involve

a journal of mathematics

msp

2020 vol. 13, no. 3

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 2020 is US $205/year for the electronic version, and $275/year (+$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP.

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

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

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

INVOLVE 13:3 (2020)
https://doi.org/10.2140/involve.2020.13.361

Hyperbolic triangular prisms with one ideal vertex

Grant S. Lakeland and Corinne G. Roth

(Communicated by Kenneth S. Berenhaut)

We classify all of the five-sided three-dimensional hyperbolic polyhedra with one ideal vertex, which have the shape of a triangular prism, and which give rise to a discrete reflection group. We show how to find each such polyhedron in the upper half-space model by considering lines and circles in the plane. Finally, we give matrix generators in $\mathrm{PSL}_2(\mathbb{C})$ for the orientation-preserving subgroup of each corresponding reflection group.

1. Introduction

A convex polyhedron in hyperbolic 3-space $\mathbb{H}^3$ generates a discrete group of isometries if the dihedral angles at which its bounding planes meet are all integer submultiples of π — that is, each angle is of the form $\frac{\pi}{n}$ radians for an integer $n \geq 2$ — and if the dihedral angles satisfy some other combinatorial criteria. The set of all such polyhedra is infinite, with some partial classifications completed. For example, the fewest sides such a polyhedron may have is four, and the 32 hyperbolic tetrahedra were found in [Lannér 1950; Vinberg 1967; Thurston 1979]. The 825 smallest-volume all-right-angled polyhedra were found in [Inoue 2015].

For certain computations, it is helpful to know matrix generators in $\mathrm{PSL}_2(\mathbb{C})$ for the orientation-preserving index-2 subgroups of these reflection groups. Such generators were found for some of the tetrahedral groups by Brunner, Lee, and Wielenberg [Brunner et al. 1985], and the second author has shown how to find these for all 32 tetrahedra [Lakeland 2010]. These matrix generators have recently been used, for example, in [Hoffman 2014] to study knot complements which cover tetrahedral orbifolds, and in [Şengün 2012] to study growth in torsion homology of subgroups of the tetrahedral groups.

In this paper, we perform the same calculations for one class of five-sided polyhedra, which when drawn schematically resemble triangular prisms. The set of all such polyhedra is infinite, and for simplicity we restrict our attention to such

MSC2010: primary 57R18; secondary 20F55.
Keywords: hyperbolic reflection group, prism, ideal polyhedron.

polyhedra which have one ideal vertex and five finite vertices. As such, all of our examples are noncompact. We find:

Theorem 1.1. *Up to relabeling by symmetry of the prism, there are* 12 *infinite families of prisms with exactly one ideal vertex, each of which has one dihedral angle which may be any integer submultiple of* π *less than or equal to* $\frac{\pi}{6}$. *In addition, there are* 78 *specific arrangements, where in each case all dihedral angles not incident to the ideal vertex are at least* $\frac{\pi}{5}$.

We note that the list of polyhedra considered in this paper overlaps with those considered by Kaplinskaja [1974], who studied finite-volume simplicial prisms in $\mathbb{H}^3$, $\mathbb{H}^4$ and $\mathbb{H}^5$ which give rise to discrete reflection groups, and listed their Coxeter graphs. Each of our prisms either appears there, or can be decomposed into two polyhedra which do. As such, our classification is not new, although we do not believe that our polyhedra have previously been listed in this way. The Coxeter graphs of these polyhedra do not immediately allow one to produce isometries which generate the orientation-preserving subgroup, and in this paper we provide a method for this.

The method used is the following. First, we use Andreev's theorem to set up the combinatorial rules which the dihedral angles must satisfy, and we find all admissible arrangements of angles which satisfy these rules. Then, we reduce the problem of finding hyperbolic planes in the upper half-space $\mathbb{H}^3$ which meet at these prescribed angles to a similar problem, involving finding lines and circles in the plane which meet at the same angles. Finally, we use this geometric data to write down matrix generators for each group.

This paper is organized as follows. After some geometric preliminaries in Section 2, in Section 3 we describe all of the possible arrangements of angles which are possible for our prisms, grouped by the possible angles at the ideal vertex. We also outline a method to locate each prism precisely in $\mathbb{H}^3$. In Section 4 we describe a general method to find each prism and write down corresponding matrices which generate the orientation-preserving subgroup of isometries of $\mathbb{H}^3$, and in Section 5 we summarize the possible angle arrangements in tables.

2. Geometric preliminaries

We recall some definitions and results about hyperbolic polyhedra. We will work in the upper half-space model for $\mathbb{H}^3$, $\{(x, y, z) \in \mathbb{R}^3 \mid z > 0\}$, and we recall that in this model, geodesic lines are vertical lines and semicircles which meet the plane $\{z = 0\}$ perpendicularly, and geodesic planes are vertical planes and hemispheres whose equators lie in the plane $\{z = 0\}$.

We first note that in order for a polyhedron to generate a discrete reflection group, all of its dihedral angles must be integer submultiples of π radians, and the integer must be no less than 2. In this paper, we will label a dihedral angle of $\frac{\pi}{n}$ by the natural number n.

Definition. A triangle is *Euclidean* if its angles p, q, and r satisfy

$$\frac{\pi}{p}+\frac{\pi}{q}+\frac{\pi}{r}=\pi \quad \text{or} \quad \frac{1}{p}+\frac{1}{q}+\frac{1}{r}=1.$$

Definition. A triangle is *spherical* if its angles p, q, and r satisfy

$$\frac{1}{p}+\frac{1}{q}+\frac{1}{r}>1.$$

Definition. A triangle is *hyperbolic* if its angles p, q, and r satisfy

$$\frac{1}{p}+\frac{1}{q}+\frac{1}{r}<1.$$

With these definitions in mind, we will appeal to Andreev's theorem [1970] for hyperbolic polyhedra, which specifies exactly what conditions a combinatorial arrangement of dihedral angles must satisfy in order for it to give rise to a hyperbolic polyhedron. For the precise statement given below, we refer to work of Dunbar, Hubbard, and Roeder [Roeder et al. 2007].

Theorem 2.1 (Andreev). *If P is a compact, finite-sided hyperbolic polyhedron with dihedral angle α_i at each edge e_i, then the following conditions hold*:

(1) *For each i, we have $\alpha_i > 0$.*

(2) *If three edges e_i, e_j, e_k meet at a vertex, then $\alpha_i+\alpha_j+\alpha_k>\pi$.*

(3) *If there exists a prismatic 3-circuit intersecting e_i, e_j and e_k, then $\alpha_i+\alpha_j+\alpha_k<\pi$.*

(4) *If there exists a prismatic 4-circuit intersecting e_i, e_j, e_k and e_l, then $\alpha_i+\alpha_j+\alpha_k+\alpha_l<2\pi$.*

(5) *For a quadrilateral face with edges enumerated successively e_1, e_2, e_3 and e_4, if e_{12}, e_{23}, e_{34}, and e_{41} are such that e_{12} is the third edge meeting at the vertex where e_1 and e_2 intersect* (*and similarly for other e_{ij}*), *then*

 (a) $\alpha_1+\alpha_3+\alpha_{12}+\alpha_{23}+\alpha_{34}+\alpha_{41}<3\pi$, *and*

 (b) $\alpha_2+\alpha_4+\alpha_{12}+\alpha_{23}+\alpha_{34}+\alpha_{41}<3\pi$.

The dihedral angle of intersection of two planes in the upper half-space model of $\mathbb{H}^3$ is the same as the angle between the respective tangent planes at any point of intersection. Since these planes are either vertical Euclidean planes or Euclidean spheres with center on the plane $\{z=0\}$, if two planes intersect, they have a common point on the plane $\{z=0\}$. In this case, the respective tangent planes are both vertical Euclidean planes, and so the dihedral angle is the angle the tangent planes make in the xy-plane.

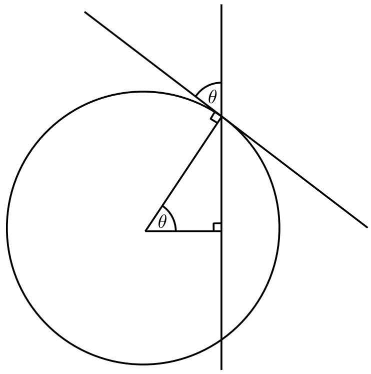


Figure 1. The vertical line is $x = r\cos\theta$.

With this in mind, we observe that the aim of finding five hyperbolic planes which intersect at prescribed angles may be reduced to finding five lines and circles in the xy-plane which intersect at the same prescribed angles. Since three of our planes intersect at an ideal vertex, we may place this ideal vertex at ∞, and thereby assume that the three planes are vertical Euclidean planes. These will correspond to Euclidean lines in the xy-plane. The remaining two sides will then correspond to circles in the xy-plane; since angles of intersection are preserved by Euclidean similarities, we may take one of the circles to be the unit circle in the xy-plane.

With these assumptions about our setup in place, when finding lines and circles which intersect at given angles, we will appeal frequently to the following results.

Lemma 2.2. *A line that intersects a circle of radius r at angle θ comes distance $r\cos(\theta)$ away from the center of the circle at its closest point.*

Proof. After translating and rotating if necessary, we may assume that the line is vertical, and further that it is of the form $x = a$ for $a \geq 0$, and that the circle is centered at the origin. The intersection points are then $(a, \sqrt{r - a^2})$ and $(a, -\sqrt{r - a^2})$; we focus on the former. The angle θ is the angle at which the line $x = a$ and the tangent line to the circle at this point intersect (see Figure 1). Since the radius of the unit circle meets the tangent line at a right angle, the triangle with vertices at $(0, 0)$, $(a, 0)$ and $(a, \sqrt{r - a^2})$ has an angle of $\frac{\pi}{2} - \theta$ at $(a, \sqrt{r - a^2})$. This triangle has a right angle at $(a, 0)$, and so must have angle θ at $(0, 0)$. Since the hypotenuse of the triangle is a radius of the unit circle, it follows that $a = \cos(\theta)$. □

When finding the precise location of the prism in the upper half-space, we will need to find three lines (corresponding to vertical planes which "meet" at the ideal vertex) and two circles (corresponding to nonvertical planes). We will take one circle to be the unit circle and then find the three lines based on that. One of the lines will be taken to be $x = c$ for some constant c; the other two lines will have

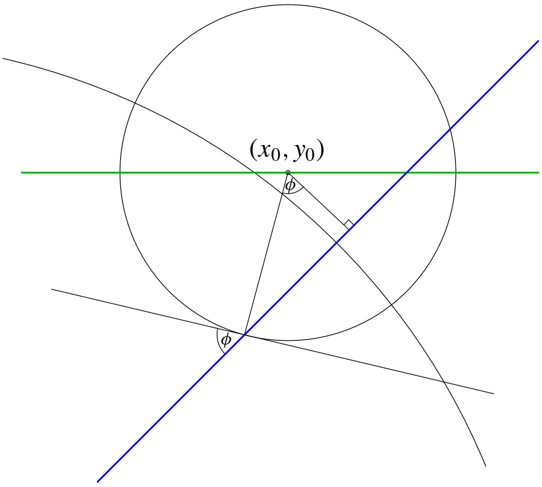


Figure 2. The circle has center (x_0, y_0) and radius r; the other arc is part of the unit circle.

the form $y = mx + b$, with slope m and y-intercept b. In order to find the second circle, we will find three equations, corresponding to the three angles at which that plane meets three other planes, in the three unknowns (x_0, y_0) and r, representing, respectively, the coordinates of the center, and the radius, of the second circle. This circle will meet the unit circle and two lines; in some circumstances, we will need an equation coming from the fact that the second circle meets a given line $y = mx + b$ at a prescribed angle ϕ.

Lemma 2.3. *If a circle with center (x_0, y_0) and radius r meets the line $y = mx + b$ at angle ϕ, and (x_0, y_0) lies on or above the line (that is, $y_0 \geq mx_0 + b$), then x_0, y_0 and r satisfy*

$$y_0 - \frac{r\cos\phi}{\sqrt{m^2+1}} = m\left(x_0 + \frac{mr\cos\phi}{\sqrt{m^2+1}}\right) + b.$$

Proof. The vector $\langle 1, m\rangle$ is parallel to the line, so the vector $\langle m, -1\rangle$ is perpendicular to the line. The unit vector parallel to this is

$$\left\langle \frac{m}{\sqrt{m^2+1}}, \frac{-1}{\sqrt{m^2+1}} \right\rangle.$$

By trigonometry (see Figure 2), we see that the point which is distance $r\cos\phi$ away from (x_0, y_0) in the direction of this vector lies on the line. Therefore, plugging the x- and y-coordinates of the vector

$$\left\langle x_0 + r\cos\phi\frac{m}{\sqrt{m^2+1}}, y_0 - r\cos\phi\frac{1}{\sqrt{m^2+1}} \right\rangle$$

into the formula $y = mx + b$ yields the required equation. □

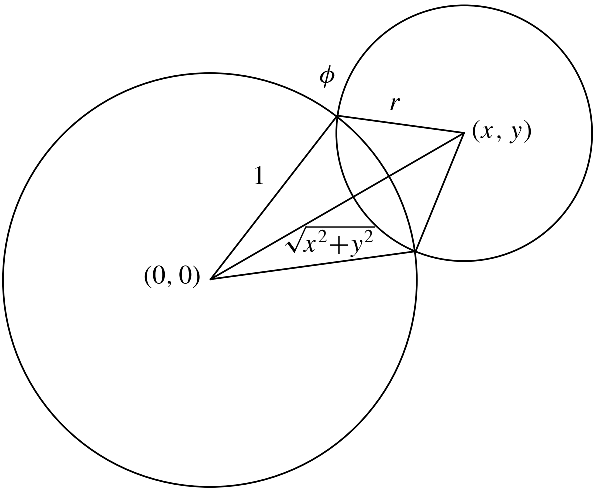


Figure 3. The circles meet at angle ϕ.

We will obtain another of our three equations by using the angle ϕ at which the second circle meets the unit circle.

Lemma 2.4. *If a circle with center (x, y) and radius r meets the unit circle at angle ϕ, then x, y, and r satisfy*

$$x^2 + y^2 = 1 + r^2 + 2(\cos\phi)r.$$

Proof. This is an application of the cosine law to the triangle whose vertices are at $(0, 0)$, (x, y), and one of the points where the circles intersect (see Figure 3). This triangle has side lengths, 1, r and $\sqrt{x^2 + y^2}$. The angle opposite the latter side is $\pi - \phi$ because the angle between the respective tangent lines at this vertex is ϕ, and the two angles between the tangent lines and their respective radii are both $\frac{\pi}{2}$. The equation follows from the fact that $\cos(\pi - \phi) = -\cos\phi$. □

3. Prisms

We now will describe the dihedral angles of all hyperbolic triangular prisms with one ideal vertex, and a way to construct each prism in the upper half-space model of $\mathbb{H}^3$. There, the dihedral angles will be the angles at which the vertical and hemispherical geodesic planes intersect. We will find these planes by considering the lines and circles where they intersect with the plane $\{(x, y, z) \mid z = 0\}$, where we find lines and circles which must intersect at the same prescribed angles.

Such a prism is specified by nine positive integers, which we will denote as a_1 through a_9, corresponding to dihedral angles π/a_i. We label the prism as in Figure 4.

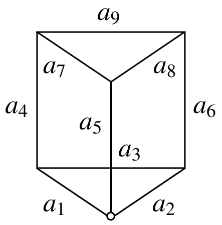


Figure 4. The labels a_1 through a_9.

We note that due to the reflectional symmetry of the prism, once we have treated one prism, we have also treated the prism one obtains by exchanging the pairs a_1 and a_2, a_4 and a_6, and a_7 and a_8.

There are restrictions on the combinations of values taken by the labels a_i which correspond to the conditions given in Theorem 2.1. Specifically:

- Condition (1) of Theorem 2.1 means that all labels must be positive (i.e., not 0 or ∞; we assume this of the a_i).
- Condition (2) states that at the five nonideal vertices, the three edges incident to the vertex must have the labels of a spherical triangle, and we add here that the three edges incident to the ideal vertex must have the labels of a Euclidean triangle.
- Condition (3) states that labels a_4, a_5 and a_6 must be the labels of a hyperbolic triangle.

We disregard the other two conditions: condition (4) does not apply because the prism has no prismatic 4-circuits, and any labeling of the prism which meets the stated conditions will already meet condition (5). This is because all dihedral angles in question are at most $\frac{\pi}{2}$, at most one of a_4, a_5 and a_6 can be 2 (the others must be larger) from condition 3, and at most one of a_1 and a_2 may be 2 from the ideal vertex condition.

3A. [2, 3, 6]-*cusp.*

Lemma 3.1. *There are eight infinite families and* 32 *specific configurations with labels* 2, 3, *and* 6 *at the ideal vertex.*

Proof. Here a_1, a_2 and a_5 take the values 2, 3 and 6. We first note that $a_5 \neq 2$. This is because whichever of a_4 or a_6 labels an edge which meets the edge labeled 6 must take the value 2, and then a_4, a_5 and a_6 are not the labels of a hyperbolic triangle.

Let $a_5 = 3$. Then, by symmetry, without loss of generality we suppose $a_1 = 2$ and $a_2 = 6$. Then $a_3 = a_6 = 2$. Since $a_6 = 2$ and $a_5 = 3$, we must have $a_4 \geq 7$, and thus that $a_7 = a_9 = 2$. The remaining label a_8 may take the values 2, 3, 4 or 5. Each of these four cases gives us one infinite family of labelings, indexed by $n \geq 7$ which corresponds to the value of a_4.

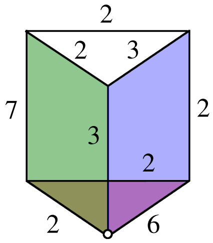


Figure 5. The left quadrilateral face is green; the right is blue; and the lower triangular face is red.

Now suppose $a_5 = 6$, and without loss of generality $a_1 = 2$ and $a_2 = 3$. Since $a_5 = 6$, we must have $a_7 = a_8 = 2$. If $a_3 = 4$ then we must have $a_6 = 2$ and $a_4 \leq 3$, in which case a_4, a_5 and a_6 are not the labels of a hyperbolic triangle. A similar argument applies if $a_3 > 4$. If $a_3 = 3$, then we must have $a_6 = 2$, and then a_4 must be 4 or 5. In each of these two cases, a_9 may be 2 or 3.

Finally, if again $a_5 = 6$, $a_1 = 2$, $a_2 = 3$, and $a_7 = a_8 = 2$, it remains to consider $a_3 = 2$. If $a_6 = 2$, then a_4 could be 4 or 5 — in each case a_9 is either 2 or 3 — or $a_4 \geq 6$, in which case $a_9 = 2$. If $a_6 = 3$, then a_4 could be 3, 4 or 5 — if $a_4 = 3$, a_9 is 2, 3, 4 or 5; if a_4 is 4 or 5, a_9 is 2 or 3 — or $a_4 \geq 6$, in which case $a_9 = 2$. If $a_6 = 4$, then a_4 could be 2, 3, 4 or 5 — in each case a_9 is either 2 or 3 — or $a_4 \geq 6$, in which case $a_9 = 2$. If $a_6 = 5$, then a_4 could be 2, 3, 4 or 5 — in each case a_9 is either 2 or 3 — or $a_4 \geq 6$, in which case $a_9 = 2$. □

We will compute explicitly the location of the prism in upper half-space for one label arrangement, where $a_1 = 2$, $a_2 = 6$, and $a_5 = 3$. Let $a_3 = 2$, $a_4 = 7$, $a_6 = 2$, $a_7 = 2$, $a_8 = 3$, and $a_9 = 2$. This is shown in Figure 5.

The values for all the edges were specifically chosen to satisfy the definitions on page 363. Using these definitions, we can examine different arrangements of the Euclidean vertex with values of 2, 3, and 6.

We will refer to the left quadrilateral face, whose edges have labels a_1, a_4, a_7 and a_5, as the red face. The left quadrilateral face, with edges a_2, a_6, a_9 and a_5, will be the blue face. The lower triangular face, with edges a_1, a_2 and a_3, will be the red face. These faces all meet at the ideal vertex, and hence correspond to vertical lines in the horizontal plane.

Making the back quadrilateral face correspond to the unit circle, we choose the red face to correspond to the line $x = 0$, which meets the unit circle at a right angle. By Lemma 2.2, the green face is the line $y = \cos\frac{\pi}{7}$, which meets the red face at a right angle, and the unit circle at $\frac{\pi}{7}$. The blue face is the line $y = \sqrt{3}x$, which meets the unit circle at a right angle and the red face at angle $\frac{\pi}{6}$ (see Figure 6).

It remains to find the last circle, corresponding to the top triangular face of the prism. Suppose this circle has center (x, y) and radius r. This circle intersects the

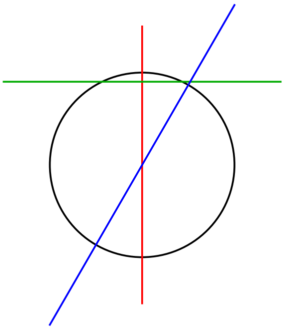


Figure 6. Three lines and the unit circle.

unit circle at an angle of $\frac{\pi}{2}$. Using the Pythagorean theorem to find an equation, we have that the equation of this intersection is

$$1 + r^2 = x^2 + y^2. \tag{1}$$

Since this circle meets the green line at a right angle, we see that the center must be on the green line, and hence that

$$y = \cos\tfrac{\pi}{7}. \tag{2}$$

We need one more equation in x, y and r, and this comes from the fact that the last circle meets the blue line at $\frac{\pi}{3}$. By Lemma 2.3,

$$y - r = \sqrt{3}x \tag{3}$$

(see Figure 7).

These equations, subject to $x > 0$ and $r > 0$, are sufficient to determine x, y and r, and thus to determine the location of the prism precisely. We find that as well as $y = \cos\frac{\pi}{7}$, we have

$$\begin{aligned} x &= \tfrac{1}{4}\big(2\sqrt{3}\cos\tfrac{\pi}{7} - \sqrt{6\sin\tfrac{3\pi}{14} - 2}\big) \approx 0.4504, \\ r &= \tfrac{1}{4}\big(\sqrt{18\sin\tfrac{3\pi}{14} - 6} - 2\cos\tfrac{\pi}{7}\big) \approx 0.1209. \end{aligned}$$

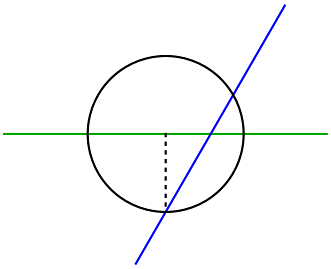


Figure 7. The point $(x, y - r)$ lies on the blue line.

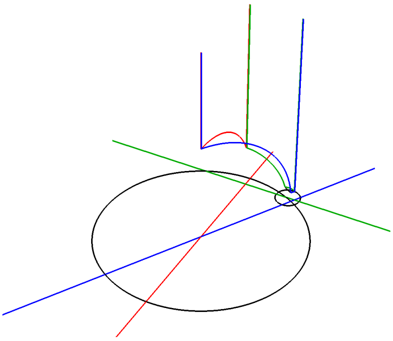


Figure 8. The lines and circles define the hyperbolic planes bounding the prism.

Now that we have found the equations of the three lines and two circles which intersect at the prescribed angles, the hyperbolic prism we seek is the region inside the triangle defined by the three straight lines, and exterior to the two spheres whose equators are the given circles (see Figure 8).

Changing the value of a_4 edge from 7 to 8, the equation of the green line changes to $y = \cos\frac{\pi}{8}$ by Lemma 2.2. And the same thing happens changing it to equal 9. This result shows that the line gets taller as the value of that side increases. That edge can be left as $a_4 = m > 6$.

3B. [2, 4, 4]-*cusp.*

Lemma 3.2. *There are four infinite families and* 24 *specific configurations with labels* 2, 4, *and* 4 *at the ideal vertex.*

Proof. First, we note that if $a_5 = 2$, then $a_1 = a_2 = 4$. Then a_4 and a_6 are both at most 3, and then a_4, a_5 and a_6 are not the labels of a hyperbolic triangle. Thus we must have $a_5 = 4$, and without loss of generality we will assume $a_1 = 2$ and $a_4 = 4$.

With these assumptions in place, we next note that a_3 must be 2 or 3. Let us first treat the case $a_3 = 3$. In this case, we must have $a_6 = 2$, and then $a_4 = 5$, because it forms a hyperbolic triangle with 2 and 4 and a spherical triangle with 2 and 3. The possible labels for (a_7, a_8, a_9) are then $(2, 2, 2)$, $(2, 2, 3)$, $(2, 3, 2)$, $(2, 3, 3)$ and $(3, 2, 2)$.

Finally, suppose $a_1 = 2$, $a_2 = 4$, $a_3 = 2$, and $a_5 = 4$. Then a_6 is either 2 or 3. If $a_6 = 2$, then $a_4 \geq 5$. If $a_4 = 5$, then as above, the possible labels for (a_7, a_8, a_9) are $(2, 2, 2)$, $(2, 2, 3)$, $(2, 3, 2)$, $(2, 3, 3)$ and $(3, 2, 2)$. If $a_4 \geq 6$, then $a_7 = a_9 = 2$, and a_8 may be 2 or 3. If $a_6 = 3$, then $a_4 \geq 3$. If $a_4 = 3$, then the possible labels for (a_7, a_8, a_9) are $(2, 2, 2)$, $(2, 2, 3)$, $(2, 2, 4)$, $(2, 2, 5)$, $(2, 3, 2)$ and $(3, 2, 2)$. If $a_4 = 4$ or 5, then the possible labels for (a_7, a_8, a_9) are $(2, 2, 2)$, $(2, 2, 3)$, $(2, 3, 2)$ and $(3, 2, 2)$. If $a_4 \geq 6$, then $a_7 = a_9 = 2$, and a_8 may be 2 or 3. □

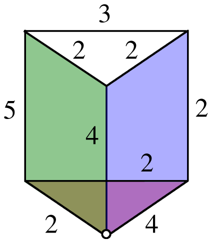


Figure 9. A [2, 4, 4]-cusp.

The first possibility of this arrangement we will work with is where $a_9 = 3$, $a_8 = 2$, $a_7 = 2$, $a_5 = 4$, $a_4 \geq 5$, $a_6 = 2$. This means $a_1 = 2$, $a_2 = 4$, and $a_3 = 2$ (see Figure 9).

With the bottom vertex being 2, 4, 4, we have a $\left(\frac{\pi}{2}, \frac{\pi}{4}, \frac{\pi}{4}\right)$ triangle, with blue, red and green sides. The back face of the prism meets green at an angle of $\frac{\pi}{5}$, and the back face meets the blue and red faces at $\frac{\pi}{2}$. The back face corresponds to the unit circle. We choose the red face to correspond to the line $x = 0$. By Lemma 2.2, the green face corresponds to the line $y = \cos\frac{\pi}{5}$. The blue line corresponds to the line $y = x$.

Finally, the top face meets blue and green at angle $\frac{\pi}{2}$, and the unit circle at angle $\frac{\pi}{3}$. Meeting blue and green at $\frac{\pi}{2}$ means the center of the last circle is on both blue and green lines. Their intersection point is $(x, y) = \left(\cos\frac{\pi}{5}, \cos\frac{\pi}{5}\right)$. Meeting the unit circle at $\frac{\pi}{3}$ means that, by Lemma 2.4, we have

$$x^2 + y^2 = 1^2 + r^2 - 2\left(\cos\tfrac{2\pi}{3}\right)r$$

and so

$$x^2 + y^2 = 1 + r^2 + r.$$

Here we find $r = \frac{1}{2}(\sqrt[4]{5} - 1)$.

As we saw, with the [2, 4, 4]-cusp the edges a_7, a_8 and a_9 are always labeled 2, 3 or 4, and they cannot all be labeled 3. We now describe what happens to the equations defining x, y and r when these labels change, noting that the red, blue and green faces, as well as the unit circle, are unaffected by these changes.

If we change a_9 to be 2, keeping $a_7 = a_8 = 2$, this means that the top face intersects the back face at $\frac{\pi}{2}$. The intersection between the last circle and the unit circle creates an angle of $\frac{\pi}{2}$. We still have $x = y = \cos\frac{\pi}{5}$, and the third equation becomes $x^2 + y^2 = 1 + r^2$.

If a_9 is kept as a 2, and we change a_8 to 3, then because the new circle meets the green line at right angles, we still have $y = \cos\frac{\pi}{5}$. Also, since $a_9 = 2$, the new circle meets the unit circle at right angles, and so $x^2 + y^2 = 1 + r^2$. By Lemma 2.3, the final equation in this case is

$$y = x + \tfrac{\sqrt{2}}{2}r.$$

The next arrangement we consider has $a_7 = a_8 = a_9 = 2$, the vertical edges satisfy $a_4 \geq 5$, $a_5 = 4$ and $a_6 = 2$, and the bottom edges are $a_1 = 2$, $a_2 = 4$, and $a_3 = 3$. This change shifts the red face from intersecting the unit circle at $x = 0$ to $x = -\frac{1}{2}$. The green line will be $y = \cos(\pi/a_4)$, and the blue will be $y = x$. Since $a_7 = a_8 = 2$, the center of the second circle is at the intersection of the blue and green lines. Letting $a_9 = 2$ results in the final equation of the last circle being $1 + r^2 = x^2 + y^2$.

Lastly, keeping $a_3 = 3$ $\big($so the red line is still $x = -\frac{1}{2}\big)$, we change a_6 to 3. By Lemma 2.2, this shifts the blue line downward from $y = x$ to $y = x - \frac{\sqrt{2}}{2}$. The equations defining x, y and r here are $y = \cos(\pi/a_4)$, $y = x - \frac{\sqrt{2}}{2}$ and $x^2 + y^2 = 1 + r^2$.

3C. [3, 3, 3]-*cusp.*

Lemma 3.3. *There are no infinite familes and* 22 *specific configurations with labels* 3, 3, *and* 3 *at the ideal vertex.*

Proof. In studying labelings here, we note that this labeling at the Euclidean vertex is symmetric, and thus we will discard some labelings as being symmetric to other labelings already listed.

We first note that if $a_3 > 2$, then $a_4 = a_6 = 2$ and we do not have a hyperbolic triangle. So $a_3 = 2$. Then by spherical triangles, a_4 and a_6 must both be one of $2, 3, 4$ or 5, but neither can be 2 because of the hyperbolic triangle, and also they cannot both be 3. Because of symmetry considerations, we suppose without loss of generality that $a_4 \leq a_6$:

- If $a_4 = 3$ and $a_6 = 4$, then (a_7, a_8, a_9) can be $(2, 2, 2)$, $(2, 2, 3)$, $(2, 3, 2)$, $(3, 2, 2)$, $(4, 2, 2)$ or $(5, 2, 2)$.
- If $a_4 = 3$ and $a_6 = 5$, then (a_7, a_8, a_9) can be $(2, 2, 2)$, $(2, 2, 3)$, $(2, 3, 2)$, $(3, 2, 2)$, $(4, 2, 2)$ or $(5, 2, 2)$.
- If $a_4 = 4$ and $a_6 = 4$, then (a_7, a_8, a_9) can be $(2, 2, 2)$, $(2, 2, 3)$ or $(2, 3, 2)$ (we discard $(3, 2, 2)$ as it is symmetric to $(2, 3, 2)$).
- If $a_4 = 4$ and $a_6 = 5$, then (a_7, a_8, a_9) can be $(2, 2, 2)$, $(2, 2, 3)$, $(2, 3, 2)$ or $(3, 2, 2)$.
- If $a_4 = 5$ and $a_6 = 5$, then (a_7, a_8, a_9) can be $(2, 2, 2)$, $(2, 2, 3)$ or $(2, 3, 2)$ (we discard $(3, 2, 2)$ as it is symmetric to $(2, 3, 2)$). □

The arrangement we examine is

$$[a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9,] = [3, 3, 2, 3, 3, 4, 2, 2, 2].$$

As before, we will let the back side of the prism be the unit circle. Since all of the admissible labelings here have $a_3 = 2$, we will fix the red line to be $x = 0$.

a_4, a_5, a_6	green line	blue line
3, 3, 4	$y=-\frac{\sqrt{3}}{3}x+\frac{\sqrt{3}}{3}$	$y=\frac{\sqrt{3}}{3}x-\frac{\sqrt{6}}{3}$
3, 3, 5	$y=-\frac{\sqrt{3}}{3}x+\frac{\sqrt{3}}{3}$	$y=\frac{\sqrt{3}}{3}x-\frac{2\sqrt{3}}{3}\cos\frac{\pi}{5}$
4, 3, 4	$y=-\frac{\sqrt{3}}{3}x+\frac{\sqrt{6}}{3}$	$y=\frac{\sqrt{3}}{3}x-\frac{\sqrt{6}}{3}$
4, 3, 5	$y=-\frac{\sqrt{3}}{3}x+\frac{\sqrt{6}}{3}$	$y=\frac{\sqrt{3}}{3}x-\frac{2\sqrt{3}}{3}\cos\frac{\pi}{5}$
5, 3, 5	$y=-\frac{\sqrt{3}}{3}x+\frac{2\sqrt{3}}{3}\cos\frac{\pi}{5}$	$y=\frac{\sqrt{3}}{3}x-\frac{2\sqrt{3}}{3}\cos\frac{\pi}{5}$

Table 1

Since the blue side meets the back side at $\frac{\pi}{4}$, the blue line intersects the unit circle at $\frac{\pi}{4}$. The green side meets the back side at $\frac{\pi}{3}$, and the red side meets the back side at $\frac{\pi}{2}$. This means that the red side creates a line that passes through the center of the unit circle at angle $\frac{\pi}{2}$. The red, green, and blue lines create an equilateral triangle since their sides meet each other at $\frac{\pi}{3}$.

If we shift the green line so that it passes through the center of the unit circle, we can find the slope of the blue line. Since this line creates an angle of $\frac{\pi}{3}$, in the fourth quadrant, there is a remaining $\frac{\pi}{6}$ in order to make a right angle. Thus the slope of the blue line is

$$\tan\frac{\pi}{6}=\frac{\sin\frac{\pi}{6}}{\cos\frac{\pi}{6}}=\frac{-\frac{1}{2}}{\frac{\sqrt{3}}{2}}=-\frac{\sqrt{3}}{3}.$$

The slope $-\frac{\sqrt{3}}{3}$ corresponds to the vector $\langle 1, -\frac{\sqrt{3}}{3}\rangle$. The vector $\langle 1, \sqrt{3}\rangle$ is orthogonal to this vector, and its corresponding unit vector is $\langle\frac{1}{2}, \frac{\sqrt{3}}{2}\rangle$. Using Lemma 2.2, we multiply the unit vector by $\cos\frac{\pi}{3}$ to get the point $\left(\frac{1}{4}, \frac{\sqrt{3}}{4}\right)$. Thus we get the line $y=\left(-\frac{\sqrt{3}}{3}\right)x+\frac{\sqrt{3}}{3}$.

The slope of the blue line is $\frac{\sqrt{3}}{3}$. The vector corresponding to this slope is $\langle 1, -\sqrt{3}\rangle$ and the unit vector is $\langle\frac{1}{2}, -\frac{\sqrt{3}}{2}\rangle$. We multiply the unit vector by $\cos\frac{\pi}{4}$ to find the point $\left(\frac{\sqrt{2}}{4}, -\frac{\sqrt{6}}{4}\right)$. Thus we get the line $y=\left(\frac{\sqrt{3}}{3}\right)x-\frac{\sqrt{6}}{3}$.

Since $a_7=a_8=a_9=2$, the last circle intersects the unit circle at right angles. This also means that the center of this last circle lies on the intersection of the green and blue lines. Since the last circle intersects the unit circle at right angles, we use the Pythagorean theorem to find that the last equation is $1+r^2=x^2+y^2$.

Table 1 lists the equations of these arrangements, where we adjust the values of a_4 and a_6.

As long as $a_7=a_8=a_9=2$, the third equation defining the last circle that intersects the unit circle does not change for all of these specific arrangements. Thus, for all of these arrangements, the equation of the last circle is $1+r^2=x^2+y^2$. When we change a_9 to 3, the equation of the last circle for the arrangements listed above is $x^2+y^2=1+r^2+r$. When a_7 or a_8 change, the equations change according to Lemmas 2.3 and 2.4.

4. Matrices

We describe a general method which produces lines and circles which intersect at the prescribed angles, which produces the same results described in the previous section. We then show how to take this geometric data of the lines and circles and use it to produce matrix generators in $\mathrm{PSL}_2(\mathbb{C})$ for the orientation-preserving subgroup of the group generated by reflections in the faces of the corresponding prism. Each group will be generated by four matrices, where each matrix acts by pairing two faces of the polyhedron one obtains by doubling the prism across one face.

We note at this point that this method also works for examples of prisms which have more than one ideal vertex, that is, where one or more of the five vertices not located at ∞ in the upper half-space is found in the complex plane. In this paper, we restrict to the case of prisms with one cusp, but the method we describe should work for some, if not all, labelings of the prism which correspond to prisms with more than one cusp.

4A. ***The case $a_3 = 2$.*** As above, we suppose that the back quadrilateral face lies on the unit sphere, or equivalently that one of the two circles is the unit circle. We further suppose that the red face lies above the imaginary axis, or equivalently that one of the straight lines is $x = 0$; this corresponds to asking that $a_3 = 2$. Lastly, we assume that the polyhedron lies to the right of the imaginary axis as we view it from above; in other words, we assume that it lies in the region of $\mathbb{H}^3$ with $x \geq 0$.

By considering Figure 1 and applying Lemma 2.2, we see that the blue line has equation

$$y = \cot\left(\frac{\pi}{a_2}\right)x - \frac{\cos(\pi/a_6)}{\sin(\pi/a_2)}$$

and, by similar reasoning, the green line has equation

$$y = -\cot\left(\frac{\pi}{a_1}\right)x + \frac{\cos(\pi/a_4)}{\sin(\pi/a_1)}.$$

The final face is on a circle with center (x, y) and radius r, which meets the blue line at angle π/a_8. By Lemma 2.3 we have one equation

$$y - \frac{r\cos(\pi/a_8)}{\sqrt{\cot^2(\pi/a_2)+1}} = \cot\left(\frac{\pi}{a_2}\right)\left(x + \frac{r\cot(\pi/a_2)\cos(\pi/a_8)}{\sqrt{\cot^2(\pi/a_2)+1}}\right) - \frac{\cos(\pi/a_6)}{\sin(\pi/a_2)},$$

which simplifies to

$$\begin{aligned} y - r\cos\left(\frac{\pi}{a_8}\right)\sin\left(\frac{\pi}{a_2}\right) \\ = \cot\left(\frac{\pi}{a_2}\right)\left(x + r\cos\left(\frac{\pi}{a_2}\right)\cos\left(\frac{\pi}{a_8}\right)\right) - \frac{\cos(\pi/a_6)}{\sin(\pi/a_2)}. \end{aligned} \tag{4}$$

By applying Lemma 2.3, with appropriate modifications, to the green line, we also have the equation

$$\begin{aligned} y + r\sin\left(\frac{\pi}{a_1}\right)\cos\left(\frac{\pi}{a_7}\right) \\ = -\cot\left(\frac{\pi}{a_1}\right)\left(x + r\cos\left(\frac{\pi}{a_1}\right)\cos\left(\frac{\pi}{a_7}\right)\right) + \frac{\cos(\pi/a_4)}{\sin(\pi/a_1)}. \end{aligned} \tag{5}$$

Finally, the two circles intersecting at angle π/a_9 yield, via the cosine law and Lemma 2.4, the equation

$$x^2 + y^2 = 1^2 + r^2 - 2(1)(r)\cos\left(\pi - \frac{\pi}{a_9}\right)$$

or

$$x^2 + y^2 = 1^2 + r^2 + 2r\cos\left(\frac{\pi}{a_9}\right). \tag{6}$$

Equations (4), (5) and (6) together define x, y and r and determine the final circle required.

With all of this work in mind, we require seven quantities to write down the matrices we seek. These quantities are $y_1 = \cos(\pi/a_4)/\sin(\pi/a_1)$ and $y_2 = -\cos(\pi/a_6)/\sin(\pi/a_2)$, the y-intercepts of the green and blue lines, the angles $\theta_1 = \pi/a_1$ and $\theta_2 = \pi/a_2$ of the ideal triangle at these points, and the center (x, y) and radius r of the second circle. Given these quantities, the matrices are

$$M_1 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

which pairs two sides which both lie on the unit sphere,

$$M_2 = \begin{pmatrix} e^{-i\theta_1} & y_1 i(e^{i\theta_1} - e^{-i\theta_1}) \\ 0 & e^{i\theta_1} \end{pmatrix},$$

which rotates counterclockwise by angle $2\theta_1$ about $(0, y_1)$,

$$M_3 = \begin{pmatrix} e^{i\theta_2} & y_2 i(e^{-i\theta_2} - e^{i\theta_2}) \\ 0 & e^{-i\theta_2} \end{pmatrix},$$

which rotates clockwise by angle $2\theta_2$ about $(0, y_2)$, and

$$M_4 = \begin{pmatrix} \frac{1}{r}(-x+yi) & \frac{1}{r}(x^2+y^2)-r \\ \frac{1}{r} & \frac{1}{r}(-x-yi) \end{pmatrix},$$

which sends the second circle to its reflection in the imaginary axis. These matrices satisfy the relations

$$\begin{aligned} M_2^{a_1} &= 1, & M_3^{a_2} &= 1, & M_1^{a_3} = M_1^2 &= 1, \\ (M_2^{-1}M_1)^{a_4} &= 1, & (M_3^{-1}M_2)^{a_5} &= 1, & (M_3^{-1}M_1)^{a_6} &= 1, \\ (M_4^{-1}M_2)^{a_7} &= 1, & (M_4^{-1}M_3)^{a_8} &= 1, & (M_4^{-1}M_1)^{a_9} &= 1. \end{aligned}$$

4B. ***The case $a_3 = 3$.*** In the case that $a_3 \neq 2$, we saw that $a_3 = 3$. In this case, we keep the back face as corresponding to the unit circle, and move the red line to $x = -\frac{1}{2}$ so that it intersects the unit circle at angle $\frac{\pi}{3}$. The equations of the green and blue lines will be the same as in the case $a_3 = 2$, and the second circle will be defined by the same three equations (4), (5) and (6). As in the previous case, we define $\theta_1 = \pi/a_1$ and $\theta_2 = \pi/a_2$, and let (x, y) and r be the center and radius of the second circle. In place of y_1 and y_2 we define

$$\begin{aligned} z_1 &= -\tfrac{1}{2} + \left(\frac{\cos(\pi/a_4)}{\sin(\pi/a_1)} + \frac{\cot(\pi/a_1)}{2} \right) i, \\ z_2 &= -\tfrac{1}{2} + \left(\frac{\cos(\pi/a_6)}{\sin(\pi/a_2)} - \frac{\cot(\pi/a_2)}{2} \right) i, \end{aligned}$$

the points where the green and blue lines meet the red line $x = -\frac{1}{2}$. Our matrices are then

$$M_1 = \begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix},$$

which pairs two sides which lie on the unit sphere and the unit sphere centered at $(-1, 0)$,

$$M_2 = \begin{pmatrix} e^{-i\theta_1} & z_1(e^{i\theta_1} - e^{-i\theta_1}) \\ 0 & e^{i\theta_1} \end{pmatrix},$$

which rotates counterclockwise by angle $2\theta_1$ about z_1,

$$M_3 = \begin{pmatrix} e^{i\theta_2} & z_2(e^{-i\theta_2} - e^{i\theta_2}) \\ 0 & e^{-i\theta_2} \end{pmatrix},$$

which rotates clockwise by angle $2\theta_2$ about z_2, and

$$M_4 = \begin{pmatrix} \frac{1}{r}(-(x+1)+yi) & \frac{1}{r}(-(x+1)+yi)(-x-yi)-r \\ \frac{1}{r} & \frac{1}{r}(-x-yi) \end{pmatrix},$$

which sends the second circle to its reflection in the line $x = -\frac{1}{2}$. These matrices satisfy the relations

$$\begin{aligned} M_2^{a_1} &= 1, & M_3^{a_2} &= 1, & M_1^{a_3} = M_1^3 &= 1, \\ (M_2^{-1}M_1)^{a_4} &= 1, & (M_3^{-1}M_2)^{a_5} &= 1, & (M_3^{-1}M_1)^{a_6} &= 1, \\ (M_4^{-1}M_2)^{a_7} &= 1, & (M_4^{-1}M_3)^{a_8} &= 1, & (M_4^{-1}M_1)^{a_9} &= 1. \end{aligned}$$

5. Results

We now list all of the possible labelings of the prism that were described in Section 3. We note that the twelve infinite families, where one label may vary, are written in bold to distinguish them.

5A. [2, 3, 6]-*cusp.*

5A1. $a_3 = 3$.

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
2	3	3	4	6	2	2	2	2
2	3	3	4	6	2	2	2	3
2	3	3	5	6	2	2	2	2
2	3	3	5	6	2	2	2	3

5A2. $a_3 = 2$.

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
2	**6**	**2**	**$n \geq 7$**	**3**	**2**	**2**	**2**	**2**
2	**6**	**2**	**$n \geq 7$**	**3**	**2**	**2**	**3**	**2**
2	**6**	**2**	**$n \geq 7$**	**3**	**2**	**2**	**4**	**2**
2	**6**	**2**	**$n \geq 7$**	**3**	**2**	**2**	**5**	**2**
2	3	2	4	6	2	2	2	2
2	3	2	4	6	2	2	2	3
2	3	2	5	6	2	2	2	2
2	3	2	5	6	2	2	2	3
2	**3**	**2**	**$n \geq 6$**	**6**	**2**	**2**	**2**	**2**
2	3	2	3	6	3	2	2	2
2	3	2	3	6	3	2	2	3
2	3	2	3	6	3	2	2	4
2	3	2	3	6	3	2	2	5
2	3	2	4	6	3	2	2	2
2	3	2	4	6	3	2	2	3
2	3	2	5	6	3	2	2	2
2	3	2	5	6	3	2	2	3
2	**3**	**2**	**$n \geq 6$**	**6**	**3**	**2**	**2**	**2**
2	3	2	2	6	4	2	2	2
2	3	2	2	6	4	2	2	3
2	3	2	3	6	4	2	2	2
2	3	2	3	6	4	2	2	3
2	3	2	4	6	4	2	2	2
2	3	2	4	6	4	2	2	3
2	3	2	5	6	4	2	2	2
2	3	2	5	6	4	2	2	3
2	**3**	**2**	**$n \geq 6$**	**6**	**4**	**2**	**2**	**2**
2	3	2	2	6	5	2	2	2
2	3	2	2	6	5	2	2	3
2	3	2	3	6	5	2	2	2
2	3	2	3	6	5	2	2	3
2	3	2	4	6	5	2	2	2
2	3	2	4	6	5	2	2	3
2	3	2	5	6	5	2	2	2
2	3	2	5	6	5	2	2	3
2	**3**	**2**	**$n \geq 6$**	**6**	**5**	**2**	**2**	**2**

5B. [2, 4, 4]-*cusp.*

5B1. $a_3 = 3$.

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
2	4	3	5	4	2	2	2	2
2	4	3	5	4	2	2	2	3
2	4	3	5	4	2	2	3	2
2	4	3	5	4	2	2	3	3
2	4	3	5	4	2	3	2	2

5B2. $a_3 = 2$.

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
2	4	2	5	4	2	2	2	2
2	4	2	5	4	2	2	2	3
2	4	2	5	4	2	2	3	2
2	4	2	5	4	2	2	3	3
2	4	2	5	4	2	3	2	2
2	**4**	**2**	$\boldsymbol{n \geq 6}$	**4**	**2**	**2**	**2**	**2**
2	**4**	**2**	$\boldsymbol{n \geq 6}$	**4**	**2**	**2**	**3**	**2**
2	4	2	3	4	3	2	2	2
2	4	2	3	4	3	2	2	3
2	4	2	3	4	3	2	2	4
2	4	2	3	4	3	2	2	5
2	4	2	3	4	3	2	3	2

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
2	4	2	3	4	3	3	2	2
2	4	2	4	4	3	2	2	2
2	4	2	4	4	3	2	2	3
2	4	2	4	4	3	2	3	2
2	4	2	4	4	3	3	2	2
2	4	2	5	4	3	2	2	2
2	4	2	5	4	3	2	2	3
2	4	2	5	4	3	2	3	2
2	4	2	5	4	3	3	2	2
2	**4**	**2**	$\boldsymbol{n \geq 6}$	**4**	**3**	**2**	**2**	**2**
2	**4**	**2**	$\boldsymbol{n \geq 6}$	**4**	**3**	**2**	**3**	**2**

5C. [3, 3, 3]-*cusp.*

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
3	3	2	3	3	4	2	2	2
3	3	2	3	3	4	2	2	3
3	3	2	3	3	4	2	3	2
3	3	2	3	3	4	3	2	2
3	3	2	3	3	4	4	2	2
3	3	2	3	3	4	5	2	2
3	3	2	3	3	5	2	2	2
3	3	2	3	3	5	2	2	3
3	3	2	3	3	5	2	3	2
3	3	2	3	3	5	3	2	2
3	3	2	3	3	5	4	2	2

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
3	3	2	3	3	5	5	2	2
3	3	2	4	3	4	2	2	2
3	3	2	4	3	4	2	2	3
3	3	2	4	3	4	2	3	2
3	3	2	4	3	5	2	2	2
3	3	2	4	3	5	2	2	3
3	3	2	4	3	5	2	3	2
3	3	2	4	3	5	3	2	2
3	3	2	5	3	5	2	2	2
3	3	2	5	3	5	2	2	3
3	3	2	5	3	5	2	3	2

Acknowledgments

We wish to thank Neil Hoffman for suggesting this project, and for help checking the results, and Charles Delman for his guidance and support. We thank Matthieu Jacquemet and the referee for helpful comments.

References

[Andreev 1970] E. M. Andreev, "On convex polyhedra in Lobachevskii spaces", *Mat. Sb.* (*N.S.*) **81** (1970), 445–478. In Russian; translated in *Math. USSR-Sb.* **10** (1970), 413-440. MR Zbl

[Brunner et al. 1985] A. M. Brunner, Y. W. Lee, and N. J. Wielenberg, "Polyhedral groups and graph amalgamation products", *Topology Appl.* **20**:3 (1985), 289–304. MR Zbl

[Şengün 2012] M. H. Şengün, "On the torsion homology of non-arithmetic hyperbolic tetrahedral groups", *Int. J. Number Theory* **8**:2 (2012), 311–320. MR Zbl

[Hoffman 2014] N. R. Hoffman, "On knot complements that decompose into regular ideal dodecahedra", *Geom. Dedicata* **173** (2014), 299–308. MR Zbl

[Inoue 2015] T. Inoue, "The 825 smallest right-angled hyperbolic polyhedra", preprint, 2015. arXiv

[Kaplinskaja 1974] I. M. Kaplinskaja, "Discrete groups generated by reflections in the faces of simplicial prisms in Lobachevskii spaces", *Mat. Zametki* **15**:1 (1974), 159–164. In Russian; translated in *Math. Notes Acad. Sci. USSR* **15**:1 (1974), 88–91. MR Zbl

[Lakeland 2010] G. S. Lakeland, "Matrix realizations of the hyperbolic tetrahedral groups in $\mathrm{PSL}_2(\mathbb{C})$", unpublished note, 2010, http://www.ux1.eiu.edu/~gslakeland/tetrahedral_realizations.pdf.

[Lannér 1950] F. Lannér, "On complexes with transitive groups of automorphisms", *Comm. Sém. Math. Univ. Lund* **11** (1950), 71. MR Zbl

[Roeder et al. 2007] R. K. W. Roeder, J. H. Hubbard, and W. D. Dunbar, "Andreev's theorem on hyperbolic polyhedra", *Ann. Inst. Fourier* (*Grenoble*) **57**:3 (2007), 825–882. MR Zbl

[Thurston 1979] W. P. Thurston, "The geometry and topology of three-manifolds", lecture notes, Princeton University, 1979, http://msri.org/publications/books/gt3m.

[Vinberg 1967] E. B. Vinberg, "Discrete groups generated by reflections in Lobachevskii spaces", *Mat. Sb.* (*N.S.*) **72 (114)** (1967), 471–488. In Russian; translated in *Math. USSR-Sb.* **1**:3 (1967), 429–444. MR Zbl

Received: 2018-10-03 Revised: 2019-12-15 Accepted: 2020-03-23

gslakeland@eiu.edu
Department of Mathematics and Computer Science,
Eastern Illinois University, Charleston, IL, United States

cgbarnett@eiu.edu
Department of Mathematics and Computer Science,
Eastern Illinois University, Charleston, IL, United States

INVOLVE 13:3 (2020)
https://doi.org/10.2140/involve.2020.13.381

On the sandpile group of Eulerian series-parallel graphs

Kyle Weishaar and James Seibert

(Communicated by Joshua Cooper)

A sandpile configuration is a representation of the current layout of theoretical sand on a graph in which every vertex is assigned a nonnegative integer value. The Abelian sandpile group is a finite group composed of the recurrent sandpile configurations of a graph. We investigate the sandpile group of graphs constructed using the composition rules of series-parallel graphs, and determine the sandpile groups of parallel compositions of path-graphs.

1. Introduction

The Abelian sandpile model was originally introduced as a simple example of self-organized criticality [Bak et al. 1988]. The model was later generalized into a form that can be expressed in terms of a rooted graph [Dhar et al. 1995]. On such a graph, nonnegative integers are assigned to each vertex to represent the current configuration of sand. Toppling happens when a vertex with an amount of sand that is at least its degree transfers one grain of sand to each of its neighboring vertices. The root (also called the sink in this context) is the only vertex that is not allowed to topple. The sink acts as an outlet for sand to escape the system. As a result, the amount of sand on the sink is inconsequential and therefore not included in the configuration of a graph. A configuration with no toppleable vertices is said to be stable. Two configurations can be added by adding the sandpiles on corresponding vertices and then toppling until stable. A configuration is recurrent if it is stable and can be reached from any starting configuration by adding some other configuration. The set of recurrent configurations makes up a finite Abelian group known as the sandpile group. The order of the sandpile group is equal to the number of spanning trees of the graph [Dhar et al. 1995]. The recently published book [Corry and Perkinson 2018, Chapters 6–9] provides a comprehensive introduction to all of these ideas, and much more.

MSC2010: 05C50.
Keywords: sandpile group, Eulerian graph, Laplacian, Smith normal form.

Previous works have classified the sandpile groups of several families of graphs [Cori and Rossin 2000; Hou et al. 2006; Levine 2009; Shen and Hou 2008; Toumpakari 2007]. Furthermore, the sandpile groups of the Cartesian products of many types of graphs have also been identified [Hou et al. 2008; Jacobson et al. 2003; Liang et al. 2008]. A better understanding of how graph compositions affect sandpile groups may allow for the classification of sandpile groups on more complex graphs. The growing catalog of classified sandpile groups can be used to view complicated graphs as the composition of graphs with known sandpile groups. In this spirit, we investigate the sandpile group of series-parallel graphs constructed from graphs with known groups.

This paper is organized as follows. In Section 2 we provide necessary background information. In Section 3 we prove the general form of series compositions involving an Eulerian graph. We present this known result (it appears as an exercise in [Corry and Perkinson 2018]) in order to prepare the reader for the new result on parallel compositions of paths in Section 4. We demonstrate a significant reduction that lends itself to several corollaries. In Section 5 we find the sandpile group of an example series-parallel graph and offer concluding remarks.

2. Background

On a directed graph, the number of edges leaving a vertex is called the outdegree. Similarly, the number of edges entering a vertex is called the indegree. A graph is said to be Eulerian if the outdegree of every vertex is equal to its indegree. A graph with v vertices can be represented by a $v \times v$ matrix known as the Laplacian matrix, L. The Laplacian of a graph is defined by

$$L_{i,j} = \begin{cases} \text{outdegree of vertex } i & \text{if } i = j, \\ -(\text{number of edges directed from } j \text{ to } i) & \text{if } i \neq j. \end{cases}$$

Column j of the Laplacian then reflects the result of toppling vertex j. Specifically, vertex j loses grains of sand equal to its outdegree (represented by the positive number on the diagonal), and each vertex connected to j gains grains of sand equal to the number of edges coming from vertex j (represented by the corresponding off-diagonal entries). The reduced Laplacian is obtained by removing the row and column that corresponds to the sink of the full Laplacian. The choice of sink does not impact the sandpile group of an Eulerian graph [Cori and Rossin 2000]. For such graphs, the reduced Laplacian can be created by removing any corresponding column and row.

The Laplacian of a graph determines its sandpile group in the following manner. We can define an equivalence relation on the additive group $\mathbb{Z}^v$ by declaring two configurations (now allowing negative as well as nonnegative integers) equivalent if one can be reached from the other by a sequence of toppling (or untoppling) moves.

Each equivalence class has a unique representative that is a recurrent configuration. (This nontrivial result is proven in several sources, including [Cori and Rossin 2000; Corry and Perkinson 2018].) The columns of the Laplacian are a complete set of relations using the standard basis vectors as a generating set of the group. In other words, the Laplacian is a presentation matrix for a finitely generated Abelian group. The torsion subgroup of this group is the sandpile group.

In order to identify the sandpile group, we can manipulate the presentation matrix of a group by using the following elementary row and columns operations [Dummit and Foote 2004]:

(1) Add an integer multiple of one row or column to another. This changes the generating set of the group and the corresponding relations.

(2) Multiply a row or column by -1. This replaces the generator with its inverse.

(3) Delete a column of zeros. This eliminates a trivial relation.

(4) Delete a row and a column that form the standard basis vector. This eliminates a redundant generator.

(5) Swapping corresponding rows and columns. This reorders the generators and adjusts the relations accordingly.

These operations do not change the group represented. If a matrix A can be changed into a matrix B by use of these operations, we will say that A is $\mathbb{Z}$-equivalent to B and we will write $A \equiv B$. Note that operations (1), (2), and (5) can be accomplished by left or right multiplication by an invertible matrix.

Every integer matrix has a unique Smith normal form, arrived at using the row and column operations above. The Smith normal form is defined as the diagonal matrix $\mathbb{Z}$-equivalent to the original in which every entry is divisible by the previous entries. For every matrix M with integer entries, there exist invertible matrices P and Q such that PMQ is in Smith normal form. The Smith normal form of the matrix can be written as an $(n+m) \times (n+m)$ matrix

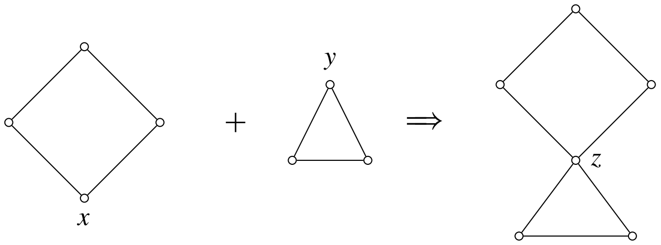
$$\begin{bmatrix} a_{1,1} & & & & & \\ & \ddots & & & & \\ & & a_{n,n} & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix}$$

in which $a_{i,i}$ divides $a_{i+1,i+1}$. This matrix represents the group

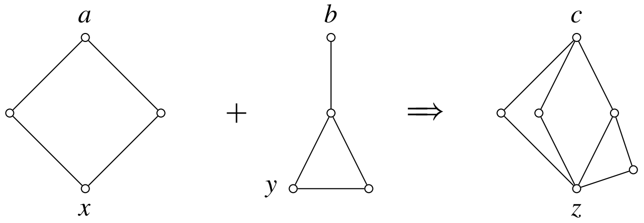
$$\left(\bigoplus_{i=1}^{n} \mathbb{Z}_{a_{i,i}}\right) \oplus \mathbb{Z}^m.$$

We arrive at the sandpile group by discarding the infinite summands and keeping the finite Abelian group that remains (the torsion subgroup). A common technique used to determine the sandpile group of a family of graphs is to show that the Laplacian matrix (or reduced Laplacian) for graphs in the family is equivalent to a predictable Smith normal form.

This paper focuses on the sandpile groups of graphs constructed using the composition rules of series-parallel graphs. A series composition of two graphs, G_1 and G_2, is formed by taking a vertex $x \in G_1$ and a vertex $y \in G_2$ and merging x with y to produce the vertex z, while all other existing vertices and connections remain unchanged:



Similarly, a parallel composition of two graphs, G_1 and G_2, is formed by taking the vertices $a, x \in G_1$ and $b, y \in G_2$, merging a with b to produce the vertex c, and merging x with y to produce the vertex z, while all other existing vertices and connections remain unchanged:



3. Series compositions

Theorem 3.1. *Let G_1 and G_2 be Eulerian graphs with the sandpile groups SG_1 and SG_2, respectively. A series composition of G_1 and G_2 will have a sandpile group of the form $SG_1 \oplus SG_2$.*

Proof. Let G_1 and G_2 be Eulerian graphs with the respective sandpile groups SG_1 and SG_2. The Laplacian of a series composition of G_1 and G_2 will be composed of the values $a_{i,j}$, which come from the Laplacian of G_1, listing the vertex to merge last, and the values $b_{i,j}$, which come from the Laplacian of G_2, listing the vertex to merge first. The merged vertex has outdegree $a_{n,n} + b_{1,1}$ and connections to G_1

and G_2 preserved. Therefore, the Laplacian is of the form

$$\begin{bmatrix} a_{1,1} & \cdots & a_{1,n} & & \\ \vdots & \ddots & \vdots & & \\ a_{n,1} & \cdots & a_{n,n}+b_{1,1} & \cdots & b_{1,m} \\ & & \vdots & \ddots & \vdots \\ & & b_{m,1} & \cdots & b_{m,m} \end{bmatrix}.$$

We will now use the elementary row operations defined in Section 2 to show this Laplacian is $\mathbb{Z}$-equivalent to a useful diagonal matrix. Since G_1 and G_2 are both Eulerian graphs, the series composition of these graphs will also be Eulerian. Therefore, we can let the sink be represented by the n-th row and column of the full Laplacian. As a result, the reduced Laplacian will have the form

$$\left[\begin{array}{ccc|ccc} a_{1,1} & \cdots & a_{1,n-1} & & & \\ \vdots & \ddots & \vdots & & & \\ a_{n-1,1} & \cdots & a_{n-1,n-1} & & & \\ \hline & & & b_{2,2} & \cdots & b_{2,m} \\ & & & \vdots & \ddots & \vdots \\ & & & b_{m,2} & \cdots & b_{m,m} \end{array}\right] = \left[\begin{array}{c|c} A & \\ \hline & B \end{array}\right].$$

Let P_1 and Q_1 be the invertible matrices that put A in Smith normal form. Similarly, let P_2 and Q_2 be the invertible matrices that put B in Smith normal form. Then

$$\left[\begin{array}{c|c} P_1 & \\ \hline & P_2 \end{array}\right] \cdot \left[\begin{array}{c|c} A & \\ \hline & B \end{array}\right] \cdot \left[\begin{array}{c|c} Q_1 & \\ \hline & Q_2 \end{array}\right]$$

would produce the matrix

$$\left[\begin{array}{ccc|ccc} f_1 & & & & & \\ & \ddots & & & & \\ & & f_{n-1} & & & \\ \hline & & & g_1 & & \\ & & & & \ddots & \\ & & & & & g_{m-1} \end{array}\right],$$

where the values $f_1, \ldots, f_{n-1}$ and $g_1, \ldots, g_{m-1}$ are the nonzero entries of the Smith normal form of the Laplacian of G_1 and G_2, respectively. This matrix presents

$$\left(\bigoplus_{i=1}^{n-1} f_i\right) \oplus \left(\bigoplus_{j=1}^{m-1} g_j\right) = SG_1 \oplus SG_2.$$

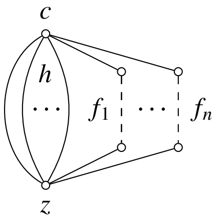
Since the reduced Laplacian of a series composition of the Eulerian graphs is $\mathbb{Z}$-equivalent to a diagonal matrix with the nonzero entries of the Smith normal forms of G_1 and G_2 as values, the sandpile group of a series composition of G_1 and G_2 is $SG_1 \oplus SG_2$. □

Corollary 3.2. *The sandpile group of any series combination of two Eulerian graphs, G_1 and G_2, will be of the form $SG_1 \oplus SG_2$, independent of which vertices are merged.*

Proof. The choice of vertices to merge simply rearranges the rows and columns using the fifth elementary operation in order to list the merged vertices in an overlapping row. The sandpile groups of G_1 and G_2 are unique, independent of the order in which the vertices are listed and independent of the vertices chosen as the sinks [Cori and Rossin 2000]. The sandpile group of the series composition of graphs will be $SG_1 \oplus SG_2$, no matter which vertices are merged. □

4. Parallel composition of paths

We now focus on a family of undirected multigraphs made of the parallel composition of h paths of length 1 and n paths of the varying lengths $f_1, \dots, f_n$, where $f_i \geq 2$. Graphs in this family have the following form:



We define a partitioned matrix that represents the general Laplacian of graphs in this family as follows. The top row of the Laplacian will correspond to the vertex c and it will have the form

$$\begin{bmatrix} n+h & D_1 & \cdots & D_n & -h \end{bmatrix},$$

where each D_i is a $1 \times (f_i - 1)$ matrix of the form $D_i = \begin{bmatrix} -1 & 0 & \cdots & 0 \end{bmatrix}$. The first entry of this row corresponds to the degree of c, which is the sum of the number of paths of length 1 and the number of paths of length 2 or greater. Each D_i matrix corresponds to the path f_i, with its vertices listed from top to bottom. We see that c is only connected to the first vertex of each path. The last entry of the top row of the Laplacian corresponds to the h paths of length 1 connecting the vertices c and z. The Laplacian of an undirected graph is symmetric. Therefore, the first column will be the transpose of the first row.

The bottom row of the Laplacian will correspond to vertex z and have the form

$$\begin{bmatrix} -h & C_1 & \cdots & C_n & n+h \end{bmatrix},$$

where each C_i is a $1 \times (f_i - 1)$ matrix of the form $C_i = \begin{bmatrix} 0 & \cdots & 0 & -1 \end{bmatrix}$. The first entry reflects the h connections from z to c, while the last represents the total degree of $n+h$. Each C_i corresponds to the path f_i and shows that z is only connected to the last vertex in this path. The last column of the Laplacian will be the transpose of the bottom row.

The vertices of the path f_i (excluding c and z, accounted for above) correspond to a diagonal block in the Laplacian, B_i. When $f_i > 2$, B_i is an $(f_i - 1) \times (f_i - 1)$ matrix of the form

$$B_i = \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}.$$

Here we see that each vertex in the path has degree 2 and is connected to the vertices immediately before and after it. When $f_i = 2$, B_i is the 1×1 matrix with entry 2, and the corresponding matrices D_i and C_i are 1×1 matrices consisting of entry -1. Putting this all together we see that the full Laplacian of the parallel composition of h paths of length 1 and n paths of the varying lengths $f_1, \ldots, f_n$, where $f_i \geq 2$, is of form

$$\left[\begin{array}{c|c|c|c|c} n+h & D_1 & \cdots & D_n & -h \\ \hline D_1^{\mathsf{T}} & B_1 & & & C_1^{\mathsf{T}} \\ \hline \vdots & & \ddots & & \vdots \\ \hline D_n^{\mathsf{T}} & & & B_n & C_n^{\mathsf{T}} \\ \hline -h & C_1 & \cdots & C_n & n+h \end{array}\right].$$

The empty partitions are zero matrices that reflect the fact that vertices from distinct paths f_i do not connect to each other. Now that we have the Laplacian for this graph of parallel paths, we can use the row operations from Section 2 to simplify it as much as possible. Since the graph is undirected (and therefore Eulerian), we can let the first row and column represent the sink. The resulting reduced Laplacian will be of the form

$$\left[\begin{array}{c|c|c|c} B_1 & & & C_1^{\mathsf{T}} \\ \hline & \ddots & & \vdots \\ \hline & & B_n & C_n^{\mathsf{T}} \\ \hline C_1 & \cdots & C_n & n+h \end{array}\right].$$

The following lemma will allow us to further reduce this matrix.

Lemma 4.1. *An $m \times m$ matrix F_m, where $m \geq 2$, of the form*

$$F_m = \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}$$

satisfies

$$F_m \equiv \begin{bmatrix} m & -1 \\ -(m-1) & 2 \end{bmatrix}.$$

Proof. We start by considering cases with small m. When $m = 2$,

$$F_2 = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ -(2-1) & 2 \end{bmatrix},$$

which is in the desired form.

When $m = 3$, we have the matrix

$$F_3 = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}.$$

We create the standard basis vector in the first row and column by adding the second column to the first, and then adding the first column to the second. The top row will now have 1 as the first entry and zeros everywhere else. Next, we use the first row to eliminate the entries in the rest of the first column:

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

We now use the fourth elementary operation to remove the first row and column, and see the desired form

$$\begin{bmatrix} 3 & -1 \\ -2 & 2 \end{bmatrix} = \begin{bmatrix} 3 & -1 \\ -(3-1) & 2 \end{bmatrix}.$$

The $m = 3$ example shows the basic process that we can iterate. If we start with a matrix of the form

$$\begin{bmatrix} k & -1 & & & \\ -(k-1) & 2 & -1 & & \\ & -1 & 2 & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{bmatrix},$$

we can add $(k-1)$ times the second column to the first column, and then add the first column to the second column to get

$$\begin{bmatrix} 1 & 0 & & \cdots & & 0 \\ k{-}1 & k{+}1 & -1 & & & \\ -(k{-}1) & -k & 2 & -1 & & \\ & & -1 & 2 & \ddots & \\ & & & \ddots & \ddots & -1 \\ & & & & -1 & 2 \end{bmatrix}.$$

Now use the first row to eliminate the rest of the entries in the first column, and we can use the fourth elementary operation to remove the first row and column. The result is a matrix of the same form, one dimension smaller, with k increased by 1:

$$\begin{bmatrix} k{+}1 & -1 & & & \\ -k & 2 & -1 & & \\ & -1 & 2 & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{bmatrix}.$$

Starting with F_m, we iterate this process $m-2$ times and the result is

$$\begin{bmatrix} m & -1 \\ -(m{-}1) & 2 \end{bmatrix}. \qquad \square$$

We will apply Lemma 4.1 to reduce the blocks B_i in the Laplacian, which leads to the following significant reduction in size. We initially deal with the case where there are three or more longer paths, and address the cases with fewer long paths separately.

Theorem 4.2. *The reduced Laplacian of an undirected graph made of the parallel composition of h paths of length* 1 *and n paths of the varying lengths $f_1, \dots, f_n$, where $f_i \geq 2$ and $n \geq 3$, is $\mathbb{Z}$-equivalent to the form*

$$\begin{bmatrix} f_3 & & & -f_1 \\ & \ddots & & \vdots \\ & & f_n & -f_1 \\ f_2 & \cdots & f_2 & f_1{+}f_2{+}h(f_1 f_2) \end{bmatrix}.$$

Proof. Recall that the reduced Laplacian of graphs in this family has the form

$$\left[\begin{array}{c|c|c|c} B_1 & & & C_1^\intercal \\ \hline & \ddots & & \vdots \\ \hline & & B_n & C_n^\intercal \\ \hline C_1 & \cdots & C_n & n{+}h \end{array}\right].$$

We begin by treating paths of length 3 or greater separately from paths of length 2. Sorting the paths by length from shortest to longest gives the matrix

$$\left[\begin{array}{c|c|c|c|c|c|c}
2 & & & & & & -1 \\ \hline
 & \ddots & & & & & \vdots \\ \hline
 & & 2 & & & & -1 \\ \hline
 & & & B_m & & & C_m^{\mathsf{T}} \\ \hline
 & & & & \ddots & & \vdots \\ \hline
 & & & & & B_n & C_n^{\mathsf{T}} \\ \hline
-1 & \cdots & -1 & C_m & \cdots & C_n & n+h
\end{array}\right],$$

where f_m through f_n are all greater than 2.

We now utilize Lemma 4.1 to condense all paths with length greater than 2 (paths f_m through f_n). We can do this because none of the row and column operations used in Lemma 4.1 involve the last row or column of the block. The entries in the corresponding rows and columns outside of the block B_i are all zero and unaffected by the operations in Lemma 4.1. We now see that the reduced Laplacian is $\mathbb{Z}$-equivalent to the matrix

$$\left[\begin{array}{c|c|c|cc|c|cc|c}
-2 & & & & & & & & -1 \\ \hline
 & \ddots & & & & & & & \vdots \\ \hline
 & & -2 & & & & & & -1 \\ \hline
 & & & -f_m-1 & -1 & & & & 0 \\
 & & & -(f_m-2) & 2 & & & & -1 \\ \hline
 & & & & & \ddots & & & \vdots \\ \hline
 & & & & & & -f_n-1 & -1 & 0 \\
 & & & & & & -(f_n-2) & 2 & -1 \\ \hline
-1 & \cdots & -1 & -0 & -1 & \cdots & 0 & -1 & n+h
\end{array}\right].$$

We can further condense each 2×2 block along the diagonal using row and column operations similar to those used to prove Lemma 4.1. Corresponding to each 2×2 block we add $(f_i - 2)$ times the second column to the first column. Next, we add the first column to the second column. Every corresponding diagonal block now has a top row with 1 as the first entry and zero everywhere else. We can utilize this to eliminate all other values of each section's first column. The first row and column of each section are now in the form of the standard basis vector and can therefore be eliminated. The resulting matrix is of the form

$$\begin{bmatrix} f_1 & & & -1 \\ & \ddots & & \vdots \\ & & f_n & -1 \\ -(f_1-1) & \cdots & -(f_n-1) & n+h \end{bmatrix}.$$

The paths of length 2 also fit into this form since such paths are represented by a diagonal value of 2 and -1 as the bottom entry. Next, we add $(f_1 - 1)$ times the rightmost column to the leftmost column to put the matrix in the form

$$\begin{bmatrix} 1 & & & & -1 \\ -(f_1-1) & f_2 & & & -1 \\ \vdots & & \ddots & & \vdots \\ -(f_1-1) & & & f_n & -1 \\ (n+h-1)(f_1-1) & -(f_2-1) & \cdots & -(f_n-1) & n+h \end{bmatrix}.$$

From here we can add the leftmost column to the rightmost column. As a result, the top row has a first entry of 1 and zeros everywhere else. Therefore, all the rest of the entries of the leftmost column can be made into zero using row operations. Note the first row and column form the standard basis vector and can therefore be eliminated. The matrix is now of the form

$$\begin{bmatrix} f_2 & & & -f_1 \\ & \ddots & & \vdots \\ & & f_n & -f_1 \\ -(f_2-1) & \cdots & -(f_n-1) & (n+h)f_1-f_1+1 \end{bmatrix}.$$

Adding the bottom row to the top row produces the matrix

$$\begin{bmatrix} 1 & -(f_3-1) & \cdots & -(f_n-1) & (n+h)f_1-2f_1+1 \\ & f_3 & & & -f_1 \\ & & \ddots & & \vdots \\ & & & f_n & -f_1 \\ -(f_2-1) & -(f_3-1) & \cdots & -(f_n-1) & (n+h)f_1-f_1+1 \end{bmatrix}.$$

We can now add $(f_2 - 1)$ times the top row to the bottom row. The first column will have 1 as its first entry and zeros everywhere else. Utilizing this, we can make the top row into the form of the standard basis vector. This allows us to once again eliminate the top row and first column. The matrix is now in the form

$$\begin{bmatrix} f_3 & & & -f_1 \\ & \ddots & & \vdots \\ & & f_n & -f_1 \\ -f_2f_3+f_2 & \cdots & -f_2f_n+f_2 & (n+h-2)(f_1f_2)+f_1+f_2 \end{bmatrix}.$$

Notice that this is an $(n-1)\times(n-1)$ matrix. We can add the first $n-2$ rows multiplied by f_2 to the bottom row, which results in

$$\begin{bmatrix} f_3 & & & -f_1 \\ & \ddots & & \vdots \\ & & f_n & -f_1 \\ f_2 & \cdots & f_2 & f_1+f_2+h(f_1f_2) \end{bmatrix}. \qquad \square$$

Theorem 4.2 is useful in computing the sandpile group of specific graphs, as it dramatically reduces the size of the Laplacian used to compute the group. This theorem can also be used to classify the sandpile group of a family of graphs, as the following corollaries show.

Corollary 4.3. *The sandpile group of the parallel composition of n paths, where $n \geq 3$, with length a, where $a \geq 2$, is of the form*

$$\left(\bigoplus_{i=1}^{n-2} \mathbb{Z}_a\right) \oplus \mathbb{Z}_{an}.$$

Proof. Since all paths are of length a and $a \geq 2$ we know by Theorem 4.2 that the reduced Laplacian is $\mathbb{Z}$-equivalent to the $(n-1)\times(n-1)$ matrix

$$\begin{bmatrix} a & & & -a \\ & \ddots & & \vdots \\ & & a & -a \\ a & \cdots & a & 2a \end{bmatrix}.$$

Adding the first $n-2$ columns to the rightmost column, and then adding -1 times the first $n-2$ rows to the bottom row produces

$$\begin{bmatrix} a & & & \\ & \ddots & & \\ & & a & \\ & & & an \end{bmatrix}.$$

This presents the group

$$\left(\bigoplus_{i=1}^{n-2} \mathbb{Z}_a\right) \oplus \mathbb{Z}_{an}. \qquad \square$$

Corollary 4.4. *The sandpile group of the parallel composition of $n-1$ paths, where $n \geq 3$, with length a and one path of length b, where $a, b \geq 2$, is of the form*

$$\left(\bigoplus_{i=1}^{n-3} \mathbb{Z}_a\right) \oplus \mathbb{Z}_{\gcd(a,b)} \oplus \mathbb{Z}_{(a^2+(n-1)ab)/\gcd(a,b)}.$$

Proof. Since $a, b \geq 2$ we know by Theorem 4.2 that the reduced Laplacian is $\mathbb{Z}$-equivalent to the $(n-1) \times (n-1)$ matrix

$$\begin{bmatrix} a & & & & -a \\ & \ddots & & & \vdots \\ & & a & & -a \\ & & & b & -a \\ a & \cdots & a & a & 2a \end{bmatrix}.$$

Adding the first $n-3$ columns to the rightmost column, and then adding -1 times the first $n-3$ rows to the bottom row produces

$$\begin{bmatrix} a & & & & \\ & \ddots & & & \\ & & a & & \\ & & & b & -a \\ & & & a & a(n-1) \end{bmatrix}.$$

We can use row operations to perform the extended Euclidean algorithm on the b and a entries of the last two rows, replacing them with $\gcd(a, b)$ and zero. To see how this works, consider the case where $a < b$. The first step of the Euclidean algorithm asks us to divide b by a,

$$b = aq_1 + r_1,$$

and replace b with the remainder r_1. We achieve this in the matrix by multiplying the last row by $-q_1$ and adding to the penultimate row. The second step of the Euclidean algorithm asks us to divide a by the remainder r_1,

$$a = r_1 q_2 + r_2,$$

and replace a with the remainder r_2. We multiply the penultimate row by $-q_2$ and add to the last row. This process continues until one remainder evenly divides the previous. At this point the last nonzero remainder is $\gcd(a, b)$, and the other entry is zero. If needed we can switch the last two rows to ensure $\gcd(a, b)$ is on the diagonal. The entries in the last column started as multiples of a, and are now linear combinations of multiples of a. We write them as $a \cdot j$ and $a \cdot k$, where j and k are some integers. This yields a matrix of the form

$$\begin{bmatrix} a & & & & \\ & \ddots & & & \\ & & a & & \\ & & & \gcd(a,b) & a\cdot j \\ & & & 0 & a\cdot k \end{bmatrix}.$$

Of course $\gcd(a,b)$ divides any multiple of a; therefore we can simplify this expression to

$$\begin{bmatrix} a & & & & \\ & \ddots & & & \\ & & a & & \\ & & & \gcd(a,b) & \\ & & & & a\cdot k \end{bmatrix}.$$

This presents the group

$$\left(\bigoplus_{i=1}^{n-3} \mathbb{Z}_a\right) \oplus \mathbb{Z}_{\gcd(a,b)} \oplus \mathbb{Z}_{ak}.$$

Recall from Section 1 that the order of a sandpile group is the number of spanning trees on its graph [Dhar et al. 1995]. The number of spanning trees in this family of graphs can be found by counting the number of ways to remove an edge from all paths except one. If we leave the path of length b intact, we would need to remove an edge from the $n-1$ paths of length a. There are a^{n-1} ways of doing this. Similarly, we can choose a path of length a to keep intact and remove an edge from the rest of the paths. There are $a^{n-2}b(n-1)$ ways to do this, making the total number of spanning trees $a^{n-1}+a^{n-2}b(n-1)$. Another way of finding the order of the group is to take the product of the components. For this group this is $a^{n-3}\times\gcd(a,b)\times ak$. Setting these values equal to each other and solving for k yields

$$k=\frac{a+b(n-1)}{\gcd(a,b)}.$$

Therefore, the group is

$$\left(\bigoplus_{i=1}^{n-3} \mathbb{Z}_a\right) \oplus \mathbb{Z}_{\gcd(a,b)} \oplus \mathbb{Z}_{(a^2+ab(n-1))/\gcd(a,b)}. \qquad \square$$

We now offer the sandpile groups of undirected graphs made of the parallel composition of h paths of length 1 and n paths, where $n=0,1,2$, of the varying lengths $f_1,\ldots,f_n$, where $f_i\geq 2$.

Theorem 4.5. *The sandpile group of the parallel composition of h paths with length* 1 *is of the form*

$$\mathbb{Z}_h.$$

Proof. The Laplacian of this family of graphs is of the form

$$\begin{bmatrix} h & -h \\ -h & h \end{bmatrix}.$$

The reduced Laplacian is a 1×1 matrix with h as the only entry, representing the group

$$\mathbb{Z}_h. \qquad \square$$

Theorem 4.6. *The sandpile group of the parallel composition of h paths with length 1 and one path with length f_1, where $f_1 \geq 2$, is of the form*

$$\mathbb{Z}_{hf_1+1}.$$

Proof. We will proceed by using a method similar to the one used to prove Lemma 4.1. The Laplacian of this family of graphs has the form

$$\left[\begin{array}{c|c|c} h+1 & D_1 & -h \\ \hline D_1^{\mathsf{T}} & B_1 & C_1^{\mathsf{T}} \\ \hline -h & C_1 & h+1 \end{array}\right].$$

Since the graphs in this family are Eulerian, the choice of sink does not affect the sandpile group. For simplicity's sake, we let the first row and column correspond to the sink. The resulting reduced Laplacian will be of the form

$$\left[\begin{array}{c|c} B_1 & C_1^{\mathsf{T}} \\ \hline C_1 & h+1 \end{array}\right].$$

Lemma 4.1 does not require adding the last column (or row) to any other column (or row). Therefore, we can utilize Lemma 4.1 to show the reduced Laplacian is $\mathbb{Z}$-equivalent to the matrix

$$\begin{bmatrix} f_1 & -1 \\ -(f_1-1) & h+1 \end{bmatrix}.$$

We now add $(f_1 - 1)$ times the right column to the left column. Then we add the left column to the right column. As a result, the top row will have 1 as the first entry and zeros everywhere else. We use the top row to eliminate all other entries in the first column. In matrix form, these steps are

$$\begin{bmatrix} 1 & -1 \\ h(f_1-1) & h+1 \end{bmatrix} \equiv \begin{bmatrix} 1 & 0 \\ h(f_1-1) & hf_1+1 \end{bmatrix} \equiv \begin{bmatrix} 1 & 0 \\ 0 & hf_1+1 \end{bmatrix}.$$

Consequently, the group presented is

$$\mathbb{Z}_{hf_1+1}. \qquad \square$$

Theorem 4.7. *The sandpile group of the parallel composition of h paths with length* 1 *and two paths with respective lengths f_1 and f_2, where $f_1, f_2 \geq 2$, is of the form*

$$\mathbb{Z}_{f_1+f_2+h(f_1 f_2)}.$$

Proof. The Laplacian of this family of graphs is of the form

$$\left[\begin{array}{c|c|c|c} h+2 & D_1 & D_2 & -h \\ \hline D_1^{\mathsf{T}} & B_1 & & C_1^{\mathsf{T}} \\ \hline D_2^{\mathsf{T}} & & B_2 & C_2^{\mathsf{T}} \\ \hline -h & C_1 & C_2 & h+2 \end{array}\right].$$

Graphs in this family are Eulerian; thus we can let the first row and column represent the sink. The resulting reduced Laplacian will be of the form

$$\left[\begin{array}{c|c|c} B_1 & & C_1^{\mathsf{T}} \\ \hline & B_2 & C_2^{\mathsf{T}} \\ \hline C_1 & C_2 & h+2 \end{array}\right].$$

We utilize Lemma 4.1 and additional steps seen in the proof of Theorem 4.2 to condense B_1 and B_2, producing the matrix

$$\begin{bmatrix} f_1 & 0 & -1 \\ 0 & f_2 & -1 \\ -(f_1-1) & -(f_2-1) & h+2 \end{bmatrix}.$$

We add $(f_1 - 1)$ times the right column to the left column. Then we add the left column to the right column and use the top row to eliminate all other entries in the first column:

$$\begin{bmatrix} 1 & 0 & -1 \\ -(f_1-1) & f_2 & -1 \\ h(f_1-1)+(f_1-1) & -(f_2-1) & h+2 \end{bmatrix} \equiv \begin{bmatrix} 1 & 0 & 0 \\ 0 & f_2 & -f_1 \\ 0 & -(f_2-1) & hf_1+f_1+1 \end{bmatrix}.$$

Observe, the top row and left column form the standard basis vector and can therefore be eliminated using the fourth elementary operation:

$$\begin{bmatrix} f_2 & -f_1 \\ -(f_2-1) & hf_1+f_1+1 \end{bmatrix}.$$

Adding the bottom row to the top will produce the matrix

$$\begin{bmatrix} 1 & hf_1+1 \\ -(f_2-1) & hf_1+f_1+1 \end{bmatrix}.$$

We now add $(f_2 - 1)$ times the top row to the bottom row. As a result, the left column has 1 as the first entry and 0 everywhere else. We use this to put the top row and column into the form of the standard basis vector:

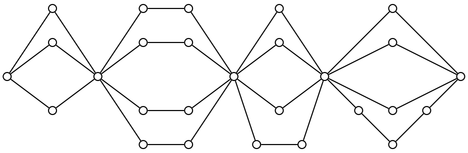
$$\begin{bmatrix}1 & 0\\ 0 & f_1+f_2+h(f_1f_2)\end{bmatrix}$$

Therefore, the group is

$$\mathbb{Z}_{f_1+f_2+h(f_1f_2)}. \qquad \square$$

5. Conclusion

The techniques of the preceding sections allow us to easily find the sandpile group of a wide range of graphs. For example, consider the following graph:



This graph is the series composition of four parallel graphs. We apply Corollary 4.3 to the first parallel graph from the left. This graph is the parallel composition of three paths of length 2, so it has a sandpile group of

$$\left(\bigoplus_{i=1}^{3-2}\mathbb{Z}_2\right)\oplus\mathbb{Z}_{2\cdot 3}=\mathbb{Z}_2\oplus\mathbb{Z}_6.$$

Observe that the second graph is the parallel composition of four paths of length 3. Therefore, applying Corollary 4.3 yields

$$\left(\bigoplus_{i=1}^{4-2}\mathbb{Z}_3\right)\oplus\mathbb{Z}_{3\cdot 4}=\mathbb{Z}_3\oplus\mathbb{Z}_3\oplus\mathbb{Z}_{12}.$$

The third graph is the parallel composition of four paths, three of length 2 and a single path of length 3. Therefore, applying Corollary 4.4 yields

$$\left(\bigoplus_{i=1}^{4-3}\mathbb{Z}_2\right)\oplus\mathbb{Z}_{\gcd(2,3)}\oplus\mathbb{Z}_{(2^2+2\cdot 3(4-1))/\gcd(2,3)}=\mathbb{Z}_2\oplus\mathbb{Z}_{22}.$$

The fourth graph is the parallel composition of four paths, three of length 2 and a single path of length 4. Therefore, applying Corollary 4.4 yields

$$\left(\bigoplus_{i=1}^{4-3}\mathbb{Z}_2\right)\oplus\mathbb{Z}_{\gcd(2,4)}\oplus\mathbb{Z}_{(2^2+2\cdot 4(4-1))/\gcd(2,4)}=\mathbb{Z}_2\oplus\mathbb{Z}_2\oplus\mathbb{Z}_{14}.$$

The entire graph is formed by using series operations to combine all four parallel graphs. By Theorem 3.1 the resulting sandpile group will be the direct sum of the sandpile groups of all four parallel graphs. Consequently, the sandpile group is

$$\mathbb{Z}_2\oplus\mathbb{Z}_6\oplus\mathbb{Z}_3\oplus\mathbb{Z}_3\oplus\mathbb{Z}_{12}\oplus\mathbb{Z}_2\oplus\mathbb{Z}_{22}\oplus\mathbb{Z}_2\oplus\mathbb{Z}_2\oplus\mathbb{Z}_{14}.$$

This example shows how the tools presented in this paper allow us to easily compute the sandpile groups of rather complicated graphs.

We have presented the completely general result for the series composition of graphs, namely that the sandpile group is the direct sum of the component sandpile groups. We initially sought a similar result for parallel composition of graphs. While a completely general result about parallel composition eluded us, we were able to prove results about the parallel composition of paths, including a significant reduction in the size of the matrix needed to compute the sandpile groups for such graphs. This work has been fascinating, and we hope to see more general results involving the sandpile groups of the parallel composition of graphs in the future.

References

[Bak et al. 1988] P. Bak, C. Tang, and K. Wiesenfeld, "Self-organized criticality", *Phys. Rev. A* (3) **38**:1 (1988), 364–374. MR Zbl

[Cori and Rossin 2000] R. Cori and D. Rossin, "On the sandpile group of dual graphs", *European J. Combin.* **21**:4 (2000), 447–459. MR Zbl

[Corry and Perkinson 2018] S. Corry and D. Perkinson, *Divisors and sandpiles: an introduction to chip-firing*, American Mathematical Society, Providence, RI, 2018. MR Zbl

[Dhar et al. 1995] D. Dhar, P. Ruelle, S. Sen, and D.-N. Verma, "Algebraic aspects of abelian sandpile models", *J. Phys. A* **28**:4 (1995), 805–831. MR Zbl

[Dummit and Foote 2004] D. S. Dummit and R. M. Foote, *Abstract algebra*, 3rd ed., John Wiley & Sons, Hoboken, NJ, 2004. MR Zbl

[Hou et al. 2006] Y. Hou, C. Woo, and P. Chen, "On the sandpile group of the square cycle C_n^2", *Linear Algebra Appl.* **418**:2-3 (2006), 457–467. MR Zbl

[Hou et al. 2008] Y. Hou, T. Lei, and C. Woo, "On the sandpile group of the graph $K_3 \times C_n$", *Linear Algebra Appl.* **428**:8-9 (2008), 1886–1898. MR Zbl

[Jacobson et al. 2003] B. Jacobson, A. Niedermaier, and V. Reiner, "Critical groups for complete multipartite graphs and Cartesian products of complete graphs", *J. Graph Theory* **44**:3 (2003), 231–250. MR Zbl

[Levine 2009] L. Levine, "The sandpile group of a tree", *European J. Combin.* **30**:4 (2009), 1026–1035. MR Zbl

[Liang et al. 2008] H. Liang, Y.-L. Pan, and J. Wang, "The critical group of $K_m \times P_n$", *Linear Algebra Appl.* **428**:11-12 (2008), 2723–2729. MR Zbl

[Shen and Hou 2008] J. Shen and Y. Hou, "On the sandpile group of $3 \times n$ twisted bracelets", *Linear Algebra Appl.* **429**:8-9 (2008), 1894–1904. MR Zbl

[Toumpakari 2007] E. Toumpakari, "On the sandpile group of regular trees", *European J. Combin.* **28**:3 (2007), 822–842. MR Zbl

Received: 2019-04-29 Revised: 2020-04-27 Accepted: 2020-04-28

kweishaar@regis.edu
Department of Mathematics, Regis University, Denver, CO, United States

jseibert@regis.edu
Department of Mathematics, Regis University, Denver, CO, United States

INVOLVE 13:3 (2020)
https://doi.org/10.2140/involve.2020.13.399

Linkages of calcium-induced calcium release in a cardiomyocyte simulated by a system of seven coupled partial differential equations

Gerson C. Kroiz, Carlos Barajas,
Matthias K. Gobbert and Bradford E. Peercy

(Communicated by Suzanne Lenhart)

Cardiac arrhythmias affect millions of adults in the U.S. each year. This irregularity in the beating of the heart is often caused by dysregulation of calcium in cardiomyocytes, the cardiac muscle cell. Cardiomyocytes function through the interplay between electrical excitation, calcium signaling, and mechanical contraction, an overall process known as calcium-induced calcium release (CICR). A system of seven coupled nonlinear time-dependent partial differential equations (PDEs), which model physiological variables in a cardiac cell, link the processes of cardiomyocytes. Through parameter studies for each component system at a time, we create a set of values for critical parameters that connect the calcium store in the sarcoplasmic reticulum, the effect of electrical excitation, and mechanical contraction in a physiologically reasonable manner. This paper shows the design process of this set of parameters and then shows the possibility to study the influence of a particular problem parameter using the overall model.

1. Introduction

The leading cause of death in the United States is currently heart disease [Mozaffarian et al. 2015]. In order to continue searching for methods to combat heart disease, it is vital that the heart and its underlying processes are understood with greater depth. The importance of having a greater understanding of the heart provides the motivation for this research.

This project focuses on a single cardiac cell and uses a mathematical model in order to represent the electrical-excitation, calcium-signaling, and mechanical-contraction components of a cardiomyocyte. The study of a single cardiac cell

MSC2010: primary 35K57, 65M08, 65Y05, 68U20, 92C30; secondary 65L04, 65M20, 68W10, 92C37, 92C50.

Keywords: cardiac arrhythmia, calcium-induced calcium release, reaction diffusion equations, finite-volume method, parallel computing.

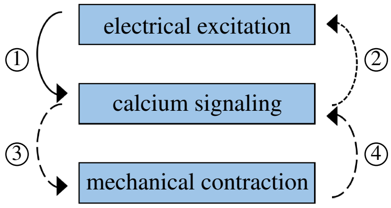


Figure 1. The three components of the model and their links labeled ① to ④.

is important as it is the basic building block of cardiac behavior. The original model for calcium-induced calcium release (CICR) was introduced in [Izu et al. 2001a; 2001b] with a three-variable model and included only calcium signaling. This original model makes up the heart of the calcium-signaling component of the system indicated in Figure 1. Figure 2 sketches the domain of the simulation. We use a hexahedron elongated in the z-direction for the shape of a cell, since the focus of CICR research is on estimating reasonable physiological parameter values. A key of the model in [Izu et al. 2001a; 2001b] is to place the calcium release units (CRUs) on a regular lattice of size, for instance, $15 \times 15 \times 31 = 6{,}975$ throughout the cell; Figure 2 shows three z-planes of 3×3 CRUs as an example. At these many locations, calcium ions are released from the calcium store in the sarcoplasmic reticulum (SR) into the cytosol of the cell, and CRUs can open ("fire") and close repeatedly over time [Izu et al. 2001a; 2001b].

The model was extended for the first time to include the electrical-excitation component in Figure 1 [Alexander et al. 2015], which implemented a one-way interaction from electrical excitation to calcium signaling indicated by link ① in Figure 1. Studies with six variables in [Angeloff et al. 2016] extended the coupling to include a two-way cycle between electrical excitation and calcium signaling by incorporating both links ① and ② in Figure 1. The formulation of the complete eight-variable model for all components in Figure 1 was introduced in [Angeloff et al. 2016], but not all model variables were used in the simulations, and the studies did not incorporate the mechanical system.

The most recent work in [Deetz et al. 2017a; 2017b] studies the introduction of the mechanical-contraction component in Figure 1 by activating the links ③ and ④ in Figure 1. This is facilitated by adding a buffer species whose concentration can be related to the contraction of the cell. Thus, this work uses a model with seven variables to represent the excitation-contraction coupling (ECC) occurring in the cardiomyocyte, in which CICR is the mechanism through which electrical excitation is coupled with mechanical contraction through calcium signaling. The initial results for the newly extended model produce simulations promising in their predictive capability.

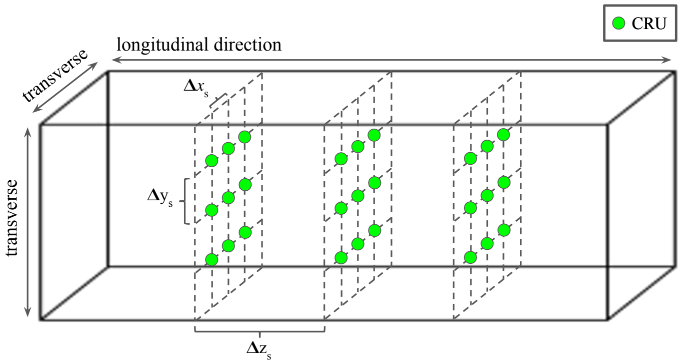


Figure 2. The CRU lattice with spacings Δx_s, Δy_s, Δz_s throughout the three-dimensional cell.

From the previous work, the linkage of the mechanical contraction was existent, but not physiologically correct with the previous set of parameters. The resulting simulations may not have provided realistic behaviors of a cardiac muscle cell. As such, it is important to create a set of parameters with physiologically reasonable simulations. To do this, we reduce the seven-variable model to the three-variable model from [Izu et al. 2001a; 2001b] by setting key parameters to zero. This removes several CICR components: the calcium store in the SR, electrical excitation, and mechanical contraction. By this process, the additional CICR components are negated as they all reduce to zero. From this base case, we reintroduce physiologically reasonable values for these parameters in successive order. First, the corresponding parameter for the calcium store is implemented, reconnecting this component to the system. The selected parameters which negated electrical excitation are then introduced. To complete the system, the key variable connecting mechanical contraction to the rest of the model is included. By creating this set of parameters that includes the CICR components, we are eventually able to study the model as a whole rather than individual components. This paper shows the design process of this set of parameters and then studies the influence of mechanical contraction on the rest of the model for simulations to an extended final time. This shows the possibility of studying the influence of a particular problem parameter using the overall model.

This paper is organized as follows: Section 2 details the seven-variable model used in this work. Section 3 specifies the numerical method used, including its parallelization and implementation. Section 4 is divided into several subsections. Sections 4.1–4.4 show the step-by-step implementation of the CICR components of the seven-variable model. With the complete construction of the set of parameters, Section 4.5 presents a sample application using the new set of parameters. Section 5 collects the conclusions of the work.

2. Dynamics of a cardiac cell

This section details the seven-variable model used in this work. The seven variables of the model are calcium in the cytosol $c(\boldsymbol{x}, t)$, a fluorescent dye $b_1^{(c)}(\boldsymbol{x}, t)$, a contractile protein (troponin) $b_2^{(c)}(\boldsymbol{x}, t)$, a contractile force $b_3^{(c)}(\boldsymbol{x}, t)$, calcium in the sarcoplasmic reticulum (SR) $s(\boldsymbol{x}, t)$, voltage $V(\boldsymbol{x}, t)$, and the potassium gating function $n(\boldsymbol{x}, t)$. These are listed in Table 1 with their units and initial values. The evolution of the system is modeled by the system of seven time-dependent, coupled, nonlinear reaction diffusion equations

$$\begin{aligned}
\frac{\partial c}{\partial t} &= \nabla \cdot (D_c \nabla c) + R_{b_1}^{(c)} + R_{b_2}^{(c)} \\
&\qquad + (J_{\mathrm{CRU}} + J_{\mathrm{leak}} - J_{\mathrm{pump}}) + \kappa J_{\mathrm{LCC}} + (J_{m_{\mathrm{leak}}} - J_{m_{\mathrm{pump}}}), && (1)\\
\frac{\partial b_1^{(c)}}{\partial t} &= \nabla \cdot (D_{b_1^{(c)}} \nabla b_1^{(c)}) + R_{b_1}^{(c)}, && (2)\\
\frac{\partial b_2^{(c)}}{\partial t} &= \nabla \cdot (D_{b_2^{(c)}} \nabla b_2^{(c)}) + R_{b_2}^{(c)}, && (3)\\
\frac{\partial b_3^{(c)}}{\partial t} &= \nabla \cdot (D_{b_3^{(c)}} \nabla b_3^{(c)}) + R_{b_3}^{(c)}, && (4)\\
\frac{\partial s}{\partial t} &= \nabla \cdot (D_s \nabla s) - \gamma (J_{\mathrm{CRU}} + J_{\mathrm{leak}} - J_{\mathrm{pump}}), && (5)\\
\frac{\partial V}{\partial t} &= \nabla \cdot (D_v \nabla V) + \tau_v \frac{1}{C} \bigg[I_{\mathrm{app}} - g_L (V - V_L) - g_K n (V - V_K) \\
&\qquad - \frac{2F}{S \tau_{\mathrm{flux}}} J_{\mathrm{LCC}} - \omega (J_{m_{\mathrm{leak}}} - J_{m_{\mathrm{pump}}}) \bigg], && (6)\\
\frac{\partial n}{\partial t} &= \nabla \cdot (D_n \nabla n) + \tau_v \lambda_n \cosh\left(\frac{V - V_3}{2V_4} \right) (n_\infty(V) - n). && (7)
\end{aligned}$$

The domain in our model is a hexahedron

$$\Omega = (-6.4, 6.4) \times (-6.4, 6.4) \times (-32.0, 32.0) \qquad (8)$$

with units μm with isotropic CRU distribution as sketched in Figure 2 with a sample of three z-planes of 3×3 CRUs. A typical cell has on the order of $15 \times 15 \times 31 = 6{,}975$ calcium release units (CRUs) throughout the cell. A cell is reasonably modeled by the hexahedral shape elongated in the z-direction, since the focus of CICR research is on estimating correct physiological parameter values. The model uses no-flux boundary conditions for all dependent variables, so that no species escapes or enters the cell through the domain boundary $\partial\Omega$; this corresponds to the goal of modeling the behavior of a single cell. To finish the well-posed statement of an initial-boundary value problem of parabolic type, such as the seven

variable	definition	initial value	
$\boldsymbol{x}$	spatial position variable (x, y, z)	μm	
t	time variable	ms	
$c(\boldsymbol{x}, t)$	calcium in the cytosol	$c_0 = 0.1\,\mu$M	A
$b_1^{(c)}(\boldsymbol{x}, t)$	free fluorescent dye in the cytosol	$45.918\,\mu$M	A
$b_2^{(c)}(\boldsymbol{x}, t)$	free troponin in the cytosol	$111.818\,\mu$M	A
$b_3^{(c)}(\boldsymbol{x}, t)$	inactive actin-myosin cross-bridges	$145.20\,\mu$M	C
$s(\boldsymbol{x}, t)$	calcium in the SR	$s_0 = 10{,}000\,\mu$M	A
$V(\boldsymbol{x}, t)$	membrane potential (voltage)	$-50\,$mV	B
$n(\boldsymbol{x}, t)$	fraction of open potassium channels	0.1	B

Table 1. Independent and dependent variables of the seven-variable model and their initial conditions. The concentration unit M is shorthand for mol/L (moles per liter). The last column indicates which system component in Figure 1 contains each dependent variable: A = calcium signaling, B = electrical excitation, C = mechanical contraction.

equations (1)–(7), values for all dependent variables need to be specified at the initial time $t = 0$, which is done by the values in Table 1.

The following text will detail the mathematical model for each of the three system components in Figure 1 containing the variables in Table 1: The calcium-signaling component contains variables c, $b_1^{(c)}$, $b_2^{(c)}$, and s. The electrical-excitation component contains V and n and is connected to the calcium signaling in both the feedforward and feedback directions, represented by links ① and ②. The mechanical-contraction component contains $b_3^{(c)}$ and is also connected to the calcium signaling in both the feedback and feedforward directions represented by links ③ and ④.

The initial value $c_0 = 0.1\,\mu$M is the basal level of calcium in the cytosol, which is a physiologically measured value for the system at rest. The initial values for all buffer species $b_1^{(c)}$, $b_2^{(c)}$, and $b_3^{(c)}$ are calculated from their reaction rates such that no reactions take place for cytosol calcium $c = c_0$ at rest. The initial value $s_0 = 10{,}000\,\mu$M for calcium in the SR is selected very high, so that the amount of calcium in the store is not a limiting factor for the release of calcium at first; this is a common technique to promote nontrivial behavior of the model in simulations. The initial values for membrane potential V and the fraction of open potassium channels n represent the resting state of the electrical system of the cell.

Tables 2 and 3 contain the parameters in the PDEs of the calcium system with their values (if fixed) and units; some of the values are varied in the experiments in Section 4. The coefficients D_c, $D_{b_i^{(c)}}$, and D_s are the 3×3 diffusivity matrices for

Ca^{2+} in the cytosol, buffer species i in the cytosol, and Ca^{2+} in the SR, respectively; these matrices are diagonal, as indicated in Table 2, with potentially different values in the x-, y-, z-directions, or potentially zero, if the species does not diffuse.

variable	definition	values/units
D_c	diffusivity matrix for $c(\boldsymbol{x}, t)$	diag(0.15,0.15,0.3) μm^2/ms
$D_{b_1^{(c)}}$	diffusivity matrix for $b_1^{(c)}$	diag(0.01,0.01,0.02) μm^2/ms
$D_{b_2^{(c)}}$	diffusivity matrix for $b_2^{(c)}$	diag(0.00,0.00,0.00) μm^2/ms
$D_{b_3^{(c)}}$	diffusivity matrix for $b_3^{(c)}$	diag(0.00,0.00,0.00) μm^2/ms
D_s	diffusivity matrix for $s(\boldsymbol{x}, t)$	diag(0.78,0.78,0.78) μm^2/ms
$R_{b_i}^{(c)}$	reactions of cytosol Ca^{2+} with buffers	μM/ms
$k^+_{b_1^{(c)}}$	forward reaction coefficient for $b_1^{(c)}$	0.080 (μM ms)$^{-1}$
$k^+_{b_2^{(c)}}$	forward reaction coefficient for $b_2^{(c)}$	0.100 (μM ms)$^{-1}$
$k^+_{b_3^{(c)}}$	forward reaction coefficient for $b_3^{(c)}$	0.040 ms^{-1}
$k^-_{b_1^{(c)}}$	reverse reaction coefficient for $b_1^{(c)}$	0.090 ms^{-1}
$k^-_{b_2^{(c)}}$	reverse reaction coefficient for $b_2^{(c)}$	0.100 ms^{-1}
$k^-_{b_3^{(c)}}$	reverse reaction coefficient for $b_3^{(c)}$	0.010 ms^{-1}
γ	ratio of volume of cytosol to SR	14
$b^{(c)}_{1,\text{total}}$	total amount of $b_1^{(c)}$ in the cytosol	50 μM
$b^{(c)}_{2,\text{total}}$	total amount of $b_2^{(c)}$ in the cytosol	123 μM
$b^{(c)}_{3,\text{total}}$	total amount of $b_3^{(c)}$ in the cytosol	150 μM
J_{leak}	calcium leak from SR	0.3209684 μM/ms
J_{pump}	calcium transfer from cytosol to SR	μM/ms
V_{pump}	maximum pump rate	4 μM/ms
K_{pump}	pump sensitivity to Ca^{2+}	0.184 μM
n_{pump}	Hill coefficient for pump function	4.0
J_{CRU}	calcium flux from SR to cytosol via CRUs	μM/ms
$\mathcal{O}$	gating function for J_{CRU}	0 or 1
J_{prob}	probability of CRU opening	0 to 1
$\boldsymbol{x}_s$	three-dimensional vector for CRU location	μm
Δx_s	CRU spacings in x	0.8 μm
Δy_s	CRU spacings in y	0.8 μm
Δz_s	CRU spacings in z	2.0 μm
$\hat{\sigma}$	maximum rate of release	110 μM μm^3/ms
$\delta(\boldsymbol{x}-\hat{\boldsymbol{x}})$	Dirac delta distribution	1/μm^3
u_{rand}	uniformly distributed random variable	0 to 1
P_{max}	maximum probability for release	0.3
K_{prob_c}	sensitivity of CRU to cytosol calcium	2 μM
n_{prob_c}	Hill coefficient for probability function	4
K_{prob_s}	sensitivity of CRU to SR calcium	550 μM
n_{prob_s}	Hill coefficient for probability function	4

Table 2. Parameters for calcium signaling.

variable	definition	values/units
D_v	diffusivity matrix for $V(\boldsymbol{x}, t)$	diag(100,100,100) μm^2/ms
D_n	diffusivity matrix for $n(\boldsymbol{x}, t)$	diag(0,0,0) μm^2/ms
τ_v	scaling factor in voltage equation	0.1 μM μm^3/ms
τ_{flux}	scaling factor in J_{LCC} term	0.1
V_1	potential at which $m_\infty = 0.5$	-1.0 mV
V_2	reciprocal of slope of voltage dependence of m_∞	15.0 mV
V_3	potential at which $n_\infty = 0.5$	10.0 mV
V_4	reciprocal of slope of voltage dependence of n_∞	14.5 mV
V_L	equilibrium potential for leak conductance	-50 mV
V_{Ca}	equilibrium potential for Ca^{2+} conductance	100 mV
V_K	equilibrium potential for K$^+$ conductance	-70 mV
C	membrane capacitance	20 μF/cm^2
I_{app}	applied current	50 μA/cm^2
g_L	max./instantaneous conductance for leak	2 mmho/cm^2
g_{Ca}	max./instantaneous conductance for Ca^{2+}	4 mmho/cm^2
g_K	max./instantaneous conductance for K$^+$	8 mmho/cm^2
m_∞	fraction of open calcium channels at steady state	0 to 1
n_∞	fraction of open potassium channels at steady state	1
λ_n	max. rate constant for opening of K$^+$ channels	0.1 ms^{-1}
J_{LCC}	influx of calcium via L-type calcium channels	μM/ms
S	surface area of the cell	3604.48 μm^2
F	Faraday constant	95484.56 C/mol
κ	scaling factor of J_{LCC}	0.1
ω	feedback strength (scaling factor) for Ca^{2+} efflux	100 μA ms/μM cm^2
$J_{m_{\text{leak}}}$	leak of calcium via L-type calcium channels	0.1739493 μM/ms
$J_{m_{\text{pump}}}$	pump of calcium via L-type calcium channels	μM/ms
$V_{m_{\text{pump}}}$	max. pump rate	0.3 μM/ms
$n_{m_{\text{pump}}}$	membrane pump Hill coefficient	4
$K_{m_{\text{pump}}}$	membrane pump Ca^{2+} sensitivity	0.18
$[XB]_0$	initial concentration of active cross-bridges	142.6805 μM
ε	shortening factor	0 to 1
$F_{\max}$	max. force by actin-myosin cross-bridges	1 μN
k_s	stiffness of actin filament	0.025 N/m

Table 3. Parameters for electrical excitation and mechanical contraction with base units of mho $= (\text{s}^3\text{A}^2)/(\text{kg m}^2)$ and F $= (\text{s}^4\text{A}^2)/(\text{kg m}^2)$.

The calcium signaling portion of the model consists of the equations (1)–(5). The reaction terms $R_{b_i}^{(c)}$ describe the reactions between calcium and the buffer species. They are the connections between (1)–(4). More precisely, we have

$$R_{b_1}^{(c)}(c, b_1^{(c)}) = -k_{b_1^{(c)}}^{+} c b_1^{(c)} + k_{b_1^{(c)}}^{-}(b_{1,\text{total}}^{(c)} - b_1^{(c)}), \tag{9}$$

for reaction between cytosol calcium and the fluorescent dye.

When troponin binds to Ca^{2+}, the protein as a whole changes shape: this not only allow actin-myosin cross-bridges to form, but also traps the calcium in its connection to the troponin so that the disassociation rate of troponin binding to Ca^{2+} decreases dramatically. To account for this, the shortening factor ε describes how the separation of troponin and calcium has been physically, but not chemically, impaired. When there are open CRUs that, as a result, release calcium in the cytosol, this chain of changes of variables increases in severity. However, after $t = 5$ ms of being open, the CRUs close and remain in a refractory state for $t = 100$ ms. During this period, the unbinding of troponin and calcium increase to the initial rate of unbinding and myosin cross-bridges become inactive. This results in these terms returning to their original states. This cycle is continuous triggered and reset by the calcium waves produced by a cardiomyocyte. Note, again, that $R_{b_2}^{(c)}$ remains a function of cytosol calcium concentration $c(\boldsymbol{x}, t)$ by its equation

$$R_{b_2}^{(c)}(c, b_2^{(c)}) = -k_{b_2^{(c)}}^{+} c b_2^{(c)} + \left(\frac{k_{b_2^{(c)}}^{-}}{\varepsilon} \right) (b_{2,\text{total}}^{(c)} - b_2^{(c)}), \tag{10}$$

with

$$\varepsilon = \exp\left(F_{\max} k_s \left(\frac{b_{3,\text{total}}^{(c)} - b_3^{(c)} - [XB]_0}{b_{3,\text{total}}^{(c)} - [XB]_0} \right) \right) \tag{11}$$

and $[XB]_0 = b_{3,\text{total}}^{(c)} - b_3^{(c)}(\boldsymbol{x}, 0)$. This shortening factor ε links ③ and ④ in Figure 1. It is important to note that the disassociation rate of troponin binding to Ca^{2+} is highest when the calcium concentration in the cytosol is lowest.

The shortening factor refers back to the concentration of $b_3^{(c)}(\boldsymbol{x}, t)$, the actin-myosin cross-bridges, and to the force that their linkage generates. It is scaled by the maximum possible contractile force $F_{\max}$, the actin stiffness k_s, and the proportion of active to inactive actin-myosin cross-bridges. As the proportion of active actin-myosin cross-bridges increases due to an increase in bounded troponin, the value of $b_{3,\text{total}}^{(c)} - b_3^{(c)} - [XB]_0$ increases. This results in a larger value of ε. As ε becomes larger, the rate at which the bounded troponin unbinds slows down. As described earlier, the lack of calcium provided by CRU resting periods results in a reset in this mechanism.

The contractile proteins in question, though considered as a single species, are the combination of actin and myosin when linked via cross-bridges. This linkage is made possible by Ca^{2+} binding to troponin, the cytosol buffer species $b_2^{(c)}(\boldsymbol{x}, t)$: it is this binding that allows the actin-myosin cross-bridges to form. The cytosol species, $b_3^{(c)}(\boldsymbol{x}, t)$, describes these actin-myosin cross-bridges and constructs a third cytosol reaction term

$$R_{b_3}^{(c)}(c, b_2^{(c)}, b_3^{(c)}) = -k_{b_3^{(c)}}^{+} \left(\frac{b_{2,\text{total}}^{(c)} - b_2^{(c)}}{b_{2,\text{total}}^{(c)}} \right)^2 b_3^{(c)} + k_{b_3^{(c)}}^{-} (b_{3,\text{total}}^{(c)} - b_3^{(c)}). \tag{12}$$

Notice that this is not the same as the generic pattern for buffer species reaction terms from the initial model. There is no immediately clear dependence on cytosolic calcium $c(\boldsymbol{x}, t)$. However, while $c(\boldsymbol{x}, t)$ is not explicitly included, it is present in the proportion involving troponin, $b_2^{(c)}(\boldsymbol{x}, t)$, which itself depends explicitly on cytosol calcium levels; $R_{b_3}^{(c)}$, like the other two reaction equations, does in fact depend on the cytosol calcium concentration. Unlike the other two reaction equations, $R_{b_3}^{(c)}$ indirectly impacts calcium levels.

These two reaction terms (10) and (12) connect the three components of our model. The calcium signaling is linked to the pseudomechanical contraction by the cross-bridge term, and the pseudo-mechanical contraction is in turn connected to the calcium signaling through the inclusion of the cytosol calcium concentration in the modified reaction equation for troponin. Thus links ①–④ in Figure 1 are established, and the three systems of the model are fully linked.

Note that in the reaction terms (9), (10), and (12), $b_1^{(c)}$, $b_2^{(c)}$, and $b_3^{(c)}$ are the amounts of *unbound* buffer known as "free" buffer. The constants $b_{i,\text{total}}^{(c)}$, $i = 1, 2, 3$, denote the *total bound and unbound* buffer, thus leaving the differences $b_{i,\text{total}}^{(c)} - b_i{}^{(c)}$ in the reaction terms to be the buffer *bound* with cytosol calcium c. Since the model uses no-flux boundary conditions, no buffer species escapes or enters the cell; thus we only need to track the "free" buffer species and use $b_{i,\text{total}}^{(c)} - b_i{}^{(c)}$ for the bound species. The values for the constants $b_{i,\text{total}}^{(c)}$, $i = 1, 2, 3$, are model parameters and specified in Table 2.

The flux terms J_{CRU}, J_{leak}, and J_{pump} in (1) describe the calcium-induced release of Ca^{2+} into the cytosol from the SR, the continuous leak of Ca^{2+} into the cytosol from the SR, and the pumping of Ca^{2+} back into the SR from the cytosol. The terms J_{LCC}, $J_{m_{\text{leak}}}$, and $J_{m_{\text{pump}}}$ describe the fluxes of calcium into and out of the cell via the plasma membrane. The coupling between (1) and (5) is achieved by the three flux terms shared by both.

More precisely, J_{LCC}, $J_{m_{\text{leak}}}$, and $J_{m_{\text{pump}}}$ in (1) describe the fluxes of calcium into and out of the cell via the plasma membrane. J_{pump} replenishes the calcium stores in the SR; it increases SR calcium concentration by decreasing cytosol calcium concentration. J_{leak} is a continuous leakage of those SR calcium stores into the cytosol; it increases cytosol concentration by decreasing SR calcium concentration. The pump term

$$J_{\text{pump}}(c) = V_{\text{pump}} \left(\frac{c^{n_{\text{pump}}}}{K_{\text{pump}}^{n_{\text{pump}}} + c^{n_{\text{pump}}}} \right) \tag{13}$$

is thus a function of cytosol calcium $c(\boldsymbol{x}, t)$. The leak term J_{leak} is a constant defined by

$$J_{\text{leak}} = J_{\text{pump}}(c_0), \tag{14}$$

which balances $J_{\text{pump}}(c)$ at basal level $c_0 = 0.1\ \mu$M of cytosol calcium. The pump term J_{pump}, a function of cytosolic calcium $c(\boldsymbol{x}, t)$, consists of the maximum pump velocity V_{pump} multiplied against the relationship between $c(\boldsymbol{x}, t)$ and the pump sensitivity K_{pump}; the exponent n_{pump} refers to the Hill coefficient (quantifying the degree of cooperative binding) for the pump function. This has the practical effect of multiplying the maximum possible pump velocity against a number between 0 and 1, exclusive. J_{leak}, which continuously leaks calcium into the cytosol from the SR, is simply J_{pump} evaluated at the basal cytosolic calcium concentration $c_0 = 0.1\,\mu$M. As noted, J_{pump} has two roles, namely to balance J_{leak} in the absence of sparking, but also to balance J_{CRU} under conditions of active calcium release.

The term J_{CRU} in (1) is the Ca^{2+} flux into the cytosol from the SR via each individual point source at which a CRU has been assigned. The effect of all CRUs is modeled as a superposition such that

$$J_{\text{CRU}}(c, s, \boldsymbol{x}, t) = \sum_{\hat{\boldsymbol{x}} \in \Omega_s} \hat{\sigma} \frac{s(\boldsymbol{x}, t)}{s_0} \mathcal{O}(c, s) \delta(\boldsymbol{x} - \hat{\boldsymbol{x}}), \tag{15}$$

with

$$\mathcal{O}(c, s) = \begin{cases} 1 & \text{if } u_{\text{rand}} \leq J_{\text{prob}}, \\ 0 & \text{if } u_{\text{rand}} > J_{\text{prob}}, \end{cases} \tag{16}$$

where

$$J_{\text{prob}}(c, s) = P_{\max} \left(\frac{c^{n_{\text{prob}_c}}}{K_{\text{prob}_c}^{n_{\text{prob}_c}} + c^{n_{\text{prob}_c}}} \right) \left(\frac{s^{n_{\text{prob}_s}}}{K_{\text{prob}_s}^{n_{\text{prob}_s}} + s^{n_{\text{prob}_s}}} \right). \tag{17}$$

The effect of each CRU is modeled here as a product of three terms:

(i) Similarly to how in J_{pump} the maximum pump rate is scaled against the concentration of available cytosol calcium, the maximum pump rate is scaled against the concentration of available cytosol calcium and the maximum rate of Ca^{2+} release $\hat{\sigma}$ is scaled here against the ratio of calcium concentration in the SR.

(ii) Following the same pattern, a maximum value multiplied against some scaling proportion between 0 and 1, the gating function $\mathcal{O}(c, s)$ has the practical effect of "budgeting" the calcium SR stores such that when the stores are low, the given CRU becomes much less likely to open; each CRU is assigned a uniformly distributed random value u_{rand}, which is compared to the single value returned by the CRU opening probability J_{prob} to determine whether or not the given CRU will open. J_{prob} is characterized by traits of CRUs CRU, which open for 5 ms and then proceed to close and remain in a refractory state for the following 100 ms.

(iii) The Dirac delta distribution $\delta(\boldsymbol{x} - \hat{\boldsymbol{x}})$ models each CRU as a point source for calcium release, which is defined by requiring $\delta(\boldsymbol{x} - \hat{\boldsymbol{x}}) = 0$ for all $\boldsymbol{x} \neq \hat{\boldsymbol{x}}$ and $\int_{\mathbb{R}^3} \psi(\boldsymbol{x}) \delta(\boldsymbol{x} - \hat{\boldsymbol{x}})\, d\boldsymbol{x} = \psi(\hat{\boldsymbol{x}})$ for any continuous function $\psi(\boldsymbol{x})$.

The linking between calcium signaling and electrical excitation consists of (6)–(7). The membrane potential of the cell depends on both the cytosol calcium ion concentration and also on the cytosol potassium ion (K^+) concentration [Banyasz et al. 2012; Morris and Lecar 1981]. In our model, the ω-term in (6) quantifies a dependence of V on c to complete the coupling from the chemical to the electrical systems in link ② in Figure 1 [Angeloff et al. 2016], after c in (1) already contains several terms that depend on V to implement link ① in Figure 1. The Ca^{2+} conductance is much faster than the K^{+} conductance, so the calcium conductance can be approximated as m_∞ or instantaneously steady-state at all times; the potassium conductance requires a separate description in (7). Needed parameter functions in (6)–(7) are

$$m_\infty(V) = \frac{1}{2}\left(1 + \tanh\left(\frac{V - V_1}{V_2}\right)\right), \tag{18}$$

$$n_\infty(V) = \frac{1}{2}\left(1 + \tanh\left(\frac{V - V_3}{V_4}\right)\right). \tag{19}$$

The connection between (1) and (6), link ① in Figure 1, the link from the electrical system to the calcium system, comes through

$$J_{\mathrm{LCC}} = \frac{\tau_{\mathrm{flux}}}{2F} S g_{\mathrm{Ca}} m_\infty(V)(V - V_{\mathrm{Ca}}), \tag{20}$$

the only calcium flux term to involve voltage. Note the parameter κ in (1), which is an external scaling factor for J_{LCC} rather than an intrinsic physiological component; if the value of κ is set to 0, the connection, link ① in Figure 1, is effectively switched off and the calcium dynamics are then modeled as though voltage were not involved. The surface area, S, of the cell is included in light of the fact that J_{LCC} describes the influx of calcium through L-type calcium channels (LCCs), which are present in the enclosing plasma membrane of the cell: the surface area of the cell is the surface area of the membrane.

We model the effect of the cytosol calcium concentration on the voltage by treating the calcium efflux term $J_{m_{\mathrm{pump}}} - J_{m_{\mathrm{leak}}}$ as equivalent to the sodium-calcium exchanger current: we are thus able to describe the current generated by the sodium-calcium exchange as a function of simple calcium loss.

The individual components of the calcium efflux term are near-duplicates in form of the earlier J_{pump} and J_{leak} functions in (13) and (14), respectively. As J_{pump} described the removal of calcium from the cytosol and its transfer into SR stores,

$$J_{m_{\mathrm{pump}}}(c) = V_{m_{\mathrm{pump}}}\left(\frac{c^{n_{m_{\mathrm{pump}}}}}{K_{m_{\mathrm{pump}}}^{n_{m_{\mathrm{pump}}}} + c^{n_{m_{\mathrm{pump}}}}}\right) \tag{21}$$

describes the removal of calcium from the cytosol and its transfer to outside the cell across the membrane. The leak term J_{leak} described a gradual leak of calcium into the cytosol from the SR, while J_{CRU} described an abrupt, high-concentration (high relative to the leak) release of calcium into the cytosol from the SR. Similarly,

$$J_{m_{\mathrm{leak}}} = J_{m_{\mathrm{pump}}}(c_0) \tag{22}$$

describes a gradual leak of calcium into the cytosol from outside the cell via the plasma membrane, while J_{LCC} describes a sudden spike of calcium release into the cytosol via the LCCs.

The seven-variable model presented in this section reduces to the three-variable model in the original sources [Izu et al. 2001a; 2001b] by the following choices: setting $\gamma = 0$, $\tau_v = 0$, $\omega = 0$, $\kappa = 0$, and $F_{\max} = 0$. With these parameters set to 0, links ①, ②, ③, and ④ in Figure 1 are all cut off.

3. Numerical method

In order to do calculations for the CICR model, we need to solve a system of time-dependent parabolic partial differential equations (PDEs). These PDEs are coupled by several nonlinear reaction and source terms. The current simulations use $n_s = 7$ physiological variables in (1)–(7). The domain in our model is a hexahedron Ω in (8) with isotropic CRU distribution as shown in Figure 2. A typical cell has on the order of $15 \times 15 \times 31 = 6{,}975$ calcium release units (CRUs) throughout the cell. A cell is reasonably modeled by the hexahedral shape elongated in the z-direction, since the focus of CICR research is on estimating correct physiological parameter values. Taking a method of lines (MOL) approach [Strikwerda 2004] to spatially discretize this model, we use the finite volume method (FVM) as the spatial discretization to ensure mass conservation on the discrete level and also to accommodate advection terms in the future, with $N = (N_x + 1)(N_y + 1)(N_z + 1)$ control volumes. Applying this to the case of the n_s PDEs results in a large system of ordinary differential equations (ODEs). A MOL discretization of a diffusion-reaction equations with second-order spatial derivatives results in a stiff ODE system. The time step size restrictions, due to the CFL condition, are considered too severe to allow for explicit time-stepping methods. This necessitates the use of a sophisticated ODE solver such as the family of implicit numerical differentiation formulas (NDFk). We use Newton's method as the nonlinear solver, and at each Newton step we use the biconjugate gradient stabilized method (BiCGSTAB) or other Krylov subspace methods as the linear solver. Complete details of the numerical method can be found in [Schäfer et al. 2015].

While the form of the PDEs in (1)–(7) is customary, the thousands of point sources at the calcium release units (CRUs) in the forcing terms imply that standard codes have difficulty. This explains why our research centered around creating a special-purpose code [Gobbert 2008; Hanhart et al. 2004; Seidman et al. 2012; Schäfer et al. 2015]. Additionally, we leverage the regular three-dimensional shape of the domain to program all methods in matrix-free form. Table 4 demonstrates thus, how despite fully implicit time-stepping, even relatively fine meshes use only reasonable amounts of memory. This provides the key benefit now for the complete

mesh resolution $N_x \times N_y \times N_z$	N	DOF n_{eq}	time steps	memory (GB) predicted	 observed
32×32×128	140,481	983,367	2667	0.125	< 1
64×64×256	1,085,825	7,600,775	3295	0.963	1.093
128×128×512	8,536,833	59,757,831	3867	7.569	8.476
256×256×1024	67,700,225	473,901,575	4470	60.024	67.103

Table 4. Sizing study with $n_s = 7$ variables using double precision arithmetic, listing the mesh resolution $N_x \times N_y \times N_z$, the number of control volumes $N = (N_x+1)(N_y+1)(N_z+1)$, the number of degrees of freedom (DOF) $n_{\text{eq}} = n_s N$, the number of time steps taken by the ODE solver up to final time $t_{\text{fin}} = 100\,\text{ms}$, and the predicted and observed memory usage in GB for a one-process run.

complex model with complete sophisticated numerical method stack, even with 7 or 8 variables, to fit into the memory of a single cluster node with extended memory. The table also shows how large the number of degrees of freedom (DOF) is, that is, the number of variables that have to be computed at every time step. With even modest meshes, there are millions of unknowns, and possibly hundreds of millions for fine meshes. This characterizes the numerical problem that needs to be tackled to resolve the large number of thousands of calcium release units (CRUs) in a cell. Note that Table 4 presents a numerical test on a shortened time frame with final time $t_{\text{fin}} = 100\,\text{ms}$, which is large enough to include significant nonlinear behavior, but small enough to allow for reasonable run times. In the following result sections, final times of $t_{\text{fin}} = 1{,}000\,\text{ms}$ and $t_{\text{fin}} = 5{,}000\,\text{ms}$ are used, as specified in the subsections as well as summarized in Table 5

The implementation of this model is done in C using MPI and OpenMP to parallelize computations. Parallelization is accomplished through the block-distribution of all large arrays to all MPI processes. We split of the mesh in the z-direction with one subdomain on each of the parallel processes. MPI commands such as `MPI_Isend` and `MPI_Irecv`, which are nonblocking point-to-point communication commands, send messages between neighboring processes. The collective command `MPI_Allreduce` is used for the computation of scalar products and norms.

4. Results

This section shows the results of simulations for the seven-variable model detailed in Section 2 with variables and their values at the initial time $t = 0\,\text{ms}$ specified in Table 1. The following subsections show physiological output that is possible through long-time computational simulations to a final time that is large enough

parameter	base case Sec. 4.1	SR store Sec. 4.2	voltage system Sec. 4.3	mechanical system Sec. 4.4	 Sec. 4.5
γ	0.0	0.01 to 1000	14.0	14.0	14.0
τ_v	0.0	0.0	0.1	0.1	0.1
$V_{m_{\text{pump}}}$	0.0	0.0	0.001 to 1	0.3	0.3
ω	0.0	0.0	0.1 to 1000	100.0	100.0
κ	0.0	0.0	0.001 to 10	0.1	0.1
$F_{\max}$	0.0	0.0	0.0	0.001 to 100	1,100
t_{fin}	1000	1000	1000	1000	5000

Table 5. Parameter differences between cases in this section.

to include several calcium waves. Tables 2 and 3 specify all model parameters, their names, and their values or ranges with units, except certain parameters that are varied in this section. Table 5 specifies for each of the subsections, what values are used. In Tables 2 and 3, the values of the variables are selected from Table 5. Section 4.1 connects to the original modeling in [Izu et al. 2001a; 2001b] by performing simulations with parameters chosen such that the seven-variable model reduces to the original three-variable model used in these papers. Sections 4.2, 4.3, and 4.4 successively vary parameters shown in Table 5, in addition to the parameter changes made in the previous subsections. This approach allows for the construction of the entire model by adding new components on top of previous ones. With the entire model pieced together, we can test the model as a whole. Section 4.5 briefly exhibits the power of the physiologically connected model by showing the impact of $F_{\max} = 1$ and $F_{\max} = 100$ on the entire system for $t = 5{,}000$ ms, a longer duration of time. These results show the power of the model with the physiologically linked components and motivate further studies using this model.

4.1. ***Base case.*** The seven-variable model in (1)–(7) reduces to the three-variable model in the original sources [Izu et al. 2001a; 2001b] by the following choices: setting $\kappa = 0$, $F_{\max} = 0$, and $\gamma = 0$. When looking at the seven equations, changing these parameters to zero is most appropriate as they result in removing the calcium store in the sarcoplasmic reticulum (SR) and the effects of mechanical contraction.

The final addition of the three-variable model, electrical excitation, can be absolutely removed by setting (6) and (7) to zero. Changing $\tau_v = 0$ achieves this. These three components are only unique to the seven-variable model. To confirm that the first equation, (1), acts identically to the first equation in [Izu et al. 2001a], we must be absolutely sure that $J_{m_{\text{leak}}} - J_{m_{\text{pump}}} = 0$. To do this, we change $V_{m_{\text{pump}}} = 0$ and $\omega = 0$, resulting in (21) negating to zero and both sides of (22) equaling zero.

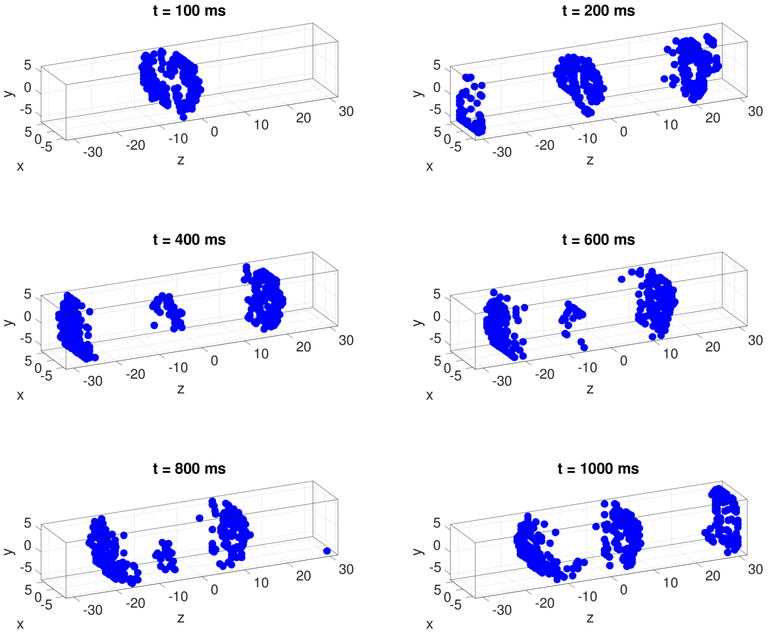


Figure 3. Open calcium release units throughout the cell.

With these additional changes, the seven-variable model functions identically to the three-variable model in the original sources [Izu et al. 2001a; 2001b].

The first set of plots, in Figure 3, display the locations of open calcium release units by a dot. The more dark dots are visible, the more CRUs are open at that specific time. More specifically, since one open CRU tends to promote neighboring CRUs to open through the diffusion of calcium, a clustering of dots indicates a region of several open CRUs. Note that we start with initial conditions in Table 5 for which there are no open CRUs and the right-hand sides of all equations in (1)–(7) are zero. Thus, these simulations represent a study in spontaneous sparking, in which some CRUs happen to open according to the probabilistic model in (15)–(17). This model embodies the effect of a higher concentration of cytosol calcium increasing the probability for a CRU to open in (17). The result of this spontaneous self-initiation is seen in the first subplot in Figure 3, where a number of CRUs in one region of the cell have opened by $t = 100\,\text{ms}$. The study of self-initiation was the original purpose of this model of calcium-induced calcium releases [Izu et al. 2001a; 2001b] and only required modest final times. The extension of the simulations to larger final times allows us to study if the opening and closing of CRUs self-organizes into a sustained wave. This is seen in the following subplots in Figure 3. The model requires CRUs to close for 100 ms after opening for 5 ms; thus the subplots indicate that the original cluster of open CRUs travels in both

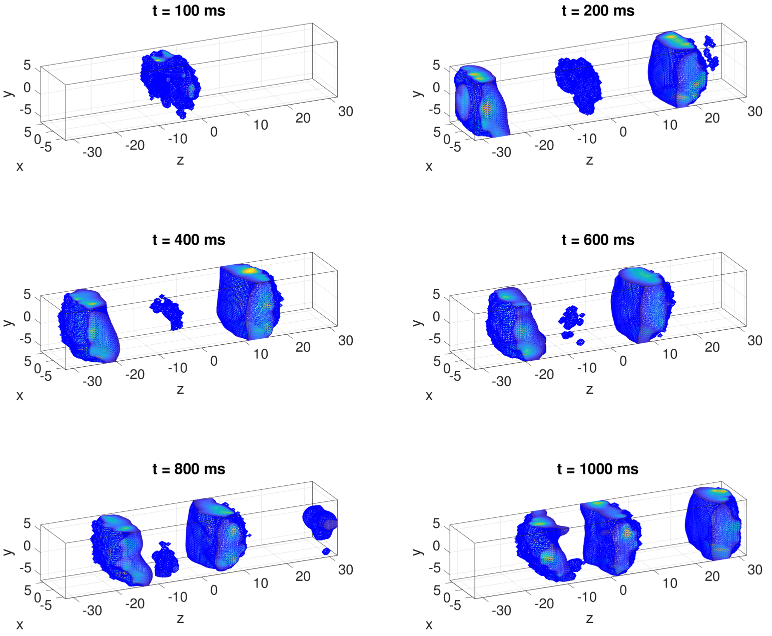


Figure 4. Concentration of $c(\boldsymbol{x}, t)$ with a critical value of 65 μM.

directions through the elongated direction of the cell and roughly repeats every 100 ms. Movies of the full results including intervening times confirm this behavior.

Figure 4 shows a collection of isosurface plots for calcium concentration in the cytosol. An isosurface plot displays the surface in the three-dimensional cell, where the species concentration is equal to a critical value, stated in the caption of the figure, here 65 μM for the concentration of cytosol calcium $c(\boldsymbol{x}, t)$. Blue represents when the species' concentration equals 65 μM. For densities lower than 65 μM, there are no markings. More extreme concentrations of each species are indicated with the spectrum from yellow to red, where darker hues are more intense in density. As we can see from the subplots in Figure 3, we have an original cluster in the middle at $t = 100$ ms that diffuses towards the ends of the cell and repeats on a 100 ms cycle. Due to the connection of the calcium in the cytosol with open CRUs through the effect of calcium-induced calcium release, calcium concentrations in the cytosol corresponds to locations of open CRUs.

This physiological concept is shown with each subplot in Figure 4, where the species' concentration colors are in agreement with the amount of open CRUs in Figure 3 at that time.

The concentrations of $b_1^{(c)}(\boldsymbol{x}, t)$ and $b_2^{(c)}(\boldsymbol{x}, t)$ have the same characteristics as the calcium levels in the cytosol since they are all influenced in the same way; therefore, we do not show them. The remaining isosurface plots for the base case are not

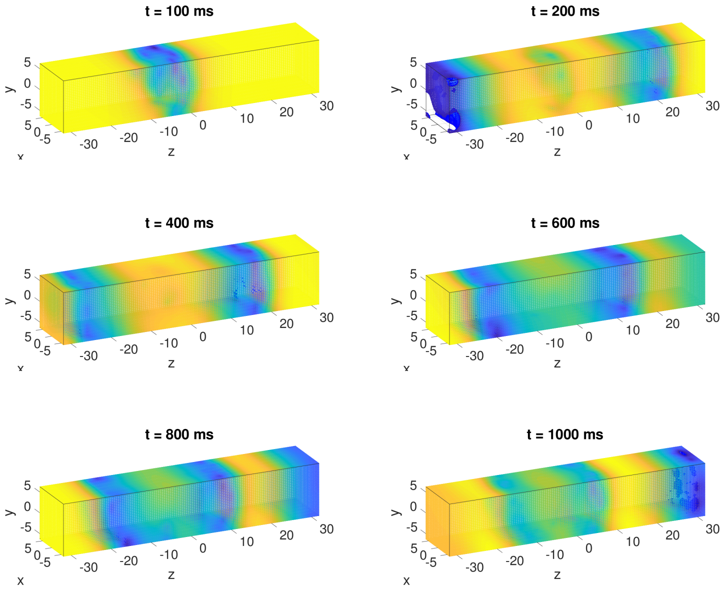


Figure 5. Concentration of $s(\boldsymbol{x}, t)$ with a critical value of 5000 μM.

included because each of the additional species' concentrations remain constant. Thus each subplot remains uneventful. As the CICR components are added to the base case in the following subsections, the respective isosurface plots will also be included.

4.2. ***Adding the calcium store in the sarcoplasmic reticulum (SR).*** To add the calcium store in the sarcoplasmic reticulum (SR), (5) must provide nonconstant concentrations of calcium. We can set $\gamma = 14$ to remove the previous reduction of this equation in Section 4.1. These values for γ and D_s are from Table 5.

We can look at the isosurface plots for calcium in the SR to determine the functionality of the calcium store in the SR with the two changed parameters. The seven equations allow for a range of concentrations for calcium in the SR from 0.01 to 10,000 μM. Figure 5 shows a collection of isosurface plots for calcium concentration in the SR with a critical value of 5,000 μM. Concentrations more extreme than the critical value are represented with a spectrum start from blue to yellow, where yellow represents concentrations closest to the maximum. When looking at Figure 5 the lighter concentrations of calcium in the SR directly correspond to the extreme concentrations of calcium throughout the cell.

Figure 5 shows that the calcium store was implemented. Unlike in Section 4.1 where the concentration of calcium in the SR remained constant at 10,000 μM, the subplots in Figure 5 contain numerous examples of varying concentration of calcium in the SR.

The choice of $\gamma = 14$ was determined through an analysis of several other cases with different values for γ. From the range that was tested, 0.01 to 1000, every value eventually resulted in continuous calcium waves. This is expected as adding to the calcium store does not hinder the continuous calcium waves. However, γ-values of 100 and above resulted in a range of calcium concentration in the SR that included negative values. The seven-equation model is limited for large γ-values as it does not account for negative concentrations, which physiologically are inconceivable. Simulations for γ-values 1 and below resulted in insignificant changes to the calcium in the SR, showing that the smaller values of γ do not sufficiently add to the decrease in calcium from the store according to the increase of calcium in the cytosol. From the parameter study, $\gamma = 10$ was idealistic, as continuous calcium waves are apparent and an appropriate calcium store was successfully implemented. In this paper, $\gamma = 14$ is used to keep consistency with previous studies that used $\gamma = 10$. No significant differences can be found between $\gamma = 14$ and $\gamma = 10$.

The CRU plots and the isosurface plots for $c(\boldsymbol{x}, t)$, $b_1^{(c)}(\boldsymbol{x}, t)$, and $b_2^{(c)}(\boldsymbol{x}, t)$ remained nearly identical to the plots in Section 4.1 This shows that the implementation of the calcium store did not heavily influence the concentrations of free fluorescent dye from (2) and free troponin from (3).

4.3. ***Adding the effect of electrical excitation.*** Implementing the effect of electrical excitation on top of the calcium store in the SR is achieved through activating (6) and (7). From the reduced version of the seven-variable model, setting $\tau_v = 0.1\,\mu\text{M}\,\mu\text{m}^3/\text{ms}$, $V_{m_{\text{pump}}} = 0.3\,\mu\text{M/ms}$, $\omega = 100$, and $\kappa = 0.1$ results in functional equations for electrical excitation. We can make several other cases with various values of each varied parameter to see how different extremes for each variable changes the resulting calcium wave.

From Section 4.2, we performed three separate one-dimensional studies in order to find values for $V_{m_{\text{pump}}}$, ω, and κ that fully add the voltage system without significant changes to the various species' concentrations in Section 4.2.

The first one-dimensional study varied $V_{m_{\text{pump}}}$ from 0.001 to 1. Changing this parameter to a fixed value allows for the functionality of the function (21) and consequently (22). Both $V_{m_{\text{pump}}}$ and $V_{m_{\text{leak}}}$ have an influence on (1) and (6). From this study, the largest value that retains the continuous calcium waves described in Section 4.1 was $V_{m_{\text{pump}}} = 0.3\,\mu\text{M/ms}$.

With this fixed value for $V_{m_{\text{pump}}}$, we can perform a one-dimensional study of ω, the feedback strength for Ca^{2+} efflux, and an additional study of κ. When varying ω, we noted that from $\omega = 0.1$ to 1000, the CRU plots are almost identical to our CRU plots from Section 4.2. As such, we can set $\omega = 100.0$, a fixed value, as all tested cases of ω simulated similar results to the plots shown in Section 4.2.

To fully implement the voltage system, we need to test values of κ to find what nonzero value of κ would allow for continuous calcium waves with a fully

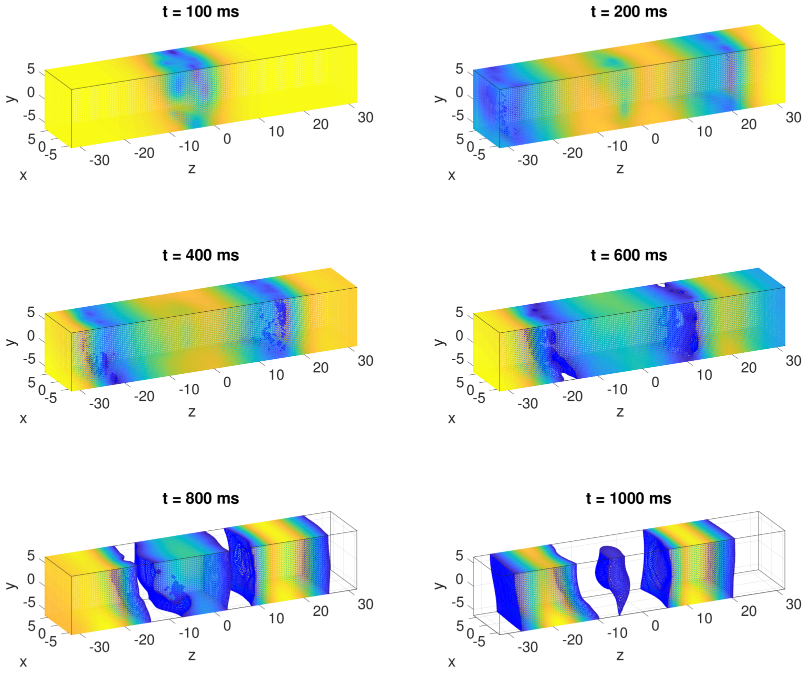


Figure 6. Concentration of $s(\boldsymbol{x}, t)$ with a critical value of 5000 μM.

implemented voltage system and calcium in the SR. When varying κ from 0.001 to 10, we noted that $\kappa = 0.1$ was the largest value in which the calcium waves were comparable to the previous subsections.

Setting $\kappa = 0.1$, $\omega = 100$, $\tau_v = 0.1\,\mu\text{M}\,\mu\text{m}^3/\text{ms}$, and $V_{m_{\text{pump}}} = 0.3\,\mu\text{M/ms}$ in addition to the predetermined values for the parameters in Table 5 allows for the full implementation of the voltage system without fundamentally changing the outflow of calcium.

The CRU plots and the isosurface plots for $c(\boldsymbol{x}, t)$, $b_1^{(c)}(\boldsymbol{x}, t)$, and $b_2^{(c)}(\boldsymbol{x}, t)$ did not significantly change with the implementation of the voltage system. However, there are more notable differences when comparing the calcium concentrations of the stores shown in Figures 6 and 5. In Figure 6, the latter plots have colorless space between each clump of calcium. As these spaces are representative of lower calcium concentrations, these plots show that the store in the SR has lower concentrations of calcium. Larger time studies may indicate that the calcium concentration in the store continues to lower past $t = 100\,\text{ms}$.

The voltage plot in Figure 7 shows that the electrical system was implemented. The voltage plots from the previous subsections remained constant at $-50\,\text{mV}$. In Figure 7, there is a clear repetition of the voltage rising up to approximately 26 mV back down to approximately $-23\,\text{mV}$. This figure confirms that external voltage is

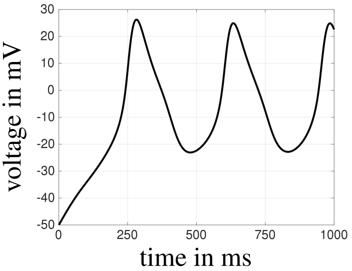


Figure 7. Voltage levels throughout the cell over period of 1000 ms.

present. Based on these comparisons, these plots maintain the same behaviors from Section 4.2 with the addition of the voltage system.

4.4. ***Adding the effect of mechanical contraction.*** In the seven-variable model, mechanical contraction is described through (3) and (4). From the parameters added in Section 4.3 shown in Table 5, a parameter study was performed on the values for $F_{\max}$. By changing this parameter to nonzero values, we can study the effects of mechanical contraction on top of the previously included calcium store and electrical excitation. In addition to these set values, we created other cases with various values for $F_{\max}$ to note the physiological changes with various values of (3) and (4).

For this parameter study, we tested a range from $F_{\max} = 0.001$ to $F_{\max} = 100$. For $F_{\max}$ values less than 1, the shortening factor remained the same, where after about 125 ms ε began to cycle between a small range and remained this way up to 1000 ms. For $F_{\max} = 100$, the shortening factor loses its period; that is, the value does not cycle. Rather, the shortening factor spikes up two times throughout the 1,000 ms. While $F_{\max} = 10$ maintains the cycling value of the shortening factor, there is an apparent delay for the cycle to begin. Unlike the start time of 125 ms for the lower values of $F_{\max}$, for $F_{\max} = 10$, the start time is 250 ms. Since the simulations are only 1,000 ms in length, the desired value for $F_{\max}$ is 1.

As in the previous subsections, the CRU plots and the isosurface plots for $c(\boldsymbol{x}, t)$, $b_1^{(c)}(\boldsymbol{x}, t)$, and $b_2^{(c)}(\boldsymbol{x}, t)$ appear to be extremely similar if not identical to plots from the previous subsections. Additionally, the concentrations shown in Figures 6 and 7 remain the same with the addition of $F_{\max}$. This shows that the implementation of the mechanical contraction did not heavily influence the other components to the extent of changing the system.

With $F_{\max}$ set to a nonzero value, the shortening factor, described in (11), is able to vary based on inactive actin-myosin cross-bridges described in (4). In the previous subsections, the lack of the variance in the shortening factor led to the inability of the continuous contractions of the cross-bridges to take place. Unlike the previous subsections, this shortening factor is able to vary in severity, allowing for the repetition of inactive cross-bridges switching to active ones. This repetitive cycle

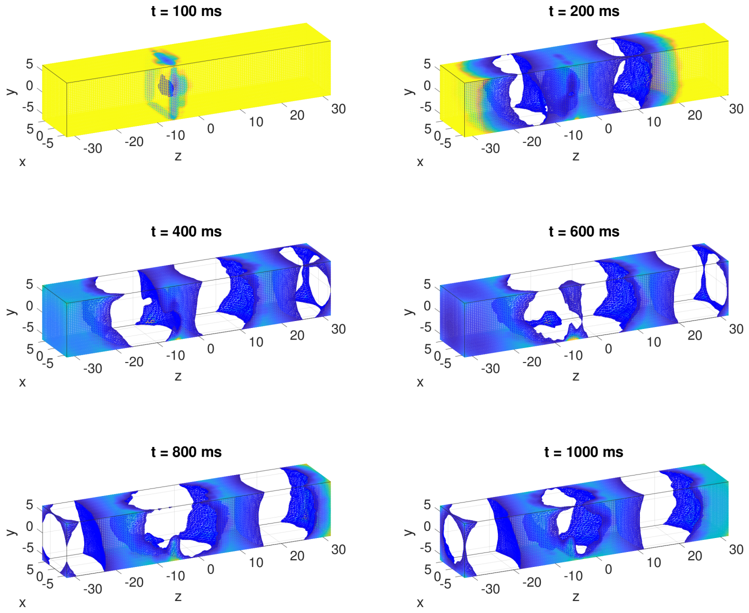


Figure 8. Concentration of $b_3^{(c)}(\boldsymbol{x}, t)$ with a critical value of 50 μM.

is displayed in Figure 8. In previous subsections, the plots for inactive actin-myosin cross-bridges remained constant throughout the 1,000 ms.

As in the previous subsections, the CRU plots and the isosurface plots for $c(\boldsymbol{x}, t)$, $b_1^{(c)}(\boldsymbol{x}, t)$, and $b_2^{(c)}(\boldsymbol{x}, t)$ appear to be extremely similar to plots from the previous subsections. Additionally, the concentrations shown in Figures 6 and 7 remain the same with the addition of $F_{\max}$ This shows that the implementation of the mechanical contraction with the so-far chosen parameters did not heavily influence the other components to the extent of changing the system.

4.5. ***Influence of $F_{\max}$ on the entire system.*** With these set parameters described in Section 4.1, we are able to test questions retaining to the entire model. One example of this is the study of $F_{\max}$ for the much longer final time $t_{\text{fin}} = 5{,}000$ ms rather than 1,000 ms on the seven-variable model. From the tested values for $F_{\max}$ from 0.001 to 100, values 1 and 100 are shown in the plots below. These values were determined through the parameter study of $F_{\max}$ from Section 4.4.

Figures 9 and 13 show plots of both the voltage V and shortening factor ε vs. time for the model. While the content is different, the structure of the voltage plots is identical to the voltage plots described in Section 4.3. The plots of the shortening factor ε vs. time track the value of the shortening factor, ε, used in (3).

There are several differences between how each $F_{\max}$ value affects the system. Figures 9–12 are representative of the same simulations as in Section 4.4, except the duration of time was extended from $t_{\text{fin}} = 1{,}000$ ms to 5,000 ms. As previously described, around 125 ms the calcium concentration in the cytosol shown in Figure 10 jumps from 0 to slightly less than 4×10^5 and slowly decreases as the calcium concentration in the store depletes. In Figure 12, the calcium concentration in the store drops at the same time. The calcium concentration in the store continues to drop until about 2,500 ms, where it then levels out. The concentration of $b_3^{(c)}(\boldsymbol{x}, t)$ in Figure 11 drops at the same time as the calcium change in the cytosol. Once the calcium in the cytosol depletes, the concentration of $b_3^{(c)}(\boldsymbol{x}, t)$ returns to its initial state.

When comparing Figures 9–12 using $F_{\text{max}} = 1$ and Figures 13–16 using $F_{\text{max}} = 100$, there are clear differences. In Figure 14, the calcium concentration jumps from its initial state similarly to the calcium concentration of Figure 10. However, it shortly returns to its initial state and remains so for several hundred milliseconds before jumping back up. This spike back downward is not characterized in Figures 9–12. The concentration around 1,000 ms jumps up to nearly 9×10^5, more than double the peak concentration of calcium in the cytosol for $F_{\text{max}} = 1$. Furthermore, the range of concentration levels in 100 ms is much larger than in Figure 10. The calcium in the store, shown in Figure 12, overall has a downward trend until around 3,000 ms, where the calcium concentration in the store levels out. The calcium concentration in the cytosol in Figure 10 correspondingly drops down to nearly 0. Similar to the first spike of calcium in Figure 10, in Figure 11, the concentration of $b_3^{(c)}(\boldsymbol{x}, t)$ drops with the spike and almost returns to its initial state before dropping a second time around 1,000 ms. The concentration remains relatively constant until the calcium in the cytosol drops. At this point the concentration of $b_3^{(c)}(\boldsymbol{x}, t)$ rises back up to the same concentration as its initial state and remains around that level until the final time $t_{\text{fin}} = 5{,}000$ ms.

There are also notable differences between the plots of the shortening factor ε vs. time in Figures 9 and 13. Both plots generally have the same characteristics as the plots of $b_3^{(c)}$. Where the value for ε is greater, the concentration of $b_3^{(c)}$ is lower. Between the plots in Figures 9 and 13, the range of oscillation is much smaller in Figure 13 than in Figure 9 during the period where the value of ε oscillates.

Unlike the other plots, the plots of voltage V vs. time in Figures 9 and 13 retain consistency throughout the entire 5,000 ms. The voltage level in both plots oscillates from a negative value to a positive value. It appears that the range of one oscillation of voltage shrinks slightly as the calcium levels in the cytosol increase. Other than this, it appears that the change in F_{max} does not have such a drastic impact on the characteristics of voltage compared to the concentrations of each species of the cell. As such, it appears that the influence of the mechanical system is directed towards calcium signaling and therefore has an indirect influence on voltage. This might be expected as there are no direct links between the mechanical system and electrical excitation.

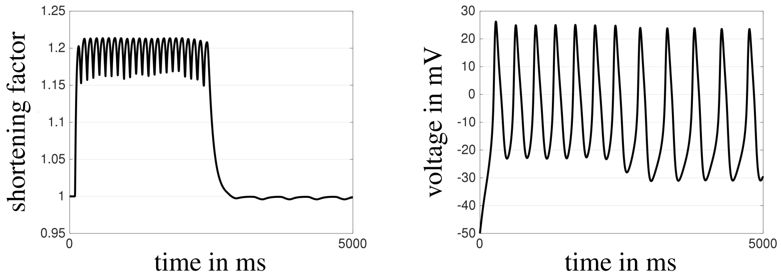


Figure 9. Plots of voltage V and shortening factor ε vs. time with $F_{\max} = 1$ for $t_{\text{fin}} = 5{,}000\,\text{ms}$.

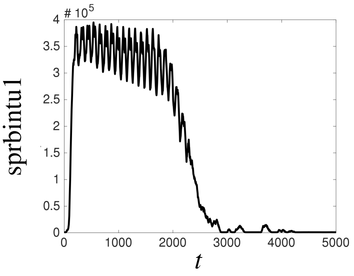


Figure 10. Total concentration of $c(\boldsymbol{x}, t)$ vs. time with $F_{\max} = 1$ for $t_{\text{fin}} = 5{,}000\,\text{ms}$.

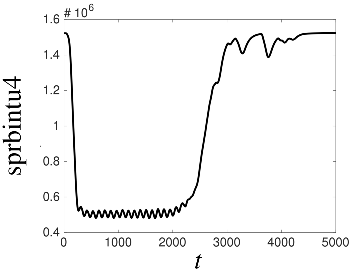


Figure 11. Total concentrations of $b_3^{(c)}(\boldsymbol{x}, t)$ vs. time with $F_{\max} = 1$ for $t_{\text{fin}} = 5{,}000\,\text{ms}$.

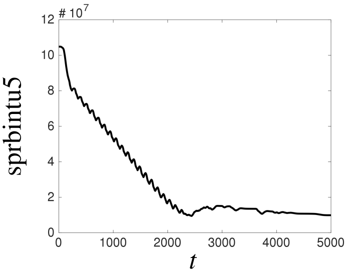


Figure 12. Total concentrations of $s(\boldsymbol{x}, t)$ vs. time with $F_{\max} = 1$ for $t_{\text{fin}} = 5{,}000\,\text{ms}$.

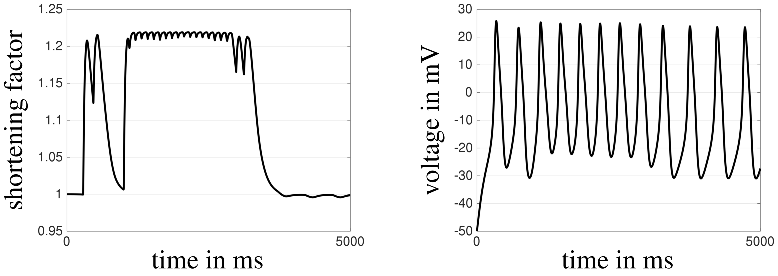


Figure 13. Plots of voltage V and shortening factor ε vs. time with $F_{\max} = 100$ for $t_{\text{fin}} = 5{,}000\,\text{ms}$.

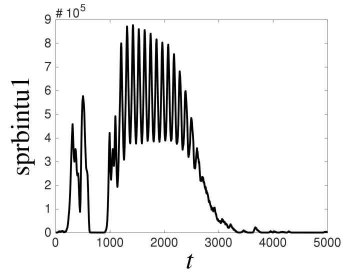


Figure 14. Total concentration of $c(\boldsymbol{x}, t)$ vs. time with $F_{\max} = 100$ for $t_{\text{fin}} = 5{,}000\,\text{ms}$.

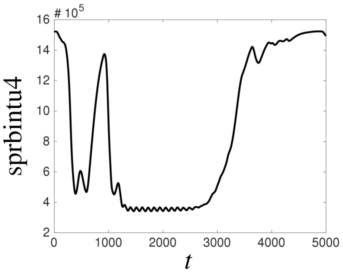


Figure 15. Total concentrations of $b_3^{(c)}(\boldsymbol{x}, t)$ vs. time with $F_{\max} = 100$ for $t_{\text{fin}} = 5{,}000\,\text{ms}$.

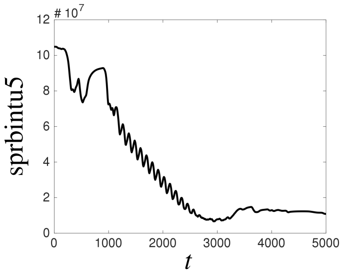


Figure 16. Total concentrations of $s(\boldsymbol{x}, t)$ vs. time with $F_{\max} = 100$ for $t_{\text{fin}} = 5{,}000\,\text{ms}$.

5. Conclusions and outlook

We successfully created a table of parameters in Section 4 that allow for a more complete functioning model. This model includes the calcium store, electrical excitation, and mechanical contraction of the cardiac muscle cell. Unlike before, this table allows for the cohesion of all components of the model. With this, we are able to study the model in its entirety. As an example, we showed some influences of $F_{\max}$, a constant of the shortening factor, on the entire system over a period of 5,000 ms. When comparing small and large values for $F_{\max}$, we noticed that the influence of $F_{\max}$ was much less extreme on the electrical excitation compared to the calcium signaling. When comparing the plots of the relevant species of the system, this conclusion becomes apparent. This is important as it allows us to grasp a better understanding of the linkage between the mechanical system and calcium signaling, shown by link ① in Figure 1.

Future research on this could look more in depth at different values of other parameters with a final time of 5,000 ms. This would allow for studies on the system for longer periods of time. Studying multiple cases for each parameter value with randomized locations for calcium sparking would allow for more concrete conclusions with a stronger representation of the physiological processes of the cell. While this is one further area of study, this complete set of parameters opens many areas to study the model in its entirety.

Acknowledgements

The hardware in the UMBC High Performance Computing Facility (HPCF) is supported by the U.S. National Science Foundation through the MRI program (grant nos. CNS–0821258, CNS–1228778, and OAC–1726023) and the SCREMS program (grant no. DMS–0821311), with additional substantial support from the University of Maryland, Baltimore County (UMBC). See hpcf.umbc.edu for more information on HPCF and the projects using its resources. Barajas was supported by UMBC as HPCF RA.

References

[Alexander et al. 2015] A. M. Alexander, E. K. DeNardo, E. Frazier, III, M. McCauley, N. Rojina, Z. Coulibaly, B. E. Peercy, and L. T. Izu, "Spontaneous calcium release in cardiac myocytes: store overload and electrical dynamics", *Spora* **1**:1 (2015), 36–48.

[Angeloff et al. 2016] K. Angeloff, C. A. Barajas, A. Middleton, U. Osia, J. S. Graf, M. K. Gobbert, and Z. Coulibaly, "Examining the effect of introducing a link from electrical excitation to calcium dynamics in a cardiomyocyte", *Spora* **2**:1 (2016), 49–73.

[Banyasz et al. 2012] T. Banyasz, B. Horvath, Z. Jian, L. T. Izu, and Y. Chen-Izu, "Profile of L-type Ca^{2+} current and Na^{+}/Ca^{2+} exchange current during cardiac action potential in ventricular myocytes", *Heart Rhythm* **9**:1 (2012), 134–142.

[Deetz et al. 2017a] K. Deetz, N. Foster, D. Leftwich, C. Meyer, S. Patel, C. Barajas, M. K. Gobbert, and Z. Coulibaly, "Developing the coupling of the mechanical to the electrical and calcium systems in a heart cell", Technical Report HPCF–2017–15, High Performance Computing Facility, University of Maryland, Baltimore County, 2017, available at http://hdl.handle.net/11603/11308.

[Deetz et al. 2017b] K. Deetz, N. Foster, D. Leftwich, C. Meyer, S. Patel, C. Barajas, M. K. Gobbert, and Z. Coulibaly, "Examining the electrical excitation, calcium signaling, and mechanical contraction cycle in a heart cell", *Spora* **3**:1 (2017), 66–85.

[Gobbert 2008] M. K. Gobbert, "Long-time simulations on high resolution meshes to model calcium waves in a heart cell", *SIAM J. Sci. Comput.* **30**:6 (2008), 2922–2947. MR Zbl

[Hanhart et al. 2004] A. L. Hanhart, M. K. Gobbert, and L. T. Izu, "A memory-efficient finite element method for systems of reaction-diffusion equations with non-smooth forcing", *J. Comput. Appl. Math.* **169**:2 (2004), 431–458. MR Zbl

[Izu et al. 2001a] L. T. Izu, J. R. H. Mauban, C. W. Balke, and W. G. Wier, "Large currents generate cardiac Ca^{2+} sparks", *Biophysical J.* **80**:1 (2001), 88–102.

[Izu et al. 2001b] L. T. Izu, W. G. Wier, and C. W. Balke, "Evolution of cardiac calcium waves from stochastic calcium sparks", *Biophysical J.* **80**:1 (2001), 103–120.

[Morris and Lecar 1981] C. Morris and H. Lecar, "Voltage oscillations in the barnacle giant muscle fiber", *Biophysical J.* **35**:1 (1981), 193–213.

[Mozaffarian et al. 2015] D. Mozaffarian et al., "Heart disease and stroke statistics — 2015 update: A report from the american heart association", *Circulation* **131**:4 (2015), e29–e322.

[Schäfer et al. 2015] J. Schäfer, X. Huang, S. Kopecz, P. Birken, M. K. Gobbert, and A. Meister, "A memory-efficient finite volume method for advection-diffusion-reaction systems with nonsmooth sources", *Numer. Methods Partial Differential Equations* **31**:1 (2015), 143–167. MR Zbl

[Seidman et al. 2012] T. I. Seidman, M. K. Gobbert, D. W. Trott, and M. Kružík, "Finite element approximation for time-dependent diffusion with measure-valued source", *Numer. Math.* **122**:4 (2012), 709–723. MR Zbl

[Strikwerda 2004] J. C. Strikwerda, *Finite difference schemes and partial differential equations*, 2nd ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, 2004. MR Zbl

Received: 2019-06-01 Revised: 2019-12-13 Accepted: 2020-01-06

gkroiz1@umbc.edu *Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD, United States*

barajasc@umbc.edu *Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD, United States*

gobbert@umbc.edu *Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD, United States*

bpeercy@umbc.edu *Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD, United States*

INVOLVE 13:3 (2020)
https://doi.org/10.2140/involve.2020.13.425

Covering numbers and schlicht functions

Philippe Drouin and Thomas Ransford

(Communicated by Kenneth S. Berenhaut)

We determine upper and lower bounds for the minimal number of balls of a given radius needed to cover the space of schlicht functions.

1. Introduction

Let (X, d) be a metric space. The *δ-covering number* of X, denoted by $N_X(\delta)$, is the minimal number of closed balls of radius δ needed to cover X. It may happen that $N_X(\delta) = \infty$, but if X is compact then always $N_X(\delta) < \infty$. The quantity $N_X(\delta)$ arises in connection with the notion of the *box-counting dimension* of X, which is defined as

$$\dim_B(X) := \lim_{\delta\to 0^+} \frac{\log N_X(\delta)}{\log(1/\delta)},$$

whenever the limit exists; see, e.g., [Falconer 2014, §2.1]. The covering number also plays an important role in the theory of universal approximation; see, e.g., [Kalmes et al. 2012]. In the latter context, the metric space X is usually a function space.

Our aim is to estimate the covering number for a particular function space, namely the class $\mathcal{S}$ of *schlicht functions*. This is the class of all holomorphic functions f on the open unit disk $\mathbb{D}$ such that f is injective and satisfies $f(0) = 0$ and $f'(0) = 1$. The class $\mathcal{S}$ arises naturally in connection with the Riemann mapping theorem. It has been studied extensively and is the subject of several books, e.g., [Duren 1983; Goodman 1983a 1983b; Pommerenke 1975; Schober 1975].

2. Statement of main result

Let $\mathrm{Hol}(\mathbb{D})$ denote the set of all holomorphic functions on the open unit disk $\mathbb{D}$. It is a Fréchet space with respect to the topology of uniform convergence on compact

MSC2010: 30C45, 54E35.
Keywords: covering number, holomorphic function, schlicht function.
Ransford was supported by grants from NSERC and from the Canada Research Chairs program.

subsets of $\mathbb{D}$. There are many choices of metric that give rise to this topology. A very common one is to take

$$d(f,g) := \sum_{j\ge 1} \lambda_j \min\bigl\{1, \max_{|z|\le r_j} |f(z)-g(z)|\bigr\}, \tag{1}$$

where (λ_j) is a positive sequence with $\sum_j \lambda_j < \infty$, and r_j is a sequence in $(0,1)$ such that $r_j \to 1$.

With respect to this metric, the set $\mathcal{S}$ is a compact subset of $\mathrm{Hol}(\mathbb{D})$, see, e.g., [Duren 1983, p.276], and therefore a compact metric space in its own right. Our aim is to estimate the covering number of this space. In order to do this, we need to impose some conditions on the sequences (λ_j) and (r_j). We shall assume that there exist constants $0 < \lambda < 1$ and $\alpha > 0$ such that

$$\lambda_j \asymp \lambda^j \quad \text{and} \quad 1 - r_j \asymp j^{-\alpha}. \tag{2}$$

These choices seem reasonable. In the literature, λ_j is nearly always taken to be 2^{-j}. Usually r_j is not specified explicitly, but in [Duren 1983, §9.1] it is taken to be $1 - 1/j$.

Here and in what follows, we write $a \lesssim b$ to signify that a, b are positive functions or sequences such that a/b remains bounded above and $a \asymp b$ to signify that $a \lesssim b$ and $b \lesssim a$ hold together.

The following theorem is our main result.

Theorem 2.1. *Let d be the metric on $\mathcal{S}$ given by* (1), *where (λ_j) and (r_j) satisfy* (2). *Then*

$$\log^2(1/\delta) \lesssim \log N_{\mathcal{S}}(\delta) \lesssim \log^{2+\alpha}(1/\delta), \quad \delta \to 0^+. \tag{3}$$

The lower bound in (3) is a generalization of a result obtained in [Kalmes et al. 2012]. It shows, in particular, that the box-counting dimension of $\mathcal{S}$ is infinite. We believe that the upper bound in (3) is new.

3. Proof of the lower bound in Theorem 2.1

Our starting point is the following sufficient condition for membership in $\mathcal{S}$. The result is well known, but we include a short proof for convenience.

Proposition 3.1. *Let $f(z) := z + \sum_{k\ge 2} a_k z^k$, where $\sum_{k\ge 2} k|a_k| \le 1$. Then $f \in \mathcal{S}$.*

Proof. We may suppose that at least one a_k is nonzero, otherwise the result is obvious. Let z, w be distinct points of $\mathbb{D}$. Then, by the triangle inequality,

$$|f(z) - f(w)| \ge |z - w| - \biggl|\sum_{k\ge 2} a_k z^k - \sum_{k\ge 2} w^k\biggr|.$$

Now

$$\left|\sum_{k\geq 2} a_k z^k - \sum_{k\geq 2} a_k w^k\right| \leq \sum_{k\geq 2} |a_k||z^k - w^k| < \sum_{k\geq 2} k|a_k||z-w| \leq |z-w|,$$

the middle inequality being strict because at least one a_k is nonzero. It follows that $f(z) \neq f(w)$. Thus f is injective, and so $f \in \mathcal{S}$. □

Let $\mathcal{A}$ be the family of functions $f(z) := z + \sum_{k\geq 2} a_k z^k$ such that $\sum_{k\geq 2} k|a_k| \leq 1$. By virtue of Proposition 3.1, we have $\mathcal{A} \subset \mathcal{S}$, and therefore $N_{\mathcal{S}}(\delta) \geq N_{\mathcal{A}}(2\delta)$. Thus, to establish the lower bound in Theorem 2.1, it suffices to prove the following theorem.

Theorem 3.2. *Let d be the metric on $\mathcal{A}$ given by* (1), *where* (λ_j) *and* (r_j) *satisfy* (2). *Then*

$$\log N_{\mathcal{A}}(\delta) \gtrsim \log^2(1/\delta), \quad \delta \to 0^+. \tag{4}$$

To prove this result, it is first convenient to establish a lemma.

Lemma 3.3. *Let $f, g \in \mathcal{A}$, say $f(z) := z + \sum_{k\geq 2} a_k z^k$ and $g(z) := z + \sum_{k\geq 2} b_k z^k$. Then*

$$d(f, g) \geq \sum_{k\geq 2} (\lambda_k r_k^k/2)|a_k - b_k|.$$

Proof. The standard Cauchy estimate for Taylor coefficients gives

$$|a_k - b_k| \leq \max_{|z|\leq r_k} \frac{|f(z) - g(z)|}{r^k}.$$

Also, since $|a_k|, |b_k|, r_k \leq 1$, we have $|a_k - b_k|r_k^k/2 \leq 1$. Combining these observations, we obtain

$$\begin{aligned} d(f, g) &= \sum_{k\geq 1} \lambda_k \min\bigl\{1, \max_{|z|\leq r_k} |f(z) - g(z)|\bigr\} \\ &\geq \sum_{k\geq 2} \lambda_k \min\bigl\{1, |a_k - b_k|r_k^k\bigr\} \geq \sum_{k\geq 2} \lambda_k |a_k - b_k| r_k^k/2. \end{aligned}$$

□

Proof of Theorem 3.2. Let $n \geq 2$, and let K be a large integer depending on n, to be chosen later. Consider the collection of functions

$$f(z) := z + \sum_{k=2}^{n} \frac{1}{kn}\frac{t_k}{K} z^k, \quad t_k \in \{1, 2, \ldots, K\},\ k = 2, \ldots, n.$$

All of these functions belong to $\mathcal{A}$. There are K^{n-1} of them. Also, by Lemma 3.3, the d-distance between any two of them, f, g say, is at least

$$d(f, g) \geq \left(\min_{2\leq k\leq n} \frac{\lambda_k r_k^k}{2}\right)\frac{1}{n^2 K}.$$

No ball of radius one third the right-hand side can contain more than one of these functions. It follows that

$$N_{\mathcal{A}}\Big(\frac{1}{3}\Big(\min_{2\le k\le n}\frac{\lambda_k r_k^k}{2}\Big)\frac{1}{n^2K}\Big)\ge K^{n-1}. \tag{5}$$

By our assumptions, there exists $\alpha > 0$ such that $1 - r_k \asymp k^{-\alpha}$. It follows that $r_k^k \ge e^{-ck^{1-\alpha}}$ for some constant $c > 0$. Since also $\lambda_k \ge C\lambda^k$ for some constant $C > 0$, we have

$$\frac{1}{3}\Big(\min_{2\le k\le n}\frac{\lambda_k r_k^k}{3}\Big)\frac{1}{n^2}\ge\frac{C}{6}\exp(-n\log(1/\lambda)-cn^{1-\alpha}-2\log n).$$

If we choose $\rho \in (0, \lambda)$, then $n\log(1/\lambda) + cn^{1-\alpha} + 2\log n \le n\log(1/\rho)$ for all sufficiently large n, and for these n we then have

$$\Big(\min_{2\le k\le n}\frac{\lambda_k r_k^k}{6}\Big)\frac{1}{n^2}\ge\rho^n. \tag{6}$$

Reducing ρ, if necessary, we can ensure that this inequality holds for all n, and also that $1/\rho$ is an integer. Note that ρ may depend on the sequences (λ_k) and (r_k), but it is independent of K. Combining the inequalities (5) and (6), we obtain

$$N_{\mathcal{A}}(\rho^n/K)\ge K^{n-1}.$$

We now choose K to be $K := \rho^{-n}$. This gives

$$N_{\mathcal{A}}(\rho^{2n})\ge\rho^{-n(n-1)}.$$

As this holds for each $n \ge 2$, we deduce that (4) holds. □

4. Proof of the upper bound in Theorem 2.1

This time, our starting point is the famous result of [de Branges 1985] that proved the Bieberbach conjecture.

Theorem 4.1. *Let $f \in \mathcal{S}$, say $f(z) = \sum_{k\ge0} a_k z^k$. Then $|a_k| \le k$ for all k.*

Special cases of this result had previously been established by various mathematicians for small values of k. Much earlier, Littlewood [1925] had proved the slightly weaker (but much easier) estimate $|a_k| \le ek$ valid for all k. In fact Littlewood's estimate would do for our purposes.

Again, it is convenient to introduce some notation. We denote by $\mathcal{B}$ the family of functions $f(z) := \sum_{k\ge1} a_k z^k$ such that $|a_k| \le k$ for all k. By virtue of Theorem 4.1, we have $\mathcal{S} \subset \mathcal{B}$, and so $N_{\mathcal{S}}(\delta) \le N_{\mathcal{B}}(\delta/2)$. Thus, to establish the upper bound in Theorem 2.1, it suffices to prove the following theorem.

Theorem 4.2. *Let d be the metric on $\mathcal{B}$ given by* (1), *where* (λ_j) *and* (r_j) *satisfy* (2). *Then*

$$\log N_{\mathcal{B}}(\delta) \lesssim \log^{2+\alpha}(1/\delta), \quad \delta \to 0^+. \tag{7}$$

For the proof, we need the following approximation lemma.

Lemma 4.3. *There exists a constant $c > 0$ with the following property*: *for each $f \in \mathcal{B}$ and $n \geq 1$, there is a polynomial $p \in \mathcal{B}$ such that* $\deg p \leq n$ *and*

$$d(f, p) \leq \exp(-cn^{1/(1+\alpha)}).$$

Proof. Let $f \in \mathcal{B}$, say $f(z) = \sum_{k \geq 1} a_k z^k$. Let $n \geq 1$ and set $p(z) := \sum_{k=1}^n a_k z^k$. Clearly $p \in \mathcal{B}$ and $\deg p \leq n$. Also

$$d(f, p) \leq \sum_{j \geq 1} \lambda_j \max_{|z| \leq r_j} \left| \sum_{k \geq n+1} a_k z^k \right| \leq \sum_{j \geq 1} \left(\lambda_j \sum_{k \geq n+1} k r_j^k \right).$$

So, to prove the lemma, it suffices to show that

$$\sum_{j \geq 1} \left(\lambda_j \sum_{k \geq n+1} k r_j^k \right) \leq \exp(-cn^{1/(1+\alpha)}), \quad n \geq 1, \tag{8}$$

where $c > 0$ is a constant independent of n.

Now, for $r \in (0, 1)$, we have

$$\sum_{k \geq n+1} k r^k = r \frac{d}{dr} \left(\sum_{k \geq n+1} r^k \right) = r \frac{d}{dr} \left(\frac{r^{n+1}}{1-r} \right) \leq (n+2) \frac{r^{n+1}}{(1-r)^2}.$$

Therefore

$$\sum_{j \geq 1} \left(\lambda_j \sum_{k \geq n+1} k r_j^k \right) \leq \sum_{j \geq 1} \lambda_j (n+2) \frac{r_j^{n+1}}{(1-r_j)^2}.$$

Recalling that $\lambda_j \asymp \lambda^j$ and $1 - r_j \asymp j^{-\alpha}$, we see that there are constants $C_1, c_1 > 0$, independent of n, such that

$$\sum_{j \geq 1} \lambda_j (n+2) \frac{r_j^{n+1}}{(1-r_j)^2} \leq C_1 n \sum_{j \geq 1} \lambda^j j^{2\alpha} \exp(-c_1 n j^{-\alpha}).$$

Now

$$\begin{aligned} \sum_{j \leq n^{1/(1+\alpha)}} \lambda^j j^{2\alpha} \exp(-c_1 n j^{-\alpha}) &\leq \sum_{j \leq n^{1/(1+\alpha)}} \lambda^j n^{2\alpha/(1+\alpha)} \exp(-c_1 n^{1/(1+\alpha)}) \\ &\leq \frac{1}{1-\lambda} n^{2\alpha/(1+\alpha)} \exp(-c_1 n^{1/(1+\alpha)}). \end{aligned}$$

Also

$$\begin{aligned}\sum_{j>n^{1/(1+\alpha)}} \lambda^j j^{2\alpha} \exp(-c_1 n j^{-\alpha}) &\le \sum_{j>n^{1/(1+\alpha)}} \lambda^j j^{2\alpha} \\ &\le \bigl(\sup_{j\ge1} \lambda^{j/2} j^{2\alpha}\bigr) \sum_{j>n^{1/(1+\alpha)}} \lambda^{j/2} \le C_2 \frac{(\sqrt{\lambda})^{n^{1/(1+\alpha)}}}{1-\sqrt{\lambda}},\end{aligned}$$

where C_2 is another constant. Combining all these inequalities, we see that (8) holds for any positive constant $c < \min\{c_1, \log(1/\sqrt{\lambda})\}$, at least for all sufficiently large n. Reducing c further, if necessary, we can ensure that (8) holds for all n. □

Remark. The estimate in Lemma 4.3 is sharp. Indeed, consider the so-called Koebe function, namely $f(z) := z/(1-z)^2 = \sum_{k\ge1} kz^k$. This belongs to $\mathcal{B}$, and even to $\mathcal{S}$. Also, for every polynomial p of degree n, whether it lies in $\mathcal{B}$ or not, we have

$$\max_{|z|\le r_j} |f(z)-p(z)|^2 \ge \frac{1}{2\pi r_j} \int_{|z|=r_j} |f(z)-p(z)|^2\,|dz| \ge \sum_{k\ge n+1} k^2 r_j^{2k} \ge r_j^{2n+2}.$$

It follows that $d(f,p) \ge \sum_{j\ge1} \lambda_j r_j^{n+1}$. Retaining just the term with $j = \lfloor n^{1/(1+\alpha)} \rfloor$, we see that

$$d(f,p) \ge \exp(-cn^{1/(1+\alpha)}),$$

where $c>0$ is a constant independent of p and n.

Proof of Theorem 4.2. Let $n\ge1$, and let K be a large integer depending on n, to be chosen later. Consider the collection of polynomials

$$q(z) := \sum_{k=1}^{n} k \frac{(s_k+it_k)}{\sqrt{2}K} z^k, \quad s_k, t_k \in \{-K, -(K-1), \dots, K\},\ k=1,\dots,n.$$

Clearly these polynomials all belong to $\mathcal{B}$, and there are $(2K+1)^{2n}$ of them. Given an arbitrary polynomial $p\in\mathcal{B}$ with $\deg p \le n$, there is a member q of this collection whose coefficients all lie within n/K of the corresponding coefficients of p, and consequently

$$d(p,q) = \sum_{j\ge1} \lambda_j \min\bigl\{1, \max_{|z|\le r_j} |p(z)-q(z)|\bigr\} \le \sum_{j\ge1} \lambda_j \frac{n^2}{K} = Cn^2/K,$$

where C is a constant independent of n. Combining this inequality with the result of Lemma 4.3, we see that, if we set $\delta := Cn^2/K + \exp(-cn^{1/(1+\alpha)})$, then the δ-balls around the polynomials q cover $\mathcal{B}$. Hence

$$N_{\mathcal{B}}(Cn^2/K + \exp(-cn^{1/(1+\alpha)})) \le (2K+1)^{2n}.$$

We now choose K to be $K := \lceil Cn^2 \exp(cn^{1/(1+\alpha)})\rceil$. This gives

$$\begin{aligned} N_{\mathcal{B}}(2\exp(-cn^{1/(1+\alpha)})) &\leq (2\lceil Cn^2 \exp(cn^{1/(1+\alpha)})\rceil + 1)^{2n} \\ &= \exp(2n \log(2\lceil Cn^2 \exp(cn^{1/(1+\alpha)})\rceil + 1)) \\ &\leq \exp(2n(\log(n^2) + cn^{1/(1+\alpha)} + O(1))) \\ &\leq \exp(2cn^{(2+\alpha)/(1+\alpha)} + O(n \log n)). \end{aligned}$$

From this it is straightforward to deduce that (7) holds. □

5. Concluding remarks and questions

(1) Among several well-known subclasses of $\mathcal{S}$, there is the class $\mathcal{C}$ of convex functions, namely the set of those $f \in \mathcal{S}$ such that $f(\mathbb{D})$ is a convex set. Theorem 2.1 holds with $\mathcal{S}$ replaced by $\mathcal{C}$. Indeed, the upper bound is obvious. As for the lower bound, it suffices to repeat the same proof, using the following result of [Goodman 1957] in place of Lemma 3.3: if $f(z) := z + \sum_{k\geq 2} a_k z^k$, where $\sum_{k\geq 2} k^2 |a_k| \leq 1$, then $f \in \mathcal{C}$.

(2) Which, if either, of the bounds in (3) is sharp? The bounds are based on the inclusions $\mathcal{A} \subset \mathcal{S} \subset \mathcal{B}$. The spaces $\mathcal{A}$ and $\mathcal{B}$ are easier to handle than $\mathcal{S}$ because they resemble infinite cartesian products. However, to answer the question above, we probably need to use finer properties of $\mathcal{S}$ to determine which of $\mathcal{A}$ or $\mathcal{B}$ better approximates $\mathcal{S}$.

Acknowledgement

The authors thank the anonymous referees for their careful reading of the manuscript, and for their corrections and helpful suggestions.

References

[de Branges 1985] L. de Branges, "A proof of the Bieberbach conjecture", *Acta Math.* **154**:1-2 (1985), 137–152. MR Zbl

[Duren 1983] P. L. Duren, *Univalent functions*, Grundlehren der Math. Wissenschaften **259**, Springer, 1983. MR Zbl

[Falconer 2014] K. Falconer, *Fractal geometry: mathematical foundations and applications*, 3rd ed., Wiley, Chichester, UK, 2014. MR Zbl

[Goodman 1957] A. W. Goodman, "Univalent functions and nonanalytic curves", *Proc. Amer. Math. Soc.* **8** (1957), 598–601. MR

[Goodman 1983a] A. W. Goodman, *Univalent functions, I*, Mariner, Tampa, FL, 1983. MR Zbl

[Goodman 1983b] A. W. Goodman, *Univalent functions, II*, Mariner, Tampa, FL, 1983. MR Zbl

[Kalmes et al. 2012] T. Kalmes, M. Nieß, and T. Ransford, "Examples of quantitative universal approximation", pp. 77–97 in *Complex analysis and potential theory* (Montréal, 2011), edited by

A. Boivin and J. Mashreghi, CRM Proc. Lecture Notes **55**, Amer. Math. Soc., Providence, RI, 2012. MR Zbl

[Littlewood 1925] J. E. Littlewood, "On inequalities in the theory of functions", *Proc. Lond. Math. Soc.* (2) **23**:7 (1925), 481–519. MR Zbl

[Pommerenke 1975] C. Pommerenke, *Univalent functions*, Stud. Math. **25**, Vandenhoeck & Ruprecht, Göttingen, Germany, 1975. MR Zbl

[Schober 1975] G. Schober, *Univalent functions*: *selected topics*, Lecture Notes in Math. **478**, Springer, 1975. MR Zbl

Received: 2019-09-04 Revised: 2019-12-11 Accepted: 2020-06-14

philippe.drouin.10@ulaval.ca *Département de mathématiques et de statistique,*
Université Laval, Quebec, QC, Canada

thomas.ransford@mat.ulaval.ca *Département de mathématiques et de statistique,*
Université Laval, Quebec, QC, Canada

INVOLVE 13:3 (2020)
https://doi.org/10.2140/involve.2020.13.433

Uniqueness of a three-dimensional stochastic differential equation

Carl Mueller and Giang Truong

(Communicated by Amarjit Singh Budhiraja)

In order to extend the study of the uniqueness property of multidimensional systems of stochastic differential equations, we look at the following three-dimensional system of equations, of which the two-dimensional case has been well studied: $dX_t = Y_t\,dt,\ dY_t = Z_t\,dt,\ dZ_t = |X_t|^\alpha\,dB_t$. We prove that if $(X_0, Y_0, Z_0) \neq (0,0,0)$ and $\frac{3}{4} < \alpha < 1$, then the system of equations has a unique solution in the strong sense.

1. Introduction and main results

The uniqueness of ordinary differential equations (ODEs) has been extensively studied; see for example [Hartman 1964]. In particular, if $F(u)$ is Lipschitz continuous, then

$$u'(t) = F(u(t)), \quad u(0) = u_0,$$

has a unique solution for all $t \geq 0$. In the case above, F, $u(t)$, and u_0 take values in $\mathbb{R}^d$, $d \geq 1$. The stochastic differential equation (SDE) realm, on the contrary, has different criteria for uniqueness of solutions; see for example [Protter 1990]. One of the most well-known results regarding strong uniqueness of SDEs is due to [Watanabe and Yamada 1971]. The result states that if $f(x)$ is locally Hölder continuous with index $\alpha \in \left[\frac{1}{2}, 1\right]$ and with linear growth, then

$$dX = f(X)\,dW, \quad X_0 = x_0,$$

has a unique strong solution for all times $t \geq 0$. Yamada and Watanabe's theory essentially focuses on one-dimensional SDEs.

MSC2010: primary 60H10; secondary 34F05.
Keywords: white noise, stochastic differential equations, uniqueness.
Mueller was supported by a Simons grant.

One motivation for studying higher-dimensional SDEs comes from the wave equation

$$\begin{aligned}\partial_t^2 u &= \Delta u,\\ u(0,x) &= u_0(x),\\ \partial_t u(0,x) &= u_1(x).\end{aligned}$$

In this equation, we have

$$\partial_t^2 u = \partial_x^2 u = \Delta u.$$

If we let

$$v = \partial_t u,$$

then we can rewrite the wave equation as the following system of equations:

$$\begin{aligned}\partial_t u &= v,\\ \partial_t v &= \Delta u.\end{aligned}$$

The original wave equation includes no noise. However, many physical systems are affected by noise. Hence, a modification of the wave equation which includes white noise is also studied:

$$\begin{aligned}\partial_t^2 u &= \Delta u + f(u)\dot{W},\\ u(0,x) &= u_0(x),\\ \partial_t u(0,x) &= u_1(x).\end{aligned} \tag{1}$$

Note $x \in \mathbb{R}$ and $\dot{W} = \dot{W}(t,x)$ is white noise.

One well-known point is that Lipschitz continuity is sufficient for the uniqueness of SDEs. Thus, many mathematicians have studied whether Hölder continuity can still ensure the uniqueness property of SDEs. Gomez, Lee, Mueller, Neuman, and Salins [Gomez et al. 2017] studied the uniqueness property of the following two-dimensional model of SDEs:

$$\begin{aligned}dX &= Y\,dt,\\ dY &= |X|^{\alpha}\,dB,\\ (X_0, Y_0) &= (x_0, y_0).\end{aligned} \tag{2}$$

The results focused on $f(x) = |x|^{\alpha}$ since it is a prototype of an equation with Hölder continuous coefficients. Moreover, (2) is a version of (1) when we drop the dependence on x, which allows us to study the modified wave equation with more simplicity. Notice that if we take the differential dY of the first derivative of X, which is Y in the system of equations, it resembles the second derivative in time in the stochastic wave equation.

It was proven in [Gomez et al. 2017] that if $\alpha > \frac{1}{2}$ and $(x_0, y_0) \neq (0, 0)$, then (2) has a unique solution in the strong sense up to the time τ at which (X_t, Y_t) first hits the origin $(0, 0)$.

Since much is still unknown about higher-dimensional SDEs, we wish to continue the study of the uniqueness property of (2) in the three-dimensional case, which is

$$\begin{aligned} dX &= Y\,dt, \\ dY &= Z\,dt, \\ dZ &= |X|^{\alpha} dB, \\ (X_0, Y_0, Z_0) &= (x_0, y_0, z_0). \end{aligned} \tag{3}$$

Theorem 1. *If* $\frac{3}{4} < \alpha < 1$ *and* $(X_0, Y_0, Z_0) \neq 0$, *then* (3) *has a unique strong solution, up to the time* τ *at which the solution* (X_t, Y_t, Z_t) *first hits the value* $(0, 0, 0)$ *or blows up.*

Moreover, we say the solution (X_t, Y_t, Z_t) blows up in finite time, with positive probability, if there is a random time $\tau < \infty$ such that

$$P\big(\lim_{t \uparrow \tau} |(X_t, Y_t, Z_t)|_{l^\infty} = \infty\big) > 0.$$

2. Proof of Theorem 1

Let (X_t^i, Y_t^i, Z_t^i), $i = 1, 2$, be two solutions to (3) with $(x_0, y_0, z_0) \neq (0, 0, 0)$; in other words, (X_t^i, Y_t^i, Z_t^i), $i = 1, 2$, have the same initial condition $(X_0, Y_0, Z_0) \neq (0, 0, 0)$. Note that since (X_t^i, Y_t^i, Z_t^i), $i = 1, 2$, have the same initial conditions, from now on, we use X_0, Y_0, Z_0 instead of $X_0^{i,n}$, $Y_0^{i,n}$, $Z_0^{i,n}$.

Let τ be the first time t that either (X_t^1, Y_t^1, Z_t^1) or (X_t^2, Y_t^2, Z_t^2) hits the origin $(0, 0, 0)$ or blows up. We let τ be infinity if there is no such time.

First, if $X_0 \neq 0$, then (3) will have Lipschitz continuity up to the time that $X_t = 0$, and thus enables uniqueness to hold. Suppose after a certain amount of time X_t hits zero, where Lipschitz continuity no longer holds; then due to the strong Markov property, we begin the process again with $X_0 = 0$. So our goal is to prove pathwise uniqueness between excursions of X up to the time τ starting from $X_0 = 0$.

For any fixed n, let τ_n be the first time that either

$$|(X_t^1, Y_t^1, Z_t^1)|_{l^\infty} \wedge |(X_t^2, Y_t^2, Z_t^2)|_{l^\infty} \leq 2^{-n}$$

or

$$|(X_t^1, Y_t^1, Z_t^1)|_{l^\infty} \vee |(X_t^2, Y_t^2, Z_t^2)|_{l^\infty} \geq 2^{n}.$$

Here, the infinity norm (also known as the L_∞-norm, l_∞-norm, max norm, or uniform norm) of a vector $\vec{v}$ is denoted by $|\vec{v}|_{l_\infty}$ and is defined as the maximum of the absolute values of its components,

$$|\vec{v}|_{l_\infty} = \max\{|v_i| : i = 1, 2, \dots, n\},$$

and

$$a \wedge b = \min(a, b),$$
$$a \vee b = \max(a, b).$$

If there is no such time, we let τ_n be infinity. Note that $\lim_{n\uparrow\infty}(\tau_n) = \tau$.

Now, for each fixed n, we will show uniqueness up to time τ_n in the system of equations

$$\begin{aligned} dX_t^{i,n} &= Y_t^{i,n}\,dt, \\ dY_t^{i,n} &= Z_t^{i,n}\,dt, \\ dZ_t^{i,n} &= |X_t^{i,n}|^\alpha\,\mathbf{1}_{[0,\tau_n]}(t)\,dB_t, \\ (X_0, Y_0, Z_0) &= (x_0, y_0, z_0). \end{aligned} \tag{4}$$

In other words, after the time τ_n, we have $dZ_t^{i,n} = 0$, which makes $Z_t^{i,n}$ become constant. Specifically, given $m, n \in N$, we need

$$(X_t^n, Y_t^n, Z_t^n) = (X_t^m, Y_t^m, Z_t^m)$$

for all $t \le \tau_n \wedge \tau_m$.

Now, before continuing the proof of uniqueness, by the method of contradiction, we show that the times that X_t hits zero do not accumulate before the time τ_n, almost surely.

For each n, let A_n be the event on which the times that $X_t^{i,n} = 0$, $i = 1$ or $i = 2$, accumulate before τ_n, and assume $P(A_n) > 0$. Then, on A_n, suppose σ_n is an accumulation point of the times t at which $X_t^{i,n} = 0$; i.e, there exists a sequence of times $\rho_{1,n} < \rho_{2,n} < \cdots$ that converges to σ_n, and $X_{\rho_{k,n}}^{i,n} = 0$. Hence, on A_n, $\lim_{k\to\infty} \rho_{k,n} = \sigma_n$.

We have $X_t^{i,n}$ is almost surely continuous, and that $X_{\rho_{k,n}}^{i,n} = 0$ on A_n, so

$$\lim_{\rho_{k,n}\to\sigma_n} X_{\rho_{k,n}}^{i,n} = X_{\sigma_n}^{i,n} = 0$$

on A_n.

Note that $dX_t^{i,n} = Y_t^{i,n}\,dt$, and $Y_t^{i,n}$ is almost surely continuous. So if $Y_{\sigma_n}^{i,n} \neq 0$ on A_n, then there exists a random interval $[\sigma_n(\omega) - \epsilon(\omega), \sigma_n(\omega)]$ of positive length for which $X_t^{i,n} \neq 0$ on $[\sigma_n(\omega) - \epsilon(\omega), \sigma_n(\omega)]$. This contradicts the hypothesis of $\rho_{k,n}$ converging to σ_n.

If $Y_{\sigma_n}^{i,n} = 0$, there are two cases,

$$\sigma_n \ge \tau_n \quad \text{and} \quad \sigma_n < \tau_n.$$

If $\sigma_n \ge \tau_n$, then it means that the times at which $X_t^{i,n}$ hit zero do not accumulate before τ_n.

If $\sigma_n < \tau_n$, then with $X^{i,n}_{\sigma_n} = 0$ and $Y^{i,n}_{\sigma_n} = 0$, we have $|Z^{i,n}_{\sigma_n}| > 2^{-n}$. Now suppose $Z_{\sigma^{i,n}_n} > 2^{-n}$, as the case $Z^{i,n}_{\sigma} < 2^{-n}$ is approached similarly due to symmetry.

If $Z^{i,n}_{\sigma_n} > 2^{-n}$, since Z_t is almost surely continuous, there exists a time interval $[\sigma_n(\omega)-\epsilon'(\omega), \sigma_n(\omega)]$ on which $Z^{i,n}_t > 2^{-n}/2$. Thus for all $t \in [\sigma_n(\omega)-\epsilon'(\omega), \sigma_n(\omega)]$ we almost surely have

$$Y^{i,n}_t = Y^{i,n}_{\sigma_n} - \int_t^{\sigma_n} Z^{i,n}_s \, ds < -\frac{2^{-n}}{2}(\sigma_n - t).$$

Hence, integrating over $X^{i,n}_t$ for $t \in [\sigma_n - \epsilon', \sigma_n]$, we have

$$\begin{aligned} X^{i,n}_t &= X^{i,n}_{\sigma_n} - \int_t^{\sigma_n} Y^{i,n}_s \, ds = -\int_t^{\sigma_n} Y^{i,n}_s \, ds > \int_t^{\sigma_n} \frac{2^{-n}}{2}(\sigma_n - s)\, ds \\ &= \frac{2^{-n}}{2}\Big(\sigma_n s - \frac{s^2}{2}\Big)\Big|_t^{\sigma_n} = \frac{2^{-n}}{4}(\sigma_n - t)^2 > 0 \end{aligned}$$

as $t < \sigma_n$. Hence, this contradicts the hypothesis of $\rho_{k,n}$ converging to σ_n on A_n. So in conclusion, $P(A_n) = 0$, which means the times at which X_t hits zero do not accumulate before the time τ_n.

One more point we need to address before continuing with the proof of uniqueness is the existence of solutions. This problem is resolved in Theorems 21.7 and 21.8 and Lemma 21.17 of [Kallenberg 2002], which prove that for all times $t \geq 0$, solutions of multidimensional SDEs exist with probability 1 provided the coefficients are continuous and bounded.

Specifically, in our problem, since $\alpha \in (0, 1)$, the coefficients of system (3) are continuous and bounded by $(2^n)^\alpha \vee 2^n = 2^n$ up to the τ_n for each n, which satisfies the condition stated in the existence theory in [Kallenberg 2002]. Hence, existence of solution holds for all $t \leq \tau_n$ for all n. Therefore, up to the time $\tau = \sup \tau_n$, existence of solutions is ensured.

From now on, $X^{i,n}_t$ means $X^{1,n}_t$ and $X^{2,n}_t$. We define $Y^{i,n}_t$ and $Z^{i,n}_t$ similarly. Back to the proof of uniqueness, we have, for all $t \in [0, \tau_n]$,

$$|Z^{i,n}_t| \vee |Y^{i,n}_t| \leq 2^n.$$

So

$$|Y^{i,n}_t| = \left| Y_0 + \int_0^t Z^{i,n}_s \, ds \right| \leq |Y_0| + \int_0^t |Z^{i,n}_s| \, ds \leq 2^n + 2^n t.$$

Therefore,

$$|X^{i,n}_t| \leq |X_0 + \int_0^t Y^{i,n}_s \, ds| \leq \int_0^t |Y^{i,n}_s| \, ds \leq \int_0^t (2^n + 2^n s)\, ds = 2^n\Big(t + \frac{t^2}{2}\Big).$$

Now let

$$t_{0,n} = \frac{2^{-2n}}{2}. \tag{5}$$

Then

$$t_{0,n}+\frac{t_{0,n}^2}{2}=\frac{2^{-2n}}{2}+\frac{2^{-2n}}{8}\le\frac{2^{-2n}}{2}+\frac{2^{-2n}}{2}=2^{-2n}.$$

So

$$2^n\left(t_{0,n}+\frac{t_{0,n}^2}{2}\right)\le 2^n\cdot 2^{-2n}=2^{-n}.$$

Since the quadratic function $2^n(t+t^2/2)$ is increasing when $t\ge 0$, we have

$$|X_t^{i,n}|\le 2^n\left(t+\frac{t^2}{2}\right)\le 2^{-n}$$

for all $t\in[0,t_{0,n}]$.

Since $|X_t^{1,n}|$ and $|X_t^{2,n}|$ belong in $[0,2^{-n}]$ for $t\in[0,t_{0,n}]$, based on the definition of τ_n above, either

$$|Y_0|\ge 2^{-n} \tag{6}$$

or

$$|Z_0|\ge 2^{-n} \tag{7}$$

for each fixed n. This is due to the fact that the solutions have the same initial condition (X_0, Y_0, Z_0) and for each time $t\in[0,t_{n,0}]$, either

$$|Y_t^{i,n}|\ge 2^{-n}\quad\text{or}\quad |Z_t^{i,n}|\ge 2^{-n}.$$

First, we deal with $Y_0>0$. Due to symmetry, we can deal with the case $Y_0<0$ with similar methods and thus omit the proof.

Now, with $Y_0>0$, we look at other subcases based on $Z_0^{i,n}$.

Case I: $Y_0>0$, $|Z_0|\le 2^{-n}$. If $|Z_0|\le 2^{-n}$, then (6) takes place. We are looking at the case $Y_0>0$, and thus $Y_0>2^{-n}$. Also, note that $dZ_t^{i,n}=0$ for all $t>\tau_n$ and $|Z_t^{i,n}|\le 2^n$ for all $t\in[0,\tau_n]$. Hence, $|Z_t^{i,n}|\le 2^n$ for all t, which means $Z_t^{i,n}\ge -2^n$ for all t. Next, we have

$$Y_t^{i,n}=Y_0+\int_0^t Z_s\,ds\ge 2^{-n}-\int_0^t 2^n\,ds=2^{-n}-2^n t.$$

If

$$0<t<t_{0,n}=\frac{2^{-2n}}{2},$$

where $t_{0,n}$ is defined as in (5), then

$$2^n t<\frac{2^{-n}}{2}.$$

Thus

$$2^{-n}-2^n t>\frac{2^{-n}}{2}$$

for all $t\in[0,t_{0,n}]$. In other words, $Y_t^{i,n}>2^{-n}/2$ for $t\in[0,t_{0,n}]$. So, for all $t\in[0,t_{0,n}]$, we have

$$Y_t^{i,n}\ge\frac{2^{-n}}{2}.$$

Hence

$$X_t^{i,n} \geq X_0^{i,n} + \int_0^t \frac{2^{-n}}{2}\, ds \geq \frac{2^{-n}}{2} t. \tag{8}$$

Furthermore, based on (8) and $|X_t^{i,n}| \leq 2^{-n}$ for all $t \in [0, t_{0,n}]$, it leads to $t_{0,n} \leq 2$, otherwise $X_t^{i,n} > 2^{-n}$, which means that $t > t_{0,n}$, a contradiction.

Note that

$$\begin{aligned} X_t^{i,n} &= X_0 + \int_0^t Y_s^{i,n}\, ds, \\ Y_s^{i,n} &= Y_0 + \int_0^s Z_k^{i,n}\, dk, \\ Z_k^{i,n} &= Z_0 + \int_0^k |X_r^{i,n}|^\alpha\, \mathbf{1}_{[0,\tau_n]}(t)\, dB_r. \end{aligned}$$

Thus

$$\begin{aligned} X_t^{i,n} &= X_0 + Y_0 t + \int_0^t \int_0^s \left(Z_0 + \int_0^k |X_r^{i,n}|^\alpha\, \mathbf{1}_{[0,\tau_n]}(t)\, dB_r \right) dk\, ds \\ &= X_0 + Y_0 t + \int_0^t \int_0^s Z_0\, dk\, ds + \int_0^t \int_0^s \int_0^k |X_r^{i,n}|^\alpha\, \mathbf{1}_{[0,\tau_n]}(t)\, dB_r\, dk\, ds \\ &= X_0 + Y_0 t + Z_0 \frac{t^2}{2} + \int_0^t \int_0^s \int_0^k |X_r^{i,n}|^\alpha\, \mathbf{1}_{[0,\tau_n]}(t)\, dB_r\, dk\, ds. \end{aligned}$$

Hence

$$(X_t^{1,n} - X_t^{2,n})^2 = \left(\int_0^t \int_0^s \int_0^k (|X_r^{1,n}|^\alpha - |X_r^{2,n}|^\alpha)\, \mathbf{1}_{[0,\tau_n]}(r)\, dB_r\, dk\, ds \right)^2.$$

Apply the Cauchy-Schwarz inequality twice, we get

$$\begin{aligned} (X_t^{1,n} - X_t^{2,n})^2 &\leq t \int_0^t \left(\int_0^s \int_0^k (|X_r^{1,n}|^\alpha - |X_r^{2,n}|^\alpha)\, \mathbf{1}_{[0,\tau_n]}(r)\, dB_r\, dk \right)^2 ds \\ &\leq t \int_0^t s \int_0^s \left(\int_0^k (|X_r^{1,n}|^\alpha - |X_r^{2,n}|^\alpha)\, \mathbf{1}_{[0,\tau_n]}(r)\, dB_r \right)^2 dk\, ds. \end{aligned}$$

Thus

$$E[(X_t^{1,n} - X_t^{2,n})^2] \leq t E \int_0^s s \int_0^s \left(\int_0^k (|X_r^{1,n}|^\alpha - |X_r^{2,n}|^\alpha)\, \mathbf{1}_{[0,\tau_n]}(r)\, dB_r \right)^2 dk\, ds.$$

By Itô's isometry,

$$\begin{aligned} E[(X_t^{1,n} - X_t^{2,n})^2] &\leq t E \int_0^t s \int_0^s \int_0^k ((|X_r^{1,n}|^\alpha - |X_r^{2,n}|^\alpha)\, \mathbf{1}_{[0,\tau_n]}(r))^2\, dr\, dk\, ds \\ &\leq t E \int_0^t t \int_0^s \int_0^k (|X_r^{1,n}|^\alpha - |X_r^{2,n}|^\alpha)^2\, dr\, dk\, ds \end{aligned}$$

$$\begin{aligned}
&\le t^2 E\int_0^t\int_0^s\int_0^k (|X_r^{1,n}|^\alpha - |X_r^{2,n}|^\alpha)^2\,dr\,dk\,ds\\
&\le t^2 E\int_0^t\int_0^t\int_0^t (|X_r^{1,n}|^\alpha - |X_r^{2,n}|^\alpha)^2\,dr\,dk\,ds\\
&= t^4 E\int_0^t (|X_r^{1,n}|^\alpha - |X_r^{2,n}|^\alpha)^2\,dr.
\end{aligned}$$

Now we apply the mean value theorem for the function $f(x) = x^\alpha,\ 0 < \alpha < 1$, and $a < b$:

$$b^\alpha - a^\alpha = \alpha c^{\alpha-1}(b-a) \le \alpha a^{\alpha-1}(b-a)$$

for $c \in (a, b)$. Then for $r \in [0, t_{0,n}]$, where t_0 is determined in (5), we apply (8):

$$\big||X_r^{1,n}|^\alpha - |X_r^{2,n}|^\alpha\big| \le \alpha\left(\frac{2^{-n}}{2}r\right)^{\alpha-1}\big||X_r^{1,n}| - |X_r^{2,n}|\big|. \tag{9}$$

Now let

$$D_t = E[(|X_t^{1,n}| - |X_t^{2,n}|)^2].$$

Since $t_{0,n} \le 2$, we have for all $t \in [0, t_{0,n}]$

$$D_t \le E[(X_t^{1,n} - X_t^{2,n})^2] \le C_n\int_0^t r^{2\alpha-2}D_r\,dr.$$

for some C_n depending on n. Since $\alpha > \frac{3}{4}$, we have $r^{2\alpha-2}$ is integrable on $[0, t_{0,n}]$. At this stage we apply Gronwall's lemma:

Lemma. *Let I denote an interval of the real line of the form $[a, \infty)$ or $[a, b]$ or $[a, b)$ with $a < b$. Let β and u be real-valued continuous functions defined on I. If u is differentiable on the interior I^o of I (the interval I without the endpoint a and possibly b) and satisfies the differential inequality*

$$u^{'}(t) \le \beta(t)u(t), \quad t \in I^o,$$

then u is bounded by the solution of the corresponding differential equation $v^{'}(t) = \beta(t)v(t)$:

$$u(t) \le u(a)\exp\left(\int_a^t \beta(s)\,ds\right)$$

for all $t \in I$.

Hence, with $D_0 = 0$, we have $D_t = 0$ for all $t \in [0, t_{0,n}]$. Therefore, (3) has unique strong solution in $[0, t_{0,n}]$.

Since $\alpha \ge \frac{3}{4}$, we have $2\alpha - 2 \ge -1$. Hence, $r^{2\alpha-2}$ is integrable on $[0, t_{0,n}]$ Note that in this case since $X_t \ge 2^{-n},\ \eta \le 1$. Applying Gronwall's lemma, with $D_0 = 0$, we have $D_t = 0$ for all $t \in [0, t_{0,n}]$. Therefore, (3) has a unique solution in strong sense up to time $t_{0,n}$.

Note that for all $t \in [0, t_{0,n}]$, we have $Y_t^{i,n} > 2^{-n}/2 > 0$. Hence $X_t^{i,n}$ is strictly increasing, which, leads to $X_{t_{0,n}}$ strictly positive. Therefore, by the strong Markov property, we have uniqueness until the time X next hits zero.

Case II: $Y_0 > 0,\ Z_0 < -2^{-n}$. Since Z_0 starts negative, $Y_t^{i,n}$ decreases for an amount of time. Since Y_0 is positive, let say $Y_0 = \beta > 0$. Note that for all t, we have $|Z_t^{i,n}| < 2^n$, which means $Z_t^{i,n} > -2^n$. First, we have

$$Y_t^{i,n} = \beta + \int_0^t Z_s \, ds \geq \beta - \int_0^t 2^n \, ds = \beta - 2^n t.$$

Let

$$t_{0,n}^{'} = \frac{\beta}{2^{n+1}}. \tag{10}$$

If

$$0 < t < t_{0,n}^{'} = \frac{\beta}{2^{n+1}},$$

then

$$2^n t < \frac{\beta}{2};$$

thus

$$\beta - 2^n t > \frac{\beta}{2}$$

for all $t \in [0, t_{0,n}{'}]$. In other words, $Y_t^{i,n} > \beta/2$ for all $t \in [0, t_{0,n}^{'}]$.

So, for all $t \in [0, t_{0,n} \wedge t_{0,n}^{'}]$, where $t_{0,n}$ and $t_{0,n}^{'}$ are determined in (5) and (10) respectively, we have

$$Y_t^{i,n} \geq \frac{\beta}{2}$$

Hence

$$X_t^{i,n} \geq X_0^{i,n} + \int_0^t \frac{\beta}{2} \, ds \geq \frac{\beta}{2} t.$$

Applying the same method (9) above, we use the mean value theorem for the new lower bound of $X_t^{i,n}$:

$$\big| |X_r^{1,n}|^\alpha - |X_r^{2,n}|^\alpha \big| \leq \alpha \left(\frac{\beta}{2} r \right)^{\alpha - 1} \big| |X_r^{1,n}| - |X_r^{2,n}| \big|.$$

Hence

$$D_t \leq E[(X_t^{1,n} - X_t^{2,n})^2] \leq C_n \int_0^t r^{2\alpha - 2} D_r \, dr.$$

Again, applying Gronwall's lemma, with $D_0 = 0$, we have $D_t = 0$ for all $t \in [0, t_{0,n} \wedge t_{0,n}^{'}]$. Therefore, (3) has a unique strong solution in $[0, t_{0,n} \wedge t_{0,n}^{'}]$. As in the previous cases, we have $Y_t^{i,n} > \beta/2 > 0$ for all $t \in [t_{0,n} \wedge t_{0,n}^{'}]$, which makes $X_t^{i,n}$ strictly increasing. So $X_{t_{0,n}} \wedge t_{0,n}^{'}$ is strictly positive. Thus, by the strong Markov property, we have uniqueness until the next time X hits zero.

Case III: $Y_0 > 0$, $Z_0 > 2^{-n}$. Now we let T_n be the first time that either $Z_t^{1,n}$ or $Z_t^{2,n}$ hits the value $2^{-n}/2$. Since both $Z_t^{1,n}$ and $Z_t^{2,n}$ are continuous, we have $T_n > 0$ with probability 1. So we now prove uniqueness up to the time $t_{0,n} \wedge T_n$, where $t_{0,n}$ is defined in (5).

Then for all t in $[0, t_{0,n} \wedge T_n]$, we have

$$Z_t^{i,n} \geq \frac{2^{-n}}{2};$$

therefore

$$Y_t^{i,n} \geq Y_0 + \int_0^t \frac{2^{-n}}{2}\, ds \geq \frac{2^{-n}}{2} t \tag{11}$$

since $Y_0^{i,n} \geq 0$.

Based on (11), for $t \in [0, t_{0,n} \wedge T_n]$

$$X_t^{i,n} \geq X_0 + \int_0^t \frac{2^{-n}}{2} s\, ds \geq \frac{2^{-n}}{4} t^2. \tag{12}$$

Now we define

$$X_t^{i,n} = \widetilde{X}_t^{i,n}, \quad Y_t^{i,n} = \widetilde{Y}_t^{i,n}, \quad Z_t^{i,n} = \widetilde{Z}_t^{i,n}$$

for $i = 1, 2$ and for $t \leq \tau_n \wedge\ T_n \wedge t_{0,n}$, where $t_{0,n}$ is defined as in (5) above.

Thus the following system of equations holds up to the stopping time $\tau_n \wedge T_n \wedge t_{0,n}$:

$$\begin{aligned} d\widetilde{X}_t^{i,n} &= \widetilde{Y}_t^{i,n}\, dt \\ d\widetilde{Y}_t^{i,n} &= \widetilde{Z}_t^{i,n}\, dt \\ d\widetilde{Z}_t^{i,n} &= |\widetilde{X}_t^{i,n}|^\alpha\, \mathbf{1}_{[0,\tau_n \wedge T_n \wedge t_{0,n}]}(t)\, dB_t, \end{aligned} \tag{13}$$

with $(\widetilde{X}_0^{i,n}, \widetilde{Y}_0^{i,n}, \widetilde{Z}_0^{i,n}) = (X_0, Y_0, Z_0)$ for $i = 1, 2$. Furthermore, using (13), $\widetilde{X}_t^{i,n}$, $\widetilde{Y}_t^{i,n}$, and $\widetilde{Z}_t^{i,n}$ can be defined for all times.

Using Itô's isometry as above with $\widetilde{X}_t^{i,n}$, $\widetilde{Y}_t^{i,n}$, and $\widetilde{Z}_t^{i,n}$,

$$\begin{aligned} E[(\widetilde{X}_t^{1,n} - \widetilde{X}_t^{2,n})^2] &\leq tE \int_0^t s \int_0^s \int_0^k \big((|\widetilde{X}_r^{1,n}|^\alpha - |\widetilde{X}_r^{2,n}|^\alpha)\, \mathbf{1}_{[0,\tau_n \wedge T_n \wedge t_{0,n}]}(r)\big)^2\, dr\, dk\, ds \\ &\leq tE \int_0^t t \int_0^s \int_0^k (|\widetilde{X}_r^{1,n}|^\alpha - |\widetilde{X}_r^{2,n}|^\alpha)^2\, \mathbf{1}_{[0,\tau_n \wedge T_n \wedge t_{0,n}]}(r)\, dr\, dk\, ds \\ &\leq t^2 E \int_0^t \int_0^s \int_0^k (|\widetilde{X}_r^{1,n}|^\alpha - |\widetilde{X}_r^{2,n}|^\alpha)^2\, \mathbf{1}_{[0,\tau_n \wedge T_n \wedge t_{0,n}]}(r)\, dr\, dk\, ds \\ &\leq t^2 E \int_0^t \int_0^t \int_0^t (|\widetilde{X}_r^{1,n}|^\alpha - |\widetilde{X}_r^{2,n}|^\alpha)^2\, \mathbf{1}_{[0,\tau_n \wedge T_n \wedge t_{0,n}]}\, dr\, dk\, ds \\ &= t^4 E \int_0^t (|\widetilde{X}_r^{1,n}|^\alpha - |\widetilde{X}_r^{2,n}|^\alpha)^2\, \mathbf{1}_{[0,\tau_n \wedge T_n \wedge t_{0,n}]}\, dr. \end{aligned}$$

Using (12) and the mean value theorem, for $r \in [0, \tau_n \wedge T_n \wedge t_{0,n}]$, we have

$$\big||\widetilde{X}_r^{1,n}|^\alpha - |\widetilde{X}_r^{2,n}|^\alpha\big| \le \alpha\left(\frac{2^{-n}}{4}r^2\right)^{\alpha-1} \big||\widetilde{X}_r^{1,n}| - |\widetilde{X}_r^{2,n}|\big|.$$

Hence

$$E[(\widetilde{X}_t^{1,n} - \widetilde{X}_t^{2,n})^2] \le t^4\alpha^2\left(\frac{2^{-n}}{4}\right)^{2(\alpha-1)} E\int_0^t r^{4(\alpha-1)}(|\widetilde{X}_r^{1,n}| - |\widetilde{X}_r^{2,n}|)^2\,dr,$$

so if we let

$$D_t = E[(|\widetilde{X}_t^{1,n}| - |\widetilde{X}_t^{2,n}|)^2],$$

then

$$D_t \le E[(\widetilde{X}_t^{1,n} - \widetilde{X}_t^{2,n})^2] \le C_n \int_0^t r^{4\alpha-4} D_r\,dr.$$

Again, applying Gronwall's lemma, with $D_0 = 0$, we have $D_t = 0$. Note that at the time $t_{0,n} \wedge T_n$, since we have $Z_t^{i,n} > 0$ for all $t \in [0, t_{0,n} \wedge T_n]$, and also $Y_0 > 0$, it leads to $Y_t^{i,n} > 0$ for all $t \in [0, t_{0,n} \wedge T_n]$. Thus $X_t^{i,n}$ is strictly increasing, which means $X_{t_{0,n}\wedge T_n}$ must be strictly greater than zero. Therefore, by the strong Markov property, we obtain uniqueness of the process until X next hits zero.

Case IV: $Y_0 = 0$. If $Y_0 = 0$, then based on the definition of τ_n, we have $|Z_0| > 2^{-n}$. We will first deal with the case $Z_0 > 2^{-n}$, and the case $Z_0 < 2^{-n}$ is approached the same way due to symmetry. As in Case III, let T_n be the first time that either $Z_t^{1,n}$ or $Z_t^{2,n}$ hits the value $2^{-n}/2$. Due to the continuity of $Z_t^{1,n}$ and $Z_t^{2,n}$, we have $T_n > 0$ with probability 1. So with for all $t \in [0, t_0 \wedge T_n]$, where t_0 is determined in (5), we have

$$Y_t^{i,n} \ge Y_0 + \int_0^t \frac{2^{-n}}{2}\,ds = \frac{2^{-n}}{2}t.$$

Then

$$X_t^{i,n} \ge X_0 + \int_0^t \frac{2^{-n}}{2}s\,ds \ge \frac{2^{-n}}{4}t^2.$$

We now apply the same method as in Case III by looking at $\widetilde{X}_t^{i,n}$, $\widetilde{Y}_t^{i,n}$, and $\widetilde{Z}_t^{i,n}$, which are defined as

$$X_t^{i,n} = \widetilde{X}_t^{i,n}, \quad Y_t^{i,n} = \widetilde{Y}_t^{i,n}, \quad Z_t^{i,n} = \widetilde{Z}_t^{i,n}$$

for $i = 1, 2$ and for $t \le \tau_n \wedge\ T_n \wedge t_{0,n}$, as $t_{0,n}$ defined as in (5) above.

Thus the following system of equations holds up to the stopping time $\tau_n \wedge T_n \wedge t_{0,n}$:

$$\begin{aligned} d\widetilde{X}_t^{i,n} &= \widetilde{Y}_t^{i,n}\,dt \\ d\widetilde{Y}_t^{i,n} &= \widetilde{Z}_t^{i,n}\,dt \\ d\widetilde{Z}_t^{i,n} &= |\widetilde{X}_t^{i,n}|^\alpha\,\mathbf{1}_{[0,\tau_n\wedge T_n\wedge t_{0,n}]}(t)\,dB_t, \end{aligned}$$

with $(\widetilde{X}_0^{i,n}, \widetilde{Y}_0^{i,n}, \widetilde{Z}_0^{i,n}) = (X_0, Y_0, Z_0)$ for $i = 1, 2$. Furthermore, using (13), $\widetilde{X}_t^{i,n}$, $\widetilde{Y}_t^{i,n}$, $\widetilde{Z}_t^{i,n}$ can be defined for all time.

Again, using the same strategy in Case III and the mean value theorem, we have

$$\big||\widetilde{X}_r^{1,n}|^\alpha - \widetilde{X}_r^{2,n}|^\alpha\big| \le \alpha\left(\frac{2^{-n}}{4}r^2\right)^{\alpha-1}\big||\widetilde{X}_r^{1,n}| - |\widetilde{X}_r^{2,n}|\big|.$$

Hence, if we let

$$D_t = E[(|\widetilde{X}_t^{1,n}| - |\widetilde{X}_t^{2,n}|)^2],$$

then

$$D_t \le E[(\widetilde{X}_t^{1,n} - \widetilde{X}_t^{2,n})^2] \le C_n \int_0^t r^{4\alpha-4} D_r\, dr.$$

Gronwall's lemma with $D_0 = 0$ yields $D_t = 0$, completing the proof of Theorem 1.

In this case, we also have $Y_t^{i,n} > 2^{-n}/2t > 0$ for all $t \in [0, t_{0,n} \wedge T_n]$. Hence $X_t^{i,n}$ is strictly increasing, which yields $X_{t_{0,n}\wedge T_n}$ strictly positive. So, by the strong Markov property, we have uniqueness up to the time X next hits zero.

Now with uniqueness proved, we actually can even strengthen the proof by showing that $\tau_n^1 = \tau_n^2$ for all n, where τ_n^1 and τ_n^2 respectively stand for the stopping times at the critical values for X_t^1 and X_t^2. Without loss of generality, suppose $\tau_n^2 > \tau_n^1$: So at the time τ_n^1, X_t^2 has not yet reached the critical values, which are 2^{-n} or 2^n, as stated above. But since we have uniqueness up to $\tau_n^1 \wedge \tau_n^2$, this implies X_t^1 has also not reached the critical value at the time τ_n^1, which is a contradiction to the definition of τ_n^1. Hence, τ_n^1 and τ_n^2 must be equal.

References

[Gomez et al. 2017] A. Gomez, J. J. Lee, C. Mueller, E. Neuman, and M. Salins, "On uniqueness and blowup properties for a class of second order SDEs", *Electron. J. Probab.* **22** (2017), art. id. 72. MR Zbl

[Hartman 1964] P. Hartman, *Ordinary differential equations*, Wiley, New York, 1964. MR Zbl

[Kallenberg 2002] O. Kallenberg, *Foundations of modern probability*, 2nd ed., Springer, 2002. MR Zbl

[Protter 1990] P. Protter, *Stochastic integration and differential equations: a new approach*, Appl. Math. (NY) **21**, Springer, 1990. MR Zbl

[Watanabe and Yamada 1971] S. Watanabe and T. Yamada, "On the uniqueness of solutions of stochastic differential equations, II", *J. Math. Kyoto Univ.* **11** (1971), 553–563. MR Zbl

Received: 2019-10-03 Revised: 2020-02-18 Accepted: 2020-05-23

carl.e.mueller@rochester.edu *Department of Mathematics, University of Rochester,*
Rochester, NY, United States

gtruong@u.rochester.edu *Department of Mathematics, University of Rochester,*
Rochester, NY, United States

INVOLVE 13:3 (2020)
https://doi.org/10.2140/involve.2020.13.445

Sharp sectional curvature bounds and a new proof of the spectral theorem

Maxine Calle and Corey Dunn

(Communicated by Frank Morgan)

We algebraically compute all possible sectional curvature values for canonical algebraic curvature tensors and use this result to give a method for constructing general sectional curvature bounds. We use a well-known method to geometrically realize these results to produce a hypersurface with prescribed sectional curvatures at a point. By extending our methods, we give a relatively short proof of the spectral theorem for self-adjoint operators on a finite-dimensional real vector space.

1. Introduction

Let V be a real vector space of finite dimension n, and let $V^* = \operatorname{Hom}(V, \mathbb{R})$ be the corresponding dual space of $\mathbb{R}$-linear maps from V to $\mathbb{R}$. An *algebraic curvature tensor* $R \in \otimes^4 V^*$ satisfies the properties below for all $x, y, z, w \in V$:

$$\begin{aligned} R(x, y, z, w) &= -R(y, x, z, w) = R(z, w, x, y), \\ R(x, y, z, w) + R(x, w, y, z) &+ R(x, z, w, y) = 0. \end{aligned} \tag{1.a}$$

Let $\mathcal{A}(V)$ denote the set of all algebraic curvature tensors. Let $S^2(V^*)$ denote the space of all symmetric bilinear forms over V, and let $\varphi \in S^2(V^*)$. Define a *canonical algebraic curvature tensor* $R_\varphi \in \mathcal{A}(V)$ as

$$R_\varphi(x, y, z, w) = \varphi(x, w)\varphi(y, z) - \varphi(x, z)\varphi(y, w).$$

If the vector space V is endowed with a nondegenerate inner product $\langle \cdot\,, \cdot \rangle$, then the triple $(V, \langle \cdot\,, \cdot \rangle, R)$ is referred to as a *model space*. In this research, all inner products are assumed to be positive definite. Throughout this work, we consider the vector space V and a positive definite inner product $\langle \cdot\,, \cdot \rangle$ to be given, and so we do not always specifically reference them. For the sake of convenience, we will refer to properties of a model space (such as sectional curvature defined below) as a property of an algebraic curvature tensor when there is no possibility of confusion.

MSC2010: primary 15A69; secondary 15A63, 53C21.
Keywords: sectional curvature, canonical algebraic curvature tensor, spectral theorem.

Let $\mathrm{Gr}_2(V)$ be the set of all 2-planes through the origin in V; this topological space is known as the *Grassmannian* of 2-planes in V. This space is compact and connected [Milnor and Stasheff 1974]. Given a model space $(V, \langle\cdot,\cdot\rangle, R)$ and a 2-plane $\pi \in V$, define the *sectional curvature* $\kappa(\pi)$ of π to be

$$\frac{R(x, y, y, x)}{\langle x, x\rangle\langle y, y\rangle - \langle x, y\rangle^2},$$

where $\pi = \operatorname{span}\{x, y\}$. This quantity is independent of the basis chosen for π.

Interest in algebraic curvature tensors stems from basic results in differential geometry. If (M, g) is a smooth manifold, one may compute the Riemann curvature tensor R_P at the point $P \in M$ using the Levi-Civita connection. It is well known that R_P satisfies the identities listed in (1.a), and thus R_P is an algebraic curvature tensor on the tangent space $T_P M$. Using the metric g_P restricted to $T_P M$, $(T_P M, g_P, R_P)$ is a model space. The converse is also true [Gilkey 2007]: given a model space $(V, \langle\cdot,\cdot\rangle, R)$, there exists a smooth manifold (M, g), a point $P \in M$, and a vector space isometry $\Psi : V \to T_P M$ so that $\Psi^* R_P = R$. The process of finding such a manifold is generally referred to as a geometric realization.

A general line of questioning is to explore exactly what one can say about a manifold that is a geometric realization of any given model space(s). In this way, the algebraic properties of the model space can influence the geometry of a geometric realization of it. Although there are many examples of this [Brozos Vázquez et al. 2012; García-Río et al. 2002; Gilkey and Nikčević 2005; Kowalski and Prüfer 1994; Tricerri and Vanhecke 1981; 1986; Tsankov 2005], we give two examples relevant to our study. The first example concerns the property of constant sectional curvature: up to local isometry, there is only one way to construct a manifold such that the model space at any of its points has constant sectional curvature.

The second example begins with the fact [Gilkey 2007] that $\{R_\varphi \mid \varphi \in S^2(V^*)\}$ is spanning set for the space of algebraic curvature tensors. For $R \in \mathcal{A}(V)$, define

$$\nu(R) = \min\bigl\{k \bigm| R = \textstyle\sum_{i=1}^k \alpha_i R_{\varphi_i}\bigr\}, \qquad \nu(n) = \max\{\nu(R) \mid R \in \mathcal{A}(V)\}.$$

This notation seems to have originated in [Gilkey 2007] and has since been studied by several authors. Díaz-Ramos and García-Río [2004] showed that $\nu(3) = 2$ and used the Nash embedding theorem [1956] to prove $\nu(n) \le n(n+1)/2$. In this way, $\nu(R)$ functions as a lower bound of the codimension of any local isometric embedding of a given manifold, where R is the algebraic curvature tensor at a given point. For this reason, subsequent work [Diaz and Dunn 2010] has aimed to discover linear dependencies in the set of canonical algebraic curvature tensors.

We have two major goals for this research. First, we establish sharp bounds on the sectional curvature values of any canonical algebraic curvature tensor. We interpret these bounds through a geometric realization result, and give a method for constructing bounds on the sectional curvature values of any algebraic curvature tensor. By an extension of our methods (inspired by R. Klinger [1991]), we meet our second goal: to give a short and self-contained proof of the spectral theorem.

More specifically, after some preliminary comments, in Section 2 we establish:

Theorem 1.1. *Let $\varphi \in S^2(V^*)$, and let $\lambda_1, \dots, \lambda_n$ be the eigenvalues of φ, repeated according to multiplicity. Let m and M, respectively, be the minimum and maximum of the set $\{\lambda_i\lambda_j \mid i \neq j\}$. The set of sectional curvatures of R_φ is precisely the interval $[m, M]$.*

We use this result to prove two corollaries. The first corollary (Corollary 2.3) uses a well-known result to geometrically realize any interval as the set of sectional curvatures of a manifold at a point. The manifolds we produce are hypersurfaces in Euclidean space. The second corollary (Corollary 2.4) provides bounds (which are not sharp, see Remarks 2.5 and 2.6) on the set of sectional curvatures of an arbitrary algebraic curvature tensor R in terms of $\nu(R)$ and the results from Theorem 1.1.

In Section 3 we consider a canonical algebraic curvature tensor to provide a short and self-contained proof of the spectral theorem:

Theorem 1.2 (the spectral theorem). *Let V be an inner product space. If $\varphi \in S^2(V^*)$, then there exists an orthonormal basis $\{f_1, \dots, f_n\}$ for V for which $\varphi(f_i, f_j) = \lambda_i\delta_{ij}$.*

2. Sectional curvature bounds

It is well known that given $\varphi \in S^2(V^*)$, there exists a self-adjoint linear map $A : V \to V$ characterized by the equation $\varphi(x, y) = \langle Ax, y\rangle$. This follows as an application of the Riesz representation theorem (see page 188 of [Axler 2015]), and we briefly pause to describe this correspondence.

If a linear map $A : V \to V$ is given, then one can define the bilinear form $\varphi(x, y) = \langle Ax, y\rangle$ for vectors $x, y \in V$. If, in addition, A is self-adjoint with respect to the inner product (i.e., $\langle Ax, y\rangle = \langle x, Ay\rangle$), then φ is symmetric. Conversely, if $\varphi \in S^2(V^*)$ is given, one could construct the (unique) self-adjoint linear map A satisfying $\varphi(x, y) = \langle Ax, y\rangle$ in the following way. First, choose an orthonormal basis $\mathcal{B} = \{f_1, \dots, f_n\}$ for V. Then compute the numbers $a_{ji} = \langle Af_i, f_j\rangle$. The matrix for A relative to this orthonormal basis is $[A]_{\mathcal{B}} = [a_{ij}]$.

Because of the bijective correspondence between symmetric bilinear forms and their corresponding self-adjoint linear maps, relevant aspects of φ are defined in terms of those same aspects in A, for example, $\ker(\varphi)$ and $\operatorname{Rank}(\varphi)$ are defined as $\ker(A)$ and $\operatorname{Rank}(A)$, respectively.

Similarly, eigenvalues and eigenvectors of a linear map $A : V \to V$ can related to the corresponding bilinear form φ. It will be helpful to do so in reference to an orthonormal basis: as described above, if $\{f_1, \dots, f_n\}$ is an orthonormal basis for V, the (j,i) matrix entry of A on this basis is equal to $\varphi(f_i, f_j)$. For this reason, f_i is an eigenvector with corresponding eigenvalue λ_i if and only if $\varphi(f_i, f_j) = \lambda_i\delta_{ij}$.

It is through this perspective that we establish our results. We will repeatedly use the following lemma; our proof uses a technique adapted from [Klinger 1991]. Translated into our terminology, the objective in that paper is to find a basis (called a Chern basis) for V that minimizes the number of nonzero entries of a given (arbitrary) algebraic curvature tensor. This is accomplished by first considering a 2-plane π with extremal sectional curvature, and then rotating this plane around an axis (within the plane) by an angle of θ to create the rotated plane π_θ, with $\pi_0 = \pi$. The sectional curvature function has an extremal value at $\theta = 0$, and so this is a critical point of the function $\theta \mapsto \kappa(\pi_\theta)$. Using the fact that the derivative of this function at $\theta = 0$ vanishes, one can uncover information about the components of the algebraic curvature tensor. We have adapted this technique (which seems to be an extension of the work in [Bishop and Goldberg 1964]) to apply to the algebraic curvature tensor R_φ to diagonalize the symmetric bilinear form φ.

Lemma 2.1. *Let $\varphi \in S^2(V^*)$, and suppose $R_\varphi \neq 0$. If π is a 2-plane whose sectional curvature is extremal, then there exists an orthonormal basis of eigenvectors for π.*

Proof. Since $R_\varphi \neq 0$, and R_φ is determined by its sectional curvatures [Lee 1997], there exists a 2-plane π of extremal nonzero sectional curvature. Let $\{e_1, e_2\}$ be any orthonormal basis for π. Let $\varphi|_\pi$ be the restriction of φ to π. We now diagonalize[1] the symmetric form $\varphi|_\pi$ to create a new orthonormal basis $\{f_1, f_2\}$ for π. For some θ, set

$$\begin{aligned} f_1 &= \cos\theta e_1 - \sin\theta e_2, \\ f_2 &= \sin\theta e_1 + \cos\theta e_2, \end{aligned}$$

where we will determine θ presently. Let $\varphi_{ij} = \varphi(e_i, e_j)$. We compute

$$\begin{aligned} \varphi(f_1, f_2) &= \cos\theta\sin\theta(\varphi_{11} - \varphi_{22}) + (\cos^2\theta - \sin^2\theta)\varphi_{12} \\ &= \tfrac{1}{2}\sin(2\theta)(\varphi_{11} - \varphi_{22}) + \cos(2\theta)\varphi_{12}. \end{aligned}$$

Choose θ so that $\varphi(f_1, f_2) = 0$. Explicitly, if $\varphi_{12} = 0$ already, then $\theta = 0$. Otherwise,

$$\theta = \tfrac{1}{2}\operatorname{arccot}\left(\frac{\varphi_{22} - \varphi_{11}}{2\varphi_{12}}\right).$$

Now extend this basis for π (in an arbitrary way) to form an orthonormal basis $\{f_1, f_2, f_3, \dots, f_n\}$ for V. Note that on this basis,

$$\begin{aligned} &\varphi(f_1, f_2) = 0, \\ &R_\varphi(f_1, f_2, f_2, f_1) = \varphi(f_1, f_1)\varphi(f_2, f_2) \quad \text{is extremal and nonzero.} \end{aligned} \tag{2.a}$$

[1] One could of course use the spectral theorem here, but since we use this result in our proof of the spectral theorem in the next section, we instead establish this directly.

We now show that $\varphi(f_1, f_j) = \varphi(f_2, f_j) = 0$ for $j \geq 3$, which will complete the proof. Choose any index $j \geq 3$, and consider the plane

$$\pi_\theta = \operatorname{span}\{\cos\theta f_1 + \sin\theta f_j,\, f_2\}.$$

Using a double angle formula and various curvature identities listed in the Introduction,

$$\kappa(\pi_\theta) = \cos^2\theta R_\varphi(f_1, f_2, f_2, f_1) + \sin^2\theta R_\varphi(f_j, f_2, f_2, f_j) \\ + \sin(2\theta) R_\varphi(f_1, f_2, f_2, f_j).$$

Since π_0 is extremal, 0 is a critical point of the function $\theta \mapsto \kappa(\pi_\theta)$, so

$$0 = \frac{d}{d\theta}[\kappa(\pi_\theta)]|_{\theta=0} = 2R_\varphi(f_1, f_2, f_2, f_j).$$

As a result,

$$0 = R_\varphi(f_1, f_2, f_2, f_j) = \varphi(f_1, f_j)\varphi(f_2, f_2) - \varphi(f_1, f_2)\varphi(f_2, f_j) \\ = \varphi(f_1, f_j)\varphi(f_2, f_2).$$

By (2.a), we know $\varphi(f_2, f_2) \neq 0$; hence $\varphi(f_1, f_j) = 0$. We can prove that $\varphi(f_2, f_j) = 0$ in a similar way by considering $\operatorname{span}\{f_1, \cos\theta f_2 + \sin\theta f_j\}$. □

Remark 2.2. Lemma 2.1 relates to Euler's theorem, which states (among other things) that at any point on a 2-dimensional manifold, the so-called principal directions are orthogonal and are eigenvectors of a certain bilinear form (provided these eigenvalues are different). This establishes Lemma 2.1 in dimension 2: the curvature tensor of a 2-dimensional manifold is of the form R_φ, where φ is this bilinear form.

We use Lemma 2.1 to establish Theorem 1.1.

Proof of Theorem 1.1. We start by proving that the maximal sectional curvature is M, the largest pairwise product of eigenvalues of φ. Since $\mathrm{Gr}_2(V)$ is compact, there exists a 2-plane Π for which

$$\kappa(\Pi) = \sup\{\kappa(L) \mid L \in \mathrm{Gr}_2(V)\}.$$

We proceed in cases: either $\kappa(\Pi) \neq 0$ or $\kappa(\Pi) = 0$. Our goal is to show $\kappa(\Pi)$ to be a pairwise product of eigenvalues. Thus, as the maximal sectional curvature value, $\kappa(\Pi) = M$.

If $\kappa(\Pi) \neq 0$, then we may use Lemma 2.1 to produce an orthonormal basis of eigenvectors $\{f_1, f_2\}$ for Π. In this case,

$$\kappa(\Pi) = R_\varphi(f_1, f_2, f_2, f_1) = \varphi(f_1, f_1)\varphi(f_2, f_2)$$

is a product of eigenvalues and hence is equal to M.

Now suppose $\kappa(\Pi) = 0$. Find an orthonormal basis $\{f_1, \dots, f_n\}$ diagonalizing φ, and note

$$\kappa(\operatorname{span}\{f_i, f_j\}) = R_\varphi(f_i, f_j, f_j, f_i) = \varphi(f_i, f_i)\varphi(f_j, f_j)$$

is a product of eigenvalues of φ. It is not possible that there are two nonzero eigenvalues of the same sign since their corresponding eigenvectors would span a 2-plane of positive sectional curvature which is contrary to our assumption that $\kappa(\Pi) = 0$ is the maximal sectional curvature. Since $\dim(V) \geq 3$, we know 0 is an eigenvalue, and therefore $0 = \kappa(\Pi) = M$ is the largest pairwise product of eigenvalues as desired.

We now prove that the minimal sectional curvature is m, the smallest pairwise product of eigenvalues of φ. Proceeding similarly, there exists a 2-plane π for which

$$\kappa(\pi) = \inf\{\kappa(L) \mid L \in \mathrm{Gr}_2(V)\}.$$

We again proceed in cases: either $\kappa(\pi) \neq 0$ or $\kappa(\pi) = 0$. If $\kappa(\pi) \neq 0$ then use Lemma 2.1 once more to produce an orthonormal basis of eigenvectors $\{f_1, f_2\}$ for π. Then

$$\kappa(\pi) = R_\varphi(f_1, f_2, f_2, f_1) = \varphi(f_1, f_1)\varphi(f_2, f_2)$$

is a product of eigenvalues and is hence equal to m.

Now suppose $\kappa(\pi) = 0$ and again diagonalize $\varphi|_\pi$ with the basis $\{f_1, f_2\}$ and extend it to an orthonormal basis $\{f_1, \dots, f_n\}$ for V. We then have

$$0 = \kappa(\pi) = R_\varphi(f_1, f_2, f_2, f_1) = \varphi(f_1, f_1)\varphi(f_2, f_2).$$

Thus, one of the two factors above is zero. As exchanging the two vectors would not change the span or disrupt the diagonalization, we may assume $\varphi(f_1, f_1) = 0$. For $j \geq 3$, since 0 is assumed to be the minimal sectional curvature, we find

$$\begin{aligned} 0 \leq \kappa(\operatorname{span}\{f_1, f_j\}) &= R_\varphi(f_1, f_j, f_j, f_1) \\ &= \varphi(f_1, f_1)\varphi(f_j, f_j) - \varphi(f_1, f_j)^2 \\ &= -\varphi(f_1, f_j)^2. \end{aligned}$$

It follows that $\varphi(f_1, f_j) = 0$ for all $j \neq 1$, and so f_1 is an eigenvector corresponding to the eigenvalue $\varphi(f_1, f_1) = 0$. Thus, 0 is an eigenvalue of φ.

To finish this portion of the proof, we prove that all other eigenvalues of φ are either zero or have the same sign, which would show 0 to be the smallest pairwise product of the eigenvalues. To do this, find an orthonormal basis $\{e_1, \dots, e_n\}$ for V that diagonalizes φ. Since 0 is minimal, we must have

$$0 \leq \kappa(\operatorname{span}\{e_i, e_j\}) = R_\varphi(e_i, e_j, e_j, e_i) = \varphi(e_i, e_i)\varphi(e_j, e_j).$$

For this reason the nonzero eigenvalues of φ cannot differ in sign, and this portion of the proof is complete.

We have proven that $\kappa(\Pi) = M$ and $\kappa(\pi) = m$ are, respectively, the maximal and minimal sectional curvatures. Since $\mathrm{Gr}_2(V)$ is connected and the mapping $L \mapsto \kappa(L)$ is continuous, the image of this mapping is connected as well, which completes the proof. □

We use a well-known geometric realization result to establish the following corollary.

Corollary 2.3. *Let* $[a, b]$ *be any interval. There exists a hypersurface* M *in Euclidean space, a smooth metric* g *on* M, *and a point* $P \in M$ *so that the set of sectional curvatures of* M *at* P *is precisely* $[a, b]$.

Proof. Choose any collection of real numbers $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ (not necessarily distinct) whose minimal and maximal pairwise products are, respectively, a and b. Define the symmetric bilinear form φ so that $\lambda_1, \dots, \lambda_n$ are its eigenvalues. One may now carry out the geometric realization procedure as outlined on page 74 of [Gilkey 2001] to produce the desired result. □

Before proving our next corollary, we note that $R_{c\varphi} = c^2 R_\varphi$, and so any linear combination of canonical algebraic curvature tensors satisfies

$$\sum \alpha_i R_{\varphi_i} = \sum \epsilon_i R_{\tilde{\varphi}_i},$$

where $\epsilon_i = \pm 1 = \mathrm{sign}(\alpha_i)$ and $\tilde{\varphi}_i = \sqrt{|\alpha_i|}\varphi_i$.

The following corollary finds bounds on the sectional curvature values of any algebraic curvature tensor and is a direct consequence of Theorem 1.1.

Corollary 2.4. *Let* $R \in \mathcal{A}(V)$, *and set* $R = \sum_{i=1}^{\nu(R)} \epsilon_i R_{\varphi_i}$, *where* $\epsilon_i = \pm 1$ *and* $\varphi_i \in S^2(V^*)$. *Let* M_i *and* m_i *be, respectively, the maximal and minimal pairwise products of the eigenvalues of* φ_i, *and let* $M = \sum \epsilon_i M_i$ *and* $m = \sum \epsilon_i m_i$. *The set of possible sectional curvature values of the model space* $(V, \langle\cdot,\cdot\rangle, R)$ *is a closed subinterval of* $[m, M]$.

Remark 2.5. We give an example which shows that the bounds presented in Corollary 2.4 are not sharp. On an orthornormal basis $\{f_1, f_2, f_3\}$, define $\varphi_1, \varphi_2 \in S^2(V^*)$ by having the nonzero entries

$$\varphi_1(f_1, f_1) = \varphi_1(f_2, f_2) = \varphi_2(f_1, f_1) = \varphi_2(f_3, f_3) = 1.$$

According to Theorem 1.1, the maximal sectional curvature of both R_{φ_1} and R_{φ_2} is 1, so Corollary 2.4 estimates the maximal sectional curvature of $R_{\varphi_1} + R_{\varphi_2}$ to be 2. A straightforward calculation shows, however, that all sectional curvatures of $R_{\varphi_1} + R_{\varphi_2}$ are strictly less than 2. □

Remark 2.6. The authors in [Bettiol and Mendes 2017] independently obtained general sectional curvature bounds. While their methods involve representation theory and differential geometric considerations, our approach provides a more basic proof (especially in the case of a canonical algebraic curvature tensor) and can be extended to give a basic proof of the spectral theorem, which we give in the next section.

3. The spectral theorem

In this section we give a relatively short proof of the well-known spectral theorem after establishing a helpful lemma.

Lemma 3.1. *Suppose φ is a symmetric bilinear form on an inner product space V with $\dim(V) \geq 2$. There exists an orthonormal basis $\{f_1, e_2, \dots, e_n\}$ for V so that f_1 is an eigenvector of φ.*

Proof. Let φ be given, and consider the algebraic curvature tensor R_φ. We break the proof into cases: either $R_\varphi \neq 0$ or $R_\varphi = 0$.

Suppose first that $R_\varphi \neq 0$. Using Lemma 2.1, we choose a 2-plane π of extremal sectional curvature and find an orthonormal basis $\{f_1, e_2\}$ of eigenvectors for π, which we extend to the orthonormal basis $\{f_1, e_2, \dots, e_n\}$ for V, establishing the result.

Now suppose $R_\varphi = 0$. We claim that 0 is an eigenvalue[2] of φ. In this case, any corresponding unit eigenvector may be chosen first as part of an orthonormal basis for V, which would again establish the result.

Suppose to the contrary that 0 is not an eigenvalue of φ. Therefore, for any orthonormal basis $\{f_1, \dots, f_n\}$, the matrix with $\varphi(f_j, f_i)$ as its (i, j)-entry must have a nonzero determinant. By expanding this determinant by cofactors, there must be at least one nonzero 2×2 minor. Thus, for some indices i_1, i_2, j_1, j_2,

$$\begin{aligned} 0 &\neq \varphi(f_{i_1}, f_{j_1})\varphi(f_{i_2}, f_{j_2}) - \varphi(f_{i_1}, f_{j_2})\varphi(f_{i_2}, f_{j_1}) \\ &= R_\varphi(f_{i_1}, f_{i_2}, f_{j_2}, f_{j_1}), \end{aligned}$$

which contradicts the assumption that $R_\varphi = 0$. □

With these short preliminary facts established, we can now give an alternative proof of the spectral theorem.

Proof of the spectral theorem. The result is trivial if $\dim(V) = 1$, so assume $\dim(V) \geq 2$. By Lemma 3.1, there exists an orthonormal basis $\{f_1, e_2, \dots, e_n\}$ for

[2] In fact, it is known [Gilkey 2007] that if $R_\varphi = 0$ then the rank of φ is less than or equal to 1, so that on a vector space of dimension 2 or more, there must be a nontrivial kernel of φ. However, we wish to give a self-contained and alternative proof here as part of our effort to produce a new proof of the spectral theorem.

V so that f_1 is an eigenvector of φ. Now consider the vector space $W = f_1^\perp = \operatorname{span}\{e_2, \dots, e_n\}$. Since our inner product is positive definite, the restriction of the inner product to W remains positive definite. In addition, the restricted curvature tensor satisfies $(R_\varphi)|_W = R_{\varphi|_W}$.

We first consider when $\dim(W) \geq 2$. We again use Lemma 3.1 to find an orthonormal basis $\{f_2, \tilde{e}_3, \dots, \tilde{e}_n\}$ for W so that f_2 is an eigenvector of $\varphi|_W$. That is,

$$\varphi|_W(f_2, \tilde{e}_k) = \varphi(f_2, \tilde{e}_k) = 0.$$

Since f_1 is an eigenvector of φ, we have $\varphi(f_1, f_2) = 0$, so f_2 is also an eigenvector of φ. We have a new orthonormal basis $\{f_1, f_2, \tilde{e}_3, \dots, \tilde{e}_n\}$ for V, where both f_1 and f_2 are eigenvectors of φ.

Now consider $W_2 = \operatorname{span}\{f_1, f_2\}^\perp$. If $\dim(W_2) > 1$, then repeat the process outlined above, and continue until there is an orthonormal basis $\{f_1, \dots, f_{n-1}, \bar{e}_n\}$ for V (this is the case if $\dim(W) = 1$ above), and for any $i \neq j$ inclusively between 1 and $n - 1$ we have

$$\varphi(f_i, f_j) = \varphi(f_i, \bar{e}_n) = 0.$$

This demonstrates that $f_n = \bar{e}_n$ is also an eigenvector, completing the proof. □

Acknowledgments

The authors would like to thank the referee for insightful suggestions that improved the quality of this work. The authors also appreciate helpful discussions with I. Stavrov concerning this work. This research was jointly funded by California State University, San Bernardino, and the NSF grant DMS-1758020.

References

[Axler 2015] S. Axler, *Linear algebra done right*, 3rd ed., Springer, 2015. MR Zbl

[Bettiol and Mendes 2017] R. G. Bettiol and R. A. E. Mendes, "Sectional curvature and Weitzenböck formulae", preprint, 2017. arXiv

[Bishop and Goldberg 1964] R. L. Bishop and S. I. Goldberg, "Some implications of the generalized Gauss–Bonnet theorem", *Trans. Amer. Math. Soc.* **112** (1964), 508–535. MR Zbl

[Brozos Vázquez et al. 2012] M. Brozos Vázquez, P. B. Gilkey, and S. Nikcevic, *Geometric realizations of curvature*, ICP Adv. Texts Math. **6**, Imperial College Press, London, 2012. MR Zbl

[Diaz and Dunn 2010] A. Diaz and C. Dunn, "The linear independence of sets of two and three canonical algebraic curvature tensors", *Electron. J. Linear Algebra* **20** (2010), 436–448. MR Zbl

[Díaz-Ramos and García-Río 2004] J. C. Díaz-Ramos and E. García-Río, "A note on the structure of algebraic curvature tensors", *Linear Algebra Appl.* **382** (2004), 271–277. MR Zbl

[García-Río et al. 2002] E. García-Río, D. N. Kupeli, and R. Vázquez-Lorenzo, *Osserman manifolds in semi-Riemannian geometry*, Lecture Notes in Math. **1777**, Springer, 2002. MR Zbl

[Gilkey 2001] P. B. Gilkey, *Geometric properties of natural operators defined by the Riemann curvature tensor*, World Sci., River Edge, NJ, 2001. MR Zbl

[Gilkey 2007] P. B. Gilkey, *The geometry of curvature homogeneous pseudo-Riemannian manifolds*, ICP Adv. Texts Math. **2**, Imperial College Press, London, 2007. MR Zbl

[Gilkey and Nikčević 2005] P. B. Gilkey and S. Ž. Nikčević, "Generalized plane wave manifolds", *Kragujevac J. Math.* **28** (2005), 113–138. MR Zbl

[Klinger 1991] R. Klinger, "A basis that reduces to zero as many curvature components as possible", *Abh. Math. Sem. Univ. Hamburg* **61** (1991), 243–248. MR Zbl

[Kowalski and Prüfer 1994] O. Kowalski and F. Prüfer, "Curvature tensors in dimension four which do not belong to any curvature homogeneous space", *Arch. Math.* (*Brno*) **30**:1 (1994), 45–57. MR Zbl

[Lee 1997] J. M. Lee, *Riemannian manifolds: an introduction to curvature*, Grad. Texts in Math. **176**, Springer, 1997. MR Zbl

[Milnor and Stasheff 1974] J. W. Milnor and J. D. Stasheff, *Characteristic classes*, Ann. of Math. Stud. **76**, Princeton Univ. Press, 1974. MR Zbl

[Nash 1956] J. Nash, "The imbedding problem for Riemannian manifolds", *Ann. of Math.* (2) **63** (1956), 20–63. MR Zbl

[Tricerri and Vanhecke 1981] F. Tricerri and L. Vanhecke, "Curvature tensors on almost Hermitian manifolds", *Trans. Amer. Math. Soc.* **267**:2 (1981), 365–397. MR Zbl

[Tricerri and Vanhecke 1986] F. Tricerri and L. Vanhecke, "Variétés riemanniennes dont le tenseur de courbure est celui d'un espace symétrique riemannien irréductible", *C. R. Acad. Sci. Paris Sér. I Math.* **302**:6 (1986), 233–235. MR Zbl

[Tsankov 2005] Y. T. Tsankov, "A characterization of n-dimensional hypersurfaces in $\mathbb{R}^{n+1}$ with commuting curvature operators", pp. 205–209 in *PDEs, submanifolds and affine differential geometry* (Będlewo, Poland, 2003), edited by B. Opozda et al., Banach Center Publ. **69**, Polish Acad. Sci. Inst. Math., Warsaw, 2005. MR Zbl

Received: 2019-10-16 Revised: 2020-03-17 Accepted: 2020-04-28

callem@reed.edu *Reed College, Portland, OR, United States*

cmdunn@csusb.edu *Mathematics Department, California State University at San Bernardino, San Bernardino, CA, United States*

INVOLVE 13:3 (2020)
https://doi.org/10.2140/involve.2020.13.455

Qualitative investigation of cytolytic and noncytolytic immune response in an HBV model

John G. Alford and Stephen A. McCoy

(Communicated by Martin J. Bohner)

We investigate a mathematical model of infection by the hepatitis B virus (HBV) that includes cytolytic and noncytolytic immune response. The model exhibits a variety of steady-state solutions depending on parameter values, including nonunique and unique equilibrium solutions and periodic behavior. The disease-free equilibrium E_f and the positive-disease equilibrium E_d are examined. The basic reproduction ratio R_0 is computed in order to examine the uniqueness and local asymptotic stability of equilibria and to understand the model's biological implications for HBV dynamics.

1. Introduction

Hepatitis B virus (HBV) is a virus that negatively impacts hundreds of millions of people worldwide by causing chronic infections [Locarnini et al. 2015] in liver cells (or hepatocytes) which may result in scarring of liver tissue and ultimately liver cancer [Lavanchy 2004]. While advances in treatment have resulted in the control or eradication of many infectious diseases, from 1993 to 2013 the rates of death and disability caused by hepatitis rose by 63% and 34%, respectively [Stanaway et al. 2016]. This makes HBV a vital public health concern across the world. Mathematical models are often used to investigate the dynamics of biological phenomena, and virology is an area of study that has been greatly impacted by these models. For example, human immunodeficiency virus (HIV) was thought to replicate at a slow rate, but applications of mathematical models suggested that HIV-1 is in fact a rapidly reproducing virus. This was confirmed by experimental observations [Perelson 2002].

MSC2010: 92C60.
Keywords: mathematical model, hepatitis B viral dynamics, immune response.

The first mathematical model of HBV is discussed in [Nowak et al. 1996] and is given by

$$x' = \lambda - dx - bvx, \tag{1-1a}$$

$$y' = bvx - ay, \tag{1-1b}$$

$$v' = ky - uv, \tag{1-1c}$$

where x is the density of uninfected cells, y is the density of infected cells, and v is the number of free virus cells. Uninfected cells are assumed to be produced at a rate, λ, which may be constant or depend on the total population size of uninfected and infected cells. Uninfected cells die at rate d, and move to the infected class at rate bv, where b is the rate constant describing the infection process. Infected cells die at rate a. Free virions are produced from infected cells at rate k and are removed at rate u.

Many models of HBV have subsequently been investigated including [Rodriguez et al. 2017; Ciupe et al. 2007; Kim et al. 2012; Pang et al. 2012; Hews et al. 2010]. All of these models include variables that specify the density of healthy and unhealthy hepatocytes as well as the density of viral cells in the liver. The dynamics for the immune response is included in [Rodriguez et al. 2017; Ciupe et al. 2007; Kim et al. 2012; Pang et al. 2012]. In general, immune response includes cytolytic (direct killing) and noncytolytic activities. Noncytolytic immune responses use chemical signaling via cytokines to stop the reproduction of HBV within the cell, "curing" the cell without having to destroy it [Vargas-De-León 2014]. Since it is the host's response to the viral infection and not an innate part of the HBV life cycle that causes hepatocyte destruction (as HBV is not pathological), the role of the noncytolytic immune response is important. In [Wodarza et al. 2002] it was found that if the virus replicates at a fast rate relative to the rate of viral cytopathicity, then a combination of both cytolytic and noncytolytic immune response is more likely to control the disease. On the other hand, if viral cytopathicity is high relative to the replication rate of the virus then cytolytic and noncytolytic immune responses may resolve the infection independently. Citing chimpanzee studies, the authors in [Ciupe et al. 2007] suggest that the noncytolytic response may be the mechanism that causes an early reduction in the HBV viral load, as this occurs much earlier than any detectable cytolytic immune response, liver T cell infiltration, or liver damage. These authors include a fifth variable in their model to account for a cell population of previously infected cells that are refractory to new infection (because of the continuing effects of a noncytolytic immune response) or a population of infected cells not producing measurable amounts of virus.

In this paper a model of HBV dynamics based on these previous models is formulated and analyzed. Here the dynamics for healthy and unhealthy hepatocytes, the virus cells, and the immune cells are included in the model equations.

While [Ciupe et al. 2007; Kim et al. 2012] are mainly concerned with validating HBV models and investigating HBV dynamics utilizing patient data and computer simulations, in this paper the model equations are mathematically analyzed to determine the reproduction ratio and give conditions that guarantee stability of the disease equilibria, as in [Rodriguez et al. 2017; Pang et al. 2012; Hews et al. 2010]. Unlike [Rodriguez et al. 2017; Hews et al. 2010], the model here includes a fourth equation for the immune response dynamics, and unlike [Pang et al. 2012] the growth dynamics for healthy and unhealthy hepatocytes are logistic. The paper is outlined as follows. In Section 2 the model equations are presented and described. In Section 3 these equations are nondimensionalized and then the disease-free equilibrium is analyzed in Section 4. The disease equilibrium is discussed in Section 5 and in Section 6 a reduced system is presented, the disease equilibrium is analyzed, and the main theorems are proven. Finally, these results are summarized and future work is suggested in the conclusion, Section 7.

2. Model equations

The following model for hepatitis B dynamics builds on the basic framework of previous HBV models, including [Rodriguez et al. 2017; Ciupe et al. 2007; Kim et al. 2012; Pang et al. 2012; Hews et al. 2010]. The state variables include healthy hepatocyte density (X), unhealthy hepatocyte density (Y), and number of virus cells (V), just as in (1-1a), (1-1b), and (1-1c). A third variable I is now included to account for the density of immune response cells. The equations are

$$X' = r_1 X\Big(1 - \frac{X+Y}{K}\Big) - \beta XV + fbYI, \tag{2-1a}$$

$$Y' = r_2 Y\Big(1 - \frac{X+Y}{K}\Big) + \beta XV - bYI, \tag{2-1b}$$

$$V' = \gamma Y - \mu V, \tag{2-1c}$$

$$I' = cY - dI + I_0. \tag{2-1d}$$

Here both the healthy hepatocyte cell population dynamics (2-1a) and the unhealthy hepatocyte cell population dynamics (2-1b) are maintained by homeostasis as they are governed by logistic dynamics, as in [Rodriguez et al. 2017; Ciupe et al. 2007]. That is, the first terms in (2-1a) and (2-1b) yield negative per-capita growth rates when $X + Y > K$, where $K > 0$ is a parameter that defines hepatocyte carrying capacity. Healthy hepatocytes and infected hepatocytes proliferate at rates $r_1 > 0$ and $r_2 > 0$. Here it will be assumed that $r_1 \geq r_2$ as in [Rodriguez et al. 2017] so that healthy cells proliferate at least as fast as unhealthy cells. The parameter $\beta > 0$ controls the rate at which healthy hepatocytes are infected by virions so that the rate of infection is directly proportional to β and the density of healthy hepatocytes (i.e., mass action), just as in [Rodriguez et al. 2017; Ciupe et al. 2007; Kim et al. 2012; Pang et al. 2012]. The last terms in (2-1a) and (2-1b) represent the immune response to disease. The parameter $b > 0$ controls the cytolytic clearance rate of infected cells and the noncytolytic "curing" of infected cells, given by the last term in (2-1a), includes a scaling parameter f. Here $f \in (0, 1]$ as in [Kim et al. 2012], where the authors refer to f as a "calibration coefficient".

The viral dynamics (2-1c) and immune response dynamics (2-1d) are linear. The parameters $\gamma > 0$ and $c > 0$ control the rates at which virions and immune cells are produced in response to the number of unhealthy hepatocytes respectively. Thus $1/\mu$ and $1/d$ specify the average lifetime of a free virus particle [Nowak and Bangham 1996] and an immune cell [Ciupe et al. 2007] respectively. In the absence of unhealthy hepatocytes, immune cells grow at a constant rate I_0, as in [Ciupe et al. 2007; Kim et al. 2012; Pang et al. 2012], and the ratio I_0/d quantifies the basal level of immune cells [Ciupe et al. 2007] in the body.

The model presented here is distinct from the various models that have been studied previously. For example, in [Nowak et al. 1996; Kim et al. 2012; Pang et al. 2012], healthy hepatocytes (in the absence of immune response) grow at a constant rate. In [Nowak et al. 1996; Pang et al. 2012; Hews et al. 2010] unhealthy hepatocyte growth occurs only by infection of healthy cells. While the investigators in [Kim et al. 2012] allow for growth of unhealthy hepatocytes in the absence of infection, the growth dynamics is linear (proportional to the density of unhealthy cells) rather than logistic.

Example parameter values determined from patient data can be found in the literature. Table 1 displays typical values or ranges of values from [Rodriguez et al.

parameter	description	value	ref.
r_1	healthy-cell proliferation rate	(0,4] day^{-1}	R
r_2	unhealthy-cell proliferation rate	(0,4] day^{-1}	R
K	carrying capacity	$1.9 \times 10^7\,\text{mL}^{-1}$	R
β	healthy-cell infection rate	$(10^{-10}, 10^{-6})\,\text{mL}\,\text{day}^{-1}$	R
f	noncytolytic scaling parameter	(0, 1]	K
b	cytolytic clearance rate	$7 \times 10^{-4}\,\text{mL}\,\text{day}^{-1}$	K
γ	viral production rate	[1.4,164] day^{-1}	K
μ	viral clearance rate	[0.18, 1] day^{-1}	R
c	CTL response rate	$0.5\,\text{day}^{-1}$	K
d	CTL death rate	$0.5\,\text{day}^{-1}$	K
I_0	CTL production rate	$9.33\,\text{mL}^{-1}\,\text{day}^{-1}$	K

Table 1. Example parameter values. Here R stands for [Rodriguez et al. 2017] and K for [Kim et al. 2012].

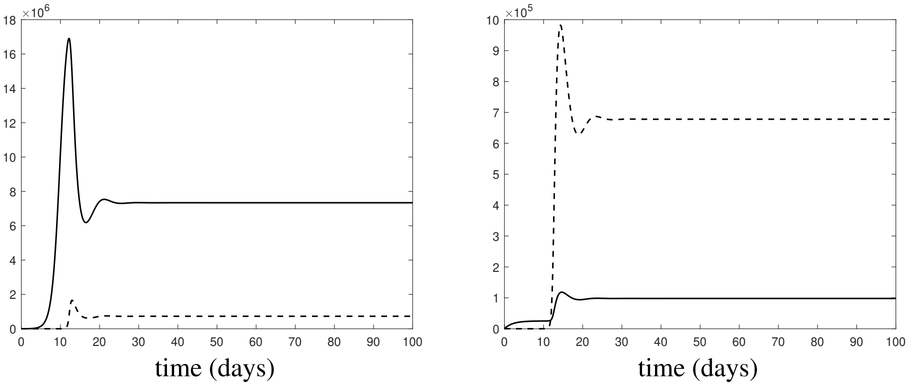


Figure 1. Simulation of (2-1a)–(2-1d). The left panel shows X (solid) and Y (dashed), where Y has been scaled by a factor of 10. The right panel shows I (solid) and V (dashed). Here $r_1 = 1$, $r_2 = 0.5$, $\beta = 10^{-6}$, $f = 0.1$, $\gamma = 6.24$, $\mu = 0.67$ and the remaining parameters are as in Table 1.

2017; Kim et al. 2012]. The last column displays the reference. The units assume that X, Y, and I are measured as the number of cells per mL, and V is measured as number of virions.

Simulations of (2-1a)–(2-1d) are shown in Figures 1 and 2. In each case, the initial conditions are that $X(0) = 1000$, $Y(0) = 100$, $V(0) = 10$, and $I(0) = 0$. Figure 2 illustrates that if the cytolytic clearance rate b is much smaller than the value listed in Table 1, the model exhibits periodic solutions as in [Hews et al. 2010].

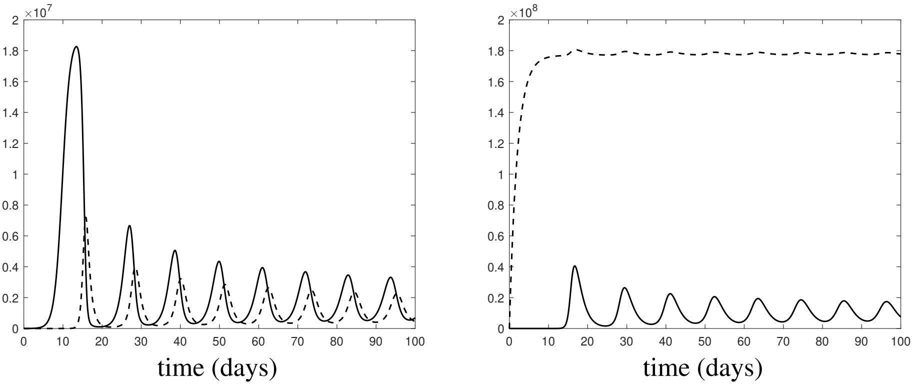


Figure 2. Simulation of (2-1a)–(2-1d). The left panel shows X (solid) and Y (dashed) and the right panel shows I (solid) and V (dashed). Here $b = 10^{-8}$, $r_1 = 1$, $r_2 = 0.5$, $\beta = 10^{-7}$, $f = 0.1$, $\gamma = 6.24$, $\mu = 0.67$ and the remaining parameters are as in Table 1.

3. Nondimensional equations

In order to analyze the steady-state solutions, it will be useful to nondimensionalize the model equations (2-1a)–(2-1d). The nondimensional parameters that will be used are

$$\hat{t} := r_2 t, \quad \hat{x} := \frac{X}{K}, \quad \hat{y} := \frac{Y}{K}, \quad \hat{v} := \left(\frac{\beta}{r_2}\right) V, \quad \hat{\alpha} := \frac{I}{K}.$$

After substituting these into (2-1a)–(2-1d) the system is transformed into

$$\hat{x}' = r\hat{x}(1 - \hat{x} - \hat{y}) - \hat{x}\hat{v} + f\varphi\hat{y}\hat{\alpha}, \tag{3-1a}$$

$$\hat{y}' = \hat{y}(1 - \hat{x} - \hat{y}) + \hat{x}\hat{v} - \varphi\hat{y}\hat{\alpha}, \tag{3-1b}$$

$$\hat{v}' = p_1\hat{y} - q_1\hat{v}, \tag{3-1c}$$

$$\hat{\alpha}' = p_2\hat{y} - q_2\hat{\alpha} + I_0 r_2 K, \tag{3-1d}$$

where

$$\varphi := bKr_2, \quad p_1 := K\gamma\beta r_2^2, \quad p_2 := cr_2, \quad q_1 := \mu r_2, \quad q_2 := dr_2. \tag{3-2}$$

Next we define

$$\omega := I_0 dK \tag{3-3}$$

and substitute $\alpha = \hat{\alpha} - \omega$ in (3-1a)–(3-1d) to get the homogeneous system

$$x' = rx(1 - x - y) - xv + f\varphi y(\alpha + \omega), \tag{3-4a}$$

$$y' = y(1 - x - y) + xv - \varphi y(\alpha + \omega), \tag{3-4b}$$

$$v' = p_1 y - q_1 v, \tag{3-4c}$$

$$\alpha' = p_2 y - q_2\alpha. \tag{3-4d}$$

If the right side of each equation (3-4a)–(3-4d) is denoted by $F_j(x_1, x_2, x_3, x_4)$, with $j = 1, 2, 3, 4$ and $x_1 = x,\ x_2 = y,\ x_3 = v,$ and $x_4 = \alpha$, the system clearly satisfies $F_k(x_1, x_2, x_3, x_4) \geq 0$ whenever $x_k = 0$ and $x_j \in [0, \infty),\ k \neq j.$ This implies solutions are nonnegative for nonnegative initial data [Thieme 2003]. The long-term behavior of the system is governed by any stable steady-states that may exist, so it is essential to understand how these solutions arise. Here only the constant steady-states (or equilibria) will be analyzed. In fact, this system has multiple equilibria. The trivial equilibrium is $E_0 = (0, 0, 0, 0)$ and the disease-free equilibrium is $E_f = (1, 0, 0, 0)$. The disease equilibrium will be denoted by $E_d = (x^*, y^*, v^*, \alpha^*)$, where $x^*, y^*, v^*, \alpha^* > 0$ since this is the sustained infection case. In the following sections the existence, uniqueness, and stability properties will be investigated for E_f and E_d. The trivial equilibrium is not considered as it represents total death of all hepatocytes. Furthermore, the equilibrium $(0, y^*, v^*, \alpha^*)$ will not be considered as it represents death of all healthy hepatocytes.

4. Disease-free equilibrium

The Jacobian of (3-4a)–(3-4d) is the matrix

$$J(x,y,v,\alpha)=\begin{bmatrix} r(1-2x-y)-v & -rx+f\varphi(\alpha+\omega) & -x & f\varphi y \\ -y+v & 1-x-2y-\varphi(\alpha+\omega) & x & -\varphi y \\ 0 & p_1 & -q_1 & 0 \\ 0 & p_2 & 0 & -q_2 \end{bmatrix}. \tag{4-1}$$

Substituting E_f into (4-1) yields the matrix

$$J(1,0,0,0)=\begin{bmatrix} -r & -r+f\varphi\omega & -1 & 0 \\ 0 & -\varphi\omega & 1 & 0 \\ 0 & p_1 & -q_1 & 0 \\ 0 & p_2 & 0 & -q_2 \end{bmatrix}, \tag{4-2}$$

which results in the following characteristic equation for the eigenvalue λ:

$$p(\lambda)=(r+\lambda)(q_2+\lambda)[\lambda^2+(q_1+\varphi\omega)\lambda+\varphi\omega q_1-p_1]. \tag{4-3}$$

Thus two eigenvalues are $\lambda=-r<0$ and $\lambda=-q_2<0$, while the remaining two are

$$\lambda=-(q_1+\varphi\omega)\pm\sqrt{(q_1-\varphi\omega)^2+4p_1}2.$$

It follows that all eigenvalues are real. Furthermore, since

$$(q_1-\varphi\omega)^2+4p_1=(q_1+\varphi\omega)^2-4(\varphi\omega q_1-p_1),$$

each eigenvalue is negative if and only if $\varphi\omega q_1-p_1>0$. It follows from [Thieme 2003; Allen 2007] that a necessary and sufficient condition for the disease-free equilibrium to be locally asymptotically stable is that $p_1/(\varphi\omega q_1)<1$. This yields the *basic reproduction ratio*, denoted by R_0, defined by

$$R_0:=p_1\varphi\omega q_1. \tag{4-4}$$

This ratio measures the average number of newly infected cells generated from one infected cell at the beginning of the infectious process [Wodarza et al. 2002; Pang et al. 2012; Thieme 2003]. Typically, for infectious disease models the basic reproduction ratio specifies a parameter grouping that provides a threshold [Diekmann et al. 1990] for infection to occur. That is, the disease will invade if $R_0>1$ or will not invade if $R_0<1$. Figure 3 shows simulations of healthy- and unhealthy-cell densities for the nondimensionalized equations (3-4a)–(3-4d), which illustrate this threshold behavior. When $R_0=0.9$, the steady-state value of healthy-cell density is 1 and the steady-state value of unhealthy-cell density is 0. However, for $R_0>1$ the disease state exists and steady-state values for healthy-cell density decrease with R_0, whereas the steady-state values for unhealthy-cell density increase with R_0.

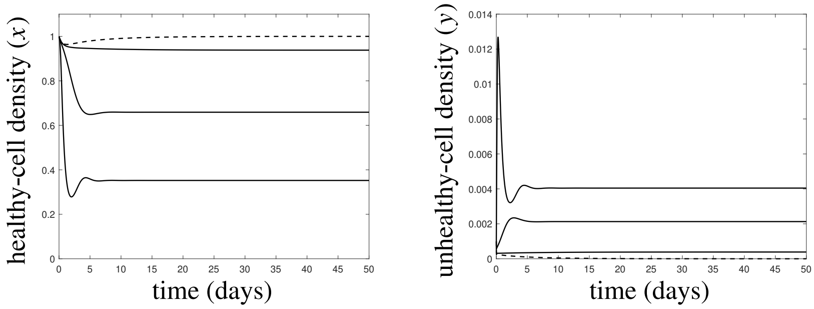


Figure 3. Simulations of (3-4a)–(3-4d) with healthy-cell density (x) on the left and unhealthy-cell density (y) on the right for $R_0 = 0.9$, $R_0 = 1.1$, $R_0 = 2$, and $R_0 = 20$, where R_0 is defined in (4-4). Dashed curves depict $R_0 = 0.9$ and solid curves correspond to values of $R_0 > 1$. The steady-state for $R_0 = 0.9$ is the disease-free equilibrium E_f, while the steady-state for $R_0 > 1$ is the disease equilibrium E_d. As R_0 increases, steady-state values of x decrease, while steady-state values of y increase.

The sections that follow investigate how the disease equilibrium E_d is affected by the parameters in the model. Section 5 establishes parameter ranges that guarantee uniqueness of E_d. In Section 6, parameter ranges are determined that guarantee the local asymptotic stability of E_d in a reduced model.

5. Disease equilibrium

In this section, the disease equilibrium $E_d = (x^*, y^*, v^*, \alpha^*)$ is computed and its parameter-dependence is investigated numerically. Assuming x^*, y^*, v^*, α^* are positive constants, then using (3-4b)–(3-4d) these constants satisfy

$$y^* = 1\theta(1-\varphi\omega-\rho x^*), \qquad v^* = \frac{p_1}{q_1}y^*, \qquad \alpha^* = p_2q_2y^*, \tag{5-1}$$

where

$$\theta := 1 + p_2q_2\varphi, \qquad \rho := 1-\frac{p_1}{q_1}. \tag{5-2}$$

Substituting these values in (3-4a) results in the quadratic equation for the healthy-cell density $\kappa_2x^2+\kappa_1x+\kappa_0 = 0$, where

$$\kappa_2 := -\rho\left(r\lambda-\frac{p_1}{\theta q_1}\right)-r\frac{p_1}{q_1}+\frac{f}{\theta}\lambda\rho^2, \tag{5-3a}$$

$$\kappa_1 := (1-\varphi\omega)\left(r\lambda-\frac{p_1}{\theta q_1}\right)+r\varphi\omega-f\frac{\rho}{\theta}[\varphi\omega+2\lambda(1-\varphi\omega)], \tag{5-3b}$$

$$\kappa_0 := \frac{f}{\theta}\{\lambda(1-\varphi\omega)^2+\varphi\omega(1-\varphi\omega)\}, \tag{5-3c}$$

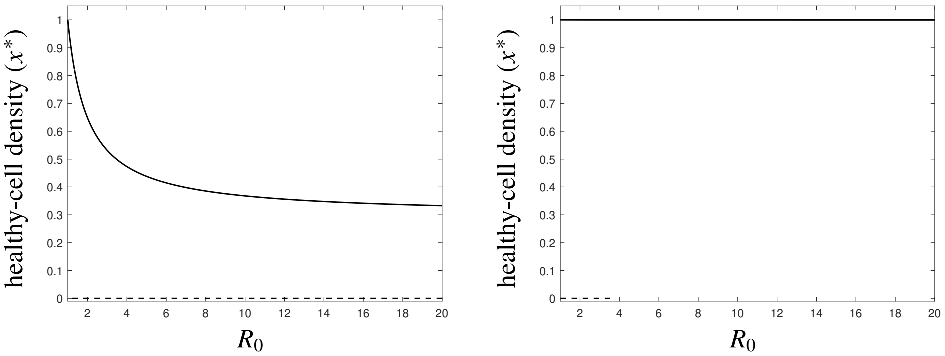


Figure 4. The healthy-cell density computed from (5-5) for the disease equilibrium E_d plotted as a function of $1 < R_0 < 20$. Here ω is varied, while all other parameters are fixed and R_0 is computed from (4-4). The healthy-cell infection rates are $\beta = 10^{-6}$ mL day^{-1} (left) and $\beta = 10^{-8}$ mL day^{-1} (right). In both panels the remaining parameter values are as in Table 1, with $r_1 = 1$, $r_2 = 0.5$, $f = 0.01$, $\gamma = 6.24$, and $\mu = 0.67$. Solid curves are stable positive equilibria, while dashed curves (near zero) are unstable positive equilibria.

with λ defined as

$$\lambda := \varphi p_2 \theta q_2 = \varphi p_2 \varphi p_2 + q_2. \tag{5-4}$$

The healthy-cell density for the disease equilibrium is then

$$x^* = -12\kappa_2\big[\kappa_1 \pm \sqrt{\kappa_1^2 - 4\kappa_2\kappa_0}\big], \tag{5-5}$$

where κ_i $(i = 0, 1, 2)$ are given by (5-3a)–(5-3c). The remaining values for the disease equilibrium are then easily computed from (5-1). After some algebra, we get from (5-3a), (5-3b), and (5-3c) that

$$\begin{aligned}\kappa_1^2 - 4\kappa_2\kappa_0 = \bigg[(1-\varphi\omega)\bigg(r\lambda - \theta^{-1}\frac{p_1}{q_1}\bigg) + r\varphi\omega + \rho\varphi\omega f\theta^{-1}\bigg]^2 \\ +4r\varphi\omega f\theta^{-1}(R_0-1)[\varphi\omega(1-\lambda)+\lambda],\end{aligned} \tag{5-6}$$

$$\kappa_2 = -r\bigg\{\frac{p_1}{q_1} + \lambda\rho\bigg\} + \theta^{-1}\rho\bigg\{\frac{p_1}{q_1} + \lambda f\rho\bigg\}, \tag{5-7}$$

$$\kappa_2 + \kappa_1 + \kappa_0 = \varphi\omega r(R_0-1)(\lambda-1) + (\varphi\omega)^2\theta^{-1}(R_0-1)[f(1-\lambda) - R_0(1-\lambda f)]. \tag{5-8}$$

Note from the definition of R_0 in (4-4) that if all parameters are fixed except ω, then $R_0 \downarrow 1$ if and only if $\omega \uparrow \bar{\omega}$, where

$$\bar{\omega} := p_1 \varphi q_1. \tag{5-9}$$

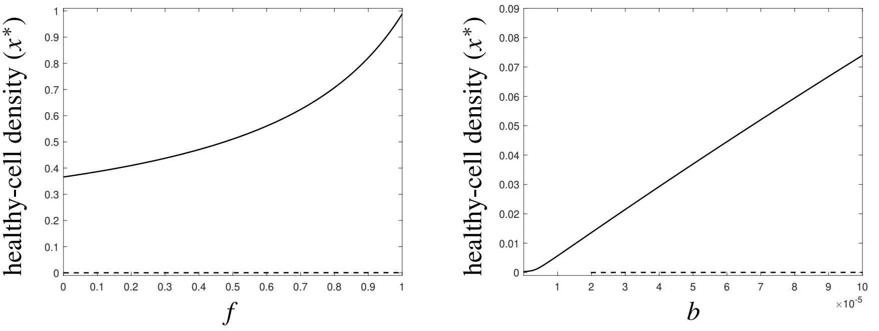


Figure 5. The healthy-cell density computed from (5-5) for the disease equilibrium E_d plotted as a function of noncytolytic scaling $f \in (0, 1]$ (left) and cytolytic clearance rate b (right) when $f = 0.1$. The healthy-cell infection rate is $\beta = 10^{-6}\,\mathrm{mL\,day}^{-1}$. In both panels $R_0 = 10$ and the remaining parameter values are as in Table 1, with $r_1 = 1$, $r_2 = 0.5$, $\gamma = 6.24$, and $\mu = 0.67$. Solid curves are stable positive equilibria, while dashed curves (near zero) are unstable positive equilibria. Note the difference in vertical scale in each plot.

In this case, $x^* \uparrow 1$, since $\kappa_1 + \kappa_2 + \kappa_0 = 0$ when $R_0 = 1$. Furthermore, $\omega \downarrow 0$ is equivalent to $R_0 \uparrow \infty$ and in this case

$$x^* \downarrow -12\kappa_2\big[\kappa_1(0) \pm \sqrt{\kappa_1(0)^2 - 4\kappa_2\kappa_0(0)}\big],$$

where $\kappa_0(0)$ and $\kappa_1(0)$ are from (5-3b) and (5-3c) respectively when $\omega = 0$.

Figure 4 shows the density of healthy cells x^* for the disease equilibrium as a function of R_0 when $1 < R_0 < 20$. Here R_0 is determined from ω as in (4-4) with all other parameters fixed. The two plots in the figure represent two different values for the healthy-cell infection rate β. If $\beta = 10^{-6}\,\mathrm{mL\,day}^{-1}$ (as in the left panel), there are two positive equilibria for each value of R_0 where the smaller is near zero and the larger decreases from a value of 1 to a value of approximately 0.33. If $\beta = 10^{-8}\,\mathrm{mL\,day}^{-1}$, there is a bifurcation at $R_0 \approx 3.5$ so that two positive equilibria exist for $R_0 < 3.5$ and there is a unique positive equilibrium for $R_0 > 3.5$. In each case, stability was determined from Matlab by computing the eigenvalues of (4-1) evaluated at E_d.

Figures 5 and 6 show the density of healthy cells x^* for the disease equilibrium when $R_0 = 10$ as it depends on both the noncytolytic scaling parameter f and the cytolytic clearance rate b. When healthy-cell infection rate β is relatively large as in Figure 5, the density of healthy cells is much more sensitive to changes in f than b. Conversely, when β is relatively small as in Figure 6, the density of healthy cells is much more sensitive to changes in b than f. The logarithm of x^* is shown

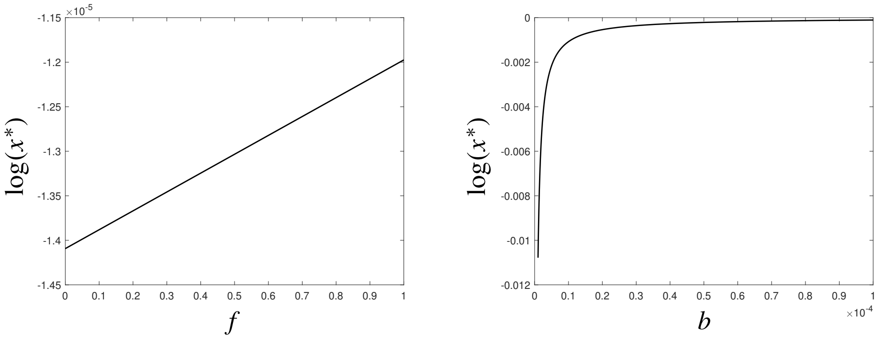


Figure 6. The logarithm (base 10) of the healthy-cell density computed from (5-5) for the disease equilibrium E_d plotted as a function of noncytolytic scaling $f \in (0, 1]$ (left) and cytolytic clearance rate b (right) when $f = 0.1$ and $R_0 = 10$. The healthy-cell infection rate is $\beta = 10^{-9}$ mL day^{-1}. In both panels the remaining parameter values are as in Table 1, with $r_1 = 1$, $r_2 = 0.5$, $\gamma = 6.24$, $\mu = 0.67$, $c = d = 0.5$. Here the parameters satisfy the conditions of Theorem 5.1.

in Figure 6 to illustrate that these solutions obey $0 < x^* < 1$. Furthermore, there is a unique value for x^* over all parameter values given in the figure. In Figure 6 the parameter ρ defined in (5-2) is positive and it will next be shown in Theorem 5.1 that this condition is sufficient so that the disease equilibrium is real, bounded, and unique when $R_0 > 1$.

Theorem 5.1. *Assume* $R_0 > 1,\ \rho > 0,\ 0 < f \le 1,$ *and* $r \ge 1$. *Then system* (3-4a)–(3-4d) *has a unique*, *positive-disease equilibrium* E_d.

Proof. If $R_0 > 1$, then from (5-6) it is clear that $\kappa_1^2 - 4\kappa_2\kappa_0 > 0$ since $0 < \lambda < 1$. From (5-7), the coefficient $\kappa_2 < 0$ if and only if

$$r\theta > \rho \frac{p_1}{q_1} + \lambda f \rho \frac{p_1}{q_1} + \lambda\rho,$$

which follows immediately from the bounds $\theta > 1,\ 0 < \lambda < 1$, and the assumptions of the theorem. Using the definition of ρ, φ, and ω from (3-2), (3-3) and (5-2) and the assumption that ρ is positive means that $p_1/q_1 < 1$ and noting that p_1/q_1 is equivalent to $\varphi\omega R_0$, it follows that $\varphi\omega < 1$ and hence $\kappa_0 > 0$. Therefore $\kappa_2 x^2 + \kappa_1 x + \kappa_0$, where κ_i $(i = 0, 1, 2)$ are given by (5-3a)–(5-3c), has two real solutions with opposite signs. From (5-8), the sum $\kappa_2 + \kappa_1 + \kappa_0 < 0$ follows easily since $f \le 1 < R_0$ and $\lambda < 1$. It then follows that the positive root of $\kappa_2 x^2 + \kappa_1 x + \kappa_0$ must be less than 1, and we get $0 < x^* < 1$. In this case, since $R_0 > 1$ is equivalent to $\rho < 1 - \varphi\omega$, inspection of (5-1) shows that $y^* > 0,\ v^* > 0$, and $\alpha^* > 0$. □

In the following section the local asymptotic stability of the disease equilibrium will be considered using a reduced system. In particular, a system of three equations is presented which mimics x, y, and v from (3-4a), (3-4b), and (3-4c) under certain parameter constraints and allows for a more tractable analytical result.

6. Reduced system

The parameter values in Table 1 have thus far provided baseline values for simulations. There is evidence in the literature that supports the investigation of smaller values for both b (cytolytic clearance rate) and c (CTL response rate). The authors in [Ciupe et al. 2007] conjecture that a patient who was immunosuppressed because of a corticosteroid drug regimen had weak cytotoxic clearance rate and low immune response rate which corresponds to reductions in the values of both b and c. In [Zapata et al. 2014] the ratio of the rate of immune response activation to the death rate of immune response (i.e., c/d in (2-1d)) is given as 0.02, which is much smaller than $c/d = 1$ from Table 1.

Inspection of (3-2) shows that decreasing b corresponds to decreasing φ and decreasing c/d corresponds to decreasing p_2/q_2. Numerical experimentation shows that as either φ or p_2/q_2 get smaller and smaller from the baseline values that are suggested by Table 1, the values x, y, and v computed from (3-4a)–(3-4d) are better and better approximated by solutions of the reduced system given by

$$x' = rx(1-x-y) - xv + f\varphi y(\alpha^* + \omega), \tag{6-1a}$$

$$y' = y(1-x-y) + xv - \varphi y(\alpha^* + \omega), \tag{6-1b}$$

$$v' = p_1 y - q_1 v, \tag{6-1c}$$

where $\alpha^* = \eta y^*$ with $\eta := p_2/q_2$, which is the equilibrium value for α obtained from (3-4d).

The disease equilibrium (x^*, y^*, v^*) for the reduced system (6-1a)–(6-1c) is equivalent to E_d from Section 5 when α^* is included. Thus, Theorem 5.1 also applies to (6-1a)–(6-1c). The Jacobian evaluated at the disease equilibrium is

$$J(x^*, y^*, v^*) = \begin{bmatrix} t_1 & t_3 & -x^* \\ -\rho y^* & t_2 & x^* \\ 0 & p_1 & -q_1 \end{bmatrix}, \tag{6-2}$$

where

$$t_1 = r(1 - 2x^* - y^*) - (1-\rho)y^*, \tag{6-3}$$

$$t_2 = 1 - x^* - 2y^* - \varphi(\eta y^* + \omega) = -(1-\rho)x^* - y^*, \tag{6-4}$$

$$t_3 = -rx^* + f\varphi(\eta y^* + \omega). \tag{6-5}$$

Here t_2 simplifies using $1 - x^* - y^* = \varphi(\alpha^* + \omega) - (1 - \rho)x^*$ from (6-1b). The characteristic equation is cubic $\lambda^3 + a_1\lambda^2 + a_2\lambda + a_3 = 0$, where

$$a_1 = q_1 - t_1 - t_2, \tag{6-6}$$

$$a_2 = t_1 t_2 - q_1(t_1 + t_2) - p_1 x^* + \rho t_3 y^*, \tag{6-7}$$

$$a_3 = q_1 t_1 t_2 + p_1 t_1 x^* + \rho q_1 t_3 y^* - \rho p_1 x^* y^*. \tag{6-8}$$

The Routh–Hurwitz conditions [Edelstein-Keshet 1988] will be used to derive necessary and sufficient conditions so that (x^*, y^*, α^*) is locally asymptotically stable. These conditions are that

$$a_1 > 0, \quad a_3 > 0, \quad a_1 a_2 > a_3. \tag{6-9}$$

Inspection of the dimensional system (2-1a)–(2-1d) shows that the constant cytolytic immune response rate I_0 will affect the steady-states for α. In the nondimensional system, I_0 controls $\omega = I_0/(dK)$ from (3-3). The following analysis uses ω as the main parameter to derive bounds on the other parameters and show that the Routh–Hurwitz conditions hold. Recall that x^* is determined from solutions of the quadratic equation

$$\kappa_2 x^2 + \kappa_1(\omega)x + \kappa_0(\omega) = 0, \tag{6-10}$$

where κ_i $(i = 0, 1, 2)$ are given by (5-3a)–(5-3c). Note that the coefficient κ_2 does not depend on ω, while κ_1 and κ_0 both depend on ω, and this has been made explicit in the notation. Therefore, x^* also depends on ω, which will be denoted as $x^*(\omega)$. Assuming the conditions in Theorem 5.1, $0 < x^*(\omega) < 1$ and $x^*(\omega)$ is defined for $\omega \in (0, \bar{\omega})$, where $\bar{\omega}$ is defined in (5-9). Furthermore, (5-3b) shows that κ_1 is a linear function of ω and that

$$\begin{aligned} \kappa_1'(\omega) &= \frac{d\kappa_1}{d\omega} = \varphi\left\{-\left(r\lambda - \frac{p_1}{\theta q_1}\right) + r - f\frac{\rho}{\theta}[1 - 2\lambda]\right\}, \\ \kappa_1(0) &= r\lambda - \frac{p_1}{\theta q_1} - 2\lambda f\frac{\rho}{\theta}. \end{aligned} \tag{6-11}$$

The following lemma derives a lower bound on $x^*(\omega)$ by replacing $\kappa_0(\omega)$ in (6-10) with a lower bound.

Lemma 6.1. *Assume all parameters are fixed except ω. Suppose that $R_0 > 1$, $\rho > 0$, $0 < f \leq 1$, and $r \geq 1$. Define*

$$\psi(\omega) := \frac{f}{\theta}\{\lambda\rho^2 + \varphi\omega\rho\}, \tag{6-12}$$

and let $x^*(\omega)$ denote the unique, positive value of the healthy-cell density predicted by Theorem 5.1 and found by solving (6-10). *Then there is a unique positive root of $\kappa_2 x^2 + \kappa_1(\omega)x + \psi(\omega) = 0$. Denote this root as $z(\omega)$. Then $0 < z(\omega) < x^*(\omega)$*

and $0 < z'(\omega) < \infty$, $|z''(\omega)| < \infty$, *and the sign of* $z''(\omega)$ *does not change and is equivalent to the sign of* ζ, *where*

$$\zeta = \kappa_2(\psi'\kappa_1')^2 - \kappa_1(0)\psi'\kappa_1' + \psi(0). \tag{6-13}$$

Proof. The assumption that $R_0 > 1$ is equivalent to the inequality $\rho < 1 - \varphi\omega$. Thus, comparing ψ with κ_0 from (5-3c) shows that $0 < \psi(\omega) < \kappa_0(\omega)$. Therefore, since $\kappa_2 < 0$ (as was shown in the proof of Theorem 5.1), the equation $\kappa_2 x^2 + \kappa_1(\omega)x + \psi(\omega) = 0$ has two real roots of opposite sign. Denoting the positive root as z, it follows that $0 < z(\omega) < x^*(\omega)$. Since κ_2 is independent of ω, differentiating both sides of $\kappa_2 x^2 + \kappa_1(\omega)x + \psi(\omega) = 0$ yields

$$z'(\omega) = -\kappa_1'(\omega)z(\omega) + \psi'(\omega)2\kappa_2 z(\omega) + \kappa_1(\omega). \tag{6-14}$$

Note that both $\kappa_1'(\omega)$ and $\psi'(\omega) > 0$ are independent of ω so the functional dependence on ω will be dropped for convenience. Since $\theta = 1 + p_2\varphi/q_2$ and $\rho = 1 - p_1/q_1$, (6-11) yields that κ_1' is positive if and only if $r + 1 > \rho + f\rho(1 - 2\lambda)$, which holds since $\lambda < 1$ and $r \geq 1,\ 0 < \rho < 1,\ 0 < f < 1$ by assumption. Since $\kappa_2 < 0$ and $z(\omega)$ is the positive root of the quadratic equation $\kappa_2 x^2 + \kappa_1(\omega)x + \psi(\omega) = 0$, it follows that the denominator in (6-14) is negative and $0 < z(\omega) < x^*(\omega)$.

Differentiating $dz/d\omega$ from (6-14) yields that

$$z''(\omega) = \kappa_1'(\kappa_1'z(\omega) + \psi') - z'(\omega)(\kappa_1'\kappa_1(\omega) - 2\kappa_2\psi')(2\kappa_2 z(\omega) + \kappa_1(\omega))^2.$$

This shows that $|z''(\omega)| < \infty$. We substitute $z'(\omega)$ from (6-14) to get that $z''(\omega)$ is positive if and only if

$$(\kappa_1'z(\omega) + \psi')\{\kappa_1' + \kappa_1'\kappa_1(\omega) - 2\kappa_2\psi'2\kappa_2 z(\omega) + \kappa_1(\omega)\} > 0,$$

and using the conditions $\kappa_2 < 0$ and $\kappa_1'z(\omega) + \psi' > 0$, it follows that $z''(\omega)$ is positive if and only if

$$z(\omega) + \kappa_1(\omega)\kappa_2 - \psi'\kappa_1' = \kappa_1(\omega)2\kappa_2 - \sqrt{\kappa_1(\omega)^2 - 4\kappa_2\psi(\omega)}2\kappa_2 - \psi'\kappa_1' > 0,$$

which is equivalent to

$$\kappa_2(\psi'\kappa_1')^2 - \kappa_1(\omega)\psi'\kappa_1' + \psi(\omega) > 0.$$

The first term is independent of ω and note that the derivative of the last two terms is zero so that the last two terms are also independent of ω. Since κ_1' and $\psi' = f\varphi\rho/\theta$ are positive, this inequality holds by the assumption (6-13). □

In order to prove local asymptotic stability of the disease equilibrium under certain parameter constraints, it will be necessary to use the properties of $z(\omega)$, $z'(\omega)$, and $z''(\omega)$ from Lemma 6.1.

Lemma 6.2. *Assume the hypotheses in Lemma 6.1 hold. Define the function $F(\omega)$ by*

$$F(\omega) := mz(\omega) + s(\omega), \quad \omega \in [0, \bar{\omega}], \tag{6-15}$$

where $\bar{\omega}$ is as in (5-9) *and*

$$m := 2r\theta - \rho(r+1-\rho)(1+\theta), \quad s(\omega) := (r+1-\rho)(1-\varphi\omega) - r\theta. \tag{6-16}$$

Then $F(\omega) \geq 0$ on $(0, \bar{\omega})$.

Proof. The constant m is positive if and only if

$$r > \rho\theta(1-\rho) + 1 - \rho\theta(1-\rho) + \theta - \rho,$$

which follows since $\theta > 1$ and $\rho < 1$. Therefore, since $s''(\omega) = 0$ and $F''(\omega) = mz''(\omega)$, we have $F''(\omega)$ is finite and has the same sign as ζ from (6-13). Since $R_0 = 1$ if and only if $\rho = 1 - \varphi\bar{\omega}$, comparing (5-3c) and (6-12) shows that $\psi(\bar{\omega}) = c(\bar{\omega})$ so that $z(\bar{\omega}) = x^*(\bar{\omega}) = 1$. Now use $\rho = 1 - \varphi\bar{\omega}$ to get

$$F(\bar{\omega}) = m + s(\bar{\omega}) = \theta(1-\rho)(r-\rho) > 0.$$

Now $s(0) = 1 - \rho - r(\theta - 1)$ so that if $r(\theta - 1) \leq 1 - \rho$ and $s(0) \geq 0$, it follows directly that $F(0) > 0$ since $z(0) > 0$. Otherwise, consider the case where $r(\theta - 1) > 1 - \rho$ and $s(0) < 0$. Note that $F(0) > 0$ if and only if $z(0) > -s(0)/m$. Since $\kappa_2 < 0$, $\psi(\omega) > 0$, and $z(\omega)$ is the positive root of $\kappa_2 x^2 + \kappa_1(\omega)x + \psi(\omega) = 0$, it follows that $F(0) > 0$ whenever

$$\kappa_2(s(0)m)^2 - \kappa_1(0)s(0)m + \psi(0) > 0. \tag{6-17}$$

After substituting $1 - \rho = p_1/q_1$ in (5-3a), (5-3b) and (6-12), we get

$$\kappa_2 = -\rho\theta^{-1}(r\theta\lambda - 1 + \rho) - r(1-\rho) + \psi(0), \quad \kappa_1(0) = \theta^{-1}(r\theta\lambda - 1 + \rho) - 2\rho\psi(0).$$

We substitute these expressions into (6-17) and multiply by θm^2 to get that $F(0) > 0$ if and only if

$$(\rho(r\theta\lambda - 1 + \rho) + r\theta(1-\rho))s(0)^2 + m(r\theta\lambda - 1 + \rho)s(0) \\ < \theta\psi(0)(s(0)^2 + 2\rho s(0)m + m^2). \tag{6-18}$$

Since $\rho < 1$, $m > 0$ and $s(0) < 0$, it follows that

$$s(0)^2 + 2\rho s(0)m + m^2 > (s(0) + m)^2 > 0,$$

and the right side of the inequality in (6-18) is positive. The left side is the product of $s(0)$ and

$$[\rho(r\theta\lambda - 1 + \rho) + r\theta(1-\rho)]s(0) + m(r\theta\lambda - 1 + \rho),$$

so that (6-18) holds whenever this quantity is positive. After substituting $s(0)$, m, and $\theta - 1 = \theta\lambda$ and then simplifying, this becomes

$$\theta(1-\rho)\{r^2(\theta-1) - r[\rho(\theta-1) - (1-\rho)] + \rho(1-\rho)\},$$

so that (6-18) holds whenever

$$r^2(\theta-1) - r[\rho(\theta-1) - (1-\rho)] + \rho(1-\rho) > 0. \tag{6-19}$$

The left side is a quadratic function of r and, after some algebra, its minimum value is computed as $-(\rho\theta-1)^2/[4(\theta-1)]$, so there are two real zeros for the left side. Therefore, since $r \geq 1$, the inequality (6-19) holds whenever the larger of these two roots is less than 1. Equivalently, when $r = 1$ the left side of (6-19) must be positive. We substitute $r = 1$ to get $(1-\rho)(\theta-\rho) > 0$, which follows since $\rho < 1$ and $\theta > 1$. Therefore, (6-18) follows and $F(0) > 0$.

Next, it will be shown that $F(\omega)$ is nonnegative on the entire interval $(0, \bar{\omega})$. First assume $\zeta \leq 0$. Then by Lemma 6.2, it follows that $F(\omega)$ is either linear or concave down on $(0, \bar{\omega})$. Since $F(0) > 0$ and $F(\bar{\omega}) > 0$, it follows that $F(\omega)$ is positive on the entire interval. Next assume $\zeta > 0$ so that $F(\omega)$ is concave up on $(0, \bar{\omega})$. If $F'(0)$ and $F'(\bar{\omega})$ have the same sign, then $F(\omega)$ is either increasing or decreasing on $(0, \bar{\omega})$, which again implies that $F(\omega)$ is positive on the entire interval. If $F'(0)$ and $F'(\bar{\omega})$ have opposite signs, there must exist $\omega^* \in (0, \bar{\omega})$ for which $F'(\omega^*) = 0$ and $F(\omega^*)$ is a minimum. If $F(\omega^*) < 0$, there exist two values $\omega_1 \in (0, \bar{\omega})$ and $\omega_2 \in (0, \bar{\omega})$, where $\omega_1 \neq \omega_2$ and $F(\omega_1) = F(\omega_2) = 0$. The corresponding values of z are given by $-s(\omega_1)/m$ and $-s(\omega_2)/m$. After squaring $F(\omega)$, it follows that these zeros for $F(\omega)$ must also obey

$$m^2 z(\omega)^2 + 2mz(\omega)s(\omega) + s(\omega)^2 = 0.$$

Since z obeys $\kappa_2 z^2 + \kappa_1(\omega)z + \psi = 0$ and $z(\omega) = -s(\omega)/m$, this equation is equivalent to

$$\kappa_2 s(\omega)^2 - m\kappa_1(\omega)s(\omega) + m^2\psi(\omega) = 0.$$

Here $\kappa_2 < 0$ and $m^2\psi > 0$ so that this equation has two real roots with opposite sign. Since $z(\omega)$ must be positive, this leads to a contradiction, and therefore $F(\omega) \geq 0$ on $(0, \bar{\omega})$. □

The next theorem uses Lemma 6.2 to show the disease equilibrium of the reduced system is locally asymptotically stable.

Theorem 6.3. *Assume $R_0 > 1$, $\rho > 0$, $0 < f \leq 1$, and $r \geq 1$. Then the disease equilibrium for the reduced system* (6-1a)–(6-1c) *is locally asymptotically stable.*

Proof. First use $1 - \rho = p_1/q_1$ with (6-4) to see that $t_2 + (p_1/q_1)x^* = -y^*$ so that

$$a_3 = q_1 y^*\{-t_1 - r\rho x^* + \rho f\varphi(\eta y^* + \omega) - \rho(1-\rho)x^*\}, \tag{6-20}$$

after substituting t_3 from (6-5). It follows that $a_3 > 0$ if and only if

$$-t_1 > r\rho x^* + \rho(1-\rho)x^* - \rho f\varphi(\eta y^* + \omega). \tag{6-21}$$

We substitute t_1 from (6-3) to get that (6-21) is equivalent to

$$r(2x^* - 1) + (r + 1 - \rho)(y^* - \rho x^*) + \rho f\varphi(\eta y^* + \omega) > 0. \tag{6-22}$$

Then, after substituting for y^* from (5-1) and rearranging terms, it follows that $a_3 > 0$ if and only if

$$mx^* + s(\omega) + \rho\theta f\varphi(\eta y^* + \omega) > 0, \tag{6-23}$$

with m and $s(\omega)$ defined in (6-16). Since $m > 0$ and $0 < z(\omega) < x^*$, it follows that a sufficient condition for (6-23) is that

$$F(\omega) + \rho\theta f\varphi(\eta y^* + \omega) > 0 \tag{6-24}$$

for all $\omega \in (0, \bar{\omega})$. Using Lemma 6.2, $F(\omega) \geq 0$ on $(0, \bar{\omega})$. Furthermore, the inequality $\rho\theta f\varphi(\eta y^* + \omega) > 0$ implies that (6-24) holds and $a_3 > 0$ on $(0, \bar{\omega})$.

To show that $a_1 > 0$ note that when (6-21) holds, this implies that $t_1 < 0$. Since $\rho < 1$, it is clear from (6-4) that $t_2 < 0$ and it follows by inspection of (6-6) that $a_1 > 0$. Finally, in order to show $a_1a_2 > a_3$, start with the identity

$$\frac{a_3}{q_1} = a_2 + q_1(t_1 + t_2) + p_1x^* + \left(\frac{p_1}{q_1}\right)t_1x^* - \rho\left(\frac{p_1}{q_1}\right)x^*y^*,$$

which is easily verified by inspection of (6-7) and (6-8) and simple algebra. Thus,

$$a_2 > \frac{a_3}{q_1} - t_1\left(q_1 + \left(\frac{p_1}{q_1}\right)x^*\right) - q_1\left(t_2 + \left(\frac{p_1}{q_1}\right)x^*\right).$$

Now use the fact that $t_1 < 0$ and $t_2 + (p_1/q_1)x^* = -y^*$ to see that

$$a_1 > q_1 \quad \text{and} \quad a_2 > \frac{a_3}{q_1},$$

which shows that $a_1a_2 > a_3$. The Routh–Hurwitz conditions (6-9) hold, and the disease-free equilibrium (x^*, y^*, α^*) is locally asymptotically stable. □

In certain parameter ranges, the reduced model may inform the full model (3-4a)–(3-4d). The Jacobian evaluated at the disease equilibrium for the full model is

$$J(x^*, y^*, v^*, \alpha^*) = \begin{bmatrix} t_1 & t_3 & -x^* & f\varphi y^* \\ -\rho y^* & t_2 & x^* & -\varphi y^* \\ 0 & p_1 & -q_1 & 0 \\ 0 & p_2 & 0 & -q_2 \end{bmatrix}. \tag{6-25}$$

In general, the eigenvalues of the reduced model determined from (6-2) cannot be used to determine the eigenvalues of (6-25). However, as discussed at the beginning of this section, the time courses for x, y, and v in the reduced model (6-1a)–(6-1c) approximate these same time courses in the full model when φ is small and it is easily seen from (6-2) and (6-25) that as $\varphi \to 0$, the eigenvalues for the full model are given by the three eigenvalues of (6-2) and a fourth eigenvalue equal to $-q_2 < 0$. These eigenvalues depend continuously on φ and this suggests that Theorem 6.3 should also hold for the full model when φ is small. A full analysis of the full model will not be presented here.

7. Conclusions

In this paper a new model of viral dynamics is presented in Section 2. These equations are given by (2-1a)–(2-1d). The healthy- and unhealthy-cell growth dynamics are logistic. In addition, this model includes both immune response behavior and the noncytolytic mediated curing of infected cells. The main parameter values are obtained from the literature and presented in Table 1. Simulations using these parameter values show that the model exhibits a range of typical behaviors including constant solutions as in Figure 1 and periodicity as in Figure 2. In Section 3, the model equations are nondimensionalized. The basic reproduction ratio R_0 is computed in Section 4 and defined in (4-4). Simulations in Figure 3 show how healthy and unhealthy-cell densities depend on R_0 and it is shown that a necessary and sufficient condition that the disease-free equilibrium E_f is locally asymptotically stable is that $R_0 < 1$. The disease equilibrium E_d is computed in Section 5 and Figures 4 and 5 illustrate that there may be critical parameter values for which E_d bifurcates from a unique positive value to two positive values. For the parameter values used in these figures it can also be seen that the unique disease equilibrium is locally asymptotically stable and in the case of nonuniqueness, the smaller equilibrium is unstable, while the larger equilibrium is locally asymptotically stable. In Section 6 a reduced system is analyzed by considering the immune response variable α to be at its steady-state value and eliminating the equation for α in the full model. This reduced system approximates the full model when the cytolytic clearance rate b or the ratio c/d of CTL response rate to CTL death rate is small. The Routh–Hurwitz conditions are then used to determine parameter values that guarantee a unique, locally asymptotically stable disease equilibrium in this reduced system.

The basic reproduction ratio can be used to understand how the parameters influence the disease state. Substituting dimensional parameters from (3-2) and (3-3) into (4-4) yields

$$R_0 = \frac{\beta K \gamma d}{\mu b I_0}.$$

Thus, R_0 decreases with bI_0/d. This ratio controls the effects of the immune response dynamics. In particular, increasing the rate at which the cytolytic immune response occurs, b, or the basal level of immune response cells, I_0/d, decreases the total density of unhealthy hepatocytes, Y, and increases the total density of healthy hepatocytes, X. When the healthy-cell infection rate β is relatively large, as in Figure 5, noncytolytic response controlled by the scaling parameter f in (2-1a) is more significant than cytolytic response b. On the other hand when β is relatively small as in Figure 6, healthy cells remain near carrying capacity for all values of f, but increase rapidly with the cytolytic clearance rate. Thus these simulations suggests that the healthy-cell infection rate may play a significant role in the effectiveness of cytolytic vs. noncytolytic immune response.

The results presented here provide only a partial analysis of the model (2-1a)–(2-1d). For example, only local asymptotic stability is considered and neither the nonunique disease equilibrium nor periodic behavior (as in Figure 2) have been investigated in detail. Thus future work should include examining global stability of E_f, analyzing local asymptotic and global stability of E_d in the full system (2-1a)–(2-1d), and further investigations of the limit cycle behavior.

Acknowledgements

The authors would like to acknowledge Sam Houston State University's College of Science and Engineering Technology for its support and funding for an undergraduate research project which led to this manuscript.

References

[Allen 2007] L. J. S. Allen, *Introduction to mathematical biology*, Pearson, Upper Saddle River, NJ, 2007.

[Ciupe et al. 2007] S. M. Ciupe, R. M. Ribeiro, P. W. Nelson, G. Dusheiko, and A. S. Perelson, "The role of cells refractory to productive infection in acute hepatitis B viral dynamics", *Proc. Natl. Acad. Sci. USA* **104**:12 (2007), 5050–5055.

[Diekmann et al. 1990] O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz, "On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations", *J. Math. Biol.* **28**:4 (1990), 365–382. MR Zbl

[Edelstein-Keshet 1988] L. Edelstein-Keshet, *Mathematical models in biology*, Random House, New York, 1988. MR Zbl

[Hews et al. 2010] S. Hews, S. Eikenberry, J. D. Nagy, and Y. Kuang, "Rich dynamics of a hepatitis B viral infection model with logistic hepatocyte growth", *J. Math. Biol.* **60**:4 (2010), 573–590. MR Zbl

[Kim et al. 2012] H. Y. Kim, H.-D. Kwon, T. S. Jang, J. Lim, and H.-S. Lee, "Mathematical modeling of triphasic viral dynamics in patients with HBeAg-positive chronic hepatitis B showing response to 24-week clevudine therapy", *PLos ONE* **7**:11 (2012), art. id. e50377.

[Lavanchy 2004] D. Lavanchy, "Hepatitis B virus epidemiology, disease burden, treatment, and current and emerging prevention and control measures", *J. Viral Hepatitis* **11**:2 (2004), 97–107.

[Locarnini et al. 2015] S. Locarnini, A. Hatzakis, D. S. Chen, and A. Lok, "Strategies to control hepatitis B: public policy, epidemiology, vaccine and drugs", *J. Hepatology* **62**:1, Suppl. (2015), S76–S86.

[Nowak and Bangham 1996] M. A. Nowak and C. R. M. Bangham, "Population dynamics of immune responses to persistent viruses", *Science* **272**:5258 (1996), 74–79.

[Nowak et al. 1996] M. A. Nowak, S. Bonhoeffer, A. M. Hill, R. Boehme, H. C. Thomas, and H. McDade, "Viral dynamics in hepatitis B virus infection", *Proc. Natl. Acad. Sci. USA* **93**:9 (1996), 4398–4402.

[Pang et al. 2012] J. Pang, J.-a. Cui, and J. Hui, "The importance of immune responses in a model of hepatitis B virus", *Nonlinear Dynam.* **67**:1 (2012), 723–734. MR Zbl

[Perelson 2002] A. S. Perelson, "Modelling viral and immune system dynamics", *Nature Rev. Immunology* **2** (2002), 28–36.

[Rodriguez et al. 2017] A. C. Rodriguez, M. Chung, and S. M. Ciupe, "Understanding the complex patterns observed during hepatitis B virus therapy", *Viruses* **9**:5 (2017), art. id. 117.

[Stanaway et al. 2016] J. D. Stanaway, A. D. Flaxman, M. Naghavi, C. Fitzmaurice, T. Vos, I. Abubakar, L. J. Abu-Raddad, R. Assadi, N. Bhala, B. Cowie, M. H. Forouzanfour, J. Groeger, K. M. Hanafiah, K. H. Jacobsen, S. L. James, J. MacLachlan, R. Malekzadeh, N. K. Martin, and G. S. Cooke, "The global burden of viral hepatitis from 1990 to 2013: findings from the Global Burden of Disease Study 2013", *The Lancet* **388**:10049 (2016), 1081–1088.

[Thieme 2003] H. R. Thieme, *Mathematics in population biology*, Princeton University Press, 2003. MR Zbl

[Vargas-De-León 2014] C. Vargas-De-León, "Global properties for a virus dynamics model with lytic and non-lytic immune responses, and nonlinear immune attack rates", *J. Biol. Systems* **22**:3 (2014), 449–462. MR Zbl

[Wodarza et al. 2002] D. Wodarza, J. P. Christensen, and A. R. Thomsen, "The importance of lytic and nonlytic immune responses in viral infections", *Trends in Immunology* **23**:4 (2002), 194–200.

[Zapata et al. 2014] H. D. T. Zapata, A. G. C. Casso, D. Bichara, and S. Lee, "Role of active and inactive cytotoxic immune response in human immunodeficiency virus dynamics", *Osong Pub. Health Res. Perspect.* **5**:1 (2014), 3–8.

Received: 2019-10-26 Revised: 2020-03-12 Accepted: 2020-03-14

jga001@shsu.edu *Department of Mathematics and Statistics,*
Sam Houston State University, Huntsville, TX, United States

smccoy10@vols.utk.edu *Department of Mathematics, University of Tennessee,*
Knoxville, TN, United States

INVOLVE 13:3 (2020)
https://doi.org/10.2140/involve.2020.13.475

A Cheeger inequality for graphs based on a reflection principle

Edward Gelernt, Diana Halikias, Charles Kenney and Nicholas F. Marshall

(Communicated by Glenn Hurlbert)

Given a graph with a designated set of boundary vertices, we define a new notion of a Neumann Laplace operator on a graph using a reflection principle. We show that the first eigenvalue of this Neumann graph Laplacian satisfies a Cheeger inequality.

1. Introduction and main result

1A. ***Introduction.*** Suppose that $G=(V, E)$ is a graph with vertices V and edges E. Let $\partial V \subseteq V$ be a designated set of boundary vertices, and $\mathring{V} := V \setminus \partial V$. We define the doubled graph G' as follows. Let $\mathring{G} = (U, F)$ be an isomorphic copy of the induced subgraph $G[\mathring{V}]$, and let f be an isomorphism from $\mathring{V}$ to U. Set

$$F' := \bigl\{\{u, v\} : u \in U,\ v \in \partial V,\ \{f^{-1}(u), v\} \in E\bigr\}.$$

Then, we define $G' := (V', E')$, where $V' := V \cup U$ and $E' := E \cup F \cup F'$. That is to say, G' is defined by making an isomorphic copy of the interior of G and attaching it to the boundary vertices ∂V as in the original graph; see Figure 1.

Definition 1.1. Let $G' = (V', E')$ be a doubled graph, and let $f : \mathring{V} \to U$ be an isomorphism as above, so that for all $w \in \partial V$ and $v \in \mathring{V}$, $\{v, w\} \in E'$ if and only if $\{f(v), w\} \in E'$. We say that a function $\varphi : V' \to \mathbb{R}$ is even with respect to ∂V if

$$\varphi(v) = \varphi(f(v)) \quad \text{for } v \in \mathring{V},$$

and we say that φ is odd with respect to ∂V if

$$\varphi(v) = -\varphi(f(v)) \quad \text{for } v \in \mathring{V} \qquad \text{and} \qquad \varphi(v) = 0 \quad \text{for } v \in \partial V.$$

Let $L' := D - A$ denote the graph Laplacian of G', where D is the degree matrix of G', and A is the adjacency matrix of G'. The following proposition characterizes the eigenvectors of L' as either even or odd.

MSC2010: primary 05C50, 05C85; secondary 15A42.
Keywords: Cheeger inequality, graph Laplacian, Neumann Laplacian.

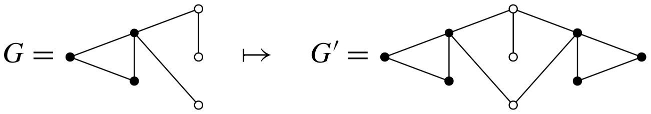


Figure 1. A graph G, and its doubled graph G', where the black and white dots denote interior and boundary vertices, respectively.

Proposition 1.2. *The graph Laplacian L' has $|V|$ eigenvectors that are even with respect to ∂V, and $|\mathring{V}|$ eigenvectors that are odd with respect to ∂V; this accounts for all eigenvectors of L'.*

1B. ***Motivation.*** We are motivated by the observation that the restrictions of the odd and even eigenvectors of L' to the graph G seem like natural Dirichlet and Neumann Laplacian eigenvectors for the graph G, given the respective odd and even behavior of Dirichlet and Neumann Laplacian eigenfunctions on manifolds. In fact, the restriction of the odd eigenvectors of L' to the graph G are eigenvectors of the Dirichlet graph Laplacian defined in [Chung 1997], and inequalities involving the eigenvalues of this operator have been investigated [Chung and Oden 2000]. However, an operator corresponding to the restriction of the even eigenvectors of L' to G has not, to our knowledge been investigated. Chung [1997] defined the Neumann graph Laplacian by enforcing a condition that a discrete derivative vanishes on the boundary nodes of the graph, which results in different eigenvectors than those arising from the even eigenvectors of L'. We note that a Cheeger inequality for Chung's definition of the Neumann graph Laplacian has recently been established in [Hua and Huang 2018].

1C. ***Odd and even eigenvectors.*** The proof of Proposition 1.2 gives some initial insight into the odd and even eigenvectors the graph Laplacian L' on the doubled graph G'.

Proof of Proposition 1.2. The proof is immediate from the block structure of the graph Laplacian L'. Indeed, let $L'(U, W)$ denote the submatrix of L' whose rows and columns are indexed by $U \subseteq V$ and $W \subseteq V$, respectively. We can write

$$L' = \begin{pmatrix} X & Y & 0 \\ Y^\top & Z & Y^\top \\ 0 & Y & X \end{pmatrix},$$

where X is the submatrix $L'(\mathring{V}, \mathring{V})$, Y is the submatrix $L'(\mathring{V}, \partial V)$, and Z is the submatrix $L'(\partial V, \partial V)$. With this notation, the eigenvectors of L' that are even with respect to ∂V are solutions to the equation

$$\begin{pmatrix} X & Y & 0 \\ Y^\top & Z & Y^\top \\ 0 & Y & X \end{pmatrix} \begin{pmatrix} u \\ v \\ u \end{pmatrix} = \mu \begin{pmatrix} u \\ v \\ u \end{pmatrix}.$$

That is to say, the vectors u and v satisfy $Xu + Yv = \mu u$ and $2Y^\top u + Zv = \mu v$. Put differently, when concatenated, u and v form an eigenvector of the matrix

$$L_R := \begin{pmatrix} X & Y \\ 2Y^\top & Z \end{pmatrix}. \tag{1}$$

Observe that L_R is similar to a symmetric matrix,

$$L_R = \begin{pmatrix} I & 0 \\ 0 & \sqrt{2}I \end{pmatrix} \begin{pmatrix} X & \sqrt{2}Y \\ \sqrt{2}Y^\top & Z \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & \sqrt{2}I \end{pmatrix}^{-1},$$

and thus by the spectral theorem, L_R has $|V|$ real eigenvectors, which give rise to $|V|$ even eigenvectors of L'. The eigenvectors of L' that are odd with respect to ∂V are solutions to the equation

$$\begin{pmatrix} X & Y & 0 \\ Y^\top & Z & Y^\top \\ 0 & Y & X \end{pmatrix} \begin{pmatrix} u \\ 0 \\ -u \end{pmatrix} = \lambda \begin{pmatrix} u \\ 0 \\ -u \end{pmatrix}.$$

Thus, each vector u such that $Xu = \lambda u$ gives rise to an odd eigenvector of L'. Let

$$L_D := X.$$

Since L_D is symmetric, it follows from the spectral theorem that it has $|\mathring{V}|$ real eigenvectors, and we conclude that L' has $|\mathring{V}|$ odd eigenvectors. □

1D. ***Contribution.*** In this paper, we study the operator L_R defined in (1), which we call the reflected Neumann graph Laplacian. This operator seems to be particularly natural on graphs approximating manifolds. For example, in Remark 1.3, we show that on the path graph, the eigenvectors of the Dirichlet graph Laplacian L_D and the reflected Neumann graph Laplacian L_R are the familiar discrete sine and cosine functions. We remark that the definition of the reflected Neumann graph Laplacian L_R has some similarities to the normalization used in the diffusion maps manifold learning method of [Coifman and Lafon 2006].

Our main result Theorem 1.4 shows that the first eigenvalue of the normalized reflected Neumann graph Laplacian $\mathcal{L}_R$ defined in (2) satisfies a Cheeger inequality. The graph cuts arising from $\mathcal{L}_R$ can differ significantly from graph cuts arising from the standard normalized graph Laplacian $\mathcal{L}$ defined in [Chung 1997]. In Figure 3, we illustrate Theorem 1.4 with an example where the first eigenvector of the Neumann graph Laplacian L_R suggests a drastically different cut than the first eigenvector of the standard graph Laplacian, and describe how the graph cut suggested by L_R is consistent with the Cheeger inequality established in Theorem 1.4. It may be interesting to investigate the analog of other classical eigenvalue inequalities involving these definitions of L_D and L_R for graphs with boundary.

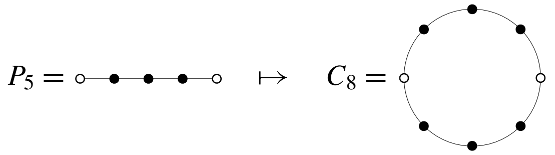


Figure 2. A path graph and its doubled graph.

Remark 1.3. The operators L_D and L_R are particularly natural on the path graph. Let $P_n = (V, E)$ denote the path graph on n vertices, where $V = \{1, \dots, n\}$ and $\{u, v\} \in E$ if and only if $|u - v| = 1$. If $\partial V := \{1, n\}$, then the doubled graph $P_n' = C_{2n-2}$ is the cycle graph on $2n - 2$ vertices; see Figure 2.

Consider L_D and L_R of the path graph P_n. The Dirichlet eigenvectors φ_k and eigenvalues λ_k, which satisfy $L_D \varphi_k = \lambda_k \varphi_k$ for $k = 1, \dots, n-2$, are of the form

$$\lambda_k = 2\left(1 - \cos\left(\frac{\pi k}{n-1}\right)\right) \quad \text{and} \quad \varphi_k(j) = \sin\left(\frac{\pi j k}{n-1}\right)$$

for $j = 1, \dots, n-2$, while the Neumann eigenvectors, ψ_k and μ_k, which satisfy $L_R \psi_k = \mu_k \psi_k$ for $k = 0, \dots, n-1$, are of the form

$$\mu_k = 2\left(1 - \cos\left(\frac{\pi k}{n-1}\right)\right) \quad \text{and} \quad \psi_k(j) = \cos\left(\frac{\pi j k}{n-1}\right)$$

for $j = 0, \dots, n-1$. Thus, the path graph doubling procedure defined in Section 1A gives the familiar sine and cosine functions, which are the Dirichlet and Neumann eigenfunctions of the Laplace operator of the unit interval.

1E. ***Notation and definitions.*** Suppose that $G = (V, E)$ is a graph with vertices V and edges E. Let $\partial V \subseteq V$ be a designated set of boundary vertices, and set $\mathring{V} = V \setminus \partial V$. We can write the adjacency matrix A of the graph G as the block matrix

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{12}^\top & A_{22} \end{pmatrix},$$

where $A_{11} = A(\mathring{V}, \mathring{V})$, $A_{12} = A(\mathring{V}, \partial V)$, and $A_{22} = A(\partial V, \partial V)$. Motivated by Proposition 1.2 we define the reflected adjacency matrix R by

$$R := \begin{pmatrix} A_{11} & A_{12} \\ 2A_{12}^\top & A_{22} \end{pmatrix}.$$

With this notation, the reflected Neumann Laplacian L_R can be defined by

$$L_R = D - R,$$

where $D = \operatorname{diag}(R\vec{1})$, where $\vec{1}$ denotes a vector whose entries are all 1, and whose dimensions are such that the matrix-vector multiplication is well-defined. We define the normalized reflected Neumann graph Laplacian $\mathcal{L}_R$ by

$$\mathcal{L}_R := D^{-1/2} L_R D^{-1/2}. \tag{2}$$

1F. ***Main result.*** In this section, we present our main result Theorem 1.4. While the matrix $\mathcal{L}_R$ is not in general symmetric, it is similar to a symmetric matrix; indeed, if

$$Q := \begin{pmatrix} I_{|\mathring{V}|} & 0 \\ 0 & \frac{1}{2} I_{|\partial V|} \end{pmatrix},$$

then the matrix $Q^{1/2} \mathcal{L}_R Q^{-1/2}$ is symmetric, positive-definite, and has the eigenvector $D^{1/2} Q^{1/2} \vec{1}$ of eigenvalue 0. It follows that the first nontrivial eigenvalue λ_R of $\mathcal{L}_R$ satisfies

$$\lambda_R := \inf_{x^\top D^{1/2} Q^{1/2} \vec{1} = 0} \frac{x^\top Q^{1/2} \mathcal{L}_R Q^{-1/2} x}{x^\top x}.$$

Let $E(U, W) := \{\{u, w\} \in E : u \in U,\ w \in W\}$; that is, $E(U, W)$ is the set of edges between U and W. We define a measure $m(U, W)$ on this set of edges by

$$m(U, W) = |E(U, W)| - \tfrac{1}{2} |E(U \cap \partial V, W \cap \partial V)|,$$

and we define the volume $\mathrm{vol}(U)$ of $U \subseteq V$ by

$$\mathrm{vol}(U) := \sum_{u \in U} m(\{u\}, V).$$

The following theorem is our main result.

Theorem 1.4. *Suppose that $G = (V, E)$ is a graph with a designated set of boundary vertices $\partial V \subseteq V$, and define the Cheeger constant h_R by*

$$h_R := \min_{S \subseteq V} \frac{m(S, V \setminus S)}{\min\{\mathrm{vol}(S), \mathrm{vol}(V \setminus S)\}}. \tag{3}$$

Then,

$$\sqrt{2\lambda_R} \geq h_R \geq \tfrac{1}{2} \lambda_R,$$

where λ_R is the first nontrivial eigenvalue of $\mathcal{L}_R$.

Recall that the standard Cheeger inequality is constructive in the sense that a cut that achieves the upper bound on the Cheeger constant can be determined from the eigenfunction corresponding to the first eigenvalue of the normalized graph Laplacian $\mathcal{L}$; see [Alon 1986; Cheeger 1970]. Specifically, a partition that achieves the upper bound can be determined by dividing the vertices into two groups based on if the value of the first eigenvector is more or less than some threshold; for a detailed exposition see for example [Chung 1997; 2007]. Similarly, the result of Theorem 1.4 is constructive in the sense that a cut which achieves the upper bound on h_R can be determined from the eigenvector ψ_R of $\mathcal{L}_R$ that corresponds to λ_R.

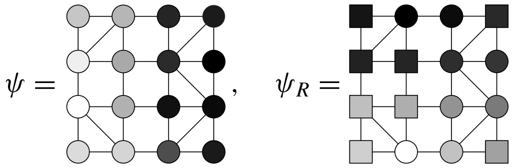


Figure 3. The same graph with vertices colored proportional to ψ (left) and colored proportional to ψ_R (right), where the squares in the right graph denote boundary vertices.

In the following remark, we present an example where the cut arising from ψ_R differs significantly from the cut arising from the first eigenvector ψ of the standard normalized graph Laplacian $\mathcal{L}$.

Remark 1.5. Graph cuts arising from ψ_R can differ significantly from graph cuts arising from ψ. Indeed, on the left of Figure 3 we illustrate a graph whose vertices are colored by greyscale values proportional to ψ. On the right of Figure 3 we illustrate the same graph except several vertices have been designated as boundary vertices (indicated by squares) and the color of the vertices is proportional to ψ_R. Observe that ψ suggests cutting the graph by a vertical line into two equal parts, while ψ_R suggests cutting the graph by a horizontal line into two equal parts.

That ψ_R suggests a horizontal cut of the graph is illustrative of Theorem 1.4. Indeed, it is straightforward to check that the horizontal cut suggested by ψ_R minimizes the cut measure $m(S, V \setminus S)/(\text{vol}(S), \text{vol}(V \setminus S))$ from (3). In contrast, the vertical cut suggested by ψ minimizes the standard cut measure, which is equivalent to the measure $m(S, V \setminus S)/(\text{vol}(S), \text{vol}(V \setminus S))$ in the case that all vertices are interior vertices. Of course, Theorem 1.4 only guarantees that the measure of the cut arising from the eigenvector ψ_R is an upper bound for h_R with value at most $\sqrt{2\lambda_R}$; however, in this simple example the cut arising from ψ_R actually obtains this minimum.

Remark 1.6. Here we visualize the first eigenfunction ψ_R of the reflected Neumann graph Laplacian $\mathcal{L}_R$ on a classic barbell shaped graph; see Figure 4. Observe that in Figure 4 the maximum and minimum value of the eigenvector occur at an interior vertex. This feature of the eigenvectors is interesting in the context of spectral

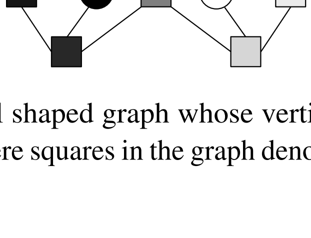


Figure 4. A barbell shaped graph whose vertices are colored proportional to ψ_R, where squares in the graph denote boundary vertices.

clustering, where extreme values of the eigenvectors often correspond to the center of clusters.

1G. ***Future directions.*** One future direction for this work is the problem of selecting boundary vertices in a principled way. How the boundary is selected may depend on the application at hand. In a social network graph, boundary vertices could correspond to individuals with many connections outside the network. In the context of manifold learning, where the vertices of the graph are points in $\mathbb{R}^n$, boundary vertices could be selected based on the number of points within some ε-neighborhood of each vertex. On the other hand, when a graph is given by sampling from a predefined manifold with boundary, vertices selected from some collar neighborhood of the boundary could be designated as boundary vertices.

Another future direction arises from generalizing the setup under which our work was done. Our graph doubling procedure inputs a graph with boundary and outputs a larger graph, containing the original graph as an induced subgraph, which has a special Z_2 symmetry. Could similar Cheeger results be proven for other reflection procedures? For example, what if $n-1$ copies of the interior vertices were attached, instead of only 1?

Finally, we note a connection between the doubled graph (defined in Section 1A) and numerical analysis that may motivate a direction for future study. Recall that for a path graph P_n the eigenfunctions of the reflected Neumann Laplacian L_R are of the form $\psi_k(j) = \cos(\pi jk/(n-1))$; see Remark 1.3. These Neumann eigenvectors are precisely the basis vectors for the discrete cosine transform (DCT) type I, as classified in [Strang 1999]. The DCT type II, which has basis vectors $\psi_k(j) = \cos\bigl(\pi\bigl(j+\frac{1}{2}\bigr)k/n\bigr)$ is also important in numerical analysis; it could be interesting to develop a graph doubling procedure whose Neumann eigenvectors on the path graph are these vectors.

2. Proof of main result

2A. ***Summary.*** The proof of Theorem 1.4 is divided into two lemmas: first, in Lemma 2.1 we show that $\lambda_R \leq 2h_R$, and second, in Lemma 2.2 we show that $h_R^2/2 \leq \lambda_R$. The structure of our argument is similar to classical Cheeger inequality proofs; see [Chung 1996; 1997].

2B. ***Proof of Theorem 1.4.***

Lemma 2.1 (trivial direction). *We have*

$$\lambda_R \leq 2h_R.$$

Proof of Lemma 2.1. Recall that

$$\mathcal{L}_R := D^{-1/2}Q^{1/2}L_R Q^{-1/2}D^{-1/2}.$$

First, we observe that QL_R can be written as

$$QL_R = L - \tfrac{1}{2}L_\partial,$$

where

$$L = \begin{pmatrix} \operatorname{diag}(A_{11}\vec{1}+A_{12}\vec{1})-A_{11} & -A_{12} \\ -A_{12}^\top & \operatorname{diag}(A_{12}^\top\vec{1}+A_{22}\vec{1})-A_{22} \end{pmatrix},$$

and

$$L_\partial := \begin{pmatrix} 0 & 0 \\ 0 & \operatorname{diag}(A_{22}\vec{1})-A_{22} \end{pmatrix}.$$

Observe that L is the standard graph Laplacian of G, while L_∂ is the graph Laplacian of the vertex induced subgraph $G[\partial V]$. Fix a subset $S \subseteq V$, and let χ_S be the indicator function for S. Define

$$x := Q^{1/2}D^{1/2}\chi_S - \frac{\chi_S^\top DQ\vec{1}}{\vec{1}^\top DQ\vec{1}} D^{1/2}Q^{1/2}\vec{1}.$$

By construction, we have $x^\top D^{1/2}Q^{1/2}\vec{1} = 0$, and it follows that

$$\begin{aligned}
\lambda_N &\le \frac{x^\top D^{-1/2}Q^{1/2}L_R Q^{-1/2}D^{-1/2}x}{x^\top x} \\
&= \frac{\chi_S^\top QL_R\chi_S}{\chi_S^\top DQ\chi_S(\vec{1}-\chi_S^\top DQ\chi_S/(\vec{1}^\top DQ1))} = \frac{\chi_S^\top\big(L-\frac{1}{2}L_\partial\big)\chi_S(\vec{1}^\top DQ\vec{1})}{(\chi_S^\top DQ\chi_S^\top)(\chi_{V\setminus S}^\top DQ\chi_{V\setminus S})} \\
&\le \frac{2\cdot\chi_S^\top\big(L-\frac{1}{2}L_\partial\big)\chi_S}{\min\{(\chi_S^\top DQ\chi_S^\top),(\chi_{V\setminus S}^\top DQ\chi_{V\setminus S})\}} = \frac{2\cdot m(S, V\setminus S)}{\min\{\operatorname{vol}(S),\operatorname{vol}(V\setminus S)\}}.
\end{aligned}$$

Since this inequality holds for all subsets $S \subseteq V$, we conclude that $\lambda_R \le 2h_R$. □

Lemma 2.2 (nontrivial direction). *We have*

$$\lambda_R \ge \tfrac{1}{2}h_R^2.$$

Proof of Lemma 2.2. Recall that

$$\lambda_R = \inf_{x^\top D^{1/2}Q^{1/2}\vec{1}=0} \frac{x^\top \mathcal{L}_R x}{x^\top x} = \inf_{y^\top DQ\vec{1}=0} \frac{y^\top QL_R y}{y^\top QDy}.$$

Let g be a vector satisfying

$$\lambda_R = \frac{g^\top QL_R g}{g^\top DQg} \quad \text{and} \quad g^\top QD\vec{1} = 0.$$

Let $\{v_1, \dots, v_n\}$ be an enumeration of the vertices V so that $g_{v_1} \le \cdots \le g_{v_n}$ and set $S_j := \{v_1, \dots, v_j\}$, for $j = 1, \dots, n$. Let p be the largest integer such that

$\operatorname{vol}(S_p) \le \frac{1}{2}\operatorname{vol}(V)$; that is,

$$p := \max\bigl\{j \in \{1, \dots, n\} : \operatorname{vol}(S_j) \le \tfrac{1}{2}\operatorname{vol}(V)\bigr\}.$$

Let g^+ and g^- denote the positive and negative parts of $g - g_{v_p}$, respectively. That is, $g_v^+ := \max\{g_v - g_{v_p}, 0\}$ and $g_v^- := \max\{g_{v_p} - g_v, 0\}$. Let $u \sim v$ denote $\{u, v\} \in E$ and $q = \operatorname{diag}(Q)$. Then

$$\begin{aligned}\lambda_R &= \frac{g^\top(L - \frac{1}{2}L_\partial)g}{g^\top DQg} = \frac{\sum_{u\sim v}(g_u - g_v)^2 - \frac{1}{2}\sum_{\substack{u\sim v\\ u,v\in\partial V}}(g_u - g_v)^2}{\sum_v g_v^2 d_v q_v}\\ &\ge \frac{\sum_{u\sim v}(g_u - g_v)^2 - \frac{1}{2}\sum_{\substack{u\sim v\\ u,v\in\partial V}}(g_u - g_v)^2}{\sum_v (g(v) - g(v_p))^2 d_v q_v},\end{aligned}$$

where the last inequality holds because we have increased the denominator. From here,

$$\lambda_R \ge \frac{\sum_{u\sim v}((g_u^+ - g_v^+)^2 + (g_u^- - g_v^-)^2) - \frac{1}{2}\sum_{\substack{u\sim v\\ u,v\in\partial V}}((g_u^+ - g_v^+)^2 + (g_u^- - g_v^-)^2)}{\sum_v ((g_v^+)^2 + (g_v^-)^2) d_v q_v}. \tag{4}$$

Recall that

$$\frac{a+b}{c+d} \ge \min\left\{\frac{a}{c}, \frac{b}{d}\right\} \tag{5}$$

for any $a, b \ge 0$ and $c, d > 0$. From (4), we can set

$$\begin{aligned} a &= \sum_{u\sim v}(g_u^+ - g_v^+)^2 - \sum_{\substack{u\sim v\\ u,v\in\partial V}}(g_u^+ - g_v^+)^2, & c &= \sum_v (g_v^+)^2 d_v q_v,\\ b &= \sum_{u\sim v}(g_u^- - g_v^-)^2 - \sum_{\substack{u\sim v\\ u,v\in\partial V}}(g_u^- - g_v^-)^2, & d &= \sum_v (g_v^-)^2 d_v q_v.\end{aligned}$$

Observe that a and b are nonnegative. Indeed,

$$a = \sum_{\substack{u\sim v\\ u\notin\partial V \text{ or } v\notin\partial V}}(g_u^+ - g_v^+)^2,$$

which has nonnegative summands, and a similar statement holds for b.

Without loss of generality, (5) implies

$$\lambda_R \ge \frac{\sum_{u\sim v}(g_u^+ - g_v^+)^2 - \frac{1}{2}\sum_{\substack{u\sim v\\ u,v\in\partial V}}(g_u^+ - g_v^+)^2}{\sum_v (g_v^+)^2 d_v q_v}.$$

To simplify notation in the following, let $f = g^+$. We begin by setting

$$\lambda := \frac{\sum_{u\sim v}(f_u - f_v)^2 - \frac{1}{2}\sum_{\substack{u\sim v\\ u,v\in\partial V}}(f_u - f_v)^2}{\sum_v f_v^2 d_v q_v}.$$

Multiplying the numerator and denominator by the same term gives

$$\lambda = \frac{\big(\sum_{u\sim v}(f_u-f_v)^2-\frac{1}{2}\sum_{\substack{u\sim v\\u,v\in\partial V}}(f_u-f_v)^2\big)\big(\sum_{u\sim v}(f_u+f_v)^2-\frac{1}{2}\sum_{\substack{u\sim v\\u,v\in\partial V}}(f_u+f_v)^2\big)}{\big(\sum_v f_v^2 d_v q_v\big)\big(\sum_{u\sim v}(f_u+f_v)^2-\frac{1}{2}\sum_{\substack{u\sim v\\u,v\in\partial V}}(f_u+f_v)^2\big)}.$$

Applying the Cauchy–Schwarz inequality in the numerator gives

$$\lambda \ge \frac{\big(\sum_{u\sim v}|f_u^2-f_v^2|-\frac{1}{2}\sum_{\substack{u\sim v\\u,v\in\partial V}}|f_u^2-f_v^2|\big)^2}{\big(\sum_v f_v^2 d_v q_v\big)\big(\sum_{u\sim v}(f_u+f_v)^2-\frac{1}{2}\sum_{\substack{u\sim v\\u,v\in\partial V}}(f_u+f_v)^2\big)}.$$

Next, we observe that

$$\sum_{u\sim v}(f_u+f_v)^2-\frac{1}{2}\sum_{\substack{u\sim v\\u,v\in\partial V}}(f_u+f_v)^2=\sum_v f_v^2 d_v q_v-\Bigg(\sum_{u\sim v}(f_u-f_v)^2-\frac{1}{2}\sum_{\substack{u\sim v\\u,v\in\partial V}}(f_u-f_v)^2\Bigg),$$

and thus it follows that

$$\lambda \ge \frac{\big(\sum_{u\sim v}|f_u^2-f_v^2|-\frac{1}{2}\sum_{\substack{u\sim v\\u,v\in\partial V}}|f_u^2-f_v^2|\big)^2}{\big(\sum_v f_v^2 d_v q_v\big)^2(2-\lambda)}.$$

We want to show

$$\sum_{u\sim v}|f_u^2-f_v^2|-\frac{1}{2}\sum_{\substack{u\sim v\\u,v\in\partial V}}|f_u^2-f_v^2|\ge\sum_{i=1}^{n}|f_{v_i}^2-f_{v_{i+1}}^2|m(S_i,V\setminus S_i).$$

We can write

$$\sum_{u\sim v}|f_u^2-f_v^2|-\frac{1}{2}\sum_{\substack{u\sim v\\u,v\in\partial V}}|f_u^2-f_v^2|=\sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(\chi_{E_{i,j}}-\frac{\chi_{\partial_i}\chi_{\partial_j}}{2}\right)(f_{v_i}^2-f_{v_j}^2),$$

where

$$\chi_{E_{i,j}}=\begin{cases}1 & \text{if } \{v_i,v_j\}\in E,\\ 0 & \text{otherwise}\end{cases}$$

is the indicator function for $\{v_i, v_j\} \in E$, and

$$\chi_{\partial_i}=\begin{cases}1 & \text{if } i\in\partial V,\\ 0 & \text{otherwise}\end{cases}$$

is the indicator function for $v_i \in \partial V$. Note that we are justified in dropping the absolute value signs because $f_{v_i}^2$ is an increasing function of i. Next we write $f_{v_i}^2 - f_{v_j}^2$ as a telescoping series

$$f_{v_i}^2-f_{v_j}^2=(f_{v_i}^2-f_{v_{i-1}}^2)+(f_{v_{i-1}}^2-f_{v_{i-2}}^2)+\cdots+(f_{v_{j+1}}^2-f_{v_j}^2),$$

and rearrange terms in the summation to conclude that

$$\sum_{i=2}^{n}\sum_{j=1}^{i-1}\biggl(\chi_{E_{i,j}}-\frac{\chi_{\partial_i}\chi_{\partial_j}}{2}\biggr)(f_{v_i}^2-f_{v_j}^2)\\ =\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{j=1}^{n}\biggl(\biggl(\chi_{E_{j,k+l}}-\frac{\chi_{\partial_j}\chi_{\partial_{k+l}}}{2}\biggr)\chi_{j\le l}\biggr)(f_{v_{l+1}}^2-f_{v_l}^2),$$

where

$$\chi_{j\le l}=\begin{cases}1 & \text{if } j\le l,\\ 0 & \text{otherwise.}\end{cases}$$

Then, to complete this step, we note that

$$\sum_{k=1}^{n}\sum_{j=1}^{n}\biggl(\biggl(\chi_{E_{j,k+l}}-\frac{\chi_{\partial_j}\chi_{\partial_{k+l}}}{2}\biggr)\chi_{j\le l}\biggr)=m(S_l,V\setminus S_l).$$

Returning to our main sequence of inequalities for λ, we have

$$\lambda\ge\frac{\bigl(\sum_i|f_{v_i}^2-f_{v_{i+1}}^2|m(S_i,V\setminus S_i)\bigr)^2}{2\bigl(\sum_v f_v^2d_vq_v\bigr)^2}\ge\frac{\bigl(\alpha\sum_{i=1}^n|f_{v_i}^2-f_{v_{i+1}}^2|\min\{\mathrm{vol}(S_i),\mathrm{vol}(V\setminus S_i)\}\bigr)^2}{2\bigl(\sum_u f(u)^2d_uq_v\bigr)^2},$$

where

$$\alpha:=\min_{1\le i\le n}\frac{m(S_i,V\setminus S_i)}{\min\{\mathrm{vol}(S_i),\mathrm{vol}(V\setminus S_i)\}}.$$

Since $f_{v_i}^2$ is nondecreasing, a rearrangement of the numerator of the previous expression gives

$$\lambda\ge\frac{\alpha^2}{2}\frac{\bigl(\sum_i(f_{v_i}^2|\min\{\mathrm{vol}(S_i),\mathrm{vol}(V\setminus S_i)\}-\min\{\mathrm{vol}(S_{i+1}),\mathrm{vol}(V\setminus S_{i+1})\}|)\bigr)^2}{\bigl(\sum_u f(u)^2d_uq_u\bigr)^2}.$$

It follows that

$$\lambda_R\ge\lambda\ge\frac{\alpha^2}{2}\frac{\bigl(\sum_i f_{v_i}^2d_{v_i}q_{v_i}\bigr)^2}{\bigl(\sum_u f_u^2d_uq_u\bigr)^2}=\frac{\alpha^2}{2}\ge\frac{h_R^2}{2},$$

which completes the proof. □

Acknowledgements

We thank the referees for their helpful comments. This research was supported by Summer Undergraduate Math Research at Yale (SUMRY) 2018. Marshall was supported in part by NSF DMS-1903015.

References

[Alon 1986] N. Alon, “Eigenvalues and expanders”, *Combinatorica* **6**:2 (1986), 83–96. MR Zbl

[Cheeger 1970] J. Cheeger, “A lower bound for the smallest eigenvalue of the Laplacian”, pp. 195–199 in *Problems in analysis* (Princeton, 1969), edited by R. C. Gunning, Princeton Univ. Press, 1970. MR Zbl

[Chung 1996] F. R. K. Chung, “Laplacians of graphs and Cheeger’s inequalities”, pp. 157–172 in *Combinatorics: Paul Erdős is eighty, II* (Keszthely, Hungary, 1993), edited by D. Miklós et al., Bolyai Soc. Math. Stud. **2**, János Bolyai Math. Soc., Budapest, 1996. MR Zbl

[Chung 1997] F. R. K. Chung, *Spectral graph theory*, CBMS Region. Conf. Ser. Math. **92**, Amer. Math. Soc., Providence, RI, 1997. MR Zbl

[Chung 2007] F. Chung, “Four Cheeger-type inequalities for graph partitioning algorithms”, pp. 751–772 in *Proceedings of the International Conference on Computational Methods* (Hiroshima, 2007), 2007.

[Chung and Oden 2000] F. R. K. Chung and K. Oden, “Weighted graph Laplacians and isoperimetric inequalities”, *Pacific J. Math.* **192**:2 (2000), 257–273. MR Zbl

[Coifman and Lafon 2006] R. R. Coifman and S. Lafon, “Diffusion maps”, *Appl. Comput. Harmon. Anal.* **21**:1 (2006), 5–30. MR Zbl

[Hua and Huang 2018] B. Hua and Y. Huang, “Neumann Cheeger constants on graphs”, *J. Geom. Anal.* **28**:3 (2018), 2166–2184. MR Zbl

[Strang 1999] G. Strang, “The discrete cosine transform”, *SIAM Rev.* **41**:1 (1999), 135–147. MR Zbl

Received: 2019-12-04 Revised: 2020-05-09 Accepted: 2020-05-23

edward.gelernt@yale.edu
Department of Mathematics, Yale University,
New Haven, CT, United States

diana.halikias@yale.edu
Department of Mathematics, Yale University,
New Haven, CT, United States

ctk47@math.rutgers.edu
Department of Mathematics, Rutgers University,
Piscataway, NJ, United States

nicholas.marshall@math.princeton.edu
Department of Mathematics, Princeton University,
Princeton, NJ, United States

INVOLVE 13:3 (2020)
https://doi.org/10.2140/involve.2020.13.487

Sets in $\mathbb{R}^d$ determining k taxicab distances

Vajresh Balaji, Olivia Edwards, Anne Marie Loftin,
Solomon Mcharo, Lo Phillips, Alex Rice and Bineyam Tsegaye

(Communicated by Anant Godbole)

We address an analog of a problem introduced by Erdős and Fishburn, itself an inverse formulation of the famous Erdős distance problem, in which the usual Euclidean distance is replaced with the metric induced by the ℓ^1-norm, commonly referred to as the *taxicab metric*. Specifically, we investigate the following question: given $d, k \in \mathbb{N}$, what is the maximum size of a subset of $\mathbb{R}^d$ that determines at most k distinct taxicab distances, and can all such optimal arrangements be classified? We completely resolve the question in dimension $d = 2$, as well as the $k = 1$ case in dimension $d = 3$, and we also provide a full resolution in the general case under an additional hypothesis.

1. Introduction

Erdős [1946] asked a now famous question: given $n \in \mathbb{N}$, what is the minimum number of distinct distances determined by n points in a plane? Denoting this minimum by $f(n)$, he proved via an elementary counting argument that $f(n) = \Omega(\sqrt{n})$, and he conjectured that the correct order of growth is $n/\sqrt{\log n}$, as attained by a $\sqrt{n} \times \sqrt{n}$ integer grid. After decades of incremental progress, this conjecture was effectively resolved in a celebrated result of Guth and Katz [2015], who established that $f(n) = \Omega(n/\log n)$.

Fifty years after Erdős's original paper, Erdős and Fishburn [1996] addressed the same question from the inverse perspective, and aspired to precise results in fixed cases rather than general asymptotic results. Specifically, they investigated the following: given $k \in \mathbb{N}$, what is the maximum number of points in a plane that determine at most k distinct distances, and can such optimal arrangements be classified? This question, which we refer to as the *Erdős–Fishburn problem*, was fully resolved by Erdős and Fishburn for $1 \le k \le 4$, then by Shinahara [2008] for $k = 5$, and Wei [2012] for $k = 6$, while it remains open for $k \ge 7$. By convention,

MSC2010: 52C10.
Keywords: Erdős distance problem, taxicab metric, discrete geometry, geometric combinatorics.

in the quoted results and throughout this paper, 0 is not counted as a distance determined by a set of points.

These questions can also be adapted to higher dimensions, and to alternative notions of distance. Here we focus on a particular, well-known alternative metric.

Definition 1.1. For $d \in \mathbb{N}$ and $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, we define the ℓ^1-*norm* of x by

$$\|x\|_1 = |x_1| + \cdots + |x_d|,$$

which in particular satisfies the *triangle inequality* $\|x + y\|_1 \leq \|x\|_1 + \|y\|_1$. Like every norm, the ℓ^1-norm induces a metric on $\mathbb{R}^d$ by defining $\|x - y\|_1$ to be the ℓ^1-*distance* between $x, y \in \mathbb{R}^d$.

The metric induced by the ℓ^1-norm is commonly referred to as the *taxicab metric*, because it measures the length of the shortest path between two points in space, under the restriction that one can only travel in directions parallel to the coordinate axes, as if in a taxicab on a grid of city streets. For example, if two people at city intersections are separated by 3 blocks horizontally and 4 blocks vertically, then, as the crow flies, they are 5 blocks apart by the Pythagorean theorem. However, to actually make the journey without cutting through buildings, they must walk 7 blocks, which is the ℓ^1-distance.

As noted in Chapters 0 and 1 of [Garibaldi, Iosevich, and Senger 2011], one can show that the minimum number of ℓ^1-distances determined by n points in $\mathbb{R}^d$ is $\Omega(n^{1/d})$, and this order of growth is attained by $\{1, 2, 3, \dots, \lceil n^{1/d} \rceil\}^d$. Therefore, in the case of the taxicab metric, the big picture asymptotic question is immediately resolved, which begs the question of whether this case can be analyzed more precisely. To begin this journey, we first consider the Erdős–Fishburn problem in the plane with $k = 1$.

We fix two points $U, V \in \mathbb{R}^2$, say $U = (-1, 0)$ and $V = (1, 0)$. With the usual notion of distance, if any additional points can be added without determining an additional distance, those points necessarily lie on the circles of radius 2 centered at U and V, respectively. Those two circles intersect in only two points, and we find that they are $Q = (0, \sqrt{3})$ and $R = (0, -\sqrt{3})$. Since the distance between Q and R is greater than 2, only one of the two can be added to $\{U, V\}$ while maintaining only a single distance. In summary, a set $P \subseteq R^2$ determining a single distance satisfies $|P| \leq 3$, and equality holds if and only if P is the set of vertices of an equilateral triangle.

However, even in this simplest case, the taxicab metric case diverges from that of the usual distance. With the taxicab metric, the "circle" (which we refer to as an ℓ^1-*circle*) of radius 2 centered at U is in fact a square, rotated 45° from axis-parallel, with the four sides connecting the points $(-3, 0)$, $(-1, 2)$, $(1, 0)$, and $(-1, -2)$. Similarly, the ℓ^1-circle of radius 2 centered at V is a square with sides connecting

$(3,0)$, $(1,-2)$, $(-1,0)$, and $(1,2)$. Like the usual-distance case, these two circles intersect in exactly two points, this time $Q=(0,1)$ and $R=(0,-1)$. The difference is that here Q and R are indeed separated by ℓ^1-distance 2, and hence the four-point configuration $\{U, V, Q, R\}$ determines a single ℓ^1-distance.

2. Main definition and results

Inspired by the four-point construction above, as well as additional trial and error, we define the following family of sets, which serve as our candidates for resolving the Erdős–Fishburn problem for the taxicab metric.

Definition 2.1. For integers $d>0$ and $k\geq 0$, we define

$$\Lambda_d(k)=\{n=(n_1,n_2,\dots,n_d)\in\mathbb{Z}^d : \|n\|_1\leq k,\ n_1+\cdots+n_d\equiv k \pmod 2\}.$$

$\Lambda_d(k)$ is the union of the integer lattice points lying on the ℓ^1-spheres (which in dimension d are 2^d-faced polytopes) centered at the origin of *every other* integer radius, starting with either 0 or 1 depending on the parity of k. The four-point configuration discussed in the Introduction is $\Lambda_2(1)$, and some additional examples are pictured in Figure 1.

In Section 3, we establish the following properties of $\Lambda_d(k)$, including the crucial fact that it determines exactly k distinct ℓ^1-distances, the primary motivation for its definition.

Theorem 2.2. *The following hold for all $d, k\in\mathbb{N}$:*

(i) $\Lambda_d(k)$ *determines exactly k distinct ℓ^1-distances, specifically* $2, 4, \dots, 2k$.

(ii) $|\Lambda_1(k)|=k+1$ *and* $|\Lambda_d(1)|=2d$.

(iii) $|\Lambda_{d+1}(k)|=|\Lambda_d(k)|+2\sum_{j=0}^{k-1}|\Lambda_d(j)|$.

Parts (ii) and (iii) of Theorem 2.2, combined with known formulas for sums of powers, allow one to determine explicit formulas for $|\Lambda_d(k)|$ for any fixed $d\in\mathbb{N}$. We include the first few examples in Table 1.

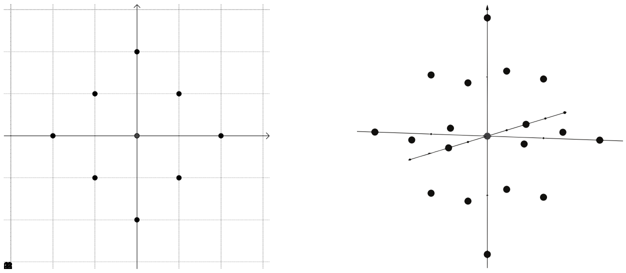


Figure 1. Left, $\Lambda_2(2)$: 9 points in $\mathbb{R}^2$ determining two ℓ^1-distances. Right, $\Lambda_3(2)$: 19 points in $\mathbb{R}^3$ determining two ℓ^1-distances.

d	$\lvert\Lambda_d(k)\rvert$
2	$(k+1)^2$
3	$\frac{2}{3}(k+1)^3+\frac{1}{3}(k+1)$
4	$\frac{1}{3}(k+1)^4+\frac{2}{3}(k+1)^2$
5	$\frac{2}{15}(k+1)^5+\frac{2}{3}(k+1)^3+\frac{1}{5}(k+1)$
6	$\frac{2}{45}(k+1)^6+\frac{4}{9}(k+1)^4+\frac{23}{45}(k+1)^2$
7	$\frac{4}{315}(k+1)^7+\frac{2}{9}(k+1)^5+\frac{28}{45}(k+1)^3+\frac{1}{7}(k+1)$

Table 1. Explicit formulas for $|\Lambda_d(k)|$.

Some of the patterns observed in Table 1 can be generalized using Faulhaber's formula for sums of powers, as seen in the following formulation, which we also prove in Section 3.

Theorem 2.3. *For each $d \in \mathbb{N}$ and each integer $k \geq 0$, we have the formula*

$$|\Lambda_d(k)| = \sum_{i=0}^{\lceil d/2\rceil -1} a_{d,i}(k+1)^{d-2i},$$

where the coefficients $a_{d,i}$ satisfy the recursive formula

$$a_{d,i} = 2\sum_{\ell=0}^{i} \frac{a_{d-1,\ell}}{d-2\ell}\binom{d-2\ell}{2(i-\ell)}B_{2(i-\ell)},$$

where B_i is the i-th Bernoulli number. In particular, we have the explicit formulas $a_{d,0} = 2^{d-1}/d!$ for all $d \in \mathbb{N}$ and $a_{d,1} = 2^{d-3}/(3(d-3)!)$ for all $d \geq 3$.

Detailed analysis of $\Lambda_d(k)$ is perhaps of independent interest, but to make headway toward our goal, we need to address the important questions: does $\Lambda_d(k)$ have maximal size amongst subsets of $\mathbb{R}^d$ determining at most k distinct ℓ^1-distances? If so, is $\Lambda_d(k)$ the only such optimal arrangement? In anticipation of the latter question, we observe that any optimal arrangement can undergo any scaling, or any transformation that preserves the ℓ^1-norm, and remain optimal, leading to the following definition.

Definition 2.4. For $d \in \mathbb{N}$, we say that two subsets of $\mathbb{R}^d$ are ℓ^1-*similar* if one can be mapped to the other via a composition of translations, reflections about coordinate hyperplanes, dilations, and coordinate permutations, as these transformations either preserve or uniformly scale collections of ℓ^1-distances.

We note that the list of transformations in Definition 2.4 does not include rotations, because, unlike the usual Euclidean metric, the taxicab metric is *not* invariant under rotation, unless the rotation can alternatively be obtained through reflection about coordinate hyperplanes and permutation of coordinates. This fact rears its head in our exploration of the taxicab metric in higher dimensions, and plays a key role in our discussions in Section 6. For now, though, the following result established in Section 4 completely resolves the taxicab analog of the Erdős–Fishburn problem in the plane.

Theorem 2.5. *If $k \in \mathbb{N}$ and $P \subseteq \mathbb{R}^2$ determines at most k distinct ℓ^1-distances, then $|P| \leq (k+1)^2$. Further, $|P| = (k+1)^2$ if and only if P is ℓ^1-similar to $\Lambda_2(k)$.*

As we discuss in Section 4, the $d = 2$ case is simplified by the fact that, for the purposes of analyzing distance sets, the ℓ^1-norm in $\mathbb{R}^2$ is effectively the same as the ℓ^∞-norm defined by $\|(x, y)\|_\infty = \max\{|x|, |y|\}$. However, this equivalence does not persist in dimension $d \geq 3$, and for this reason, our proof strategy does not immediately generalize to higher dimensions. (Although, for the interested reader, the proof does generalize to show that if $P \subseteq \mathbb{R}^d$ determines at most k distinct ℓ^∞-distances, then $|P| \leq (k+1)^d$, and equality holds if and only if P is ℓ^1-similar to $\{0, 1, 2, \dots, k\}^d$.)

With considerable additional effort, we successfully get our foot into the higher-dimensional door in Section 5, which assures us that the unique optimality of $\Lambda_d(k)$ is not completely dependent on a connection to the ℓ^∞-norm.

Theorem 2.6. *If $P \subseteq \mathbb{R}^3$ determines a single ℓ^1-distance, then $|P| \leq 6$. Further, $|P| = 6$ if and only if P is ℓ^1-similar to $\Lambda_3(1)$.*

Remark on previous work for $k = 1$. After the initial posting of this paper to the arXiv server, we were alerted to previous work done in the $k = 1$ case (referred to as *equilateral sets*) in a variety of metric spaces, including $\mathbb{R}^d$ with the taxicab metric (referred to as *rectilinear space*). Specifically, Theorem 2.6 above follows from Corollary 4.2 of [Bandelt, Chepoi, and Laurent 1998], while Koolen, Laurent, and Schrijver [2000] showed that if $P \subseteq \mathbb{R}^4$ determines a single ℓ^1-distance, then $|P| \leq 8 = |\Lambda_4(1)|$. This partially settles a question of Kusner (see Problem 0 in [Guy 1983]), who asked if $|P| \leq 2d = |\Lambda_d(1)|$ holds for subsets of $\mathbb{R}^d$ determining a single ℓ^1-distance, and this remains open for $d \geq 5$. Conjecture 2.7 below can be thought of as a precise, multidistance generalization of Kusner's question. While the conclusion of Theorem 2.6 was known previously, we believe our alternative, elementary proof given in Section 5 remains of interest.

In Section 6, we explore the question of what additional hypotheses are required to prove the optimality of $\Lambda_3(k)$ for all $k \in \mathbb{N}$, or even $\Lambda_d(k)$ in full generality. We find that the proof of Theorem 2.5 can be fully adapted with a seemingly mild additional assumption, leading us to make the following general conjecture.

Conjecture 2.7. *If $d, k \in \mathbb{N}$ and $P \subseteq \mathbb{R}^d$ determines at most k distinct ℓ^1-distances, then $|P| \leq |\Lambda_d(k)|$. Further, $|P| = |\Lambda_d(k)|$ if and only if P is ℓ^1-similar to $\Lambda_d(k)$.*

3. Properties of $\Lambda_d(k)$: proofs of Theorems 2.2 and 2.3

We begin this section by proving the essential properties of $\Lambda_d(k)$ that make it a worthy candidate for resolving the Erdős–Fishburn problem for the taxicab metric.

Proof of Theorem 2.2. Fix $k \in \mathbb{N}$. For (i), fix $d \in \mathbb{N}$; note that by the definition of $\Lambda_d(k)$, we have $\|n\|_1 \le k$ for all $n \in \Lambda_d(k)$. In particular, for any $n, m \in \Lambda_d(k)$, by the triangle inequality we have

$$\|n - m\|_1 \le \|n\|_1 + \|m\|_1 \le k + k = 2k.$$

Further, $\|n - m\|_1 = |n_1 - m_1| + \cdots + |n_d - m_d|$ is certainly an integer, and by the definition of $\Lambda_d(k)$, and the fact that an integer is congruent to its absolute value modulo 2, we have

$$\begin{aligned} |n_1 - m_1| + \cdots + |n_d - m_d| &\equiv n_1 - m_1 + \cdots + n_d - m_d \\ &\equiv (n_1 + \cdots + n_d) - (m_1 + \cdots + m_d) \\ &\equiv k - k \equiv 0 \pmod 2. \end{aligned}$$

Therefore, the only possible nonzero values of $\|n - m\|_1$ are $2, 4, \dots, 2k$, and for each $1 \le j \le k$ the distance $2j$ is attained between the points $(j, 0, \dots, 0)$ and $(-j, 0, \dots, 0)$ if $j \equiv k \pmod 2$, or between $(j, 1, \dots, 0)$ and $(-j, 1, \dots, 0)$ if $j \not\equiv k \pmod 2$.

For (ii), we first see that

$$\Lambda_1(k) = \begin{cases} \{-k, -k+2, \dots, -1, 1, \dots, k-2, k\} & \text{if } k \text{ is odd,} \\ \{-k, -k+2, \dots, -2, 0, 2, \dots, k-2, k\} & \text{if } k \text{ is even.} \end{cases}$$

In particular, $|\Lambda_1(k)| = 2\lceil k/2 \rceil = k+1$ if k is odd and $|\Lambda_1(k)| = 2(k/2)+1 = k+1$ if k is even. Secondly, we see that $\Lambda_d(1)$ is precisely $\{\pm e_i : 1 \le i \le d\}$, where $\{e_i\}$ is the standard basis for $\mathbb{R}^d$.

For (iii), we see that the possible values of the final coordinate for elements of $\Lambda_{d+1}(k)$ are integers satisfying $-k \le x_{d+1} \le k$. Further, for a fixed value $x_{d+1} = c$, the intersection of this hyperplane with $\Lambda_{d+1}(k)$ is

$$\begin{aligned} \{(n_1, \dots, n_d, c) \in \mathbb{Z}^{d+1} : |n_1| + \cdots + |n_d| \le k - |c|, \\ n_1 + \cdots + n_d \equiv k - c \equiv k - |c| \pmod 2\}, \end{aligned}$$

which is in natural bijection with $\Lambda_d(k - |c|)$. Therefore,

$$|\Lambda_{d+1}(k)| = \sum_{c=-k}^{k} |\Lambda_d(k - |c|)| = |\Lambda_d(k)| + 2\sum_{j=0}^{k-1} |\Lambda_d(j)|. \qquad \square$$

We continue by establishing a detailed formula for $|\Lambda_d(k)|$, which in particular guarantees that it has the correct order of magnitude $\Omega(k^d)$.

Proof of Theorem 2.3. We first note that by Theorem 2.2(ii), we have $|\Lambda_1(k)| = k+1$ for all $k \geq 0$. We now fix $d \geq 2$, let $h = \lceil d/2 \rceil - 1$, and make the inductive hypothesis

$$|\Lambda_{d-1}(k)| = a_{d-1,0}(k+1)^{d-1} + a_{d-1,1}(k+1)^{d-3} + \cdots + a_{d-1,h}(k+1)^{d-1-2h} \quad (1)$$

for all $k \geq 0$. Faulhaber's formula gives

$$F_p(n) = \sum_{j=1}^{n} j^p = \frac{n^{p+1}}{p+1} + \frac{n^p}{2} + \frac{1}{p+1}\sum_{i=0}^{p-1}\binom{p+1}{i} B_{p+1-i}n^i \quad (2)$$

for all $n, p \in \mathbb{N}$, where B_i is the i-th Bernoulli number. By Theorem 2.2(iii), we have

$$|\Lambda_d(k)| = 2\sum_{j=0}^{k-1} |\Lambda_{d-1}(j)| + |\Lambda_{d-1}(k)| = 2\sum_{j=0}^{k} |\Lambda_{d-1}(j)| - |\Lambda_{d-1}(k)|, \quad (3)$$

which combines with (1) to yield

$$\begin{aligned}
|\Lambda_d(k)| &= 2\left(a_{d-1,0}\sum_{j=0}^{k}(j+1)^{d-1} + \cdots + a_{d-1,h}\sum_{j=0}^{k}(j+1)^{d-1-2h}\right) - |\Lambda_{d-1}(k)| \\
&= 2\left(a_{d-1,0}\sum_{j=1}^{k+1} j^{d-1} + \cdots + a_{d-1,h}\sum_{j=1}^{k+1} j^{d-1-2h}\right) - |\Lambda_{d-1}(k)| \\
&= 2(a_{d-1,0}F_{d-1}(k+1) + \cdots + a_{d-1,h}F_{d-1-2h}(k+1)) - |\Lambda_{d-1}(k)|.
\end{aligned}$$

This tells us that we can indeed write $|\Lambda_d(k)|$ as a polynomial in $k+1$, but we wish to establish the claimed explicit and recursive formulas for the coefficients, as well as the fact that every other coefficient is zero. First we consider the $(k+1)^d$ coefficient, which only arises from the term $2a_{d-1,0}F_{d-1}(k+1)$. Since the n^{p+1} coefficient of $F_p(n)$ is $1/(p+1)$, we have $a_{d,0} = 2a_{d-1,0}/d$. Using the base case $a_{1,0} = 1$, we have by induction that $a_{d,0} = 2^{d-1}/d!$, as claimed.

Next we consider the $(k+1)^{d-1}$ coefficient, which arises from two sources: the $(k+1)^{d-1}$ coefficients of $2a_{d-1,0}F_{d-1}(k+1)$ and $-|\Lambda_{d-1}(k)|$, respectively. The former is $2a_{d-1,0}\left(\frac{1}{2}\right) = a_{d-1,0}$, while the latter is $-a_{d-1,0}$, which means that the $(k+1)^{d-1}$ coefficient of $|\Lambda_d(k)|$ is indeed 0. More generally, for other coefficients corresponding to terms of the form $(k+1)^{d-1-2i}$, we use the following three facts: the $(k+1)^{d-1-2i}$ coefficient of $2a_{d-1,i}F_{d-1-2i}(k+1)$ is $a_{d-1,i}$ by the same logic as above, the $(k+1)^{d-1-2i}$ coefficient of $-|\Lambda_{d-1}(k)|$ is $-a_{d-1,i}$, and the $(k+1)^{d-1-2i}$ coefficient of $F_{d-1-2\ell}(k+1)$ is 0 for all $\ell < i$, because $B_n = 0$ for all odd $n \geq 3$. Therefore, all $(k+1)^{d-1-2i}$ coefficients of $|\Lambda_d(k)|$ are 0.

For the $(k+1)^{d-2}$ coefficient, we begin by noting that a direct calculation using (2) and (3) yields $|\Lambda_3(k)| = \frac{2}{3}(k+1)^3 + \frac{1}{3}(k+1)$; hence $a_{3,1} = \frac{1}{3}$, which serves as the base case for another induction. Fixing $d \geq 4$ and assuming the claimed formula

$a_{d-1,1} = 2^{d-4}/(3(d-4)!)$ holds, the $(k+1)^{d-2}$ coefficient of $|\Lambda_d(k)|$ is formed by two contributions, from $2a_{d-1,0}F_{d-1}(k+1)$ and $2a_{d-1,1}F_{d-3}(k+1)$, respectively. The former is given by

$$2a_{d-1,0}\left(\frac{1}{d}\right)\binom{d}{d-2}B_2 = 2\cdot\frac{2^{d-2}}{(d-1)!}\cdot\frac{1}{d}\cdot\frac{d(d-1)}{2}\cdot\frac{1}{6} = \frac{2^{d-3}}{3(d-2)!},$$

while the latter is given by

$$2a_{d-1,1}\cdot\frac{1}{d-2} = 2\cdot\frac{2^{d-4}}{3(d-4)!}\cdot\frac{1}{d-2} = \frac{2^{d-3}}{3(d-4)!(d-2)}.$$

Therefore, we have

$$\begin{aligned} a_{d,1} &= \frac{2^{d-3}}{3(d-2)!} + \frac{2^{d-3}}{3(d-4)!(d-2)} \\ &= \frac{2^{d-3}+(d-3)2^{d-3}}{3(d-2)!} = \frac{2^{d-3}(d-2)}{3(d-2)!} = \frac{2^{d-3}}{3(d-3)!}, \end{aligned}$$

as claimed.

More generally, by (2) and (3), we see that the $(k+1)^{d-2i}$ coefficient of $|\Lambda_d(k)|$ receives a contribution from $2a_{d-1,\ell}F_{d-1-2\ell}(k+1)$ for each $0 \le \ell \le i$. Specifically, that contribution is

$$2a_{d-1,\ell}\cdot\frac{1}{d-2\ell}\cdot\binom{d-2\ell}{d-2i}\cdot B_{2i-2\ell} = \frac{2a_{d-1,\ell}}{d-2\ell}\binom{d-2\ell}{2(i-\ell)}B_{2(i-\ell)},$$

and the recursive formula for $a_{d,i}$ follows. □

4. Optimality in two dimensions: proof of Theorem 2.5

In this section, we prove the unique optimality of $\Lambda_2(k)$, in that it is the unique subset of $\mathbb{R}^2$, up to ℓ^1-similarity, of maximal size amongst sets determining at most k distinct ℓ^1-distances. As referenced in Section 2, the proof is in part enabled by an equivalence between the ℓ^1-norm and the ℓ^∞-norm on $\mathbb{R}^2$. We frame our discussion entirely in the context of the ℓ^1-norm, but the connection is implicit in our proof, particularly the following lemma.

Lemma 4.1. *Let $v_1 = (1, 1)$ and $v_2 = (-1, 1)$. If $x \in \mathbb{R}^2$ with $x = c_1v_1 + c_2v_2$, then*

$$\|x\|_1 = 2\max\{|c_1|, |c_2|\}.$$

Proof. Let $v_1 = (1, 1)$, $v_2 = (-1, 1)$, fix $x \in \mathbb{R}^2$, and write x uniquely as $x = c_1v_1 + c_2v_2 = (c_1 - c_2, c_1 + c_2)$. By potentially reflecting over the diagonal $x_1 = x_2$ and/or replacing x by $-x$, both of which preserve the ℓ_1-norm, we can assume without loss of generality that $|c_1| \ge |c_2|$ and $c_1 \ge 0$. In this case,

$$\|x\|_1 = |c_1 - c_2| + |c_1 + c_2| = c_1 - c_2 + c_1 + c_2 = 2c_1 = 2\max\{|c_1|, |c_2|\}. \quad \square$$

Remark. As an alternative approach to Section 4, one could treat the connection between the ℓ^1 and ℓ^∞ norms on $\mathbb{R}^2$ in a more explicit way. Namely, Lemma 4.1 can be reframed as the statement that the map $f : (\mathbb{R}^2, \|\cdot\|_1) \to (\mathbb{R}^2, \|\cdot\|_\infty)$ defined by $f(x, y) = (x+y, x-y)$ is a linear isomorphism satisfying $\|(x, y)\|_1 = \|f(x, y)\|_\infty$. Therefore, any results related to the ℓ^∞-norm can be immediately transferred to the ℓ^1-norm via this isomorphism. In particular, the proofs that follow could be rewritten in a slightly cleaner way in the ℓ^∞ context. However, in order to maintain our hands-on approach with the taxicab metric, we have chosen to leave the proofs in their ℓ^1 form.

Our main strategy for proving Theorem 2.5 is inspired by [Erdős and Fishburn 1996]. Specifically, we suppose that $P \subseteq \mathbb{R}^2$ determines at most k distinct ℓ^1-distances, and we seek an upper bound on the number of points we must remove from P in order to eliminate the largest ℓ^1-distance, hence reducing to the case of $k-1$ distinct ℓ^1-distances and allowing us to invoke an inductive hypothesis. The following sequence of lemmas formalizes this strategy. Here we define an *ℓ^1-ball* in the expected way, as the region bounded by an ℓ^1-sphere, which for $d = 2$ is an ℓ^1-circle.

Lemma 4.2. *Suppose $P \subseteq \mathbb{R}^2$ is finite. If D is the largest ℓ^1-distance determined by P, then P is contained in a closed ℓ^1-ball of diameter D.*

Proof. Suppose $P \subseteq \mathbb{R}^2$ is finite. Let $v_1 = (1, 1)$ and $v_2 = (-1, 1)$. Since $\{v_1, v_2\}$ forms a basis for $\mathbb{R}^2$, every $x \in P$ can be written uniquely as $x = c_1v_1 + c_2v_2$. Choose $x_1, x_2, x_3, x_4 \in P$ such that x_1 maximizes c_1, x_2 minimizes c_1, x_3 maximizes c_2, and x_4 minimizes c_2. Call these values $c_{1,\max}$, $c_{1,\min}$, $c_{2,\max}$, and $c_{2,\min}$, respectively. These choices contain P inside of a rectangle R, rotated 45° from axis parallel, determined by the inequalities $c_{1,\min} \le c_1 \le c_{1,\max}$ and $c_{2,\min} \le c_2 \le c_{2,\max}$.

Let $w_1 = c_{1,\max} - c_{1,\min}$ and $w_2 = c_{2,\max} - c_{2,\min}$, and assume without loss of generality that $w_1 \ge w_2$. By Lemma 4.1, we have $\|x_1 - x_2\|_1 = 2w_1$ and $\|p_1 - p_2\|_1 \le 2w_1$ for all $p_1, p_2 \in R$, so $D = 2w_1$ is the largest ℓ_1-distance determined by P. Let $c_{2,\text{new}} = c_{2,\max} - w_1 \le c_{2,\min}$, and let $B \supseteq R \supseteq P$ be defined by the inequalities $c_{1,\min} \le c_1 \le c_{1,\max}$ and $c_{2,\text{new}} \le c_2 \le c_{2,\max}$. B is a square rotated 45° from axis parallel, or in other words a closed ℓ^1-ball, of diameter D, as required. □

Lemma 4.3. *If $P \subseteq \mathbb{R}^2$ is contained in a closed ℓ^1-ball B of diameter D, then the ℓ^1-distance D can be eliminated from P by removing the points of P contained in any two adjacent sides of the boundary of B.*

Proof. Suppose $P \subseteq \mathbb{R}^2$ is contained in a closed ℓ^1-ball B of diameter D.

Let a_1, a_2 be the left and right vertices of B, respectively, so in particular $\|a_1 - a_2\|_1 = D$. Let U denote the closed (including a_1, a_2) upper ℓ^1-semicircle connecting a_1 and a_2, and let L denote the open (not including a_1, a_2) lower ℓ^1-semicircle connecting a_1 and a_2. Since the ℓ^1-norm is invariant under 90° rotation, it suffices to establish the conclusion of the lemma for removing the points of P lying in U. Suppose $x_1, x_2 \in P \setminus U$.

Case 1: At least one of x_1, x_2 lies in $B \setminus (U \cup L)$, which is an open ℓ^1-ball of radius $D/2$.

Assume without loss of generality that $x_1 \in B \setminus (U \cup L)$, and let c be the center of B. Therefore, $\|x_1 - c\|_1 < D/2$ and $\|x_2 - c\|_1 \leq D/2$. By the triangle inequality, $\|x_1 - x_2\|_1 \leq \|x_1 - c\|_1 + \|c - x_2\|_1 < D/2 + D/2 = D$.

Case 2: $x_1, x_2 \in L$. After possibly reflecting, assume without loss of generality that x_1 is to the left of x_2 and $\|x_1 - a_1\|_1 \leq \|x_2 - a_2\|_1$, so x_1 is positioned at least as high as x_2, By replacing x_1 with a_1, we move up and to the left, so both the horizontal and vertical components of the ℓ^1-distance to x_2 get larger; hence

$$\|x_1 - x_2\|_1 < \|a_1 - x_2\|_1 = D.$$

In both cases, all distances amongst points in $P \setminus U$ are strictly less than D, and the lemma follows. □

Lemma 4.4. *If $P \subseteq \mathbb{R}^d$ is contained in a line and determines at most k distinct ℓ^1-distances, then $|P| \leq k+1$. Further, if $|P| = k+1$, then P is an arithmetic progression, meaning the ℓ^1-distances are $\lambda, 2\lambda, \dots, k\lambda$ for some $\lambda > 0$.*

Proof. Since ℓ^1-distance along a straight line in $\mathbb{R}^d$ is just a constant multiple, depending on the direction of the line, of the standard Euclidean distance, it suffices to establish the lemma with $d = 1$, for which we induct on k.

The base case $k = 1$ is trivial, as three points $x_1 < x_2 < x_3$ in $\mathbb{R}$ automatically determine two distances $x_2 - x_1 < x_3 - x_1$, and any two points form an arithmetic progression.

Now, fix $k \geq 2$, and assume that if $Q \subseteq \mathbb{R}$ determines at most $k-1$ distances, then $|Q| \leq k$, and further, if $|Q| = k$, then Q is an arithmetic progression. Now suppose $P \subseteq \mathbb{R}$ determines at most k distances.

Let $P = \{x_1 < x_2 < \cdots < x_n\}$. The $n-1$ distances $x_2 - x_1 < x_3 - x_1 < \cdots < x_n - x_1$ are all distinct; hence $n-1 \leq k$, or in other words $n \leq k+1$. Further, suppose $n = k+1$. By removing x_{k+1}, we also remove the longest distance $x_{k+1} - x_1$, so the set $Q = \{x_1, \dots, x_k\}$ determines $k-1$ distances. By our inductive hypothesis, Q must be an arithmetic progression; in other words $Q = \{x_1, x_1 + \lambda, x_1 + 2\lambda, \dots, x_1 + (k-1)\lambda\}$.

If $x_{k+1} < x_1 + k\lambda$, then both $x_{k+1} - x_1 > (k-1)\lambda$ and $x_{k+1} - x_k < \lambda$ are new distances not determined by Q. If $x_{k+1} > x_1 + k\lambda$, then both $x_{k+1} - x_1 > k\lambda$ and $x_{k+1} - x_2 > (k-1)\lambda$ are new distances not determined by Q. In either case, P determines at least $k+1$ distinct distances, contradicting the assumption that it determines at most k distances. Therefore, x_{k+1} must be $x_1 + k\lambda$, and the lemma follows. □

Lemma 4.5. *If $S \subseteq \mathbb{R}^2$ is contained in the union of two adjacent sides of an ℓ^1-circle and determines at most k distinct ℓ^1-distances, then* $|S| \le 2k+1$. *Further, if* $|S| = 2k+1$, *then the points of S on each side form an arithmetic progression containing the shared vertex.*

Proof. Suppose $S \subseteq \mathbb{R}^2$ is contained in the union of two adjacent sides of an ℓ^1-circle and determines at most k distinct ℓ^1-distances. Assume without loss of generality that the two adjacent sides are the closed upper semicircle. We know from Lemma 4.4 that there are at most $k+1$ points on each of the two sides.

Further, if $|S| \ge 2k+1$, then there are exactly $k+1$ points on one side, assume the left, and at least k points on the right side. Let $x_1, \dots, x_{k+1}$ denote the points of P on the left side, ordered left to right, and let y be any point of P on the right side. We note that

$$\|x_1 - x_2\|_1 < \|x_1 - x_3\|_1 < \cdots < \|x_1 - x_{k+1}\|_1 \le \|x_1 - y\|_1,$$

and $\|x_1 - x_{k+1}\|_1 = \|x_1 - y\|_1$ is only possible if x_{k+1} is the vertex shared by the two sides. In particular, if the shared vertex is not included amongst the $k+1$ points on the left side, then at least $k+1$ distinct ℓ^1-distances occur from the leftmost point, contradicting our assumption.

Therefore, if $|S| \ge 2k+1$, it must be the case that there are exactly $k+1$ points on both the left and right sides, including the shared vertex, meaning in fact $|S| = 2k+1$. Finally, by Lemma 4.4, we know that the $k+1$ points on each side must form an arithmetic progression. □

We are now fully armed to show the unique optimality of $\Lambda_2(k)$.

Proof of Theorem 2.5. We induct on k. For our base case, consider $k = 0$. In order for a set to determine zero ℓ^1-distances (as always, not including 0), it can contain at most $1 = (0+1)^2$ point, and if it contains a point, then it is trivially a translation of $\Lambda_2(0) = \{(0, 0)\}$.

Now, fix $k \in \mathbb{N}$, assume the conclusion of the theorem holds for $k-1$, and suppose $P \subseteq \mathbb{R}^2$ determines at most k distinct ℓ^1-distances. By Lemma 4.2, P is contained in a closed ℓ^1-ball B of diameter D, where D is the largest ℓ^1-distance determined by P. By Lemma 4.3, we can remove the distance D by removing the points of P that lie on the closed upper ℓ^1-semicircle U on the boundary of B. Since D has been removed as an ℓ^1-distance, we know that $T = P \setminus U$ determines at most $k-1$ distinct ℓ^1-distances. By our inductive hypothesis, $|T| \le k^2$, and if $|T| = k^2$, then T is ℓ^1-similar to $\Lambda_2(k-1)$.

Further, by Lemma 4.5, we know that $S = P \cap U$ satisfies $|S| \le 2k+1$, and if $|S| = 2k+1$, then S consists of two $(k+1)$-term arithmetic progressions, one on each side of U, which meet at the shared vertex. Therefore, $|P| \le |T| + |S| \le k^2 + 2k + 1 = (k+1)^2$, and $|P| = (k+1)^2$ if and only if T is ℓ^1-similar to $\Lambda_2(k-1)$ and S is a union of two arithmetic progressions meeting at the shared vertex. Finally, the only way these two sets can be combined without creating additional ℓ^1-distances is for $S \cup T$ to be ℓ^1-similar to $\Lambda_2(k)$. □

5. Single ℓ^1-distance in three dimensions: proof of Theorem 2.6

Without analogs of Lemmas 4.1 and 4.2 in dimension $d \geq 3$, our strategy for proving Theorem 2.5 does not naturally generalize to higher dimensions. However, in the case of $k = 1$, we make the observation that if $P \subseteq \mathbb{R}^d$ determines a single ℓ^1-distance, then all but the "southernmost" point (the point minimizing the last coordinate) of P lie on a single closed upper ℓ^1-hemisphere. The following sequence of lemmas provide a detailed investigation into how ℓ^1-distance behaves between points on a single upper ℓ^1-hemisphere in $\mathbb{R}^3$, which consists of four flat faces, one for each quadrant determined by the first two coordinates, intersecting at a single northernmost point. The three lemmas correspond to the cases where the points lie on the same face, opposite faces, or neighboring faces, respectively.

Lemma 5.1. *Suppose $V, W \in \mathbb{R}^3$, with $V = (x_1, y_1, z_1)$ and $W = (x_2, y_2, z_2)$. If $\|V\|_1 = \|W\|_1$ and $x_1x_2, y_1y_2, z_1z_2 \geq 0$, then*

$$\|V - W\|_1 = 2\max\{|x_1 - x_2|, |y_1 - y_2|, |z_1 - z_2|\}.$$

Proof. Suppose $V, W \in \mathbb{R}^3$, with $V = (x_1, y_1, z_1)$, $W = (x_2, y_2, z_2)$, $\|V\|_1 = \|W\|_1 = \lambda$, and $x_1x_2,\ y_1y_2,\ z_1z_2 \geq 0$. After reflections about coordinate planes, coordinate permutations, and relabeling V and W (which all preserve both sides of the equation in the conclusion of the lemma), we can assume without loss of generality that all coordinates are nonnegative and $x_1 - x_2 \geq |y_1 - y_2| \geq |z_1 - z_2|$. Since

$$\|V\|_1 = x_1 + y_1 + z_1 = \|W\|_1 = x_2 + y_2 + z_2 = \lambda,$$

we have in particular that $(x_1 - x_2) + (y_1 - y_2) + (z_1 - z_2) = 0$. Since the largest coordinate distance is in the x-direction, and $x_1 \geq x_2$, we must have $y_1 \leq y_2$ and $z_1 \leq z_2$. Therefore

$$\begin{aligned}\|V - W\|_1 &= (x_1 - x_2) + (y_2 - y_1) + (z_2 - z_1)\\ &= x_1 - x_2 + y_2 - y_1 + (\lambda - x_2 - y_2) - (\lambda - x_1 - y_1)\\ &= 2(x_1 - x_2),\end{aligned}$$

and the lemma follows. □

Lemma 5.2. *Suppose $V, W \in \mathbb{R}^3$, with $V = (x_1, y_1, z_1)$ and $W = (x_2, y_2, z_2)$. If $\|V\|_1 = \|W\|_1 = \lambda,\ x_1x_2 \leq 0,\ y_1y_2 \leq 0$, and $z_1, z_2 \geq 0$, then*

$$\|V - W\|_1 = 2(\lambda - \min\{z_1, z_2\}).$$

Proof. Suppose $V, W \in \mathbb{R}^3$, with $V = (x_1, y_1, z_1)$, $W = (x_2, y_2, z_2)$, $\|V\|_1 = \|W\|_1 = \lambda$, $x_1x_2 \le 0$, $y_1y_2 \le 0$, and $z_1, z_2 \ge 0$. After reflections about coordinate planes and relabeling V and W, we can assume without loss of generality that $x_1, y_1 \ge 0$, $x_2, y_2 \le 0$, and $z_1 \le z_2$.

Therefore, $x_1 + y_1 = \lambda - z_1$, while $-x_2 - y_2 = \lambda - z_2$; hence

$$\begin{aligned}\|V - W\|_1 &= (x_1 - x_2) + (y_1 - y_2) + (z_2 - z_1)\\ &= \lambda - z_1 + \lambda - z_2 + z_2 - z_1\\ &= 2(\lambda - z_1),\end{aligned}$$

and the lemma follows. □

Lemma 5.3. *Suppose* $V, W \in \mathbb{R}^3$, *with* $V = (x_1, y_1, z_1)$, $W = (-x_2, y_2, z_2)$, $\|V\|_1 = \|W\|_1 = \lambda$, *and* $x_1x_2, y_1y_2, z_1z_2 \ge 0$. *If* $\|V - W\|_1 = \lambda$, *then* $|x_1| \le \lambda/2$.

Proof. Suppose $V, W \in \mathbb{R}^3$, with $V = (x_1, y_1, z_1)$, $W = (-x_2, y_2, z_2)$, $\|V\|_1 = \|W\|_1 = \lambda$, $x_1x_2, y_1y_2, z_1z_2 \ge 0$. After reflecting about coordinate planes and scaling, we can assume $x_1, x_2, y_1, y_2, z_1, z_2 \ge 0$, and $\lambda = 2$. If $\|V - W\|_1 = 2$, then the largest possible value of $y_2 + z_2$ is $y_1 + z_1 + 2 - (x_1 + x_2)$. However, since $\|W\|_1 = 2$, we must have $y_2 + z_2 = 2 - x_2$; hence $2 - x_2 \le y_1 + z_1 + 2 - (x_1 + x_2)$, which rearranges to $x_1 \le y_1 + z_2 = 2 - x_1$. Thus $x_1 \le 1$, as required. □

We now establish the unique optimality of $\Lambda_3(1)$ by conducting a case analysis based on the concentration of the points of P, apart from the southernmost point, on the four faces of a single closed upper ℓ^1-hemisphere.

Proof of Theorem 2.6. Suppose $P \subseteq \mathbb{R}^3$ determines a single ℓ^1-distance λ, and choose a point $c \in P$ that minimizes the z-coordinate. By translating and dilating, we can assume without loss of generality that $c = (0, 0, 0)$ and $\lambda = 2$, and hence the remaining elements of P are all contained in the closed upper ℓ^1-hemisphere H of radius 2 centered at $(0, 0, 0)$. We note that the southernmost point of $\Lambda_3(1)$ is $(0, 0, -1)$, so our end goal in this proof is to show that $|P| < 6$ unless P is $\Lambda_3(1)$ shifted up by 1.

We consider the different ways that P can be concentrated on the faces of H. To this end, we define

$$\begin{aligned}H_{++} &= \{(x, y, z) \in H : x, y \ge 0\},\\ H_{+-} &= \{(x, y, z) \in H : x \ge 0, y \le 0\},\end{aligned}$$

with analogous definitions for H_{-+} and H_{--}. We refer to the pair H_{++}, H_{--} as *opposite* faces, and likewise for H_{+-}, H_{-+}. The three lemmas proven at the beginning of this section allow us to make the following assertions:

(i) For any pair of distinct points $U = (x_1, y_1, z_1)$, $V = (x_2, y_2, z_2) \in P \cap H$, with U and V lying on opposite faces, we have by Lemma 5.2 that $\min\{z_1, z_2\} = 1$.

(ii) For any pair of distinct points $U = (x_1, y_1, z_1)$, $V = (x_2, y_2, z_2) \in P \cap H$, with U and V lying on the same face, we have by Lemma 5.1 that $\max\{|x_1 - x_2|, |y_1 - y_2|, |z_1 - z_2|\} = 1$.

(iii) For distinct points $U = (x_1, y_1, z_1)$, $V = (x_2, y_2, z_2) \in P \cap H$, with $U \in H_{++}$ and $V \in H_{-+}$, we have by Lemma 5.3 that $x_1 \leq 1$. Similarly, by permuting coordinates, if $U \in H_{++}$ and $V \in H_{+-}$, then $y_1 \leq 1$.

If $|P| \geq 6$, then at least five points of P lie on H, and in particular the sizes of the four intersections of P with the respective faces of H must add to at least 5. Therefore, the only possible arrangements of $P \cap H$ include either three points on a single face, or two points on one face and a point on the opposite face. Further, the proof is greatly simplified in the case that the "north pole" $(0, 0, 2)$ lies in P, so we divide the argument into the following three cases:

Case 1: $(0, 0, 2) \in P$.

Case 2: $(0, 0, 2) \notin P$, and P contains three points $U, V, W \in H$ such that U and V lie on the same face, and W lies on the opposite face.

Case 3: $(0, 0, 2) \notin P$, and there exists a face of H containing at least three points of P.

Proof for Case 1. Let $V = (0, 0, 2)$. For $Q = (x, y, z) \in (P \cap H) \setminus \{V\}$, we have by (i) that $z = 1$. In particular, the elements of P other than $(0, 0, 0)$ and $(0, 0, 2)$ take the form $(x, y, 1)$ with $|x| + |y| = 1$, and all pairs are separated by ℓ^1-distance 2. By Theorem 2.5, there can be at most four such elements, and the only choice of four that works is $(1, 0, 1)$, $(-1, 0, 1)$, $(0, 1, 1)$, and $(0, -1, 1)$. The resulting arrangement is $\Lambda_3(1)$ translated up by 1, which establishes Theorem 2.6 in this case.

Proof for Case 2. After reflecting about coordinate planes, we can assume that $U, V \in H_{++}$ and $W \in H_{--}$. Letting $U = (x_0, y_0, z_0)$, $V = (x_1, y_1, z_1)$, and $W = (x_2, y_2, z_2)$, we have by (i) and (ii) that

$$\max\{|x_0 - x_1|, |y_0 - y_1|, |z_0 - z_1|\} = \min\{z_0, z_2\} = \min\{z_1, z_2\} = 1.$$

In particular, all three z-coordinates are at least 1, and since $(0, 0, 2) \notin P$, we have $|z_0 - z_1| < 1$. Therefore, we simultaneously know that $0 \leq x_0, x_1, y_0, y_1 \leq 1$ and $\max\{|x_0 - y_0|, |x_1 - y_1|\} = 1$.

This implies that (after potentially relabeling) either $U = (1, 0, 1)$ and $V = (0, y, 2 - y)$ for some $0 < y \leq 1$ or $U = (x, 0, 2 - x)$ for some $0 < x \leq 1$ and $V = (0, 1, 1)$. In either case, $U \in H_{++} \cap H_{+-}$, and $V \in H_{++} \cap H_{-+}$, so P contains at least one element on every face of H. Therefore, by (i), all points of P lying on H have z-coordinate at least 1. Further, by the same reasoning as above, there are at most two points of P on each face, and the only way two points of P can lie on the same face is if they lie on opposite sides of the boundary, as with U and V. In particular, at most four points of P lie on H, and hence P contains at most five points in total.

Proof for Case 3. This case gets a bit stickier, because, as some trial and error reveals, there are a variety of possible arrangements of three points on a single face of H that are all separated by ℓ^1-distance 2.

Focusing on H_{++} for the sake of exposition, we see that our desired configuration of $\{(1, 0, 1), (0, 1, 1), (0, 0, 2)\}$ is merely one member of a family of arrangements obtained from the following process:

- Choose $x_0, y_0, z_0 \geq 0$ with $x_0 + y_0 + z_0 \leq 1$, and let $\alpha = 1 - (x_0 + y_0 + z_0)$.

- Starting from (x_0, y_0, z_0), construct a point by adding 1 to one coordinate and α to another coordinate (so the coordinates add to 2), then produce two additional points in a similar way by rotating the original choice of coordinates. For example, the initial choice of $U = (x_0+1, y_0+\alpha, z_0)$ uniquely determines the two additional points $V = (x_0, y_0+1, z_0+\alpha)$ and $W = (x_0+\alpha, y_0, z_0+1)$. All of these points lie on H_{++}, and by (ii) they are all separated by ℓ^1-distance 2. In fact, the only other possible set of three points yielded by this process (up to labeling) is $U = (x_0 + 1, y_0, z_0 + \alpha)$, $V = (x_0 + \alpha, y_0 + 1, z_0)$, and $W = (x_0, y_0 + \alpha, z_0 + 1)$. For additional clarity, a specific example is $x_0 = 0.1$, $y_0 = 0.3$, $z_0 = 0.4$; hence $\alpha = 0.2$, which could yield the three-point arrangements $\{(1.1, 0.5, 0.4), (0.1, 1.3, 0.6), (0.3, 0.3, 1.4)\}$ or $\{(1.1, 0.3, 0.6), (0.3, 1.3, 0.4), (0.1, 0.5, 1.4)\}$.

We hope to demystify the situation by arguing that the arrangements discussed above are in fact the only possible arrangements. To this end, after reflections, we can assume P contains three points $U, V, W \in H_{++}$, and we settle Case 3 with the following steps:

- Step 1: Show that U, V, W take the form discussed above. In particular, after specifying the minimum values of each coordinate and a single point, the second and third points are uniquely determined; hence there cannot be a fourth point in $P \cap H_{++}$.
- Step 2: Show that P can contain at most one point in $(H_{+-} \cup H_{-+}) \setminus H_{++}$ before necessarily reducing to Case 2. This means that any hypothetical fifth point of $P \cap H$ necessarily lies on H_{--}, which itself reduces the argument back to Case 2.

Step 1: Let x_0, y_0, and z_0 be the minimum x, y, and z-coordinates, respectively, attained by U, V, and W. In what follows, we repeatedly appeal to (ii), which tells us that for every pair of points in $\{U, V, W\}$, the maximum coordinate distance is exactly 1. In particular, the maximum x, y, and z-coordinates attained by U, V, and W are at most $x_0 + 1$, $y_0 + 1$, and $z_0 + 1$, respectively, and we begin by arguing that this inequality must be equality in all three coordinates.

Suppose that this inequality is strict in at least one coordinate. By permuting coordinates and relabeling points we may assume that $U = (x_0, y, z)$, and neither of V and W has x-coordinate $x_0 + 1$. Therefore, the maximum coordinate distance of 1 required by (ii) must occur in either the y or z-coordinates, and since x_0 is the minimum x-coordinate, V and W must both take one of the following forms: $(x_0 + \alpha, y - 1, z + (1 - \alpha))$ for some $0 \leq \alpha < 1$, or $(x_0 + \beta, y + (1 - \beta), z - 1)$ for some $0 \leq \beta < 1$. However, no combination of these choices for V and W have a maximum coordinate distance of 1 from each other, so this arrangement is impossible. Therefore, all the maxima $x_0 + 1$, $y_0 + 1$, and $z_0 + 1$ are indeed achieved. For the remainder of the proof, we will refer to the respective coordinate values x_0, y_0, and z_0 as *minimum coordinates*, and we will similarly refer to the respective coordinate values $x_0 + 1$, $y_0 + 1$, $z_0 + 1$ as *maximum coordinates*. We complete Step 1 by considering the following three subcases.

Subcase A: two maximum coordinates appear simultaneously in a single point. Since all the points have ℓ^1-norm 2, this subcase necessitates that $x_0 = y_0 = z_0 = 0$, and we assume without loss of generality that $U = (1, 1, 0)$. Since the minimum x- and y-coordinates of 0 must be attained, $\{V, W\}$ contains points of the form $(x, 0, 2 - x)$ and $(0, y, 2 - y)$, respectively, for some $0 \leq x, y \leq 1$. However, since the maximum z-coordinate is 1, the only admissible choices are $x = y = 1$; hence the three points are $(1, 1, 0)$, $(0, 1, 1)$, and $(1, 0, 1)$, which take the required form with $\alpha = 1$.

Subcase B: two minimum coordinates appear simultaneously in a single point. Assume without loss of generality that $U = (x_0, y_0, z)$. Since x_0 and y_0 are minimum coordinates, each of V and W must take the form $(x_0 + \alpha, y_0 + \beta, z - (\alpha + \beta))$ for some $\alpha, \beta \geq 0$, and by (ii) we must have $\alpha + \beta = 1$. Further, since the z-coordinates of V and W are both $z - 1$ (which is hence the minimum coordinate z_0), the maximum coordinate distance of 1 must occur in the first two coordinates, meaning that $\{V, W\} = \{(x_0 + 1, y_0, z_0), (x_0, y_0 + 1, z_0)\}$. In particular, the arrangement takes the required form with $\alpha = 0$.

Subcase C: exactly one minimum coordinate and one maximum coordinate occurs in each point. After permuting coordinates and relabeling points we assume $U = (x_0 + 1, y_0 + \alpha, z_0)$, where $0 < \alpha = 1 - (x_0 + y_0 + z_0) < 1$. In order to meet the subcase conditions, have a maximum coordinate distance of 1 from U, and have ℓ^1-norm 2, the options for V and W are $(x_0, y_0 + 1, z_0 + \alpha)$, $(x_0, y_0 + \alpha, z_0 + 1)$, and $(x_0 + \alpha, y_0, z_0 + 1)$. Of these three possibilities, there is only one pair that are separated by ℓ^1-distance 2 from each other; hence

$$\{V, W\} = \{(x_0, y_0 + 1, z_0 + \alpha), (x_0 + \alpha, y_0, z_0 + 1)\},$$

as required.

Step 2: Suppose P contains a point $Q \in H_{-+} \setminus H_{++}$ (the argument is completely analogous for $Q \in H_{+-} \setminus H_{++}$). By (iii), in order for Q to be separated from $U = (x_0 + 1, y_0 + \alpha, z_0)$ by ℓ^1-distance 2, we must have $x_0 + 1 \leq 1$, and hence $x_0 = 0$. In particular, $V = (0, y_0 + 1, z_0 + \alpha) \in P \cap (H_{-+} \cap H_{++})$, so P contains at least two points on H_{-+}. This means that, in order to avoid reducing to Case 2, P cannot contain any elements of H_{+-}.

If instead P contains a second point $R \in H_{-+} \setminus H_{++}$, and hence a third point in H_{-+}, then we fall back to our previous analysis of three points on a single face, adapted by taking negatives of all x-coordinates. In particular, because V has x-coordinate 0, which minimizes the x-coordinate in absolute value among the points in $P \cap H_{-+}$, either Q or R must maximize the x-coordinate in absolute value and have x-coordinate -1. Assuming Q has x-coordinate -1, in order for Q to be separated from $U = (1, y_0 + \alpha, z_0)$ by ℓ^1-distance 2, we must have $Q = (-1, y_0 + \alpha, z_0)$. In order for $\{V, Q, R\}$ to meet the required form for three points of P on H_{-+} established in Step 1, we must have $R = (-\alpha, y_0, z_0 + 1)$. However, in this case we see that $\|U - R\|_1 = 2 + 2\alpha = 2$; hence $\alpha = 0$, which contradicts the assumption that $R \notin H_{++}$. □

6. Conditional results in higher dimensions

In the remainder of our discussion, we use the terms ℓ^1-*sphere* and ℓ^1-*ball* as before, defined analogously to regular spheres and balls in $\mathbb{R}^d$, with the usual distance replaced by ℓ^1-distance. In an effort to establish results in higher dimensions, we make the following observations, heavily inspired by our journey thus far:

(a) As noted at the beginning of Section 5, our proof of Theorem 2.5 does not naturally generalize to higher dimensions, because in dimension $d \geq 3$ it is not necessarily the case that if the largest ℓ^1-distance determined by a finite set $P \subseteq \mathbb{R}^d$ is λ, then P is contained in a closed ℓ^1-ball of diameter λ. However, the argument in Lemma 4.3 does generalize to all dimensions: the distance λ can be removed from an ℓ^1-ball of diameter λ by removing the closed upper ℓ^1-hemisphere. In particular, if we somehow could capture our set inside such a ball, then by mimicking the proof of Theorem 2.5, the problem is reduced to determining maximal configurations of points arranged on single closed upper ℓ^1-hemisphere, which would then facilitate an induction on the number of distinct distances.

(b) Suppose $P \subseteq \mathbb{R}^d$ is a finite set determining at most k distinct ℓ^1-distances, with largest ℓ^1-distance λ. By translating and scaling, we can assume that $\lambda = 2k$ and the "southernmost point" of P, minimizing the x_d-coordinate, is $-ke_d$, where e_d is the d-th standard basis vector. An enticing observation, particularly in juxtaposition with (a), is the following: if ke_d is also in P, then, since $2k$ is the largest ℓ^1-distance, P is contained in the intersection of the closed ℓ^1-ball of radius $2k$ centered at

$-ke_d$ and the closed ℓ^1-ball of radius $2k$ centered at ke_d, which is conveniently the closed ℓ^1-ball of radius k centered at the origin.

(c) Inspired by the simplicity of Case 1 in the proof of Theorem 2.6, we see that if U lies on an upper ℓ^1-hemisphere $H \subseteq \mathbb{R}^d$, then the ℓ^1-distance between U and the "north pole" of H is determined entirely by the x_d-coordinate of U. More specifically, if H is the closed upper ℓ^1-hemisphere of radius k centered at the origin and $U = (x_1, \dots, x_d) \in H$, then

$$\|U - ke_d\|_1 = |x_1| + \cdots + |x_{d-1}| + k - x_d = 2(k - x_d).$$

In particular, if $ke_d \in P$ and P determines only k distinct ℓ^1-distances $\{\lambda_i\}_{i=1}^k$, then the points of $(P \cap H) \setminus ke_d$ are restricted to the hyperplanes $\{x_d = c_i\}$ for $1 \le i \le k$, where $c_i = k - \lambda_i/2$. Further, the intersection of H with the hyperplane $\{x_d = c_i\}$ is

$$\{(x_1, \dots, x_{d-1}, c_i) : |x_1| + \cdots + |x_{d-1}| = k - c_i\},$$

which is a copy of the ℓ^1-sphere of radius $k - c_i$ centered at the origin in $\mathbb{R}^{d-1}$. This would allow us to analyze $P \cap H$ by inducting on dimension, analogous to the invocation of Theorem 2.5 during Case 1 in the proof of Theorem 2.6.

These three items combine to a clear aspirational reality: given a finite set $P \subseteq \mathbb{R}^d$ that determines at most k distinct ℓ^1-distances, the largest of which is λ, letting H denote the closed upper ℓ^1-hemisphere of radius λ centered at the "southernmost" point of P, we could fully adapt the proof of Theorem 2.5 and induct on both d and k, if only we could assume that the "north pole" of H is also in P. If we were considering the usual Euclidean distance, this would be no obstruction at all, as we could rotate our set and assume without loss of generality that the largest distance λ occurs parallel to the x_d-axis. However, since ℓ^1-distance is not invariant under rotation, we require an additional assumption to establish a conditional version of Conjecture 2.7. The following definition, conjecture, and theorem fully formalize this conditional result, after which we conclude our discussion.

Definition 6.1. Given $P \subseteq \mathbb{R}^d$ and an ℓ^1-distance $\lambda > 0$, we say that λ *occurs in an axis-parallel direction* if there exists $x \in P$ and $1 \le i \le d$ such that $x + \lambda e_i \in P$, where e_i is the i-th standard basis vector. Further, if P is bounded, we say that P is *axis-parallel* if the largest ℓ^1-distance determined by P occurs in an axis-parallel direction.

Conjecture 6.2. *Suppose $d, k \in \mathbb{N}$. If P is of maximal size amongst subsets of $\mathbb{R}^d$ determining at most k distinct ℓ^1-distances, then P is axis-parallel. The same holds within the class of sets contained in an ℓ^1-sphere in $\mathbb{R}^d$.*

Theorem 6.3. *Conjecture 6.2 implies Conjecture 2.7.*

Proof. We proceed via two inductions, one on the dimension d and another on the number of distinct ℓ^1-distances k. We streamline the argument by defining the following propositions for each $d \in \mathbb{N}$ and each nonnegative integer k:

<u>Opt(d, k)</u>: $\Lambda_d(k)$ is the unique set, up to ℓ^1-similarity, of maximal size amongst subsets of $\mathbb{R}^d$ determining at most k distinct ℓ^1-distances. Conjecture 2.7 is precisely the statement that Opt(d, k) holds for all $d, k \in \mathbb{N}$.

<u>S-Opt(d, k)</u>: $\Lambda_d(k) \setminus \Lambda_d(k-2)$ is the unique set, up to ℓ^1 similarity, of maximal size amongst sets contained in an ℓ^1-sphere in $\mathbb{R}^d$ determining at most k distinct ℓ^1-distances. For $k = 0$ or 1, we take the convention that $\Lambda_d(-1) = \Lambda_d(-2) = \varnothing$.

<u>H-Opt(d, k)</u>: Let H denote the closed upper ℓ^1-hemisphere of radius k centered at the origin in $\mathbb{R}^d$. If $ke_d \in E \subseteq H$ and E determines at most k distinct ℓ^1-distances, the largest of which is $2k$, then $|E| \le |\Lambda_d(k) \cap H|$, and $|E| = |\Lambda_d(k) \cap H|$ if and only if $E = \Lambda_d(k) \cap H$.

For the necessary base cases, we note that S-Opt$(1, k)$ and H-Opt$(1, k)$ trivially hold for all $k \in \mathbb{N}$ as ℓ^1-spheres and ℓ^1-hemispheres in $\mathbb{R}$ contain just two points and one point, respectively. Also, Opt$(d, 0)$ holds trivially for all $d \in \mathbb{N}$ because a set determining no ℓ^1-distances contains at most a single point. Under the assumption that Conjecture 6.2 holds, we verify Conjecture 2.7 by establishing the following implications:

(1) S-Opt$(d-1, k)$ for all $k \in \mathbb{N} \Longrightarrow$ H-Opt(d, k) and S-Opt(d, k) for all $k \in \mathbb{N}$, so S-Opt(d, k) and H-Opt(d, k) hold for all $d, k \in \mathbb{N}$.

(2) Opt$(d, k-1)$ and H-Opt$(d, k) \Longrightarrow$ Opt(d, k), so Opt(d, k) holds for all $d, k \in \mathbb{N}$, as required.

Proof of (2). Fix $d, k \in \mathbb{N}$, and suppose Opt$(d, k-1)$ and H-Opt(d, k) hold. Suppose $P \subseteq \mathbb{R}^d$ determines at most k distinct ℓ^1-distances, and has maximal size amongst sets with this property. By scaling we can assume that the largest ℓ^1-distance determined by P is $2k$, which by Conjecture 6.2 we know occurs in an axis-parallel direction. After permuting coordinates and translating, we can assume that $-ke_d, ke_d \in P$. As noted in (b), this implies that P is contained the closed ℓ^1-ball of radius k centered at the origin. To verify this, suppose $U = (x_1, \dots, x_d) \in P$ with $\|U\|_1 > k$. If $x_d \ge 0$, then $\|U - (-ke_d)\|_1 = \|U\|_1 + k > 2k$, while if $x_d \le 0$, the same holds for the ℓ^1-distance between U and ke_d, contradicting the fact that $2k$ is the largest ℓ^1-distance determined by P.

As noted in (a), the ℓ^1-distance $2k$ can be eliminated from P by removing the points in $P \cap H$, where H is the closed upper ℓ^1-hemisphere of radius k centered at the origin. This is because, within a closed ℓ^1-ball of radius k, the ℓ^1-distance $2k$ only occurs between pairs of points on opposite faces of the boundary. By H-Opt(d, k), we know that $|P \cap H| \le |\Lambda_d(k) \cap H|$, and further $|P \cap H| = |\Lambda_d(k) \cap H|$ if and only if $P \cap H = \Lambda_d(k) \cap H$, in which case the ℓ^1-distances determined by P are $2, 4, \dots, 2k$.

Because the ℓ^1-distance $2k$ does not occur in $P \setminus H$, we have that $P \setminus H$ determines at most $k - 1$ distinct ℓ^1-distances. By $\operatorname{Opt}(d, k-1)$, we know that $|P \setminus H| \le |\Lambda_d(k-1)|$, and further $|P \setminus H| = |\Lambda_d(k-1)|$ if and only if $P \setminus H$ is ℓ^1-similar to $\Lambda_d(k-1)$. In order for both $P \cap H$ and $P \setminus H$ to attain their maximum possible sizes, the ℓ^1-distances determined by $P \setminus H$ must be $2, 4, \dots, 2k-2$. In this case, since an ℓ^1-similar copy of $\Lambda_d(k-1)$ is uniquely determined by its "south pole" and its largest distance, we know that $P \setminus H$ must be $\Lambda_d(k-1)$ shifted down by 1. In other words,

$$\begin{aligned} P \setminus H &= \{(x_1, \dots, x_d - 1) \in \mathbb{Z}^d : |x_1| + \cdots + |x_d| \le k-1,\ x_1 + \cdots + x_d \equiv k-1 \pmod 2\} \\ &= \{(x_1, \dots, x_d) \in \mathbb{Z}^d : |x_1| + \cdots + |x_d + 1| \le k-1,\ x_1 + \cdots + x_d \equiv k \pmod 2\}. \end{aligned}$$

The latter description ensures that $P \setminus H \subseteq \Lambda_d(k) \setminus H$, and conversely, if $U = (x_1, \dots, x_d) \in \Lambda_d(k) \setminus H$, then either $x_d < 0$ or $\|U\|_1 \le k - 2$. In either case $|x_1| + \cdots + |x_d + 1| \le k - 1$, and hence $U \in P \setminus H$. Bringing everything together, we have that if $P \cap H$ and $P \setminus H$ both attain their maximum possible size, then $P \cap H = \Lambda_d(k) \cap H$ and $P \setminus H = \Lambda_d(k) \setminus H$; hence $P = \Lambda_d(k)$, so $\operatorname{Opt}(d, k)$ holds.

Proof of (1). Fix $d \ge 2$, suppose $\operatorname{S-Opt}(d-1, k)$ holds for all $k \in \mathbb{N}$, and fix $k \in \mathbb{N}$. Suppose $P \subseteq \mathbb{R}^d$ is contained in the ℓ^1-sphere S of radius k centered at the origin, and that P has maximal size amongst all such sets determining at most k distinct ℓ^1-distances. To establish $\operatorname{S-Opt}(d, k)$, we must show that $P = \Lambda_d(k) \setminus \Lambda_d(k-2)$. Thanks to our inductive hypothesis, we can assume P determines exactly k distinct ℓ^1-distances, not fewer, and we denote those ℓ^1-distances by $\lambda_1 < \cdots < \lambda_k$. By the ℓ^1-sphere component of Conjecture 6.2, we know that λ_k occurs in an axis-parallel direction. By permuting coordinates, we assume that λ_k occurs in the last coordinate direction; in other words $(x_1, \dots, x_d), (x_1, \dots, x_d + \lambda_k) \in P$ for some $x_1, \dots, x_d \in \mathbb{R}$ with $|x| + \cdots + |x_d| = |x_1| + \cdots + |x_d + \lambda_k| = k$, which in particular forces $x_d = -\lambda_k/2$ and $|x_1| + \cdots + |x_{d-1}| = k - \lambda_k/2$. This transformation is allowable because our end goal, $\Lambda_d(k) \setminus \Lambda_d(k-2)$, is invariant under coordinate permutation.

We argue (informally for the moment) that the only reasonable choice is $\lambda_k = 2k$ and $x_1 = \cdots = x_{d-1} = 0$, meaning $ke_d, -ke_d \in P$. This is because, since λ_k is the largest ℓ^1-distance, P is contained in the intersection of the closed ℓ^1-balls of radius λ_k centered at $(x_1, \dots, x_{d-1}, -\lambda_k/2)$ and $(x_1, \dots, x_{d-1}, \lambda_k/2)$, respectively, and this intersection is the ℓ^1-ball of radius $\lambda_k/2$ centered at $(x_1, \dots, x_{d-1}, 0)$. However, P is also contained in S, so if it is not the case that $\lambda_k = 2k$, then P would in fact be contained in the intersection of an ℓ^1-sphere with a closed ℓ^1-ball of a smaller radius, which is at most a closed ℓ^1-hemisphere. The idea that a maximal subset of an ℓ^1-sphere determining at most k distinct ℓ^1-distances could actually be contained in a closed ℓ^1-hemisphere is intuitively suspect, and we return to this issue near the end of the proof. For now, we assume $-ke_d, ke_d \in P$.

We denote by H the closed upper ℓ^1-hemisphere of S, and we establish H-Opt(d,k) along the way. As discussed in (c), all the points of $(P \cap H) \setminus ke_d$ have x_d-coordinates in the list $c_1 > c_2 > \cdots > c_k$, where $c_i = k - \lambda_i/2$. For each c_i, the points of H with x_d-coordinate equal to c_i take the form $(x_1, \dots, x_{d-1}, c_i)$, where $|x_1| + \cdots + |x_{d-1}| = k - c_i$, and we refer to the set of such points as S_i. With regard to ℓ^1-distances, S_i is equivalent to an ℓ^1-sphere in $\mathbb{R}^{d-1}$, centered at the origin with radius $k - c_i$. All ℓ^1-distances determined by $P \cap S_i$ are at most $2(k - c_i) = \lambda_i$, so $P \cap S_i$ determines at most i distinct ℓ^1-distances. By our inductive hypothesis, $|P \cap S_i| \le |\Lambda_{d-1}(i)| - |\Lambda_{d-1}(i-2)|$, with equality holding if and only if the projection $P \cap S_i$ onto the first $d-1$ coordinates is ℓ^1-similar to $\Lambda_{d-1}(i) \setminus \Lambda_{d-1}(i-2)$, so in particular $\lambda_1, \dots, \lambda_i$ form an arithmetic progression. Since $k - z_i = \lambda_i/2$ and $\lambda_k = 2k$, this equality holds for all $1 \le i \le k$ if and only if $P \cap S_i = \{(x, k-i) : x \in \Lambda_{d-1}(i) \setminus \Lambda_{d-1}(i-2)\}$ for all $1 \le i \le k$. Here we note that if it were not the case that $ke_d, -ke_d \in P$ as previously assumed, then $P \cap S_i$ would be, at most, equivalent to an ℓ^1-hemisphere in $\mathbb{R}^{d-1}$, in which case our inductive hypothesis would prohibit it from having $|\Lambda_{d-1}(i)| - |\Lambda_{d-1}(i-2)|$ elements.

In summary,

$$|P \cap H| \le \sum_{i=0}^{k} |\Lambda_{d-1}(i)| - |\Lambda_{d-1}(i-2)|, \tag{4}$$

taking $\Lambda_{d-1}(-1)$ and $\Lambda_{d-1}(-2)$ to be empty, and equality holds if and only if

$$P \cap H = \bigcup_{i=0}^{k} \{(x, k-i) : x \in \Lambda_{d-1}(i) \setminus \Lambda_{d-1}(i-2)\} = \Lambda_d(k) \cap H,$$

which establishes H-Opt(d,k).

Further, $P \cap H$ can contain at most $\sum_{i=0}^{k-1} |\Lambda_{d-1}(i)| - |\Lambda_{d-1}(i-2)|$ points with $x_d > 0$. Letting H' denote the closed lower ℓ^1-hemisphere of S, we employ reasoning identical to the above to yield the same upper bound (4) on $|P \cap H'|$, with equality holding if and only if

$$P \cap H' = \bigcup_{i=0}^{k} \{(x, i-k) : x \in \Lambda_{d-1}(i) \setminus \Lambda_{d-1}(i-2)\} = \Lambda_d(k) \cap H.$$

Further, $P \cap H'$ can contain at most $\sum_{i=0}^{k-1} |\Lambda_{d-1}(i)| - |\Lambda_{d-1}(i-2)|$ points with $x_d < 0$. Putting all this together, we have

$$\begin{aligned} |P| &= |P \cap (H \setminus H')| + |P \cap (H' \setminus H)| + |P \cap H \cap H'| \\ &\le 2\left(\sum_{i=0}^{k-1} |\Lambda_{d-1}(i)| - |\Lambda_{d-1}(i-2)|\right) + (|\Lambda_{d-1}(k)| - |\Lambda_{d-1}(k-2)|), \end{aligned}$$

and equality holds if and only if

$$P = \bigcup_{i=-k}^{k} \{(x, i) : x \in \Lambda_{d-1}(k - |i|) \setminus \Lambda_{d-1}(k - |i| - 2)\} = \Lambda_d(k) \setminus \Lambda_d(k-2).$$

Therefore, S-Opt(d, k) holds, and the induction on dimension is complete. □

Remark. If one is specifically interested in dimension $d = 3$, then, because we have fully resolved the problem in dimension $d = 2$, no inductive hypothesis is needed for dimension, just for the number of ℓ^1-distances. In other words, if Opt$(3, k-1)$ holds, then $\Lambda_3(k)$ is the unique set, up to ℓ^1-similarity, of maximal size amongst axis-parallel subsets of $\mathbb{R}^3$ determining at most k distinct ℓ^1-distances. In particular, because Theorem 2.6 tells us that O(3, 1) holds, we know that $\Lambda_3(2)$, which contains 19 points and is pictured in Figure 1, right, is uniquely optimal amongst axis-parallel sets determining only two ℓ^1-distances. However, we cannot make the analogous claim for $\Lambda_3(3)$, because we cannot exclude the possibility of a non-axis-parallel set determining two ℓ^1-distances that contains more than 19 points (or that contains exactly 19 points but is not ℓ^1-similar to $\Lambda_3(2)$). This possibility disables our bridge from two ℓ^1-distances to three, and exemplifies the need in assuming Conjecture 6.2 if we wish to glean additional information in dimension $d \geq 3$.

Acknowledgements

This research was initiated during the Summer 2019 Kinnaird Institute Research Experience at Millsaps College. All authors were supported during the summer by the Kinnaird Endowment, gifted to the Millsaps College Department of Mathematics. At the time of submission, all authors except Alex Rice were Millsaps College undergraduate students. The authors would like to thank Alex Iosevich for his helpful references, and Tomasz Tkocz for alerting us to previous research done in the $k = 1$ case. Finally, the authors would like to thank the anonymous referee for encouraging comments and helpful recommendations, particular regarding Sections 5 and 6.

References

[Bandelt, Chepoi, and Laurent 1998] H.-J. Bandelt, V. Chepoi, and M. Laurent, "Embedding into rectilinear spaces", *Discrete Comput. Geom.* **19**:4 (1998), 595–604. MR Zbl

[Erdős 1946] P. Erdős, "On sets of distances of n points", *Amer. Math. Monthly* **53** (1946), 248–250. MR Zbl

[Erdős and Fishburn 1996] P. Erdős and P. Fishburn, "Maximum planar sets that determine k distances", *Discrete Math.* **160**:1-3 (1996), 115–125. MR Zbl

[Garibaldi, Iosevich, and Senger 2011] J. Garibaldi, A. Iosevich, and S. Senger, *The Erdős distance problem*, Student Math. Lib. **56**, Amer. Math. Soc., Providence, RI, 2011. MR Zbl

[Guth and Katz 2015] L. Guth and N. H. Katz, "On the Erdős distinct distances problem in the plane", *Ann. of Math.* (2) **181**:1 (2015), 155–190. MR Zbl

[Guy 1983] R. K. Guy, "An olla-podrida of open problems, often oddly posed", *Amer. Math. Monthly* **90**:3 (1983), 196–200. MR

[Koolen, Laurent, and Schrijver 2000] J. Koolen, M. Laurent, and A. Schrijver, "Equilateral dimension of the rectilinear space", *Des. Codes Cryptogr.* **21**:1-3 (2000), 149–164. MR Zbl

[Shinohara 2008] M. Shinohara, "Uniqueness of maximum planar five-distance sets", *Discrete Math.* **308**:14 (2008), 3048–3055. MR Zbl

[Wei 2012] X. Wei, "A proof of Erdős–Fishburn's conjecture for $g(6) = 13$", *Electron. J. Combin.* **19**:4 (2012), art. id. 38. MR Zbl

Received: 2019-12-05 Revised: 2020-05-14 Accepted: 2020-05-23

balajv@millsaps.edu *Department of Mathematics, Millsaps College, Jackson, MS, United States*

edwarof@millsaps.edu *Department of Mathematics, Millsaps College, Jackson, MS, United States*

loftiam@millsaps.edu *Department of Mathematics, Millsaps College, Jackson, MS, United States*

mcharsk@millsaps.edu *Department of Mathematics, Millsaps College, Jackson, MS, United States*

philllg@millsaps.edu *Department of Mathematics, Millsaps College, Jackson, MS, United States*

riceaj@millsaps.edu *Department of Mathematics, Millsaps College, Jackson, MS, United States*

tsegabl@millsaps.edu *Department of Mathematics, Millsaps College, Jackson, MS, United States*

mathematical sciences publishers msp

INVOLVE 13:3 (2020)
https://doi.org/10.2140/involve.2020.13.511

Total difference chromatic numbers of graphs

Ranjan Rohatgi and Yufei Zhang

(Communicated by Joseph A. Gallian)

Inspired by graceful labelings and total labelings of graphs, we introduce the idea of total difference labelings. A k-total labeling of a graph G is an assignment of k distinct labels to the edges and vertices of a graph so that adjacent vertices, incident edges, and an edge and its incident vertices receive different labels. A k-total difference labeling of a graph G is a function f from the set of edges and vertices of G to the set $\{1, 2, \dots, k\}$ that is a k-total labeling of G and for which $f(\{u, v\}) = |f(u) - f(v)|$ for any two adjacent vertices u and v of G with incident edge $\{u, v\}$. The least positive integer k for which G has a k-total difference labeling is its total difference chromatic number, $\chi_{\mathrm{td}}(G)$. We determine the total difference chromatic number of paths, cycles, stars, wheels, gears and helms. We also provide bounds for total difference chromatic numbers of caterpillars, lobsters, and general trees.

1. Introduction

Graph labelings have been widely studied for over half a century, as evidenced by the sheer quantity of results contained in Gallian's regularly updated survey [1998] of the subject. We define the *total difference chromatic number* of a graph, inspired by graceful and graceful-like labelings (see, for example, [Bi et al. 2017] for some recent work on the subject) and total labelings.

A graph labeling is an assignment of integers to the vertices and/or edges of a graph which follows certain conditions. For example, in a graceful labeling (introduced as a "β-valuation" in [Rosa 1967] and first referred to as a "graceful labeling" in [Golomb 1972]) of a graph with m edges, each vertex is assigned an integer from the set $\{0, 1, 2, \dots, m\}$ and each edge gets as its label the absolute value of the difference of the labels on its incident vertices. If the resulting edge labels are distinct, the graph is said to have a graceful labeling.

MSC2010: 05C15, 05C78.

Keywords: graph labelings, graceful graphs, total colorings.

Zhang was supported by the Sister Miriam Cooney, CSC '51 Endowed Grant for Undergraduate Student Research in Mathematics.

A (proper) *k-labeling* of a graph is an assignment of k labels to the vertices of a graph so that adjacent vertices have receive different labels. Similarly, a (proper) *k-edge-labeling* is an assignment of k labels to the edges of a graph so that incident edges receive different labels. Combining both ideas, a (proper) *k-total-labeling* is an assignment of k labels to the edges and vertices of a graph so that adjacent vertices, incident edges, and an edge and its incident vertices receive different labels. The minimum number of labels for which a k-labeling (respectively, k-edge-labeling, k-total-labeling) of a graph G exists is called the *chromatic number* (respectively, *edge chromatic number*, *total chromatic number*) and is denoted by $\chi(G)$ (respectively, $\chi'(G)$, $\chi''(G)$). It is also common to think of labels as colors, and therefore call labelings "colorings" instead. As our labels are integers, we stick with the "labeling" terminology. Unless otherwise specified, all labelings are assumed to be proper.

Combining the mechanism of labeling the edges of a graph from graceful labelings with total labelings, we define a *k-total difference labeling* and the *total difference chromatic number* of a graph in Section 2, along with some preliminary observations. In Sections 3, 4, and 5 we determine the total difference chromatic number for arbitrary paths, cycles, stars, and wheels. In the final section, we prove lower and upper bounds for the total difference chromatic numbers of caterpillars, lobsters, and general trees.

2. Preliminaries

We present a few definitions, including that of the *total difference chromatic number* of a graph, and preliminary observations related to the total difference chromatic number of a graph in this section. We use $V(G)$ and $E(G)$ to denote the vertex set and edge set, respectively, of a graph G. For integers $a < b$ we denote by $[a, b]$ the set of integers from a to b, inclusive.

Definition 2.1. A *labeling* of a graph G is a function $c : V(G) \to \mathbb{Z}^+$, where $\mathbb{Z}^+$ is the set of positive integers, such that $c(u) \neq c(v)$ for any two adjacent vertices u and v. If the maximum label assigned to any vertex is k, we say that c is a *k-labeling* of G.

Definition 2.2. A *total labeling* of a graph G is a function $f : V(G) \cup E(G) \to \mathbb{Z}^+$ such that $f(u) \neq f(v)$ for any two adjacent vertices u and v, $f(\{u, v\}) \neq f(\{v, w\})$ for any two incident edges $\{u, v\}$ and $\{v, w\}$, and $f(u) \neq f(\{u, v\})$ for any edge $\{u, v\}$ and an incident vertex u. If the maximum label assigned to any edge or vertex is k, we say that f is a *k-total labeling* of G.

Definition 2.3. Let f be a k-total labeling of G. We say that f is a *k-total difference labeling* if $f(\{u, v\}) = |f(u) - f(v)|$ whenever $\{u, v\} \in E(G)$. That is, the label assigned to an edge is the absolute difference of the labels assigned to its incident vertices.

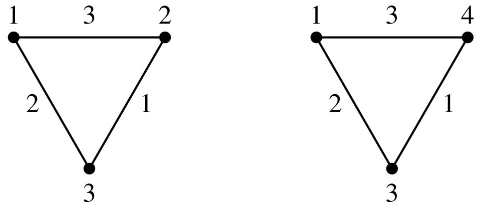


Figure 1. Upper bounds for $\chi''(K_3)$ (left) and $\chi_{\text{td}}(K_3)$ (right).

Definition 2.4. The *total difference chromatic number* of a graph G, denoted by $\chi_{\text{td}}(G)$, is the smallest integer k for which G has a k-total difference labeling.

We first prove that the total difference chromatic number is well-defined.

Proposition 2.5. *Given a graph G with n vertices*, $\chi''(G) \le \chi_{\text{td}}(G) \le 3^{n-1}$.

Proof. The first inequality follows from the observation that a total difference labeling is precisely a total labeling with an additional condition specifying how the edges are labeled.

To show that $\chi_{\text{td}}(G) \le 3^{n-1}$, arbitrarily label the n vertices with distinct elements from the set $\{3^0, 3^1, 3^2, \dots, 3^{n-1}\}$. All edge labels will be of the form $3^i - 3^j$ for distinct integers $i, j \in [0, n-1]$ with $i > j$. Notice that $3^i - 3^j = 3^j(3^{i-j} - 1)$, which implies that all edge labels will be distinct and different from every vertex label. □

Example 2.6. As an example to show that the total chromatic number and total difference chromatic number can be different, we consider the complete graph K_3. As shown in Figure 1, $\chi''(K_3) \le 3$. If we attempted to totally label K_3 with only two colors, adjacent vertices would have the same label; hence $\chi''(K_3) = 3$.

On the other hand, $\chi_{\text{td}}(K_3) = 4$. Figure 1 gives 4 as an upper bound. Notice that only one of 1 and 2 can be used as a vertex label since all vertices are adjacent and the edge between the two vertices with these labels would also have label 1. Therefore, 4 is also a lower bound for $\chi_{\text{td}}(K_3)$.

The following definition and propositions help us construct k-total difference labelings of several graphs.

Definition 2.7. Let $c : V(G) \to [1, k]$ be a k-labeling of a graph G:

(a) A *double* is a pair of adjacent vertices u and v such that $c(u) = 2c(v)$. We write $(c(u), c(v))$-*double* if we want to specify the labels of the vertices that form the double.

(b) A *triple* is a set of three vertices u, v, w with $\{u, v\}$,$\{v, w\} \in E(G)$ such that $|c(u) - c(v)| = |c(v) - c(w)|$. We write $(c(w), c(v), c(u))$-*triple* if we want to specify the labels of the vertices that form the triple.

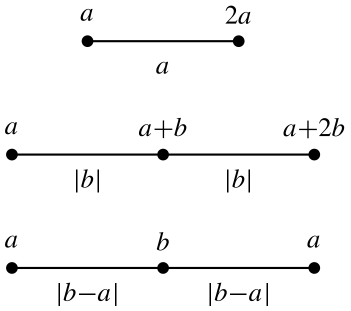


Figure 2. In a *double*, an induced edge label is equal to the label on an incident vertex. In a *triple*, two incident edges receive the same label.

See Figure 2 for an example of a $(2a, a)$-double, an $(a+2b, a+b, a)$-triple and an (a, b, a)-triple.

Proposition 2.8. *If a graph G with n vertices has* $\operatorname{diam}(G) \leq 2$, *then* $\chi_{\mathrm{td}}(G) \geq n$ *and in any k-total difference labeling of G all vertices must have different labels.*

Proof. We prove this proposition by contradiction. Assume that a graph G with n vertices has $\operatorname{diam}(G) \leq 2$ and $\chi_{\mathrm{td}}(G) = k < n$. Then there must exist two vertices u and v of G such that $f(u) = f(v)$, where f is any k-total difference labeling of G. Since f is a k-total difference labeling of G, the vertices u and v cannot be adjacent. Since $\operatorname{diam}(G) \leq 2$, there must exist a vertex w adjacent to both u and v. But then u, v, and w form a triple as $|f(u) - f(w)| = |f(v) - f(w)|$, contradicting that $\chi_{\mathrm{td}}(G) < n$. □

Lemma 2.9. *Let c be a proper k-labeling of a graph G and let f be the (not necessarily proper) total labeling of G defined by $f(v) = c(v)$ for a vertex v of G, and $f(\{u, v\}) = |c(u) - c(v)|$ for an edge $\{u, v\}$ of G. Then f is a k-total difference labeling of G if and only if c does not contain any doubles or triples.*

Proof. Suppose that f is a k-total difference labeling of G. By Definition 2.3, f cannot assign a vertex and an incident edge the same label, nor can it assign two incident edges the same label. Hence, c cannot contain a double or triple.

Now suppose that c does not contain doubles or triples. We must show that f does not assign the same label to adjacent vertices, a vertex and an incident edge, or two incident edges.

First, as c is a proper labeling of G, no adjacent vertices of G receive the same label. That is, $f(u) \neq f(v)$ for adjacent vertices u and v.

Next, assume, by contradiction, that a vertex v and an incident edge $\{u, v\}$ are assigned the same label by f. This implies that $f(v) = f(\{u, v\}) = |f(u) - f(v)|$.

Solving this equation for $f(u)$, we see that either $f(u) = 2f(v)$ or $f(u) = 0$. As $f(v) = c(v)$ for any vertex v of G, we get that either $c(u) = 2c(v)$, which contradicts the assumption that c does not contain a double, or $c(u) = 0$, which contradicts the assumption that c is a labeling of G.

Finally, assume, by contradiction that for two incident edges, $\{u, v\}$ and $\{v, w\}$, we have $f(\{u, v\}) = f(\{v, w\})$. An analysis similar to that in the previous case shows that u, v, and w form a triple under the labeling c. □

Remark 2.10. Lemma 2.9 provides a method of determining if a proposed k-total difference labeling actually is one only by looking at the vertex labels. Therefore, when constructing k-total difference labelings of families of graphs throughout this paper, we only provide a (proper) vertex labeling and check that it does not contain doubles or triples.

We conclude this section with a result relating the total difference chromatic number of a graph to those of its subgraphs.

Proposition 2.11. *If G' is a subgraph of G, then $\chi_{\text{td}}(G') \le \chi_{\text{td}}(G)$.*

Proof. Let f be a k-total difference labeling of G. As we are removing vertices or edges from G to get G', the restriction of f to G' is an ℓ-total difference labeling of G' for some $\ell \le k$. □

3. Paths and cycles

We determine the total difference chromatic numbers of paths and cycles. Lemma 2.9 enables us to consider solely the labels on the vertices of these graphs.

Theorem 3.1. *For any path P_n with $n \ge 4$, $\chi_{\text{td}}(P_n) = 4$.*

Proof. We denote the vertices in P_n by $v_1, v_2, \dots, v_n$ from left to right as in Figure 3. We label v_i with 1 if $i \equiv 1 \pmod 3$, with 4 if $i \equiv 2 \pmod 3$, and with 3 if $i \equiv 0 \pmod 3$. It is straightforward to see that this labeling does not create any doubles or triples, and hence $\chi_{\text{td}}(P_n) \le 4$.

We now show that $\chi_{\text{td}}(P_n) \ge 4$ by contradiction. Assume we can 3-total difference label P_n. To avoid (2, 1)-doubles, we must label every other vertex with 3, in which case we form a (3, 2, 3)- or (3, 1, 3)-triple. □

We now turn our attention to cycles. The total difference chromatic number of C_n depends on n, the number of vertices in the cycle.

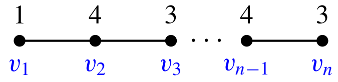


Figure 3. We denote the vertices in P_n by $v_1, v_2 \dots, v_n$ and provide a construction that shows $\chi_{\text{td}}(P_n) \le 4$.

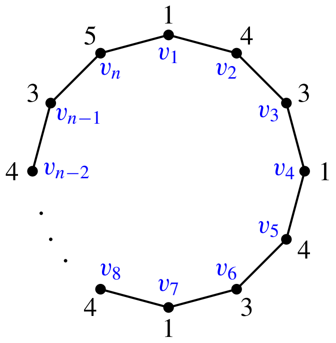


Figure 4. A construction that shows $\chi_{\text{td}}(C_n) \leq 5$ when $n \equiv 1 \pmod 3$.

Theorem 3.2. *The total difference chromatic number of a cycle is given by* $\chi_{\text{td}}(C_n) = 4$ *if* $n \equiv 0 \pmod 3$ *and* $\chi_{\text{td}}(C_n) = 5$ *otherwise.*

Proof. Note that C_n is constructed by adding an edge between the two vertices of degree 1 in P_n. Therefore, by Proposition 2.11, $\chi_{\text{td}}(C_n) \geq \chi_{\text{td}}(P_n) = 4$. We denote the vertices by $v_1, v_2, \ldots, v_n$ in this cyclic order.

If $n \equiv 0 \pmod 3$ we can label each vertex v_i in C_n exactly as we labeled v_i in P_n, and hence $\chi_{\text{td}}(C_n) = 4$ in this case.

If $n \equiv 1 \pmod 3$, we label all vertices, except v_n, as in the case in which $n \equiv 0 \pmod 3$. We label v_n with 5. See Figure 4 for an example. To show $\chi_{\text{td}}(C_n) \geq 5$, we assume $\chi_{\text{td}}(C_n) = 4$ by contradiction. First notice that if any vertex gets label 2, then in avoiding (4, 2)- and (2, 1)-doubles, we form a (3, 2, 3)-triple. Therefore, all vertices must have labels 1, 3, and 4. Without loss of generality, suppose we label v_1 with 1. Then the label on v_2 is either 3 or 4 and v_3 gets the remaining label. To avoid doubles and triples, this sequence of labels must repeat, ending with vertex v_{n-1}. But then v_n is forced to have a label greater than 4, completing the proof that $\chi_{\text{td}}(C_n) = 5$ when $n \equiv 1 \pmod 3$.

If $n \equiv 2 \pmod 3$, we label vertices $v_1, v_2, \ldots, v_{n-5}$ as in the previous two cases. We then label $v_{n-4}, v_{n-3}, v_{n-2}, v_{n-1}, v_n$ with 5, 1, 4, 3, 5 respectively. The argument that shows $\chi_{\text{td}}(C_n) \geq 5$ for $n \equiv 2 \pmod 3$ is similar to that in the $n \equiv 1 \pmod 3$ case. □

4. Stars

We now consider stars, denoted by $K_{1,m}$. After determining the total difference chromatic number for an arbitrary star, we explicitly determine the different labels that the maximum-degree vertex of a star can receive in a k-total difference labeling for certain k. For all stars $K_{1,m}$ throughout this paper, we denote the vertex of degree m by v_0 and the remaining vertices by $v_1, v_2, \ldots, v_m$.

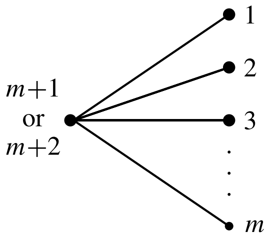


Figure 5. If m is even then $\chi_{\mathrm{td}}(K_{1,m}) \le m+1$, and if m is odd then $\chi_{\mathrm{td}}(K_{1,m}) \le m+2$.

Theorem 4.1. *Let $K_{1,m}$ be a star. Then*

$$\chi_{\mathrm{td}}(K_{1,m}) = \begin{cases} m+1, & m \text{ is even}, \\ m+2, & m \text{ is odd}. \end{cases}$$

Proof. Consider the star $K_{1,m}$. Since $\operatorname{diam}(K_{1,m})=2$, by Proposition 2.8, $\chi_{\mathrm{td}}(K_{1,m}) \ge m+1$.

We first consider the case where m is even. Give v_0 the label $m+1$ and v_i the label i for $1 \le i \le m$. Since v_0 has the greatest label and is odd and $v_1, \dots, v_m$ have distinct labels, $K_{1,m}$ has no doubles or triples.

If m is odd, then $\chi_{\mathrm{td}}(K_{1,m}) \le m+2$ if we label v_0 with $m+2$ and the rest as in the even case. Note that there are no doubles or triples using the same argument as before. (See Figure 5 for the construction.)

To prove $\chi_{\mathrm{td}}(K_{1,m}) \ge m+2$, we assume $\chi_{\mathrm{td}}(K_{1,m}) = m+1$ by contradiction. In this case, each label in $[1, m+1]$ appears on exactly one v_i. Note that we cannot use any integer in $[2, m]$ to label v_0: in this case there must be two leaves whose labels, along with the label for v_0, will form a $(\ell+1, \ell, \ell-1)$-triple for some $\ell \in [2, m]$. We also cannot use 1 or $m+1$ to label v_0 because we get a $(2, 1)$- or $(m+1, (m+1)/2)$-double. Thus, when m is odd, $\chi_{\mathrm{td}}(K_{1,m}) = m+2$. □

Remark 4.2. One can verify that if m is even, v_0 must receive the label $m+1$, and that if m is odd, v_0 must receive the label 1 or $m+2$.

We now determine the possible labels for v_0 in an $(m+r)$-total difference labeling of $K_{1,m}$ for $1 \le r \le m$. We will use this result in Section 6 to provide an upper bound for the total difference chromatic number of general trees.

Lemma 4.3. *Consider an $(m+r)$-total difference labeling of the star $K_{1,m}$ with $1 \le r \le m$. Then v_0 can have any label in the set $[1, r-1] \cup [m+2, m+r]$. If m is even or if m is odd and $r = (m+3)/2$, then v_0 can additionally receive the label $m+1$.*

Proof. Throughout this proof, we let ℓ be the label given to v_0. The greatest label on a vertex in $K_{1,m}$ must be $m+r$ since we are considering $(m+r)$-total difference labelings of $K_{1,m}$. Since $K_{1,m}$ has $m+1$ vertices, which require $m+1$ distinct labels from $[1, m+r]$, we must use all but $r-1$ integers in $[1, m+r-1]$. We look at four cases, based on the possible values of ℓ, namely: $\ell \in [1, r-1]$, $\ell = r$, $\ell \in [r+1, m]$, and $\ell \in [m+1, m+r]$. Note that if $r = 1$, we have only three distinct cases.

If $1 \le \ell \le r-1$, then ℓ could be the middle value in the $\ell - 1$ triples of the form $(\ell+i, \ell, \ell-i)$, where $1 \le i \le \ell - 1$. Additionally, ℓ could form a double with 2ℓ or with $\ell/2$ if ℓ is even. Therefore there are ℓ doubles and triples which include ℓ but are disjoint otherwise. (Notice that the $(\ell, \ell/2)$-double is not disjoint with the $(3\ell/2, \ell, \ell/2)$-triple). Since $\ell \le r-1$, we must avoid using at most $r-1$ integers, as desired. Therefore, v_0 can be labeled with any integer in $[1, r-1]$.

Suppose $\ell = r$; then again ℓ could be the middle value in the same $\ell - 1$ triples (of the form $(\ell+i, \ell, \ell-i)$ as above) with the integers in $[1, m+r]$, and a double with $2\ell = 2r \le m+r$. Therefore, we have $\ell = r$ labels that cannot be used, giving us only m possible labels for $m+1$ vertices. So, r cannot be used to label v_0.

Assume $r+1 \le \ell \le m$. Then ℓ could be the middle value in r triples (and perhaps more) with the integers in $[1, m+r]$, namely, $(\ell+i, \ell, \ell-i)$, where $1 \le i \le r$. As in the previous case, v_0 cannot be labeled using any element of $[r+1, m]$.

If $m+1 \le \ell \le m+r$, there are $m+r-\ell$ triples possible: $(\ell+i, \ell, \ell-i)$ for $i \in [1, m+r-\ell]$. Therefore, if $\ell = m+j$ for $j \in [1, r]$, there are $r-j$ possible triples. In addition to these triples, ℓ may also be part of a double. Therefore we must avoid $r-j+1$ labels for $j \in [1, r]$. If $j > 1$ then we are able to label v_0 with $m+j$ as there are at most $r-1$ forbidden labels. If $j = 1$ and m is even, ℓ could not be part of a double as it is odd, and hence ℓ could be $m+1$. If m is odd we could have up to r forbidden labels ($r-1$ triples and a double) and so ℓ could not be $m+1$. In all of these cases except when $r = (m+3)/2$, the triples and double are disjoint and hence ℓ cannot be $m+1$. The exception is described in the next paragraph.

Consider the case where m is odd and $r = (m+3)/2$. Then for $\ell = m+1$, we will have $r-1$ triples of the form $(\ell+i, \ell, \ell-i)$ with i in $[1, r-1]$, or equivalently in $[1, (m+1)/2]$. We now consider doubles. Notice that the $(\ell, \ell/2)$-double (equivalently, $(m+1, (m+1)/2)$-double) and the triple $(\ell+i, \ell, \ell-i)$ with $i = (m+1)/2$ both contain the label $(m+1)/2$. We can avoid both the double and triple simply by not using the label $(m+1)/2$; hence the possible double does not give us an additional label that must be avoided. It is also easy to check that the $(2\ell, \ell)$-double is not possible since $2\ell = 2m+2 > (3m+3)/2 = m+(m+3)/2 = m+r$ exceeds our bound. Hence, v_0 can have label $m+1$ as there are only $r-1$ possible violations of Lemma 2.9. □

5. Wheels, gears, and helms

We create the wheel, W_n, from the cycle C_{n-1} by adding a vertex adjacent to all other vertices in the cycle. We use Theorem 4.1 to determine the total difference chromatic number of wheels.

Theorem 5.1. *For $n \geq 4$, $\chi_{\mathrm{td}}(W_n) = \chi_{\mathrm{td}}(K_{1,n})$ except when n is 4 or 5. Explicitly,*

$$\chi_{\mathrm{td}}(W_n) = \begin{cases} 8, & n = 4, \\ 7, & n = 5, \\ n+1, & n \text{ is even and } n \geq 6, \\ n, & n \text{ is odd and } n \geq 7. \end{cases}$$

Proof. We denote by v_0 the vertex with degree $n-1$ and the remainder of the vertices by $v_1, \ldots, v_{n-1}$ in this cyclic order. Notice that $K_{1,n-1}$ is a subgraph of W_n so Proposition 2.11 implies $\chi_{\mathrm{td}}(W_n) \geq \chi_{\mathrm{td}}(K_{1,n-1})$.

Suppose $n = 4$. Note that in W_4, all vertices are adjacent to each other, so $W_4 = K_4$. We obtain a total difference labeling of W_4 by labeling the vertices with $1, 5, 7, 8$ (see the graph in the top left in Figure 6). It is straightforward to show that $\chi_{\mathrm{td}}(W_4) \geq 8$ by case analysis.

For W_5, we label v_0 with 7 and v_1, v_2, v_3, v_4 with $1, 3, 2, 5$, respectively. The graph in the top right in Figure 6 shows this construction.

Since W_5 has diameter 2, we use Proposition 2.8 to see that $5 \leq \chi_{\mathrm{td}}(W_5) \leq 7$. We will now show that it cannot be 5 or 6. Suppose that $\chi_{\mathrm{td}}(W_5) = 5$. If v_0 is labeled with 2, 3, or 4, then a triple must be formed. If v_0 gets the label 1, it must be adjacent to the vertex that gets labeled with 2. Finally, if v_0 is labeled with 5, one of the vertices with label 1 or 4 must be adjacent to the vertex with label 2. We can show similarly that $\chi_{\mathrm{td}}(W_5) \neq 6$ by ruling out possible labels for v_0.

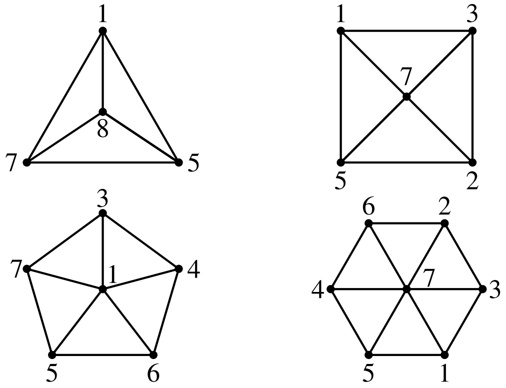


Figure 6. A construction that shows $\chi_{\mathrm{td}}(W_4) \leq 8$ and $\chi_{\mathrm{td}}(W_5)$, $\chi_{\mathrm{td}}(W_6)$, $\chi_{\mathrm{td}}(W_7) \leq 7$.

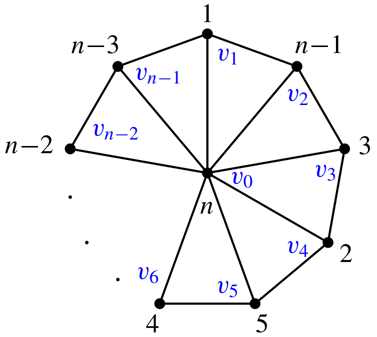


Figure 7. A construction that shows $\chi_{\text{td}}(W_n) \leq n$ when n is odd and $n \geq 7$.

For $\chi_{\text{td}}(W_6)$ and $\chi_{\text{td}}(W_7)$, constructions are provided in the bottom two graphs in Figure 6. Notice that we do indeed get $\chi_{\text{td}}(W_6) = \chi_{\text{td}}(W_7) = 7$ due to Theorem 4.1 about the total difference chromatic number of stars and Proposition 2.11.

We construct total difference labelings of W_n for the general case, in which $n \geq 8$. If n is odd, we label v_0 with n and the vertices $v_1, v_3, v_5, \dots, v_{n-2}$ with the consecutive odd integers $1, 3, 5, \dots, n-2$, respectively. We label $v_2, v_4, v_6, \dots, v_{n-1}$ with $n-1, 2, 4, 6, \dots, n-3$, respectively. See Figure 7 for the construction.

Notice that, except for vertices v_2 and v_{n-1}, vertices with even labels are adjacent to vertices with odd labels greater than their own. This immediately shows that these vertices cannot be in a double, and can be used to show, along with Proposition 2.8, that none of these vertices can be in a triple. It is then straightforward to check the remaining cases: that no triple can involve v_2, v_{n-1}, or only vertices with odd labels.

If n is even, our construction is nearly identical: vertices $v_1, v_3, v_5, \dots, v_{n-1}$ are labeled with $1, 3, 5, \dots, n-1$, respectively and $v_2, v_4, v_6, \dots, v_{n-2}$ with $n-2$, $2, 4, 6, \dots, n-4$. The only difference is that v_0 gets the label $n+1$. Appealing again to Theorem 4.1 and Proposition 2.11 completes the proof. □

We turn our attention to two families of graphs related to wheels: gears and helms. For $n \geq 4$, the gear G_n is formed by subdividing each edge on the outer circuit of W_n into two edges. The helm H_n is created from W_n by appending a leaf to each vertex on the outer circuit of W_n. See Figure 8 for several examples of gears and Figure 10 for an example of a helm.

Theorem 5.2. *For gears G_n with $n \geq 4$,*

$$\chi_{\text{td}}(G_n) = \begin{cases} 6, & n = 4, 5, \\ n+1, & n \text{ is even and } n \geq 6, \\ n, & n \text{ is odd and } n \geq 7. \end{cases}$$

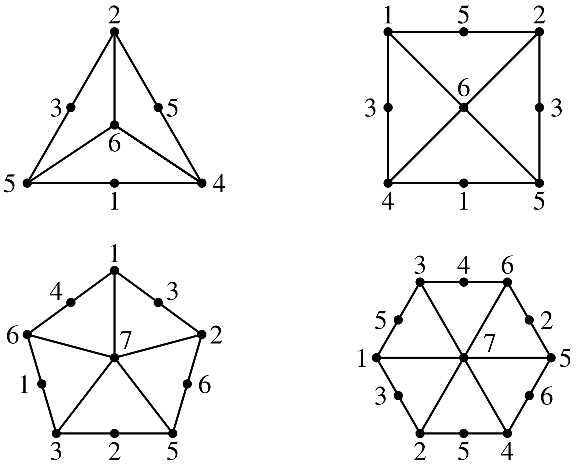


Figure 8. Constructions that show $\chi_{\text{td}}(G_4), \chi_{\text{td}}(G_5) \leq 6$ and $\chi_{\text{td}}(G_6), \chi_{\text{td}}(G_7) \leq 8$.

Proof. The constructions in Figure 8 show that the theorem holds when $4 \leq n \leq 7$. Observe that $K_{1,n-1}$ is a subgraph of G_n, so that $\chi_{\text{td}}(G_n) \geq \chi_{\text{td}}(K_{1,n-1})$ by Proposition 2.11. Theorem 4.1 implies that $\chi_{\text{td}}(K_{1,n-1}) = n$ if n is odd and $\chi_{\text{td}}(K_{1,n-1}) = n+1$ if n is even, so it remains to show by construction that $\chi_{\text{td}}(G_n) \leq \chi_{\text{td}}(K_{1,n-1})$ for $n \geq 8$. We break up the construction into two cases based upon the parity of n.

First suppose that n is even and $n \geq 8$. We let v_0 be the vertex of degree $n-1$ in G_n. We denote the remaining vertices by $v_1, v_2, \ldots, v_{2n-2}$ in this cyclic order, choosing v_1 to be any vertex adjacent to v_0. Notice that v_{2i} has degree 2, while v_{2i-1} has degree 3 for $1 \leq i \leq n-1$. We label v_0 with $n+1$ and $v_1, v_3, \ldots, v_{2n-3}$ with $1, 2, \ldots, n-1$, respectively. We then label $v_2, v_4, \ldots, v_{2n-2}$ with $n-2$, $n-1, 5, 6, \ldots, n-1, 2, 3$, respectively. It is straightforward to check that this does indeed give an $(n+1)$-total difference labeling of G_n when $n \geq 8$ is even. This construction is shown in Figure 9.

If n is odd and $n \geq 9$, we use a similar construction to show that $\chi_{\text{td}}(G_n) = n$: the only difference is that v_0 gets label n rather than $n+1$. □

Theorem 5.3. *For $n \geq 4$, we have $\chi_{\text{td}}(H_n) = \chi_{\text{td}}(W_n)$ except when n is 6 or 7. In these cases, $\chi_{\text{td}}(H_6) = \chi_{\text{td}}(H_7) = 8$.*

We do not prove Theorem 5.3 in detail. By Proposition 2.11 it is clear that $\chi_{\text{td}}(H_n) \geq \chi_{\text{td}}(W_n)$. For $n \geq 8$, a construction that shows $\chi_{\text{td}}(H_n) \leq \chi_{\text{td}}(W_n)$ can be obtained from the total difference labelings for W_n described in Theorem 5.1 by determining labels for the leaves of H_n that avoid doubles and triples. In Figure 10, we show that $\chi_{\text{td}}(H_7) \leq 8$.

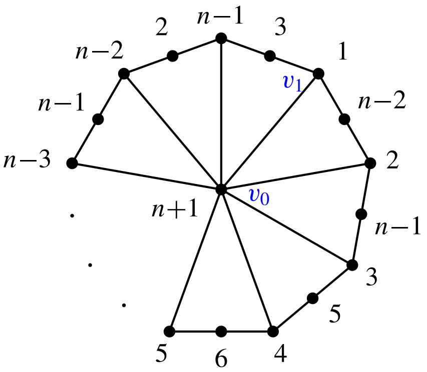


Figure 9. Vertices v_0 and v_1 (with labels n and 1, respectively) are named in the figure. Vertex v_2 has label $n-2$ and is adjacent to v_1, and $v_3, v_4, \ldots, v_{2n-2}$ continue clockwise. This construction shows $\chi_{\text{td}}(G_n) \le n+1$ when n is even and $n \ge 8$.

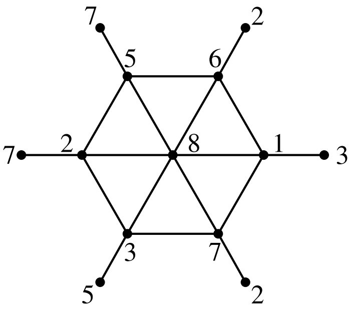


Figure 10. A construction that shows $\chi_{\text{td}}(H_7) \le 8$.

6. Trees

We first determine the total difference chromatic numbers of caterpillars. We then provide lower and upper bounds for the total difference chromatic numbers of lobsters, and end with an upper bound for the total difference chromatic numbers of general trees.

Definition 6.1. A *caterpillar* is a tree with the property that removal of its vertices of degree 1 and their incident edges leaves a path.

Consider the path formed by the removal of the degree-1 vertices in a given caterpillar. The degrees of the endpoints of this path in the caterpillar must be at least 2 (else they would not be part of the path as they would have been removed).

Definition 6.2. A caterpillar is a tree in which all vertices are at most distance 1 from a central path.

Theorem 6.3. *If G is a caterpillar with maximum degree Δ, then $\Delta+1 \le \chi_{\text{td}}(G) \le \Delta+3$.*

Proof. Let G be a caterpillar with maximum degree Δ. Notice that $K_{1,\Delta}$ is a subgraph of G so Proposition 2.11 and Theorem 4.1 imply $\chi_{\text{td}}(G) \ge \chi_{\text{td}}(K_{1,\Delta}) \ge \Delta+1$. If Δ is 1 or 2, then note that G is a path and we can refer to Theorem 3.1 to determine $\chi_{\text{td}}(G)$. Therefore, we assume $\Delta \ge 3$ throughout the proof.

If there are multiple options for the choice of central path in G, choose one arbitrarily and call it P. We denote the vertices on P by $v_1, v_2, \dots, v_n$ consecutively, so that $\deg(v_1) = \deg(v_n) = 1$ and $\deg(v_i) > 1$ for $i \ne 1, n$. We label v_i on P with 1 if $i \equiv 1 \pmod 3$, with $\Delta+3$ if $i \equiv 2 \pmod 3$, and with $\Delta+2$ if $i \equiv 0 \pmod 3$. We now consider labelings of the neighbors of an arbitrary vertex on P based upon its label.

Suppose the vertex v_i on P gets label 1 so that v_{i-1} and v_{i+1} (if they exist) have labels $\Delta+2$ and $\Delta+3$, respectively. Then, since the subgraph induced by v_i and its neighbors is a star $K_{1,d}$ for some $d \le \Delta$, the remaining vertices adjacent to v_i can be labeled with distinct integers from $[3, \Delta+1]$ without creating a double or triple.

Similarly, if v_i gets label $\Delta+3$, then the neighbors of v_i not on P (of which there are at most $\Delta-2$) can be labeled arbitrarily with distinct integers from $[2, \Delta+1]$, excluding $(\Delta+3)/2$ if Δ is odd. Likewise, if v_i gets label $\Delta+2$ we can label its neighbors not on P arbitrarily with distinct integers from $[2, \Delta]$, excluding $(\Delta+2)/2$ if Δ is even. (Note that we avoid a $(\Delta+3, \Delta+2, \Delta+1)$-triple by disallowing the use of $\Delta+1$ as a label.)

Hence, by looking at all possible constructions we have proven that $\Delta+1 \le \chi_{\text{td}}(G) \le \Delta+3$. □

In fact, we can classify caterpillars according to their total difference chromatic numbers, though the proofs are omitted here.

Theorem 6.4. *If G is a caterpillar, then $\chi_{\text{td}}(G) = \Delta+1$ if and only if*

(1) *Δ is even,*

(2) *the distance between vertices of degree Δ is at least* 3,

(3) *no three consecutive vertices have degrees at least $\Delta-1$, and*

(4) *there are no five consecutive vertices with first and last having degree Δ and second and fourth having degree $\Delta-1$.*

Theorem 6.5. *If G is a caterpillar, $\chi_{\text{td}}(G) = \Delta+3$ if and only if Δ is odd and there are at least three vertices with degree Δ in a row.*

Using Theorems 6.3, 6.4, and 6.5 we can determine the total difference chromatic number of any caterpillar. In particular, any caterpillar not satisfying the conditions set forth in Theorem 6.4 or 6.5 must have total difference chromatic number $\Delta+2$.

We now turn our attention to the trees known as lobsters, a natural extension of caterpillars.

$r\downarrow\ s\rightarrow$	2	3	4	5	6	7	8	9
1		11	12	13	14	15	12	7
10	9	11	12		14	15	12	
11	9	10	12	13	14	15	12	6

Table 1. The values of $m_{8,7}(r, s)$ for all $(r, s) \in R \times S$. The table entries for invalid (r, s) pairs are left blank.

Definition 6.6. A *lobster* is a tree in which all vertices are at most distance 2 from a central path.

In other words, removing the leaves of a lobster forms a caterpillar.

We call the vertices on the central path the *primary vertices*. Vertices adjacent to primary vertices which are not themselves primary are called *secondary vertices*, and leaves adjacent to secondary vertices are called *tertiary vertices*. We define Δ_1 to be the maximum degree of the primary vertices and Δ_2 to be the maximum degree of the secondary level vertices. A *maximal lobster* is one in which every primary vertex with degree at least 2 has degree Δ_1 and every secondary vertex has degree Δ_2. We will prove an upper bound on the total difference chromatic number of maximal lobsters and hence, by Proposition 2.11, all lobsters.

The first step in determining our upper bound for the total difference chromatic number of an arbitrary maximal lobster is to label the central path, P in the same way we labeled it for a caterpillar. Namely, we label v_i on P with 1 if $i \equiv 1 \pmod 3$, with $\Delta_1 + 3$ if $i \equiv 2 \pmod 3$, and with $\Delta_1 + 2$ if $i \equiv 0 \pmod 3$. Let $R = \{1, \Delta_1 + 2, \Delta_1 + 3\}$ By Theorem 6.3 we know that all of the secondary vertices can have labels from $[1, \Delta_1 + 3]$. To avoid triples involving two primary vertices and one secondary vertex, we restrict our possible secondary vertex labels to the set $S = [2, \Delta_1 + 1]$.

For each possible pair $(r, s) \in R \times S$ we consider the set of tertiary vertices adjacent to a secondary vertex with label s, which itself is adjacent to a primary vertex with label r. We greedily label this set of vertices and record the maximum label used. For given values of Δ_1 and Δ_2, we call this value $m_{\Delta_1,\Delta_2}(r, s)$. (See Table 1 for an example with $\Delta_1 = 8$ and $\Delta_2 = 7$.) Notice that $m_{\Delta_1,\Delta_2+1}(r, s) \geq m_{\Delta_1,\Delta_2}(r, s) + 1$.

Definition 6.7. We call the least value of Δ_2 at which all subsequent unit increases of Δ_2 cause unit increases of $m_{\Delta_1,\Delta_2}(r, s)$ the *stabilization point* of (r, s).

Lemma 6.8. *Let G be a lobster and let $r \in R = \{1, \Delta_1 + 2, \Delta_1 + 3\}$ and $s \in S = [2, \Delta_1 + 1]$.*

(a) *If $2 \leq s < (\Delta_1 + 4)/2$, then $m_{\Delta_1,\Delta_1+3-s}(r, s) = \Delta_1 + 4$, and the stabilization point of (r, s) is at most $\Delta_1 + 3 - s$.*

(b) *If* $(\Delta_1+4)/2 \le s \le \Delta_1+1$, *then* $m_{\Delta_1,s}(r,s) = 2s+1$, *and the stabilization point of* (r,s) *is at most* s.

Proof. We first prove part (b). Suppose that $(\Delta_1+4)/2 \le s \le \Delta_1+1$, and consider a secondary vertex, v, with label s and degree s adjacent to a primary vertex with label $r \in \{1, \Delta_1+2, \Delta_1+3\}$. We can label $s-1$ of the s vertices adjacent to v by using exactly one element from each set $\{i, 2s-i\}$ for $1 \le i \le 2s-1$. Notice that the primary vertex adjacent to v has a label from one of these sets, since 1, Δ_1+2, and Δ_1+3 are all less than $2s$. The remaining tertiary vertex adjacent to v cannot be labeled with any unused positive integer less than $2s$ since that would create a $(2s-i, s, i)$-triple, and labeling it with $2s$ would create a double. Hence, the least possible label for this one remaining vertex is $2s+1$, so $m_{\Delta_1,s}(r,s) = 2s+1$. Notice that tertiary vertices adjacent to v can be labeled with any integer $2s+k$ for $k>0$: these labels are larger than $2s$, and hence cannot form a double, and larger than $2s-1$ and hence cannot form a triple involving s. Therefore, the stabilization point of (r,s) is at most s.

We prove part (a) similarly to part (b). Again consider a secondary vertex v with label s and degree Δ_1+3-s for some s with $2 \le s < (\Delta_1+4)/2$. We can label $s-1$ of the vertices adjacent to v as before: by using exactly one element from each set $\{i, 2s-i\}$ for $1 \le i \le s-1$. We still must label $\Delta_1+3-s-(s-1) = \Delta_1+4-2s$ vertices adjacent to v, but this can be done using all integers in $[2s+1, \Delta_1+4]$ (if $r \ne 1$ then r is part of this set). Observe that $\Delta_2 = \Delta_1+3-s$ is greater than or equal to the stabilization point of (r,s) since if $\Delta_2 > \Delta_1+3-s$ we can label additional tertiary vertices with integers greater than $\max\{2s, r\}$. □

Corollary 6.9. *Assume* $\Delta_2 \ge \Delta_1+1$. *Then*

(a) $m_{\Delta_1,\Delta_2}(r,s) \ge m_{\Delta_1,\Delta_2}(r,s')$ *if* $s \ge s'$, *and*

(b) $m_{\Delta_1,\Delta_2}(r,s) = m_{\Delta_1,\Delta_2}(r',s)$ *for any* $r, r' \in R$,

when these values exist.

Proof. By Lemma 6.8, we know that $\Delta_2 = \Delta_1+1$ is greater than or equal to the stabilization points of all $(r,s) \in R \times S$, and hence we can compute each $m_{\Delta_1,\Delta_2}(r,s)$ exactly when $\Delta_2 \ge \Delta_1$. It is then straightforward to check that both parts of the claim hold. □

Theorem 6.10. *For any lobster* G *which is not a path or caterpillar*, $\Delta+1 \le \chi_{\mathrm{td}}(G) \le \Delta_1+\Delta_2+1$.

Proof. To prove that the claimed lower bound holds, we observe that a lobster contains as an induced subgraph the star $K_{1,\Delta}$ and then appeal to Proposition 2.11.

Consider a maximal lobster in which all primary vertices (except the two of degree 1) have degree Δ_1 and all secondary vertices have degree Δ_2. First assume

$\Delta_2 = \Delta_1 + 1$. By Lemma 6.8 and Corollary 6.9(a), when $\Delta_2 = \Delta_1 + 1$, the largest tertiary vertex label occurs when $s = \Delta_1 + 1$ and $r = 1, \Delta_1 + 3$ and is $2(\Delta_1 + 1) + 1 = \Delta_1 + \Delta_2 + 2$. By Lemma 6.8, all pairs (r, s) have reached their stabilization points. Therefore, when $\Delta_2 \geq \Delta_1 + 1$ and $k \geq 0$,

$$m_{\Delta_1,\Delta_2+k}(r, s) = m_{\Delta_1,\Delta_2}(r, s) + k,$$
$$m_{\Delta_1,\Delta_2-k}(r, s) \leq m_{\Delta_1,\Delta_2}(r, s) - k.$$

Hence for any value of Δ_2, the largest tertiary vertex label is at most $\Delta_1 + \Delta_2 + 2$.

We now show that we can decrease this upper bound by 1. For the primary vertices with degree greater than 1, we know, by our construction, the labels of the two adjacent primary vertices. Therefore we need only $\Delta_1 - 2$ distinct secondary vertex labels (i.e., values of s), and hence, for each $r \in \{1, \Delta_1 + 2, \Delta_1 + 3\}$ we can choose the least $\Delta_1 - 2$ values of $s \in [2, \Delta_1 + 1]$ that do not create any doubles or triples.

We check all possible pairs $(r, s) \in R \times S$ for doubles and triples. If $r = 1$, then s cannot be 2, or we would have a double. If $r = \Delta_1 + 2$, then s cannot be $\Delta_1 + 1$ otherwise we would have a $(\Delta_1 + 3, \Delta_1 + 2, \Delta_1 + 1)$-triple involving two primary vertices. Finally, depending on the parity of Δ_1, we must avoid either $s = (\Delta_1 + 2)/2$ (for $r = \Delta_1 + 2$) or $s = (\Delta_1 + 3)/2$ (for $r = \Delta_1 + 3$). See the empty boxes in Table 1 for (r, s) pairs that create doubles or triples in the case where Δ_1 is even.

For each value of r, we need only the least possible $\Delta_1 - 2$ values of s as secondary vertex labels. If $r = 1$, since the only forbidden value of s is 2, we can use all of the integers in $[3, \Delta_1]$. For one of the remaining two possible values of r we cannot have $s = r/2$. For this value of r, we use as our secondary labels all integers in $[2, \Delta_1]$ except $r/2$. For the other value of r, we can use all integers in $[2, \Delta_1 - 1]$. Notice that we need not label any of the secondary vertices in our maximal lobster with $\Delta_1 + 1$.

Hence, when $\Delta_2 = \Delta_1 + 1$, the largest relevant secondary vertex label is $s = \Delta_1$ and $m_{\Delta_1,\Delta_1+1}(r, \Delta_1) = 2\Delta_1 + 2 = \Delta_1 + \Delta_2 + 1$ by Lemma 6.8. As all pairs (r, s) have reached their stabilization points we know that for any value of Δ_2, and for any lobster, the largest tertiary vertex label is at most $\Delta_1 + \Delta_2 + 1$. □

We conclude by finding bounds for the total difference chromatic number of an arbitrary tree in terms of its maximum degree. We first define a *uniform full Δ-ary tree*, denoted by $T_{\Delta,h}$. For any integer $\Delta \geq 2$, the tree $T_{\Delta,1}$ is defined to be the star $K_{1,\Delta}$. We let v_0 be the vertex in $T_{\Delta,1}$ with maximal degree, and each other vertex is a leaf. For each integer $h \geq 2$ we define $T_{\Delta,h}$ to be the tree obtained from $T_{\Delta,h-1}$ by appending $\Delta - 1$ new leaves to each leaf of $T_{\Delta,h-1}$. Therefore, $T_{\Delta,h}$ is a rooted tree in which the distance from the root v to every other vertex is at most h.

It is maximal in the sense that every vertex whose distance from v is less than h has degree Δ, while those at distance h have degree 1.

Notice that every tree is a subgraph of $T_{\Delta,h}$ for some choice of Δ and h. Therefore, if we find an upper bound for $\chi_{\text{td}}(T_{\Delta,h})$ the same bound works for an arbitrary tree T with maximum degree Δ and appropriately chosen h.

Before providing such a bound, we determine the total difference chromatic number for uniform full Δ-ary trees with height 2.

Lemma 6.11. *For a uniform full Δ-ary tree with height* 2,

$$\chi_{\text{td}}(T_{\Delta,2}) = \left\lfloor \frac{3\Delta+3}{2} \right\rfloor.$$

Proof. The root v_0 of $T_{\Delta,2}$ is adjacent to Δ vertices, denoted by $v_1, v_2, \ldots, v_\Delta$, each of which is adjacent to a further Δ vertices (including v_0). First assume Δ is odd. By Lemma 4.3, for any $1 \le r \le \Delta$ with $r \ne (\Delta+3)/2$, there are exactly $2r-2$ possible labels for each v_i so that the maximum label on each $K_{1,\Delta}$ is exactly $\Delta + r$.

Notice that $v_0, v_1, \ldots, v_\Delta$ all must have different labels since they are most distance 2 apart. Therefore, we must choose r so that $2r-2 \ge \Delta+1$. In particular we must have $r \ge (\Delta+3)/2$. (In the exceptional $r = (\Delta+3)/2$ case, this inequality is still satisfied.)

If we label v_0 with $\Delta + r$, then each of these $\Delta+1$ copies of $K_{1,\Delta}$ will have maximum label $\Delta + r$ and therefore it is possible to give the entire $T_{\Delta,2}$ a $(\Delta+r)$-total difference labeling.

To minimize $\Delta+r$, we choose $r = (\Delta+3)/2$. By construction, we cannot choose a smaller value of r as there would not be enough distinct labels for $v_0, v_1, \ldots, v_\Delta$. Therefore, for odd Δ,

$$\chi_{\text{td}}(T_{\Delta,2}) = \Delta + \frac{\Delta+3}{2} = \frac{3\Delta+3}{2}.$$

If Δ is even, Lemma 4.3 implies that there are $2r-1$ options for the labels of $v_0, v_1, \ldots, v_\Delta$. Therefore $2r-1 \ge \Delta+1$, or, equivalently, $r \ge (\Delta+2)/2$. A similar argument to the one in the previous case implies that, for even Δ,

$$\chi_{\text{td}}(T_{\Delta,2}) = \Delta + \frac{\Delta+2}{2} = \frac{3\Delta+2}{2}.$$

Combining the two cases gives us the desired result. □

Theorem 6.12. *For any uniform full Δ-ary tree $T_{\Delta,h}$ with $h \ge 2$,*

$$\left\lfloor \frac{3\Delta+3}{2} \right\rfloor \le \chi_{\text{td}}(T_{\Delta,h}) \le 2\Delta+1.$$

Proof. The lower bound is a result of Lemma 6.11 and Proposition 2.11.

To prove the upper bound, first consider the subgraph T' induced by the root v_0 and its neighbors (which we denote $v_1, v_2, \dots, v_\Delta$). The labels on these vertices must all be distinct by Proposition 2.8. Choose arbitrarily a label $\ell \in [1, 2\Delta+1]$ for v_0. We show that for any choice of ℓ, at most Δ numbers in $[1, 2\Delta+1]$ cannot be labels of other vertices of T', which leave the remaining numbers, of which there are at least Δ, free to label the Δ neighbors of v_0.

First, suppose $\ell < \Delta+1$. Then, to avoid triples, we may use at most one element from each set $\{\ell-i, \ell+i\}$ for $1 \le i \le \ell-1$ (though, if ℓ is even we must not use the element $\ell/2$ from the set $\{\ell/2, 3\ell/2\}$ to avoid a double). Further, we may not use the number 2ℓ as a label. All other numbers are valid for labeling vertices of T' and, as we have only ruled out $\ell \le \Delta$ options, there are at least enough to label the remaining Δ vertices of T'. An analogous argument shows that if $\ell > \Delta+1$, we also have enough numbers to label all vertices of T'.

If $\ell = \Delta+1$ then, to avoid triples, we again may use at most one element from each set $\{\ell-i, \ell+i\}$ for $1 \le i \le \ell-1$ (and again, one of these includes $\ell/2$ if ℓ is even). In this case, however, $2\ell \notin [1, 2\Delta+1]$ so we still have enough numbers to label the remaining vertices of T'.

In each case, all vertices of T' can be labeled; suppose that vertex v_j gets label ℓ_j for $1 \le j \le \Delta$. By the same argument as above, there are enough available numbers in $[1, 2\Delta+1]$ to label all neighbors of each v_j. We can give these vertices labels and then repeat the process until the vertices of the entire tree $T_{\Delta,h}$ have been labeled with integers in $[1, 2\Delta+1]$. □

The following corollary follows directly from Theorem 6.12.

Corollary 6.13. *For any tree T, we have $\Delta+1 \le \chi_{\mathrm{td}}(T) \le 2\Delta+1$.*

References

[Bi et al. 2017] Z. Bi, A. Byers, S. English, E. Laforge, and P. Zhang, “Graceful colorings of graphs”, *J. Combin. Math. Combin. Comput.* **101** (2017), 101–119. MR Zbl

[Gallian 1998] J. A. Gallian, “A dynamic survey of graph labeling”, *Electron. J. Combin.* **5** (1998), art. id. 6. MR Zbl

[Golomb 1972] S. W. Golomb, “How to number a graph”, pp. 23–37 in *Graph theory and computing* (Kingston, Jamaica, 1969), edited by R. C. Read, Academic Press, New York, 1972. MR Zbl

[Rosa 1967] A. Rosa, “On certain valuations of the vertices of a graph”, pp. 349–355 in *Theory of graphs* (Rome, 1966), edited by P. Rosenstiehl, Gordon and Breach, New York, 1967. MR Zbl

Received: 2019-12-23 Revised: 2020-03-12 Accepted: 2020-04-28

rrohatgi@saintmarys.edu
Department of Mathematics and Computer Science,
Saint Mary's College, Notre Dame, IN, United States

yzhang01@saintmarys.edu
Department of Mathematics and Computer Science,
Saint Mary's College, Notre Dame, IN, United States

INVOLVE 13:3 (2020)
https://doi.org/10.2140/involve.2020.13.529

Decline of pollinators and attractiveness of the plants

Leila Alickovic, Chang-Hee Bae,
William Mai, Jan Rychtář and Dewey Taylor

(Communicated by Suzanne Lenhart)

Flowering plants rely mostly on pollinators to reproduce. A decline of pollinators puts an evolutionary pressure on the allocation of plants' resources towards attracting the few remaining pollinators. This may result in fewer resources available for the plants' survival and actual seed production. Moreover, due to the "magnet effect", attractive plants generally attract pollinators to all plants in their neighborhood, even the less attractive ones. To better understand the allocation trade-offs, we built a computer simulation and studied the evolution of resource allocation towards attracting pollinators. We observed that when pollinators are relatively abundant, there is not much incentive for the plants to allocate more energy to attract them. Only when pollinators are below a certain critical threshold is a relatively large investment in attracting the pollinators suddenly favored. The value of the critical threshold is quite low and further decreases with the increasing seed dispersal distance and the plant population size.

1. Introduction

Flowering plants, including plants that yield crops, are pollinated mainly by insects [Williams 2002]. Attractiveness of a plant's flowers is a major factor in the plant's reproductive success [Kunin and Iwasa 1996]. Flowers of different species may vary vastly in color, pattern, and size but they all retain a similar structure in order to appeal to the available insect pollinators [Endress 1994].

Pollinators thrive and depend on a nectar produced by the flowers [Gardener and Gillman 2002]. The nectar serves as an incentive to pick one plant over another [Faegri and Van Der Pijl 1979]. When a pollinator visits a flower to harvest its nectar, it helps the plant to reproduce: it deposits pollen collected earlier from other

MSC2010: 91A22, 92B05.
Keywords: plant-pollinator interactions, plant-pollinator mutualism, magnet effect, resource allocation.

plants and it collects the plant's own pollen to be deposited in other flowers in the future [Johri 1984].

There is competition between plants for the same pollinators [Levin and Anderson 1970]. When the pollinators are rare, the plants tend to produce more and larger flowers and they allocate more resources to attract the pollinators [Ågren et al. 2013]. At the same time, when a plant is attractive, it increases the amount of pollinators in its immediate neighborhood [Faegri and Van Der Pijl 1979]. As a result, less attractive plants within a short distance of an attractive plant will also be visited and pollinated [Laverty 1992]. This "magnet effect" is observed across many species [Torices et al. 2018].

We study how the abundance of pollinators affects the evolution of plants' resource allocation. In Section 2, we describe a computer simulation to quantify the costs and benefits of allocating more resources to attract pollinators. The simulation follows a field of native plants that is invaded by a small number of potentially more attractive plants. We are interested in finding out the best allocation strategy for the invading plants. The results are presented in Section 3. We show that there is a critical threshold: when the number of pollinators is above the threshold, there is only a very small tendency to increase the attractiveness; when the number is below the threshold, a very large allocation to attract the pollinators is favored. The threshold is quite sharp, low and decreasing with the plant field size and seed dispersal distance. We conclude the paper in Section 4.

2. Methods: computer simulation

The basic structures and algorithm of the simulation are described below; see also Figure 2. The actual simulation was written in Matlab and the computations were done on the "teal" cluster in VCU's Center for High Performance Computing.

2.1. ***Plants.*** The plants are assumed to be annual. They grow from a seed, bloom, produce seeds, and die in one growing season. The plants can grow only from the seeds that were produced during the last season.

2.2. ***Plant field.*** The plants are assumed to grow in patches. The patches are arranged in an $N \times N$ square grid with periodic boundaries; see Figure 1. All patches on the field are identical and differ only in the number (and type) of plants that grow on them. Each patch can sustain up to C plants of any type. For simplicity, we assumed that each patch is either empty or consists of C plants.

2.3. ***Plant's attractiveness and seed production.*** We assume all plants have the same amount of resources (1 unit without loss of generality). The resources can be allocated to blooming and attracting pollinators, seed production, and survival. A plant's strategy is given by $a \in [0, 1]$, a proportion of resources allocated towards blooming and attracting the pollinators. When a plant using a strategy a is pollinated, it produces $S(a) = a(1-a)$ units of seeds. Plants growing from those seeds will have exactly the same allocation strategy a.

The function $S(a)$ was chosen as it is a simple function and the plants will not produce many seeds if either they do not allocate enough resources to blooming and attracting the pollinators or they allocate too much so that they will not have enough resources left for the actual seed production and their own survival.

As an artifact of choosing this function $S(a) = a(1-a)$, the pollinated plants achieve maximum seed production using strategy $a = 0.5$. However, as seen later, it does not necessarily mean that the strategy $a = 0.5$ is always the best strategy to use.

2.4. ***Patch attractiveness.*** We assume that the pollinators are attracted towards patches of plants rather than individual plants. The more attractive the patch is, the more likely it is to attract the pollinators. If a given patch P consists of $S_{P,\text{nat}}$ plants each allocating a_{nat} and $S_{P,\text{inv}}$ plants each allocating a_{inv} towards attractiveness, the attractiveness of a patch will be given by

$$A_P = S_{P,\text{nat}}\, a_{\text{nat}} + S_{P,\text{inv}}\, a_{\text{inv}} + r_P, \tag{1}$$

where $r_P \in (0, 0.001)$ is a small random number that is different for different patches and also changing every season. The purpose of r_P is to add a natural variation among patches occupied by plants using exactly the same strategies.

2.5. ***Pollinators.*** We assume that there is only one pollinator species. The abundance of the pollinators is represented by $p \in [0, 1]$. We assume that only pN^2 patches can be pollinated during one season. The pollinators will prefer patches with higher attractiveness. This is achieved by sorting all patches by their attractiveness and assuming that only the top pN^2 patches are visited. Once a patch is visited, all the plants on the patch are assumed to be pollinated regardless of their individual allocation strategies. This assumption allows us to account for the magnet effect.

2.6. ***Seed dispersal.*** Once a plant is pollinated, it produces seeds. The seeds then disperse over the field. We assume the seeds can disperse up to d patches away from their patch of origin. For simplicity, we assume that the seeds are equally likely to travel any distance $0, 1, \dots, d$ and move in any direction. This means that the probability of the seed staying at the origin is $1/(d+1)$. There are 8δ patches that are exactly $\delta \in \{1, 2, \dots, d\}$ away from the patch of the origin and so the probability that a seed will end up on a patch δ away is given by

$$\pi(\delta) = \begin{cases} (d+1)^{-1} & \text{if } \delta = 0, \\ (8\delta(d+1))^{-1} & \text{if } \delta \in \{1, 2, \dots, d\}, \\ 0 & \text{if } \delta > d; \end{cases} \tag{2}$$

see Figure 1.

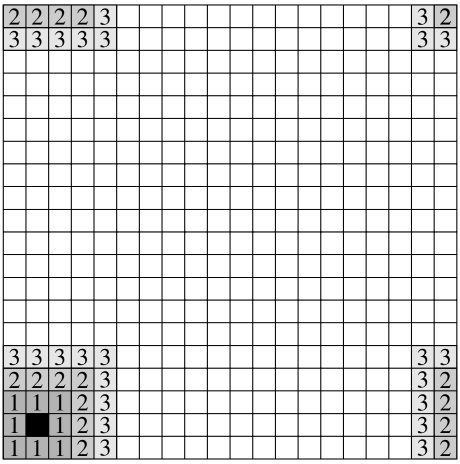


Figure 1. Visualization of the field and seed dispersal for $N = 20$ and $d = 3$. Each square represents a patch. In this case, the seeds originate from the black square. A fourth of all seeds produced there will stay there. The squares with the same shade of gray (and the same label representing its distance from the black square) will get the same number of seeds. In the sequence of increasing distance, there will be $\frac{1}{32}$, $\frac{1}{64}$ and $\frac{1}{96}$ of all seeds produced on the black patch. The boundaries are periodic.

Let $\rho(P, P_0)$ denote the distance between patches P and P_0 and let S_P be the number of seeds produced on patch P in a given season. Then, at the end of the season, after the seed dispersal, a patch P_0 will have

$$S'_{P_0} = \sum_{\delta \geq 0} \pi(\delta) \sum_{P, \rho(P, P_0)=\delta} S_P \tag{3}$$

seeds.

We assume the plants produce a large amount of seeds. With this assumption, if there is at least one seed on a patch, there are many more than C seeds on the patch. Since a patch can support only C plants, we renormalize: if on a given patch P_0 there is at least one seed and $S'_{P_0,\,\mathrm{nat}}$ seeds come from native plants and $S'_{P_0,\,\mathrm{inv}}$ seeds come from invasive plants, we will assume that there will be $C(S_{P'_0,\mathrm{nat}}/(S'_{P_0,\mathrm{nat}} + S'_{P_0,\mathrm{inv}}))$ seeds of native plants and $C(S_{P'_0,\mathrm{inv}}/(S'_{P_0,\mathrm{nat}} + S'_{P_0,\mathrm{inv}}))$ seeds of invasive plants that will germinate and grow into plants in the next season.

2.7. ***Summary of the simulation of one season.*** We will consider only a situation with two types of plants: native and invasive. Quantities related to native plants will have a subscript "nat", quantities related to invasive plants will have subscript "inv".

At the beginning of a season, each of the $N \times N$ patches P contains $S_{P,\text{nat}}$ ($S_{P,\text{inv}}$) seeds that will germinate and grow into $S_{P,\text{nat}}$ ($S_{P,\text{inv}}$) plants allocating a_{nat} (a_{inv}) towards attracting the pollinators where $0 \le S_{P,\text{nat}}$, $0 \le S_{P,\text{inv}}$, and $S_{P,\text{nat}} + S_{P,\text{inv}} \le C$.

The attractiveness of the patch P_0 is given by $A_{P_0} = S_{P_0,\text{nat}}\, a_{\text{nat}} + S_{P_0,\text{inv}}\, a_{\text{inv}} + r_{P_0}$, where $r_{P_0} \in (0, 0.001)$ is a small random number.

All plants on the patch P_0 are pollinated if and only if A_{P_0} is in the top pN^2 values of $\{A_P\}_{P=1}^{N^2}$. Only the plants that are pollinated produce seeds; new plants growing from those seeds during the next season will have exactly the same allocation strategy as the parent plant. Plants that are not pollinated do not produce any seeds.

The seeds disperse through the field. At the end of the season, all existing plants die and only the seeds remain on the patch to grow in the next season. We renormalize the seed count to keep the number of seeds, and consequently the plant count on each patch either 0 or C. If there was any seed on the patch at the end of the season, there will be C seeds after the at the end of the next season. We can consider this procedure to model the fact that out of possibly many seeds on the patch only some will survive to the next season and germinate and the patch itself can support at most C plants.

2.8. ***Overall simulation.*** The parameters of the simulation are as follows:

(1) The size, N, of the field. The field will have N^2 patches; each patch will have either 0 or C plants.

(2) The abundance, p, of the pollinators. Only pN^2 patches (the most attractive ones) will be visited by the pollinators. All plants on the visited patches will be pollinated.

(3) The allocation strategy of the native (a_{nat}) and invasive (a_{inv}) plants. A plant with strategy a will produce $S(a) = a(1-a)$ seeds if pollinated.

(4) The maximal distance, d, the seeds can disperse. The field is assumed to have periodic boundaries; see Figure 1.

For every set of the parameters specified above, we initialize the field by letting every patch be occupied by only native plants (using strategy a_{nat}). We let the simulation run for ten seasons to allow the population of plants to stabilize at the appropriate level. The number of occupied patches correlates with the pollinators' abundance p; see Figure 3.

At the end of the tenth season, we introduce invasive plants into one of the patches (we assume all plants on that patch now use strategy a_{inv}). We then let the simulation run for 100 seasons. At the end of the 100th season, we record the number of invasive and native plants in the field; see Figure 2. If there are no native plants left, we say that the invasive plants replaced the native plants.

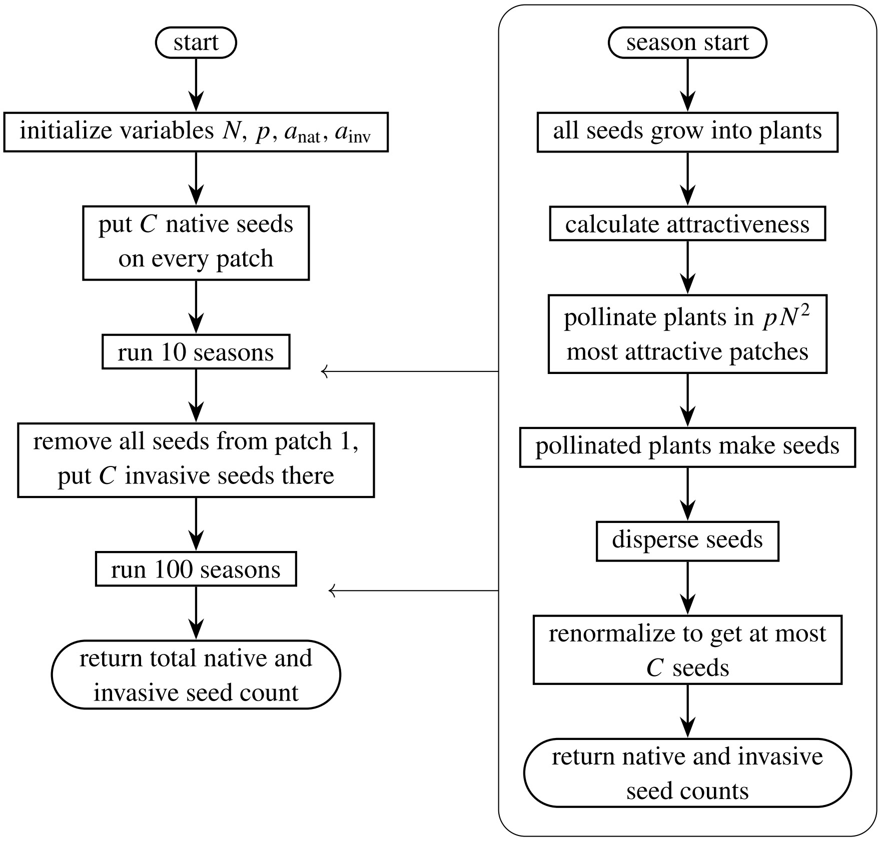

start
initialize variables N, p, a_{nat}, a_{inv}
put C native seeds on every patch
run 10 seasons
remove all seeds from patch 1, put C invasive seeds there
run 100 seasons
return total native and invasive seed count
season start
all seeds grow into plants
calculate attractiveness
pollinate plants in pN^2 most attractive patches
pollinated plants make seeds
disperse seeds
renormalize to get at most C seeds
return native and invasive seed counts

Figure 2. Scheme of the simulation described in more detail in Section 2.8.

We repeat the simulation 1000 times with the same parameters and record the probability of replacement and the average abundance of the invasive plants at the end of the 100th season.

2.9. ***Experiments.*** We are interested in what is the best invading strategy, i.e., a strategy that gives the highest chance of replacing the native plant population. If there are two (or more such values), we restrict our attention only to those values and the one yielding the largest final abundance is said to be the best. We run the following two experiments: we fixed $d = 1$ and used $N \in \{20, 30, 40, 50\}$ and we fixed $N = 20$ and used $d \in \{1, 2, 3, 4, 5\}$. For each such a pair N and d we used p between 0.01 and 0.4 in increments of 0.01. For given N, d and p, we run the simulation for all values a_{nat} from 0.5 to 1 (in increments of 0.01) and different values a_{inv} from a_{nat} to 1 (in increments 0.01).

To save computational time, we did not use $p > 0.4$ because our initial testing did not reveal any significant difference in the results between $p = 0.4$ and $p > 0.4$.

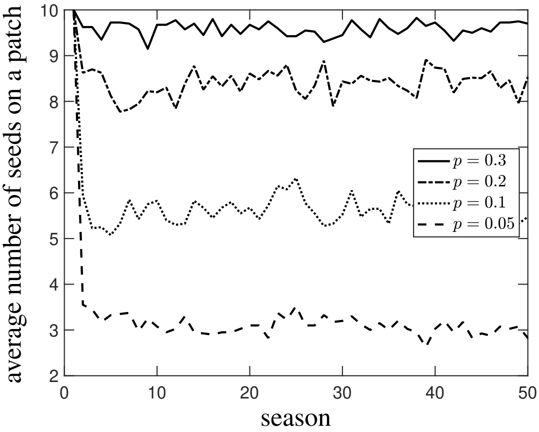


Figure 3. The time evolution of the seed count of native plants ($a_{\text{nat}} = 0.5$, $N = 20, d = 1$). The average seed count stabilizes within few seasons, we consider it stable at season 10.

We also did not consider $a_{\text{nat}} < 0.5$ because such native plants would be always replaced by invasive plants using $a_{\text{inv}} = 0.5$. Indeed, if $a_{\text{nat}} < 0.5$ and $a_{\text{inv}} = 0.5$, then patches with invasive plants will be more attractive than patches with native plants only. Moreover, the invasive plants will produce more seeds than the native plants. Finally, we did not use $a_{\text{inv}} < a_{\text{nat}}$, because such invasive plants never stay in the population for too long (their initial patch is the least attractive patch in the field so it very rarely gets chosen by the pollinators).

For each of the above combinations of $N, d, p, a_{\text{nat}}, a_{\text{inv}}$ we ran the simulation 1000 times. Each time, we recorded the total number of native and invasive plants at the end of the 100th season. We calculated $\text{Repl}(N, d, p, a_{\text{nat}}, a_{\text{inv}})$, the probability of replacement as the proportion of simulations that ended with some invasive but no native plants. We then identified a set

$$A_{\text{best}}(N, d, p, a_{\text{nat}}) = \text{argmax}_{a_{\text{inv}}}(\text{Repl}(N, d, p, a_{\text{nat}}, a_{\text{inv}})) \tag{4}$$

of values that yielded the highest replacement probability. The best invasive strategy, $a_{\text{best}} = a_{\text{best}}(N, d, p, a_{\text{nat}})$, was determined as a strategy in $A_{\text{best}}(N, d, p, a_{\text{nat}})$ that yielded the maximal final abundance of invasive plants.

3. Results

We identified $a_{\text{best}} = a_{\text{best}}(N, d, p, a_{\text{nat}})$, the best invading strategy. We found that there is a critical value $p_{\text{crit}} = p_{\text{crit}}(N, d)$ and constants $M_{N,d}$ and $B_{N,d}$ such that

$$a_{\text{best}}(N, d, p, a_{\text{nat}}) \approx \begin{cases} a_{\text{nat}} + \Delta & \text{if } p \geq p_{\text{crit}}(N, d), \\ M_{N,d}a_{\text{nat}} + B_{N,d} & \text{if } p < p_{\text{crit}}(N, d), \end{cases} \tag{5}$$

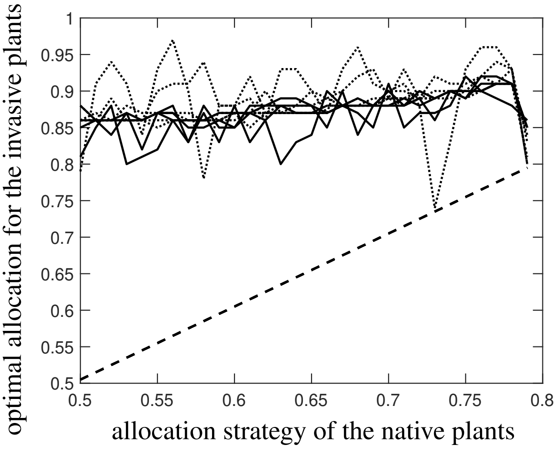


Figure 4. The optimal allocation strategy $a_{\text{best}}(N, d, p, a_{\text{nat}})$ as a function of the allocation of the native plants a_{nat}. Full lines are for $p \leq 0.1$, dotted lines for p between 0.11 and 0.18, dashed line is for $p \geq 0.19$. In this figure, $N = 20,\ d = 1$.

where $\Delta = 0.01$ is the increment in our values of a_{inv}. As seen in Figure 4, regardless of the value of p, $a_{\text{best}}(N, d, p, a_{\text{nat}})$ exhibits a linear trend in a_{nat}. For $p \geq p_{\text{crit}}(N, d)$, the best strategy follows the line $a_{\text{nat}}+\Delta$ exactly. For $p < p_{\text{crit}}(N, d)$, the slope of the linear trend is around 0.13 and the outcome is much noisier. While the coefficients $M_{N,d}$ and $B_{N,d}$ varied in N and d ($M_{N,d}$ was generally slightly decreasing in N and d; $B_{N,d}$ was generally slightly increasing in N and d), the variation could be contributed to random fluctuation and was not substantial. Overall, $M_{N,d} \approx 0.13$ and $B_{N,d} \approx 0.79$ are good approximations (with R^2 around 0.8). Since the significance of formula (5) is the existence of a critical value of p_{crit} and the two distinct behaviors of a_{best}, we did not investigate the best values of the coefficients $M_{N,d}$ and $B_{N,d}$ in greater detail.

When the pollinators are relatively abundant ($p \geq p_{\text{crit}}$), there is very little incentive for the plants to allocate too much energy towards attracting them. If the invasive plants attract more pollinators, the patches with them will be pollinated with very high probability. If the invasive plants allocate a'_{inv} instead of a_{inv} (for $a'_{\text{inv}} > a_{\text{inv}} > a_{\text{nat}}$), the increase in attractiveness will not increase the probability of their patches to be visited. However, many of these patches will contain native plants as well. The native plants will be pollinated too and will, in fact, produce more seeds than the invasive plants (because for $0.5 < a_{\text{nat}} < a_{\text{inv}} < a'_{\text{inv}}$ and $S(a_{\text{nat}}) > S(a_{\text{inv}}) > S(a'_{\text{inv}})$). This means that in the next generation, there will again be a lot of patches with a mix of native and invasive plants. Moreover, if the invasive plants allocate a'_{inv} instead of a_{inv} (for $a'_{\text{inv}} > a_{\text{inv}} > a_{\text{nat}}$), there will be even fewer invasive plants in the next generation (because $S(a_{\text{nat}}) > S(a_{\text{inv}}) > S(a'_{\text{inv}})$).

d	$p_{\text{crit}}(20, d)$
1	0.19
2	0.14
3	0.12
4	0.10
5	0.07

Table 1. The critical value p_{crit} as a function of the seed displacement distance d. Here, $N = 20$.

Consequently, the invasive plants should allocate as little as possible above the native plants' allocation.

When the pollinators are rare ($p < p_{\text{crit}}$), the pollinators are visiting only very few patches. This provides a strong incentive for the invasive plants to invest many more resources to attracting the pollinators because then only the patches with no (or very few) native species get visited.

What makes our result surprising is that there is a relatively sharp cut off point between the two kinds of allocation strategies; see Figure 4.

The critical value $p_{\text{crit}}(N, d)$ is decreasing in the seed dispersal distance d; see Table 1. As d increases (and p stays constant), the number of patches containing only invasive plants decreases and many patches contain a mix of invasive and native plants. If the plants on a mixed patch are all pollinated, a native plant will produce more seeds than an invasive plant. Consequently, the proportion of invasive plants can be decreasing in time. However, for a fixed d, as p decreases, only the patches with decreasing proportion of native plants get pollinated, which means that it is actually the proportion of native patches that is decreasing in time.

The critical value $p_{\text{crit}}(N, d)$ also decreases with the increasing field size, N; see Table 2. We do not fully understand the reason for this decrease. However, we believe that the critical value decreases to $p = 1/(d+1)^2$. For such p and large N, pollinating $N^2/(d+1)^2$ patches means that in the next generation, almost all patches will be occupied again. More simulation trials are needed to confirm or reject this hypothesis.

field size N	$p_{\text{crit}}(N, 1)$
20	0.19
30	0.16
40	0.14
50	0.12

Table 2. The critical value p_{crit} as a function of N. Here $d = 1$.

4. Conclusions

The lives of pollinators and plants are interconnected in many ways [Corbet 1996]. A decline of pollinators puts many plant species at a serious risk of extinction [Lernnartsson 2002] and under pressure to evolve strategies that will help them deal with the decline.

We built a computer simulation to study the effect of pollinators' decline on the plants' resource allocation. We saw that plants with more attractive flowers are able to replace plants with less attractive flowers; however, we saw this happening only when pollinators were truly very rare. When the pollinators' numbers are above this relatively sharp threshold, the incentive to evolve more attractive flowers is still there but it seems to be kept in check by the associated costs, such as being left with fewer resources towards seed production and survival. This finding seems to be in line with results of previous studies such as [Rathcke 1983; Kunin and Iwasa 1996].

We note that evolving more attractive flowers is not the only option. Plants can evolve a decreased dependence on the pollinators by allowing their flowers to be pollinated by wind and other means [Culley et al. 2002].

Acknowledgements

Alickovic, Bae and Mai worked on this manuscript as part of the course MATH/BIOL 380: Introduction to mathematical biology. They wish to acknowledge the help and support of their classmates, and particularly the help of S. Lee.

References

[Ågren et al. 2013] J. Ågren, F. Hellström, P. Toräng, and J. Ehrlén, "Mutualists and antagonists drive among-population variation in selection and evolution of floral display in a perennial herb", *Proc. Nat. Acad. Sci.* **110**:45 (2013), 18202–18207.

[Corbet 1996] S. A. Corbet, "Role of pollinators in species preservation, conservation, ecosystem stability and genetic diversity", pp. 219–230 in *International Symposium on Pollination, VII* (Lethbridge, AB, 1996), edited by K. W. Richards and I. H. Williams, Acta Horticult. **437**, Int. Soc. Horticult. Sci., Leuven, Belgium, 1996.

[Culley et al. 2002] T. M. Culley, S. G. Weller, and A. K. Sakai, "The evolution of wind pollination in angiosperms", *Trends Ecol. Evol.* **17**:8 (2002), 361–369. Correction in **17**:10 (2002), 491.

[Endress 1994] P. K. Endress, "Floral structure and evolution of primitive angiosperms: recent advances", *Plant Systemat. Evol.* **192**:1-2 (1994), 79–97.

[Faegri and Van Der Pijl 1979] K. Faegri and L. Van Der Pijl, *Principles of pollination ecology*, Elsevier, 1979.

[Gardener and Gillman 2002] M. C. Gardener and M. P. Gillman, "The taste of nectar: a neglected area of pollination ecology", *Oikos* **98**:3 (2002), 552–557.

[Johri 1984] B. M. Johri (editor), *Embryology of angiosperms*, Springer, 1984.

[Kunin and Iwasa 1996] W. Kunin and Y. Iwasa, "Pollinator foraging strategies in mixed floral arrays: density effects and floral constancy", *Theoret. Popul. Biol.* **49**:2 (1996), 232–263. Zbl

[Laverty 1992] T. M. Laverty, “Plant interactions for pollinator visits: a test of the magnet species effect”, *Oecologia* **89**:4 (1992), 502–508.

[Lernnartsson 2002] T. Lernnartsson, “Extinction thresholds and disrupted plant-pollinator interactions in fragmented plant populations”, *Ecology* **83**:11 (2002), 3060–3072.

[Levin and Anderson 1970] D. A. Levin and W. W. Anderson, “Competition for pollinators between simultaneously flowering species”, *Amer. Naturalist* **104**:939 (1970), 455–467.

[Rathcke 1983] B. Rathcke, “Competition and facilitation among plants for pollination”, pp. 305–330 in *Pollination biology*, edited by L. Real, Academic Press, Orlando, FL, 1983.

[Torices et al. 2018] R. Torices, J. M. Gómez, and J. R. Pannell, “Kin discrimination allows plants to modify investment towards pollinator attraction”, *Nature Commun.* **9** (2018), art. id. 2018.

[Williams 2002] I. H. Williams, “Insect pollination and crop production: a European perspective”, pp. 59–65 in *The conservation link between agriculture and nature* (Sao Paulo, 1998), edited by P. G. Kevan and V. L. Imperatriz-Fonseca, Minist. Environ., Brasilia, 2002.

Received: 2020-01-16 Revised: 2020-04-02 Accepted: 2020-04-28

alickovicl@vcu.edu
Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA, United States

baech@vcu.edu
Department of Biology, Virginia Commonwealth University, Richmond, VA, United States

maiaw@vcu.edu
Department of Biology, Virginia Commonwealth University, Richmond, VA, United States

rychtarj@vcu.edu
Department of Mathematics and Applied Mathematics, Virginia Commonwealth University, Richmond, VA, United States

dttaylor2@vcu.edu
Department of Mathematics and Applied Mathematics, Virginia Commonwealth University, Richmond, VA, United States

mathematical sciences publishers msp

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

2020 vol. 13 no. 3

1944-4176(2020)13:3;1-R