Online at `https://digitalresearch.bsu.edu/mathexchange`

ISSN 1550-1736

Cover design by Patrick Foley.

A Word from the Editor

The editorial board of the *Mathematics Exchange* is pleased to present this latest issue, a collection of nine enjoyable articles addressing a range of mathematical topics of interest to a broad audience at the undergraduate level. We appreciate the authors' efforts to disseminate their new discoveries as well as appreciate how they inspire and motivate our readership to follow their example in sharing their love for mathematics. And we hope you enjoy the fruits of their labor. We believe that getting students involved in publishing mathematics is a true milestone in helping them find their (permanent) place in the mathematical community and we are thrilled to be a part of that endeavor.

The first article gives a detailed treatment of generalized trigonometric functions defined on the unit $p-$ circle $|x|^p + |y|^p = 1$. It covers the existing foundational material well, leads ideas in some interesting directions, provides a variety of new results, and lists some open questions for future work.

The second article considers quantities associated to a graph G called the "$k-$diameter component vertex connectivity parameter" and "$k-$diameter component connectivity function", denoted respectively as $CV_k(G)$ and $CM_k(G;p)$, and computes their values for several families of graphs. The quantity $CV_k(G)$ is the minimal number of vertices that needs to be removed from G so that no component of the remaining graph has diameter $\geq k$. The mixed parameter $CM_k(G;p)$ is similar but considers the minimal number of edges that need to be removed allowing for any p vertices to first be removed from G. The article is well written, and it should be quite accessible to undergraduates with some modest exposure to graph theory.

The third article provides a good example of mathematical modeling. It takes a standard stability analysis of an ordinary differential equations model approach to determine if a zombie apolocalypse is theoretically possible with realistic parameter values taken from empirical sources. The mathematics it presents is appropriate and reasonable and the application is fun.

It is well known that the sum of the entries along a slope-2 diagonal through Pascal's triangle is a Fibonacci number. For integers $h \leq 2$, the fourth article considers the sequence $d_h(n)$ of sums along slope-h diagonals, derives a recurrence and generating function for $d_h(n)$, and uses the generating function to obtain an approximation to $d_h(n)$.

A known characterization for entire functions that preserve all nonnegative matrices of order two is shown to characterize polynomials that preserve nonnegative matrices of order two. The fifth article gives a new characterization for polynomials that preserve nonnegative circulant matrices of order two.

Is it possible to dilate (or shrink) each side of a polygon $P \subset \mathbb{R}^2$ by a factor of positive t to get a new polygon P' while preserving the unit normal vectors to the edges? The sixth article draws a connection with Viviani's theorem and equiangular polygons, and uses the Minkowski's existence and uniqueness theorem for polytopes to show that such P' exists if and only if P satisfies the constant Viviani sum.

The classical Lotka-Volterra equation models the interaction between two species competing for limited resources; the seventh article explores its extension in which a general nonlinear relationship models the effects of each species on the other.

The eighth article gives a brief exposition of two different well-known metrics on the Heisenberg group, an extremely well-studied object in analysis, proves that there exist minimal geodesics for the Koranyi metric as a consequence of the Arzelà-Ascoli theorem, and shows that lengths of (horizontal) curves are the same when computed in either metric.

The final article investigates the numerical range of 3×3 matrices over finite fields, particularly when the matrix is strictly triangular. The article is both novel and deep. The reviewer rated this article highly, believing that it will have an impact on the field of research and be cited by future publications.

We hope that you will enjoy reading this issue of the *Mathematics Exchange*. As always, we welcome and encourage ideas on how we can better serve our readers.

Yayuan Xiao

10.25.2022

Call for Papers

We are always soliciting contributions for future issues of this journal. Contributions are accepted from all undergraduate students who have worked on a project beyond the classroom in any mathematical area (e.g., pure, applied, actuarial, and education). Appropriate papers from other departments and other institutions are also welcome. Often the articles are written by undergraduates individually, working in teams, or working with faculty. On occasion we also include articles written solely by faculty or graduate students as long as they are accessible to undergraduates.

To submit an article, please select ONE member from the editorial board, and forward your material in PDF form, usually prepared by LaTeX (preferred) or Microsoft Word, to the editor you selected. We use double anonymized peer review, the identities of both reviewers and authors are concealed from each other throughout the review. To facilitate this, please remove any identifying information, such as authors' names or affiliations, from your manuscript before submission. Please ensure that the title page (that include all authors' names and affiliations, a complete address of the corresponding author including an email address, acknowledgements, and conflict of interest statement) is present in your submission as a separate file. If authors are undergraduate students, please include your advisor's name and contact information in the title page. Review and selection of articles is handled by the editorial committee. Editorial changes of accepted articles are communicated through students' advisors, when appropriate.

More information, including links to all previous issues, are available online at
`https://digitalresearch.bsu.edu/mathexchange`.

Contents

Ball State Undergraduate Mathematics Exchange
`https://digitalresearch.bsu.edu/mathexchange`
Vol. 16, No. 1 (Fall 2022)
Pages 2 – 22

Trigonometric Functions in the p-norm

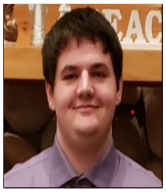
Sunil Chebolu , Andrew Hatfield, Riley Klette, Christopher Moore, Elizabeth Warden*

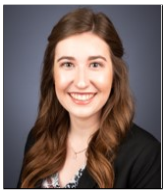
Sunil Chebolu received his Ph.D. from the University of Washington in 2005. He is a professor at Illinois State University and his research interests lie in algebra and number theory. In his spare time, he enjoys playing his guitar or observing deep-sky objects through his telescope.

Andrew Hatfield worked on this paper as an undergraduate studying mathematics at Illinois State University. He is currently at Illinois State working towards his master's degree and wishes to attend a Ph.D. program, where he would like to research algebraic geometry.

Riley Klette is a junior at Illinois State University majoring in Secondary Education in Mathematics. After graduation, she plans on teaching in a high school mathematics classroom.

Christopher Moore is a third-year student at Illinois State University majoring in Computer Science and minoring in Mathematics. After graduating, he plans to go into the field of software engineering and further his education by getting a master's degree.





Elizabeth Warden is a senior at Illinois State University, where she is majoring in secondary mathematics education. As a teacher, she hopes to utilize her research experience to inspire her students to think like mathematicians and explore unsolved problems.

***Corresponding author**: schebol@ilstu.edu

Abstract

Trigonometry is the study of circular functions, which are functions defined on the unit circle $x^2+y^2=1$, where distances are measured using the Euclidean norm. When distances are measured using the L_p-norm, we get generalized trigonometric functions. These are parametrizations of the unit p-circle $|x|^p+|y|^p=1$. Investigating these new functions leads to interesting connections involving double angle formulas, norms induced by inner products, Stirling numbers, Bell polynomials, Lagrange inversion, gamma functions, and generalized π values.

1 Introduction

It is a well-known fact that trigonometric functions are periodic: if $f(x)$ is any trigonometric function, then $f(x+2\pi)=f(x)$ for all values of x in the domain of f. Therefore, it is natural to define trigonometric functions on the unit circle, where all multiples of 2π are identified when we wrap the real line onto the circle. Because of this definition, trigonometric functions are also called circular functions. In this setting, the trigonometric functions $\sin t$ and $\cos t$ are just the unit circle's parametrization with respect to arc length.

Recall that the unit circle is the locus of all points in the plane $\mathbb{R}^2$ that are at a distance of one unit from the origin, where distances are measured using the standard Euclidean norm: $\|\vec{x}\|=(x_1^2+x_2^2)^{1/2}$. What if we switch to an L_p-norm: $\|\vec{x}\|_p=(|x_1|^p+|x_2|^p)^{1/p}$, $(p\geq 1)$? We then get a new family of curves defined by the equations $|x|^p+|y|^p=1$. These are called unit p-circles and are shown in the figure below. Because these curves are in between a square and a circle, they are also called squircles.

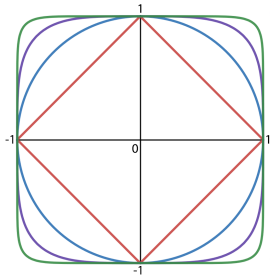


Figure 1: p-circles for $p=1,2,4$, and 10 from inside to outside, respectively

Can we parametrize these p-circles to get p-trigonometric functions $x=\sin_p t$ and $y=\cos_p t$ such that, when $p=2$, we recover the standard trigonometric functions? What properties and identities do these generalized trigonometric functions have? Can we do calculus over these curves? What can be said about the periods of these functions? How does the curvature change along a p-circle? What is the area it encloses? What are the rational points on p-circles? Note that for any $p\geq 1$, L_p, as defined above, gives a norm, but this norm is induced by an inner product only when $p=2$ ([9]). Therefore, $p=2$ is a special case of interest; however, all of the aforementioned questions are well defined for any $p\geq 1$. The goal of this paper is to investigate these questions. Our

primary reference for this research is [7]. While we follow the general outline given in [7], we also do some independent investigation.

There are at least three ways to generalize trigonometric functions. These correspond to 3 different parametrizations of the unit p-circle: areal, arc length, and angular. It turns out that these three parametrizations are equivalent only when $p = 2$! The parametrization we will be working with corresponds to the areal parametrization. Our investigation of these generalized trigonometric functions and their inverses led to several interesting connections involving double angle formulas, norms induced by inner products, Stirling numbers, Bell polynomials, Lagrange inversion, gamma functions, and generalized π values.

These p-trigonometric functions have several applications, specifically in design. Rather than using rounded rectangles, Apple uses p-circles for their icons, as the curvature continuity leads to a more sleek look, unifying the design of their hardware and icons [10]. Another design application can be found in squircular dinner plates, designed to allow a greater surface area for food while taking up the same amount of cabinet space as their circular counterparts [4].

The paper is organized as follows. In Section 2, we define p-trigonometric functions using a differential equations approach and derive some basic properties of these functions. We show that for any positive integer k, the well-known double angle formula for $\sin(2x)$ holds for $\sin_k(2x)$ if and only if $k = 2$. In Section 3, we focus on the successive derivatives of $\sin_p(x)$. This revealed a connection between the coefficients of the terms in the derivatives and Stirling numbers of the first kind. We derive the Taylor series of $\sin_p^{-1} x$ using Newton's binomial series and then find the Taylor series of its inverse using Lagrange inversion theorem. It is shown that both $\sin_p x$ and $\sin_p^{-1} x$ are analytic functions at $x = 0$. Our work gave rise to the concept of rigidity of functions, which deals with the simultaneous vanishing of the derivatives of a function and its inverse. A generalization of π for p-circles, π_p, and its properties are examined in Section 4 using beta and gamma functions. Furthermore, we use a Monte Carlo method to compute π_p. In Section 5, we determine the value of p for which the unit p-circle is halfway between the unit circle and the square that contains it from the lenses of area, perimeter, and curvature. Rational points on p-circles are determined in Section 6. We end the paper with some questions for future work in Section 7.

2 p-trigonometric Functions

Unless stated otherwise, p will denote a positive real number that is at least 1.

2.1 Coupled Initial Value Problem

The standard trigonometric functions sine and cosine that parametrize the unit circle are famously coupled by the derivative relation $\sin' t = \cos t$, $\cos' t = -\sin t$. If we take $x(t) = \cos t$ and $y(t) = \sin t$, we see that the pair is one of many solutions to the system of differential equations

$$x'(t) = -y(t), \quad y'(t) = x(t).$$

However, with the inclusion of the initial conditions

$$x(0) = 1,\ y(0) = 0,$$

differential equation theory guarantees that the sine and cosine functions are, in fact, the only solutions to this system [2], better known as the Coupled Initial Value Problem (CIVP).

For $p \geq 1$, a natural extension of the CIVP considers the functions $x(t), y(t)$ satisfying

$$x'(t) = -y(t)^{p-1},\ y'(t) = x(t)^{p-1},\ x(0) = 1,\ y(0) = 0.$$

The motivation for this extension comes from that fact that any functions $x(t)$ and $y(t)$ that satisfy the above CIVP parametrize the curve $x^p + y^p = 1$. This is seen by differentiating $h(t) := x(t)^p + y(t)^p$ with respect to t, to get $h'(t) = px(t)^{p-1}x'(t) + py(t)^{p-1}y'(t)$. Substituting $x'(t) = -y(t)^{p-1},\ y'(t) = x(t)^{p-1}$, will show that $h'(t) = 0$. This means $h(t)$ is a constant function. Using the initial conditions, we can conclude that $h(t) = 1$, i.e., $x^p + y^p = 1$, as desired.

Again, from the general theory of differential equations, the above CIVP has a unique solution. We can define $\cos_p t = x(t)$ and $\sin_p t = y(t)$ as the unique solution to the generalized CIVP. But these functions do not parametrize p-circles in general. For instance, when p is an odd positive integer, these functions parametrize p-circles only in the first quadrant where x and y are both positive. To circumvent this issue, we restrict the domain of the solutions of the CIVP and then extend them to functions on the real line using symmetry and periodicity. This is done in the next three subsections.

Once we have $\sin_p t$ and $\cos_p t$ in place, we may then define the other trigonometric functions $\tan_p t := \frac{\sin_p t}{\cos_p t}$, $\csc_p t := \frac{1}{\sin_p t}$, $\sec_p t := \frac{1}{\cos_p t}$, and $\cot_p t := \frac{1}{\tan_p t}$ such that the familiar inverse relations are maintained.

2.2 Inverse p-trigonometric Functions

Starting with the equation $x = \sin_p y$, we use the CIVP to find $\sin_p^{-1} x$. Differentiating both sides with respect to y and simplifying, we find:

$$\begin{aligned}
\frac{dx}{dy} &= \cos_p^{p-1} y \\
&= \left(\cos_p^p y\right)^{\frac{p-1}{p}} \\
&= \left(1 - \sin_p y^p\right)^{\frac{p-1}{p}} \\
&= \left(1 - x^p\right)^{\frac{p-1}{p}}.
\end{aligned}$$

This is a separable differential equation. To solve it, we separate and integrate both sides. This gives:

$$\begin{aligned}
\frac{dx}{dy} &= \left(1 - x^p\right)^{\frac{p-1}{p}} \\
\int \frac{dx}{\left(1 - x^p\right)^{\frac{p-1}{p}}} &= \int dy \\
\int_0^x \frac{dt}{\left(1 - t^p\right)^{\frac{p-1}{p}}} &= y = \sin_p^{-1} x.
\end{aligned}$$

We can do the same for $x = \cos_p y$ to get

$$\cos_p^{-1} x = \int_x^1 \frac{dt}{(1-t^p)^{\frac{p-1}{p}}}.$$

2.3 Areal Parametrization of p-circles

The unit circle has a useful property that a sector with angle measure θ in radians has an area of $\theta/2$. We can use this property to find sine and cosine in terms of area where $x = \cos(2a)$, $y = \sin(2a)$, and a is the area of the sector made by the points $(1,0)$ and (x,y). It is then natural to ask if this property extends to all p-circles.

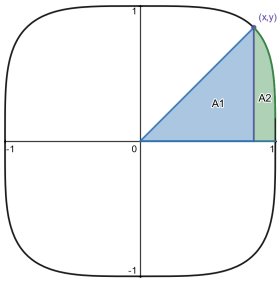


Figure 2: Area of a p-sector

Proposition 1. *Let (x,y) be a point in the first quadrant of the unit p-circle, and a be the area of the sector made by the points $(1,0)$ and (x,y). It holds that $x = \cos_p(2a)$ and $y = \sin_p(2a)$.*

Proof. This argument is in the spirit of Levin [3]. Working in the first quadrant, the area of the sector in a p-circle can be given by the area of $A1 + A2$ as denoted in Figure 2. This can be given by $a = \frac{1}{2}x(1-x^p)^{\frac{1}{p}} + \int_x^1 (1-t^p)^{\frac{1}{p}}\,dt$. We can differentiate both sides with respect to x and simplify to get the following:

$$\begin{aligned}
\frac{da}{dx} &= \frac{1}{2}\left((1-x^p)^{\frac{1}{p}} + x\frac{1}{p}(1-x^p)^{\frac{1}{p}-1}(-px^{p-1})\right) - (1-x^p)^{\frac{1}{p}} \\
&= (1-x^p)^{\frac{1}{p}}\left(\frac{1}{2} - \frac{x^p}{2}(1-x^p)^{-1} - 1\right) \\
&= (1-x^p)^{\frac{1}{p}}\left(\frac{(1-x^p) - x^p - 2(1-x^p)}{2(1-x^p)}\right) \\
&= -\frac{(1-x^p)^{\frac{1}{p}-1}}{2}.
\end{aligned}$$

Using the fundamental theorem of calculus, we can write this as $a = \int_x^1 \frac{(1-t^p)^{\frac{1}{p}-1}}{2}dt + c$. When $a = 0$ and $x = 1$, we get $c = 0$. From here, we can conclude that $a = \frac{1}{2}\arccos_p x$. Solving for x gives $x = \cos_p(2a)$.

We can do the same thing in terms of y to get $a = \int_0^y \frac{(1-t^p)^{\frac{1}{p}-1}}{2}dt$+c. When $a = 0$ and $y = 0$, we get $c = 0$. From here, this equation has been shown to be $a = \frac{1}{2}\arcsin_p y$ and thus $y = \sin_p(2a)$. As such, this shows that this property does extend to all unit p-circles. □

2.4 Definition and Graphs of $\sin_p x$ and $\cos_p x$

To generalize the formula $\pi/2 = \sin^{-1}(1)$, we first set $\pi_p/2 := \sin_p^{-1}(1)$. Since we have shown that the p-trigonometric functions can be parametrized by area, we can now extend then to functions defined on the entire real line as follows. We first restrict them to $[0, \sin_p^{-1}(1)] = [0, \pi_p/2]$ and then extend the domain to $[0, 2\pi_p]$ using symmetry:

$$\sin_p t := \begin{cases} \sin_p(\pi_p - t) & \pi_p/2 < t \leq \pi_p, \\ -\sin_p(2\pi_p - t) & \pi_p < t < 2\pi_p. \end{cases}$$

We then periodically extend that it $(-\infty, \infty)$ by setting $\sin_p(t + 2\pi_p k) = \sin_p(t)$ for any integer k. The definition of $\cos_p(t)$ is similar. The resulting graphs are shown below.

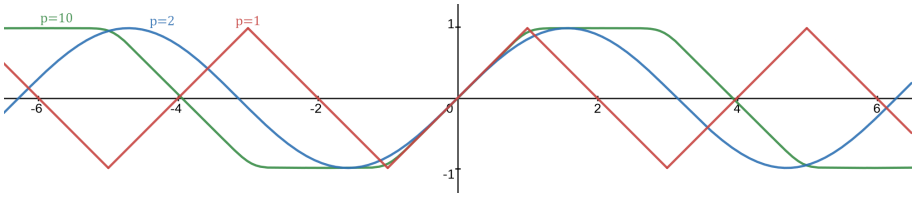


Figure 3: Graph of $\sin_p x$ for $p = 1, 2$, and 10

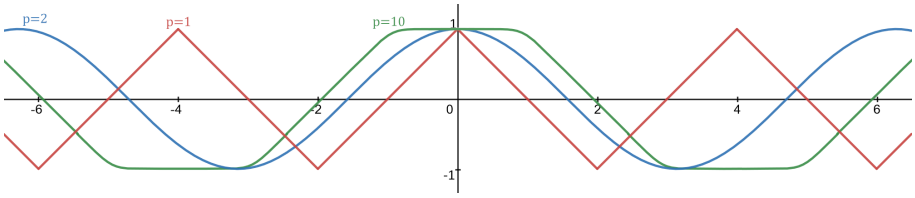


Figure 4: Graph of $\cos_p x$ for $p = 1, 2$, and 10

2.5 Trigonometric Identities

While we may have defined the generalized CIVP in a manner similar to the original, there is no guarantee that $\sin_p t, \cos_p t$ thus defined satisfy familiar trigonometric properties and identities. In this section, we explore a few identities of the p-trigonometric functions.

Lemma 2 (p-Pythagorean Equation)**.** *[7, p. 268] The functions* $\sin_p t, \cos_p t$ *satisfy* $|\sin_p t|^p + |\cos_p t|^p = 1$ *for all real* t.

Proof. This is clear from the definition of these functions using the CIVP and extension using symmetry and periodicity. □

It is clear from the p-Pythagorean Equation that the functions $\sin_p t, \cos_p t$ are bounded and $|\sin_p t| \leq 1, |\cos_p t| \leq 1$. Dividing all terms of the p-Pythagorean equation by $|\sin_p t|^p$ and $|\cos_p t|^p$ gives the identities $1 + |\cot_p t|^p = |\csc_p t|^p$ and $|\tan_p t|^p + 1 = |\sec_p t|^p$, respectively.

Lemma 3. *The functions* $\sin_p t$ *and* $\cos_p t$ *are odd and even, respectively.*

Proof. The functions $\alpha(t) := -\sin_p(-t)$ and $\beta(t) = \cos_p(-t)$ satisfy $\alpha'(t) = -(-\sin_p'(-t)) = \cos_p^{p-1}(-t) = \beta(t)^{p-1}$ and $\beta'(t) = -\cos_p'(-t) = \sin_p^{p-1}(-t) = -\alpha(t)^{p-1}$. Note that $\alpha(0) = -\sin_p(0) = 0$ and $\beta(0) = \cos_p(0) = 1$; thus, the functions α, β satisfy the generalized CIVP. Then, by the uniqueness of solutions, we must have $\sin_p t = -\sin_p(-t)$ and $\cos_p t = \cos_p(-t)$. □

However, not all standard 2-trigonometric identities are satisfied. For instance, we show that for positive integer values of p, $\sin_p(2t) = 2\sin_p t \cos_p t$ is satisfied if and only if $p = 2$. A double angle formula for generalized trigonometric functions is still sought after [1, 8].

Proposition 4. *Let* $k \in \mathbb{Z}_+$. *Then* $\sin_k(2t) = 2\sin_k(t)\cos_k(t)$ *if and only if* $k = 2$.

Proof. The desired identity is well known for $k = 2$. We suppose the identity holds for $k \geq 1$ and show that k must be 2. We consider the cases $k = 1$ and $k > 1$ separately. For $k = 1$, we note that the CIVP gives the unique solution $\sin_1(t) = t$ and $\cos_1(t) = 1 - t$. Then $\sin_1(2t) = 2t \neq 2t(1-t) = 2\sin_1 t \cos_1 t$. If $k > 1$, then by Lemma 2, the functions $\sin_k$ and $\cos_k$ satisfy $|\sin_k t|^k + |\cos_k t|^k = 1$. By the Intermediate Value Theorem, there exists some t_0 in $[0, \pi_p/2]$ such that $\sin_k(t_0) = \cos_k(t_0)$. As $t_0 \geq 0$, the substitution $\sin_k(t_0) = \cos_k(t_0)$ into the p-Pythagorean identity gives $2\sin_k^k(t_0) = 1$, therefore $\sin_k^k(t_0) = \frac{1}{2}$. Then, by the assumption that $\sin_k(2t) = 2\sin_k t \cos_k t$ is satisfied for all t, we may raise all terms to the power k and evaluate at the point t_0 to obtain $\sin_k^k(2t_0) = 2^k(\frac{1}{2})\frac{1}{2} = 2^{k-2}$. Since $\sin_k^k t$ is bounded above by 1, we obtain $2^{k-2} \leq 1$, which implies that $k \leq 2$. Together with the assumption that $k > 1$, we obtain $k = 2$. □

It is known that the L_p norm is induced by an inner product if and only if $p = 2$ [9]. Then together with Proposition 4, we make the following remark.

Remark 5. *The following are equivalent for* $k \in \mathbb{Z}_+$*:*

- L_k *is a norm induced by an inner product,*
- $\sin_k(2t) = 2\sin_k t \cos_k t$, *and*
- $k = 2$.

3 Taylor Series

Now that we have defined p-trigonometric functions and their derivatives by the CIVP, it is natural to study the higher derivatives of these functions. We begin by observing that for any $p > 1$, all the successive derivatives of $\sin_p x$ and $\cos_p x$ are defined for all values of x. In this section, we provide an algorithm for differentiating these functions, demonstrate some patterns and connections present in their successive derivatives, and formulate the Taylor series for $\sin_p^{-1} x$ and $\sin_p x$. The Taylor series representations of these functions provide a tool to express all of the derivatives of the p-trigonometric functions in one formula.

3.1 Higher Derivatives and the Bracket Notation

Because of the simplicity and utility of the closed formulas for differentiation of $\sin_2 x$, $\cos_2 x$, it is natural to wonder about higher derivatives of $\sin_p x$, $\cos_p x$. We find these higher derivatives by utilizing the definition given by the CIVP in Section 2.1. However, these derivatives become complex rather quickly. To help address this, we introduce a notation that will be used throughout this section in relation to higher derivatives of these p-trigonometric functions: $[m,n]_p := \cos_p^m(x)\sin_p^n(x)$.

Lemma 6. *The derivative of* $\cos_p^m(x)\sin_p^n(x)$ *satisfies* $\frac{d}{dx}[m,n]_p = -m[m-1,n+p-1]_p + n[m+p-1,n-1]_p$.

Proof. Applying the standard rules of differentiation, we get the following.

$$\begin{aligned}\frac{d}{dx}[m,n]_p &= \frac{d}{dx}(\cos_p^m(x)\sin_p^n(x))\\ &= -m\cos_p^{m-1}(x)\sin_p^{p-1}(x)\sin_p^n(x) + \cos_p^m(x)\cdot n\sin_p^{n-1}(x)\cos_p^{p-1}(x)\\ &= -m\cos_p^{m-1}(x)\sin_p^{n+p-1}(x) + n\cos_p^{m+p-1}(x)\sin_p^{n-1}(x).\\ &= -m[m-1,n+p-1]_p + n[m+p-1,n-1]_p.\square\end{aligned}$$

Although we do not have a closed formula for finding derivatives of these functions, Lemma 6 serves as a recursive algorithm for computing successive derivatives, as demonstrated in the following example.

Example 7. *Lemma 6 can be iteratively applied to* $\sin_p x$ *to find the first few derivatives:*

$$\begin{aligned}\sin_p x &= [0,1]_p\\ \frac{d}{dx}\sin_p x &= 0 + 1[p-1,0]_p\\ \frac{d^2}{dx^2}\sin_p x &= 0 + (-p+1)[p-2,p-1]_p + 0\\ \frac{d^3}{dx^3}\sin_p x &= 0 + -(p-1)(-(p-2)[p-3,2p-2]_p + (p-1)[2p-3,p-2]_p) + 0\\ &= 0 + (p^2-3p+2)[p-3,2p-2]_p + (-p^2+2p-1)[2p-3,p-2]_p + 0.\end{aligned}$$

There seems to be no clear pattern that arises from these derivatives like there is for $\sin x$. However, in the next subsection, we will see one pattern in the coefficients of the first terms of these derivatives.

3.2 Connection to Stirling Numbers

For any variable x and a non-negative integer n, the falling factorial is defined as follows.

$$(x)_n := \begin{cases} 1 & \text{if } n = 0,\\ x(x-1)(x-2)\cdots(x-n+1) & \text{if } n \geq 1.\end{cases}$$

For $n \geq 1$, $(x)_n$ is a non-constant polynomial of degree n whose coefficients are the

Stirling numbers of the first kind. More precisely, we set:

$$(x)_n = \sum_{k=1}^{n} s(n,k)x^k.$$

We will now show a connection between the successive derivatives of $\sin_p x$ and Stirling numbers. Building a tower from the coefficients in Example 7, we get:

$$\begin{array}{ccccccc} & & & 1 & & & \\ & & 0 & | & \underline{1} & & \\ & 0 & | & \underline{-p+1} & | & 0 & \\ 0 & | & \underline{p^2-3p+2} & | & -p^2+2p-1 & | & 0 \end{array}$$

We observed that the coefficients of the polynomials in the second column (underlined) can be expressed using Stirling numbers of the first kind $s(n,k)$. For instance, corresponding to the polynomial p^2-3p+2 (corresponding to the 3rd derivative of $\sin_p x$), we have $s(3,3)=1$, $s(3,2)=-3$ and $s(3,1)=2$. To prove this, we need the following lemma.

Lemma 8. *For any $n \geq 1$, the first term of $\frac{d^n}{dx^n}(\sin_p(x))$ is given by*

$$(-1)^{n-1}(p-1)_{n-1}[p-n,(n-1)(p-1)]_p.$$

Proof. We prove this using mathematical induction. For $n=1$, $\frac{d}{dx}(\sin_p(x)) = \cos_p^{p-1}(x)$ $= 1[p-1,0]_p$, which agrees with the answer obtained with $n=1$ in the given expression. Having proved the base case, let us assume that the result is true for $n=k$. Differentiating the first term of $\frac{d^k}{dx^k}(\sin_p(x))$ using the chain rule, and only picking the first term of the resulting expression will give us

$$\begin{aligned}&(-1)^{k-1}(p-1)_{k-1}\left((-1)(p-k)[p-(k+1),k(p-1)]\right)\\ &\quad = (-1)^k(p-1)_{k-1}(p-1)[p-(k+1),k(p-1)].\end{aligned}$$

The recursive nature of the falling factorial tells us that $(p-1)_{k-1}(p-1) = (p-1)_k$. This shows that the first term of $\frac{d^{k+1}}{dx^{k+1}}(\sin_p(x))$ is given by $(-1)^k(p-1)_k[p-(k+1),k(p-1)]$. By the principle of mathematical induction, the result is true for all $n \geq 1$. □

The connection to Stirling numbers and the successive derivatives of the $\sin_p(x)$ is now clear. Simplifying the coefficient of the first term of $\frac{d^n}{dx^n}(\sin_p(x))$ obtained from the above lemma gives:

$$(-1)^{n-1}(p-1)_{n-1} = (-1)^{n-1}\frac{(p)_n}{p} = \frac{(-1)^{n-1}}{p}\sum_{k=1}^{n} s(n,k)p^k.$$

3.3 Newton's Binomial Series

Let p be any integer that is greater than 1. As the previous section demonstrates, finding a formula for the successive derivatives of $\sin_p x$ to compute its Taylor series is complicated. Instead, we examine $\sin_p^{-1} x$, whose Taylor series at $x=0$ is more manageable, and use this to find the Taylor series of $\sin_p x$ at $x=0$ through the Lagrange inversion theorem. To do this, we apply Newton's binomial series to derive the Taylor series of $\sin_p^{-1} x$. Newton's binomial series tells us the following for any exponent a

and $|x| < 1$:

$$\begin{aligned}(1-x)^{-a} &= 1 + ax + \frac{a(a+1)}{2!}x^2 + \frac{a(a+1)(a+2)}{3!}x^3 + \ldots \\ &= \sum_{k=0}^{\infty} \frac{a^{(k)}x^k}{k!},\end{aligned}$$

where $a^{(k)} = a(a+1)(a+2)\cdots(a+k-1)$ is the rising factorial [12, p. 742]. Note that, by convention, $a^{(0)} = 1$.

Proposition 9. *We can express* $\sin_p^{-1} x$ *as the following Taylor series:*

$$\sin_p^{-1} x = \sum_{k=0}^{\infty} \left(\frac{p-1}{p}\right)^{(k)} \frac{x^{kp+1}}{k!(kp+1)}.$$

Proof. Beginning with the integral form of $\sin_p^{-1} x$ derived in Section 2.2, we apply Newton's binomial series:

$$\begin{aligned}\sin_p^{-1} x &= \int_0^x (1-t^p)^{-(\frac{p-1}{p})} dt \\ &= \int_0^x \left(\sum_{k=0}^{\infty} \left(\frac{p-1}{p}\right)^{(k)} \frac{t^{kp}}{k!}\right) dt.\end{aligned}$$

Power series have the property that they can be integrated term by term within the interval of convergence. Thus, when we integrate and apply the fundamental theorem of calculus, the result follows. □

Example 10. *Applying Proposition 9 for* $p = 2$ *gives the following well-known result:*

$$\sin_2^{-1} x = x + \frac{1}{6}x^3 + \frac{3}{40}x^5 + \frac{5}{112}x^7 + \cdots + \binom{2n}{n}\frac{x^{2n+1}}{2^{2n}(2n+1)} + \cdots.$$

Similarly, when $p = 4$*, we get the first few terms as follows:*

$$\sin_4^{-1} x = x + \frac{3}{20}x^5 + \frac{7}{96}x^9 + \frac{77}{1664}x^{13} + \cdots.$$

It would be helpful to have a closed-form solution for these higher derivatives. In the next section, we introduce some tools and discuss what this will look like.

3.4 $\sin_p^{-1} x$ through the Gamma Function

We now introduce a special function to shed light on $\sin_p^{-1} x$. The gamma function, $\Gamma(z)$, is defined as $\Gamma(z) = \int_0^\infty e^{-t}t^{z-1}\,dt$, for $z > 0$. This converges for any real number $z > 0$, and it is an extension of the factorial function: $\Gamma(n) = (n-1)!$. It is well-known that $\Gamma(1/2) = \sqrt{\pi}$. Two important properties of the gamma function are

$$\Gamma(x+1) = x\Gamma(x) \quad \text{and} \quad \Gamma(x)\Gamma(1-x) = \frac{\pi}{\sin(\pi x)}.$$

Using the gamma function, for any integer $p > 1$, we can further simplify the Taylor series for $\sin_p^{-1}(x)$ as follows. We begin by the formula from Proposition 9 which states that

$$\sin_p^{-1} x = \sum_{k=0}^{\infty} \left(\frac{p-1}{p}\right)^{(k)} \frac{x^{kp+1}}{k!(kp+1)}.$$

Then we have the following:

$$\begin{aligned}
\sin_p^{-1} x &= \sum_{k=0}^{\infty} \left(\frac{p-1}{p}\right)^{(k)} \frac{x^{kp+1}}{k!(kp+1)} \\
&= \sum_{k=0}^{\infty} \left(1-\frac{1}{p}\right)^{(k)} \frac{x^{kp+1}}{k!(kp+1)} \\
&= \sum_{k=0}^{\infty} \left(1-\frac{1}{p}\right)\left(2-\frac{1}{p}\right)\left(3-\frac{1}{p}\right)\cdots\left(k-\frac{1}{p}\right) \frac{x^{kp+1}}{k!(kp+1)} \\
&= \sum_{k=0}^{\infty} \Gamma\left(1-\frac{1}{p}\right)\left(1-\frac{1}{p}\right)\left(2-\frac{1}{p}\right)\left(3-\frac{1}{p}\right)\cdots\left(k-\frac{1}{p}\right) \frac{1}{\Gamma(1-\frac{1}{p})} \frac{x^{kp+1}}{k!(kp+1)} \\
&= \sum_{k=0}^{\infty} \Gamma\left(2-\frac{1}{p}\right)\left(2-\frac{1}{p}\right)\left(3-\frac{1}{p}\right)\cdots\left(k-\frac{1}{p}\right) \frac{1}{\Gamma(1-\frac{1}{p})} \frac{x^{kp+1}}{k!(kp+1)} \\
&\vdots \\
&= \sum_{k=0}^{\infty} \Gamma\left(k-\frac{1}{p}\right) \frac{1}{\Gamma(1-\frac{1}{p})} \frac{x^{kp+1}}{k!(kp+1)} \\
&= \sum_{k=0}^{\infty} \frac{\Gamma\left(k-\frac{1}{p}\right)}{\Gamma\left(1-\frac{1}{p}\right)} \frac{x^{kp+1}}{k!(kp+1)}.
\end{aligned}$$

Theorem 11. *Let $n > 1$ be a positive integer. Then for any positive integer l, let k and r be the integers given by the division algorithm: $l = nk + r$ where $k \geq 0$ and $0 \leq r \leq n-1$. Then we have*

$$\left(\frac{d^l}{dx^l} \sin_n^{-1}(x)\right)\Big|_{x=0} = \begin{cases} \frac{\Gamma\left(k-\frac{1}{n}\right)}{\Gamma\left(1-\frac{1}{n}\right)} \frac{(kn)!}{k!}, & \text{if } r = 1, \\ 0, & \text{if } r \neq 1. \end{cases}$$

Proof. The Taylor series for $\sin_n^{-1}(x)$ at $x = 0$ has the form

$$\sin_n^{-1}(x) = h(x) = \sum_{m=0}^{\infty} \frac{h^{(m)}(0)}{m!} x^m.$$

On the other hand, from the above calculation, we know that

$$\sin_n^{-1}(x) = \sum_{k=0}^{\infty} \frac{\Gamma\left(k-\frac{1}{n}\right)}{\Gamma\left(1-\frac{1}{n}\right)} \frac{x^{kn+1}}{k!(kn+1)} = \sum_{k=0}^{\infty} \left(\frac{\Gamma\left(k-\frac{1}{n}\right)}{\Gamma\left(1-\frac{1}{n}\right)} \frac{(kn)!}{k!}\right) \frac{1}{(kn+1)!} x^{kn+1}.$$

Equating the coefficients of like-powers of x in both these series, we get the theorem. □

Now that we have derived the Taylor series of $\sin_p^{-1} x$, we can apply the Lagrange inversion theorem as outlined in the next section.

3.5 Lagrange Inversion

A function $z = f(w)$ is said to be analytic at c if it is infinitely differentiable at c and if the Taylor series for $f(w)$ at $w = c$ converges to $f(w)$ for all w in a neighborhood of c.

For an equation $z = f(w)$, where f is analytic at c and $f'(c) \neq 0$, the Lagrange inversion theorem can be used to find the equation's inverse, $w = g(z)$, in a neighborhood of 0. This inverse is given by the formula [5, Chapter 3]:

$$g(z) = c + \sum_{n=1}^{\infty} g_n \frac{(z-f(c))^n}{n!}, \text{ where}$$
$$g_n = \lim_{w\to c} \frac{d^{n-1}}{dw^{n-1}} \left[\left(\frac{w-c}{f(w)-f(c)} \right)^n \right].$$

For power series, this theorem takes a slightly different form. Specifically, when f and g are formal power series expressed as

$$f(w) = \sum_{k=0}^{\infty} f_k \frac{w^k}{k!} \quad \text{and} \quad g(z) = \sum_{k=0}^{\infty} g_k \frac{z^k}{k!},$$

with $f_0 = 0$ and $f_1 \neq 0$, applying the Lagrange inversion theorem gives us the following [5]:

$$g(z) = c + \sum_{n=1}^{\infty} g_n \frac{(z-f(c))^n}{n!}, \quad \text{with}$$

$$g_n = \frac{1}{f_1^n} \sum_{k=1}^{n-1} (-1)^k n^{(k)} B_{n-1,k}(\hat{f}_1, \hat{f}_2, ..., \hat{f}_{n-k}), \quad n \geq 2, \quad \text{where}$$

$$\hat{f}_k = \frac{f_{k+1}}{(k+1)f_1}, \quad g_1 = \frac{1}{f_1}, \quad n^{(k)} = n(n+1)\cdots(n+k-1), \text{ and}$$

$$B_{n,k}(x_1, x_2, ..., x_{n-k+1}) = \sum \frac{n!}{j_1! j_2! ... j_{n-k+1}!} \left(\frac{x_1}{1!}\right)^{j_1} \left(\frac{x_2}{2!}\right)^{j_2} ... \left(\frac{x_{n-k+1}}{(n-k+1)!}\right)^{j_{n-k+1}},$$

where this sum is taken over all sequences $j_1, j_2, j_3, ..., j_{n-k+1}$ of non-negative integers that satisfy $j_1 + j_2 + ... + j_{n-k+1} = k$ and $j_1 + 2j_2 + 3j_3 + ... + (n-k+1)j_{n-k+1} = n$. These are the Bell polynomials.

The Taylor series expansion of $\sin_p x$ is obtained when the above theorem is applied to

$$\sin_p^{-1} x = \sum_{k=0}^{\infty} \left(\frac{p-1}{p} \right)^{(k)} \frac{x^{kp+1}}{k!(kp+1)},$$

which was derived in the previous section. We are able to apply this theorem to $\sin_p^{-1} x$, as it meets the initial conditions given above: $f_0 = 0$ and $f_1 \neq 0$.

Example 12. *When $p = 2$, we can apply Lagrange Inversion Theorem with $c = 0$, as $f(c) = 0$ and $f'(c) = 1$. To do so, we must calculate f_k for the first few terms. Expanding $\sin_2^{-1} x$, we find*

$$f_0 = 0,\ f_1 = 1,\ f_2 = 0,\ f_3 = 1,\ f_4 = 0,\ f_5 = 9,\ f_6 = 0.$$

Using these values, we can find $\hat{f}_k$:

$$\hat{f}_1 = 0,\ \hat{f}_2 = \frac{1}{3},\ \hat{f}_3 = 0,\ \hat{f}_4 = \frac{9}{5},\ \hat{f}_5 = 0.$$

We may now use these values to find the first few g_n using the formulas given above. To this end, we record a couple of special Bell polynomials that will be used below: $B_{n,n}(x_1) = (x_1)^n$ and $B_{n,n-1}(x_1, x_2) = \binom{n}{2}(x_1)^{n-2}x_2$. These are obtained by simplifying the general Bell polynomial given above.

When $n=1$, $g_1 = \frac{1}{f_1} = \frac{1}{1} = 1$. When $n=2$, we have $g_2 = (-1)^1 \cdot 2^{(1)} B_{1,1}(0) = 0$. Similarly, when $n=3$, we have

$$\begin{aligned} g_3 &= \frac{1}{f_1^3}\left((-1)^1 3^{(1)} B_{2,1}(\hat{f}_1, \hat{f}_2) + (-1)^2 3^{(2)} B_{2,2}(\hat{f}_1)\right) \\ &= \frac{1}{1^3}\left(-3B_{2,1}(0,1/3) + 12B_{2,2}(0)\right) \\ &= -3\binom{2}{2}\frac{1}{3} + 12(0^2) = -1. \end{aligned}$$

In the same manner, applying this formula to the next few values of n, we find that $g_4 = 0$ and $g_5 = 1$.

Substituting these values into the formula for $g(z)$ given by Lagrange Inversion Theorem above, we have:

$$g(z) = 0 + \sum_{n=1}^{\infty} g_n \frac{(z-0)^n}{n!}.$$

$$\sin_2(z) = z - \frac{z^3}{3!} + \frac{z^5}{5!} + \cdots.$$

When $p=4$, these computations get more tedious. Using SageMath, we find that

$$\sin_4 x = x - \frac{18}{5!}x^5 + \frac{14364}{9!}x^9 - \cdots.$$

The above ideas prove the following theorem.

Theorem 13. *For any integer $p > 1$, the functions $\sin_p^{-1} x$ and $\sin_p x$ are analytic at $x = 0$.*

It is well-known that $\sin x/x \to 1$ as $x \to 0$. We now generalize this result.

Corollary 14. *Let $p > 1$ be an integer. Then we have*

$$\lim_{x\to 0} \frac{\sin_p x}{x} = 1.$$

Proof. By Theorem 13, we know that $\sin_p x$ is analytic at $x = 0$, and moreover, from the CIVP, $\sin_p 0 = 0$. Therefore, we can express $\sin_p x$ as a power series whose constant term is 0:

$$\sin_p x = a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n + \cdots.$$

Differentiating both sides and invoking the CIVP gives:

$$(\cos_p x)^{p-1} = a_1 + 2a_2 x + 3a_3 x^2 + \cdots + n x^{n-1} + \cdots.$$

Since $\cos_p(0) = 1$, setting $x = 0$ in the above equation tells us that $a_1 = 1$. Finally, we have

$$\lim_{x\to 0} \frac{\sin_p x}{x} = \lim_{x\to 0} \frac{x + a_2 x^2 + a_3 x^3 + \cdots}{x} = \lim_{x\to 0} 1 + a_2 x + a_3 x^2 + \cdots = 1.$$

□

Note that from this result it also follows that $\tan_p x/x \to 1$ as $x \to 0$. One can also prove these limits using l'Hôpital's rule. In the same vein, one can also show the following.

Corollary 15.

$$\lim_{x\to 0}\frac{\sin_4 x - x}{x^5} = -\frac{18}{5!}.$$

3.6 Rigidity

Note that the missing terms of the Taylor series for $\sin_4(x)$ are exactly the ones that were also missing in $\sin_4^{-1}(x)$; see Example 10. In fact, for both functions, the non-zero terms in the Taylor series correspond to powers of x that form an arithmetic progression of the form $4m+1$. We proved this fact in Theorem 11 for $\sin_n^{-1}(x)$. We now conjecture that this is also true for $\sin_n(x)$.

Conjecture 16. *Let n be a positive integer. Then*

$$\left(\frac{d^l}{dx^l}\sin_n(x)\right)\Big|_{x=0} \neq 0 \iff l \equiv 1 \mod n.$$

This led to the following, more general question in analysis.

Question: Suppose $f(x)$ is a real-valued function that is infinitely differentiable at $x=a$ such that $f'(a)\neq 0$. Let $f(a)=b$ and let $g(x)$ be the local inverse of $f(x)$ (this exists by the inverse function theorem) at $x=a$. Is it true that for every positive integer n, the nth derivative of $f(x)$ at $x=a$ is non-zero if and only if the nth derivative of $g(x)$ at $x=b$ is non-zero?

It turns out that, in general, the above answer is no. Take for example $f(x)=x^2$. We have $f(1)=1$ and $f'(1)=2\neq 0$. At $x=1$, the local inverse of $f(x)$ is $g(x)=\sqrt{x}$. Note that for all $k\geq 3$, $f^{(k)}(1)=0$ but $g^{(k)}(1)\neq 0$. On the other hand, for the function $f(x)=\sin(x)$, the above question has an affirmative answer because the Taylor series for $\sin x$ and $\sin^{-1}x$, have only odd terms. This leads naturally to the following definition.

Definition 17. Let $y=f(x)$ be a function that is infinitely differentiable at $x=a$ such that $f'(a)\neq 0$. We say that $f(x)$ is rigid at $x=a$ if for any positive integer k, $f^{(k)}(a)\neq 0$ if and only if $g^{(k)}(b)\neq 0$, where $g(x)$ is the local inverse of $f(x)$ at $x=a$ and $b=f(a)$.

In this terminology, $f(x)=x^2$ is not rigid at $x=1$ but $f(x)=\sin x$ is rigid at $x=0$. Conjecture 16 can now be restated as follows. For any positive integer k, $\sin_k x$ is rigid at $x=0$.

Question: What are necessary and sufficient conditions for a function $y=f(x)$ that is infinitely differentiable at $x=a$ to be rigid at a?

4 Generalized π values

4.1 Organic Definition

As we generalize trigonometric functions in the p-norm, we must also take into consideration generalizing the value of π. Recall that $\pi = 2\sin^{-1}(1)$. Using this as our inspiration, we can organically define π_p as $\pi_p := 2\sin_p^{-1}(1)$. Using our $\sin_p^{-1}x$ formula

we derived in Section 2.2 and letting $x = 1$, we get

$$\pi_p = 2\int_0^1 \frac{1}{(1-t^p)^{\frac{p-1}{p}}}\,dt. \tag{1}$$

Note that, unless otherwise indicated, when we refer to π, we are referring to π_2.

When $p = 2$, we find that the area of the unit circle is equal to π. It is then natural to wonder if π_p has any relation to the area of a unit p-circle.

Proposition 18. *The area of a unit p-circle is π_p, when $p \geq 1$.*

Proof. In Proposition 1, we found that the area of the sector of the p-circle that connects the points (x,y) and $(1,0)$ is given as a function of y by $a(y) = \frac{1}{2}\int_0^y (1-t^p)^{\frac{1}{p}-1}\,dt$. If we let (x,y) be the point $(0,1)$, we get the area of the unit p-circle in the first quadrant, given by $a(1) = \frac{1}{2}\int_0^1 (1-t^p)^{\frac{1}{p}-1}\,dt$. Since the unit p-circle has 4-fold symmetry, we can multiply both sides of the equation by four to find the area of the entire p-circle:

$$4a(1) = 2\int_0^1 (1-t^p)^{\frac{1}{p}-1}\,dt.$$

From Section 2.2, we know that $2\int_0^x (1-t^p)^{\frac{1}{p}-1}\,dt = 2\sin_p^{-1} x$. Therefore, we know that the right hand side of the equation is $2\sin_p^{-1}(1)$, which is equal to π_p as shown in Equation (1). We have also already established that $a(1)$ is the area of the quarter unit p-circle, so $4 \cdot a(1)$ gives us the area of the entire unit p-circle. Therefore, we find that the area of the unit p-circle is π_p. □

Corollary 19. *For any $p \geq 1$, we have $2 \leq \pi_p < 4$.*

Proof. As shown above, π_p is the area of a unit p-circle. When $p = 1$, we get the region bounded by the square $|x| + |y| = 1$, which has area 2. Similarly, since the p-circle is inscribed in a square of side length two, we know that the area of the p-circle is bounded by the square's area, which is 4. This shows that for any $p \geq 1$, we have $2 \leq \pi_p < 4$. □

4.2 A Formula for π_p

We now show how we can compute π_p in terms of the gamma function. To this end, we need another special function called the beta function, $\beta(x,y)$, which is closely related to the gamma function and can be defined as $\beta(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt$, for any two real numbers x, y such that $x > 0$ and $y > 0$. We can put the beta function in terms of gamma using the property

$$\beta(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}. \tag{2}$$

Proposition 20. *For any $p \geq 1$, we have*

$$\pi_p = \frac{2\Gamma^2(\frac{1}{p})}{p\Gamma(\frac{2}{p})}.$$

In particular, π_p is a differentiable function of p.

Proof. Referring back to our definition for π_p, if we let $u = t^p$, we can express π_p in terms of the beta function as follows:

$$\pi_p = 2\int_0^1 \frac{1}{(1-t^p)^{\frac{p-1}{p}}}\,dt = \frac{2}{p}\int_0^1 (1-u)^{\frac{1}{p}-1}\cdot u^{\frac{1}{p}-1}\,du = 2\beta(1/p,1/p).$$

We can then use Equation (2) to put π_p in terms of the gamma function, and that gives the formula stated in the proposition. Since $\Gamma(x)$ is a differentiable function and compositions and quotients of differentiable functions are again differentiable, it follows that π_p is differentiable. □

Example 21. *Using the above equation, we can numerically approximate π_p for any p. For $p = 2,3$ and 4, we get:*

$$\pi_2 = \frac{2\Gamma^2(\frac{1}{2})}{2\Gamma(1)} \approx 3.1415, \quad \pi_3 = \frac{2\Gamma^2(\frac{1}{3})}{3\Gamma(\frac{2}{3})} \approx 3.533 \quad \textit{and} \quad \pi_4 = \frac{2\Gamma^2(\frac{1}{4})}{4\Gamma(\frac{2}{4})} \approx 3.708.$$

4.3 Properties of π_p

We have already seen that π_p is a differentiable function of p for all $p > 0$. Is it a monotonic function? Example 21 suggests that π_p increases with p. We now prove that fact.

Proposition 22. *π_p is an increasing function on $(0,\infty)$.*

Proof. Recall that π_p is the area of a unit p-circle. Since p-circles have a 4-fold symmetry, we get $\pi_p = 4\int_0^1 (1-x^p)^{\frac{1}{p}}\,dt$. We will be done if we can show that, for any fixed value of x in $(0,1)$, $(1-x^p)^{\frac{1}{p}}$ is an increasing function in p. This is because if $(1-x^{p_1})^{\frac{1}{p_1}} < (1-x^{p_2})^{\frac{1}{p_2}}$ for all x in $(0,1)$ and $p_1 < p_2$, then

$$4\int_0^1 (1-x^{p_1})^{\frac{1}{p_1}}\,dx < 4\int_0^1 (1-x^{p_2})^{\frac{1}{p_2}}\,dx,$$

showing that $\pi_{p_1} < \pi_{p_2}$ whenever $0 < p_1 < p_2$.

To this end, let $\psi(p) := (1-x^p)^{\frac{1}{p}}$, where x is a fixed number in $(0,1)$. Taking the natural logarithm and differentiating with respect to p on both sides, we get

$$\begin{aligned}
\ln(\psi(p)) &= \frac{\ln(1-x^p)}{p},\\
\frac{\psi'(p)}{\psi(p)} &= \frac{-\ln(1-x^p)}{p^2} + \frac{-\ln(x)x^p}{p(1-x^p)},\\
\psi'(p) &= (1-x^p)^{\frac{1}{p}}\left(\frac{-\ln(1-x^p)}{p^2} + \frac{-\ln(x)x^p}{p(1-x^p)}\right).
\end{aligned}$$

For $0 < x < 1$ and $p > 0$, note that $0 < 1-x^p < 1$. Therefore, $\ln(x)$ and $\ln(1-x^p)$ are both negative. This shows that all parts of the derivative are positive. Therefore, $\psi'(p) > 0$, which means $\psi(p)$ is an increasing function. □

Having shown above that π_p is an increasing function and that it has an upper bound of 4 in Corollary 19, we know that a limit exists. It is then only natural to wonder what

the limit of π_p is.

Proposition 23. $\lim_{p\to\infty} \pi_p = 4$.

Proof. Using $\pi_p = \frac{2\Gamma^2(\frac{1}{p})}{p\Gamma(\frac{2}{p})}$, we can take the limit of π_p as p approaches infinity. Note that we have not yet stated Legendre's duplication formula, $\Gamma(2z) = \frac{\Gamma(z)\Gamma(z+\frac{1}{2})}{2^{1-2z}\sqrt{\pi}}$.

$$\begin{aligned}
\pi_p &= \frac{2}{p}\cdot\frac{\Gamma^2(\frac{1}{p})}{\Gamma(\frac{2}{p})} = \frac{2}{p}\cdot\frac{\Gamma^2(\frac{1}{p})\cdot 2^{1-\frac{2}{p}}\sqrt{\pi}}{\Gamma(\frac{1}{p})\Gamma\left(\frac{1}{p}+\frac{1}{2}\right)} \\
&= \frac{2}{p}\cdot\frac{\Gamma(\frac{1}{p})\cdot 2^{1-\frac{2}{p}}\sqrt{\pi}}{\Gamma\left(\frac{1}{p}+\frac{1}{2}\right)} = \frac{2\cdot\Gamma(\frac{1}{p}+1)\cdot 2^{1-\frac{2}{p}}\sqrt{\pi}}{\Gamma\left(\frac{1}{p}+\frac{1}{2}\right)}. \\
\lim_{p\to\infty}\pi_p &= \frac{2\cdot\Gamma(1)\cdot 2\sqrt{\pi}}{\Gamma(\frac{1}{2})} = \frac{1\cdot 4\sqrt{\pi}}{\sqrt{\pi}} = 4.
\end{aligned}$$

□

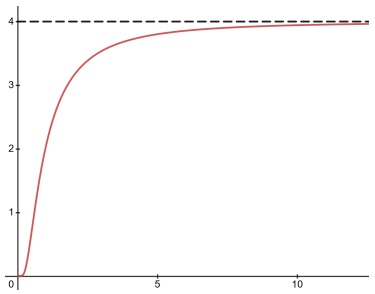


Figure 5: Graph of π_p

Recall that it is well-known that $\pi_2 = \pi$ is an irrational number (a number that is not the ratio of two integers). On the other hand, $\pi_1 = 2$, a rational number. ($\pi_1 = 2$ because it is the area enclosed by the square $|x| + |y| = 1$ of side length $\sqrt{2}$.) It is natural to ask for what values of p, is π_p irrational? This is a hard question. Since π_p is a continuous function, it takes rational and irrational values infinitely often; see Figure 5.

4.4 π_p with the Monte Carlo Method

Since we have shown that π_p can be described as the area of a p-circle, we can use a rather fun technique to approximate the value of π_p. Given a p-circle shaped dartboard inscribed inside a square, what is the probability that a uniformly random throw will land on the dartboard (assuming that the dart must land inside the square)? The probability is the ratio of the area of the board to the area of the box. Therefore, if we have n throws where t of them land on the dartboard, the probability would be $\frac{t}{n} = \frac{\pi_p}{4}$. We can solve for π_p to get $\pi_p = \frac{4t}{n}$. Because of the law of large numbers, when $n \to \infty$, the ratio goes to the true ratio, and we find the true value of π_p. Writing a simple program to do this for us, at $n = 1{,}000{,}000{,}000$, we get $\pi_3 \approx 3.53324$ and $\pi_4 \approx 3.7081$.

5 Optimal Unit p-circles

One question that naturally arises when examining p-circles is, "At what value of p is the corresponding squircle halfway between a unit circle ($p = 2$) and a square ($p \to \infty$)?" This question was examined from three lenses: area, perimeter, and curvature.

5.1 Area

We sought to find the value of p for which the area enclosed by the p-circle $|x|^p + |y|^p = 1$ is $\frac{\pi+4}{2}$, which is the average of the areas of the unit circle and the square that the unit circle is inscribed in. Because the p-circle is symmetric, we can examine the first quadrant only, resulting in the following equation:

$$\int_0^1 \sqrt[p]{1-x^p}\, dx = \frac{\pi+4}{8}.$$

Using SageMath, the root of this equation can be found, giving the approximation $p \approx 3.162038$. As such, we can conclude that the value of p for which the area of the p-circle is exactly halfway between the areas of the unit circle ($p = 2$) and the square in which it is inscribed is $p \approx 3.162038$.

5.2 Perimeter

Next, we want to find the value of p for which the perimeter of a unit p-circle is halfway between those of a unit circle and the square that the unit circle is inscribed in. The circumference of a unit circle is 2π, and the perimeter of a square that contains the unit circle is 8. Therefore, we have to find the value of p for which the perimeter of a unit p-circle is $\pi + 4$. To find the perimeter of the unit p-cirle, we apply the Euclidean arc length formula to the defining equation of a p-circle. We equate the resulting integral to $\pi + 4$ to obtain the equation:

$$\pi + 4 = 4\int_0^1 \sqrt{1 + (1-x^p)^{2(1-p)/p} x^{2(p-1)}}\, dx.$$

We solved this equation numerically using SageMath to find that $p \approx 4.667489$.

5.3 Curvature

Finally, we wish to find p such that the curvature of the unit p-circle is halfway between that of a square (here said to have curvature 0) and the 2-unit circle (which has curvature 1). For a given smooth curve C in $\mathbb{R}^2$, the curvature is a measure of how different our curve is from a circle at a given point. While there are many equivalent formulations of the curvature of a given curve, the following gives the curvature for a curve defined implicitly by $F(x,y) = 0$:

$$\kappa = \frac{|F_y^2 F_{xx} - 2F_x F_y F_{xy} + F_x^2 F_{yy}|}{(F_x^2 + F_y^2)^{\frac{3}{2}}}.$$

Using the relation $F(x,y) = x^p + y^p - 1 = 0$ for the unit p-circle, we obtain

$$\kappa = (p-1)\frac{x^p y^{2p} + x^{2p} y^p}{(x^{2p} y^2 + y^{2p} x^2)^{\frac{3}{2}}}(xy).$$

If we investigate the curvature of the unit p-circle at the point $x = y$, we find that

$$\kappa = (p-1)x^2 \frac{2x^{3p}}{(2x^{2p+2})^{\frac{3}{2}}} = \frac{p-1}{\sqrt{2}x}.$$

When $x = y$, we can write the relation for the unit circle as $2x^p = 1$, which gives $x = 2^{-\frac{1}{p}}$. Substituting for x gives

$$\kappa = \frac{p-1}{\sqrt{2} \cdot 2^{-\frac{1}{p}}} = \frac{p-1}{2^{\frac{1}{2}-\frac{1}{p}}} = (p-1)2^{\frac{1}{p}-\frac{1}{2}}.$$

Therefore, if we solve for p such that the unit p-circle has curvature 1/2, we find that $p \approx 1.43643264$.

5.4 Resulting Graphs

Graphing these 3 p-circles gives Figure 6 where the unit circle and square are dashed, and the p-circle is solid. For the optimal curvature, we also have $p = 1$ since both $p = 1$ and $p \to \infty$ have the same curvature.

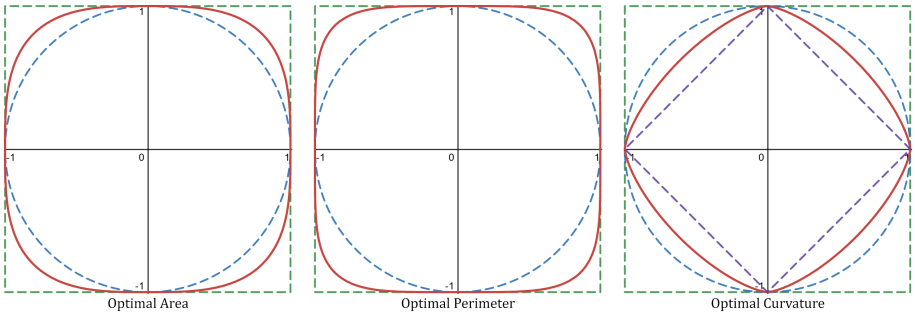


Figure 6: All 3 optimal p-circles

6 Rational Points on p-circles

6.1 2-circles and Pythagorean Triples

Right triangles (and as a result, Pythagorean triples) have long been objects of mathematical interest, studied intensely by the Babylonians even more than a thousand years before Pythagoras [6]. Given any Pythagorean triple (x, y, z) satisfying $x^2 + y^2 = z^2$, we may divide all parts by z^2 to obtain $\frac{x^2}{z^2} + \frac{y^2}{z^2} = 1$. Then the point $(\frac{x}{z}, \frac{y}{z})$ is a rational point which lies on the 2-unit circle defined by $x^2 + y^2 = 1$. On the other hand, given any rational number $\frac{v}{u}$, we obtain the Pythagorean triple $(u^2 - v^2, 2uv, u^2 + v^2)$ [11]. In this manner, we may translate contexts between rational points on the unit circle and right triangles with integer side lengths.

6.2 p-circles and Fermat's Last Theorem

We can generalize the known results for rational points on 2-circles and ask the same question for p-circles where p is an integer greater than 2. There are certainly 4 trivial rational points along the axes of the graph, which are the points $(0,1),(1,0),(0,-1),(-1,0)$. To find the others, we may look at the rational solutions in the first quadrant and use symmetry to extend our answers to the entire unit p-circle.

Let p be an integer greater than 2 and let $P = (\frac{p_1}{q_1}, \frac{p_2}{q_2}) \in \mathbb{Q}^2$ be a rational point on the unit p-circle lying in the first quadrant. As P lies on the unit p-circle and P is in the first quadrant, we must have $(\frac{p_1}{q_1})^p + (\frac{p_2}{q_2})^p = 1$ and $\frac{p_1}{q_1} > 0, \frac{p_2}{q_2} > 0$. We then find $(p_1q_2)^p + (p_2q_1)^p = (q_1q_2)^p$. However, by Fermat's Last Theorem, there are no positive integers p_1, p_2, q_1, q_2 that satisfy this relation. Thus, there exist no rational solutions in the first quadrant. Then, by symmetry, we see that the only rational points on the circle are exactly those along the axes.

7 Future Research

The results in this paper seem to indicate that p-trigonometric functions have interesting but complex behavior. For instance, even basic formulas such as the double-angle formulas for $\sin(2x)$ and $\cos(2x)$ seem to have no straightforward generalization. Similarly, understanding higher derivatives of $\sin_p(x)$ at $x = 0$ looks very difficult; see Conjecture 16 and the questions following it. There are several other open questions. We list a few that we think merit further study.

1. We know the derivatives of $\sin_p x$ and $\cos_p x$. What about $\int \sin_p x\,dx$ and $\int \cos_p x\,dx$? Using the Taylor series for $\sin_p x$ and $\cos_p x$, one can evaluate these integrals as series. But are there closed-form answers for these integrals?

2. The parametrization of p-circles we considered in this paper are with respect to area. We can also parametrize these curves with respect to arc length. These give yet another generalization of the p-trigonometric functions. What properties do these functions have?

3. Can we extend this work for (p,q)-trigonometric functions that come from looking at the curves $|x|^p + |y|^q = 1$? Parametrizing these curves will give us $\sin_{p,q} x$ and $\cos_{p,q} x$. What can be said about these functions?

4. So far, we have been working in $\mathbb{R}^2$. Can we extend this work to $\mathbb{R}^3$? To this end, we should look at the unit p-sphere $|x|^p + |y|^p + |z|^p = 1$. For $p = 2$, this is the standard unit sphere, and as p goes to infinity, we get a cube that encloses the unit sphere. These surfaces can be called sphubes (p-spheres), analogous to our squircles (p-circles). It opens gates to a whole new area of research. What are the parametric equations of these surfaces? Can we do sphubical trigonometry that is similar to spherical trigonometry? What are the volume and surface areas of the regions enclosed by these surfaces? What is the Gaussian curvature function of these surfaces?

Acknowledgement

This paper is the outcome of the MAT 268 (Introduction to Undergraduate Research in Mathematics) course taught by the first author to the remaining authors at Illinois State University in Spring 2021. We want to thank the department of mathematics for providing us with the necessary resources for this research. Discussions with Anindya Sen led to the notion of rigidity in Section 2. Pisheng Ding raised several interesting questions and comments after reading this paper. We thank both of them for their interest and input in this paper. Finally, we are grateful to an anonymous referee for many comments and suggestions.

Bibliography

[1] Edmunds, David E. and Gurka, Petr and Lang, Jan. Properties of generalized trigonometric functions. Journal of Approximation Theory, 164, 47-56 (2012).

[2] Blanchard, P. and Devaney, Robert L. and Hall, Glen R.. Differential Equations. Brooks/Cole, 3rd edition, 2005.

[3] Levin, A.. A Geometric Interpretation of an Infinite Product for the Lemniscate Constant. The American Mathematical Monthly, 113(6), 510-520 (2006). https://doi.org/10.2307/27641976.

[4] Peter Lynch. Squircles. ThatsMaths, (2016) 7. https://thatsmaths.com/2016/07/14/squircles/.

[5] A. I. Markushevich. Theory of functions of a complex variable. Vol. II. Prentice-Hall, Inc., Eaglewood Cliffs, N.J., 1965. Revised English edition translated and edited by Richard A. Silverman.

[6] Neugebauer, O. The Exact Sciences in Antiquity. Dover Publications (1969).

[7] Wood, William E. and Poodiack, Robert D. Squigonometry: Trigonometry in the p-norm. A project-Based Guide to Undergraduate Research in Mathematics, Birkhauser, ISBN 2520-1212, 263-286.

[8] Sato, Shota and Takeuchi, Shingo. Two double-angle formulas of generalized trigonometric functions. Journal of Approximation Theory, 250, 105322 (2020). https://medium.com/minimal-notes/rounded-corners-in-the-apple-ecosystem-1b3f45e18fcc.

[9] Saxe, Karen. Beginning functional analysis. Undergraduate Texts in Mathematics. Springer-Verlag, New York, 2002.

[10] Arthur Van Siclen. Rounded Corners in the Apple Ecosystem. Medium, (2020) 6.

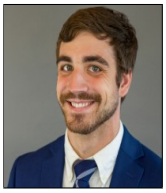
[11] Silverman, Joseph H. A Friendly Introduction to Number Theory. Pearson Education, Inc. (1997).

[12] James Stewart. Calculus Early Transcendentals, 6e, Thomson Brooks/Cole, ISBN 978-0-495-01166-8 (2008) 742.

Ball State Undergraduate Mathematics Exchange
`https://digitalresearch.bsu.edu/mathexchange`
Vol. 16, No. 1 (Fall 2022)
Pages 23 – 35

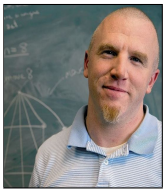
Vertex and Mixed k-Diameter Component Connectivity

*Adam Buzzard, Nathan Shank**



Adam Buzzard graduated from Moraivan College in 2016 with a bachelor's degree in Mathematics. This work was part of a summer research project in 2015 which led into his senior Honors thesis. After graduation, Adam worked with Lutron Electronics.

Nathan Shank is a Professor of Mathematics at Moravian University in Bethlehem Pennsylvania. His research interests lie at the interaction of graph theory and probability theory and he is particularly interested in theoretical applications to network reliability.



Abstract

In the k-diameter component connectivity model a network is consider operational if there is a component with diameter at least k. Therefore, a network is in a failure state if every component has diameter less than k. In this paper we find the vertex variant of the k-diameter component connectivity parameter, which is the minimum number of vertex deletions in order to put a network into a failure state, for particular classes of graphs. We also show the mixed variant by allowing vertex and edge failures within the network. We show results for paths, cycles, complete, and complete bipartite graphs for both variants as well as perfect r-ary trees for the vertex variant.

1 Introduction

Many different network structures can be modeled through graph theory. We think of nodes, hubs, people, stations, objects, etc. as vertices of a graph and the communication or connection between them as the edges of the graph. For many reasons, accidental

*__Corresponding author__:shankn@moravian.edu

and deliberate, these networks break or fail. Therefore, understanding the reliability and vulnerability of a network is crucial to maintaining and building a reliable network.

When considering network reliability there are two issues to understand: what are the minimum requirements to maintain an operational network and what pieces of the network may fail. In a network modeled as a graph we can have edges, vertices, or both (mixed) fail. To keep a network operational we often consider what characterizes a failure state for a network. Several different network reliability models have been studied and Harary [6] provided the general framework for these network reliability models by considering a property P and a network, G. He defined a network to be operational if there is a component of G which has property P and therefore, if no component of G has property P the network is in a failure state. Therefore, the network reliability of G based on property P is the minimum number of failures so that no component of G has property P.

For example, the *component order edge connectivity* [2] considers the minimum number of edges that need to be removed from a graph so that all the components have order less that some specific bound. Similarly the *component order vertex connectivity* [3] considers vertex removal. For a survey of results see [4].

In this paper we will consider a graph operational if it contains a component of diameter at least k for some fixed positive integer k. In practice, a network may need to have a component with a minimum diameter for reliability testing of a particular function, the spread of information must travel a minimum distance before it is deemed valuable, or a virus which is transmitted to neighbors may stay dormant until it has passed through a specific number of hosts or nodes. These, and other examples, motivate the need to consider networks which contain a component of minimum diameter. In [5], the authors considered the instance where edges fail, or are removed. In this paper we consider vertex failure as well as the mixed failure case for certain graph classes.

2 Background and Definitions

We will be using common graph theory notation found in [7]. Throughout we will assume that $G = (V,E)$ is a finite simple graph with vertex set V and edge set E. For any edge set $D \subseteq E$, let $G - D$ denote the spanning subgraph of G containing the vertex set V and the edge set $E - D$. For any vertex set $H \subseteq V$, let $G - H$ denote the subgraph of G induced by $V - H$. Similarly, if $V' \subseteq V$ and $E' \subseteq E(G - V')$ we will write $G - V' - E'$ to denote $(G - V') - E'$. For any set A, let $|A|$ denote the cardinality of A.

If $u, v \in V$, let $d_G(u,v)$ denote the distance between u and v in G (length of the shortest $u - v$ path in G). Note that if u and v are not connected (there is no $u - v$ path), we will define $d_G(u,v) = \infty$. If the graph G is clear, we will denote $d_G(u,v) = d(u,v)$. If $d(u,v) = k$ for some positive integer k, then we will say that u is a *k-neighbor* of v, $\{u,v\}$ a *$k-pair$*, and a $u - v$ path of length k will be a *k-path*. If there exists a vertex $x \in V$ so that $d(u,x) = k$, then we will say that *u has a $k-neighbor$* in G. If G is a connected graph then the *diameter of G* is the maximum distance between any two

vertices. If G is not connected, the diameter is defined to be infinite. A component of a graph G is a connected and induced subgraph of G, call it H, so that no other vertex in $v \in V(G-H)$ is adjacent to a vertex in H. Clearly, if G has more than one component, then the diameter of G is infinite.

Throughout we will consider k to be a positive integer and we will consider a network to be *operational* if there is a component of diameter at least k. Thus, a network is in a *failure state* if every component has diameter less than k. We can easily render a network into a failure state by removing all of the vertices or all of the edges. However, we are interested in finding the minimum number of vertex or edge deletions to produce a failure state.

Clearly, if $k \geq 2$ and if a graph contains a k-pair, then the graph also contains a $(k-1)$-pair. This leads to the following lemma which will be useful for our considerations of vertex and edge connectivity.

Lemma 1. *Let $G = (V,E)$ be a graph and k be a positive integer. The graph G is in a failure state if and only if there does not exist a k-pair in G.*

Thus, in the process of making a failure state we must remove vertices or edges that impact each k-path. The following lemma shows that vertex disjoint k-paths can not both be impacted by only one vertex or edge removal. This will help us to show bounds on the number of vertex or edge removals we need to render our network into a failure state.

Lemma 2. *Let $G = (V,E)$ be a graph and k be a positive integer. If there exists m vertex disjoint k-paths in G, then for any $v \in V$ or $e \in E$, $G-\{v\}$ and $G-\{e\}$ each have at least $m-1$ vertex disjoint k-paths.*

Proof. Let $G = (V,E)$ be a graph and k be a positive integer. Assume G has m vertex disjoint k-paths. Assume by way of contradiction that there exists some $v \in V$ such that $G-\{v\}$ has less than $m-1$ vertex disjoint k-paths. Then v was a vertex in at least two of the vertex disjoint k-paths. Hence, G did not contain m vertex disjoint k-paths. A similar argument holds for edge removals. □

It is often the case that edges are the object that fails. The *k-diameter component edge connectivity* of the graph was introduced in [5] and results were shown for path, complete, and complete bipartite graphs as well as perfect r-ary trees. In this paper we focus on the *k-diameter component vertex connectivity* parameter as well as the mixed parameter which allows vertex and edge deletions.

Definition 3. Let $G = (V,E)$ be a graph and k be a positive integer. A set $V' \subseteq V$ is a *k-diameter component vertex disconnecting set* if $G-V'$ has no vertex with a k-neighbor.

This means that a vertex set V' is a k-diameter component vertex disconnecting set if every component of $G-V'$ has diameter less than k. If V' is a k-diameter component vertex disconnecting set, then $G-V'$ is in a failure state.

Recall we are interested in finding the minimum number of vertices that can be removed to produce a failure state. This motivates the definition of the *k-diameter component vertex connectivity parameter*.

Definition 4. Given a graph $G = (V,E)$ and a positive integer k, the *k-diameter component vertex connectivity parameter* of G, denoted $CV_k(G)$, is the size of the smallest k-diameter component vertex disconnecting set.

Thus, the k-diameter component vertex connectivity parameter is the size of the smallest vertex set V' so that $G-V'$ is in a failure state.

Similarly, we will consider edge disconnecting sets which will be used in our mixed deletion case.

Definition 5. Let $G = (V,E)$ be a graph and k be a positive integer. A set $E' \subseteq E$ is a *k-diameter component edge disconnecting set* if $G-E'$ has no vertex with a k-neighbor.

This means that an edge set E' is a k-diameter component edge disconnecting set if every component of $G-E'$ has diameter less than k. If E' is ak-diameter component edge disconnecting set, then $G-E'$ is in a failure state.

As with vertex deletions, we can also consider the minim number of edges whose removal produces a failure state. This motivates the definition of the *k-diameter component edge connectivity parameter*.

Definition 6. Given a graph $G = (V,E)$ and a positive integer k, the *k-diameter component edge connectivity parameter* of G, denoted $CE_k(G)$, is the size of the smallest k-diameter component edge disconnecting set.

Thus, the k-diameter component edge connectivity parameter is the minimum size of an edge set E' so that $G-E'$ is in a failure state.

It is often the case that vertices and edges fail which is investigated through the mixed deletion case. This was first introduced by Beineke and Harary [1]. The following definitions address the k-diameter component connectivity function, which is a mixed version of the k-diameter component connectivity involving both vertex and edge deletions. As is standard we will remove vertices first then remove edges.

Definition 7. Let $G = (V,E)$ be a graph, k be a positive integer, and $p \in \{0,1,...,CV_k(G)\}$. Then the *$k$-diameter component connectivity function* of G is defined as $CM_k(G,p) = \min\{CE_k(G-V') : V' \subseteq V, |V'| = p\}$.

So $CM_k(G,p)$ is the minimum number of edges that must be removed to render the graph into a failure state assuming we can also remove any p vertices in the graph. Note that we must remove the p vertices first then remove the least amount of edges.

Definition 8. Let $G = (V,E)$ be a graph and k be a positive integer. A *k-diameter component connectivity pair* of G for each $p \in \{0,1,...,CV_k(G)\}$ is an ordered pair (p,q), such that $CM_k(G,p) = q$.

Two obvious connectivity pairs of G are $(0,CE_k(G))$ and $(CV_k(G),0)$. For each value of p where $0 \leq p \leq CV_k(G)$, there is a unique k-diameter component connectivity pair.

When trying to find the value of $CM_k(G,p)$ it is often useful to consider which p vertices we need to remove to minimize $CE_k(G-V')$. This motivates the following definition of a *optimal p-set*.

Definition 9. Let $G = (V,E)$ be a graph, k be a positive integer, and p be a nonnegative integer. Let $V' \subseteq V$ such that $|V'| = p$. We say V' is an *optimal p-set* if $CM_k(G,p) = CE_k(G-V')$.

The following lemma will prove valuable for providing lower bounds. The lemma shows disjoint k-paths provide a lower bound for the number of vertex deletions, edge deletions, or mixed deletions needed to produce a failure state.

Lemma 10. *Let $G = (V,E)$ be a graph and let k be a positive integer. If there exists M vertex disjoint k-paths in G, then $CV_k(G) \geq M$ and $CE_k(G) \geq M$. Furthermore, if $CM_k(G,p) = q$, then $p+q \geq M$.*

Proof. Let $G = (V,E)$ be a graph and let k be a positive integer. Assume there exists M vertex disjoint k-paths in G. Let $V' \subseteq V$ such that $G-V'$ is in a failure state. Let $E' \subseteq E$ such that $G-E'$ is in a failure state. Then Lemma 1 and multiple iterations of Lemma 2 implies that $|V'| \geq M$ and $|E'| \geq M$.

Let $V^* \subseteq V$ and $E^* \subseteq E(G-V^*)$ such that $G-V^*-E^*$ is in a failure state. Then Lemma 1 and Lemma 2 implies that $|V^*|+|E^*| \geq M$. Hence, if $CM_k(G,p) = q$, then $p+q \geq M$.

□

3 Vertex Deletion Results

In this section we will consider only vertex deletions; we compute $CV_k(G)$ for specific graphs G. We provide results for path graphs, cycles, complete graphs, complete bipartite graphs, and perfect r-ary trees. Note that if $k = 1$, then CV_k is the minimum number of vertex deletions whose removal results in an edgeless graph. Therefore, we will always assume $k \geq 2$.

3.1 Path Graphs

Consider the path graph on n vertices, denoted P_n. Label the vertices consecutively from 1 to n starting at a pendant vertex. Since any path of length k has $k+1$ vertices, there are $\lfloor \frac{n}{k+1} \rfloor$ vertex disjoint k-paths in P_n. By Lemma 10, $CV_k(P_n) \geq \lfloor \frac{n}{k+1} \rfloor$.

If we delete every vertex whose label is a multiple of $k+1$, then all of the remaining components have k vertices, with the exception of at most one component which could have fewer than k vertices. Therefore, the diameter of each remaining component will be less than k. This results in a total of $\lfloor \frac{n}{k+1} \rfloor$ deletions. Hence, $CV_k(P_n) \leq \lfloor \frac{n}{k+1} \rfloor$. These two observations imply the following:

Theorem 11. *For every positive integer n,*

$$CV_k(P_n) = \left\lfloor \frac{n}{k+1} \right\rfloor.$$

3.2 Cycle Graphs

Consider the cycle graph on n vertices, denoted C_n. Since $diam(C_n) = \lfloor \frac{n}{2} \rfloor$, if $k > \lfloor \frac{n}{2} \rfloor$, then C_n is already in a failure state and no deletions are necessary. If $k \leq \lfloor \frac{n}{2} \rfloor$, then at least one deletion must be made. Notice that the deletion of any single vertex from

C_n leaves a path graph on $n-1$ vertices. Then, by Theorem 11, $CV_k(C_n) = \left\lfloor \frac{n-1}{k+1} \right\rfloor + 1$. Hence, we have the following:

Theorem 12. *For every positive integer n,*

$$CV_k(C_n) = \begin{cases} 0 & \text{if } k > \left\lfloor \frac{n}{2} \right\rfloor \\ \left\lfloor \frac{n+k}{k+1} \right\rfloor & \text{if } k \leq \left\lfloor \frac{n}{2} \right\rfloor. \end{cases}$$

3.3 Complete Graphs

Consider the complete graph on n vertices, denoted K_n. Since the diameter of a complete graph is 1 and $k \geq 2$, K_n is already in a failure state. Thus, we see the following obvious result:

Theorem 13. *For any positive integer n,*

$$CV_k(K_n) = 0.$$

3.4 Complete Bipartite Graph

Now we will consider a complete bipartite graph $K_{a,b} = (V,E)$ with parts A and B where $V = A \cup B$, $A \cap B = \emptyset$, $|A| = a > 0$, and $|B| = b > 0$.

Theorem 14. *For any positive integer a and b,*

$$CV_k(K_{a,b}) = \begin{cases} 0 & \text{if } a = b = 1, \\ 0 & \text{if } k > 2 \text{ and } \max\{a,b\} \geq 2, \\ min\{a,b\} & \text{if } k = 2 \text{ and } \max\{a,b\} \geq 2. \end{cases}$$

Proof. Let $K_{a,b} = (V,E)$ be a complete bipartite graph with parts A and B where $V = A \cup B$, $A \cap B = \emptyset$, $|A| = a > 0$, and $|B| = b > 0$. The diameter of a complete bipartite graph is 2 unless $a = b = 1$, in which case the diameter is 1 and $K_{1,1}$ is already in a failure state for all $k > 1$. Consider when $a \geq 2$ or $b \geq 2$. If $k > 2$, then $K_{a,b}$ is already in a failure state. Now consider when $k = 2$. The only induced subgraphs of $K_{a,b}$ which are in a failure state are $K_{1,1}$, subgraphs of A, and subgraphs of B. To produce $K_{1,1}$, we must delete all but two vertices: one vertex from A and one vertex from B. Thus, the resulting number of vertex deletions is $(a-1)+(b-1)$. Since A is the subgraph of A with the most vertices, we only need to consider deleting vertices to produce A. In order to produce A, we must delete all vertices from B and, therefore, the resulting number of vertex deletions is b. Similarly, to produce B, a vertex deletions are necessary. If either $a \geq 2$ or $b \geq 2$, then it is easily seen that $(a-1)+(b-1) \geq \min\{a,b\}$. □

3.5 Perfect r-ary Trees

Throughout we will assume r and l are positive integers and let $T_{r,l} = (V,E)$ be a perfect $r-ary$ tree with height l. This means that $T_{r,l}$ has $\frac{r^{l+1}-1}{r-1}$ vertices and $\frac{r(r^{l+1}-1)}{r-1} - r$

edges. We can enumerate the vertices and edges of $T_{r,l}$ as follows:

$$V = \{v_{i,j} : 1 \leq i \leq l+1, 1 \leq j \leq r^{(l+1)-i}\}, \text{ and}$$
$$E = \{(v_{i,j}, v_{i-1,m}) : 2 \leq i \leq l+1, 1 \leq j \leq r^{(l+1)-I},$$
$$(j-1)r+1 \leq m \leq jr\}.$$

We will say that vertex $v_{i,j} \in V$ is on *level i*. Notice that the root vertex is on level $l+1$ and the leaves are on level 1.

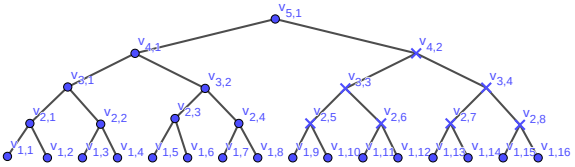


Figure 1: A perfect 2-ary tree with height 3 ($r = 2, l = 4$) with vertices in $T^2_{v_{4,2}}$ indicated by $\times$.

Let h be a positive integer. Fix a vertex $v_{i,j} \in V$, with $i > h$ and consider the subtree $T^h_{v_{i,j}}$ of $T_{r,l}$ induced by $v_{i,j}$ and all of its descendants at a distance at most h. Notice in $T^h_{v_{i,j}}$, the degree of $v_{i,j}$ is r and any vertex x of degree 1 in $T^h_{v_{i,j}}$ will satisfy $d(x, v_{i,j}) = h$. Also notice that $T^h_{v_{i,j}}$ is a perfect r-ary tree of height h.

The following lemma establishes a set V' of vertices which will form our minimum k-diameter component vertex disconnecting set for $T_{r,l}$. The cardinality of this set is shown so that we can prove it is in fact the minimum k-diameter component vertex disconnecting set.

Lemma 15. *Let $T_{r,l} = (V,E)$. Let $V' \subset V$ such that*

$$V' = \left\{ v_{m(\lceil \frac{k}{2} \rceil + 1), j} \in V : 1 \leq m \leq \left\lfloor \frac{l+1}{\lceil \frac{k}{2} \rceil + 1} \right\rfloor, 1 \leq j \leq r^{l+1-m(\lceil \frac{k}{2} \rceil + 1)} \right\}.$$

Then,

$$|V'| = \frac{r^{l+1} - r^{l+1-\left\lfloor \frac{l+1}{\lceil \frac{k}{2} \rceil + 1} \right\rfloor (\lceil \frac{k}{2} \rceil + 1)}}{r^{\lceil \frac{k}{2} \rceil + 1} - 1}.$$

Proof. Consider V' as defined above. Then by summing over all possible choices of m we see

$$|V'| = \sum_{m=1}^{\left\lfloor \frac{l+1}{\lceil \frac{k}{2} \rceil + 1} \right\rfloor} r^{l+1-m(\lceil \frac{k}{2} \rceil + 1)}.$$

Simplifying the previous expression, we see:

$$
\begin{aligned}
|V'| &= r^{l+1} \sum_{m=1}^{\left\lfloor \frac{l+1}{\lceil \frac{k}{2} \rceil + 1} \right\rfloor} r^{-m\left(\lceil \frac{k}{2} \rceil + 1\right)} \\
&= r^{l+1} \left(\frac{1 - r^{-\left\lfloor \frac{l+1}{\lceil \frac{k}{2} \rceil + 1} \right\rfloor \left(\lceil \frac{k}{2} \rceil + 1\right)}}{r^{\lceil \frac{k}{2} \rceil + 1} - 1} \right) \\
&= \frac{r^{l+1} - r^{l+1-\left\lfloor \frac{l+1}{\lceil \frac{k}{2} \rceil + 1} \right\rfloor \left(\lceil \frac{k}{2} \rceil + 1\right)}}{r^{\lceil \frac{k}{2} \rceil + 1} - 1}.
\end{aligned}
$$

□

Now we will show $CV_k(T_{r,l}) = |V'|$. For each $v_{i,j} \in V'$, $T_{v_{i,j}}^{\lceil \frac{k}{2} \rceil}$ is not in a failure state. We will also show that for any $v_{i,j}, v_{i',j'} \in V'$ with $v_{i,j} \neq v_{i',j'}$, $T_{v_{i,j}}^{\lceil \frac{k}{2} \rceil}$ and $T_{v_{i',j'}}^{\lceil \frac{k}{2} \rceil}$ are disjoint. Thus, for each vertex in V', we need at least one vertex deletion to produce a subgraph of $T_{r,l}$ which is in a failure state. We will also show that $T_{r,l} - V'$ is in a failure state and, therefore, V' forms a minimum k-diameter disconnecting set. For the sake of simplicity, we will denote $T_{v_{i,j}}^{\lceil \frac{k}{2} \rceil}$ with $T_{v_{i,j}}$.

Theorem 16. *Let r, l and k be positive integers. Let $T_{r,l} = (V, E)$ be a perfect $r-$ary tree with height l. Then,*

$$
CV_k(T_{r,l}) = \frac{r^{l+1} - r^{l+1-\left\lfloor \frac{l+1}{\lceil \frac{k}{2} \rceil + 1} \right\rfloor \left(\lceil \frac{k}{2} \rceil + 1\right)}}{r^{\lceil \frac{k}{2} \rceil + 1} - 1}.
$$

Proof. Let r, l, and k be positive integers. Let $T_{r,l} = (V, E)$.

First we will establish that

$$
CV_k(T_{r,l}) \geq \frac{r^{l+1} - r^{l+1-\left\lfloor \frac{l+1}{\lceil \frac{k}{2} \rceil + 1} \right\rfloor \left(\lceil \frac{k}{2} \rceil + 1\right)}}{r^{\lceil \frac{k}{2} \rceil + 1} - 1}
$$

by finding a set of disjoint k-pairs. Let

$$
V' = \left\{ v_{m\left(\lceil \frac{k}{2} \rceil + 1\right), j} \in V : 1 \leq m \leq \left\lfloor \frac{l+1}{\lceil \frac{k}{2} \rceil + 1} \right\rfloor, 1 \leq j \leq r^{l+1-m\left(\lceil \frac{k}{2} \rceil + 1\right)} \right\}.
$$

For each $v_{i,j} \in V'$ let $T_{v_{i,j}}$ be the subgraph induced on $v_{i,j}$ and all of its descendants at a distance at most $\lceil \frac{k}{2} \rceil$. Notice $i > \lceil \frac{k}{2} \rceil$ for each $v_{i,j} \in V'$ and $T_{v_{i,j}}$ is a perfect r-ary tree with height $\lceil \frac{k}{2} \rceil$ for all $v_{i,j} \in V'$. Since the diameter of a perfect r-ary tree of height a is $2a$, we have $diam(T_{v_{i,j}}) = 2\lceil \frac{k}{2} \rceil$ for each $v_{i,j} \in V'$. This implies, since k is an integer,

$$
k + 1 \geq diam(T_{v_{i,j}}) \geq k.
$$

Since $diam(T_{v_{i,j}}) \geq k$, $T_{v_{i,j}}$ contains at least one k-pair and, thus, is not in a failure

state.

Let $v_{i,j}, v_{i,j'} \in V'$ with $j \neq j'$. We will show that all $T_{v_{i,j}}$ and $T_{v_{i,j'}}$ are disjoint. Since $v_{i,j}$ and $v_{i,j'}$ have a common ancestor, they cannot share any decendants, or $T_{r,l}$ would contain a cycle and is not a tree. Therefore, $T_{v_{i,j}}$ and $T_{v_{i,j'}}$ are disjoint.

Consider $v_{i,j}, v_{i',j'} \in V'$ where $i \neq i'$. We will show that $T_{v_{i,j}}$ and $T_{v_{i',j'}}$ are disjoint. By the definition of V', vertices on different levels in V' are at a distance of at least $\lceil \frac{k}{2} \rceil + 1$ from each other, so $v_{i,j} \notin T_{v_{i',j'}}$ and $v_{i',j'} \notin T_{v_{i,j}}$. Since $v_{i,j}$ and $v_{i',j'}$ are the roots of $T_{v_{i,j}}$ and $T_{v_{i',j'}}$ respectively, and $T_{v_{i,j}}$ and $T_{v_{i',j'}}$ are trees of height $\lceil \frac{k}{2} \rceil$, this implies $T_{v_{i,j}}$ and $T_{v_{i,j'}}$ are disjoint.

Consider $v_{i,j}, v_{i',j'} \in V'$ where $i \neq i'$. We will show that $T_{v_{i,j}}$ and $T_{v_{i',j'}}$ are disjoint. Without loss of generality, assume $i' < i$. By the definition of V', vertices on different levels in V' are at a distance of at least $\lceil \frac{k}{2} \rceil + 1$ from each other. So if, $v_{i',j'}$ is a descendent of $v_{i,j}$, then $T_{v_{i',j'}}$ and $T_{v_{i,j}}$ will be disjoint since $T_{v_{i,j}}$ has height $\lceil \frac{k}{2} \rceil$. If $v_{i',j'}$ is not a descendent of $v_{i,j}$, then there exists a $v_{i,j} - v_{i',j'}$ path which goes through a vertex $v_{a,b}$ with $a > i$ which is not in $T_{v_{i,j}}$ or $T_{v_{i',j'}}$. Therefore, $T_{v_{i,j}}$ and $T_{v_{i',j'}}$ must be disjoint since trees are acyclic.

Therefore, we have shown that $\{T_{v_{i,j}} : v_{i,j} \in V'\}$ are pairwise disjoint.

Since $T_{v_{i,j}}$ contains a k-pair for each $v_{i,j} \in V'$, there are at least $|V'|$ disjoint k-pairs in $T_{r,l}$. Therefore, by Corollary 10 and Lemma 15,

$$CV_k(T_{r,l}) \geq \frac{r^{l+1} - r^{l+1-\left\lfloor \frac{l+1}{\lceil \frac{k}{2} \rceil + 1} \right\rfloor (\lceil \frac{k}{2} \rceil + 1)}}{r^{\lceil \frac{k}{2} \rceil + 1} - 1}.$$

Now we will show that

$$CV_k(T_{r,l}) \leq \frac{r^{l+1} - r^{l+1-\left\lfloor \frac{l+1}{\lceil \frac{k}{2} \rceil + 1} \right\rfloor (\lceil \frac{k}{2} \rceil + 1)}}{r^{\lceil \frac{k}{2} \rceil + 1} - 1}$$

by showing $T_{r,l} - V'$ is in a failure state.

Consider $T_{r,l} - V'$. By deleting all vertices in V' from $T_{r,l}$, we are deleting entire levels of vertices. In fact, we are deleting all vertices which are on a level which is an integer multiple of $\lceil \frac{k}{2} \rceil + 1$. Notice, then, that $T_{r,l} - V'$ is a disconnected graph where each component is a perfect r-ary tree. All of these trees have height $\lceil \frac{k}{2} \rceil - 1$, except the tree containing the root vertex, which could have less. Then, $diam(T_{r,\lceil \frac{k}{2} \rceil - 1}) = 2(\lceil \frac{k}{2} \rceil - 1) \leq 2(\frac{k+1}{2}) - 2 \leq k - 1$. Therefore, each component has diameter less than k, and $G - V'$ is in a failure state. Hence,

$$CV_k(T_{r,l}) \leq \frac{r^{l+1} - r^{l+1-\left\lfloor \frac{l+1}{\lceil \frac{k}{2} \rceil + 1} \right\rfloor (\lceil \frac{k}{2} \rceil + 1)}}{r^{\lceil \frac{k}{2} \rceil + 1} - 1}.$$

Combining these two inequalities, we see that

$$CV_k(T_{r,l}) = \frac{r^{l+1} - r^{l+1-\left\lfloor \frac{l+1}{\lceil \frac{k}{2} \rceil + 1} \right\rfloor \left(\lceil \frac{k}{2} \rceil + 1 \right)}}{r^{\lceil \frac{k}{2} \rceil + 1} - 1}.$$

□

4 Mixed Deletion Results

Now we will investigate the k-diameter component connectivity function of a few simple graph classes. We provide results for path graphs, cycles, complete graphs, and complete bipartite graphs. As with the previous section, we will assume throughout that $k \geq 2$ is a positive integer.

4.1 Path Graphs

To decompose path graphs into failure states, we will create path graphs components of maximum length which are in a failure state.

Theorem 17. *Let P_n be the path on n vertices. For any nonnegative integer $p \leq CV_k(P_n)$,*

$$CM_k(P_n, p) = \begin{cases} 0 & \text{if } p = \left\lfloor \frac{n}{k+1} \right\rfloor \\ \left\lfloor \frac{n-p(k+1)-1}{k} \right\rfloor & \text{if } p < \left\lfloor \frac{n}{k+1} \right\rfloor. \end{cases}$$

Proof. Let n be a positive integer. Consider P_n and let $p \leq CV_k(P_n)$ be a nonnegative integer. If $p = \left\lfloor \frac{n}{k+1} \right\rfloor$, then by Theorem 11, $CM_k(P_n, p) = 0$.

Assume $p < \left\lfloor \frac{n}{k+1} \right\rfloor$. Label vertices of P_n consecutively starting at a pendant vertex so that $V = \{v_i : i \in \mathbb{Z}, 1 \leq i \leq n\}$. Let $V' \subset V$ such that $V' = \{v_{j(k+1)} : j \in \mathbb{Z}, 1 \leq j \leq p\}$. Then $P_n - V'$ has p components with diameter $k-1$ and one component, denoted C, which is a path on $n - p(k+1)$ vertices. By Theorem 3.1 in [5], $CE_k(C) = \left\lfloor \frac{n-p(k+1)-1}{k} \right\rfloor$. Let E' be the set of $\left\lfloor \frac{n-p(k+1)-1}{k} \right\rfloor$ edges removed from component C to produce a failure state. Then $P_n - V' - E'$ is in a failure state. Therefore, $CM_k(P_n, p) \leq \left\lfloor \frac{n-p(k+1)-1}{k} \right\rfloor$.

Assume we are going to delete a set, P, of vertices where $|P| = p$ and a set, Q, of edges where $|Q| = q$ so that $P_n - P - Q$ is in a failure state. Since P_n is a tree, any edge or vertex deletion creates at most one new component. Therefore, $P_n - P - Q$ can have at most $p + q + 1$ components. If $P_n - P - Q$ is in a failure state, then each component contains at most k vertices. Thus, the number of vertices in our original graph is the sum of all the vertices in the failed components plus the p vertices we had to delete from P. Therefore, the number of vertices in our original graph would be bounded above by $(p+q+1)k + p$. We will now use this fact to show that $CM_k(P_n, p) = \left\lfloor \frac{n-p(k+1)-1}{k} \right\rfloor$.

Assume by way of contradiction that there exists some set $V^* \subseteq V$ with $|V^*| = p$ and $E^* \subseteq E$ with $|E^*| = \left\lfloor \frac{n-p(k+1)-1}{k} \right\rfloor - 1$ such that $P_n - V^* - E^*$ is in a failure state. As shown above, if $P_n - V^* - E^*$ is in a failure state, then the number of vertices in P_n is bounded above by

$$|V(P_n)| \leq \left(p + \left(\left\lfloor \frac{n-p(k+1)-1}{k} \right\rfloor - 1 \right) + 1 \right) k + p.$$

Simplifying this expression, we see

$$\begin{aligned} |V(P_n)| &\leq p(k+1) + k \left\lfloor \frac{n-p(k+1)-1}{k} \right\rfloor \\ &\leq p(k+1) + k \left(\frac{n-p(k+1)-1}{k} \right) \\ &= n-1. \end{aligned}$$

Hence, we have a contradiction, because $|V(P_n)| = n$.

□

4.2 Cycle Graphs

A cycle is only one vertex deletion away from becoming a path graph. Therefore, we can use our results for path graphs to analyze cycle graphs.

Theorem 18. *Consider C_n, the cycle on n vertices. For any positive integer n and nonnegative integer $p \leq CV_k(C_n)$,*

$$CM_k(C_n, p) = \begin{cases} 0 & \text{if } k > \left\lfloor \frac{n}{2} \right\rfloor \\ 0 & \text{if } k \leq \left\lfloor \frac{n}{2} \right\rfloor \text{ and } p = \left\lfloor \frac{n+k}{k+1} \right\rfloor \\ \left\lfloor \frac{n-p(k+1)-1}{k} \right\rfloor + 1 & \text{if } k \leq \left\lfloor \frac{n}{2} \right\rfloor \text{ and } 0 < p < \left\lfloor \frac{n+k}{k+1} \right\rfloor. \end{cases}$$

Proof. Let C_n be the cycle graph on n vertices. Note that $diam(C_n) = \left\lfloor \frac{n}{2} \right\rfloor$, so if $k > \left\lfloor \frac{n}{2} \right\rfloor$, C_n is already in a failure state and requires no deletions. Consider when $p = \left\lfloor \frac{n+k}{k+1} \right\rfloor$. Then by Theorem 12, there exists a set V' of p vertices such that $C_n - V'$ is in a failure state. Hence, if $p = \left\lfloor \frac{n+k}{k+1} \right\rfloor$, then $CM_k(C_n, p) = 0$.

Consider when $k \leq \left\lfloor \frac{n}{2} \right\rfloor$ and $0 < p < \left\lfloor \frac{n+k}{k+1} \right\rfloor$. Note that any vertex deletion leaves a path on $n-1$ vertices, hence, $CM_k(C_n, p) = CM_k(P_{n-1}, p-1)$. Therefore, by Theorem 17 $CM_k(C_n, p) = \left\lfloor \frac{(n-1)-(p-1)(k+1)-1}{k} \right\rfloor$. Simplifying this expression, we see

$$CM_k(C_n, p) = \left\lfloor \frac{n-p(k+1)-1}{k} \right\rfloor + 1.$$

□

4.3 Complete Graphs

Consider the complete graph on n vertices, denoted K_n. Since any complete graph has diameter 1, we see K_n is already in a failure state since we are assuming $k \geq 2$. Thus, we see the following obvious result:

Theorem 19. *For every positive integer n,*

$$CM_k(K_n, p) = 0.$$

4.4 Complete Bipartite Graph

Theorem 20. *Consider the complete bipartite graph $K_{(a,b)}$ where a and b are positive integers with $a \leq b$. Then for any nonnegative integer $p \leq CV_k(K_{a,b})$,*

$$CM_k(K_{a,b}, p) = \begin{cases} 0 & \text{if } a = b = 1 \\ 0 & \text{if } k > 2 \text{ and } b > 1 \\ (a-p)(b-1) & \text{if } k = 2 \text{ and } b > 1. \end{cases}$$

Proof. Let k be a positive integer. Consider the complete bipartite graph, denoted $K_{a,b} = (V,E)$ with parts A and B where $V = A \cup B$, $A \cap B = \emptyset$, $|A| = a > 0$, and $|B| = b > 0$. Furthermore, assume without loss of generality that $a \leq b$.

If $a = b = 1$, then $diam(K_{a,b}) = 1$ and $K_{a,b}$ is in a failure state. If $b \geq 2$, then $diam(K_{a,b}) = 2$. If $k > 2$, then $K_{a,b}$ is in a failure state. If $k = 2$, then $K_{a,b}$ is not in the failure state. Therefore, fix $k = 2$ for the remainder of the proof.

Note that $\min\{a,b\} = a$, so Theorem 14 implies that $CM_k(K_{a,b}, a) = 0$. Thus, assume $p < a$. Let V' be a set of vertices deleted from $K_{a,b}$ such that $|V'| = p$. We claim that $V' \subseteq A$ is an optimal p-set. In other words, it is optimal to delete all p vertices from part A of $K_{a,b}$. Then by, Theorem 3.3 of [5], $CM_k(K_{(a,b)}) = CE_k(K_{a-p,b}) = (a-p)(b-1)$.

Assume by way of contradiction that $V' \subseteq A$ is not an optimal p-set. Then there exists an optimal p-set $V^* = V_A \cup V_B$, where $V_A \subseteq A$, $V_B \subseteq B$, $|V_A| = x$, $|V_B| = y$, $y \geq 1$ and $x + y = p$. In other words, it is optimal to delete some vertices from part A and some from part B of $K_{a,b}$. Then, we have the following two cases.

Case 1: Assume $a - x \leq b - y$. Then if V^* is an optimal p-set, by Theorem 3.3 of [5], $CM_k(K_{a,b}, p) = CE_k(K_{a-x,b-y}) = (a-x)(b-y-1)$.

Note that $y \geq 1$ implies

$$y(a-x) \leq y(b-1).$$

Then, by substituting $y = p - x$ on the right side of this inequality, we see

$$x(b-1) + y(a-x) \leq p(b-1).$$

Adding $a(b-1)$ to both sides of the above inequality and simplifying, we see

$$(a-p)(b-1) \leq (a-x)(b-y-1).$$

Hence, $CE_k(K_{a-p,b}) \leq CE_k(K_{a-x,b-y})$.

Case 2: Assume $b - y < a - x$. Then if V^* is an optimal p-set, by Theorem 3.3 of [5], $CM_k(K_{a,b}, p) = CE_k(K_{a-x,b-y}) = (b-y)(a-x-1)$.

Note that $a \leq b$ implies that $a - x \leq b$. Multiplying by $(y-1)$ and then adding $b(a-x)$ to both sides of this inequality, we see

$$(a-x)b + y(a-x) - (a-x) \leq (a-x)b + y(b-1) - b + y.$$

Simplifying this expression, we see

$$(a-x-y)(b-1) \leq (b-y)(a-x-1).$$

Then substituting $p = x+y$, we see

$$(a-p)(b-1) \leq (b-y)(a-x-1).$$

Hence, $CE_k(K_{a-p,b}) \leq CE_k(K_{a-x,b-y})$.

In either of these two cases, $CE_k(K_{a-p,b}) = (a-p)(b-1) \leq CE_k(K_{a-x,b-y})$. Hence, V' is an optimal p-set and

$$CM_k(K_{a,b},p) = \begin{cases} 0 & \text{if } a=b=1 \\ 0 & \text{if } k>2 \text{ and } b>1 \\ (a-p)(b-1) & \text{if } k=2 \text{ and } b>1. \end{cases}$$

□

Bibliography

[1] L. Beineke and F. Harary. *The connectivity function of a graph*. Mathematika 14 (1967), 197–202.

[2] F. Boesch, D. Gross, W. Kazmierczak, C. Suffel, and A. Suhartomo. *Component order edge connectivity—an introduction*. Proceedings of the Thirty-Seventh Southeastern International Conference on Combinatorics, Graph Theory and Computing - Conger. Numen. 178 (2006), 7–14.

[3] F. Boesch, D. Gross, and C. Suffel, *Component order connectivity*. Proceedings of the Twenty-ninth Southeastern International Conference on Combinatorics. Graph Theory and Computing - Conger. Numen. 131 (1998), 145–155.

[4] F. Boesch, A. Satyanarayana, and C. Suffel, *A survey of some network reliability analysis and synthesis results*. Networks 54 (2009), no. 2, 99–107.

[5] A. Buzzard and N. Shank. *The k-diameter component edge connectivity parameter*. Involve 11 (2018), no. 5, 848–856.

[6] F. Harary. *Conditional connectivity*. Networks 13 (1983), 347–357.

[7] D. West. *Introduction to graph theory*. 2 ed., Prentice Hall, Upper Saddle River, NJ 07458, 2001.

Ball State Undergraduate Mathematics Exchange
`https://digitalresearch.bsu.edu/mathexchange`
Vol. 16, No. 1 (Fall 2022)
Pages 36 – 49

On the Origin of Zombies: A Modeling Approach

*Alisha Kumari, Elijah Reece, Kursad Tosun, Scott Greenhalgh**

Alisha Kumari is an applied physics major on the mechanical engineering track at Siena College. After graduating in 2024, she plans on pursuing a master's degree in mechanical engineering and pursuing a job in this field.

Elijah Reece is an undergrad in biology at Siena College set to graduate in 2022. After obtaining his Bachelor's degree, he plans to enroll in medical school with the goal of becoming a pediatrician. Eli credits his family and his passion for the health of children as his inspiration.

Kursad Tosun is an applied mathematician/statistician, interested in medical research. He received his Ph.D. in mathematics with a focus on probability theory from Southern Illinois University Carbondale. Currently, Kursad Tosun is teaching in the Department of Mathematics at Siena College. He held positions in the Faculty of Medicine at Mugla Sitki Kocman University, Mathematics and Statistics Department at Vassar College, Mathematics Department at Southern Illinois University, and Cancer Institute at West Virginia University.

Scott Greenhalgh is a 5th year Assistant Professor of Mathematics at Siena College. He received his PhD in applied mathematics from the University of Guelph, and now specializes in disease modeling. Outside of academics, Scott loves hockey, pizza, and regularly plays "the floor is lava" with his two-year-old.

***Corresponding author**: sgreenhalgh@siena.edu

Abstract

A zombie apocalypse is one pandemic that would likely be worse than anything humanity has ever seen. However, despite the mechanisms for zombie uprisings in pop culture, it is unknown whether zombies, from an evolutionary point of view, can actually rise from the dead. To provide insight into this unknown, we created a mathematical model that predicts the trajectory of human and zombie populations during a zombie apocalypse. We parametrized our model according to the demographics of the US, the zombie literature, and then conducted an evolutionary invasion analysis to determine conditions that permit the evolution of zombies. Our results indicate a zombie invasion is theoretically possible, provided the ratio of transmission rate to the zombie death rate is sufficiently large. While achieving this ratio is uncommon in nature, the existence of zombie ant fungus illustrates it is possible and thereby suggests that a zombie apocalypse among humans could occur.

1 Introduction

The world is continuously at risk from epidemics, with COVID-19, SARS, and Ebola serving as recent examples of their devastating impacts. As time progresses, diseases capable of starting another pandemic are more than likely to occur [13]. One important potential pandemic that would likely be worse than anything humanity has ever seen is a zombie apocalypse. While this may seem far-fetched for humanity, in South America among other regions, a fungus exists that can turn ants into zombie ants [5], which implies such an outbreak among humans is within the realm of biological possibilities.

What is biologically possible constitutes all species, most of which exhibit enormous diversity of traits [3, 16]. Through examining these traits, specifically the trade-offs between them [16], the direction of evolution can be inferred, which can provide a glimpse as to what may be in store for a species' future. Typically, such an evolution is caused by the occurrence of a rare mutant, or a patient zero in the case of a novel disease [15, 19], which can feature some form of trait advantage in reproductive ability, size, speed, susceptibility to disease, or survival rate, among others.

While patient zero is ubiquitous in many pop culture movies and tv shows as the first individual to become a zombie [24] the mechanism by which the first zombie is created is often relatively unknown. Classically, many possible scenarios lead to the uprising of patient zero, and ultimately a full-blown zombie apocalypse. For instance, consumption of the mutated zombie ant fungi could infect humans, causing them to seek out nutrients by cannibalism, and thereby further spread fungal spores through their saliva [1]. Alternatively, medical experimentation is often a culprit in causing patient zero, with cross-transmission events from monkeys to humans [8, 9], and side-effects of untested vaccines [11] standing as common causes. However, despite these mechanisms for zombie uprisings, and numerous works on modeling zombie outbreaks [12, 20, 21], it is unknown whether zombies, from an evolutionary point of view, can

actually rise from the dead. So, to provide insight into this unknown we created a mathematical model that predicts the trajectory of human and zombie populations during a zombie apocalypse.

Using this mathematical model, we apply stability analysis to estimate the long-term prognosis of the United States, conduct an evolutionary invasion analysis to infer conditions that allow the zombie apocalypse to occur, or invade from another country, and investigate the potential for an endless human-zombie war through Hopf bifurcation analysis. Our main findings show an uprising of zombies requires the fungus to transmit their spores to more than 0.023 humans per day, and would likely lead to an oscillating struggle that decreases over thousands of years between humans and zombies, as both try to overwhelm the other.

2 Methods

To determine the conditions that permit the biological evolution of zombies, we developed a mathematical model of zombie transmission in a human population. We calibrate our model to the demographics of the US and then apply stability [14], evolutionary invasion [17], and Hopf-bifurcation analyzes [17] to inform on the potential outcomes for humanity.

2.1 Mathematical Model

To begin, we created a mathematical model that predicts the long-term population of the US. We then extend the model to include zombies and proceed to investigate the model's behavior.

2.1.1 The Resident System

We first consider a resident system of humans split into two compartments. One compartment represents the population of humans in the United States (N), which we assume is governed by logistic growth and corresponds to the population of susceptible humans, and the second represents the number of deceased humans due to natural causes, which have yet to completely decompose (D). The rates governing the transition between these compartments is given by

$$\begin{aligned} \tfrac{dN}{dt} &= b(1-\tfrac{N}{K})N - \delta N, \\ \tfrac{dD}{dt} &= \delta N - \mu D, \end{aligned} \tag{1}$$

where b is the birth rate, δ is the mortality rate of people in the US during the year 2020, K is the limiting capacity of humans in the US, and μ is the rate at which dead bodies decompose.

Table 1. Parameters, base values, and sources.

Constant	Parameter	Value	Citation
β	Transmission rate	0.35 per day	[18]
b	Birth rate	0.000065753 per day	[2]
δ	Mortality rate	0.00003562 per day	[2]
μ	Decomposition rate of a human body	0.00595 per day	[6]
K	Population capacity	$4,672,507,360$ people	Section 2.2
γ	Zombie death rate	0.01 corpses per day	[7]

2.1.2 The Zombie Equation

The zombie equation. We also consider a third compartment that tracks the number of humans that have been turned into zombies (Z). This compartment is governed by the differential equation,

$$\frac{dZ}{dt} = \frac{\beta}{K}NZ - \gamma Z, \tag{2}$$

where γ is the zombie death rate, and β is the transmission rate of zombism. Note, it is assumed that zombies that are killed, cannot rise again.

2.1.3 The Extended System

The extended system is a combination of the resident system and the zombie equation. The equations are linked by including a transmission rate to capture the spread of zombism and a mortality rate that reflects patient zero naturally rising from the dead. As is common in the analysis of traits [10], we assume that the transmission rate is a function of virulence, specifically the mortality rate. Altogether, this yields an $S-I-I$ type model

$$\begin{aligned}
\tfrac{dN}{dt} &= b(1-\tfrac{N}{K})N - \delta N - \tfrac{\beta(\delta)}{K}DN - \delta_M N - \tfrac{\beta(\delta_M)}{K}ZN,\\
\tfrac{dD}{dt} &= (\delta + \tfrac{\beta(\delta)}{K}D)N - \mu D,\\
\tfrac{dZ}{dt} &= (\delta_M + \tfrac{\beta(\delta_M)}{K}Z)N - \gamma Z,
\end{aligned} \tag{3}$$

In addition, we assume when $\delta = 1/77/365\ day^{-1}$ that $\beta(\delta) = 0\ day^{-1}$ because natural death is not transmittable, and that δ_M is the trait value of death that leads to zombism.

2.2 Parameter Estimation

We estimated the population capacity, K, from publicly available data [2], and determined the value of transmission rate, β, based on the spread of zombie outbreaks from the literature [18]. In addition, we also obtained γ from the literature [7]. Details of model parameters, including values and their sources are available in Table 1.

2.2.1 Population Capacity

To estimate the population capacity K in the US, we applied linear regression using a least-squares method. This method used the population of the US from 1960, 1980, 2000, and 2020 [2] (Figure 1), using the average lifespan of a person within the US, 77 years [22], in conjunction with predictions of the US population from the resident

model (Figure 2). Through this procedure, the K value that had the least-squares error is $K = 558,075,379$.

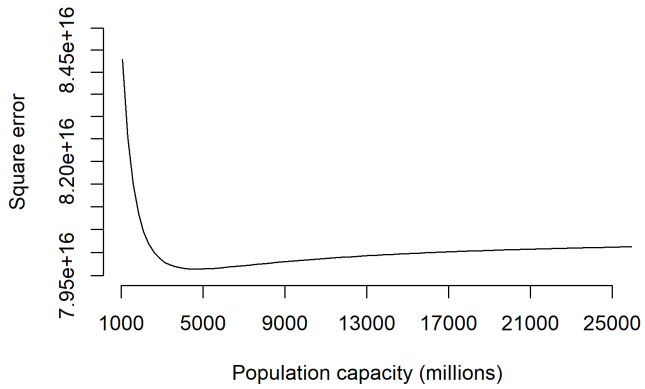


Figure 1: Square error of resident model from United States population. The square error for estimating K for the predictions of the resident model and US population data for the given value of population capacity.

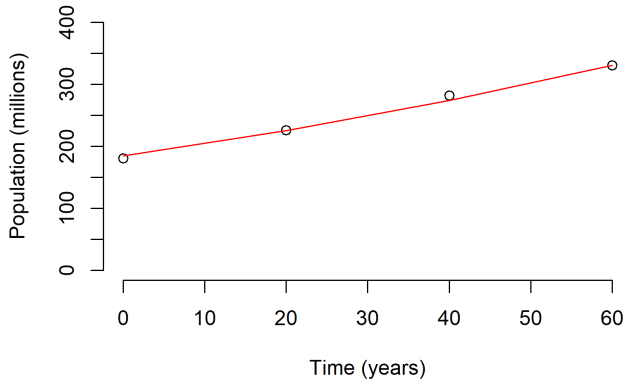


Figure 2: Resident model vs US population. The resident model with $K = 558,075,379$ (red line) and the US population from 1960 to 2020 (black points).

2.3 Equilibria and Stability Analysis of the Resident Model

Here, we determine equilibria of the resident model and apply stability analysis to evaluate its long-term behavior. We evaluated the Jacobian at the non-extinction and extinction equilibria to determine the long-term behavior of the resident system, as characterized by its eigenvalues.

To begin, the the non-extinction equilibrium of system (1) is

$$\hat{N} = (1 - \frac{\delta}{b})K \text{ and } \hat{D} = \frac{\delta}{\mu}(1 - \frac{\delta}{b})K \tag{4}$$

Evaluating the Jacobian of system (1) at the this non-extinction equilibrium, we have

$$J_{res}|_{N=(1-\frac{\delta}{b})K,\, D=\frac{\delta}{\mu}(1-\frac{\delta}{b})K} = \begin{pmatrix} -b+\delta & 0 \\ \delta & -\mu \end{pmatrix} \tag{5}$$

which yields the eigenvalues of $\lambda_1 = -b + \delta$, and $\lambda_2 = -\mu$. Thus, the non-extinction equilibrium is locally stable provided the death rate is lower than the birth rate, $\delta < b$. For the extinction equilibrium of (1), we have that $\tilde{N} = 0$ and $\tilde{D} = 0$. Thus, the Jacobian of the system (1) when $N = 0$ and $D = 0$ simplifies to

$$J_{res}|_{N=0,\, D=0} = \begin{pmatrix} b-\delta & 0 \\ \delta & -\mu \end{pmatrix} \tag{6}$$

The associated eigenvalues are $\lambda_1 = b - \delta$, and $\lambda_2 = -\mu$. The extinction equilibrium is thus locally stable when the death rate is greater than the birth rate, $\delta > b$.

2.4 Analysis of Zombie Invasion and Evolution from the Dead

To determine whether zombies could biologically evolve and become the dominant form of death, we extend the resident model to include a Zombie class, Z. We then provide details of the Jacobian of the extended model to illustrate conditions that permit zombies to invade a population and their potential evolution from the dead.

2.4.1 Conditions for a Zombie Invasion

For the extended system (3), when $\beta(\delta) = 0$ and $\delta_M = 0$, we have the non-extinction and zombie-free equilibrium

$$\hat{N} = (1 - \frac{\delta}{b})K,\ \hat{D} = \frac{\delta}{\mu}(1 - \frac{\delta}{b})K, \text{ and } \hat{Z} = 0. \tag{7}$$

The Jacobian of (3) at this equilibrium is

$$J|_{N=(1-\frac{\delta}{b})K,\, D=\frac{\delta}{\mu}(1-\frac{\delta}{b})K,\, Z=0} = \begin{pmatrix} -b+\delta & 0 & -\beta(\delta_M)(1-\frac{\delta}{b}) \\ \delta & -\mu & 0 \\ 0 & 0 & \beta(\delta_M)(1-\frac{\delta}{b}) - \gamma \end{pmatrix}. \tag{8}$$

Thus, given $\lambda_1 = -b + \delta < 0, \lambda_2 = -\mu < 0$, it follows that zombies cannot invade provided

$$\lambda_3 = \beta(\delta_M)(1 - \frac{\delta}{b}) - \gamma < 0, \tag{9}$$

as natural death would be an evolutionary stable state [17].

2.4.2 Conditions that Prevent the Zombie Uprising

To examine the potential uprising of patient zero, we now consider the Jacobian of the extended system for δ_m close, but not equal, to zero. Specifically, the Jacobian

evaluated at $\hat{N} = (1-\frac{\delta}{b})K$, $\hat{D} = \frac{\delta}{b}\left(1-\frac{\delta}{b}\right)$, and $\hat{Z} = 0$ is

$$J_{ext}|_{N=(1-\frac{\delta}{b})K,\, D=\frac{\delta}{\mu}(1-\frac{\delta}{b})K,\, Z=0} = \begin{pmatrix} b\left(1-\frac{2\hat{N}}{K}\right)-\delta-\delta_M & 0 & -\frac{\beta(\delta_M)\hat{N}}{K} \\ \delta & -\mu & 0 \\ \delta_M & 0 & \beta(\delta_M)\frac{\beta(\delta_M)\hat{N}}{K}-\gamma \end{pmatrix}. \tag{10}$$

It follows that the eigenvalues are:

$$\lambda_1(\delta,\delta_M) = -\mu,$$

$$\lambda_2(\delta,\delta_M) = \frac{1}{2}\left((1-\frac{\delta}{b})\beta(\delta_M) - b + \delta - \delta_M - \gamma + \sqrt{(1-\frac{\delta}{b})\beta(\delta_M)^2 + 2(1-\frac{\delta}{b})(b-\gamma-\delta_M-\delta)\beta(\delta_M) + (\delta_M-\delta-\gamma+b)^2}\right),$$

$$\lambda_3(\delta,\delta_M) = \frac{1}{2}\left((1-\frac{\delta}{b})\beta(\delta_M) - b + \delta - \delta_M - \gamma - \sqrt{(1-\frac{\delta}{b})\beta(\delta_M)^2 + 2(1-\frac{\delta}{b})(b-\gamma-\delta_M-\delta)\beta(\delta_M) + (\delta_M-\delta-\gamma+b)^2}\right). \tag{11}$$

Thus, for the zombie-free equilibrium to be an evolutionary stable state [17], we require

$$\lambda_2(\delta^*,\delta_M^*) \geq \lambda_2(\delta^*,\delta_M) \tag{12}$$

where $\delta^* = 1/77/365\ day^{-1}$, and $\delta_M \approx 0\ day^{-1}$.
For values of δ_M close to 0, we have that

$$\lambda_2(\delta^*,\delta_M) \approx \lambda_2(\delta^*,0) + \frac{\partial\lambda(\delta,0)}{\partial\delta_M}(\delta_M - 0) \tag{13}$$

where $\lambda_2(\delta^*,0) = -\gamma$ and $\frac{\partial\lambda(\delta,\delta_M)}{\partial\delta_M}|_{\delta_M=0} = \frac{d\beta(0)}{d\delta_M}(1-\frac{\delta^*}{b})$. Therefore, for δ_M close to 0, the zombie-free equilibrium is an evolutionary stable state provided

$$\frac{\partial\lambda_2(\delta^*,\delta_M)}{\partial\delta_M}|_{\delta_M=0} = 0 \Leftrightarrow \frac{d\beta(0)}{d\delta_M} = 0. \tag{14}$$

and

$$\frac{\partial^2\lambda(\delta^*,\delta_M)}{\partial^2\delta_M}|_{\delta_M=0} < 0. \tag{15}$$

2.5 Periodic Behavior

We now examine the potential for periodic behavior in the dynamics between humans and zombies by means of Hopf bifurcation analysis.

To begin, we assume $\delta_M \approx 0$. Thus, the extended system has the non-extinction and zombie endemic equilibria:

$$\bar{N} = \frac{\gamma}{\beta}K, \bar{D} = \frac{\gamma}{\beta}\frac{\delta}{\mu}K, \bar{Z} = \frac{(b-\delta)\beta-\gamma b}{\beta^2}K. \tag{16}$$

Rearranging the order of the system, computing the Jacobian and evaluating it at the

non-extinction and zombie endemic equilibrium, we have that

$$J_{N=\bar{N},\, D=\bar{D},\, Z=\bar{Z}} = \begin{pmatrix} -\mu & \delta & 0 \\ 0 & b(1-2\frac{2\bar{N}}{K}) - \delta - \frac{\beta}{K}\bar{Z} & -\frac{\beta}{K} \\ 0 & \frac{\beta}{K}\bar{Z} & \frac{\beta}{K}\bar{N} - \gamma \end{pmatrix} \tag{17}$$

It follows that the eigenvalues of $J_{N=\bar{N},D=\bar{D},Z=\bar{Z}}$ are

$$\lambda_1 = -\mu \text{ and } \lambda_{2,3} = -\frac{\gamma b}{2\beta} \pm \frac{1}{2}\sqrt{(\frac{\gamma b}{\beta})^2 - 4\gamma(b-\delta-\frac{\gamma b}{\beta})}. \tag{18}$$

For periodic behavior to occur we require purely imaginary eigenvalues, and so $\frac{\gamma b}{\beta} = 0$. If $\gamma = 0$ then $\lambda_{2,3} = 0$, which implies periodic behavior does not occur. If $b = 0$ then $\lambda_{2,3} = \pm\sqrt{\gamma\delta}$. Thus, for $\delta \geq 0$ and $\gamma \geq 0$ periodic behavior does not occur.

3 Results

To illustrate our predictions on the likelihood of a zombie apocalypse, and its effect on human populations, we parameterized our model according to the demographics of the US, and the zombie literature. Furthermore, to illustrate the potential outcomes for humanity and zombies, we evaluate the trajectory of our model for $\beta = 0.35$ based on the literature [18], in addition to $\beta = 0.023$ and $\beta = 0.015$ solely for the purposes of illustrating the long-term behavior of the model through stability, evolutionary invasion analysis, and Hopf-bifurcation analysis.

In the absence of zombies, the US population converges towards maximum capacity, as nothing is hindering population growth. When zombies are included, the behavior of the system depends critically on the values of β and δ_M. For instance, given $\delta_M \approx 0$, the value of β must be greater than 0.023 for zombies to disrupt the stability of the non-extinction equilibrium (Figure 4). Similarly, when $\beta = 0.35$ it is required that $\delta_M < 0.1575$ for zombies to be able to disrupt the stability of the non-extinction equilibrium (Figure 3).

The phases of the extended model show the pattern the outbreak could take, depending on how fast or slow zombies spread (Figure 6). For the zombie apocalypse to occur, the β value must be greater than 0.023. A value of β less than 0.023 causes the non-extinction and zombie endemic equilibrium to be unstable. For example, with a low value of β, such as 0.015, the zombies die out and the humans converge to their carrying capacity, K (Figure 5A, D, G). When β is slightly above 0.023, for example, 0.029, the system approaches a non-extinction and zombie endemic equilibrium, implying zombie and human populations end up coexisting (Figure 5B, E, H). For higher values of β, such as 0.35, shows more frequent decreasing oscillations between human and zombie populations, implying both populations will battle it out for dominance (Figure 5 C, F, I, and Figure 6).

To determine if the current form of death is an evolutionary stable state, we examine the largest eigenvalue $\lambda_2(\delta^*, \delta_M^*)$ when $\delta_M^* \approx 0$ (Figure 3). Specifically, for $\delta_M > \delta_M^*$, we have that $\lambda_2(\delta^*, \delta_M^*) > \lambda_2(\delta^*, \delta_M)$ (Figure 3). This means that natural death without zombies is the dominant form of death for humans, which implies that zombies cannot evolve from the dead.

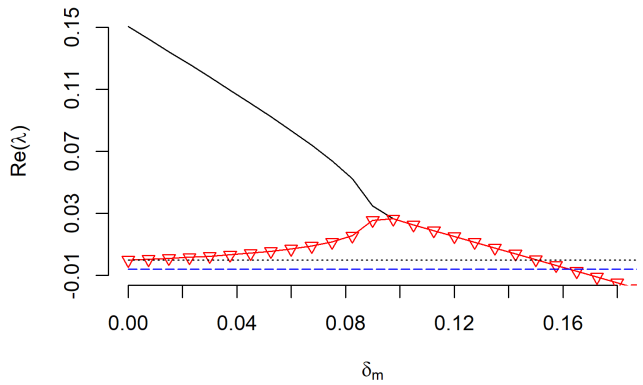


Figure 3: The change in the real part of the eigenvalues of the extended system with respect to δ_M. The black dotted line corresponds to $\lambda = 0$, with the blue dashed line being the eigenvalue $-\mu$, and the red line with triangles and black solid line representing the real parts of the eigenvalues $\lambda_{2,3}$, respectively. When $Re(\lambda) > 0$ for any eigenvalue, a zombie outbreak can occur.

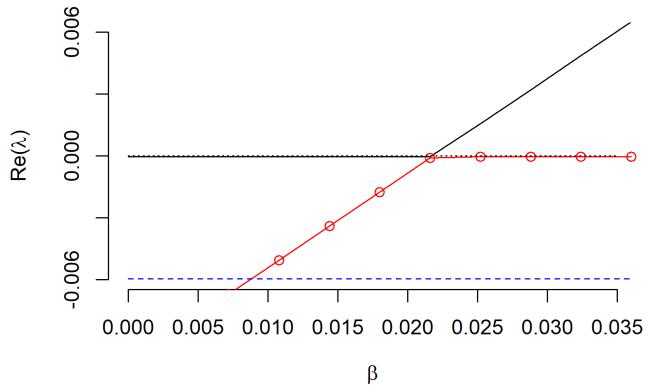


Figure 4: The change in eigenvalues of the extended system with respect to β. The value of $\lambda_1 = -\mu$ is shown by the blue line with circles. The real part is represented by the red line with triangles and black lines, respectively. The critical point on this graph is where λ_2 is greater than 0, which occurs when $\beta \approx 0.023$ per day.

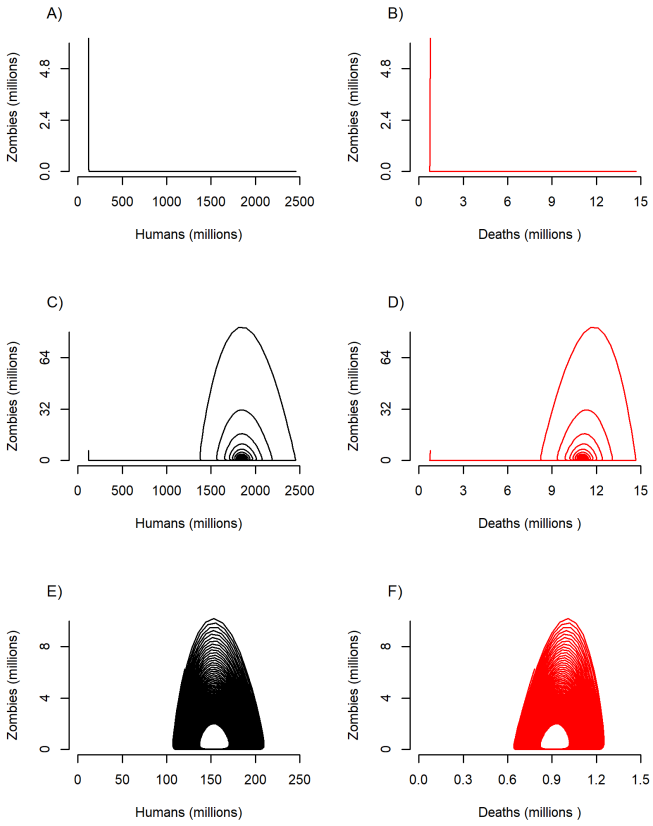


Figure 5: Phase portraits of the extended model. From top to bottom, the rows have β values of 0.015, 0.029, and 0.35. The first row shows what happens when the zombie population dies off after initial infection as the US human population continues to grow towards K. The second row results in an endemic equilibrium where each population never reaches K, but never falls to 0. This leads to both species eventually coexisting with one another. The third row illustrates more chaotic behavior as both populations rise and dip over the years showing a constant struggle for survival.

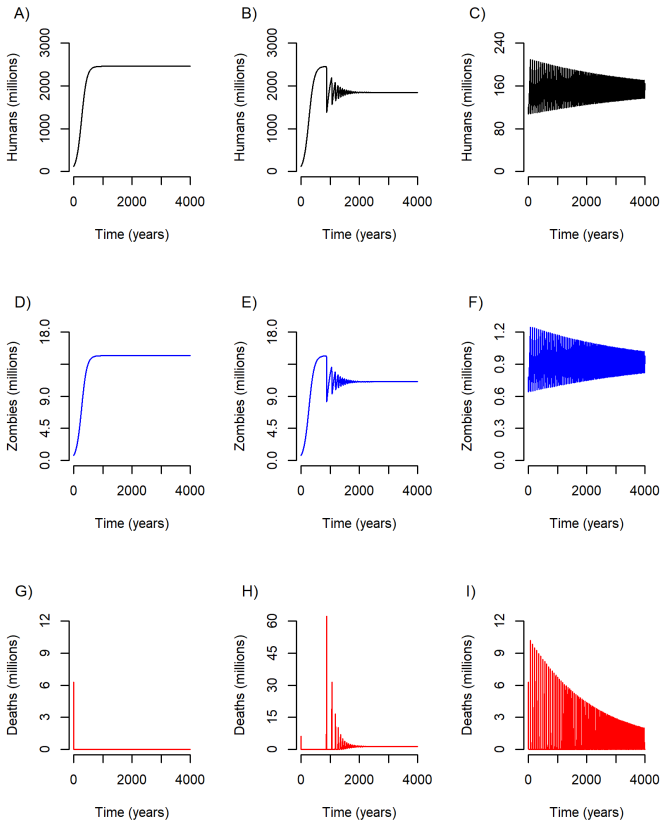


Figure 6: The trajectory of the extended model. The left, middle, and right columns have β values of 0.015, 0.029, and 0.35, respectively. The top and bottom rows correspond to plots of zombies vs. humans, and deaths vs. zombies, respectively.

4 Discussion

We analyzed a mathematical model using stability, evolutionary invasion, and Hopf-bifurcation analyses to determine the long-term prognosis of the United States and the likelihood of a potential zombie uprising. According to our model, the prognosis of the United States remains positive, so long as its birth rate continues to exceed its mortality rate, and no country imports any form of zombie infection. Importantly, our evolutionary invasion analysis shows that an invasion is likely only possible if the ratio of the zombie transmission rate to the zombie death rate is less than the ratio of alive humans to alive and non-decayed dead humans. Unfortunately, if zombies can invade the United States, our stability analysis shows that we would likely have to learn to coexist with zombies, at least until some form of public health intervention is implemented to eradicate them.

According to our results, zombie invasions are theoretically possible, provided a suffi-

ciently large ratio of transmission rate to the zombie death rate. While achieving this ratio is uncommon in nature, a single ant infected with zombie ant fungus can potentially infect entire colonies by seeking elevated locations that promote transmission in tropical climates, such as Brazil, Africa, and Thailand [4]. If a human zombie followed such behavior, this suggests they would seek out a more densely populated area, which would increase the chances of people being infected.

While our work focused on showing the theoretical conditions required for zombies to evolve or invade the United States, there exist many potential future directions. For instance, we could calibrate our model to the transmission cycle of zombie ant fungus and ants to inform the dynamics of zombie evolution. Furthermore, we could also generalize our model to account for additional traits, such as zombie speed or intelligence, or additional zombification stages, such as latent or asymptomatic infection, to gauge their effects on the likelihood of an uprising.

As with all mathematical models, our work has several limitations. To begin, there is a lack of available and reliable data on zombie outbreaks, and our analysis hinged on the functional form of the human mortality rate. Furthermore, research studies on zombie evolution are limited, although recent trends in studying zombie ant fungus are on the rise [23, 25]. Other important factors from our work include simplifying assumptions on the demographics of zombies and humans alike. Specifically, people with underlying health conditions, disabled people, the elderly, and the young would likely be at high risk of becoming zombies, which could stand to influence the speed that zombism transmits, and its capacity for invasion. Having stated this, the likely advances in science and public health from a zombie outbreak would help to offset such health inequalities, in addition to improving humanity's ability to combat epidemics and eradicate zombies.

Even though zombie ants exist, our main finding indicates human zombies are impossible, from an evolutionary standpoint. Furthermore, upon a situation where human zombies do rise, our work further highlights that the US would likely survive, either by promoting conditions that discourage a zombie invasion or by learning to coexist with zombies in some form of steady-state, at least until the time that medicine or some massive public health intervention turns the tide in humanities' favor.

Bibliography

[1] Naughty Dog. The last of us. Sony Interactive Entertainment `https://www.playstation.com/en-us/games/the-last-of-us-remastered/`. Accessed: 2021-7-19.

[2] Population, total - united states. `https://data.worldbank.org/indicator/SP.POP.TOTL?locations=US`. Accessed: 2021-6-28.

[3] Michael Rosenzweig. Coevolution of habitat diversity and species diversity. *Species Diversity in Space and Time*, 151–189, 1995.

[4] Zombie ants - real world sci-fi horror story, October 2020. `https://youtu.be/BOMNpw7kOLQ`

[5] Zombie ant. https://insects.factsdiet.com/ant/zombie-ant/, March 2021. Accessed: 2021-6-28.

[6] Vernard Adams. Dying: What Happens to the Body After Death. *Dying and Death in Oncology*, 23–30, 2017.

[7] Archyde.com. Corona 19 fatality rate, only 0.01% for people in their 20s... is it still '4 steps of distancing'? - archyde. https://www.archyde.com/, July 2021. Accessed: 2021-7-21.

[8] Danny Boyle and Alex Garland. 28 days later, June 2003.

[9] Juan Carlos Fresnadillo and Rowan Joffe. 28 weeks later, May 2007.

[10] Amy Hurford, Daniel Cownden, and Troy Day. Next-generation tools for evolutionary invasion analyses. *J. R. Soc. Interface*, 7(45):561–571, April 2010.

[11] Francis Lawrence, Mark Protosevich, and Akiva Goldsman. I am legend, December 2007.

[12] Eric T Lofgren, Kristy M Collins, Tara C Smith, and Reed A Cartwright. Equations of the end: Teaching mathematical modeling using the zombie apocalypse. *J. Microbiol. Biol. Educ.*, 17(1):137–142, March 2016.

[13] Marco Marani, Gabriel G Katul, William K Pan, and Anthony J Parolari. Intensity and frequency of extreme novel epidemics. *Proc. Natl. Acad. Sci. U. S. A.*, 118(35), August 2021.

[14] Maia Martcheva. *An introduction to mathematical epidemiology*, Springer 2015.

[15] Richard A McKay. *Patient Zero and the Making of the AIDS Epidemic*. University of Chicago Press, November 2017.

[16] Martin A Nowak. *Evolutionary Dynamics: Exploring the Equations of Life*. Harvard University Press, September 2006.

[17] Sarah P Otto and Troy Day. *A biologist's guide to mathematical modeling in ecology and evolution*, Princeton University Press 2011.

[18] Jacob Pacheco, Roman Ahmani, Kursad, Tosun, and Scott Greenhalgh.

Guns, zombies, and steelhead axes: cost-effective recommendations for surviving human societies, 2021. doi 10.13140/RG.2.2.13146.88005

[19] Marilee Peters. *Patient Zero: Solving the Mysteries of Deadly Epidemics*. Annick Press, June 2014.

[20] Robert Smith. *Mathematical Modelling of Zombies*. University of Ottawa Press, October 2014.

[21] Tara C Smith. Zombie infections: epidemiology, treatment, and prevention. *BMJ*, 351:h6423, December 2015.

[22] Laurel Wamsley. American life expectancy dropped by a full year in 1st half of 2020. *NPR*, February 2021.

[23] whyevolutionistrue. Recent data on how the "ant zombie" fungus works. https://whyevolutionistrue.com/2020/06/02/new-data-on-how-the-ant-zombie-fungus-works/, June 2020. Accessed: 2021-7-23.

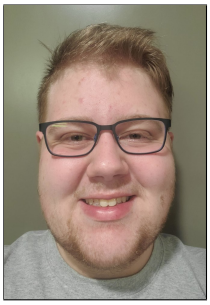
[24] Wikipedia contributors. World war Z (film). https://en.wikipedia.org/w/index.php?title=World_War_Z_(film)&oldid=1032172597, July 2021. Accessed: 2021-7-7.

[25] Carl Zimmer. After this fungus turns ants into zombies, their bodies explode. *The New York Times*, October 2019.

Ball State Undergraduate Mathematics Exchange
`https://digitalresearch.bsu.edu/mathexchange`
Vol. 16, No. 1 (Fall 2022)
Pages 50 –57

Sums of Diagonals in Pascal's Triangle

Jamisen McCrary , Russell May*



Jamisen McCrary is an undergraduate student at the University of Kentucky, studying mechanical engineering. He was previously a student in the Craft Academy at Morehead State University. His inspiration for the research topic was his prior appreciation for Pascal's Triangle and finding patterns. Apart from mathematics, Jamisen enjoys listening to music, creating art, working with electronics, programming, 3D modeling, cooking, and deep thinking. His overarching goal in life is to become a professor of engineering at a secondary education institution and build up the next generation of engineers to their fullest potential.

Russell May teaches math at Morehead State University and coaches the problem-solving club for the Craft Academy, a dual-credit academy for talented high school students in Kentucky. He is interested in combinatorics, especially problems with nice generating function solutions.

Abstract

We analyze sums of entries on diagonals of integer slope in Pascal's triangle, obtain a recurrence relation that these diagonal sums obey, and compute their generating function. We use the generating function to approximate the exponential growth of the diagonal sums.

1 Introduction

A question from a national high school math competition poses (problem 3, page 274 of [4]): how many subsets, $d(n)$, of the integers in the interval $1 \ldots n+1$ have the property that their least element coincides with their cardinality? Here is a solution: let $k \geq 1$ be the common value of the least element and cardinality of the subset. Then, besides the least element, the subset must contain a choice of the remaining $k-1$ of the $n+1-k$ integers in the interval $k+1 \ldots n+1$. So, the solution is $d(n) = \sum_{k \geq 1} \binom{n+1-k}{k-1}$.

***Corresponding author**: Jamisen.McCrary@uky.edu

However, this sum can be changed to a more telling form. By the symmetry property for binomial coefficients $\binom{a}{b} = \binom{a}{a-b}$, each $d(n)$ is equivalent to $\sum_{k\geq 1}\binom{n+1-k}{n+2-2k}$. With the substitution $r = \lfloor n/2 \rfloor + 1 - k$, we can change the subtraction in the binomial coefficients to addition. So,

$$d(n) = \sum_{r\geq 0} \binom{\lceil \frac{n}{2} \rceil \quad + \quad r}{(-n) \bmod 2 \quad + \quad 2r}. \tag{1}$$

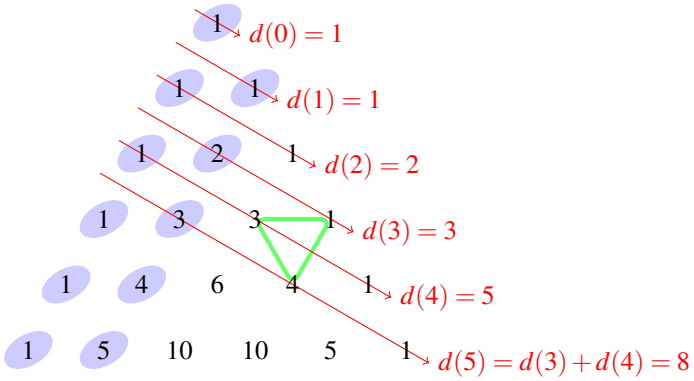


Figure 1: Sums of entries in Pascal's triangle on diagonals of slope 2 in red and intercepts in indigo. These sums coincide with the Fibonacci numbers. The triangle in green indicates how the additive identity for Pascal's triangle leads to the recurrence relation for the $d(n)$'s.

The sums in Equation 1 are depicted in Figure 1 as diagonals in Pascal's triangle, where $d(n)$ is the sum of the entries on the n^{th} diagonal from the top. These diagonals have a *slope* of 2 in the sense that if a diagonal passes through an entry in some row and position, then the diagonal also passes through the entry in the next row whose position is two more than the previous one. Each diagonal also has an *intercept*, i.e., the entry in the uppermost row with non-negative position through which the diagonal passes. The n^{th} diagonal has an intercept at $\binom{\lceil n/2 \rceil}{(-n) \bmod 2}$, meaning that the uppermost row that the diagonal passes through is $\lceil n/2 \rceil$ and the position in this row is $(-n) \bmod 2$. It is both well-known and easy to prove that the sequence $\langle d(n) : n \geq 0 \rangle$ coincides with the Fibonacci sequence beginning with two one's: 1, 1, 2, 3, 5, 8, 13, ... (see [5]). Each entry on a diagonal in Pascal's triangle is the sum of the two entries directly above it. Since these two entries lie on the two previous diagonals, then $d(n+2) = d(n) + d(n+1)$, which is the same recurrence relation that the Fibonacci numbers satisfy.

The problem from this high school competition inspires many questions. What are the sums of the entries on diagonals of slope greater than two? Are these sums on steeper diagonals also famous sequences, like the Fibonacci sequence? The terms in the Fibonacci sequence grow like powers of the golden ratio—what is the growth rate of the sums on steeper diagonals?

The general question we consider here is to determine the sum of the entries on the n^{th} diagonal of slope h in Pascal's triangle. We denote this sum by $d_h(n)$. Historically, sums

of entries on diagonals of various slopes have already been considered, for instance, in [1] and [2]. In fact, [1] even considers slopes with *rational* values and obtains recurrence relations for these sums. However, we extend their analyses by computing the generating function for the diagonal sums and then use the generating function to approximate their exponential growth.

As a brief review, we highlight a couple basic properties of Pascal's triangle. Pascal's triangle is depicted in a hexagonal lattice in a half-plane with a numerical entry in each cell. The cells in the rows are indexed by integers, and the rows are indexed by non-negative integers. The position of an entry in one row is the same as the entry to the left in the row below. The entry in the r^{th} position of the n^{th} row coincides with the binomial coefficient $\binom{n}{r}$. If $0 \leq r \leq n$, then $\binom{n}{r}$ is positive. Otherwise, it is zero for $r < 0$ or $r > n$. The additive identity asserts that $\binom{n}{r}$ is the sum of the entries in the row above to the left $\binom{n-1}{r-1}$ and right $\binom{n-1}{r}$.

To be definitive, we say a *diagonal* in Pascal's triangle is a line which passes through entries in the triangle. For a diagonal to have slope h means that if the diagonal passes through $\binom{a}{b}$ in one row, then it also passes through $\binom{a+1}{b+h}$ in the next row. We say the *intercept* of a diagonal is the uppermost row and the non-negative position in this row which the diagonal passes through. Thus, if $a \geq 0$ and $0 \leq b < h$, then $\sum_{r \geq 0} \binom{a+r}{b+hr}$ represents the sum of the entries on the diagonal in Pascal's triangle with slope h and intercept $\binom{a}{b}$. Diagonals are enumerated from the top down. So, if the diagonal through an entry has index n, then the diagonal through the entry to the left in the same row has index $n+1$, and the diagonal through the entry to the left in the next row has index $n+h$. There are h diagonals of slope h with intercepts in each row, except for the top row which just has a single intercept.

2 Diagonals of Slope Three

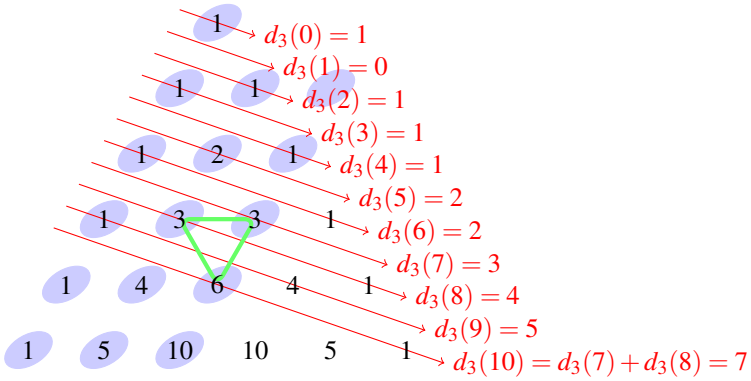


Figure 2: Pascal's triangle with sums of diagonals of slope 3 in red and intercepts in indigo. The sums of the diagonals coincide with the Padovan numbers. The triangle in green shows how the additive identity for Pascal's triangle leads to the recurrence relation for the $d_3(n)$'s

Now let's consider diagonals of slope 3 in Pascal's triangle, as depicted in Figure 2. Let $d_3(n)$ denote the sum of the entries on the n^{th} diagonal of slope 3. Note that the initial diagonal for $n = 0$ goes through the apex of Pascal's triangle, and so $d_3(0) = 1$, but the next diagonal for $n = 1$ only passes through the intercept at $\binom{1}{2}$, and so $d_3(1) = 0$. In general, the intercepts of diagonals snake through the first three positions of each row from right to left, except for the top row which only includes one position. So, the intercept of the n^{th} diagonal goes through row $\lceil n/3 \rceil$ at position $(-n) \bmod 3$. Since the slope of each diagonal is 3, the position of each entry on a diagonal is three more than the previous row. Therefore, the sum of the entries on the n^{th} diagonal of slope 3 is

$$d_3(n) = \sum_{r \geq 0} \binom{\lceil \frac{n}{3} \rceil \quad + \quad r}{(-n) \bmod 3 \quad + \quad 3r}.$$

The sums of the diagonals of slope 3 coincide with the so-called Padovan sequence $\langle p_n : n \geq 0 \rangle = 1, 0, 1, 1, 1, 2, 2, 3, 4, 5, 7, 9, 12, \ldots$. The Padovan sequence satisfies the recurrence relation $p_{n+3} = p_n + p_{n+1}$ for all $n \geq 0$, and the terms p_n in the Padovan sequence grow asymptotically as αr^n, where $r \approx 1.3247$ is the real root of the polynomial $x^3 - x - 1$ and $\alpha = 1/(2r+3)$ (see [6]). The fact that the sequence $\langle d_3(n) : n \geq 0 \rangle$ of diagonals of slope 3 satisfies the same recurrence relation is a direct consequence of the additive identity for Pascal's triangle $\binom{a+1}{b} = \binom{a}{b-1} + \binom{a}{b}$. Suppose the diagonal that passes through entry $\binom{a}{b}$ has index n. Then the diagonal that passes through the entry immediately to the left $\binom{a}{b-1}$ has index $n+1$, and the diagonal that passes through the entry $\binom{a+1}{b}$ to the left in the row below has index $n+3$. Since the additive identity holds uniformly for all entries on these diagonals, then $d_3(n+3) = d_3(n) + d_3(n+1)$.

3 Diagonals of Integer Slope

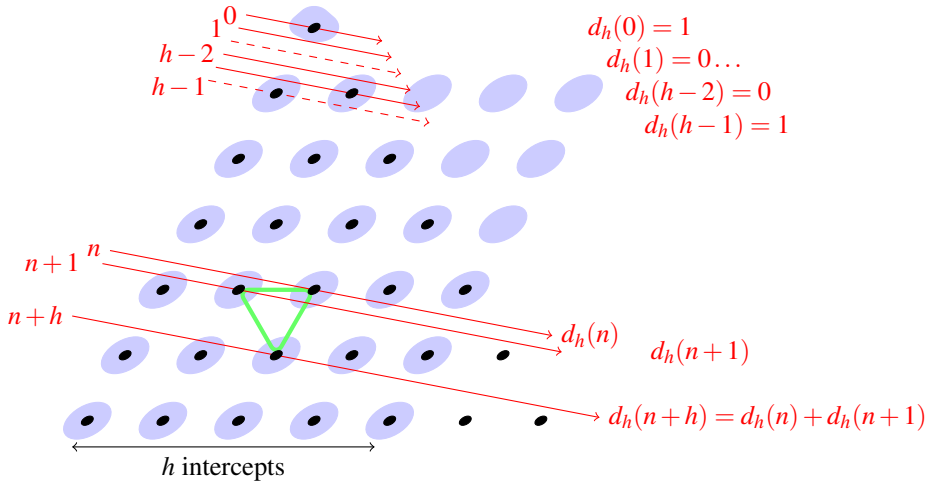


Figure 3: The first rows of Pascal's triangle with sums of diagonals of integer slope h in red and intercepts in indigo. The triangle in green shows how the additive identity for Pascal's triangle leads to the recurrence relation for the $d_h(n)$'s.

For the general case, we consider diagonals of integer slope h in Pascal's triangle, enumerated from the top down. We denote the sum of the entries on the n^{th} diagonal of slope h by $d_h(n)$. Figure 3 displays these diagonals. Note that the initial diagonal for $n = 0$ goes through the apex of Pascal's triangle, and so $d_h(0) = 1$, but the next diagonals for $n = 1 \ldots h-2$ have intercepts at $\binom{1}{r}$ in the first row with $r > 1$, and so $d_h(1) = \cdots = d_h(h-2) = 0$. The diagonal of index $h-1$ only passes through the intercept at $\binom{1}{1}$, and so $d_h(h-1) = 1$. Together, these give the initial conditions

$$d_h(0) = 1, \quad d_h(1) = \cdots = d_h(h-2) = 0, \quad d_h(h-1) = 1. \tag{2}$$

The intercepts of the remaining diagonals continue to snake through the first h positions of each row from right to left. Therefore, the row of the intercept of the n^{th} diagonal is $\lceil n/h \rceil$, and its position in this row is $(-n) \bmod h$. The position of each entry in a row on the n^{th} diagonal is h more than the previous row. Therefore,

$$d_h(n) = \sum_{r \geq 0} \binom{\lceil \frac{n}{h} \rceil \quad + \quad r}{(-n) \bmod h \quad + \quad hr}.$$

This is an explicit representation of $d_h(n)$ as a sum. For the purposes of approximation, however, it is more useful to have a recursive representation. A recurrence relation for the $d_h(n)$'s is a direct consequence of the additive identity for Pascal's triangle that $\binom{a+1}{b} = \binom{a}{b-1} + \binom{a}{b}$. Suppose the n^{th} diagonal of slope h passes through entry $\binom{a}{b}$. Then, the diagonal that passes through the entry immediately to the left $\binom{a}{b-1}$ has index $n+1$, and the diagonal that passes through the entry in the next row to the left $\binom{a+1}{b}$ has index $n+h$. Since the additive identity holds uniformly for all entries on the diagonals, then

$$d_h(n+h) = d_h(n) + d_h(n+1), \quad \text{for all } n \geq 0. \tag{3}$$

This recurrence relation is linear, has constant coefficients, and is of degree h. We obtain an asymptotic approximation for $d_h(n)$ in the Section 5. Figure 4 shows a graph of the logarithm of $d_h(n)$ for $h = 20$ and $n = 1 \ldots 1200$, based on Equation 3. This graph prominently shows damped oscillations of period h, but modulo these oscillations the graph shows simple exponential growth for the diagonal sums.

4 Generating Function

An often-used tool for analyzing combinatorial sequences is the generating function. The ordinary generating function of the sequence $\langle a_n : n \geq 0 \rangle$ is $\sum_{n \geq 0} a_n x^n$. It can be thought of as a formal power series or, wherever it converges, a function of complex numbers. Wilf in section 1.2 of [3] gives a five-step method for converting a recurrence relation describing a sequence to its generating function: clarify the set of valid values of the free variable in the recurrence relation, name the generating function, multiply each instance of the recurrence by an appropriate power of the variable of the generating function and sum over the valid values, express both sides of the resulting equation in terms of the generating function, and finally solve the resulting equation for the generating function. The initial conditions in Equation 2 give the

values of $d_h(n)$ for $n = 0 \ldots h-1$, and the general recurrence relation in Equation 3 determines all the rest of the values for $n \geq h$. We define the generating function $D_h(x) = \sum_{n \geq 0} d_h(n) x^n$. Multiplying each term in the initial conditions of Equation 2 and recurrence relation of Equation 3 by the corresponding power of x and summing, we get $(D_h(x) - 1 - x^{h-1})/x^h = D_h(x) + (D_h(x) - 1)/x$. Solving for the generating function results in an amazingly simple expression:

$$D_h(x) = \frac{1}{1 - x^{h-1} - x^h}. \tag{4}$$

5 Approximation of Sums

The generating function in Equation 4 for the sequence $\langle d_h(n) : n \geq 0 \rangle$ is first and foremost a rational function, i.e., a ratio of polynomials. In this case the rational function has a numerator $f(x) = 1$ and denominator $g(x) = 1 - x^{h-1} - x^h$. Using basic tools of calculus, it is straightforward to approximate the exponential growth of any sequence whose generating function is rational. The first step is to determine the partial fraction decomposition of the generating function. Let R_h denote the collection of roots of the denominator $g(x)$. Since $g(x)$ and its derivative $g'(x) = -x^{h-2}(h - 1 + hx)$ have no common roots, then none of the roots of $g(x)$ are repeated. So, the partial fraction decomposition of the generating function has the form

$$D_h(x) = \sum_{r \in R_h} \frac{a_r}{x - r},$$

where $a_r = \frac{f(r)}{g'(r)}$. Each term in the partial fraction decomposition represents a geometric series, as follows:

$$\frac{a_r}{x - r} = -\frac{a_r}{r(1 - \frac{x}{r})} = -\frac{a_r}{r} \sum_{m \geq 0} \frac{x^m}{r^m}$$

We can extract the coefficient of each term of the decomposition with the coefficient extraction operator. By definition, if $f(x) = \sum_{n \geq 0} a_n x^n$, then $[x^n] f(x) = a_n$. Then,

$$\begin{aligned} & [x^n] D_h(x) \\ &= \textstyle\sum_{r \in R_h} [x^n] \sum_{m \geq 0} -\frac{a_r}{r} \frac{x^m}{r^m} \\ &= \textstyle\sum_{r \in R_h} -\frac{f(r)}{r g'(r)} r^{-n}. \end{aligned} \tag{5}$$

The terms with the largest contribution to $d_h(n)$ in Equation 5 are the ones whose roots have the smallest modulus. In this case, we will shortly see that there is a single real root $\tilde{r}$ of $g(x)$ with the smallest modulus, and this root is called the *dominant singularity* of the generating function. The exponential growth approximation for $d_h(n)$ concentrates solely on this singularity, that is,

$$d_h(n) \approx -\frac{f(\tilde{r})}{\tilde{r} g'(\tilde{r})} \tilde{r}^{-n}. \tag{6}$$

To find the dominant singularity, first consider the case when the slope h of the diagonals is even. When $g(x) = 1 - x^{h-1} - x^h$ is graphed on the real line, we see that $g(-\infty) = g(\infty) = -\infty$, $g(-1) = g(0) = 1$, and $g(1) = -1$. Because of the sign

changes, there are real roots in the intervals $(-\infty,-1)$ and $(0,1)$. Since the derivative $g'(x) = -x^{h-2}(h-1+hx) > 0$ iff $x < -\frac{h-1}{h}$, then these are the only real roots. To see that $\tilde{r}$ has a smaller modulus than any of the complex roots, note that if $|z| < \tilde{r}$, then $|z^h + z^{h-1}| \leq |z^h| + |z^{h-1}| < \tilde{r}^h + \tilde{r}^{h-1} = 1$, excluding the possibility that such z's could be a root of g. So, the dominant singularity $\tilde{r}$ of this generating function must be the real root in the interval $(0,1)$. By a similar analysis, when h is odd, the dominant singularity is still in $(0,1)$. For $h \geq 5$, it is impossible to express $\tilde{r}$ in terms of radicals. However, it is easy to approximate for large values of h. If h is large, then $h-1$ and h are both nearly equal to $h-1/2$, and so $g(x) \approx 1 - 2x^{h-1/2}$. Therefore, the dominant singularity is approximately $\tilde{r} \approx 2^{-1/(h-1/2)}$. At this value, $f(\tilde{r}) = 1$ and $g'(\tilde{r}) \approx -h + 1/2$. Plugging in these values into Equation 6 results in

$$d_h(n) \approx \frac{1}{h-1/2} 2^{\frac{n+1}{h-1/2}} \tag{7}$$

The graph of the logarithm of this approximation appears as a line in Figure 4 and shows close agreement to the graph of the logarithm of $d_h(n)$. However, it is inherent that since only an approximation was used for the dominant singularity, then this approximating line must eventually diverge from the exact values. Of course, the exponential growth in Equation 7 does not account for the oscillations that are prominent in the graph of $d_h(n)$ in Figure 4. These oscillations are the result of the complex roots in the partial fractions decomposition of the generating function and will be the object of further study.

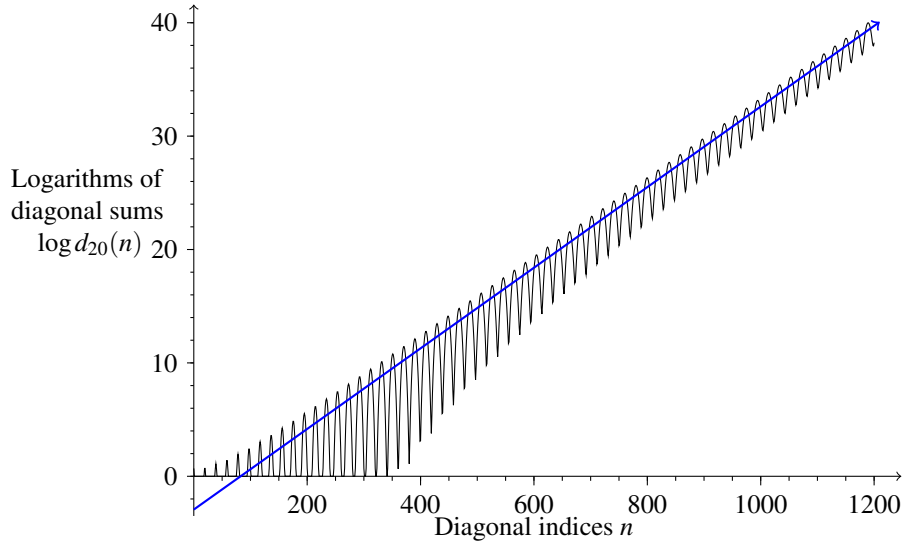


Figure 4: A graph of the logarithms of the sums of the entries on the first 1200 diagonals with slope $h = 20$ in Pascal's triangle, along with the linear approximation from Equation 7.

Bibliography

[1] Green, Thomas M. Recurrent Sequences and Pascal's Triangle. Mathematics Magazine, 41(1), 13-21, (1968). http://www.jstor.org/stable/2687953

[2] Joseph A. Raab. A Generalization of the Connection Between the Fibonacci Sequence and Pascal's Triangle. The Fibonacci Quarterly, 3(3), 21-31, (1963) https://www.fq.math.ca/Scanned/1-3/raab.pdf

[3] Herbert S. Wilf. Generatingfunctionology. A K Peters/CRC Press, (2005).

[4] Paul J. Karafiol et al. American Regions Math League Power & Local Contests 2009–2014. ARML, (2015).

[5] N. J. A. Sloane. Online Encyclopedia of Integer Sequences. http://oeis.org/A000045

[6] N. J. A. Sloane. Online Encyclopedia of Integer Sequences. http://oeis.org/A000045

Ball State Undergraduate Mathematics Exchange
`https://digitalresearch.bsu.edu/mathexchange`
Vol. 16, No. 1 (Fall 2022)
Pages 58 – 65

Polynomials that Preserve Nonnegative Matrices of Order Two

*Benjamin J. Clark, Pietro Paparella**

Benjamin J. Clark finished his post baccalaureate degree in Mathematics at the University of Washington Bothell after previously completing his Computer Science degree at the same school. He is starting graduate study in Mathematics at Washington State University. Outside of Math Benjamin enjoys programming, chess, biking, and video games.

Pietro Paparella received the Ph.D. degree in mathematics from Washington State University in 2013 and is currently an Associate Professor of mathematics in the Division of Engineering and Mathematics at the University of Washington Bothell. His research interests are in nonnegative matrix theory, combinatorial matrix theory, discrete geometry, and the geometry of polynomials. His hobbies include guitar playing, charcoal drawing, and oil painting.

Abstract

A known characterization for entire functions that preserve all nonnegative matrices of order two is shown to characterize polynomials that preserve nonnegative matrices of order two. Equivalent conditions are derived and used to prove that $\mathscr{P}_3 \subset \mathscr{P}_2$, which was previously unknown. A new characterization is given for polynomials that preserve nonnegative circulant matrices of order two.

1 Introduction

In 1979, Loewy and London [3] posed the problem of characterizing

$$\mathscr{P}_n := \{p \in \mathbb{C}[x] \mid p(A) \geq 0, \forall A \in \mathsf{M}_n(\mathbb{R}), A \geq 0\},$$

*__Corresponding author__: pietrop@uw.edu

for every positive integer n. In particular, and for practical purposes, necessary and sufficient conditions are sought in terms of the coefficients of the polynomials belonging to $\mathscr{P}_n$.

The characterization of $\mathscr{P}_1$ is known as the Pólya–Szegö theorem (see, e.g., Powers and Reznick [5, Proposition 2]), which asserts that $p \in \mathscr{P}_1$ if and only if

$$p(x) = \left(f_1(x)^2 + f_2(x)^2\right) + x\left(g_1(x)^2 + g_2(x)^2\right).$$

Bharali and Holtz [1] gave partial results for the set

$$\mathscr{F}_n := \{f \text{ entire} \mid f(A) \geq 0, \forall A \in \mathsf{M}_n(\mathbb{R}), A \geq 0\} \supset \mathscr{P}_n$$

and characterized entire functions that preserve certain structured nonnegative matrices, including upper-triangular matrices and circulant matrices. In addition, they gave necessary and sufficient conditions for an entire function f to belong to $\mathscr{F}_2$. Specifically, they showed that an entire function f belongs to $\mathscr{F}_2$ if and only if

$$f(x+y) - f(x-y) \geq 0, \ \forall x, y \geq 0, \tag{1}$$

and

$$(x+y-z)f(x-y) + (z-x+y)f(x+y) \geq 0, \ \forall x \geq z \geq 0, y \geq x-z, \tag{2}$$

or, equivalently, if f satisfies (1) and

$$(x+y)f(x-y) + (y-x)f(x+y) \geq 0, \ \forall y \geq x \geq 0. \tag{3}$$

More recently, Clark and Paparella [2] gave partial results for $\mathscr{P}_n$ in terms of the coefficients of the polynomials in $\mathscr{P}_n$. While it is known that $\mathscr{P}_{n+1} \subseteq \mathscr{P}_n$, $\forall n \in \mathbb{N}$, Clark and Paparella proved that $\mathscr{P}_2 \subset \mathscr{P}_1$ and conjectured that $\mathscr{P}_{n+1} \subset \mathscr{P}_n$, $\forall n \in \mathbb{N}$.

In this work, it is shown that the characterization for $\mathscr{F}_2$ established by Bharali and Holtz also characterizes $\mathscr{P}_2$. Our demonstration, which utilizes the definition of matrix function via *Jordan canonical form*, directly establishes that (1) and (3) are necessary and sufficient whereas Bharali and Holtz establish (1) and (2) (via the definition of matrix function via *interpolating polynomial*) and proceed to show that (2) is equivalent to (3). Equivalent conditions are derived for (1) that are used to prove that $\mathscr{P}_3 \subset \mathscr{P}_2$, which was previously unknown. A new characterization is given for polynomials that preserve nonnegative circulant matrices of order two.

2 Notation and Background

The set of m-by-n matrices with entries from a field $\mathbb{F}$ is denoted by $\mathsf{M}_{m\times n}(\mathbb{F})$. If $m = n$, then $\mathsf{M}_{m\times n}(\mathbb{F})$ is abbreviated to $\mathsf{M}_n(\mathbb{F})$. The set of all n-by-1 column vectors is identified with the set of all ordered n-tuples with entries in $\mathbb{F}$ and thus denoted by $\mathbb{F}^n$.

If $A \in \mathsf{M}_n(\mathbb{F})$, then a_{ij} denotes the (i,j)-entry of A. If $\mathbb{F} = \mathbb{R}$ and $a_{ij} \geq 0$ ($a_{ij} > 0$), $1 \leq i, j \leq n$, then A is called *nonnegative* (respectively, *positive*) and this is denoted by $A \geq 0$ (respectively, $A > 0$).

Unless otherwise stated,

$$p(x) = \sum_{k=0}^{m} a_k x^k \in \mathbb{C}[x],$$

where $a_m \neq 0$. If n is a positive integer less than or equal to m, then the coefficients $a_0, a_1, \ldots, a_{n-1}$ are called the *first n terms of p* and the coefficients $a_{m-n+1}, \ldots, a_{m-1}, a_m$ are called the *last n terms of p*.

3 Basic Observations

Lemma 1. *If D is a positive diagonal matrix, then $p(A) \geq 0$ if and only if $p(D^{-1}AD) \geq 0$.*

Proof. The result follows immediately by observing that $p(D^{-1}AD) = D^{-1}p(A)D$. □

Lemma 2. *If P is a permutation matrix, then $p(A) \geq 0$ if and only if $p(P^{\top}AP) \geq 0$.*

Proof. Similar to the proof of Lemma 1. □

We briefly digress to present the following result which, to the best of our knowledge, has not previously been addressed in the literature.

Theorem 3. *If $p \in \mathscr{P}_1$, then $a_k \in \mathbb{R}$, $\forall k \in \{0, 1, \ldots, m\}$.*

Proof. It is known that if f is an analytic function defined on a *self-conjugate* domain $\mathscr{D} \subseteq \mathbb{C}$ (i.e., $\mathscr{D}$ is symmetric with respect to the real-axis in the complex-plane) and $f(x) \in \mathbb{R}$, $\forall x \in \mathscr{I} := \mathscr{D} \cap \mathbb{R}$, then $f^{(k)}(x) \in \mathbb{R}$, $\forall x \in \mathscr{I}$ (see, e.g., Paparella [4, Lemma 4.7]). In particular, $p^{(k)}(x) \in \mathbb{R}$, $\forall x \geq 0$. The result follows by noting that $a_k = p^{(k)}(0)/k! \in \mathbb{R}$. □

Corollary 4. *If $p \in \mathscr{P}_n$, then $a_k \in \mathbb{R}$, $\forall k \in \{0, 1, \ldots, m\}$.*

Proof. Since $\mathscr{P}_{n+1} \subseteq \mathscr{P}_n$, $\forall n \in \mathbb{N}$ [1, Lemma 1], it follows that $p \in \mathscr{P}_1$. The result is now immediate from Theorem 3. □

4 A Characterization of $\mathscr{P}_2$

Lemma 5. *Let $A \in \mathsf{M}_2(\mathbb{R})$ and suppose that $A > 0$. If $\sigma(A) = \{\rho, \mu\}$, with $\rho > |\mu|$, then A is similar to a matrix of the form*

$$\frac{1}{1+\alpha}\begin{bmatrix} \alpha\rho + \mu & \rho - \mu \\ \alpha(\rho - \mu) & \alpha\mu + \rho \end{bmatrix},$$

where $\alpha > 0$.

Proof. By the Perron–Frobenius theorem for positive matrices, there is a positive vector x such that $Ax = \rho x$. If $D = \operatorname{diag}(x_1, x_2)$, then the positive matrix

$$B := D^{-1}AD$$

has row sums equal to ρ. Thus, there is an invertible matrix $\hat{S} = \begin{bmatrix} 1 & \hat{a} \\ 1 & \hat{b} \end{bmatrix}$ such that

$$B = \hat{S} \begin{bmatrix} \rho & 0 \\ 0 & \mu \end{bmatrix} \hat{S}^{-1}.$$

Notice that $\hat{a} \neq 0$ and $\hat{b} \neq 0$: for contradiction, if $\hat{a} = 0$, then

$$B \begin{bmatrix} 0 \\ \hat{b} \end{bmatrix} = \mu \begin{bmatrix} 0 \\ \hat{b} \end{bmatrix},$$

but

$$B \begin{bmatrix} 0 \\ \hat{b} \end{bmatrix} = \hat{b} \begin{bmatrix} b_{12} \\ b_{22} \end{bmatrix}.$$

Thus, $b_{12} = 0$, but this is a contradiction since $B > 0$. A similar calculation demonstrates that $\hat{b} \neq 0$.

If

$$S := \hat{S} \begin{bmatrix} 1 & 0 \\ 0 & 1/\hat{a} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & a \end{bmatrix},$$

where $a = \hat{b}/\hat{a}$, then

$$B = S \begin{bmatrix} \rho & 0 \\ 0 & \mu \end{bmatrix} S^{-1}.$$

Furthermore, $a < 0$ (otherwise,

$$B = \frac{1}{1-a} \begin{bmatrix} a\rho - \mu & \mu - \rho \\ a(\rho - \mu) & a\mu - \rho \end{bmatrix}$$

and $b_{12} < 0$). Thus,

$$S = \begin{bmatrix} 1 & 1 \\ 1 & -\alpha \end{bmatrix}, \ \alpha > 0,$$

and

$$B = \frac{1}{1+\alpha} \begin{bmatrix} \alpha\rho + \mu & \rho - \mu \\ \alpha(\rho - \mu) & \alpha\mu + \rho \end{bmatrix}, \tag{4}$$

as desired. □

To simplify the main result, we rely on the following result [1, Lemma 4].

Lemma 6. *If $p \in \mathbb{R}[x]$, then $p \in \mathscr{P}_n$ if and only if $p(A) \geq 0$ whenever $A > 0$.*

Proof. Follows from the continuity of p and the fact that the set of positive matrices of order n is dense in the set of all nonnegative matrices of order n. □

Theorem 7 (cf. [1, Theorem 13])**.** *If $p \in \mathbb{R}[x]$, then $p \in \mathscr{P}_2$ if and only if*

$$p(\rho) \geq |p(\mu)|, \ \forall \rho, \mu \in \mathbb{R}, \rho \geq |\mu| \tag{5}$$

and

$$\rho p(-\mu) + \mu p(\rho) \geq 0, \tag{6}$$

whenever $0 < \mu \leq \rho$.

Proof. If $p \in \mathscr{P}_2$, then the necessity of (5) follows by noting that if

$$A := \frac{1}{2}\begin{bmatrix} \rho+\mu & \rho-\mu \\ \rho-\mu & \rho+\mu \end{bmatrix} \geq 0, \tag{7}$$

with $\rho \geq |\mu|$, then

$$p(A) = \frac{1}{2}\begin{bmatrix} p(\rho)+p(\mu) & p(\rho)-p(\mu) \\ p(\rho)-p(\mu) & p(\rho)+p(\mu) \end{bmatrix} \geq 0. \tag{8}$$

Let ρ and μ be real numbers such that $0 < \mu \leq \rho$. If

$$A := \begin{bmatrix} 0 & \rho \\ \mu & \rho-\mu \end{bmatrix} \geq 0$$

then

$$\begin{aligned} A &= \begin{bmatrix} \rho & 1 \\ -\mu & 1 \end{bmatrix}\begin{bmatrix} -\mu & 0 \\ 0 & \rho \end{bmatrix}\begin{bmatrix} \rho & 1 \\ -\mu & 1 \end{bmatrix}^{-1} \\ &= \frac{1}{\rho+\mu}\begin{bmatrix} \rho & 1 \\ -\mu & 1 \end{bmatrix}\begin{bmatrix} -\mu & 0 \\ 0 & \rho \end{bmatrix}\begin{bmatrix} 1 & -1 \\ \mu & \rho \end{bmatrix}, \end{aligned}$$

and

$$p(A) = \frac{1}{\rho+\mu}\begin{bmatrix} \rho p(-\mu)+\mu p(\rho) & \rho(p(\rho)-p(-\mu)) \\ \mu(p(\rho)-p(-\mu)) & \rho p(\rho)+\mu p(-\mu) \end{bmatrix} \geq 0.$$

i.e., p satisfies (6).

Conversely, suppose that p satisfies (5) and (6). In view of Lemma 6, it suffices to show that p maps positive matrices of order two to nonnegative matrices of order two. To this end, suppose that A is a positive matrix of order two with spectrum $\{\rho,\mu\}$. Without loss of generality, assume that $\rho > |\mu|$.

By Lemma 5, A is similar to a matrix of the form

$$B = \frac{1}{1+\alpha}\begin{bmatrix} \alpha\rho+\mu & \rho-\mu \\ \alpha(\rho-\mu) & \alpha\mu+\rho \end{bmatrix},$$

where $\alpha > 0$. In view of Lemmas 1 and 2, it suffices to show that $p(B) \geq 0$. By (4), notice that

$$p(B) = Sp(D)S^{-1} = \frac{1}{1+\alpha}\begin{bmatrix} \alpha p(\rho)+p(\mu) & p(\rho)-p(\mu) \\ \alpha(p(\rho)-p(\mu)) & \alpha p(\mu)+p(\rho) \end{bmatrix}.$$

Since p satisfies (5), it follows that $p \in \mathscr{P}_1$. Thus, $p(B) \geq 0$ whenever $\mu \geq 0$.

If $\mu < 0$, then, since $B > 0$, it follows that $\alpha > -\mu/\rho = |\mu|/\rho$. Thus,

$$\alpha p(\rho)+p(\mu) > \frac{|\mu|}{\rho}p(\rho)+p(-|\mu|) = \frac{|\mu|p(\rho)+\rho p(-|\mu|)}{\rho} \geq 0$$

by (6). The remaining entries of $p(B)$ are nonnegative by (5). □

5 Equivalent Conditions for $\mathscr{P}_2$

Proposition 8. *If $p \in \mathbb{R}[x]$, then*

$$p(x) \geq |p(y)|,\ \forall x,y \in \mathbb{R}, x \geq |y| \tag{9}$$

if and only if

$$p' \in \mathscr{P}_1 \tag{10}$$

and

$$p(x) \geq |p(-x)|, \ \forall x \geq 0. \tag{11}$$

Proof. First, note that $p \in \mathscr{P}_1$ whenever p satisfies (9) or (11).

If (9) holds, then (11) clearly holds. To demonstrate (10), for contradiction, let $x \geq 0$ and $h > 0$. By (9), $p(x+h) \geq |p(x)| \geq p(x)$. Hence, $p(x+h) - p(x) \geq 0$. Dividing by h and letting $h \to 0^+$ shows that $p' \in \mathscr{P}_1$.

Assume that $a, b \in \mathbb{R}$, with $a \geq |b|$. By assumption, $p' \in \mathscr{P}_1$ and so p is increasing on $[0, \infty)$. Using this and (11), we obtain $p(a) \geq p(|b|) \geq |p(b)|$. □

Recall that if $f : \mathbb{C} \longrightarrow \mathbb{C}$, then

$$f_e(x) := \frac{f(x) + f(-x)}{2}$$

is called the *even-part of* f and

$$f_o(x) := \frac{f(x) - f(-x)}{2}$$

is called the *odd-part of* f.

Proposition 9. *If $p : \mathbb{C} \longrightarrow \mathbb{C}$, then p satisfies* (11) *if and only if $p_e, p_o \in \mathscr{P}_1$.*

Proof. Notice that

$$\begin{aligned} p_e, p_o \in \mathscr{P}_1 &\iff \frac{p(x) + p(-x)}{2} \geq 0 \text{ and } \frac{p(x) - p(-x)}{2} \geq 0, \ \forall x \geq 0 \\ &\iff p(x) + p(-x) \geq 0 \text{ and } p(x) - p(-x) \geq 0, \ \forall x \geq 0 \\ &\iff p(x) \geq |p(-x)|, \ \forall x \geq 0. \square \end{aligned}$$

Theorem 10. *Conditions* (5) *and* (6) *are independent.*

Proof. If $p(x) = x^5 - 2x^3 + 2x$, $\rho = 1$, and $\mu = .5$, then

$$\rho p(-\mu) + \mu p(\rho) = -0.78125 < 0,$$

i.e., p does not satisfy equation (6).

Clearly, $p_e \in \mathscr{P}_1$ since $p_e(x) = 0$. Since

$$p'(x) = 5x^4 - 6x^2 + 2 = 5\left(x^2 - \frac{3}{5}\right)^2 + \frac{1}{5}$$

and

$$p_o(x) = p(x) = x((x^2 - 1)^2 + 1^2),$$

it follows that $p', p_o \in \mathscr{P}_1$. Thus, p satisfies (5) by Propositions 8 and 9.

If $p(x) = -x$, then p does not satisfy (5). If $0 < \mu \leq \rho$, then

$$\rho p(-\mu) + \mu p(\rho) = \rho\mu - \rho\mu = 0,$$

i.e., p satisfies (6). □

Theorem 11. *If $p \in \mathbb{R}[x]$, then $p \in \mathscr{P}_2$ if and only if $p', p_e, p_o \in \mathscr{P}_1$ and p satisfies* (6).

Proof. Immediate from Theorem 7 and Propositions 8 and 9. □

Remark 12. *The preceding arguments also apply to* entire functions*; as such, $f \in \mathscr{F}_2$ if and only if $f', f_e, f_o \in \mathscr{P}_1$ and f satisfies* (6)*.*

Proposition 13. *If p is a polynomial such that p_o satisfies* (6) *and $p_e \in \mathscr{P}_1$, then p satisfies* (6)*.*

Proof. The result follows with the observation that

$$\begin{aligned}\mu p(\rho) + \rho p(-\mu) &= \mu(p_e(\rho) + p_o(\rho)) + \rho(p_e(-\mu) + p_o(-\mu)) \\ &= (\mu p_e(\rho) + \rho p_e(\mu)) + (\mu p_o(\rho) + \rho p_o(-\mu)),\end{aligned}$$

which is nonnegative by the hypotheses. □

Clark and Paparella [2, Conjecture 5.2] conjectured that $\mathscr{P}_{n+1} \subset \mathscr{P}_n$, $\forall n \in \mathbb{N}$ and showed that $\mathscr{P}_2 \subset \mathscr{P}_1$. The following result settles the conjecture when $n = 2$.

Theorem 14. $\mathscr{P}_3 \subset \mathscr{P}_2$.

Proof. Consider the polynomial $p(x) = x^4 - x^2 + x + 1$. It is known that if

$$p(x) = \sum_{k=0}^{m} a_k x^k \in \mathscr{P}_n,\ a_m \neq 0,$$

and $m \geq n-1$, then $a_k \geq 0$, $\forall k \in \{0, 1, \ldots, n-1\}$ (see Bharali and Holtz [1, Proposition 2] or Clark and Paparella [Corollary 4.2][2]). Thus, $p \notin \mathscr{P}_3$.

Since $p_o(x) = x$, it is clear that $p_o \in \mathscr{P}_1$. Notice that $p', p_e \in \mathscr{P}_1$ since

$$p'(x) = 4x^3 - 2x^2 + 1 = x[(2x-1)^2 + 1^2] + [(2x-1)^2 + 0^2]$$

and

$$p_e(x) = x^4 - x^2 + 1 = \left(x^2 - \frac{1}{2}\right)^2 + \frac{3}{4}.$$

We also have that p_o satisfies (6) since

$$\rho p_o(-\mu) + \mu p_o(\rho) = -\rho\mu + \mu\rho = 0.$$

By Proposition 13, p satisfies (6). Thus, $p \in \mathscr{P}_2$ by Theorem 11. □

We conclude by providing a novel characterization for polynomials that preserve all nonnegative circulant matrices or order two. If

$$A = \begin{bmatrix} a & b \\ b & a \end{bmatrix} = \frac{1}{2}\begin{bmatrix} (a+b)+(a-b) & (a+b)-(a-b) \\ (a+b)-(a-b) & (a+b)+(a-b) \end{bmatrix}, \tag{12}$$

then by (7) and (8)

$$p(A) = \frac{1}{2}\begin{bmatrix} p(\rho)+p(\mu) & p(\rho)-p(\mu) \\ p(\rho)-p(\mu) & p(\rho)+p(\mu) \end{bmatrix} \geq 0,$$

where $\rho = a + b$ and $\mu = a - b$. We immediately obtain the following result.

Theorem 15 (cf. [1, Theorem 10])**.** *If $p \in \mathbb{R}[x]$, then p preserves all two-by-two nonnegative circulant matrices of the form* (12) *if and only if $p', p_e, p_o \in \mathscr{P}_1$.*

Acknowledgement

We thank the anonymous referee for their careful review and thoughtful suggestions that greatly improved this work.

Bibliography

[1] G. Bharali and O. Holtz. Functions preserving nonnegativity of matrices. *SIAM J. Matrix Anal. Appl.*, 30(1):84–101, 2008.

[2] B. J. Clark and P. Paparella. Polynomials that preserve nonnegative matrices. *Linear Algebra Appl.*, 637:110–118, 2022.

[3] R. Loewy and D. London. A note on an inverse problem for nonnegative matrices. *Linear and Multilinear Algebra*, 6(1):83–90, 1978/79.

[4] P. Paparella. Matrix functions that preserve the strong Perron-Frobenius property. *Electron. J. Linear Algebra*, 30:271–278, 2015.

[5] V. Powers and B. Reznick. Polynomials that are positive on an interval. *Trans. Amer. Math. Soc.*, 352(10):4677–4692, 2000.

Ball State Undergraduate Mathematics Exchange
https://digitalresearch.bsu.edu/mathexchange
Vol. 16, No. 1 (Fall 2022)
Pages 66 – 72

Viviani's Theorem, Minkowski's Theorem and Equiangular Polygons

Elie Alhajjar , Michael Nasta*

Elie Alhajjar is an Associate Professor in the department of mathematical sciences at the United States Military Academy in West Point, NY. He teaches and mentors cadets from all academic disciplines. He is also a senior research scientist at the Army Cyber Institute where his work focuses on mathematical applications in cybersecurity.

Michael Nasta is a second lieutenant in the US Army within the Cyber Branch. He is currently working on his masters degree in engineering at MIT through the Lincoln Lab Fellowship.

Abstract

Consider a polygon $P \subset \mathbb{R}^2$ and a positive real number t. The action of dilating (or shrinking) P by a factor of t is equivalent to dilating (or shrinking) each side of P by t, while preserving the unit normal vectors to the edges. A possible variation to this task is to consider elongating or shortening each side of P by t, also keeping the unit normal vectors intact. It is not clear a priori that such a task can always be accomplished. The current work addresses this adaptation and draws a connection with Viviani's theorem and equiangular polygons. The main purpose of the paper is to highlight a famous theorem of Minkowski from convex geometry that makes this connection possible and gives a generalization to higher dimensions.

***Corresponding author**: elie.alhajjar@westpoint.edu

1 Introduction

Let P be a regular polygon of side length s. Then, dilating P by a factor $t > 1$ is the same as adding $(t-1)s$ to each edge, and shrinking P by a factor of $t < 1$ is the same as subtracting $(1-t)s$ from each edge. Take the square S of edge length 3 as an example. For $t = 2$, the square $2S$ has edge length $6 = 3 + (2-1)(3)$ and for $t = \frac{1}{3}$, the square $\frac{1}{3}S$ has edge length $1 = 3 - (1 - \frac{1}{3})(3)$.

It is not hard to see that the two problems are generally equivalent in the case of regular polygons. What happens when P is not regular? If P is the trapezoid with side lengths $5, 5, 5, 11$, then it is impossible to add a real number t to each of the sides while keeping the edge normal vectors intact (the reader is encouraged to try it on their own).

To this end, we aim at connecting two seemingly unrelated theorems from differents historical eras of mathematics: Viviani's theorem and Minkowski's theorem. The former dates back to the mid 17-th century; it asserts that no matter where you place a point inside a regular polygon, the sum of the distances from the point to the sides of the polygon remains constant. The latter is due to Hermann Minkowski from the early 1900's; it states that every polygon (or polytope in general) is uniquely determined, up to translation, by the directions and measures of its sides (or facets in general).

2 Viviani's Theorem

Viviani's theorem states that the sum of the distances from any interior point to the sides of an equilateral triangle is independent of the position of the point. In particular, this sum is equal to the length of the height of the triangle.

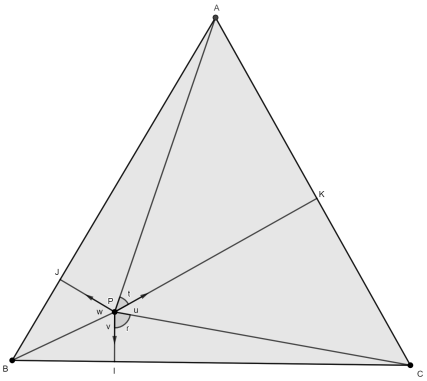


Figure 1: Equilateral Triangle ABC with an interior point P

There are many proofs and generalizations of Viviani's theorem in the literature. We survey some of them below and provide two additional elementary proofs as well. Historically, the problem of finding the Fermat point of the vertices A, B, C of a triangle ABC, i.e., the point that minimizes the sum of the distances to the vertices, was first

proposed by Fermat in a private letter to Torricelli. Torricelli solved the problem and his solution was published by his student Viviani in 1659. The solution uses the fact that the sum of the distances from any point inside an equilateral triangle to its sides is constant, which is commonly known today as Viviani's theorem.

Viviani's original proof [8] (Appendix, pp. 143-150) uses areas as follows. Let ABC be an equilaterial triangle of side length s and height length h. Let P be an interior point. The area of the triangle ABC $\left(\frac{sh}{2}\right)$ is equal to the sum of the areas of the triangles ABP, BPC, and CPA. Since $AB = AC = BC = s$, then we conclude that $PJ + PI + PK = h$ (see Figure 1). In fact, Viviani proved a bit more, namely that the sum of the distances from any point inside a regular polygon to its sides is constant, and is less than the sum from any point outside the regular polygon.

Using rotations of smaller triangles inside the equilateral triangle, Kawasaki [6] proved Viviani's theorem as illustrated in Figure 2.

Chen and Liang [3] proved the converse of Viviani's theorem: if the sum of the distances from an interior point of a triangle to its sides is independent of the location of the point, then the triangle is equilateral. Moreover, they showed that the sum of the distances from an interior point to the sides of a quadrilateral is constant if and only if the quadrilateral is a parallelogram.

The area method highlighted in Viviani's original proof can be extended to show that the theorem holds for all regular polygons as well. Likewise, by a volume argument, a similar result holds for regular polyhedra in $\mathbb{R}^3$: the sum of the distances from any point inside a regular polyhedron to its faces is independent of the location of the point.

Abboud [1] defines a polygon to have the *constant Viviani sum (CVS) property* if the sum of the distances from any interior point to the sides of the polygon is constant. He then shows that a necessary and sufficient condition for a convex polygon to have such property is the existence of three non-collinear interior points with equal sums of distances. His proof relies on ideas from linear programming.

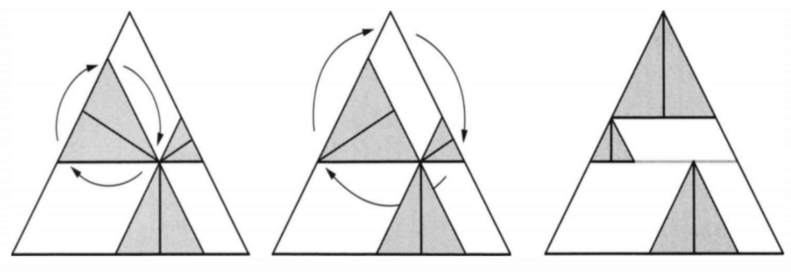


Figure 2: Kawasaki's proof using rotations [6].

We conclude this section with two new proofs of Viviani's theorem, based on simple geometric arguments.

First Proof. With the notation set above, recall that the normal vectors to the edges of the equilateral triangle ABC satisfy

$$\overrightarrow{u} + \overrightarrow{v} + \overrightarrow{w} = \overrightarrow{0}. \tag{1}$$

After taking the dot product with the vector $\overrightarrow{PA}$, we get:

$$\begin{aligned}
\overrightarrow{PA}.(\overrightarrow{u} + \overrightarrow{v} + \overrightarrow{w}) &= \overrightarrow{PA}.\overrightarrow{0} \\
\overrightarrow{PA}.\overrightarrow{u} + \overrightarrow{PA}.\overrightarrow{v} + \overrightarrow{PA}.\overrightarrow{w} &= 0 \\
PK + PA\Big(\cos\Big(\frac{2\pi}{3} + \angle APK\Big)\Big) + PJ &= 0 \\
PK + PA\Big(-\frac{1}{2}\cos\angle APK - \frac{\sqrt{3}}{2}\sin\angle APK\Big) + PJ &= 0 \\
PK - \frac{1}{2}PA\cos\angle APK - \frac{\sqrt{3}}{2}PA\sin\angle APK + PJ &= 0 \\
PK - \frac{1}{2}PK - \frac{\sqrt{3}}{2}AK + PJ &= 0.
\end{aligned}$$

This leads to the following result

$$\frac{1}{2}PK + PJ - \frac{\sqrt{3}}{2}AK = 0. \tag{2}$$

Similarly, multiplying Equation (1) by the vector $\overrightarrow{PC}$, we obtain

$$\frac{1}{2}PK + PI - \frac{\sqrt{3}}{2}CK = 0. \tag{3}$$

Finally, adding Equations (2) and (3), we get

$$\begin{aligned}
\frac{1}{2}PK + PJ - \frac{\sqrt{3}}{2}AK + \frac{1}{2}PK + PI - \frac{\sqrt{3}}{2}CK &= 0 \\
PK + PJ + PI &= \frac{\sqrt{3}}{2}(AK + CK) \\
PK + PJ + PI &= \frac{\sqrt{3}}{2}AC \\
PK + PJ + PI &= h.
\end{aligned}$$

□

Next we embed Figure 1 in cartesian coordinates and provide yet another proof of Viviani's theorem.

Second Proof. Without loss of generality, we may assume that $B(0,0)$ and $C(s,0)$ for some positive real number s. Since ABC is an equilateral triangle, the point A has coordinates $\left(\frac{s}{2}, \frac{s\sqrt{3}}{2}\right)$ and the line segments BC, AB, AC have equations $y = 0$, $\sqrt{3}x - y = 0$, $\sqrt{3}x + y - \sqrt{3}s = 0$, respectively. Let $P(x,y)$ be a point inside the triangle ABC. Using the formula for the distance from a point to a line, we get $PI = y$, $PK = \frac{-\sqrt{3}x - y + \sqrt{3}s}{2}$, and $PJ = \frac{\sqrt{3}x - y}{2}$. Adding the three lengths together leads to $PK + PJ + PI = \frac{\sqrt{3}}{2}s = h$. □

As a side note, we mention a remarkable application of Viviani's theorem in chemistry.

Consider a mixture of three chemical components represented by the vertices of an equilateral triangle. If the height of the triangle is taken as unity and the mixture is depicted by a point inside the triangle, then the distances from this point to the sides correspond to the proportions of the components in the mixture. The same principles can be applied to a system of four components: within a regular tetrahedron whose vertices represent the pure components, the distances from an interior point to the faces again sum to a constant, and may be used to represent the proportions. For further details, the reader is referred to the book [4] (Chapter 8).

3 Minkowski's Theorem

Polytopes are the generalization of polygons in higher dimensions. Formally, a convex polytope is the convex hull of a finite set of points in $\mathbb{R}^n$, or equivalently, the intersection of a finite number of hyperplanes.

The Minkowski problem for polytopes concerns the following specific question. Given a collection $\overrightarrow{u_1},\ldots,\overrightarrow{u_k}$ of unit vectors and $a_1,\ldots,a_k>0$, under what condition(s) does there exist a polytope P having the $\overrightarrow{u_i}$'s as its facet normal vectors and the a_i's as its facet areas? The answer to this question is known as the Minkowski's existence and uniqueness theorem for polytopes.

Theorem 1 (Minkowski)**.** *Let $\overrightarrow{u_1},\ldots,\overrightarrow{u_k}$ be unit vectors that span $\mathbb{R}^n$, and $a_1,\ldots,a_k>0$. Then there exists a polytope P in $\mathbb{R}^n$ having facet unit normal vectors $\overrightarrow{u_1},\ldots,\overrightarrow{u_k}$ and corresponding facet areas $a_1,\ldots,a_k$ if and only if*

$$a_1\overrightarrow{u_1}+\cdots+a_k\overrightarrow{u_k}=\overrightarrow{0}. \tag{4}$$

Moreover, this polytope is unique up to translation.

Minkowski's original proof involves two steps. First, the existence of a polytope satisfying the given facet data is demonstrated by a linear optimization argument. In the second step, the uniqueness of that polytope (up to translation) is shown by a generalized isoperimetric inequality for mixed volumes. Alternative proofs, generalizations, and applications of Minkowski's theorem are abundant in the literature. We refer the reader to [7] and the references therein for a good exposition on this topic.

Note that in the 2-dimensional Euclidean space, the facet areas of a polygon are simply the edge lengths of the polygon. In the case of equilateral triangles, it is clear that Equation (4) is equivalent to Equation (1).

A special family of polygons are the *equiangular* polygons. These are characterized by having equal angles without necessarily having congruent edges. For a set of positive real numbers $a_1,\ldots,a_k$, it is well known [2] that there exists an equiangular polygon with side lengths $a_1,\ldots,a_k$ if and only if the polynomial $a_1+a_2x+\cdots+a_kx^{k-1}$ vanishes at $e^{\frac{2\pi}{k}i}$. Hence, for example, equilateral triangles are the only equiangular triangles and rectangles are the only equiangular quadrilaterals.

We prove this result using Minkowski's theorem as follows. Let P be a polygon in $\mathbb{R}^2$ with side lengths $a_1,\ldots,a_k$ and interior angle measures $\theta_1,\ldots,\theta_k$ ($k\geq 3$). Recall that $\theta_1+\cdots+\theta_k=(k-2)\pi$ for any k-gon. Consider the following polynomial in $k-1$

variables

$$p(x_1,x_2,\ldots,x_{k-1}) := a_1 + a_2x_1 + \cdots + a_kx_1\ldots x_{k-1}$$
$$= a_1 + \sum_{i=1}^{k-1} a_{i+1}x_1\ldots x_i.$$

With the definition above, Minkowski's theorem in dimension 2 can be written in algebraic form. Observe that the angle formed by the vectors $\overrightarrow{u_j}$ and $\overrightarrow{u_{j+1}}$ is equal to $\pi-\theta_j$, for each j. Since each $\overrightarrow{u_j}$ is a unit vector, we can then write the vector $\overrightarrow{u_j} = e^{i[(\pi-\theta_1)+(\pi-\theta_2)+\cdots+(\pi-\theta_{j-1})]}$ for $j=2,\ldots,k$ (we consider $\overrightarrow{u_1}$ the vector of reference here). By substituting the latter expression of $\overrightarrow{u_j}$ in Equation (4), we get the following theorem.

Theorem 2. *Let $a_1,\ldots,a_k$ and $\theta_1,\ldots,\theta_k$ be positive real numbers such that $\theta_1+\cdots+\theta_k=(k-2)\pi$. Then, there exists a polygon with edge lengths $a_1,\ldots,a_k$ and interior angle measures $\theta_1,\ldots,\theta_k$ if and only if the polynomial $p(x_1,x_2,\ldots,x_{k-1})$ vanishes at $(e^{i(\pi-\theta_1)},e^{i(\pi-\theta_2)},\ldots,e^{i(\pi-\theta_{k-1})})$.*

If P is equiangular, then $\theta_1=\cdots=\theta_k=\frac{k-2}{k}\pi$. This implies that $\pi-\theta_i=\frac{2\pi}{k}$ for $i=1,\ldots,k$. The following can then be deduced.

Corollary 3. *There exists an equiangular polygon with edge lengths $a_1,\ldots,a_k>0$ if and only if $a_1+a_2e^{\frac{2\pi}{k}i}+a_3e^{\frac{4\pi}{k}i}+\cdots+a_ke^{\frac{2(k-1)\pi}{k}i}=0$.*

4 Viviani Polytopes

Similar to the CVS property defined above, Zhou [9] introduced the notion of Viviani polytopes as follows. Let $p_1,\ldots,p_k$ be distinct hyperplanes enclosing a convex polytope $P\subset\mathbb{R}^n$, and $\overrightarrow{u_1},\ldots,\overrightarrow{u_k}$ the outward unit normal vectors to each p_i, respectively. For a point $T\in\mathbb{R}^n$, denote by d_i the signed distance from T to the hyperplane p_i and let $v(P):=\sum_{i=1}^k d_i$. We call P a *Viviani polytope* if v is a constant function, i.e. independent of the choice of the point T.

The main result in [9] is a geometric characterization of Viviani polytopes in any dimension. An algebraic characterization using linear programming was previously derived in [1].

Theorem 4 (Theorem 1 in [9])**.** *With the above notation, a polytope $P\subset\mathbb{R}^n$ is Viviani if and only if*

$$\overrightarrow{u_1}+\cdots+\overrightarrow{u_k}=\overrightarrow{0}. \tag{5}$$

In light of Theorem 2, a polynomial formulation for Viviani polygons can be derived as follows. Given a set of positive real numbers $\theta_1,\ldots,\theta_k$ that add up to $(k-2)\pi$, there exists a polygon with interior angle measures $\theta_1,\ldots,\theta_k$ if and only if $(e^{i(\pi-\theta_1)},e^{i(\pi-\theta_2)},\ldots,e^{i(\pi-\theta_{k-1})})$ is a root of the polynomial $1+x_1+x_1x_2+\cdots+x_1x_2\ldots x_{k-1}$. In particular, we get the following corollary.

Corollary 5. *Equilateral triangles are the only Viviani triangles and parallelograms are the only Viviani quadrilaterals. Moreover, equiangular polygons are Viviani for any number of sides.*

As mentioned in the first section, it can be shown that regular polygons in $\mathbb{R}^2$ and regular polyhedra in $\mathbb{R}^3$ are Viviani using an area and a volume argument, respectively. Along the same line of thought, it was shown in [5] that any polyhedron with faces of equal area is Viviani. We extend this result to all dimensions using Minkowski's theorem.

Consider a polytope $P \subset \mathbb{R}^n$ with facet unit normal vectors $\overrightarrow{u_1}, \ldots, \overrightarrow{u_k}$. If the facets of P have equal area (i.e. $(n-1)$-dimensional volume), then $a_1 = \cdots = a_k$ in the statement of Theorem 1, which implies that $\overrightarrow{u_1} + \cdots + \overrightarrow{u_k} = \overrightarrow{0}$. By Theorem 3, one can deduce that the polytope P is Viviani. Thus, we proved the following general result.

Theorem 6. *Any polytope whose facets have equal area is Viviani.*

Finally, we turn back to our original question. Assume $P \subset \mathbb{R}^2$ is a polygon with side lengths $s_1, \ldots, s_k$ and unit normal vectors $\overrightarrow{u_1}, \ldots, \overrightarrow{u_k}$. The goal is to find another polygon P' with the same unit normal vectors but with side legths $s_1 \pm t, \ldots, s_k \pm t$. Applying Minkowski's theorem to P and P', we get $s_1\overrightarrow{u_1} + \cdots + s_k\overrightarrow{u_k} = \overrightarrow{0}$ and $(s_1 \pm t)\overrightarrow{u_1} + \cdots + (s_k \pm t)\overrightarrow{u_k} = \overrightarrow{0}$, respectively. Combining the two equations, we obtain $\pm t(\overrightarrow{u_1} + \cdots + \overrightarrow{u_k}) = \overrightarrow{0}$. This is equivalent to P (or P') being Viviani!

Bibliography

[1] E. Abboud, Viviani's theorem and its extension, *College Math. J.* 41 (2010) 203-211.

[2] M. Bras-Amoros and M. Pujol, Side Lengths of Equiangular Polygons (as seen by a coding theorist), *The American Mathematical Monthly* 122(5) (2015) 476-478.

[3] Z. Chen and T. Liang, The converse of Viviani's theorem, *College Math. J.* 37 (2006) 390–391.

[4] B. W. Darvell, Materials Science for Dentistry, *Woodhead Publishing Series in Biomaterials, tenth edition* (2009) 197-213.

[5] M. De Villiers, 3D Generalisations of Viviani's theorem, *The Mathematical Gazette*, 97 (2013) 441-445.

[6] K. Kawasaki, Proof without words: Viviani's theorem, *Math. Mag.* 78 (2005) 213.

[7] D. A. Klain, The Minkowski problem for polytopes, *Advances in Mathematics* 185 (2004) 270–288.

[8] V. Viviani, De Maximis et Minimis, (1659) available at *http://www.math.uni-bielefeld.de*.

[9] L. Zhou, Viviani Polytopes and Fermat Points, *College Math. J.* 43 (2012) 309-312.

Ball State Undergraduate Mathematics Exchange
`https://digitalresearch.bsu.edu/mathexchange`
Vol. 16, No. 1 (Fall 2022)
Pages 73 – 85

Nonlinear Lotka-Volterra Competition Models

*Mara Smith**

Mara Smith is an undergraduate student in her final year at Indiana Wesleyan University. She is studying mathematics and honors humanities. This research was performed during her junior year under the supervision of Dr. Melvin Royer.

Abstract

The classical Lotka-Volterra equations that model the interactions between two species competing for a limited resource have many potential modifications to improve biological accuracy; this paper explores modifications to the exponent of the competition term. After an introduction to the behavior of the classical Lotka-Volterra model is given, a nonlinear modification to this model by Taylor and Crizer is discussed. In section 2, an extension of this modification is proposed, in which the population variable of the competition term is raised first to the power of positive real numbers and, next, small integers. A proof is offered that at most 3 coexistent equilibrium points exist for any positive exponent values, and additional proofs further limit the number of equilibria for certain exponent and parameter values. In section 3, we prove that, in such models, the stability of the equilibria alternates between stable and unstable when considered in a northwest to southeast configuration. Combining these results allows us to describe the equilibrium behavior of a broad class of competition models.

1 Introduction

Competition models consider scenarios involving two species that compete for the same limited prey or other vital resources. In 1925, American biophysicist Alfred Lotka and Italian mathematician Vito Volterra proposed one of the first valid competition models to describe cases of coexistence or competitive exclusion [4].
A competition model implies a reciprocal, negative interaction between the two species. Further, the Lotka-Volterra model treats the competition as density-dependent, and the

***Corresponding author**: mara.smith@myemail.indwes.edu

equations include terms for both intraspecific and interspecific competition:

$$\begin{cases} \dfrac{dx}{dt} = \beta_1 x(K_1 - x - \mu_1 y) \\ \dfrac{dy}{dt} = \beta_2 y(K_2 - y - \mu_2 x), \end{cases} \tag{1}$$

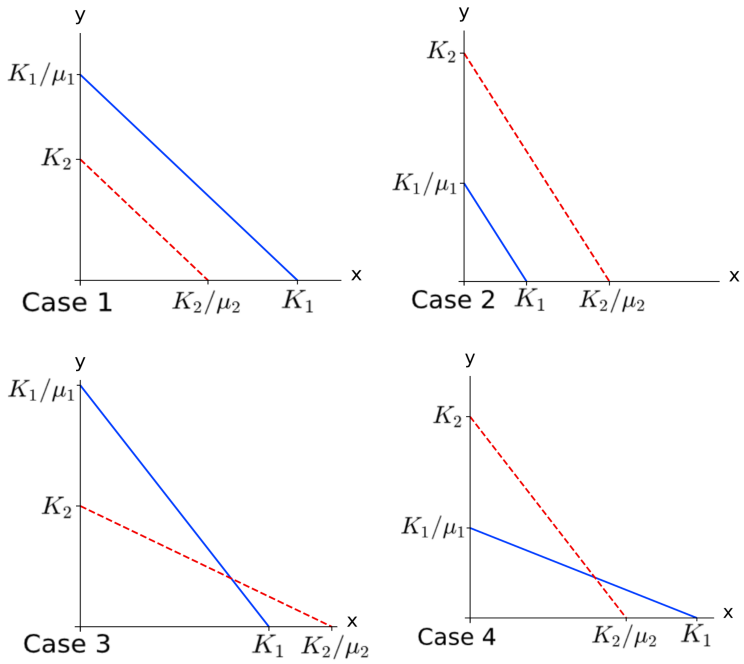


Figure 1: The isocurves of the classical Lotka-Volterra equations.

where β_i, K_i, and μ_i are positive constants for $i = 1,2$ [3]. Next, $\frac{dx}{dt}$ and $\frac{dy}{dt}$ denote the growth rates of populations x and y at time t. The β_i constants are the respective intrinsic growth rates, the K_i constants are the carrying capacities, and the μ_i constants are the competition coefficients, which represent the negative effect of one species on the other. The isocurves are linear and fall into one of four cases, depending upon parameter relationships, as depicted in Figure 1. As shown, the relative values of parameters K_1 versus K_2/μ_2 and K_2 versus K_1/μ_1 determine the relative positions of the $x-$ and $y-$intercepts of the isocurves, consequently impacting the number of possible intersection points.

These classical Lotka-Volterra equations have been modified in various studies [1][4][5]. Taylor and Crizer introduce a nonlinear relationship to model the effects of each species on the other,

$$\begin{cases} \dfrac{dx}{dt} = \beta_1 x(K_1 - x - \mu_1 y^2) \\ \dfrac{dy}{dt} = \beta_2 y(K_2 - y - \mu_2 x^2), \end{cases} \tag{2}$$

where β_i, K_i, and μ_i again are positive constants for $i = 1,2$ [5]. Because the isocurves

are nonlinear, they are not limited to a maximum of one intersection point, as addressed in [4].
In this paper, we examine the more general nonlinear relationship

$$\begin{cases} \dfrac{dx}{dt} = \beta_1 x(K_1 - x - \mu_1 y^{w_1}) \\ \dfrac{dy}{dt} = \beta_2 y(K_2 - y - \mu_2 x^{w_2}), \end{cases} \tag{3}$$

where β_i, K_i, and μ_i are positive constants and w_i is any positive real number for $i = 1,2$. We first determine that the four cases of parameter relationships contain several subcases, which will be investigated in Section 2. We establish the number of possible intersection points for any positive real exponents on the competition terms. Next, we prove that isocurves for any exponents are limited to a maximum of three intersection points, with an even smaller number allowed for small positive integer exponents. Finally, we detail the stability patterns of these equilibria and their biological implications.

2 Number of Intersection Points

2.1 Exponents as Positive Real Numbers

Given the modified Lotka-Volterra equations in equation (3), the equilibrium points are given by $(0,0),(K_1,0),(0,K_2)$ and positive solutions to the following system:

$$\begin{cases} x + \mu_1 y^{w_1} = K_1 \\ y + \mu_2 x^{w_2} = K_2. \end{cases} \tag{4}$$

Defining $v_1 = \sqrt[w_1]{K_1/\mu_1}$ and $v_2 = \sqrt[w_2]{K_2/\mu_2}$, we have the following case divisions:
Case 1: $K_1 > v_2$ and $K_2 < v_1$
Case 2: $K_1 < v_2$ and $K_2 > v_1$
Case 3: $K_1 < v_2$ and $K_2 < v_1$
Case 4: $K_1 > v_2$ and $K_2 > v_1$.
We consider these cases as they mirror the four cases seen in the original Lotka-Volterra equations, defining the relative intercept positions of the isocurves. We also define

$$F(x,y) = \beta_1(K_1 - x - \mu_1 y^{w_1})$$
$$G(x,y) = \beta_2(K_2 - y - \mu_2 x^{w_2})$$

and let f and g denote the curves $F(x,y) = 0$ and $G(x,y) = 0$, respectively.

Lemma 1. Isocurves f and g are monotonically decreasing.

Proof. Using equation (4), we see that the two terms on the left hand side of each equation add to K_i, a fixed constant. Hence, in both equations, as x increases, y decreases, causing both curves to decrease monotonically. □

We then solve the second equation from the set in (4) for y and substitute the result into $x + \mu_1 y^{w_1} = K_1$. Defining this result as a function of x, we obtain $h(x) = 0$ where

$$h(x) = K_1 - x - \mu_1(K_2 - \mu_2 x^{w_2})^{w_1}.$$

The roots of this equation give the x-coordinate of any intersection points of the isocurves f and g. Taking the first derivative with respect to x, we have

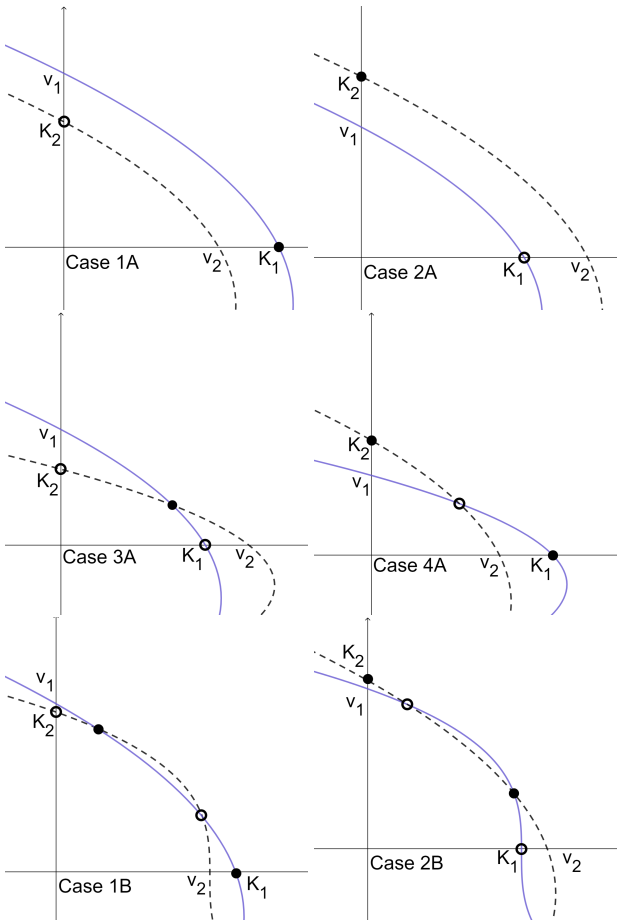
$$h'(x) = -1 + \mu_1\mu_2 w_1 w_2 x^{w_2-1}(K_2 - \mu_2 x^{w_2})^{w_1-1}.$$

Again taking the derivative, we have

$$h''(x) = \mu_1\mu_2 w_1 w_2 x^{w_2-2}(K_2 - \mu_2 x^{w_2})^{w_1-2}[K_2(w_2-1) - \mu_2 x^{w_2}(w_1 w_2 - 1)].$$

The roots and undefined points of $h''(x)$ give the x-coordinate of any inflection points of $h(x)$, allowing us to determine the maximum number of critical points and therefore zeros of $h(x)$.

Theorem 1. *For any w_1 and w_2, f and g have a maximum of 3 intersection points in the interior of the first quadrant.*



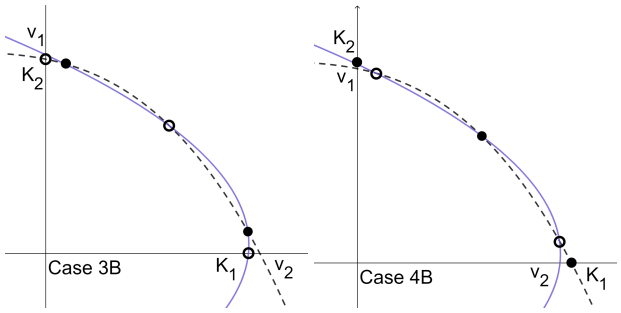


Figure 2: The 4 possible cases, each with one subcase, where f is dashed and g is solid. As shown in Section 3, solid points are stable equilibria, and open points are unstable equilibria. The cases marked with 'A' indicate the subcase of each numbered case with 0 or 1 intersection points, and the cases marked with 'B' indicate the subcase with 2 or 3 intersection points.

Proof. As a consequence of Rolle's Theorem, to show f and g can have at most 3 intersection points, we need only to show that $h(x)$ has at most one inflection point. We see that $h''(x) = 0$ or $h''(x)$ is undefined when $x = 0$, $\sqrt[w_2]{K_2/\mu_2}$, or $\sqrt[w_2]{\frac{K_2(w_2-1)}{\mu_2(w_1w_2-1)}}$, depending on the values of w_1 and w_2. Because $x = 0$ and $\sqrt[w_2]{K_2/\mu_2}$ mark points where f or g would intersect a coordinate axis, the only inflection point that could produce a coexistent equilibria is at $x = \sqrt[w_2]{\frac{K_2(w_2-1)}{\mu_2(w_1w_2-1)}}$. Hence, there is a maximum of one inflection point for $h(x)$. □

Theorem 2. *There are 8 possible configurations of the graphs of f and g, shown in Figure 2.*

Proof. Because the isocurves are continuous and can intersect a maximum of 3 times in the first quadrant, the geometric positions of the intercepts K_i and v_i determine which of the 8 configurations are possible, and their limited intersections limit the number of configurations. □

2.2 Cases of Small Integer Exponents

Theorem 3. *If $w_1 = 2$ and if w_2, written from here on as w for simplicity, is any integer greater than or equal to 2,*
(1) In case 1, equation (4) has in the first quadrant either 0 or 2 solutions.
(2) In case 2, equation (4) has in the first quadrant either 0 or 2 solutions.
(3) In case 3, equation (4) has in the first quadrant either 1 or 3 solutions.
(4) In case 4, equation (4) has in the first quadrant exactly 1 solution.

Proof. Following the work of [4], to find the number of intersection points of the isocurves, we begin by obtaining the polynomials, derived from equation (4), which are satisfied by the equilibrium solutions. Starting with $y + \mu_2 x^w = K_2$, we solve for y and substitute into $x + \mu_1 y^2 = K_1$ to obtain $x + \mu_1(K_2 - \mu_2 x^w)^2 = K_1$, which expands to:

$$\mu_1\mu_2^2 x^{2w} - 2\mu_1\mu_2 K_2 x^w + x + \mu_1 K_2^2 - K_1 = 0. \tag{5}$$

Similarly, we next start with $x+\mu_1 y^2 = K_1$, solve for x, and substitute the result into $y+\mu_2 x^w = K_2$ to obtain $y+\mu_2(K_1-\mu_1 y^2)^w - K_2 = 0$, which, using the Binomial Theorem, expands to:

$$y+\mu_2(K_1-\mu_1 y^2)^w - K_2 = y - K_2 + \sum_{k=0}^{w} \binom{w}{k} \mu_2 K_1^{w-k}(-\mu_1 y^2)^k$$

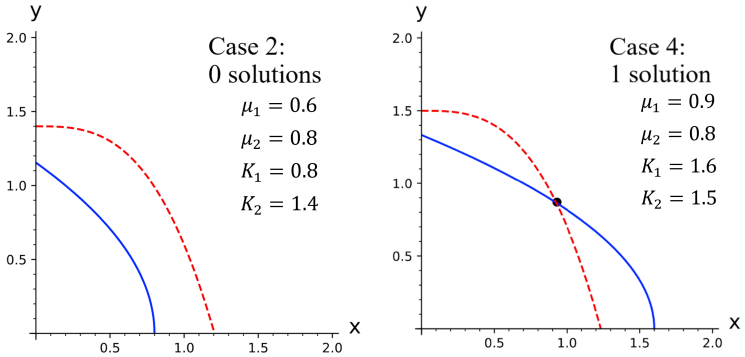
$$= \sum_{k=1}^{w} \binom{w}{k} \mu_2 K_1^{w-k}(-\mu_1 y^2)^k + y + \mu_2 K_1^w - K_2 = 0, \tag{6}$$

where $k \in \mathbb{Z}$. Because the term $\mu_1 y^2$ is preceded by a negative sign, the sign of each term from the summation will alternate. As we will be using Descartes' Rule of Signs later in this proof, it is notable that the polynomial in equation (5) has either 1 or 2 sign changes depending on the sign of $\mu_1 K_2^2 - K_1$, and the polynomial in equation (6) has either w or $w+1$ sign changes, dependent upon the sign of $\mu_2 K_1^w - K_2$. It is also notable that the only coefficients whose sign is dependent upon the parameter values are $\mu_2 K_1^w - K_2$ for any w. Hence, the only information required to determine the number of sign changes in the polynomial defined by equation (6) are the signs of $\mu_1 K_2^2 - K_1$ and $\mu_2 K_1^w - K_2$.

From Theorem 2, we see that, for cases 1, 2, and 3, all 6 possible ways of intersection could occur, as Descartes' Rule of Signs eliminates no possibilities.

In case 4, $\mu_2 K_1^w - K_2 > 0$ and $\mu_1 K_2^2 - K_1 > 0$. Then equation (5) has 2 sign changes and equation (6) has w sign changes. Hence, by Descartes' Rule of Signs, equation (4) has at most min(2, w) solutions in the interior of the first quadrant. The positions of the intercepts indicate that the curves must intersect an odd number of times, so equation (4) has exactly one solution, eliminating case 4B. □

As seen, the smaller of w_1 and w_2 serves as a limiting factor in determining the number of intersection points of the isocurves. Thus, when $w_1 = 2$, any integer value of $w_2 \geq 2$ will yield the same results as would $w_1 = w_2 = 2$. Further, the case in which $w_2 = 2$ and $w_1 \geq 2$ is symmetric and yields the same number of intersection points in the symmetric cases.



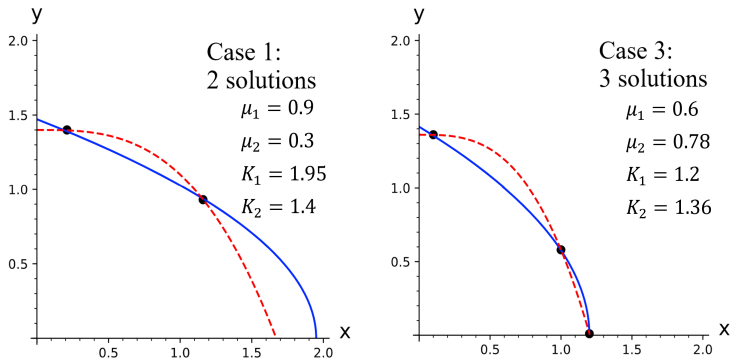


Figure 3: Examples of the subcases for $w_1 = 2$, $w_2 = 3$, and $\beta_1 = \beta_2 = 1$ resulting in 0, 1, 2, and 3 intersection points.

Theorem 4. *If $w_1 = 1$ and if w_2, written from here on as w for simplicity, is any integer greater than or equal to 1,*
(1) In case 1, equation (4) has in the first quadrant either 0 or 2 solutions.
(2) In case 2, equation (4) has in the first quadrant 0 solutions.
(3) In case 3, equation (4) has in the first quadrant exactly 1 solution.
(4) In case 4, equation (4) has in the first quadrant exactly 1 solution.

Proof. Following the arguments in the proof for Theorem 3, solutions to equation (4) under the above conditions will satisfy the following equations:

$$-\mu_1\mu_2 x^w + x + \mu_1 K_2 - K_1 = 0, \tag{7}$$

and

$$y + \mu_2(K_1 - \mu_1 y)^w - K_2 = 0. \tag{8}$$

(1) In case 1, we see from Theorem 2 that either case 1A or 1B could occur.

(2) In case 2, $\mu_2 K_1^w - K_2 < 0$ and $\mu_1 K_2 - K_1 > 0$. This yields 1 sign change from equation (7) and either w or $w+1$ sign changes from equation (8). Thus, equation (4) has at most min(1, w) or min(1, $w+1$) solutions, or at most 1. From the intercepts, there must be an even number of intersection points, so there are 0 positive solutions, eliminating case 2B.

(3) In case 3, $\mu_2 K_1^w - K_2 < 0$ and $\mu_1 K_2 - K_1 < 0$. This yields 2 sign changes from equation (7) and either w or $w+1$ sign changes from equation (8). Thus, equation (4) has at most min (2, w) or min(2, $w+1$) solutions, or at most 2. From the intercepts, there must be an odd number of intersection points, so there is exactly 1 positive solution, eliminating case 3B.

(4) In case 4, $\mu_2 K_1^w - K_2 > 0$ and $\mu_1 K_2 - K_1 > 0$. This yields 1 sign change from equation (7) and either n or $w+1$ sign changes from equation (8). Thus, equation (4) has at most min(1, w) or min(1, $w+1$) solutions, or at most 1. From the intercepts, there must be an odd number of intersection points, so there is exactly 1 positive solution, eliminating case 4B. □

The case where one of the exponents is zero is well-known and will not be discussed here.

While Theorem 1 indicates that any parameter combination results in a maximum of 3 intersections and Theorems 3 and 4 show how possible cases may be eliminated, we can perform further analysis to determine further restrictions on the number of intersection points by case divisions.

Theorem 5. *The table below indicates when $h(x)$ has one inflection point inside the interval of interest, which indicates when f and g can have 3 intersection points. Letting $m = \sqrt[w_2]{\frac{K_2(w_2-1)}{\mu_2(w_1w_2-1)}}$:*

	Cases 1 and 4	**Cases 2 and 3**
A: $w_1 < 1, w_2 < 1$	always	$m < K_1$
B: $w_1 > 1, w_2 > 1$	always	$m < K_1$
C: $w_1 < 1, w_2 > 1$	never	never
D: $w_1 > 1, w_2 < 1$	never	never

Table 1: The conditions under which $h(x)$ has one inflection point.

Proof. Recall that when $h(x)$ has one inflection point, it can have at most 3 zeros and, accordingly, f and g can intersect at most 3 times. When $h(x)$ has no inflection points, f and g are limited to a maximum of 2 intersection points. Also note that we are only interested in intersection points with the $x-$coordinate in the interval $0 < x < \min(K_1, v_2)$.

(A) We see that $\sqrt[w_2]{\frac{K_2(w_2-1)}{\mu_2(w_1w_2-1)}}$ will under the radical have a negative numerator and negative denominator, resulting in a positive radicand and a real x value. Then since $\frac{w_2-1}{w_1w_2-1} < 1$, $m < \sqrt[w_2]{\frac{K_2}{\mu_2}}$ and is therefore inside the interval of interest for cases 1 and 4 of the divisions of parameter relationships. Hence, there is one positive inflection point in cases 1 and 4. In cases 2 and 3, only the values of $m < K_1$ are inside the interval of interest. Hence, there is one inflection point in cases 2 and 3 when $m < K_1$ and zero inflection points when $m > K_1$.

(B) We see that the radicand is positive, resulting in a real x value. Since $\frac{w_2-1}{w_1w_2-1} < 1$, $m < \sqrt[w_2]{\frac{K_2}{\mu_2}}$ and is therefore inside the interval of interest for all cases. Hence, there is one inflection point. Following the same reasoning as above, there is one inflection point in cases 1 and 4. In cases 2 and 3, there is one inflection point when $m < K_1$ and zero inflection points when $m > K_1$.

(C) The radicand may be either positive or negative. If $w_1w_2 < 1$, the radicand is negative and the solution has no real parts, resulting in zero inflection points. If $w_1w_2 > 1$, the radicand is positive and $\frac{w_2-1}{w_1w_2-1} > 1$, meaning that $m > \sqrt[w_2]{\frac{K_2}{\mu_2}}$ and is always outside our interval of interest. Hence, both possibilities yield zero inflection points.

(D) The radicand may be either positive or negative. If $w_1w_2 < 1$, the radicand is negative and there are zero inflection points. If $w_1w_2 > 1$, the radicand is positive

and $\frac{w_2-1}{w_1w_2-1} > 1$, meaning that $m > \sqrt[w_2]{\frac{K_2}{\mu_2}}$ and is always inside our interval of interest. Hence, there are again zero inflection points. □

This exploration of the numbers of inflection points places restrictions on the number of possible intersection points of the two isocurves. We have shown above that for any combination of positive exponent values, there is a maximum of one inflection point for the polynomial whose zeros give the $x-$coordinate of intersection points of f and g; hence, this polynomial has a maximum of three roots, indicating that f and g are limited to a maximum of 3 intersection points in the interior of the first quadrant. From above, we now know that the appearance of 3 equilibria depends on the relative values of $\sqrt[w_2]{\frac{K_2(w_2-1)}{\mu_2(w_1w_2-1)}}$ and K_1 in cases 2 and 3. Further, there will never be 3 equilibria in cases when both $w_1 < 1$ and $w_2 > 1$ or $w_1 > 1$ and $w_2 < 1$. In the cases that result in zero inflection points, the isocurves have a maximum of 2 intersection points. While the complicated parameter relationships make it difficult to offer distinct value ranges for w_1 and w_2 that yield specific numbers of intersection points, we have found restrictions for the maximum numbers of equilibria and the cases in which they may occur.

3 Stability of Equilibria

The dynamic stability of equilibria is significant as only stable equilibria are realistic points where the populations can be maintained in equilibrium. The sample population trajectories in Figure 4 illustrate that coexistent equilibria can either be stable or unstable. Following the work of Hirsch, Smale, and Devaney in [2], we offer a proof regarding the stability of equilibrium points in the interior of the first quadrant.

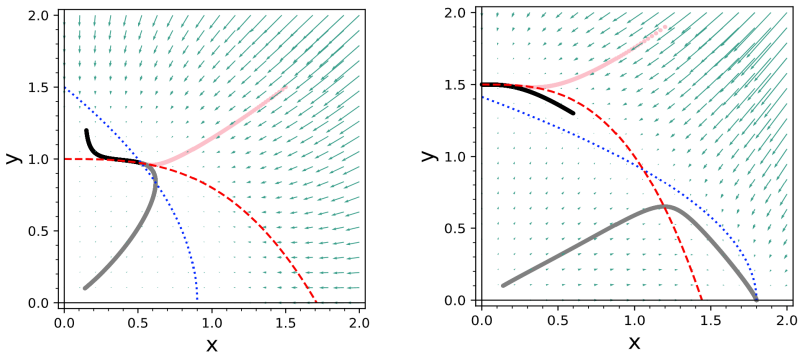


Figure 4: Left: Sample trajectories for case 1.
Right: Sample trajectories for case 2.

We note the following facts regarding F and G:

$F1$. The populations of the two species x and y are inversely related; if the population of one increases, then the growth rate of the other decreases. Thus, $F_y < 0$ and $G_x < 0$.

$F2$. If either population reaches a large value, then both populations decrease. In particular, letting $K = \max\{K_1, K_2, (K_1/\mu_1)^{1/w_1}, (K_2/\mu_2)^{1/w_2}\}$, we have that $F(x,y) <$

0 and $G(x,y) < 0$ if $x \geq K$ or $y \geq K$.

$F3$. If the population of one species is zero, then the other species has a positive growth rate to a certain population value and a negative growth rate beyond it. In particular, $F(x,0)$ is positive when $x < K_1$ and negative when $x > K_1$, and $G(0,y)$ is positive when $y < K_2$ and negative when $y > K_2$.

Theorem 6. *Each intersection point of isocurves f and g in the interior of the first quadrant yields a locally stable equilibrium if and only if f is above g to the left of the intersection and f is below g to the right.*

Proof. Any coexistent equilibria of this system modeled by equation (3) are given by the intersection(s) of the isocurves in the interior of the first quadrant. At an intersection point, the slope of $f = -\frac{F_x}{F_y}$ and the slope of $g = -\frac{G_x}{G_y}$ by the Implicit Function Theorem. We know that any intersection points occur under one of three cases:

Case A. f is above g to the left of the intersection, and f is below g to the right.

Case B. g is above f to the left of the intersection, and g is below f to the right.

Case C. g and f are tangent to each other and touch at a point without crossing at that point.

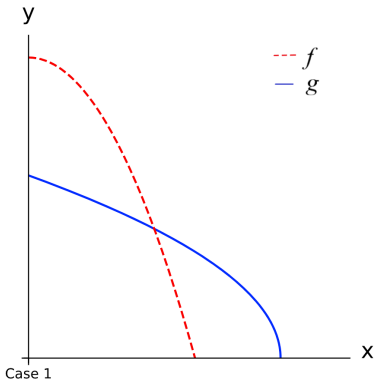


Figure 5: Case 1, where the slope of f is steeper than that of g.

(A) As shown in Figure 5, the slope of f is steeper than that of g, meaning $-\frac{F_x}{F_y} < -\frac{G_x}{G_y} < 0$, as both curves are monotonically decreasing. From fact $F1$, $F_y < 0$ and $G_x < 0$, and we conclude that $F_x < 0$ and $G_y < 0$, as we have $-\frac{F_x}{F_y} < 0$ and $-\frac{G_x}{G_y} < 0$.

To determine the local stability at this critical point, we next seek the eigenvalues of the Jacobian matrix,

$$\begin{bmatrix} F_x & F_y \\ G_x & G_y \end{bmatrix} = \begin{bmatrix} \beta_1 K_1 - 2\beta_1 x - \beta_1 \mu_1 y^{w_1} & -w_1 \beta_1 \mu_1 x y^{w_1 - 1} \\ -w_2 \beta_2 \mu_2 y x^{w_2 - 1} & \beta_2 K_2 - 2\beta_2 y - \beta_2 \mu_2 x^{w_2} \end{bmatrix}.$$

Along the isocurves, $F(x,y) = 0$ and $G(x,y) = 0$, so we substitute these values into the matrix, yielding

$$\begin{bmatrix} -\beta_1 x & -w_1 \beta_1 \mu_1 x y^{w_1 - 1} \\ -w_2 \beta_2 \mu_2 y x^{w_2 - 1} & -\beta_2 y \end{bmatrix}.$$

The trace of the Jacobian matrix is $-\beta_1 x - \beta_2 y < 0$. The determinant is $\beta_1\beta_2(xy - w_1w_2\mu_1\mu_2 x^{w_2}y^{w_1})$, which we see, generally, is $xy(F_xG_y - F_yG_x)$. In case A, $-\frac{F_x}{F_y} < -\frac{G_x}{G_y}$, meaning $\frac{F_x}{F_y} > \frac{G_x}{G_y}$ and consequently $F_xG_y > F_yG_x$, so the determinant is positive. The eigenvalues are the roots of the characteristic polynomial $p(\lambda)$ of the matrix, which, for a standard 2x2 matrix, is given by

$$p(\lambda) = \det(A - \lambda I) = \lambda^2 - \mathrm{tr}(A)\lambda + \det(A).$$

Using the quadratic formula, the eigenvalues of the general matrix are

$$\lambda = \frac{\mathrm{tr}(A) \pm \sqrt{\mathrm{tr}(A)^2 - 4\det(A)}}{2}.$$

Returning to our Jacobian matrix, we see that, since the determinant is positive, $|\mathrm{tr}(A)| > \sqrt{\mathrm{tr}(A)^2 - 4\det(A)}$, meaning that the real part of both eigenvalues must always be negative as the trace is negative. Hence, both eigenvalues have negative real parts, indicating a locally stable equilibrium point.

(B) We begin with $-\frac{G_x}{G_y} < -\frac{F_x}{F_y} < 0$. Again, the trace of the matrix is $-\beta_1 x - \beta_2 y < 0$. The determinant is $xy(F_xG_y - F_yG_x)$, which in this case is negative.

From our Jacobian matrix, we see that both the trace and determinant are negative. Hence, both eigenvalues will be real. Further, since $\sqrt{\mathrm{tr}(A)^2 - 4\det(A)} > \mathrm{tr}(A)$, one eigenvalue must be negative and the other must be positive, indicating an unstable equilibrium point.

(C) In this case, the two isocurves are tangent and touch without crossing. While this case is highly biologically improbable, we show that this tangent point yields an unstable equilibrium.

For the two curves to touch without crossing, their slopes must be equal at the point we consider; thus, we begin with $-\frac{G_x}{G_y} = -\frac{F_x}{F_y} < 0$. Since $F_xG_y = F_yG_x$, the determinant is zero. We then have

$$\lambda = \frac{\mathrm{tr}(A) \pm \sqrt{\mathrm{tr}(A)^2}}{2}.$$

The eigenvalue $\lambda = 0$ implies an unstable equilibrium at this intersection point, as at least one of the eigenvalues of this matrix has a nonnegative real part. □

Since both isocurves are monotonically decreasing, the equilibria have a well-defined order moving from a northwest to southeast direction. In this order, any one of the two possible cases of intersection with crossing isocurves cannot occur twice in a row. In other words, if more than one intersection point exists on the interior of the first quadrant, the stability of adjacent equilibria will alternate between stable and unstable when the equilibria are being considered in a northwest to southeast configuration. Therefore, knowing the stability of just one equilibrium point in these models is sufficient to determine the local stability of the rest.

Though we do not do so here, facts $F2$ and $F3$ can be used to show that the locally stable equilibria are actually globally stable, as done in [2]. To demonstrate the degree of information we can now readily obtain from modified Lotka-Volterra systems in the form of equation (3), we now discuss a numerical example. Letting $\mu_1 = 0.59, \mu_2 = 0.74, K_1 = 1.21, K_2 = 1.36, w_1 = 2$, and $w_2 = 3$, we see that $v_1 \approx 1.43$ and $v_2 \approx 1.22$.

Because $K_1 < v_2$ and $K_2 < v_1$, these parameters place us in case 3. Using the table from Theorem 5, we have that $m \approx 0.74 < K_1$ and hence these parameters yield isocurves with 3 intersection points in the first quadrant, which is case 3B. Using Theorem 6 and the aid of Figure 2, we see that the relative positions of the intercepts indicate that the first equilibrium point when considered in a northwest to southeast configuration, which is $(0, K_2)$, must be unstable. Continuing down the isocurves in this direction, the first intersection is stable, the second is unstable, the third is stable, and $(K_1, 0)$ is unstable. Hence, we have determined the number of equilibria and their stability with minimal calculations.

4 Areas of Further Research

While these explorations of the models have added insight into Lotka-Volterra modifications, there is much more to be explored. Placing additional restrictions on the parameters of Table (1), for example, would enable more efficient and clear determination of the possible number of intersection points, as the parameter relationships are clearly complicated. Additionally, the number of intersection points in Theorems 1, 3, and 4 is dependent upon parameter relationships. Further exploration of the cases in these theorems could reveal which number of intersection points actually occurs for more specific parameter values. Moreover, investigating relationships between w_1 and w_2 could offer additional restrictions on when certain numbers of equilibria occur.

Acknowledgement

I would like to thank Dr. Melvin Royer for his guidance throughout this research.

Bibliography

[1] Gavina, Maica Krizna A., Takeru Tahara, Kei-ichi Tainaka, Hiromu Ito, Satoru Morita, Genki Ichinose, Takuya Okabe, Tatsuya Togashi, Takashi Nagatani, and Jin Yoshimura (2018). “Multi-Species Coexistence in Lotka-Volterra Competitive Systems with Crowding Effects," *Scientific Reports*, 8(1), 1–8.

[2] Hirsch, Morris W., Stephen Smale and Robert L. Devaney (2004). *Differential Equations, Dynamical Systems, and an Introduction to Chaos.* San Diego, CA: Academic Press (Pure and Applied Mathematics; a Series of Monographs and Textbooks).

[3] Lotka, A. J. (1927). “Fluctuations in the Abundance of a Species considered Mathematically," *Nature*, 119 (2983), Article 12.

[4] Stover, Christopher, Jemal Mohammed-Awel, and Andreas Lazari (2009). “Investigation of the Qualitative Behavior of the Equilibrium Points for a Modified Lotka-Volterra Model," *Georgia Journal of Science*, Vol. 67, No. 2, Article 5.

[5] Taylor, Austin and Crizer, Amy (2005). “A Modified Lotka-Volterra Competition Model with a Non-Linear Relationship Between Species," *Rose-Hulman Undergraduate Mathematics Journal*: Vol. 6, Iss. 2 , Article 8.

Ball State Undergraduate Mathematics Exchange
`https://digitalresearch.bsu.edu/mathexchange`
Vol. 16, No. 1 (Fall 2022)
Pages 85 – 103

Carnot-Carathéodory and Korányi-Geodesics in the Heisenberg Group

Josh Ascher , Armin Schikorra*

Josh Ascher is an undergraduate student studying math and computer science at the University of Pittsburgh. He is in his senior year and plans to attend graduate school where he will conduct research in theoretical computer science. Josh hopes to one day become a professor so that he can continue pursuing research, as well as teaching the next generation of researchers.

Armin Schikorra received his Ph.D. from RWTH Aachen University in 2010 and is currently associate professor at the University of Pittsburgh. He works in partial differential equations which are often motivated from the geometric calculus of variations, one example being harmonic maps between manifolds and various generalizations thereof. In particular he is interested in regularity theory of such local or nonlocal equations.

Abstract

We discuss the Heisenberg group $\mathbb{H}_1$, the three-dimensional space $\mathbb{R}^3$ equipped with one of two equivalent metrics, the Korányi- and Carnot-Carathéodory metric. We show that the notion of length of curves for both metrics coincide, and that shortest curves, so-called geodesics, exist.

1 Introduction

The Heisenberg group $\mathbb{H}_1$ is a subject of intensive study, as a special case of sub-Riemannian manifolds or Carnot groups, see [2] or [1].

From the point of view of Analysis, $\mathbb{H}_1$ consists of all the points $p = (p^1, p^2, p^3) \in \mathbb{R}^3$, where $\mathbb{R}^3$ denotes the usual Euclidean three-dimensional space. However, the distance

***Corresponding author**: joa71@pitt.edu

between two points $p,q \in \mathbb{R}^3$ is given by a non-Euclidean metric $d(p,q)$. Actually, there are two typical metrics used in the Heisenberg group $\mathbb{H}_1$, and we begin by describing the first one, the *Carnot-Carathéodory*-metric $d_{cc}(p,q)$ of $\mathbb{H}_1$: Take any (for now continuously differentiable) curve $\gamma : [0,1] \to \mathbb{R}^3$ with $\gamma(0) = q$ and $\gamma(1) = p$. From calculus we know that the length of a curve is given by

$$\mathscr{L}(\gamma) = \int_{[0,1]} |\dot{\gamma}(t)| dt, \tag{1}$$

where $\dot{\gamma}$ denotes the derivative of γ. If we consider the minimal possible length of curves $\gamma : [0,1] \to \mathbb{R}^3$ that are continuously differentiable and connect p to q in the sense that $\gamma(0) = p$ and $\gamma(1) = q$, then one can show that this minimal length is exactly the Euclidean distance $|p-q|$,

$$|p-q|_{\mathbb{R}^3} = \inf_{\gamma \in X(p,q)} \mathscr{L}(\gamma)$$

where

$$X(p,q) = \left\{ \gamma : [0,1] \to \mathbb{R}^3 : \quad \text{continuously differentiable, } \gamma(0) = p,\ \gamma(1) = q \right\}.$$

The Carnot-Carathéodory metric is also the infimum of the lengths of curves connecting p and q, however those curves have to be *horizontal*, meaning that $\dot{\gamma}(t)$ has to belong to the horizontal space $H_{\gamma(t)}\mathbb{H}_1$ for each $t \in (0,1)$, which is spanned by the vectors

$$H_p\mathbb{H}_1 = \operatorname{span}\left\{ \begin{pmatrix} 1 \\ 0 \\ 2p^2 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ -2p^1 \end{pmatrix} \right\}.$$

That is for each $t \in (0,1)$ there must be some $\lambda_1(t)$ and $\lambda_2(t)$ such that

$$\dot{\gamma}(t) = \lambda_1(t) \begin{pmatrix} 1 \\ 0 \\ 2\gamma^2(t) \end{pmatrix} + \lambda_2(t) \begin{pmatrix} 0 \\ 1 \\ -2\gamma^1(t) \end{pmatrix},$$

or, taking $\lambda_1(t) = \dot{\gamma}^1(t)$ and $\lambda_2(t) = \dot{\gamma}^2(t)$, equivalently,

$$\dot{\gamma}^3(t) = -2\gamma^1(t)\dot{\gamma}^2(t) - \gamma^2(t)\dot{\gamma}^1(t) \quad \forall t \in (0,1). \tag{2}$$

For such curves we define the length

$$\mathscr{L}_{cc}(\gamma) := \int_{[0,1]} \sqrt{|\lambda_1(t)|^2 + |\lambda_2(t)|^2} dt \equiv \int_{[0,1]} \sqrt{|\dot{\gamma}_1(t)|^2 + |\dot{\gamma}_2(t)|^2} dt.$$

The Carnot-Carathéodory length $d_{cc}(p,q)$ is then given by

$$d_{cc}(p,q) = \inf_{\gamma \in Y(p,q)} \mathscr{L}_{cc}(\gamma) \tag{3}$$

where

$$Y(p,q) = \left\{ \gamma : [0,1] \to \mathbb{R}^3 : \quad \text{continuously differentiable, } \gamma(0) = p,\ \gamma(1) = q, \text{ (2) holds} \right\}.$$

Observe that this is very similar to curves γ into a Riemannian manifold $\mathscr{M} \subset \mathbb{R}^3$: any differentiable curve $\gamma : [0,1] \to \mathscr{M}$ satisfies $\gamma(t) \in T_{\gamma(t)}\mathscr{M}$, where $T_p\mathscr{M}$ is the tangent space of the manifold $\mathscr{M}$, and if we want to find the distance between two points p

and q on the manifold, it makes sense to define this distance as the minimal length of curves tangent to the manifold at every point and connecting p and q. So from this perspective, the Heisenberg group is $\mathbb{R}^3$ with a "strange" tangent plane distribution (and since it is strange we call it horizontal plane distribution instead) – the strangeness of the Heisenberg group is that its horizontal plane distribution cannot be written as a tangent space of any manifold $\mathscr{M}$, that is the horizontal plane distribution is not integrable in the sense of the Frobenius' theorem. Here is actually where the "group" of the Heisenberg group enters, the vectors spanning the horizontal space $H_p\mathbb{H}_1$ are left-invariant vector fields for a group structure – but we will not pursue this point of view further here.

It is known that for each $p,q \in \mathbb{R}^3$ the infimum in (3) is attained, i.e. there exists a shortest curve γ, called *geodesic* such that

$$\mathscr{L}_{cc}(\gamma) = d_{cc}(p,q),$$

see e.g. Hajłasz-Zimmerman [3, (1.3)]. In particular between any two points $p,q \in \mathbb{R}^3$ there exist horizontal curves. Let us remark that for more general sub-Riemannian geometry it a very deep result, called Chow–Rashevskii theorem, that $d_{cc}(p,q)$ is even *finite* for all points p,q, cf. [4].

While the above notion of distance $d_{cc}(p,q)$ is attractive from a geometric point of view, it is not easily computable (given p and q we first need to find the shortest curve γ between then, then compute its length).

The other metric we want to consider, the *Korányi-metric*, is much easier to compute. It simply is given by

$$d_K(p,q) := (|p_1 - q_1|^2 + |p_2 - q_2|^{2^2} + |p_3 - q_3 + 2(p_2q_1 - p_1q_2)|^2)^{\frac{1}{4}}$$

There is also a more group-theoretic motivation for $d_K(p,q) = \|p^{-1} * q\|_{\mathbb{H}_1}$, but we will also not pursue this aspect further here, we refer the interested reader to [1].

Any metric space naturally is equipped with a notion of length of curves, see Definition 5, which gives us the notion of a Korányi-length $\mathscr{L}_K(\gamma)$.

We will first prove the following result.

Theorem 1. *Let $p,q \in \mathbb{R}^3$. Then there exists a shortest continuous curve (i.e. a geodesic) $\gamma : [0,1] \to \mathbb{R}^3$, $\gamma(0) = p$, $\gamma(1) = q$ such that*

$$\mathscr{L}_K(\gamma) = \inf_{\tilde{\gamma} \in \tilde{X}(p,q)} \mathscr{L}_K(\tilde{\gamma}),$$

where

$$\tilde{X}(p,q) := \left\{ \gamma : [0,1] \to \mathbb{R}^3 : \quad \textit{continuous, } \gamma(0) = p,\, \gamma(1) = q \right\}.$$

Observe the difference to $X(p,q)$ above is that curves do not need to be differentiable.

The above theorem follows from a general principle using the Arzelá-Ascoli theorem and holds true in much more generality.

More specifically to the Heisenberg group we will show that although the metric d_K differs from d_{cc}, the Korányi-length $\mathscr{L}_K$ equals the Carnot-Carathéodory length $\mathscr{L}_{cc}$.

Theorem 2. *Let $\gamma\colon [0,1] \to \mathbb{R}^3$ be twice continuously differentiable. If γ is horizontal (i.e.* (2) *holds) and $\mathscr{L}_{cc}(\gamma) < \infty$ then $\mathscr{L}_K(\gamma) < \infty$ and we have*

$$\mathscr{L}_K(\gamma) = \mathscr{L}_{cc}(\gamma).$$

From Theorem 2 we actually can conclude that $(\mathbb{R}^3, d_K)$ is not a length space: By the definition of length of a curve in a metric space (X,d), see Definition 5, for any p,q and any curve $\gamma\colon [0,1] \to X$, $\gamma(0) = p$, $\gamma(1) = q$ we have the inequality

$$\mathscr{L}(\gamma) \geq d(p,q).$$

If for any $p,q \in X$ there exists a curve $\gamma\colon [0,1] \to X$, $\gamma(0) = p$, $\gamma(1) = q$ such that we have equality

$$\mathscr{L}(\gamma) = d(p,q),$$

then we call X a length space. The following example shows that $(\mathbb{R}^3, d_K)$ is not a length space (this is in contrast to the Carnot-Carathéodory metric where the corresponding equality holds by definition (3)).

Example 3. *The following is the shortest curve between $p := (0,0,0)$ and $q := (0,0,\frac{1}{4\pi})$*

$$\gamma(t) = \begin{pmatrix} (1-\cos(2\pi t)) \\ \sin(2\pi t) \\ \frac{1}{4\pi}(t - \frac{\sin(2\pi t)}{2\pi}) \end{pmatrix}.$$

See [3, Theorem 2.1]. It can be checked by a direct computation that $\mathscr{L}_K(\gamma) = \mathscr{L}_{cc}(\gamma) > d_K(p,q)$

The outline of the remaining part of the paper is as follows: in Section 2 we discuss preliminary results on metric spaces, in particular Arzelá-Ascoli's theorem. In Section 3 we discuss properties of horizontal curves that we need for both theorems. In Section 4 we establish the existence of shortest curves with respect to $\mathscr{L}_K$ in the Heisenberg group. In Section 5 we prove Theorem 2. Let us remark that the results in this work are probably well-known to experts, the purpose of this paper is to provide a detailed account making this exciting field accessible to non-experts, students and early career researchers.

2 Some Preliminary Statements from Analysis: Metric Spaces

Let X be a metric space with metric d. A *curve* γ is simply a continuous map $\gamma\colon I \to X$, where $I = [a,b]$ is any closed finite interval.

We say that a curve $\gamma\colon [a,b] \to X$ connects two points $p,q \in X$ if $\gamma(a) = p$ and $\gamma(b) = q$.

We now want to define the length of a curve $\gamma\colon [a,b] \to X$, however observe that γ may not be differentiable. Indeed, we may not even know what differentiability of γ means since X is not a linear space! So a formula such as (1) does not make sense. But recall from Calculus how we obtained the formula (1), we used polygonal approximation of a curve. We will do the same in metric spaces.

Definition 4 (Partition)**.** Given an interval $[a,b]$, a partition of size n is the set $\{x_0,x_1,\ldots,x_n\}$ where

$$a = x_0 < x_1 < \cdots < x_n = b$$

With the notion of partition we can "approximate" curves by a discrete path through the points $\gamma(a),\gamma(x_1),\ldots,\gamma(b)$. Then we use the metric to define the length of these "polygon"-lines.

Definition 5 (Length of curve)**.** Given a metric space (X,d) and a curve $\gamma\colon [a,b]\to X$. The length of γ is given by

$$\mathscr{L}(\gamma) = \sup_{p\in P}\sum_{i=1}^{n} d(\gamma(t_i),\gamma(t_{i-1})),$$

where the supremum is taken over all partitions p of $[a,b]$ (i.e. P is the collection of all partitions of $[a,b]$).

Observe that the length of a curve $\mathscr{L}(\gamma)$ is always nonnegative, indeed since $\{a,b\}$ is a partition of $[a,b]$, we have

$$\mathscr{L}(\gamma) \geq d(\gamma(a),\gamma(b)). \tag{4}$$

In general, even if $d(\gamma(a),\gamma(b)) < \infty$ the length $\mathscr{L}(\gamma)$ could be $+\infty$. We call any curve γ with finite length $\mathscr{L}(\gamma) < \infty$ *rectifiable*.

It is worth noting the following

Lemma 6. *Given a metric space (X,d), let $\gamma\colon [a,b]\to X$ be a curve of finite length, $\mathscr{L}(\gamma) < \infty$. Then for any $s_0 \in [a,b]$, the restricted curves*

$$\gamma\Big|_{[s_0,b]} : [s_0,b]\to X, \quad [s_0,b]\ni t\mapsto \gamma(t)$$

and

$$\gamma\Big|_{[a,s_0]} : [a,s_0]\to X, \quad [a,s_0]\ni t\mapsto \gamma(t)$$

are curves of finite length. Moreover

$$[a,b]\ni s\mapsto \mathscr{L}\left(\gamma\Big|_{[a,s]}\right)$$

and

$$[a,b]\ni s\mapsto \mathscr{L}\left(\gamma\Big|_{[s,b]}\right)$$

are continuous monotone increasing maps.

Proof. Finiteness and monotonicity are easy to obtain from the definition of the curve. For the continuity, we observe that for $a\leq s_1 < s_2 \leq b$

$$\mathscr{L}\left(\gamma\Big|_{[a,s_2]}\right) - \mathscr{L}\left(\gamma\Big|_{[a,s_1]}\right) = \mathscr{L}\left(\gamma\Big|_{[s_1,s_2]}\right)$$

and

$$\mathscr{L}\left(\gamma\Big|_{[s_1,b]}\right) - \mathscr{L}\left(\gamma\Big|_{[s_2,b]}\right) = \mathscr{L}\left(\gamma\Big|_{[s_1,s_2]}\right).$$

So what we need to show is that for any $\varepsilon > 0$ and any $s_1 \in [a,b]$ there exists $\delta > 0$ such that

$$\mathscr{L}\left(\gamma\Big|_{[s_1,s_2]}\right) < \varepsilon \quad \forall s_2 : |s_1 - s_2| < \delta.$$

Fix $\varepsilon > 0$ and $s_1 \in [a,b]$. By continuity of γ we find $\delta_1 > 0$ such that

$$d(\gamma(\tilde{s}), \gamma(\tilde{t})) < \varepsilon \quad \forall |\tilde{s} - s_1|,\ |\tilde{t} - s_1| < \delta_1. \tag{5}$$

Since $\mathscr{L}(\gamma) < \infty$ there exists a partition

$$a = t_0 < t_1 \ldots < t_n = b$$

such that

$$\mathscr{L}(\gamma) - \varepsilon \leq \sum_{i=1}^{n} d(\gamma(t_i), \gamma(t_{i-1})).$$

Set

$$\delta_2 := \inf_{i=1,\ldots,n} |t_i - t_{i-1}|.$$

Set $\delta := \min\{\delta_1, \delta_2\}$ and fix any $s_2 \in [a,b]$ with $|s_1 - s_2| < \frac{\delta}{2}$.

W.l.o.g. $s_1 < s_2$. We then may assume that $t_{i_0-1} < s_1 < t_{i_0} < s_2 < t_{i_0+1}$ for some $i_0 \in \mathbb{N}$ (all other cases follow by an easy adaptation). We now consider the new partition $\tilde{t}_i$,

$$\tilde{t}_i = \begin{cases} t_i & i \leq i_0 - 1 \\ s_1 & i = i_0 \\ t_{i_0} & i = i_0 + 1 \\ s_2 & i = i_0 + 2 \\ t_{i-2} & i \geq i_0 + 3. \end{cases}$$

Then, by triangular inequality,

$$\mathscr{L}(\gamma) - \varepsilon \leq \sum_{i=1}^{n+2} d(\gamma(\tilde{t}_i), \gamma(\tilde{t}_{i-1})). \tag{6}$$

Now let $s_1 = r_0 < r_1 < \ldots = r_m = s_2$ be any partition of $[s_1, s_2]$. Then

$$\begin{aligned} \sum_{j=0}^{m} d(\gamma(r_j), \gamma(r_{j-1})) = &\sum_{i \neq i_0+1, i_0+2} d(\gamma(\tilde{t}_i), \gamma(\tilde{t}_{i-1})) + \sum_{j=0}^{m} d(\gamma(r_j), \gamma(r_{j-1})) \\ &- \sum_{i=1}^{n+2} d(\gamma(\tilde{t}_i), \gamma(\tilde{t}_{i-1})) \\ &+ d(\gamma(\tilde{t}_{i_0+1}), \gamma(\tilde{t}_{i_0})) + d(\gamma(\tilde{t}_{i_0+2}), \gamma(\tilde{t}_{i_0+1})) \end{aligned}$$

Since we can combine the partitions $\tilde{t}_i$, $i \neq i_0 + 1, i_0 + 2$ with r_j to obtain a partition of $[a,b]$, we have by the definition of length,

$$\sum_{i \neq i_0+1, i_0+2} d(\gamma(\tilde{t}_i), \gamma(\tilde{t}_{i-1})) + \sum_{j=0}^{m} d(\gamma(r_j), \gamma(r_{j-1})) \leq \mathscr{L}(\gamma).$$

By (6) we have

$$-\sum_{i=1}^{n+2} d(\gamma(\tilde{t}_i), \gamma(\tilde{t}_{i-1})) \leq -\mathscr{L}(\gamma) + \varepsilon.$$

By (5) which we can apply since $s_1 < t_{i_0} < s_2$ and thus $|s_1 - s_2|, |t_{i_0} - s_2| < \delta_1$,

$$d(\gamma(\tilde{t}_{i_0+1}), \gamma(\tilde{t}_{i_0})) + d(\gamma(\tilde{t}_{i_0+2}), \gamma(\tilde{t}_{i_0+1})) = d(\gamma(t_{i_0}), \gamma(s_1)) + d(\gamma(s_2), \gamma(t_{i_0})) \leq 2\varepsilon.$$

So we have shown

$$\sum_{j=0}^{m} d(\gamma(r_j), \gamma(r_{j-1})) \leq 3\varepsilon.$$

This holds for any partition (r_j) of $[s_1, s_2]$ and thus

$$\mathscr{L}\left(\gamma\Big|_{[s_1,s_2]}\right) < 3\varepsilon.$$

Since ε was arbitrary, we can conclude. □

For simplicity, we will often restrict our attention to curves defined on $I = [0,1]$, which we can do without loss of generality. Indeed any curve

$$\gamma : [a,b] \to X$$

can be reparametrized to a curve

$$\tilde{\gamma} : [0,1] \to X$$

by simply setting

$$\tilde{\gamma}(t) := \gamma(tb + (1-t)a).$$

Similarly any curve $\gamma : [0,1] \to X$ can be reparametrized to a curve $\tilde{\gamma} : [a,b] \to X$. The length of the curve γ and $\tilde{\gamma}$ above are the same, $\mathscr{L}(\gamma) = \mathscr{L}(\tilde{\gamma})$. Indeed, the length of curves is invariant under reparametrization.

[Reparametrization] Let $\gamma : [a,b] \to X$ be a curve. Let $\tau : [c,d] \to [a,b]$ be a continuous bijection with continuous inverse (i.e. a homeomorphism) such that $\tau(c) = a$ and $\tau(d) = b$. Then, τ is a reparametrization of γ.

Lemma 7. *Let $\gamma : [a,b] \to X$ be a curve and $\tau : [c,d] \to [a,b]$ be a reparametrization. Then if we set $\tilde{\gamma}(t) := \gamma(\tau(t))$ we get that $\tilde{\gamma} : [c,d] \to X$ is a curve and*

$$\mathscr{L}(\gamma) = \mathscr{L}(\tilde{\gamma})$$

We leave the proof as an exercise, but observe that τ maps any partition for $[c,d]$ into a partition of $[a,b]$, and τ^{-1} maps any partition of $[a,b]$ into a partition of $[c,d]$.

Now we want to find geodesics, i.e. shortest curves between two points p and q in X. A curve $\gamma : I \to X$ is called the *shortest curve* or (minimizing) geodesic from p to q if it connects p and q and for any other curve $\tilde{\gamma} : \tilde{I} \to X$ which connects p and q we have we have $\mathscr{L}(\gamma) \leq \mathscr{L}(\tilde{\gamma})$.

In general metric spaces X there is no reason that there exists such a shortest curve γ. As a side-note a shortest curve in general is not unique: think of the many shortest curves connecting the north pole and the south pole of a sphere. In order to conduct in the following chapters our analysis of the Heisenberg group, we conclude this section with a few important notions and facts on maps (possibly) on metric spaces.

The first result from Analysis is the Arzelá-Ascoli theorem – the proof can be found in essentially all Advanced Calculus books. Recall that a set $E \subset X$ is compact, if any sequence $(x_n)_{n\in\mathbb{N}} \subset E$ has a subsequence $(x_{n_i})_{i\in\mathbb{N}}$ and a point $x \in E$ such that $d(x_{n_i}, x) \xrightarrow{i\to\infty} 0$.

Theorem 8 (Arzelá-Ascoli)**.** *Let (X,d) be a metric space and $E \subset X$ be compact. Assume there is a sequence of maps $\gamma_k : [0,1] \to E$ which are* equicontinuous*, i.e. for any $\varepsilon > 0$ there exists $\delta > 0$ such that $\sup_{k\in\mathbb{N}} d(\gamma_k(t),\gamma_k(s)) < \varepsilon$ for all $s,t \in [0,1]$ with $|s-t| < \delta$.*

Then, there exists a subsequence $(\gamma_{k_i})_{i\in\mathbb{N}}$ and a continuous limit function $\gamma : [0,1] \to X$ such that γ_{k_i} uniformly converge to γ in the sense that

$$\sup_{t\in[0,1]} d(\gamma_{k_i}(t),\gamma(t)) \xrightarrow{i\to\infty} 0.$$

We will use later that uniform Lipschitz continuity implies equicontinuity. Namely if there exists $\Lambda > 0$ such that

$$\sup_{k\in\mathbb{N}} d(\gamma_k(s),\gamma_k(t)) \le \Lambda|s-t| \quad \text{for all } s,t \in [0,1]$$

then the equicontinuity condition in Theorem 8 is satisfied.

We now show that any curve with finite length can be parametrized so that it is Lipschitz continuous (so curves with uniformly bounded length are uniformly Lipschitz continuous, and thus equicontinuous).

Proposition 9 (Monotone Reparametrization)**.** *Let $\gamma : [a,b] \to X$ be a curve of finite length, $\mathscr{L}(\gamma) < \infty$.*

Then γ admits a Lipschitz reparameterization in the following sense.

There exists $\tilde{\gamma} : [0,1] \to X$ with the following properties

- *$\tilde{\gamma}(0) = \gamma(a)$ and $\tilde{\gamma}(1) = \gamma(b)$*
- *$\tilde{\gamma}([0,1]) = \gamma([a,b])$ (in the sense of sets in X)*
- *$\mathscr{L}(\gamma) = \mathscr{L}(\tilde{\gamma})$,*
- *$|\tilde{\gamma}(s) - \tilde{\gamma}(t)| \le \mathscr{L}(\gamma)|s-t| \quad \forall s,t \in [0,1]$.*

Proof. Without loss of generality, $[a,b] = [0,1]$. Let $\gamma : [0,1] \to (X,d)$ be a curve of finite length.

Define $\tau(t) := \mathscr{L}(\gamma|_{[0,t]}) : [0,1] \to [0,\mathscr{L}(\gamma)]$, which by Lemma 6 is continuous and monotone increasing.

We would like to set $\hat{\gamma} := \gamma \circ \tau^{-1} : [0,\mathscr{L}(\gamma)] \to X$. The issue is that τ may not be strictly monotone, so τ may not be invertible.

However $\hat{\gamma}$ is still well-defined. Observe that if for some $0 \le r \le \tilde{r} \le 1$ we have $\tau(r) = \tau(\tilde{r})$, then

$$0 = \mathscr{L}(\gamma|_{[0,\tilde{r}]}) - \mathscr{L}(\gamma|_{[0,r]}) = \mathscr{L}(\gamma|_{[r,\tilde{r}]}),$$

that is $\mathscr{L}(\gamma|_{[r,\tilde{r}]}) = 0$ and from the definition of the length $\mathscr{L}$ we conclude that $d(\gamma(s),\gamma(t)) = 0$ for all $s,t \in [r,\tilde{r}]$.

That is $\tau(r) = \tau(\tilde{r})$ implies that γ is constant on $[\tilde{r},r]$, in particular $\gamma(r) = \gamma(\tilde{r})$.

So we can still define $\hat{\gamma} := \gamma \circ \tau^{-1}$ in the following sense: for a given $t \in [0, \mathscr{L}(\gamma)]$ take any $r \in [0,1]$ such that $\tau(r) = t$. Such a r exists by the intermediate value theorem since τ is continuous, $\tau(0) = 0$ and $\tau(1) = \mathscr{L}(\gamma)$. Then we set

$$\hat{\gamma}(t) := \gamma(r).$$

If we were to pick any other $\tilde{r}$ with $\tau(\tilde{r}) = t$ then by the above observation we have $\gamma(r) = \gamma(\tilde{r})$ and $\hat{\gamma}(t)$ still has the same value.

We now claim that $\hat{\gamma}$ is continuous. Fix $t_0 \in [0, \mathscr{L}(\gamma)]$ and $\varepsilon > 0$. Take $R \subset [0,1]$ such that $\tau(r) = t_0$ for all $r \in R$. By the above observation, whenever $r, \tilde{r} \in R$ we have $[r, \tilde{r}] \subset R$. On the other hand if $(r_k)_{k \in \mathbb{N}} \subset [0,1]$ such that $\tau(r_k) = t_0$ for all $k \in \mathbb{N}$ then if $r = \lim_{k \to \infty} r_k$ we have $\tau(r) = t_0$, by continuity of τ. Combining this with monotonicity of τ we find that for some $r_0 \leq r_1$

$$R = [r_0, r_1], \quad \text{and} \quad \tau(r) < t_0 \quad \text{if } r < r_0, \quad \text{and} \quad \tau(r) > t_0 \quad \text{if } r > r_1.$$

By continuity of γ, there exists an $\delta_1 > 0$ such that $|\gamma(r) - \hat{\gamma}(t_0)| < \varepsilon$ whenever $r \in (r_0 - \delta_1, r_1 + \delta_1)$. Let now $\delta_2 := \min\{\tau(r_0) - \tau(r_0 - \delta_1), \tau(r_1 + \delta_1) - \tau(r_0)) > 0$. Recall that $t_0 = \tau(r_0) = \tau(r_1)$. So whenever t satisfies $|t - t_0| < \delta_2$ then we have $t \in (\tau(r_0 - \delta_1), \tau(r_1 + \delta_1))$, and thus by monotonicity, $t \in \tau(r_0 - \delta_1, r_1 + \delta_1)$ which implies that $|\hat{\gamma}(t) - \hat{\gamma}(t_0)| < \varepsilon$. That is, we have shown continuity of $\hat{\gamma}$.

With the same observation as above, it is now not too difficult to show that $\mathscr{L}(\gamma) = \mathscr{L}(\hat{\gamma})$ – since the only points where τ is not invertible are points where no length is added. Indeed, let $0 = r_0 < r_1 < \ldots < r_n = 1$ be a partition of $[0,1]$. Set $t_0 = 0$ and $t_n = \mathscr{L}(\gamma)$ and set $t_i := \tau(r_i)$ for $i = 1, \ldots, n-1$. Then $\hat{\gamma}(t_i) = \gamma(r_i)$. By monotonicity of τ we have $0 = t_1 \leq t_2 \leq \ldots \leq t_n = \mathscr{L}(\gamma)$. It might happen that we have equality $t_i = t_{i-1}$ but then $\tau(r_i) = \tau(r_{i-1})$ which by the argument above means $\hat{\gamma}(t_i) = \hat{\gamma}(t_{i-1})$ and thus $d(\hat{\gamma}(t_i), \hat{\gamma}(t_{i-1})) = 0$. Consequently we have

$$\sum_{i=1}^{n} d(\gamma(r_i), \gamma(r_{i-1})) = \sum_{i=1}^{n} d(\hat{\gamma}(t_i), \hat{\gamma}(t_{i-1})) \leq \mathscr{L}(\hat{\gamma}).$$

Taking the supremum of all partitions of $[0,1]$ we have

$$\mathscr{L}(\gamma) \leq \mathscr{L}(\hat{\gamma}). \tag{7}$$

For the other direction let $0 = t_0 < t_1 < \ldots < t_n = \mathscr{L}(\gamma)$ be any partition of $[0, \mathscr{L}(\gamma)]$. We now create a new partition $0 = r_0 < \ldots < r_i < \ldots < r_n = 1$ such that $\tau(r_i) = t_i$, and thus by the definition of $\hat{\gamma}$, $\gamma(r_i) = \hat{\gamma}(t_i)$. We set $r_0 := 0$ and $r_n := 1$. We define r_i to be any $r_i \in (0,1)$ such that $\tau(r_i) = t_i$, this choice of r_i may not be unique but from the intermediate value theorem at least one such r_i must exists. Since $t_{i-1} < t_i$ for all i, from the monotonicity of τ we conclude that $r_{i-1} < r_i$ for all i, and thus $0 = r_0 < r_1 < \ldots < r_n = 1$ is the desired new partition of $[0,1]$. We then have

$$\sum_{i=1}^{n} d(\hat{\gamma}(t_i), \hat{\gamma}(t_{i-1})) = \sum_{i=1}^{n} d(\gamma(r_i), \gamma(r_{i-1})) \leq \mathscr{L}(\gamma).$$

Taking the supremum over all partitions of $[0, \mathscr{L}(\gamma)]$ we conclude

$$\mathscr{L}(\hat{\gamma}) \leq \mathscr{L}(\gamma). \tag{8}$$

Together, (7) and (8) imply

$$\mathscr{L}(\hat{\gamma}) = \mathscr{L}(\gamma).$$

Next, we observe that the definition of the length of a curve implies

$$d(\gamma(t),\gamma(s)) \overset{(4)}{\leq} \mathscr{L}\left(\gamma\,|_{[s,t]}\right) = \left|\mathscr{L}\left(\gamma|_{[0,t]}\right) - \mathscr{L}\left(\gamma|_{[0,s]}\right)\right| = |\tau(t)-\tau(s)|.$$

Let $\hat{s},\hat{t} \in [0,\mathscr{L}(\gamma)]$ and take any $s,t \in [0,1]$ such that $\tau(s)=\hat{s}$, $\tau(t)=\hat{t}$. Then

$$d(\hat{\gamma}(\hat{t}),\hat{\gamma}(\hat{s}))d(\gamma(t),\gamma(s)) \leq |\tau(t)-\tau(s)| = |\hat{t}-\hat{s}|.$$

Thus, $\hat{\gamma}$ is Lipschitz continuous, albeit with the wrong constant, which is easy to fix.

Set

$$\tilde{\gamma}(s) := \hat{\gamma}(\mathscr{L}(\gamma)s), \quad s \in [0,1].$$

Then we have

$$d(\tilde{\gamma}(s),\tilde{\gamma}(t)) \leq \mathscr{L}(\gamma)|s-t| \quad \forall s,t \in [0,1].$$

□

The Arzelá-Ascoli theorem, Theorem 8, will play a crucial role in constructing a *candidate* for a shortest curve in the Heisenberg group. Another important ingredient is the following lower semicontinuity of the length.

Proposition 10 (Lower semicontinuity of the length functional)**.** *Let (X,d) be a metric space, and $\{\gamma_n\}_{n\in\mathbb{N}}$ be a sequence of curves into X. If γ_n converges pointwise to a curve, γ, in X, then*

$$\mathscr{L}(\gamma) \leq \liminf_{n\to\infty} \mathscr{L}(\gamma_n)$$

Proof. As discussed above, without loss of generality we can assume that all curves $\gamma_n : [0,1] \to X$.

Let $\varepsilon > 0$ be arbitrary. Since

$$\mathscr{L}(\gamma) = \sup_{p\in P} \sum_{i\geq 1} (d(\gamma(t_i),\gamma(t_{i-1}))$$

where P is the set of partitions of $[0,1]$, we can find a specific partition, $\mu = (t_0,t_1,\ldots,t_m)$, such that

$$\mathscr{L}(\gamma) < \left(\sum_{t_i\in\mu, i\geq 1} d(\gamma(t_i),\gamma(t_{i-1}))\right) + \frac{\varepsilon}{2}.$$

By pointwise convergence $\gamma_n(t) \xrightarrow{n\to\infty} \gamma(t)$ for each fixed t, we can find $N \in \mathbb{N}$ such that

$$d(\gamma_n(t_i),\gamma(t_i)) < \frac{\varepsilon}{4m} \quad \forall i = 0,\ldots,m, \quad \forall n \geq N.$$

Then,

$$\begin{aligned} d(\gamma(t_i),\gamma(t_{i-1}) &\leq d(\gamma(t_i),\gamma_n(t_i)) + d(\gamma_n(t_i),\gamma_n(t_{i-1})) + d(\gamma_n(t_{i-1}),\gamma(t_{i-1})) \\ &< \frac{\varepsilon}{4m} + d(\gamma_n(t_i),\gamma_n(t_{i-1})) + \frac{\varepsilon}{4m} \\ &= d(\gamma_n(t_i),\gamma_n(t_{i-1})) + \frac{\varepsilon}{2m}. \end{aligned}$$

Thus,

$$\mathscr{L}(\gamma) < \left(\sum_{t_i \in \mu, i \geq 1} d(\gamma_n(t_i), \gamma_n(t_{i-1})) \right) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2}.$$

Finally, since

$$\mathscr{L}(\gamma_n) = \sup_{\mu \in \mathscr{P}} \sum_{t_i \in \mu, i \geq 1} (d(\gamma_n(t_i), \gamma_n(t_{i-1}))$$

we have

$$\sum_{t_i \in \mu, i \geq 1} d(\gamma_n(t_i), \gamma_n(t_{i-1})) \leq \mathscr{L}(\gamma_n).$$

Thus we have shown,

$$\mathscr{L}(\gamma) < \mathscr{L}(\gamma_n) + \varepsilon, \quad \forall n \geq N.$$

In particular

$$\mathscr{L}(\gamma) < \liminf_{n \to \infty} \mathscr{L}(\gamma_n) + \varepsilon.$$

This holds for any $\varepsilon > 0$, letting $\varepsilon \to 0$ we conclude

$$\mathscr{L}(\gamma) \leq \liminf_{n \to \infty} \mathscr{L}(\gamma_n).$$

□

From Arzelá-Ascoli theorem, Theorem 8, and the observations above we obtain the existence of shortest curves in the following sense.

Theorem 11. *Let (X,d) be any* complete *metric space and $E \subset X$ be a compact set. Let $p \neq q \in E$ such that there exists a continuous curve $\gamma_0 : [0,1] \to E$ of finite length $\mathscr{L}(\gamma_0) < \infty$ and $\gamma_0(0) = p$ and $\gamma_0(1) = q$. Then there exists a geodesic between p and q, i.e. a curve $\gamma : [0,1] \to E$ such that $\gamma(0) = p$ and $\gamma(1) = q$ and such that*

$$\mathscr{L}(\gamma) = \inf_{\tilde{\gamma}} \mathscr{L}(\tilde{\gamma})$$

where the infimum is taken over all continuous curves $\tilde{\gamma} : [0,1] \to E$ with $\tilde{\gamma}(0) = p$ and $\tilde{\gamma}(1) = q$.

It is important to note that above the notion of "shortest curve" is with respect to E not with respect to X, and this might lead to a different notion of what is a shortest curve. Take for example a compact banana-shaped set E in $\mathbb{R}^3$. The straight line from top to bottom of the banana E is likely to not lie within E, so it is not the shortest curve in E!

Proof of Theorem 11. For simplicity we assume $X = E$. Since there exists one curve connecting p and q with finite length we have

$$I := \inf_{\tilde{\gamma}} \mathscr{L}(\tilde{\gamma}) \in [0, \infty).$$

Since there exists one curve connecting p and q there also must be a "minimizing sequence"

$$\gamma_k : [0,1] \to X \quad \text{of finite length, } \mathscr{L}(\gamma_0) < \infty, \text{ and } \gamma_k(0) = p \text{ and } \gamma_k(1) = q$$

such that

$$\mathscr{L}(\gamma_k) \xrightarrow{k \to \infty} I.$$

We may even assume that

$$I \leq \mathscr{L}(\gamma_k) \leq I + \frac{1}{k} \quad \forall k.$$

By Proposition 9 we may assume without loss of generality (otherwise use $\tilde{\gamma}_k$ instead of γ_k)

$$|\gamma_k(x) - \gamma_k(y)| \leq \left(I + \frac{1}{k}\right)|x-y| \quad \forall x, y \in [0,1], \quad k \in \mathbb{N}.$$

By Arzelá-Ascoli, Theorem 8, we may assume that we have uniform convergence to some continuous $\gamma : [0,1] \to X$, otherwise we could pass yet again to a subsequence.

Then, by lower semicontinuity of the length, Proposition 10, we have

$$\mathscr{L}(\gamma) \leq \liminf_{k\to\infty} \mathscr{L}(\gamma_k)$$

This means

$$I \leq \mathscr{L}(\gamma) \leq \liminf_{k\to\infty} \mathscr{L}(\gamma_k) = I.$$

So γ is a shortest curve. □

3 Horizontal Curves in the Heisenberg Group

Recall that a differentiable curve $\gamma : [0,1] \to \mathbb{R}^3$ is called horizontal if (2) holds. In this section we compute important properties of horizontal curves that we will use in the proofs of both our main theorems.

Proposition 12. *If $\gamma \in C^2([0,1])$ and (2) holds. Then*

$$\lim_{s\to t} \frac{\frac{\gamma^3(t)-\gamma^3(s)}{t-s} + 2\left(\frac{\left(\gamma^2(t)-\gamma^2(s)\right)}{t-s}\gamma^1(s) - \frac{\left(\gamma^1(t)-\gamma^1(s)\right)}{t-s}\gamma^2(s)\right)}{t-s} = 0.$$

The convergence rate is uniform in t.

Proof. Since γ is C^2, we have

$$\gamma(s) = \gamma(t) + (s-t)\dot{\gamma}(t) + \frac{1}{2}\ddot{\gamma}(t)(s-t)^2 + o(|t-s|^2).$$

and o is uniform in the domain of γ.

Then,

$$\begin{aligned}
&\frac{\frac{\gamma^3(t)-\gamma^3(s)}{t-s} + 2\left(\frac{\left(\gamma^2(t)-\gamma^2(s)\right)}{t-s}\gamma^1(s) - \frac{\left(\gamma^1(t)-\gamma^1(s)\right)}{t-s}\gamma^2(s)\right)}{t-s} \\
&= \frac{\dot{\gamma}^3(t) - \frac{1}{2}\ddot{\gamma}^3(t)(t-s) + 2\left(\left(\dot{\gamma}^2(t) - \frac{1}{2}\ddot{\gamma}^2(t)(t-s)\right)\gamma^1(s) - \left(\dot{\gamma}^1(t) - \frac{1}{2}\ddot{\gamma}^1(t)(t-s)\right)\gamma^2(s)\right)}{t-s} \\
&\quad + o(1)
\end{aligned}$$

$$
\begin{aligned}
=&\frac{-\frac{1}{2}\ddot{\gamma}^3(t)(t-s)+2\left(\left(-\frac{1}{2}\ddot{\gamma}^2(t)(t-s)\right)\gamma^1(s)-\left(-\frac{1}{2}\ddot{\gamma}^1(t)(t-s)\right)\gamma^2(s)\right)}{t-s}\\
&+\frac{\dot{\gamma}^3(t)+2\left(\left(\dot{\gamma}^2(t)\right)\gamma^1(s)-\left(\dot{\gamma}^1(t)\right)\gamma^2(s)\right)}{t-s}\\
&+o(1)\\
=&-\frac{1}{2}\ddot{\gamma}^3(t)+2\left(-\frac{1}{2}\ddot{\gamma}^2(t)\gamma^1(t)+\frac{1}{2}\ddot{\gamma}^1(t)\gamma^2(t)\right)+o(1)\\
&+\frac{\dot{\gamma}^3(t)+2\left(\left(\dot{\gamma}^2(t)\right)\gamma^1(s)-\left(\dot{\gamma}^1(t)\right)\gamma^2(s)\right)}{t-s}\\
&+o(1)
\end{aligned}
$$

We define

$$f(s):=\dot{\gamma}^3(t)+2\left(\left(\dot{\gamma}^2(t)\right)\gamma^1(s)-\left(\dot{\gamma}^1(t)\right)\gamma^2(s)\right)$$

and, we observe that by horizontality, $f(t)=0$. Thus

$$\frac{f(s)}{t-s}=-\frac{f(s)-f(t)}{s-t}=-f'(t)+o(1)$$

Then, we have

$$f'(s)=2\left(\left(\dot{\gamma}^2(t)\right)\dot{\gamma}^1(s)-\left(\dot{\gamma}^1(t)\right)\dot{\gamma}^2(s)\right)$$

So

$$f'(t)=2\left(\left(\dot{\gamma}^2(t)\right)\dot{\gamma}^1(t)-\left(\dot{\gamma}^1(t)\right)\dot{\gamma}^2(t)\right)$$

Consequently,

$$
\begin{aligned}
&\frac{\frac{\gamma^3(t)-\gamma^3(s)}{t-s}+2\left(\frac{\left(\gamma^2(t)-\gamma^2(s)\right)}{t-s}\gamma^1(s)-\frac{\left(\gamma^1(t)-\gamma^1(s)\right)}{t-s}\gamma^2(s)\right)}{t-s}\\
=&-\frac{1}{2}\ddot{\gamma}^3(t)+2\left(-\frac{1}{2}\ddot{\gamma}^2(t)\gamma^1(t)+\frac{1}{2}\ddot{\gamma}^1(t)\gamma^2(t)\right)+o(1)\\
&-\left(2\left(\left(\dot{\gamma}^2(t)\right)\dot{\gamma}^1(t)-\left(\dot{\gamma}^1(t)\right)\dot{\gamma}^2(t)\right)\right)+o(1)\\
&+o(1)\\
=&-\frac{1}{2}\frac{d}{dt}\dot{\gamma}^3(t)+2\left(\dot{\gamma}^2(t)\gamma^1(t)-\dot{\gamma}^1(t)\gamma^2(t)\right)\\
&-\frac{1}{2}-2\left(\dot{\gamma}^2(t)\dot{\gamma}^1(t)-\dot{\gamma}^1(t)\dot{\gamma}^2(t)\right)\\
&-\left(+2\left(\left(\dot{\gamma}^2(t)\right)\dot{\gamma}^1(t)-\left(\dot{\gamma}^1(t)\right)\dot{\gamma}^2(t)\right)\right)\\
&+o(1)\\
=&0\\
&-\left(+1\left(\left(\dot{\gamma}^2(t)\right)\dot{\gamma}^1(t)-\left(\dot{\gamma}^1(t)\right)\dot{\gamma}^2(t)\right)\right)\\
&+o(1)\\
=&o(1).
\end{aligned}
$$

Then

$$\lim_{s\to t}\frac{\frac{\gamma^3(t)-\gamma^3(s)}{t-s}+2\left(\frac{\left(\gamma^2(t)-\gamma^2(s)\right)}{t-s}\gamma^1(s)-\frac{\left(\gamma^1(t)-\gamma^1(s)\right)}{t-s}\gamma^2(s)\right)}{t-s}=0.$$

as desired.

Then,

$$\begin{aligned}
&\lim_{s\to t}\frac{\left[|\gamma^1(t)-\gamma^1(s)|^2+|\gamma^2(t)-\gamma^2(s)|^2+\left|\gamma^3(t)-\gamma^3(s)+2\left(\left(\gamma^2(t)-\gamma^2(s)\right)\gamma^1(s)-\left(\gamma^1(t)-\gamma^1(s)\right)\gamma^2(s)\right)\right|\right]^2}{|t-s|^4}\\
&=\lim_{s\to t}\frac{|\gamma^1(t)-\gamma^1(s)|^2+|\gamma^2(t)-\gamma^2(s)|^{2^2}}{|t-s|^4}\\
=&\lim_{s\to t}\frac{|\gamma^1(t)-\gamma^1(s)|^4+2|\gamma^1(t)-\gamma^1(s)|^2|\gamma^2(t)-\gamma^2(s)|^2+|\gamma^2(t)-\gamma^2(s)|^4}{|t-s|^4}\\
=&\lim_{s\to t}\left(\frac{|\gamma^1(t)-\gamma^1(s)|}{|t-s|}\right)^4+2\left(\frac{|\gamma^1(t)-\gamma^1(s)|}{|t-s|}\right)^2\left(\frac{|\gamma^2(t)-\gamma^2(s)|}{|t-s|}\right)^2+\left(\frac{|\gamma^2(t)-\gamma^2(s)|}{|t-s|}\right)^4\\
=&\dot\gamma^1(t)^4+2\dot\gamma^1(t)^2\dot\gamma^2(t)^2+\dot\gamma^4(t)=\left(\dot\gamma^1(t)^2+\dot\gamma^2(t)^2\right)^2
\end{aligned}$$

and the convergence is uniformly in t by the above considerations. □

From Proposition 12 we readily obtain

Corollary 13. *If $\gamma\in C^2([0,1],\mathbb{R}^3)$ and (2) holds*

$$\frac{d_K(\gamma(t),\gamma(s))}{|t-s|}\xrightarrow{s\to t}\sqrt{\dot\gamma^1(t)^2+\dot\gamma^2(t)^2}$$

The convergence is uniform in t. In particular we have

$$\mathscr{L}_K(\gamma)<\infty.$$

4 Existence of Shortest Curves in the Heisenberg Group

In this section we want to show Theorem 1.

Of course we would like to apply Theorem 11, however we need to be careful with the compactness assumption in that theorem, since $\mathbb{H}_1$ is not compact. However, one could justifiably believe that any curve $\gamma:[0,1]\to\mathbb{H}_1$ which goes too far away from p and q is not a good candidate for shortest curve. We need to quantify this and for this we compare the Korányi metric locally with the Euclidean metric.

Lemma 14. *Let $K\subset\mathbb{R}^3$ be compact (in the sense of the Euclidean metric). Then $K\subset\mathbb{H}_1$ is compact (in the sense of the Korányi metric).*

Proof. Since K is compact as Euclidean set $\mathbb{R}^3$ it is bounded and thus there must be some $\Lambda>0$ such that

$$\max\{|p_1|,|p_2|,|p_2|\}<\Lambda\quad\forall p=(p_1,p_2,p_3)\in K.$$

Using repeatedly Young's inequality $2ab \leq a^2 + b^2$ we find that for $p, q \in K$

$$\begin{aligned} d_K(q,p) =& (|p_1 - q_1|^2 + |p_2 - q_2|^{2^2} + |p_3 - q_3 + 2(p_2 q_1 - p_1 q_2)|^2)^{\frac{1}{4}} \\ \leq& (|p_1 - q_1|^2 + |p_2 - q_2|^{2^2} + 2\,|p_3 - q_3|^2 + 2\,|2(p_2 q_1 - p_1 q_2)|^2)^{\frac{1}{4}} \\ =& (|p_1 - q_1|^2 + |p_2 - q_2|^{2^2} + 2\,|p_3 - q_3|^2 + 2\,|2(p_2 - q_2)q_1 + (q_1 - p_1)q_2)|^2)^{\frac{1}{4}} \\ \leq& (|p_1 - q_1|^2 + |p_2 - q_2|^{2^2} + 2\,|p_3 - q_3|^2 + 8\,(|p_2 - q_2|\Lambda + |q_1 - p_1|\Lambda))^2)^{\frac{1}{4}} \end{aligned}$$

We conclude that for each $\varepsilon > 0$ there exists $\delta > 0$ such that if $p, q \in K$ and $|p - q| < \delta$ (in the Euclidean sense) then $d_K(p,q) < \varepsilon$.

In particular any (Euclidean) converging sequence in K also converges in the sense of the Korányi metric d_K. Thus K is also compact in the Korányi sense. □

The following lemma shows that "far away" in the Euclidean sense implies "far away" in the Korányi sense.

Lemma 15. *Fix $q \in \mathbb{R}^3$. For any $\Lambda > 0$ there exists $\Theta > 0$ such that the following is true: if for some $p \in \mathbb{R}^3$ we have*

$$|p - q| > \Theta$$

then

$$d_K(p,q) > \Lambda.$$

Proof. Observe that for any $p, q \in \mathbb{R}^3$

$$\begin{aligned} d_K(p,q)^2 \geq |p_3 - q_3 + 2(p_2 q_1 - p_1 q_2)| &= |p_3 - q_3 + 2((p_2 - q_2)q_1 + q_1 q_2 - (p_1 - q_1)q_2 \\ &\quad - q_1 q_2)| \\ &= |p_3 - q_3 + 2((p_2 - q_2)q_1 - (p_1 - q_1)q_2| \\ &\geq (|p_3 - q_3| - 2\,|q_1(p_2 - q_2) - q_2(p_1 - q_1)|) \\ &\geq (|p_3 - q_3| - 2\,(|q_1||p_2 - q_2| + |q_2||p_1 - q_1|)) \\ &\geq (|p_3 - q_3| - 2\,(|q_1||p_2 - q_2| + |q_2||p_1 - q_1|)) \end{aligned}$$

Now fix $q = (q_1, q_2, q_3) \in \mathbb{R}^3$ and $\Lambda > 0$ and set

$$\Gamma := |q_1| + |q_2|.$$

Take $\Theta > 0$ so that the following conditions are satisfied: $\Theta > \sqrt{3}\Lambda$ and $\frac{1}{\sqrt{3}}\Theta - 2\Gamma\Lambda > \Lambda^2$.

Now take $p = (p_1, p_2, p_3) \in \mathbb{R}^3$ such that

$$|p - q| > \Theta.$$

Then

$$\max\{|p_1 - q_1|, |p_2 - q_2|, |p_3 - q_3|\} > \frac{1}{\sqrt{3}}\Theta.$$

Then either

$$\max\{|p_1 - q_1|, |p_2 - q_2|\} > \Lambda$$

or

$$|p_3 - q_3| > \frac{1}{\sqrt{3}}\Theta.$$

From the above estimates we have

$$d_K(p,q) \geq \max\left\{|p_1-q_1|, |p_2-q_2|, (|p_3-q_3| - 2\left(|q_1||p_2-q_2| + |q_2||p_1-q_1|\right))^{\frac{1}{2}}\right\}$$

In the case that $\max\{|p_1-q_1|, |p_2-q_2|\} > \Lambda$ we conclude that

$$d_K(p,q) > \Lambda,$$

and we are done. If on the other hand both $|p_1-q_1|$ or $|p_2-q_2| < \Lambda$ then we have $|p_3-q_3| > \frac{1}{\sqrt{3}}\Theta$ and thus

$$\begin{aligned} d_K(p,q)^2 \geq & |p_3-q_3| - 2\left(|q_1||p_2-q_2| + |q_2||p_1-q_1|\right) \\ \geq & \frac{1}{\sqrt{3}}\Theta - 2\Gamma\Lambda \end{aligned}$$

Again in this case, by the choice of Θ we find that

$$d_K(p,q)^2 > \Lambda^2,$$

and we can conclude $d_K(p,q) > \Lambda$ as desired. □

Proof of Theorem 1. Fix $p, q \in \mathbb{R}^3$. There exists a smooth horizontal curve $\tilde{\gamma}$ connecting p and q, take for example the $\mathscr{L}_{cc}$-geodesic from [3], and in view of Corollary 13 $\tilde{\gamma}$ has finite length: $\mathscr{L}_K(\tilde{\gamma}) < \infty$.

Let $R > 0$ such that for any $r \in \mathbb{R}^3$ with $|p-r| > R$ we have in view of Lemma 15

$$d_K(p,r) > \mathscr{L}_K(\tilde{\gamma}).$$

This implies that any continuous curve $\gamma\colon [0,1] \to \mathbb{R}^3$ with $\gamma(0) = p$ and $\gamma(1) = q$ and $|\gamma(t) - p| > R$ for any $t \in (0,1)$ we have

$$\mathscr{L}_K(\gamma) > \mathscr{L}_K(\tilde{\gamma}).$$

Set $E := \{r \in \mathbb{R}^3 : |r-p| \leq R\}$ which is a compact set in the Euclidean sense, and thus in view of Lemma 14 also in the Korányi sense. Then we have shown that

$$\inf_{\gamma:[0,1]\to E} \mathscr{L}_K(\gamma) = \inf_{\gamma:[0,1]\to\mathbb{R}^3} \mathscr{L}_K(\gamma),$$

where both infima are taken over continuous curves γ with $\gamma(0) = p$ and $\gamma(1) = q$. Now we can finally apply Theorem 11. Thus, there is a shortest curve between p and q. □

5 Length of Curves in the Heisenberg Group – Proof of Theorem 2

In this section we show that

$$\mathscr{L}_{cc}(\gamma) = \mathscr{L}_K(\gamma),$$

whenever $\gamma \in C^2$ is a horizontal curve, i.e. whenever γ satisfies (2).

Proof of Theorem 2. From (2) in particular,

$$\frac{d}{dt}\left(\dot{\gamma}^3(t) + 2\left(\dot{\gamma}^2(t)\gamma^1(t) - \dot{\gamma}^1(t)\gamma^2(t)\right)\right) = 0$$

We apply Proposition 12 and obtain

$$\lim_{s\to t}\left(\frac{d_K(\gamma(t),\gamma(s))}{|t-s|}\right)^4 = \lim_{s\to t}\frac{|\gamma^1(t)-\gamma^1(s)|^2+|\gamma^2(t)-\gamma^2(s)|^{2^2}}{|t-s|^4}$$

$$=\lim_{s\to t}\frac{|\gamma^1(t)-\gamma^1(s)|^4+2|\gamma^1(t)-\gamma^1(s)|^2|\gamma^2(t)-\gamma^2(s)|^2+|\gamma^2(t)-\gamma^2(s)|^4}{|t-s|^4}$$

$$=\lim_{s\to t}\left(\frac{|\gamma^1(t)-\gamma^1(s)|}{|t-s|}\right)^4+2\left(\frac{|\gamma^1(t)-\gamma^1(s)|}{|t-s|}\right)^2\left(\frac{|\gamma^2(t)-\gamma^2(s)|}{|t-s|}\right)^2+\left(\frac{|\gamma^2(t)-\gamma^2(s)|}{|t-s|}\right)^4$$

$$=\dot\gamma^1(t)^4+2\dot\gamma^1(t)^2\dot\gamma^2(t)^2+\dot\gamma^4(t)=\left(\dot\gamma^1(t)^2+\dot\gamma^2(t)^2\right)^2$$

Then, taking the fourth root,

$$\lim_{s\to t}\frac{d_K(\gamma(t),\gamma(s))}{|t-s|}=\sqrt{\dot\gamma^1(t)^2+\dot\gamma^2(t)^2}$$

□

Proof of $\mathscr{L}_{cc}(\gamma)=\mathscr{L}_K(\gamma)$ if γ is horizontal. From Corollary 19, we see that

$$\lim_{s\to t}\frac{d_K(\gamma(t),\gamma(s))}{|t-s|}=\sqrt{\dot\gamma^1(t)^2+\dot\gamma^2(t)^2}$$

uniformly in t. Then, from the above limit, given some $\varepsilon>0$ choose $\delta>0$, such that when $|t_i-t_{i-1}|<\delta$, we have

$$\left|\frac{d(\gamma(t_i),\gamma(t_{i-1}))}{|t_i-t_{i-1}|}-\sqrt{|\dot\gamma^1(t_i)|^2+|\dot\gamma^2(t_i)|^2}\right|<\varepsilon.$$

Multiplying by $|t_i-t_{i-1}|$,

$$\left|\frac{d(\gamma(t_i),\gamma(t_{i-1}))}{|t_i-t_{i-1}|}|t_i-t_{i-1}|-\sqrt{|\dot\gamma^1(t_i)|^2+|\dot\gamma^2(t_i)|^2}|t_i-t_{i-1}|\right|<\varepsilon|t_i-t_{i-1}|.$$

Now, let $\mathscr{P}$ be the set of partitions of $[0,1]$ such that for any $\mu\in\mathscr{P}$, we have $|t_i-t_{i-1}|<\delta$ for each t_i in μ. Then, for a given $\mu\in\mathscr{P}$,

$$\sum_{t_i\in\mu,i\geq 1}\left|\frac{d(\gamma(t_i),\gamma(t_{i-1}))}{|t_i-t_{i-1}|}|t_i-t_{i-1}|-\sqrt{|\dot\gamma^1(t_i)|^2+|\dot\gamma^2(t_i)|^2}|t_i-t_{i-1}|\right|<\varepsilon\underbrace{\sum_{t_i\in\mu,i\geq 1}|t_i-t_{i-1}|}_{=1}=\varepsilon$$

So, we get

$$\left|\sum_{t_i\in\mu,i\geq 1}\frac{d(\gamma(t_i),\gamma(t_{i-1}))}{|t_i-t_{i-1}|}|t_i-t_{i-1}|-\sum_{t_i\in\mu,i\geq 1}\sqrt{|\dot\gamma^1(t_i)|^2|\dot\gamma^2(t_i)|^2}|t_i-t_{i-1}|\right|$$

$$=\left|\sum_{t_i\in\mu,i\geq 1}d(\gamma(t_i),\gamma(t_{i-1}))-\sum_{t_i\in\mu,i\geq 1}\sqrt{|\dot\gamma^1(t_i)|^2+|\dot\gamma^2(t_i)|^2}|t_i-t_{i-1}|\right|<\varepsilon$$

Then,

$$\begin{aligned}\varepsilon &> \sup_{\mu\in\mathscr{P}}\left|\sum_{t_i\in\mu,i\geq 1} d(\gamma(t_i),\gamma(t_{i-1})) - \sum_{t_i\in\mu,i\geq 1}\sqrt{|\dot\gamma^1(t_i)|^2+|\dot\gamma^2(t_i)|^2}|t_i-t_{i-1}|\right| \\ &\geq \sup_{\mu\in\mathscr{P}}\left|\sum_{t_i\in\mu,i\geq 1} d(\gamma(t_i),\gamma(t_{i-1}))\right| - \sup_{\mu\in\mathscr{P}}\left|\sum_{t_i\in\mu,i\geq 1}\sqrt{|\dot\gamma^1(t_i)|^2+|\dot\gamma^2(t_i)|^2}|t_i-t_{i-1}|\right| \\ &= \mathscr{L}_K(\gamma) - \sup_{\mu\in\mathscr{P}}\left|\sum_{t_i\in\mu,i\geq 1}\sqrt{|\dot\gamma^1(t_i)|^2+|\dot\gamma^2(t_i)|^2}|t_i-t_{i-1}|\right|\end{aligned}$$

Similarly,

$$\begin{aligned}\varepsilon &> \sup_{\mu\in\mathscr{P}}\left|\sum_{t_i\in\mu,i\geq 1} d(\gamma(t_i),\gamma(t_{i-1})) - \sum_{t_i\in\mu,i\geq 1}\sqrt{|\dot\gamma^1(t_i)|^2+|\dot\gamma^2(t_i)|^2}|t_i-t_{i-1}|\right| \\ &\geq \sup_{\mu\in\mathscr{P}}\left|\sum_{t_i\in\mu,i\geq 1}\sqrt{|\dot\gamma^1(t_i)|^2+|\dot\gamma^2(t_i)|^2}|t_i-t_{i-1}|\right| - \sup_{\mu\in\mathscr{P}}\left|\sum_{t_i\in\mu,i\geq 1} d(\gamma(t_i),\gamma(t_{i-1}))\right| \\ &= \sup_{\mu\in\mathscr{P}}\left|\sum_{t_i\in\mu,i\geq 1}\sqrt{|\dot\gamma^1(t_i)|^2+|\dot\gamma^2(t_i)|^2}|t_i-t_{i-1}|\right| - \mathscr{L}_K(\gamma)\end{aligned}$$

That is,

$$\varepsilon > \left|\sup_{\mu\in\mathscr{P}}\left|\sum_{t_i\in\mu,i\geq 1}\sqrt{|\dot\gamma^1(t_i)|^2+|\dot\gamma^2(t_i)|^2}|t_i-t_{i-1}|\right| - \mathscr{L}_K(\gamma)\right|$$

So, we have

$$\begin{aligned}&\left|\mathscr{L}_K(\gamma) - \sup_{\mu\in\mathscr{P}}\left|\sum_{t_i\in\mu,i\geq 1}\sqrt{|\dot\gamma^1(t_i)|^2+|\dot\gamma^2(t_i)|^2}|t_i-t_{i-1}|\right|\right| \\ \leq &\left|\mathscr{L}_K(\gamma) - \sup_{\mu\in\mathscr{P}}\sum_{t_i\in\mu,i\geq 1}\sqrt{|\dot\gamma^1(t_i)|^2+|\dot\gamma^2(t_i)|^2}|t_i-t_{i-1}|\right| < \varepsilon\end{aligned}$$

Note that, since $\dot\gamma$ is continuous, the function

$$\sqrt{|\dot\gamma^1(t_i)|^2+|\dot\gamma^2(t_i)|^2}$$

is continuous and hence integrable. So,

$$\sum_{t_i\in\mu,i\geq 1}\sqrt{|\dot\gamma^1(t_i)|^2+|\dot\gamma^2(t_i)|^2}|t_i-t_{i-1}|$$

is a Riemann Sum, and

$$\sup_{\mu\in\mathscr{P}}\sum_{t_i\in\mu,i\geq 1}\sqrt{|\dot\gamma^1(t_i)|^2+|\dot\gamma^2(t_i)|^2}|t_i-t_{i-1}| = \int_0^1\sqrt{|\dot\gamma^1(t)|^2+|\dot\gamma^2(t)|^2}dt$$

Then,

$$\left| \mathscr{L}_K(\gamma) - \int_0^1 \sqrt{|\dot{\gamma}^1(t)|^2 + |\dot{\gamma}^2(t)|^2} dt \right| < \varepsilon$$

This holds for any $\varepsilon > 0$, so letting $\varepsilon \to 0$ we conclude

$$\mathscr{L}_K(\gamma) = \int_0^1 \sqrt{|\dot{\gamma}^1(t)|^2 + |\dot{\gamma}^2(t)|^2} dt.$$

This proves $\mathscr{L}_{cc}(\gamma) = \mathscr{L}_K(\gamma)$ which in particular implies Theorem 2. □

Acknowledgement

The authors would like to thank the anonymous referee for valuable suggestions on the article. This is part of a Undergraduate Research Project with Dr. Schikorra. Funding was provided by NSF Career DMS-2044898.

Bibliography

[1] L. Capogna, D. Danielli, S. D. Pauls, and J. T. Tyson. *An introduction to the Heisenberg group and the sub-Riemannian isoperimetric problem*, volume 259 of Progress in Mathematics. Birkh¨auser Verlag, Basel, 2007.

[2] M. Gromov. Carnot-Carathéodory spaces seen from within. In *Sub-Riemannian geometry*, volume 144 of Progr. Math., pages 79–323. Birkh¨auser, Basel, 1996.

[3] P. Hajłlasz and S. Zimmerman. Geodesics in the Heisenberg group. *Anal. Geom. Metr. Spaces,* 3(1):325–337, 2015.

[4] R. Montgomery. *A tour of subriemannian geometries, their geodesics and applications, volume 91 of Mathematical Surveys and Monographs.* American Mathematical Society, Providence, RI, 2002.

Ball State Undergraduate Mathematics Exchange
`https://digitalresearch.bsu.edu/mathexchange`
Vol. 16, No. 1 (Fall 2022)
Pages 104 – 120

Numerical Range of Strictly Triangular Matrices over Finite Fields

*Ariel Russell**

Ariel Russell produced this work during her time as an undergraduate student at Taylor University. She graduated in 2021 with a B.S. in Mathematics and a B.A. in Computer Science. She now enjoys teaching high school math and programming in Colorado Springs, CO.

Abstract

In this paper we investigate the numerical range of 3×3 matrices over finite fields, particularly when the matrix is strictly triangular. We provide a conjecture for this case that extends to $n \times n$ matrices for $n \geq 3$ and also provide sample code for generating the numerical range.

1 Introduction

Numerical ranges of matrices over $\mathbb{C}$ have been studied extensively, most notably by Hausdorff, Toeplitz, and Kippenhahn. Investigation into numerical ranges over finite fields was initiated in [3] and has been continued in several papers of Ballico (see e.g. [1], [2]). Here we require our field to be of characteristic p where $p \equiv 3 \mod 4$, ensuring that the element -1 is not a quadratic residue in $\mathbb{Z}_p$, so that i has a proper analog to its use in $\mathbb{C}$.

Let p be a prime congruent to $3 \mod 4$, and define $\mathbb{Z}_p[i]$ as the Galois Field of order p^2 in the form $\{a+bi : a,b \in \mathbb{Z}\}$. $M_n(\mathbb{Z}_p[i])$ denotes the set of $n \times n$ matrices with entries in $\mathbb{Z}_p[i]$. The numerical range of matrix $M \in M_n(\mathbb{Z}_p[i])$ is defined as $W(M) = \{x^*Mx : x \in \mathbb{Z}_p[i]^n, \|x\|^2 = x^*x = 1\}$ with x^* representing the conjugate transpose of x. Thus, $W(M)$ forms a set in $\mathbb{Z}_p[i]$. The authors in [3] also introduce the concept of the k-th numerical range, $W_k(M) = \{x^*Mx : x \in \mathbb{Z}_p[i]^n, \|x\|^2 = x^*x = k \in \mathbb{Z}_p\}$. (Here, then, $W(M) = W_1(M)$.)

***Corresponding author**: ariel.j.russell@gmail.com

In [3], work is primarily focused on upper triangular 2×2 matrices. In no 2×2 matrix do we see a numerical range that includes every element of $\mathbb{Z}_p[i]$. It seems one more dimension is needed: in all of our testing, every strictly triangular matrix of dimension 3 or higher had $W(M) = \mathbb{Z}_p[i]$. The goal of this paper is to make as much progress towards that conjecture as possible.

2 Preliminaries

Our proofs in the following sections depend on some key tools. In particular, we frequently attempt to remove one of the entries of an input vector x from the expression x^*Mx, so that the missing entry can ensure that $\|x\|^2 = 1$. The validity of this technique comes from the following two lemmas; the first justifies the second.

Lemma 1. [3, Lemma 2.1] *For all primes p congruent to $3 \mod 4$, and for all $k \in \mathbb{Z}_p$, there exists $t, s \in \mathbb{Z}_p$ for which $t^2 + s^2 = k$.*

Lemma 2. [6, Lemma 5] *Let p be a prime congruent to $3 \mod 4$. For all $k \in \mathbb{Z}_p$ and all $x \in \mathbb{Z}_p[i]$, there exists a $y \in \mathbb{Z}_p[i]$ for which $|x|^2 + |y|^2 \equiv k \mod p$.*

More generally, we will often use unitary equivalence, scaling, and shifting to simplify our calculations. In particular, since for all of our work the resulting numerical range is all of $\mathbb{Z}_p[i]$, any scaling or shifting leaves the result invariant.

Definition 3. [3, Definition 2.5] Let p be a prime congruent to 3 modulo 4 and let $U \in M_n(\mathbb{Z}_p[i])$. We call U a **unitary** matrix if $U^*U = I$.

Lemma 4. [3, Lemma 2.6] *Let $M, U \in M_n(\mathbb{Z}_p[i])$ with U unitary and p a prime congruent to $3 \mod 4$. Then, $W(M) = W(U^*MU)$.*

Lemma 5. [3, Lemma 2.7] *Let p be a prime congruent to $3 \mod 4$ and let $M \in M_n(\mathbb{Z}_p[i])$. For any $a, b \in \mathbb{Z}_p[i]$ we have $W(aM + bI) = aW(M) + b$.*

3 A 0 Entry Above the Diagonal

The following two lemmas appear in an oversimplified form in [6] and in a setting too complex for our needs in [2], and so are reconstructed here. They also have farther-reaching implications than noted in either of those papers, as seen later in this section.

Lemma 6. *For all primes $p \equiv 3 \mod 4$, $W(M) = \mathbb{Z}_p[i]$ where $M \in M_3(\mathbb{Z}_p[i])$ is given by*

$$M = \begin{pmatrix} 0 & 0 & a \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

with $a \neq 0$ in $\mathbb{Z}_p[i]$, or any other 3×3 matrix with a single non-zero entry in $\mathbb{Z}_p[i]$ off of the main diagonal.

Proof. First, assume

$$M = \begin{pmatrix} 0 & 0 & a \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Define $x = (x_1\ x_2\ x_3)^T$, and let $x^*Mx = ax_3\overline{x_1}$ represent elements in the numerical range. (*Note:* $\overline{x_1}$ represents the conjugate of x_1.) For an arbitrary element $k \in \mathbb{Z}_p[i]$, let $x_1 = 1$, and $x_3 = a^{-1}k$, so that $x^*Mx = k$. By Lemma 2, there exists $x_2 \in \mathbb{Z}_p[i]$ such that $|x_2|^2 + |x_3|^2 \equiv 0 \mod p$, so that $|x_1|^2 + |x_2|^2 + |x_3|^2 = 1$. Since k can be any element of $\mathbb{Z}_p[i]$ we have $W(M) = \mathbb{Z}_p[i]$.

If a is in one of the other five spots off of the main diagonal, there is a permutation matrix P so that P^*MP has a in the top-right corner. Since permutation matrices are unitary, by Lemma 4, we still have $W(M) = \mathbb{Z}_p[i]$. □

Lemma 7. *For all primes $p \equiv 3 \mod 4$, $W(M) = \mathbb{Z}_p[i]$ where $M \in M_3(\mathbb{Z}_p[i])$ is given by*

$$M = \begin{pmatrix} 0 & a & 0 \\ 0 & 0 & c \\ 0 & 0 & 0 \end{pmatrix}$$

where $a, c \neq 0$, or any other 3×3 matrix with exactly two non-zero entries, either both above the main diagonal, or both below the main diagonal.

Proof. First assume

$$M = \begin{pmatrix} 0 & a & 0 \\ 0 & 0 & c \\ 0 & 0 & 0 \end{pmatrix}.$$

Consider $x^*Mx = ax_2\overline{x_1} + cx_3\overline{x_2}$. We will again consider a subset of the numerical range by stipulating that $x_2 = 1$.

First, we show that there is a non-zero element in this set. Letting $x_3 = c^{-1}$, we have that $|x_1|^2 \equiv -|c^{-1}|^2$. By Lemma 1, there exists $A, B \in \mathbb{Z}_p$ such that $A^2 + B^2 \equiv -|c^{-1}|^2$, so we will let $x_1 = A + Bi$. Then $x^*Mx = a(A - Bi) + 1$. This is only 0 if $-a^{-1} = (A - Bi)$, in which case we can instead begin by choosing $x_3 = -c^{-1}$, and use the same choice for x_1.

Now, let $a\overline{x_1} + cx_3$ be a fixed non-zero quantity with the constraint that $|x_1|^2 + |x_3|^2 \equiv 0$. Let us now consider $\overline{k}x_1$ and kx_3 where k is an arbitrary element of $\mathbb{Z}_p[i]$. Note that $|\overline{k}x_1|^2 + |kx_3|^2 = |k|^2|x_1|^2 + |k|^2|x_3|^2 = |k|^2(|x_1|^2 + |x_3|^2) = |k|^2(0) = 0$, which satisfies the constraint. Then, the output becomes $ak\overline{x_1} + kcx_3 = k(a\overline{x_1} + cx_3)$. Since $a\overline{x_1} + cx_3$ is fixed and k varies over all of $\mathbb{Z}_p[i]$, we have that $k(a\overline{x_1} + cx_3)$ maps to every element of $\mathbb{Z}_p[i]$, since $k \to \overline{a}^{-1}k$ is an automorphism of $\mathbb{Z}_p[i]$ (where $\overline{a}^{-1} = a\overline{x_1} + cx_3 \in \mathbb{Z}_p[i]^*$). Therefore, $W(M) = \mathbb{Z}_p[i]$.

Now, if M has its two non-zero elements in other entries off of the main diagonal, the roles of x_1, x_2, x_3 can be adjusted accordingly to achieve the same result. For example, if $M = \begin{pmatrix} 0 & a & c \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, let $x_1 = 1$ and apply the same argument to $ax_2 + cx_3$. Similarly,

if $M = \begin{pmatrix} 0 & 0 & a \\ 0 & 0 & c \\ 0 & 0 & 0 \end{pmatrix}$, let $x_3 = 1$ and apply the argument to $a\overline{x_1} + c\overline{x_2}$. If M is instead strictly lower triangular, then conjugation by the standard exchange matrix (which is unitary) will change it to a strictly upper triangular matrix, while preserving the numerical range, so that the prior results may be applied. □

In [2], Lemmas 6 and 7 are considered only as 3×3 without considering higher dimensions, and in [6] little variance is given to where the entries appear and with what values, although higher dimensions are considered. However by using submatrices, these results extend to matrices of arbitrary size.

Lemma 8. *Suppose M_i is a principal submatrix of M created by deleting the ith row and ith column of M. Then $W(M_i) \subseteq W(M)$.*

Proof. If x' is generated from $x \in \mathbb{Z}_p[i]^{n-1}$ by inserting a 0 in position i, then $\|x'\| = \|x\|$ and $\langle x', Mx' \rangle = \langle x, M_i x \rangle$. □

Theorem 9. *Suppose $M \neq 0$ is an $n \times n$ triangular matrix with elements in $\mathbb{Z}_p[i]$ and $n \geq 3$, and a constant diagonal. Suppose also that at least one element above the main diagonal (if M is upper triangular) or below the main diagonal (if M is lower triangular) is 0. Then $W(M) = \mathbb{Z}_p[i]$.*

Proof. If M is strictly lower triangular, it is unitarily equivalent by a permutation matrix (the standard exchange matrix) to a strictly upper triangular matrix, so we will assume M is strictly upper triangular without loss of generality.

We will proceed by induction on n. For the base case ($n = 3$), the statement is a direct corollary of Lemmas 6 and 7.

Suppose $n \geq 4$, and assume for any strictly upper triangular, non-zero $(n-1) \times (n-1)$ matrix with at least one 0 above the main diagonal and all entries in $\mathbb{Z}_p[i]$, the numerical range is $\mathbb{Z}_p[i]$.

If the constant diagonal is not 0, then we may use Lemma 5 and achieve the same result.

For an $n \times n$ matrix M with the same hypotheses, consider the principal submatrix M_i by deleting a row and corresponding column which does not remove all of the zeroes above the diagonal. If M_i is the zero matrix, since $n \geq 4$, there are other rows and corresponding columns that can be instead deleted so that M_i is not 0, while keeping at least one 0 above the diagonal. Once M_i is correctly chosen, by our inductive hypothesis, $W(M_i) = \mathbb{Z}_p[i]$. Then by Lemma 8, $\mathbb{Z}_p[i] = W(M_i) \subseteq W(M)$, so $W(M) = \mathbb{Z}_p[i]$. □

There is a clear third case missing: what if all three entries above the diagonal in a 3×3 matrix are non-zero? Unfortunately, this problem has proved particularly vexing. Testing indicates that all strictly triangular matrices M have $W(M) = \mathbb{Z}_p[i]$, but we are unable to resolve the last piece of the puzzle. In the next section, we will achieve some results in this situation for 4×4 matrices and higher.

It is also worth noting that we are considering *strictly* triangular matrices for another reason beyond simplicity of calculations. In, [3, Example 4.1] a block-reduced upper triangular 3×3 matrix is shown with $W(M) \neq \mathbb{Z}_p[i]$.However, while maintaining some 0 entries above the diagonal, we are able to allow some more variance along the diagonal. This next result can be conjugated by permutations to give results for other similar 3×3 matrices, but this specific form will be useful in the next section, so we leave it as is.

Conjecture 3.5 Suppose M is a 3×3 matrix of the form

$$M = \begin{pmatrix} a & b & 0 \\ 0 & a & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

with all elements in $\mathbb{Z}_p[i]$, $b \neq 0$. Then $W(M) = \mathbb{Z}_p[i]$.

Though we have not completed a proof for this conjecture, we believe such a proof would be possible. As we will show in the next section, the consequences of this conjecture could expand our results to new dimensions.

4 No 0 Entries Above the Diagonal

Here we are able to make progress on some strictly triangular 4×4 matrices with no 0 entries above the diagonal, and then generalize to higher dimensions. The work depends on results about 2×2 matrices.

Theorem 10. *Suppose*

$$M = \begin{pmatrix} 0 & a & b & c \\ 0 & 0 & d & e \\ 0 & 0 & 0 & f \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

with $a,b,c,d,e,f \neq 0$ belonging to $\mathbb{Z}_p[i]$. Suppose further that the 2×2 matrix

$$T = \begin{pmatrix} -\overline{a}^{-1}\overline{b}d & (f - \overline{a}^{-1}\overline{b}e) \\ -\overline{a}^{-1}\overline{c}d & -\overline{a}^{-1}\overline{c}e \end{pmatrix}$$

is such that any element of $\mathbb{Z}_p[i]$ can be represented as y^*Ty where $y \in \mathbb{Z}_p[i]^2$. Then $W(M) = \mathbb{Z}_p[i]$.

Proof. Keep in mind that if $x = (x_1\ x_2\ x_3\ x_4)^T$ is a vector with entries in $\mathbb{Z}_p[i]$, then a typical numerical range element looks like:

$$x^*Mx = \overline{x_1}(ax_2 + bx_3 + cx_4) + \overline{x_2}(dx_3 + ex_4) + f\overline{x_3}x_4.$$

Since $a \neq 0$, it is invertible, we can let $x_2 = -a^{-1}(bx_3 + cx_4)$. Then, the expression becomes

$$\overline{x_1}(ax_2 + bx_3 + cx_4) + \overline{x_2}(dx_3 + ex_4) + f\overline{x_3}x_4 = \tag{1}$$

$$0 - \overline{a}^{-1}(\overline{bx_3} + \overline{cx_4})(dx_3 + ex_4) + f\overline{x_3}x_4 = \tag{2}$$

$$-\overline{a}^{-1}\overline{b}d|x_3|^2 - \overline{a}^{-1}\overline{c}e|x_4|^2 + (f - \overline{a}^{-1}\overline{b}e)\overline{x_3}x_4 - \overline{a}^{-1}\overline{c}d\overline{x_4}x_3. \tag{3}$$

And this final expression represents y^*Ty if $y = (x_3\ x_4)^T$ and

$$T = \begin{pmatrix} -\overline{a}^{-1}\overline{b}d & (f - \overline{a}^{-1}\overline{b}e) \\ -\overline{a}^{-1}\overline{c}d & -\overline{a}^{-1}\overline{c}e \end{pmatrix}.$$

Note that while we are assuming x_2 has a specific form, we have made no assumptions about x_1, x_3, x_4. We can let x_3, x_4 be any values in $\mathbb{Z}_p[i]$, and by Lemma 2, x_1 can always be chosen so that $|x_1|^2 + |x_2|^2 + |x_3|^2 + |x_4|^2 = 1$. Since we assume that y^*Ty can represent any element of $\mathbb{Z}_p[i]$ when x_3 and x_4 can be freely chosen, we are done. □

The assumption of representation in the Theorem 10 is equivalent to requiring that the 2×2 matrix T satisfies $\bigcup_{k \in \mathbb{Z}_p} W_k(T) = \mathbb{Z}_p[i]$. Prior work in [6] and [3] help answer this question.

Lemma 11 ([6, Lemma 10])**.** *Let $A \in M_n(\mathbb{Z}_p[i])$ and let B be the block matrix $\begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix}$. Then $W(B) = \bigcup_{k \in \mathbb{Z}_p} W_k(A)$.*

In [3]; 2×2 numerical ranges are largely reduced to a few specific cases; we will consider those as parts of 3×3 block matrices in the following proof.

Proposition 12 (Corollary of Conjecture 3.5)**.** *Suppose $M \in M_2(\mathbb{Z}_p[i])$ has a single (repeated) eigenvalue in $\mathbb{Z}_p[i]$, with corresponding eigenvectors $v \in \mathbb{Z}_p[i]^2$ satisfying $\|v\|^2 \neq 0$, and M is irreducible. Then $\bigcup_{k \in \mathbb{Z}_p} W_k(M) = \mathbb{Z}_p[i]$.*

Proof. By Lemma 11, we need only show that for any such M, $W(B) = \mathbb{Z}_p[i]$ where $B = \begin{pmatrix} M & 0 \\ 0 & 0 \end{pmatrix}$. By [3, Theorem 1.2], M is unitarily equivalent to an upper triangular matrix; since it has a single eigenvalue, we can write $M = \begin{pmatrix} a & b \\ 0 & a \end{pmatrix}$. By Conjecture 3.5 and Lemma 4.2, $W(B) = \bigcup_{k \in \mathbb{Z}_p} W_k(M) = \mathbb{Z}_p[i]$. □

Putting these together, we can see a clearer form of Theorem 10.

Proposition 13 (Corollary of Conjecture 3.5)**.** *Suppose*

$$M = \begin{pmatrix} 0 & a & b & c \\ 0 & 0 & d & e \\ 0 & 0 & 0 & f \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

with $a, b, c, d, e, f \neq 0$ belonging to $\mathbb{Z}_p[i]$. Suppose further that the 2×2 matrix

$$T = \begin{pmatrix} -\overline{a}^{-1}\overline{b}d & (f - \overline{a}^{-1}\overline{b}e) \\ -\overline{a}^{-1}\overline{c}d & -\overline{a}^{-1}\overline{c}e \end{pmatrix}$$

has a single (repeated) eigenvalue in $\mathbb{Z}_p[i]$, with all eigenvectors v satisfying $\|v\|^2 \neq 0$, and T is irreducible. Then $W(M) = \mathbb{Z}_p[i]$.

Proof. The proof follows immediately from Theorem 10 and Proposition 12. □

Corollary 14. *Suppose M is an $n \times n$ matrix, $n \geq 4$, with a constant diagonal and all entries above the diagonal constant (possibly different from the diagonal). Then $W(M) = \mathbb{Z}_p[i]$.*

Proof. If $n \geq 4$, consider a 4×4 submatrix M' of M. By Lemma 5, we need only consider

$$M' = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The corresponding 2×2 matrix in Proposition 13 is $T = \begin{pmatrix} -1 & 0 \\ -1 & -1 \end{pmatrix}$. This matrix has a single, repeated eigenvalue in $\mathbb{Z}_p[i]$, with eigenvector $v = (0\ 1)^T$ satisfying $\|v\|^2 \neq 0$, so the result follows from Proposition 13. □

Of course, having results for 4×4 then generalizes to higher dimensions.

Corollary 15. *Suppose that $M \in M_n(\mathbb{Z}_p[i]), n \geq 5$, has a principal submatrix that satisfies the conditions of Proposition 13. Then $W(M) = \mathbb{Z}_p[i]$.*

Proof. This follows directly from Theorem 10 and Proposition 12. □

Unfortunately, though Theorem 10 and Proposition 12 are sufficient to prove Corollary 15, they are not necessary. The matrix

$$M = \begin{pmatrix} 0 & 1 & 4+2i & 4+4i \\ 0 & 0 & 1+6i & 1+6i \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

over $\mathbb{Z}_7[i]$ has a full numerical range, but the corresponding 2×2 matrix

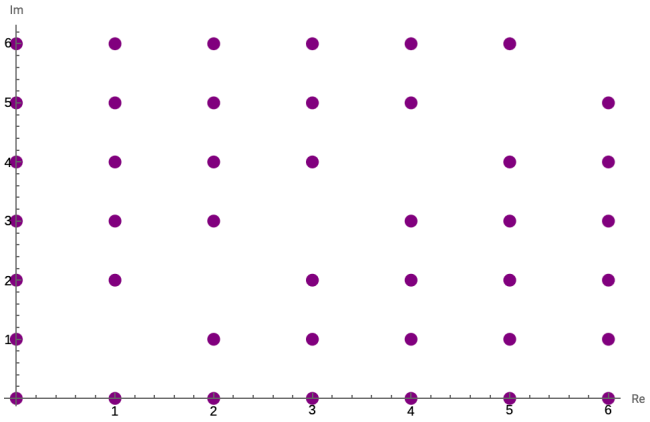
$$T = \begin{pmatrix} 5+6i & 6i \\ i & i \end{pmatrix}$$

does not have $\bigcup_{k \in \mathbb{Z}_7} W_k(T) = \mathbb{Z}_7[i]$. By Lemma 11, this can be seen by viewing $W(B)$ where

$$B = \begin{pmatrix} 5+6i & 6i & 0 \\ i & i & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

That image is shown in Figure 1. It is noteworthy that the eigenvalues of T, $\frac{1}{2}((5+7i) \pm \sqrt{-24+50i})$, do not belong to $\mathbb{Z}_7[i]$.

Figure 1: $W(M) = \mathbb{Z}_7[i]$, but $\bigcup_{k\in\mathbb{Z}_7} W_k(T) \neq \mathbb{Z}_7[i]$.



5 Future Work

Our biggest concern is finishing the case of 3×3 strictly triangular matrices. We feel fairly confident in the following conjecture.

Conjecture. *If $M \in M_n(\mathbb{Z}_p[i]), n \geq 3$ is strictly triangular, then $W(M) = \mathbb{Z}_p[i]$.*

More broadly, we believe, based on our preliminary explorations, that the numerical range of all 3×3 matrices over $\mathbb{Z}_p[i]$ can be classified into one of a few finite categories. Much work still needs done to identify the criteria for determining the size of the numerical range of a given matrix, but examples of each of these numerical range shapes may be found in Appendix .

Conjecture. *If $M \in M_3(\mathbb{Z}_p[i])$, then $W(M)$ contains either 1 element, p elements, p^2-1 elements, $p^2-(p-1)$ elements, or p^2 elements.*

Beyond that, in [3, Propositon 3.4], a variation of Schur's Theorem for 2×2 matrices over $\mathbb{Z}_p[i]$ is established. If that generalizes to higher dimensions, then we have the following conjecture.

Conjecture. *If $M \in M_n(\mathbb{Z}_p[i]), n \geq 3$ has a single eigenvalue, belonging to $\mathbb{Z}_p[i]$, with all eigenvectors v satisfying $\|v\|^2 \neq 0$, then $W(M) = \mathbb{Z}_p[i]$.*

Acknowledgements

I would like to thank the Taylor University Women's Giving Circle funding this research. Special thanks to Derek Thompson for his continual guidance and mentorship. I also thank Jordan Crawford, Edoardo Ballico, and Patrick X. Rault for their input.

Bibliography

[1] Ballico, E. (2018). "The Hermitian Null-range of a Matrix over a Finite Field", *Electronic Journal of Linear Algebra*, Volume 34, pp.205-216. DOI:https://doi.org/10.13001/1081-3810,1537-9582.3416

[2] Ballico, E. (2019). Numerical range over finite fields: Restriction to subspaces, *Linear Algebra and its Applications*, Volume 571, pp. 1-13.

[3] Coons, J.I., Jenkins, J., Knowles, D., Luke, R.A., Rault, P.X. (2016). Numerical Ranges Over Finite Fields. *Elsevier: Linear Algebra and its Applications, 501*. Retrieved From https://www.sciencedirect.com/science/article/pii/S0024379516300106

[4] Gallier, J. (2011). *Geometric Methods and Applications For Computer Science and Engineering*. New York, NY. Springer Science+ Business Media.

[5] Martiǹez-Avendanõ, R.A., Rosenthal, P. (2007). *An Introduction to Operators on the Hardy-Hilbert Space*. New York, NY. Springer Science+ Business Media.

[6] Guillaume, M., Mishra, A., Thompson, D. (2020). Numerical Range of Toeplitz Matrices over Finite Fields. *Proceedings of the ACMS*, Volume 22, pp.151-158.

[7] Zachlin, P.F., Hochstenback, M.E. (2007). *On the numerical range of a matrix.* (CASA-report; Vol.0702). Eindhoven: Technische Universiteit Eidehoven.

Appendices

1 Examples of Various Sizes of Numerical Ranges

1.1 $W(M) = \mathbb{Z}$

As established, we have reason to believe that strictly triangular matrices in $\mathbb{Z}$ have $W(M) = \mathbb{Z}$. However, we can see by these examples that such matrices are not the *only* matrices to satisfy this.

Consider the following examples of matrices $M \in \mathbf{Z}_7[i]$ which satisfy $W(M) = \mathbb{Z}$.

As in Figure 2, each of the following matrices $M \in \mathbf{Z}_7[i]$ satisfies $W(M) = \mathbb{Z}$.

- $M = \begin{pmatrix} 1+4i & 5i & 4+5i \\ 1+2i & 2i & 6+2i \\ 2+i & 3+5i & 1+6i \end{pmatrix}$
- $M = \begin{pmatrix} 1+i & 2+6i & 6+i \\ 2+3i & 5+6i & 3+3i \\ 5+i & 4 & i \end{pmatrix}$
- $M = \begin{pmatrix} 4+6i & 3 & 1+2i \\ 2+i & 4+6i & 1+3i \\ 2+4i & 2+4i & 4i \end{pmatrix}$
- $M = \begin{pmatrix} 4+6i & 1+3i & 4+2i \\ 4+5i & 5+4i & 0 \\ 3 & 6+i & 3+6i \end{pmatrix}$
- $M = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$
- $M = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}$

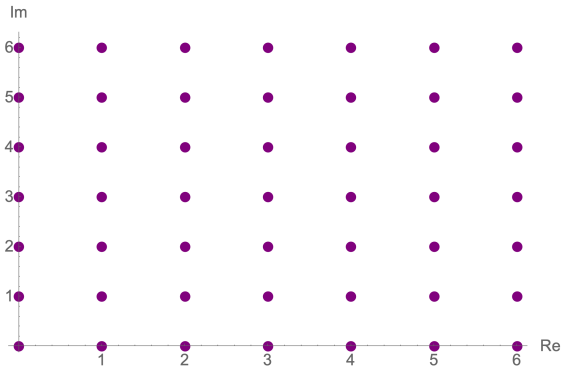


Figure 2: The numerical range of M when $W(M) = \mathbb{Z}_7[i]$

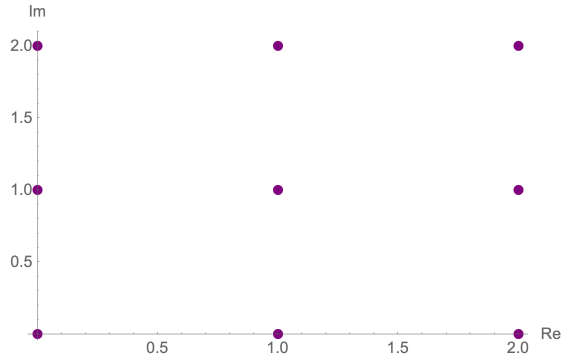


Figure 3: The numerical range of M when $W(M) = \mathbb{Z}_3[i]$

Below, we also have a selection of examples of $M \in \mathbf{Z}_3[i]$ which satisfy $W(M) = \mathbf{Z}_3[i]$, as in Figure 3

- $M = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}$
- $M = \begin{pmatrix} 1+i & 2 & 2i \\ 1 & 0 & 1 \\ 1 & 0 & 1+i \end{pmatrix}$
- $M = \begin{pmatrix} 1+i & 2 & 2i \\ 1 & i & i \\ 2 & 2i & 1+i \end{pmatrix}$
- $M = \begin{pmatrix} 1 & i & 1 \\ i & 1 & i \\ 1 & i & 1 \end{pmatrix}$

1.2 $|W(M)| = p^2 - 1$

We have no current conjecture regarding how to classify these matrices, but in our exploration we identified several instances of where $W(M)$ contains all but one element of $\mathbb{Z}$. A selection of examples are given below.

Example. $M \in \mathbf{Z}_7[i], M = \begin{pmatrix} 0 & 6+5i & 3i \\ 4+3i & 0 & 6+5i \\ 2+5i & 4+3i & 0 \end{pmatrix}$ with $W(M)$ as shown in Figure

4.

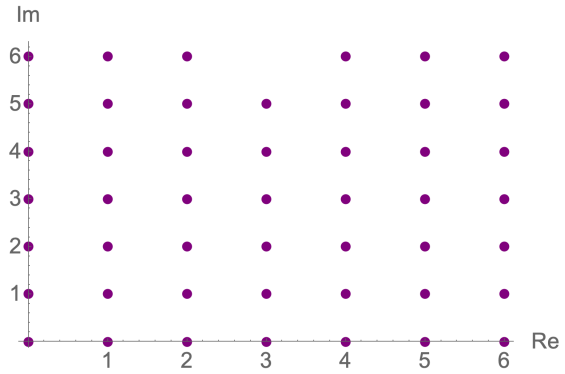


Figure 4: $W(M)$ with $M = \begin{pmatrix} 0 & 6+5i & 3i \\ 4+3i & 0 & 6+5i \\ 2+5i & 4+3i & 0 \end{pmatrix}$ over $\mathbf{Z}_7[i]$

Example. $M \in \mathbf{Z}_7[i], M = \begin{pmatrix} 4+i & 5+3i & 5i \\ 2+3i & 1+4i & 4i \\ 1+i & 4i & 4+3i \end{pmatrix}$ with $W(M)$ as shown in Figure 5.

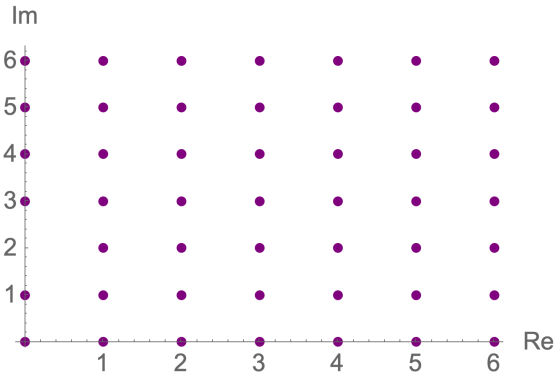


Figure 5: $W(M)$ with $M = \begin{pmatrix} 4+i & 5+3i & 5i \\ 2+3i & 1+4i & 4i \\ 1+i & 4i & 4+3i \end{pmatrix}$ over $\mathbf{Z}_7[i]$

Example. $M \in \mathbf{Z}_3[i], M = \begin{pmatrix} 2+2i & 2i & 2i \\ 1+2i & 0 & 2+2i \\ 1 & 2 & 1+2i \end{pmatrix}$ with $W(M)$ as shown in Figure 6.

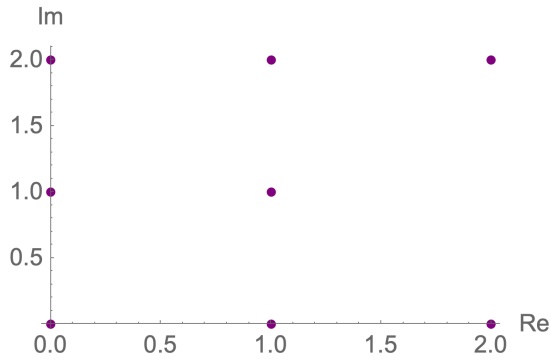


Figure 6: $W(M)$ with $M = \begin{pmatrix} 2+2i & 2i & 2i \\ 1+2i & 0 & 2+2i \\ 1 & 2 & 1+2i \end{pmatrix}$ over $\mathbf{Z}_3[i]$

Example. $M \in \mathbf{Z}_3[i], M = \begin{pmatrix} 1 & 2+i & i \\ 2 & 0 & 1+i \\ 2+2i & i & 0 \end{pmatrix}$ with $W(M)$ as shown in Figure 7.

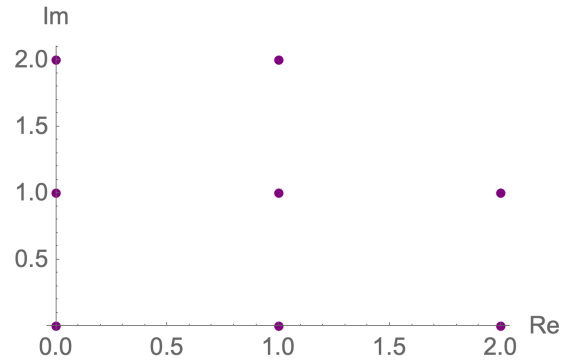


Figure 7: $W(M)$ with $M = \begin{pmatrix} 1 & 2+i & i \\ 2 & 0 & 1+i \\ 2+2i & i & 0 \end{pmatrix}$ over $\mathbf{Z}_3[i]$

1.3 $|W(M)| = p^2 - (p-1)$

We also provide here a selection of examples where $W(M)$ is missing $p-1$ elements of $\mathbb{Z}$. Intuitively, this means that there is nearly a "line" missing.

Example. $M \in \mathbf{Z}_7[i], M = \begin{pmatrix} 0 & 3 & 5 \\ 6 & 0 & 3 \\ 4 & 6 & 0 \end{pmatrix}$ with $W(M)$ as shown in Figure 8. You can see that all elements $6+bi, b \neq 0$ are excluded from the numerical range.

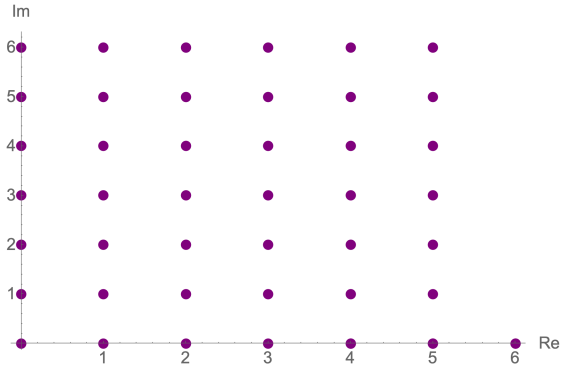


Figure 8: $W(M)$ with $M = \begin{pmatrix} 0 & 3 & 5 \\ 6 & 0 & 3 \\ 4 & 6 & 0 \end{pmatrix}$ over $\mathbf{Z}_7[i]$

Example. $M \in \mathbf{Z}_3[i], M = \begin{pmatrix} 0 & 2+3i & 1+5i \\ 4+6i & 0 & 2+3i \\ 5+4i & 4+6i & 0 \end{pmatrix}$ with $W(M)$ as shown in Figure 9. This matrix is the same as the previous matrix, scaled by a factor of $3+i$. Transformations like this rotate the numerical range according to the factor it was scaled by. In Figure 9, the "missing line" is still visible, identifiable with a "slope" of 4.

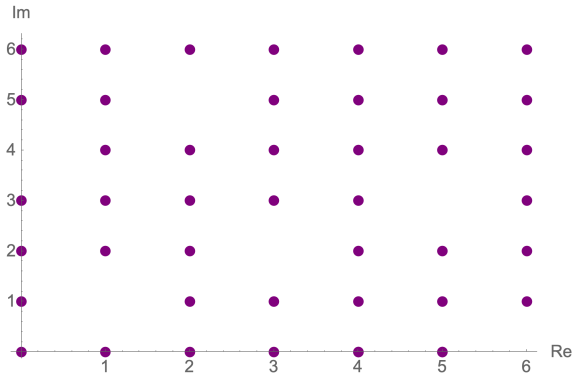


Figure 9: $W(M)$ with $M = \begin{pmatrix} 0 & 2+3i & 1+5i \\ 4+6i & 0 & 2+3i \\ 5+4i & 4+6i & 0 \end{pmatrix}$ over $\mathbf{Z}_7[i]$

Example. $M \in \mathbf{Z}_3[i], M = \begin{pmatrix} 1+i & 2 & i \\ 2+2i & 2i & 0 \\ i & 2+i & 2 \end{pmatrix}$ with $W(M)$ as shown in Figure 10. We see that the elements $1+2i$ and $2+i$ are not included in the numerical range.

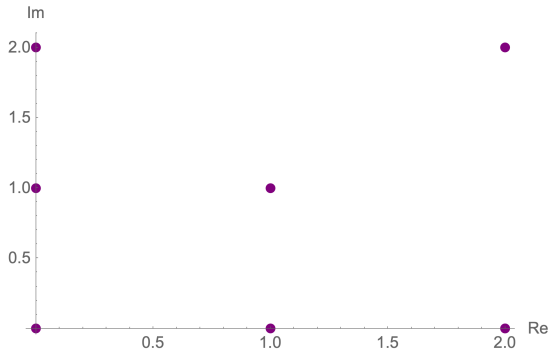


Figure 10: $W(M)$ with $M = \begin{pmatrix} 1+i & 2 & i \\ 2+2i & 2i & 0 \\ i & 2+i & 2 \end{pmatrix}$ over $\mathbf{Z}_3[i]$

1.4 $|W(M)| = p$

Matrices which are equal to their own conjugate transpose have a numerical range of $\mathbf{Z}_p$. Multiples of such matrices have numerical ranges with p elements, rotated off of the $\mathbf{Z}_p$ line according the the factor the matrix was scaled by.

Example. $M \in \mathbf{Z}_7[i], M = \begin{pmatrix} 0 & 5i & 3i \\ 2i & 0 & 5i \\ 4i & 2i & 0 \end{pmatrix}$ with $W(M)$ as shown in Figure 11.

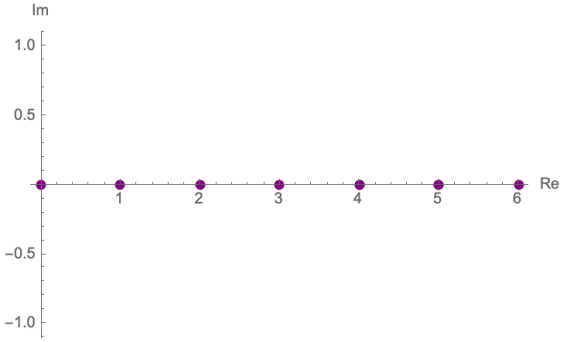


Figure 11: $W(M)$ with $M = \begin{pmatrix} 0 & 5i & 3i \\ 2i & 0 & 5i \\ 4i & 2i & 0 \end{pmatrix}$ over $\mathbf{Z}_7[i]$

Example. $M \in \mathbf{Z}_7[i], M = \begin{pmatrix} 0 & 2+5i & 4+3i \\ 5+2i & 0 & 2+5i \\ 3+4i & 5+2i & 0 \end{pmatrix}$ with $W(M)$ as shown in Figure 12. This matrix is the previous example, scaled by a factor of $1+i$.

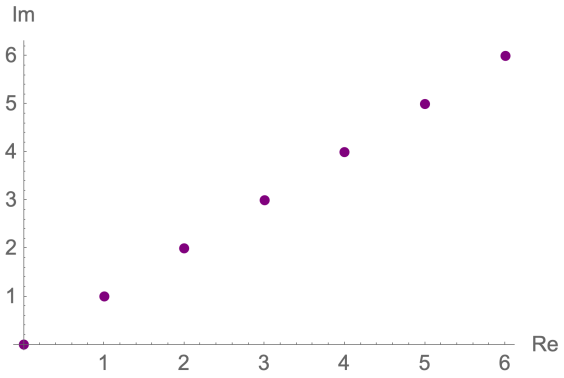


Figure 12: $W(M)$ with $M = \begin{pmatrix} 0 & 2+5i & 4+3i \\ 5+2i & 0 & 2+5i \\ 3+4i & 5+2i & 0 \end{pmatrix}$ over $\mathbf{Z}_7[i]$

1.5 $|W(M)| = 1$

The zero matrix (or a shifted zero matrix) has only 1 element in its numerical range.

Example. $M \in \mathbf{Z}_3[i], M = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ with $W(M)$ as shown in Figure 13.

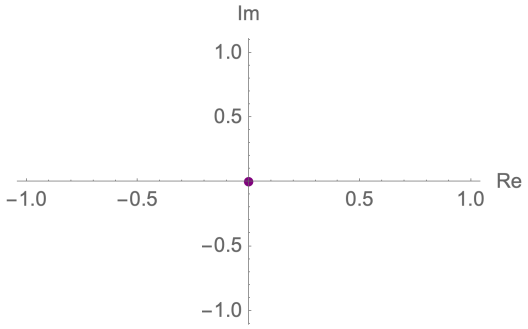


Figure 13: $W(M)$ with $M = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ over $\mathbf{Z}_3[i]$

2 Mathematica Code to Replicate Results

In our research, we relied heavily on computation to explore and verify results. We include key aspects of our Mathematica code here for the purposes of replication. Great thanks to Amish Mishra for writing the original version of the code, which we have adapted to be what is included here.

2.1 Preliminaries

Several functions and variables must be defined in order to calculate and plot the numerical range of a matrix over a finite field. Since the built-in functions do not account for finite fields, we must build our own.

```
p = 3; (* Change for different size finite field *)

plotNumericalRange[numRange_] := (
plotPoints = {};
Do[
AppendTo[
plotPoints, {Re[numRange[[each]][[1]][[1]]],
        Im[numRange[[each]][[1]][[1]]]}];
, {each, 1, Length[numRange]}];
ListPlot[plotPoints,
PlotStyle -> Directive[Purple, PointSize[.02]],
AxesLabel -> {Re, Im}]
)

zpi = {};
Do[
Do [
AppendTo[zpi, a + b*I];
, {b, 0, p - 1}];
, {a, 0, p - 1}];

ZpiArray3x3[p_, n_]  := (
Module[{Z1, num, Zpi},
Z1 = ConstantArray[0, {p, p, p}]; (* TODO: make this dynamic,
maybe with another ConstantArray? *)
Do[
Clear[num];
Do[
num = a + b*I;
Z1[[a + 1, b + 1]] = num;
, {b, 0, p - 1}]
, {a, 0 , p - 1}];
Zpi = Tuples[Flatten[Z1], {n}];
Zpi
]
)
numericalRange3x3[k_, M_, p_, n_] := (
Module[{numRange, Zpi},
numRange = {};
Zpi = ZpiArray3x3[p, n];
Do[
x = Zpi[[idx]];
```

```
norm = Mod[x.{x}\[ConjugateTranspose], p][[1]];
If[norm == k,
numRangeElem =
Mod[Transpose[{x}]\[ConjugateTranspose].M.Transpose[{x}], p];
If[! MemberQ[numRange, numRangeElem],
AppendTo[numRange, numRangeElem]
]
];
, {idx, 1, Length[Zpi]}];
(* This portion makes all terms of the numerical range the \
positive modulo p. *)
Do[
If[Re[numRange[[idx]]][[1]][[1]] < 0,
numRange[[idx]][[1]][[1]] = numRange[[idx]][[1]][[1]] + p;
];
If[Im[numRange[[idx]]][[1]][[1]] < 0,
numRange[[idx]][[1]][[1]] = numRange[[idx]][[1]][[1]] + p*I;
];
, {idx, 1, Length[numRange]}];
nr = {};
Do[
If[! MemberQ[nr, numRange[[i]]],
AppendTo[nr, numRange[[i]]];
];
, {i, 1, Length[numRange]}];
nr]
)
```

2.2 Plotting a numerical range

Once we have our functions defined, we can utilize them to calculate and plot a numerical range.

```
(* Change M to a 3x3 matrix here. To get an
imaginary symbol, type esc i i esc *)

M = {{1, 0, 0}, {0, 1, 0}, {0, 0, 1}};

Print["M: ", MatrixForm[M]];
plotNumericalRange[numericalRange3x3[0, M, p, 3]]
```