k$, a graph G being H-linked is equivalent to G being k-ordered. A graph G is *k-ordered* if for every ordered set of vertices S such that $|S| = k$, the graph G contains a cycle C encountering the vertices S in the given order. When $H = P_k$, a graph G being H-linked is equivalent to G being k-ordered connected. A graph G is *k-ordered connected* if for every ordered set of vertices $S = \{v_1, v_2, \dots, v_k\}$, the graph G contains a path P from v_1 to v_k encountering S in the given order. Note that by forcing the cycle (or path) that encounters the vertices of S in order to be a hamiltonian cycle (or a hamiltonian path), we get the property k-ordered hamiltonian (k-ordered hamiltonian connected). The concept of k-ordered was introduced by Ng and Schutz [1997].

Using Theorem 1.2 and Theorem 4.3, we have the following corollary.

Corollary 4.4. *Let G be a graph on n vertices and let $k \geq 3$.*

(i) *If $n > 20k$ and $\sigma_2(G) \geq n + \lceil \frac{1}{2}(3k-9) \rceil$, then G is k-ordered.*

(ii) *If $n > 20(k-1)$ and $\sigma_2(G) \geq n + \lceil \frac{1}{2}(3k-10) \rceil$, then G is k-ordered connected.*

The bounds on σ_2 in both cases are best possible.

To the best of our knowledge, the above σ_2 conditions for k-ordered and k-ordered connected are not explicitly stated in the literature, although it is implied in several sources that the σ_2 conditions for k-ordered and k-ordered connected should be the same as the σ_2 conditions for k-ordered hamiltonian and k-ordered hamiltonian connected.

There are a number of degree-sum results for k-ordered hamiltonian and k-ordered hamiltonian connected graphs. Ng and Schultz [1997] proved a sharp σ_2 condition for any graph on $n \geq 3$ vertices to be k-ordered hamiltonian. J. Faudree et al. [2000] proved the bound for σ_2 could be reduced for graphs on n vertices with $n \geq 53k^2$. The same σ_2 condition in [Faudree et al. 2000] was shown to work for $n \geq 2k$ by R. Faudree et al. [2003b]. Note that the sharpness example they construct in [Faudree et al. 2003b] is neither k-ordered hamiltonian nor k-ordered.

Theorem 4.5 [Faudree et al. 2003b]. *Let k be an integer with $3 \leq k \leq \frac{1}{2}n$, and let G be a graph of order n. If $\sigma_2(G) \geq n + \frac{1}{2}(3k-9)$, then G is k-ordered hamiltonian. The bound on $\sigma_2(G)$ is sharp.*

For k-ordered hamiltonian connected, a σ_2 condition for large n is mentioned (without proof) in [Faudree et al. 2003a]. A stronger and sharp σ_2 condition for k-ordered hamiltonian connected was proven by Nicholson and Wei [2015].

Theorem 4.6 [Nicholson and Wei 2015]. *If G is a graph on n vertices with $\sigma_2(G) \geq n + \frac{1}{2}(3k-10)$, where $4 \leq k \leq \frac{1}{2}(n+1)$, then G is k-ordered hamiltonian connected.*

Overall, while the concepts of k-ordered and k-ordered hamiltonian are distinct, the σ_2 condition is the same for both when n is large as shown in Corollary 4.4. Similarly, the σ_2 conditions for k-ordered connected and k-ordered hamiltonian connected are the same for large n. Corollary 4.4 has higher bounds on n than the optimal known results, but by utilizing Theorem 1.2, the proofs are much less technical.

5. Complete graphs and complete bipartite graphs

Our last results provide $a(H)$ when H is the complete graph K_k or the complete bipartite graph $K_{r,s}$. First, we consider the complete graph K_k.

Theorem 5.1. *For any integer $k \geq 3$, we have $a(K_k) = \lfloor \frac{1}{4}k^2 \rfloor$.*

Proof. Suppose $|A| = t$ and $|B| = k-t$. Then, $a(A, B) = (k-t)(t) + (k-t) - (k-t) = kt - t^2$. Let $f(t) = kt - t^2$. Then, $f'(t) = -2t + k$. So, $f'(t) = 0$ implies that $t = \frac{1}{2}k$. Since $f''(t) < 0$, $f(t)$ has a global maximum at $t = \frac{1}{2}k$. If k is even, then $a(K_k) = \frac{1}{4}k^2$. If k is odd, then either

$$a(K_k) = \left[k - \left(\tfrac{1}{2}(k-1)\right)\right]\left(\tfrac{1}{2}(k-1)\right) \quad \text{or} \quad a(K_k) = \left[k - \tfrac{1}{2}(k+1)\right]\left(\tfrac{1}{2}(k+1)\right).$$

In both cases, $a(K_k) = \frac{1}{4}(k^2-1)$. Therefore, $a(K_k) = \lfloor \frac{1}{4}k^2 \rfloor$ for any integer $k \geq 3$. □

Now, we prove the value of $a(H)$ when is the complete bipartite graph $K_{r,s}$. Note that $K_{1,1}$ and $K_{1,2}$ are covered by previous results in this article.

Theorem 5.2. *For $r \geq s \geq 2$, we have $a(K_{r,s}) = rs$.*

Proof. Let $(A, B) \in \mathbb{P}(K_{r,s})$ such that $a(A, B) = a(K_{r,s})$. Using the canonical bipartition of $K_{r,s}$, let X and Y be the partite sets. Let $X_A = X \cap A$, $Y_B = Y \cap B$, $Y_A = Y \cap A$, and $X_B = X \cap B$. Additionally, let $|X_A| = x_A$, $|X_B| = x_B$, $|Y_A| = y_A$, and $|Y_B| = y_B$.

Suppose that exactly one of the sets X_A, X_B, Y_A, and Y_B is empty. Assume without loss of generality that the partite sets of $K_{r,s}$ are labeled X and Y so that either X_A or X_B is empty. Assume first that only $X_A = \varnothing$. Let $v \in Y_B$. Consider the partition (A', B') starting with (A, B) and moving v from B to A. That is, we have $X_{B'} = X_B = X$, $Y_{A'} = Y_A \cup \{v\}$, and $Y_{B'} = Y_B - \{v\}$. Then,

$$a(A', B') = a(A, B) + |X_B| - 1 = a(A, B) + |X| - 1.$$

As $|X| \geq 2$, we have $a(A', B') > a(A, B)$, which contradicts our choice of (A, B).

Assume now that X_B is the only empty set among X_A, X_B, Y_A, and Y_B. Let $w \in Y_A$. Consider the partition (A', B') starting with (A, B) and moving w from B to A as in the first case. Then,

$$a(A', B') = a(A, B) + |X_A| + 1 - 1 = a(A, B) + |X|.$$

Since $|X| \geq 2$, we have $a(A', B') > a(A, B)$, which contradicts our choice of (A, B).

Assume now that each of X_A, X_B, Y_A, and Y_B is nonempty. Then,

$$a(A, B) = (x_A)(y_B) + (y_A)(x_B) + x_B + y_B - \max\{x_B, y_B\}.$$

Note that $x_B + y_B - \max\{x_B, y_B\} = \min\{x_B, y_B\}$.

Consider the partition (A', B') formed by starting with (A, B) and moving the vertices of X_B from B to A and moving the vertices of Y_A from A to B. Then,

$$a(A', B') = (x_A)(y_B) + (x_B)(y_A) + (x_A)(y_A) + (x_B)(y_B).$$

Note that $a(A', B') \geq a(A, B)$ whenever $(x_A)(y_A) + (x_B)(y_B) > \min\{x_B, y_B\}$. However, since x_B, x_A, y_A, and y_B are all at least one, $(x_A)(y_A) + (x_B)(y_B)$ is strictly larger than $\min\{x_B, y_B\}$. Thus, $a(A', B') \geq a(A, B)$, which contradicts our choice of (A, B).

Consequently, the only remaining possibility for (A, B) is either $A = X$ and $B = Y$, or $A = Y$ and $B = X$. In either case, $a(A, B) = rs$ and thus, $a(K_{r,s}) = rs$. □

These results together with Theorem 1.2, we get the following corollary.

Corollary 5.3. *Let G be a graph on n vertices.*

(i) *Let $k \geq 3$. If $n > 20k$ and $\sigma_2(G) \geq n + \lfloor \frac{1}{4}k^2 \rfloor - 2$, then G is K_k-linked.*

(ii) *Let $r \geq s \geq 2$. If $n \geq 20rs$ and $\sigma_2(G) \geq n + rs - 2$, then G is $K_{r,s}$-linked.*

6. Final observations

For all of the classes of graphs examined above, an optimal partition (A, B) always exists where the vertex of maximum degree is in A. However, this is not always the case. Consider the graph J with $V(J) = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7\}$ and

$$E(J) = \{v_1v_2, v_1v_3, v_1v_5, v_2v_3, v_2v_4, v_2v_6, v_2v_7, v_3v_4, v_4v_6, v_4v_7, v_5v_7, v_6v_7\}.$$

Note that $\Delta(J) = 5$ and v_2 is the vertex of maximum degree. By checking all possible partitions of the vertex set (possibly with the aid of a computer), it can be shown that J has a unique optimal partition (A, B) given by $A = \{v_1, v_4, v_7\}$. From this partition, we have $a(J) = 10$. However, in this optimal partition, the vertex of maximum degree (i.e., v_2) is in B and $\Delta_B(A) = 3$. So, it is not always the case that a graph has an optimal partition (A, B) where the vertex of maximum degree is in A.

We conclude by making an observation about optimal partitions of the union of graphs. Consider two graphs M_1 and M_2 and assume an optimal partition of both graphs is known. We note that the union of these two optimal partitions is not necessarily an optimal partition for the union of M_1 and M_2. As an example, let M_1 be the five cycle $v_1v_2v_3v_4v_5$ with the additional edges v_2v_4 and v_3v_5. Let M_2 be the graph on the set $\{w_1, w_2, w_3, w_4, w_5\}$ where w_1, w_2, w_3, and w_4 form a K_4 and the only other edge is w_4w_5.

Now, the graph M_1 has exactly two optimal partitions which both give $a(M_1) = 6$. One of the optimal partitions of M_1, which we denote by (A_{M_1}, B_{M_1}), has $A_{M_1} = \{v_2, v_3\}$. The graph M_2 has nine different optimal partitions which give $a(M_2) = 5$. One of these optimal partitions of M_2, which we denote by (A_{M_2}, B_{M_2}), has $A_{M_2} = \{w_1, w_3\}$. Let M be the graph formed by the union of M_1 and M_2 and consider the partition

$$(A_M, B_M) = (A_{M_1} \cup A_{M_2}, B_{M_1} \cup B_{M_2}).$$

This partition gives $a(A_M, B_M) = 13$. However, $a(M) = 14$, which can be achieved using the optimal partition of M_1 given above and a different optimal partition of M_2 such as the partition (A'_{M_2}, B'_{M_2}), where $A'_{M_2} = \{w_1, w_4\}$. So, finding an optimal partition for a union of graphs is not simply a matter of taking any optimal partition of the graphs individually and forming the union of these partitions.

7. Acknowledgments

This research was partially funded by Samford University's Arts and Sciences Program for Independent Research (ASPIRE). The authors would like to thank an anonymous referee for giving several suggestions which were helpful in preparing the final manuscript.

References

[Dirac 1960] G. A. Dirac, "In abstrakten Graphen vorhandene vollständige 4-Graphen und ihre Unterteilungen", *Math. Nachr.* **22** (1960), 61–85. MR Zbl

[Faudree et al. 2000] J. R. Faudree, R. J. Faudree, R. J. Gould, M. S. Jacobson, and L. Lesniak, "On k-ordered graphs", *J. Graph Theory* **35**:2 (2000), 69–82. MR Zbl

[Faudree et al. 2003a] J. R. Faudree, R. J. Gould, F. Pfender, and A. Wolf, "On k-ordered bipartite graphs", *Electron. J. Combin.* **10** (2003), RP11, 12 pp. MR Zbl

[Faudree et al. 2003b] R. J. Faudree, R. J. Gould, A. V. Kostochka, L. Lesniak, I. Schiermeyer, and A. Saito, "Degree conditions for k-ordered Hamiltonian graphs", *J. Graph Theory* **42**:3 (2003), 199–210. MR Zbl

[Ferrara et al. 2006] M. Ferrara, R. Gould, G. Tansey, and T. Whalen, "On H-linked graphs", *Graphs Combin.* **22**:2 (2006), 217–224. MR Zbl

[Ferrara et al. 2012] M. Ferrara, R. Gould, M. Jacobson, F. Pfender, J. Powell, and T. Whalen, "New Ore-type conditions for H-linked graphs", *J. Graph Theory* **71**:1 (2012), 69–77. MR Zbl

[Gould and Whalen 2006] R. J. Gould and T. C. Whalen, "Distance between two k-sets and path-systems extendibility", *Ars Combin.* **79** (2006), 211–228. MR Zbl

[Gould et al. 2006] R. J. Gould, A. Kostochka, and G. Yu, "On minimum degree implying that a graph is H-linked", *SIAM J. Discrete Math.* **20**:4 (2006), 829–840. MR Zbl

[Jung 1970] H. A. Jung, "Eine Verallgemeinerung des n-fachen Zusammenhangs für Graphen", *Math. Ann.* **187** (1970), 95–103. MR Zbl

[Kawarabayashi et al. 2006] K.-i. Kawarabayashi, A. Kostochka, and G. Yu, "On sufficient degree conditions for a graph to be k-linked", *Combin. Probab. Comput.* **15**:5 (2006), 685–694. MR Zbl

[Kostochka and Yu 2005] A. Kostochka and G. Yu, "An extremal problem for H-linked graphs", *J. Graph Theory* **50**:4 (2005), 321–339. MR Zbl

[Kostochka and Yu 2008] A. V. Kostochka and G. Yu, "Ore-type degree conditions for a graph to be H-linked", *J. Graph Theory* **58**:1 (2008), 14–26. MR Zbl

[Ng and Schultz 1997] L. Ng and M. Schultz, "k-ordered Hamiltonian graphs", *J. Graph Theory* **24**:1 (1997), 45–57. MR Zbl

[Nicholson and Wei 2015] E. W. Nicholson and B. Wei, "Degree sum condition for k-ordered Hamiltonian connected graphs", *Graphs Combin.* **31**:3 (2015), 743–755. MR Zbl

[West 1996] D. B. West, *Introduction to graph theory*, Prentice Hall, Upper Saddle River, NJ, 1996. MR Zbl

Received: 2016-05-05 Revised: 2016-07-01 Accepted: 2016-07-11

least@samford.edu
Smiths Station High School, 4228 Lee Rd. 430,
Smiths Station, AL 36877, United States

jspowel1@samford.edu
Department of Mathematics and Computer Science,
Samford University, 800 Lakeshore Drive,
Birmingham, AL 35229, United States

Guidelines for Authors

Submissions in all mathematical areas are encouraged. All manuscripts accepted for publication in *Involve* are considered publishable in quality journals in their respective fields, and include a minimum of one-third student authorship. Submissions should include substantial faculty input; faculty co-authorship is strongly encouraged. Authors may submit manuscripts in PDF format on-line at the Submission page at the Involve website.

Originality. Submission of a manuscript acknowledges that the manuscript is original and 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, 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

2017 vol. 10 no. 4

1944-4176(2017)10:4;1-9