

involve

a journal of mathematics

Editorial Board

Kenneth S. Berenhaut, *Managing Editor*

Colin Adams
Arthur T. Benjamin
Martin Bohner
Amarjit S. Budhiraja
Scott Chapman
Joshua N. Cooper
Michael Dorff
Joel Foisy
Amanda Folsom
Stephan R. Garcia
Anant Godbole
Ron Gould
Sat Gupta
Jim Haglund
Glenn H. Hurlbert
Michael Jablonski
Nathan Kaplan
David Larson

Suzanne Lenhart
Chi-Kwong Li
Robert B. Lund
Gaven J. Martin
Steven J. Miller
Frank Morgan
Mohammad Sal Moslehian
Ken Ono
Jonathon Peterson
Vadim Ponomarenko
Bjorn Poonen
József H. Przytycki
Javier Rojo
Filip Saidak
Ann Trenk
Ravi Vakil
John C. Wierman



involve

msp.org/involve

INVOLVE YOUR STUDENTS IN RESEARCH

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

MANAGING EDITOR

Kenneth S. Berenhaut Wake Forest University, USA

BOARD OF EDITORS

Colin Adams	Williams College, USA	Suzanne Lenhart	Univ. of Tennessee, USA
Arthur T. Benjamin	Harvey Mudd College, USA	Chi-Kwong Li	College of William and Mary, USA
Martin Bohner	Missouri Univ. of Science and Tech., USA	Robert B. Lund	Clemson Univ., USA
Amarjit S. Budhiraja	Univ. of North Carolina, Chapel Hill, USA	Gaven J. Martin	Massey Univ., New Zealand
Scott Chapman	Sam Houston State Univ., USA	Steven J. Miller	Williams College, USA
Joshua N. Cooper	Univ. of South Carolina, USA	Frank Morgan	Williams College, USA
Michael Dorff	Brigham Young Univ., USA	Mohammad Sal Moslehian	Ferdowsi Univ. of Mashhad, Iran
Joel Foisy	SUNY Potsdam, USA	Ken Ono	Univ. of Virginia, Charlottesville
Amanda Folsom	Amherst College, USA	Jonathon Peterson	Purdue Univ., USA
Stephan R. Garcia	Pomona College, USA	Vadim Ponomarenko	San Diego State Univ., USA
Anant Godbole	East Tennessee State Univ., USA	Bjorn Poonen	Massachusetts Institute of Tech., USA
Ron Gould	Emory Univ., USA	József H. Przytycki	George Washington Univ., USA
Sat Gupta	Univ. of North Carolina, Greensboro, USA	Javier Rojo	Oregon State Univ., USA
Jim Haglund	Univ. of Pennsylvania, USA	Filip Saidak	Univ. of North Carolina, Greensboro, USA
Glenn H. Hurlbert	Virginia Commonwealth Univ., USA	Ann Trenk	Wellesley College, USA
Michael Jablonski	Univ. of Oklahoma, USA	Ravi Vakil	Stanford Univ., USA
Nathan Kaplan	Univ. of California, Irvine, USA	John C. Wierman	Johns Hopkins Univ., USA
David Larson	Texas A&M Univ., USA		

PRODUCTION

Silvio Levy, Scientific Editor


Cover: Alex Scorpan

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2024 is US \$255/year for the electronic version, and \$340/year (+\$45, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP.

Involve (ISSN 1944-4184 electronic, 1944-4176 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840, is published continuously online.

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

PUBLISHED BY

 **mathematical sciences publishers**

nonprofit scientific publishing

<http://msp.org/>

© 2024 Mathematical Sciences Publishers

An r^p -weighted local energy approach to global existence for null form semilinear wave equations

Michael Facci, Alex McEntarrfer and Jason Metcalfe

(Communicated by Kenneth S. Berenhaut)

We revisit the proof of small-data global existence for semilinear wave equations that satisfy a null condition. This new approach relies on a weighted local energy estimate that is akin to those of Dafermos and Rodnianski. Using weighted Sobolev estimates to obtain spatial decay and arguing in the spirit of the work of Keel, Smith, and Sogge, we are able to obtain global existence while only relying on translational and (spatial) rotational symmetries.

1. Introduction

We shall examine systems of semilinear wave equations in (1+3)-dimensions of the form

$$\begin{cases} \square u^I := (\partial_t^2 - \Delta)u^I = Q^I(\partial u), & (t, x) \in \mathbb{R}_+ \times \mathbb{R}^3, \quad I = 1, 2, \dots, M, \\ u^I(0, \cdot) = f^I, \quad \partial_t u^I(0, \cdot) = g^I. \end{cases} \quad (1-1)$$

Here $\partial u = (\partial_t u, \nabla u)$ is the space-time gradient, and each component Q^I is a smooth function that vanishes to second order at the origin. As we shall only consider small data, the long-time behavior is dictated by the lowest-order terms, and, as such, we will truncate Q to the quadratic level.

As the linear wave equation decays like $t^{-(n-1)/2}$ in n -spatial dimensions and as this factor is integrable at infinity when $n \geq 4$, it has long been known that global existence of solutions to (1-1) for sufficiently small initial data is guaranteed in these dimensions. When $n = 3$, however, a logarithmic blow up is instead encountered, and only almost global existence, which states that the lifespan of the solution grows exponentially as the size of the initial data shrinks, is available generically; see, e.g., [Sogge 2008].

MSC2020: 35L05, 35L71.

Keywords: semilinear wave equations, null condition, global existence, local energy estimate.

Metcalfe gratefully acknowledges the support of a Simons Foundation Collaboration Grant (711724) and National Science Foundation grants DMS-2054910 and DMS-2135998.

When the nonlinearity is assumed to satisfy a null condition, it was discovered in [Christodoulou 1986; Klainerman 1986] that sufficiently small initial data always produce global solutions in three dimensions. In the current setting, assuming that our quadratic nonlinearity is of the form

$$Q^I(\partial u) = A_{JK}^{\alpha\beta, I} \partial_\alpha u^J \partial_\beta u^K,$$

the null condition requires that

$$A_{JK}^{\alpha\beta, I} \xi_\alpha \xi_\beta = 0, \quad \text{when } \xi_0^2 - \xi_1^2 - \xi_2^2 - \xi_3^2 = 0. \quad (1-2)$$

Here we are using the summation convention with α, β running from 0 to 3 and the common conventions that $\partial_0 u = \partial_t u$, $\partial_j u = \partial_{x_j} u$. We are also allowing repeated capital indices to sum from 1 to M .

A common approach for establishing such long-time existence results relies on the method of invariant vector fields and the Klainerman–Sobolev inequality [Klainerman 1985]. Due to the unbounded normal component on the boundary, the Lorentz boosts $x_k \partial_t + t \partial_k$ are inappropriate when studying such nonlinear equations, say, exterior to a compact obstacle with Dirichlet boundary conditions. In response, [Keel et al. 2002] developed a method of establishing long-time existence for three-dimensional semilinear wave equations that only relies upon the generators of translations and spatial rotations:

$$\Omega_{ij} = x_i \partial_j - x_j \partial_i, \quad Z = (\partial_1, \partial_2, \partial_3, \Omega_{23}, \Omega_{13}, \Omega_{12}).$$

Here the authors depended on the integrated local energy estimate, which will be introduced in Section 2, and a weighted Sobolev estimate [Klainerman 1986] that provided decay in $|x|$ rather than t but only requires the vector fields Z . This method was adapted to the quasilinear setting in [Metcalf and Sogge 2006] by exploring local energy estimates for perturbations of the d’Alembertian. The desire for a method that did not necessitate the use of the Lorentz boosts was also motivated by wanting to understand multiple speed systems of wave equations and the equations of elasticity; see, e.g., [Klainerman and Sideris 1996; Sideris 2000].

Here we shall explore small-data global existence for null-form wave equations. Many approaches exist for establishing such global existence, see, e.g., [Klainerman 1986; Christodoulou 1986; Sideris and Tu 2001; Metcalfe and Sogge 2007; Katayama and Kubo 2008; Lindblad et al. 2013]. Unlike many of the preceding results, our method shall only rely on the time-independent vector fields Z .

The key to our argument is to replace the use of the local energy estimate with a variant, specifically a type of r^p -weighted local energy estimate of [Dafermos and Rodnianski 2010]. See [Moschidis 2016] for some generalizations of this method. This estimate has been applied in a number of nonlinear settings such as [Luk 2013; Yang 2015a; 2015b; Keir 2018]. Typically it is used to derive decay in t . Such decay is then used to control the integral within the energy inequality and thus

provides long-time existence. We believe our approach to be more straightforward, though those preceding results were all in much more complicated settings.

The r^p -weighted local energy estimate only controls the “good” derivatives $\partial = (\partial_t + \partial_r, \nabla)$, where $\nabla = \nabla - (x/r)\partial_r$ are the angular derivatives. These are the directions that are tangent to the light cone and for which better decay is known. The r^p -weighted estimate is particularly well-suited to null form wave equations as the algebraic cancellation condition (1-2) precisely guarantees that in each quadratic term of $Q(\partial u)$ one of the two factors is a good derivative.

Our main result is:

Theorem 1.1. *Suppose that $f, g \in (C^\infty(\mathbb{R}^3))^M$. And let $0 < p < 1$. Then, for any $\varepsilon > 0$ sufficiently small, if*

$$\|(1+r)^{p/2} Z^{\leq 10} f\|_{L^2(\mathbb{R}^3)} + \|(1+r)^{p/2} Z^{\leq 9} g\|_{L^2(\mathbb{R}^3)} \leq \varepsilon, \quad (1-3)$$

then (1-1) with nonlinearity satisfying (1-2) has a unique global solution $u \in C^\infty(\mathbb{R}_+ \times \mathbb{R}^3)$.

Here, and throughout, we shall use the abbreviation $Z^{\leq N} u = \sum_{|\alpha| \leq N} Z^\alpha u$.

To keep the exposition as accessible as possible, we have only focused on semilinear equations on Minkowski space. We expect that the argument can readily be extended to, e.g., quasilinear equations and equations on exterior domains, and these topics will be explored subsequently.

Our proof of [Theorem 1.1](#) most resembles [\[Lindblad et al. 2013\]](#). There an alternate local energy estimate that relies upon $t-r$ weights, which is from [\[Lindblad and Rodnianski 2005; Alinhac 2001\]](#), was used. In order to achieve the decay in $t-r$, the authors called upon decay estimates of [\[Klainerman and Sideris 1996\]](#), but these in turn required the use of the time-dependent vector fields. The current argument is much more directly reminiscent of [\[Keel et al. 2002\]](#).

2. Integrated local energy estimates

The integrated local energy estimate first appeared in [\[Morawetz 1968\]](#). Through subsequent refinements, on $\mathbb{R}_+ \times \mathbb{R}^n$, $n \geq 3$, we know that

$$\begin{aligned} \|\partial u\|_{L_t^\infty L_x^2}^2 + \sup_{R \geq 1} R^{-1} \|\partial u\|_{L_t^2 L_x^2(\mathbb{R}_+ \times \{(x) \approx R\})}^2 + \sup_{R \geq 1} R^{-3} \|u\|_{L_t^2 L_x^2(\mathbb{R}_+ \times \{(x) \approx R\})}^2 \\ \lesssim \|\partial u(0, \cdot)\|_{L^2}^2 + \int_0^\infty \int |\square u| (|\partial u| + \langle x \rangle^{-1} |u|) dx dt. \end{aligned} \quad (2-1)$$

The most robust proof of this estimate pairs the equation $\square u$ with a multiplier of the form

$$C \partial_t u + \frac{r}{r+R} \partial_r u + \frac{n-1}{2} \frac{1}{r+R} u$$

and follows from integration by parts; see, e.g., [\[Sterbenz 2005; Metcalfe and Sogge 2006\]](#). Related estimates are known to hold for stationary, nontrapping perturbations

and for sufficiently small nonstationary perturbations. See [Metcalf et al. 2020] for a more complete history and the most general results in the nontrapping setting.

Our first task will be to prove the following r^p -weighted estimate, which first appeared in [Dafermos and Rodnianski 2010].

Proposition 2.1. *Suppose $u \in C^\infty(\mathbb{R}_+ \times \mathbb{R}^3)$ and that for every T there is an R so that $u(t, x) = 0$ for $t \in [0, T]$ and $|x| > R$. Then, for $0 < p < 1$,*

$$\|r^{(p-1)/2} \not\partial u\|_{L_t^2 L_x^2}^2 + \|r^{(p-3)/2} u\|_{L_t^2 L_x^2}^2 + \sup_t \tilde{E}[u](t) \lesssim \tilde{E}[u](0) + \|r^{(p+1)/2} \square u\|_{L_t^2 L_x^2}^2, \quad (2-2)$$

where

$$\tilde{E}[u](t) = \frac{1}{2} \int r^{p-2} |\not\partial(ru(t, x))|^2 dx + \frac{p}{2} \int r^{p-2} u^2(t, x) dx.$$

The local energy estimate (2-1) has an ℓ^∞ -summation over the annuli, which we may take to be dyadic, in the left side. In [Keel et al. 2002], the difference between this and having ℓ^2 -summability accounts for a logarithm, which in turn corresponds to the exponential within the notion of almost global existence. While restricted only to the good directions, the above estimate has the desired ℓ^2 -summability, and as such, it will yield global existence so long as the equation permits its application on each term, which the null condition exactly provides.

Proof. For any $0 \leq p \leq 2$, we first consider

$$\begin{aligned} \int_0^T \int \square u \cdot r^p \left(\partial_t u + \partial_r u + \frac{1}{r} u \right) dx dt \\ = \int_0^T \int \int r^p (\partial_t^2 - \partial_r^2 - \not\partial \cdot \not\partial)(ru) (\partial_t + \partial_r)(ru) d\sigma dr dt. \end{aligned}$$

Using integration by parts and the fact that $[\not\partial, \partial_r] = (1/r)\not\partial$, the right side is equal to

$$\begin{aligned} \frac{1}{2} \int_0^T \int \int r^p (\partial_t - \partial_r) [(\partial_t + \partial_r)(ru)]^2 d\sigma dr dt \\ + \frac{1}{2} \int_0^T \int \int r^p (\partial_t + \partial_r) |\not\partial(ru)|^2 d\sigma dr dt + \int_0^T \int \int r^{p-1} |\not\partial(ru)|^2 d\sigma dr dt. \end{aligned}$$

Further integrating by parts yields

$$\begin{aligned} \int_0^T \int \square u \cdot r^p \left(\partial_t u + \partial_r u + \frac{1}{r} u \right) dx dt \\ = \frac{1}{2} \int \int r^p \{ [(\partial_t + \partial_r)(ru)]^2 + |\not\partial(ru)|^2 \} d\sigma dr \Big|_{t=0}^T \\ + \frac{p}{2} \int_0^T \int \int r^{p-1} [(\partial_t + \partial_r)(ru)]^2 d\sigma dr dt \\ + \left(1 - \frac{p}{2} \right) \int_0^T \int \int r^{p-1} |\not\partial(ru)|^2 d\sigma dr dt. \quad (2-3) \end{aligned}$$

For simplicity, we now restrict to $0 \leq p < 1$. We then observe that

$$\begin{aligned} \frac{P}{2} \int_0^T \iint r^{p-1} [(\partial_t + \partial_r)(ru)]^2 d\sigma dr dt &= \frac{P}{2} \int_0^T \int r^{p-1} (\partial_t u + \partial_r u)^2 r^2 d\sigma dr dt \\ &\quad + \frac{P}{2} \int_0^T \iint r^p (\partial_t + \partial_r) u^2 d\sigma dr dt \\ &\quad + \frac{P}{2} \int_0^T \iint r^{p-1} u^2 d\sigma dr dt, \end{aligned}$$

which upon a last integration by parts and reverting back to rectangular coordinates gives

$$\frac{P}{2} \int r^{p-2} u^2 dx \Big|_{t=0}^T + \frac{P}{2} \int_0^T \int r^{p-1} (\partial_t u + \partial_r u)^2 dx dt + \frac{P(1-p)}{2} \int_0^T \int r^{p-3} u^2 dx dt.$$

Making this replacement in (2-3) and applying the Schwarz inequality gives

$$\begin{aligned} \frac{P}{2} \|r^{(p-1)/2} (\partial_t + \partial_r) u\|_{L_t^2 L_x^2}^2 + \frac{2-p}{2} \|r^{(p-1)/2} \not\partial u\|_{L_t^2 L_x^2}^2 \\ + \frac{P(1-p)}{2} \|r^{(p-3)/2} u\|_{L_t^2 L_x^2}^2 + \tilde{E}[u](T) \\ \leq \tilde{E}[u](0) + \|r^{(p+1)/2} \square u\|_{L_t^2 L_x^2} (\|r^{(p-1)/2} (\partial_t + \partial_r) u\|_{L_t^2 L_x^2} + \|r^{(p-3)/2} u\|_{L_t^2 L_x^2}). \end{aligned}$$

Bootstrapping the last factor of the last term completes the proof. We moreover note that the implicit constant is independent of T , and thus we may take the supremum over all T to obtain (2-2). \square

In the sequel, we shall require a version of (2-2) that permits the application of the invariant vector fields, which is presented in the next proposition.

Proposition 2.2. *Let $0 < p < 1$ and fix any $N \in \mathbb{N}$. Suppose $u \in C^\infty(\mathbb{R}_+ \times \mathbb{R}^3)$ and that, for every T , there is an R so that $u(t, x) = 0$ for $t \in [0, T]$ and $|x| > R$. Then,*

$$\begin{aligned} \|Z^{\leq N} \partial u\|_{L_t^\infty L_x^2}^2 + \|(1+r)^{(p-1)/2} Z^{\leq N} \not\partial u\|_{L_t^2 L_x^2}^2 + \|(1+r)^{(p-3)/2} Z^{\leq N} u\|_{L_t^2 L_x^2}^2 \\ \lesssim \|(1+r)^{p/2} Z^{\leq N} \partial u(0, \cdot)\|_{L^2}^2 + \|(1+r)^{(p+1)/2} Z^{\leq N} \square u\|_{L_t^2 L_x^2}^2. \quad (2-4) \end{aligned}$$

Proof. We first note that

$$\begin{aligned} \int_0^\infty \int |\square u| (|\partial u| + \langle x \rangle^{-1} |u|) dx dt \\ \leq \|(1+r)^{(p+1)/2} \square u\|_{L_t^2 L_x^2} (\|(1+r)^{-(p+1)/2} \partial u\|_{L_t^2 L_x^2} + \|(1+r)^{-(p+3)/2} u\|_{L_t^2 L_x^2}), \end{aligned}$$

and that

$$\begin{aligned} \|(1+r)^{-(p+1)/2} \partial u\|_{L_t^2 L_x^2} + \|(1+r)^{-(p+3)/2} u\|_{L_t^2 L_x^2} \\ \lesssim \sup_{j \geq 0} 2^{-j/2} \|\partial u\|_{L_t^2 L_x^2(\mathbb{R}_+ \times \{(x) \approx 2^j\})} + \sup_{j \geq 0} 2^{-3j/2} \|u\|_{L_t^2 L_x^2(\mathbb{R}_+ \times \{(x) \approx 2^j\})}. \end{aligned}$$

Thus, by bootstrapping this factor into the left side of (2-1), we see from (2-1) that

$$\|\partial u\|_{L_t^\infty L_x^2} \lesssim \|\partial u(0, \cdot)\|_{L^2} + \|(1+r)^{(p+1)/2} \square u\|_{L_t^2 L_x^2}.$$

Since $[\square, Z] = 0$ and since $[Z, \partial] \in \text{span}(\partial)$, the bound for the first term in (2-4) follows by replacing u by $Z^{\leq N} u$.

Since

$$\begin{aligned} [\partial_i, \partial_t + \partial_r] &= \frac{1}{r} \not\partial_i, & [\partial_i, \not\partial_j] &= \frac{1}{r} \left(-\delta_{ij} + \frac{x_i x_j}{r^2} \right) \partial_r - \frac{1}{r} \frac{x_j}{r} \not\partial_i, \\ [\Omega_{ij}, \not\partial_k] &= \delta_{jk} \not\partial_i - \delta_{ik} \not\partial_j, & [\Omega_{ij}, \partial_t + \partial_r] &= 0 \end{aligned}$$

and since $|\not\partial u| \leq (1/r)|\Omega u|$, we have that $|[Z, \not\partial]u| \leq (1/r)|Zu|$. Thus the remainder of the proof follows upon replacing u by $Z^{\leq N} u$ in (2-2). We may readily replace r by $1+r$ in the $L_t^2 L_x^2$ -terms since the powers in the left are negative, while powers in the right are positive. We also note that, due to a Hardy-type inequality,

$$\tilde{E}[u](t) \lesssim \|(1+r)^{p/2} \partial u(t, \cdot)\|_{L^2}^2. \quad \square$$

3. Proof of Theorem 1.1

The decay that we require will be obtained from the following weighted Sobolev estimate of [Klainerman 1986]. This estimate only provides decay in $|x|$, but simultaneously it does not necessitate the use of any time-dependent vector fields.

Lemma 3.1. *For $h \in C^\infty(\mathbb{R}^3)$ and $R \geq 1$,*

$$\|h\|_{L^\infty(\{R/2 < \langle x \rangle < R\})} \lesssim R^{-1} \|Z^{\leq 2} h\|_{L^2(\{R/4 < \langle x \rangle < 2R\})}. \quad (3-1)$$

The bound (3-1) follows, after localizing appropriately, from applying Sobolev estimates in the r - and ω -variables separately and comparing the volume element $dr d\sigma(\omega)$ with that of \mathbb{R}^3 in spherical coordinates: $r^2 dr d\sigma(\omega)$.

As mentioned earlier, the null condition (1-2) guarantees that at least one of the two factors in each nonlinear term is a ‘‘good’’ derivative. In fact, using a product rule argument, we have

$$|Z^{\leq 10} Q(\partial u)| \lesssim |Z^{\leq 5} \partial u| |Z^{\leq 10} \not\partial u| + |Z^{\leq 5} \not\partial u| |Z^{\leq 10} \partial u|. \quad (3-2)$$

This is well known; we refer the reader to, e.g., [Lindblad et al. 2013, Lemma 2.3].

We will use an iteration to solve (1-1). We let $u_{-1} \equiv 0$ and let u_k solve

$$\begin{cases} \square u_k = Q(\partial u_{k-1}), \\ u_k(0, \cdot) = f, \quad \partial_t u_k(0, \cdot) = g. \end{cases}$$

Boundedness: Our first step is to show an appropriate boundedness of this iteration. To this end, we shall set

$$M_k = \|Z^{\leq 10} \partial u_k\|_{L_t^\infty L_x^2} + \|(1+r)^{(p-1)/2} Z^{\leq 10} \not\partial u_k\|_{L_t^2 L_x^2} + \|(1+r)^{(p-3)/2} Z^{\leq 10} u_k\|_{L_t^2 L_x^2}.$$

Due to (2-4) and (1-3), there is a constant C_0 so that

$$M_0 \leq C_0 \varepsilon.$$

We shall argue inductively that for every k

$$M_k \leq 2C_0 \varepsilon. \quad (3-3)$$

To show (3-3), we use (2-4), which provides the bound

$$M_k \leq C_0 \varepsilon + C \|(1+r)^{(p+1)/2} Z^{\leq 10} \mathcal{Q}(\partial u_{k-1})\|_{L_t^2 L_x^2}.$$

Applying (3-2) and (3-1) we obtain

$$\begin{aligned} & \|(1+r)^{(p+1)/2} Z^{\leq 10} \mathcal{Q}(\partial u_{k-1})\|_{L_t^2 L_x^2} \\ & \lesssim \|(1+r)^{(p+1)/2} |Z^{\leq 5} \partial u_{k-1}| |Z^{\leq 10} \mathcal{P} u_{k-1}|\|_{L_t^2 L_x^2} \\ & \quad + \|(1+r)^{(p+1)/2} |Z^{\leq 5} \mathcal{P} u_{k-1}| |Z^{\leq 10} \partial u_{k-1}|\|_{L_t^2 L_x^2} \\ & \lesssim \|Z^{\leq 7} \partial u_{k-1}\|_{L_t^\infty L_x^2} \|(1+r)^{(p-1)/2} Z^{\leq 10} \mathcal{P} u_{k-1}\|_{L_t^2 L_x^2} \\ & \quad + \|(1+r)^{(p-1)/2} Z^{\leq 7} \mathcal{P} u_{k-1}\|_{L_t^2 L_x^2} \|Z^{\leq 10} \partial u_{k-1}\|_{L_t^\infty L_x^2}. \end{aligned}$$

Thus, using the inductive hypothesis, it follows that

$$M_k \leq C_0 \varepsilon + C(M_{k-1})^2 \leq C_0 \varepsilon + C \cdot C_0^2 \varepsilon^2.$$

And if $\varepsilon < 1/(C \cdot C_0)$, (3-3) results as desired.

Cauchy: We complete the proof by showing that the sequence is Cauchy in an appropriate norm. By completeness, the sequence must converge and by standard results the limiting function solves (1-1) as desired.

To this end, we set

$$\begin{aligned} A_k &= \|Z^{\leq 10} \partial(u_k - u_{k-1})\|_{L_t^\infty L_x^2} + \|(1+r)^{(p-1)/2} Z^{\leq 10} \mathcal{P}(u_k - u_{k-1})\|_{L_t^2 L_x^2} \\ & \quad + \|(1+r)^{(p-3)/2} Z^{\leq 10} (u_k - u_{k-1})\|_{L_t^2 L_x^2}. \end{aligned}$$

We note that

$$\begin{aligned} & Q^I(\partial u_{k-1}) - Q^I(\partial u_{k-2}) \\ &= A_{JK}^{\alpha\beta, I} \partial_\alpha (u_{k-1}^J - u_{k-2}^J) \partial_\beta u_{k-1}^K + A_{JK}^{\alpha\beta, I} \partial_\alpha u_{k-2}^J \partial_\beta (u_{k-1}^K - u_{k-2}^K). \end{aligned}$$

Thus, as in (3-2), we obtain

$$\begin{aligned} & |Z^{\leq 10} \mathcal{Q}(\partial u_{k-1}) - Z^{\leq 10} \mathcal{Q}(\partial u_{k-2})| \\ & \lesssim |Z^{\leq 5} \mathcal{P}(u_{k-1} - u_{k-2})| (|Z^{\leq 10} \partial u_{k-1}| + |Z^{\leq 10} \partial u_{k-2}|) \\ & \quad + (|Z^{\leq 5} \partial u_{k-1}| + |Z^{\leq 5} \partial u_{k-2}|) |Z^{\leq 10} \mathcal{P}(u_{k-1} - u_{k-2})| \\ & \quad + |Z^{\leq 5} \partial(u_{k-1} - u_{k-2})| (|Z^{\leq 10} \mathcal{P} u_{k-1}| + |Z^{\leq 10} \mathcal{P} u_{k-2}|) \\ & \quad + (|Z^{\leq 5} \mathcal{P} u_{k-1}| + |Z^{\leq 5} \mathcal{P} u_{k-2}|) |Z^{\leq 10} \partial(u_{k-1} - u_{k-2})|. \quad (3-4) \end{aligned}$$

As above, we apply (3-1) to the lower-order factor in each term to see that

$$\begin{aligned}
& \| (1+r)^{(p+1)/2} (Z^{\leq 10} Q(\partial u_{k-1}) - Z^{\leq 10} Q(\partial u_{k-2})) \|_{L_t^2 L_x^2} \\
& \lesssim \| (1+r)^{(p-1)/2} Z^{\leq 7} \not\partial(u_{k-1} - u_{k-2}) \|_{L_t^2 L_x^2} \\
& \quad \times \left(\| Z^{\leq 10} \partial u_{k-1} \|_{L_t^\infty L_x^2} + \| Z^{\leq 10} \partial u_{k-2} \|_{L_t^\infty L_x^2} \right) \\
& + \left(\| Z^{\leq 7} \partial u_{k-1} \|_{L_t^\infty L_x^2} + \| Z^{\leq 7} \partial u_{k-2} \|_{L_t^\infty L_x^2} \right) \\
& \quad \times \| (1+r)^{(p-1)/2} Z^{\leq 10} \not\partial(u_{k-1} - u_{k-2}) \|_{L_t^2 L_x^2} \\
& + \| Z^{\leq 7} \partial(u_{k-1} - u_{k-2}) \|_{L_t^\infty L_x^2} \\
& \quad \times \left(\| (1+r)^{(p-1)/2} Z^{\leq 10} \not\partial u_{k-1} \|_{L_t^2 L_x^2} + \| (1+r)^{(p-1)/2} Z^{\leq 10} \not\partial u_{k-2} \|_{L_t^2 L_x^2} \right) \\
& + \left(\| (1+r)^{(p-1)/2} Z^{\leq 7} \not\partial u_{k-1} \|_{L_t^2 L_x^2} + \| (1+r)^{(p-1)/2} Z^{\leq 7} \not\partial u_{k-2} \|_{L_t^2 L_x^2} \right) \\
& \quad \times \| Z^{\leq 10} \partial(u_{k-1} - u_{k-2}) \|_{L_t^\infty L_x^2}.
\end{aligned}$$

From (2-4) it then follows that

$$A_k \leq C(M_{k-1} + M_{k-2})A_{k-1} \leq C \cdot C_0 \varepsilon A_{k-1}.$$

So long as, say, $\varepsilon < 1/(2C \cdot C_0)$, we obtain

$$A_k \leq \frac{1}{2} A_{k-1} \quad \text{for all } k,$$

which implies that the sequence is Cauchy and completes the proof.

References

- [Alinhac 2001] S. Alinhac, “The null condition for quasilinear wave equations in two space dimensions, I”, *Invent. Math.* **145**:3 (2001), 597–618. [MR](#) [Zbl](#)
- [Christodoulou 1986] D. Christodoulou, “Global solutions of nonlinear hyperbolic equations for small initial data”, *Comm. Pure Appl. Math.* **39**:2 (1986), 267–282. [MR](#) [Zbl](#)
- [Dafermos and Rodnianski 2010] M. Dafermos and I. Rodnianski, “A new physical-space approach to decay for the wave equation with applications to black hole spacetimes”, pp. 421–432 in *XVIIth International Congress on Mathematical Physics*, edited by P. Exner, World Sci., Hackensack, NJ, 2010. [MR](#) [Zbl](#)
- [Katayama and Kubo 2008] S. Katayama and H. Kubo, “An alternative proof of global existence for nonlinear wave equations in an exterior domain”, *J. Math. Soc. Japan* **60**:4 (2008), 1135–1170. [MR](#) [Zbl](#)
- [Keel et al. 2002] M. Keel, H. F. Smith, and C. D. Sogge, “Almost global existence for some semilinear wave equations”, *J. Anal. Math.* **87** (2002), 265–279. [MR](#) [Zbl](#)
- [Keir 2018] J. Keir, “The weak null condition and global existence using the p-weighted energy method”, preprint, 2018. [arXiv 1808.09982](#)
- [Klainerman 1985] S. Klainerman, “Uniform decay estimates and the Lorentz invariance of the classical wave equation”, *Comm. Pure Appl. Math.* **38**:3 (1985), 321–332. [MR](#) [Zbl](#)
- [Klainerman 1986] S. Klainerman, “The null condition and global existence to nonlinear wave equations”, pp. 293–326 in *Nonlinear systems of partial differential equations in applied mathematics*,

Part 1 (Santa Fe, NM, 1984), edited by B. Nicolaenko et al., *Lectures in Appl. Math.* **23**, Amer. Math. Soc., Providence, RI, 1986. [MR](#) [Zbl](#)

- [Klainerman and Sideris 1996] S. Klainerman and T. C. Sideris, “On almost global existence for nonrelativistic wave equations in 3D”, *Comm. Pure Appl. Math.* **49**:3 (1996), 307–321. [MR](#) [Zbl](#)
- [Lindblad and Rodnianski 2005] H. Lindblad and I. Rodnianski, “Global existence for the Einstein vacuum equations in wave coordinates”, *Comm. Math. Phys.* **256**:1 (2005), 43–110. [MR](#) [Zbl](#)
- [Lindblad et al. 2013] H. Lindblad, M. Nakamura, and C. D. Sogge, “Remarks on global solutions for nonlinear wave equations under the standard null conditions”, *J. Differential Eq.* **254**:3 (2013), 1396–1436. [MR](#) [Zbl](#)
- [Luk 2013] J. Luk, “The null condition and global existence for nonlinear wave equations on slowly rotating Kerr spacetimes”, *J. Eur. Math. Soc. (JEMS)* **15**:5 (2013), 1629–1700. [MR](#) [Zbl](#)
- [Metcalf and Sogge 2006] J. Metcalf and C. D. Sogge, “Long-time existence of quasilinear wave equations exterior to star-shaped obstacles via energy methods”, *SIAM J. Math. Anal.* **38**:1 (2006), 188–209. [MR](#) [Zbl](#)
- [Metcalf and Sogge 2007] J. Metcalf and C. D. Sogge, “Global existence of null-form wave equations in exterior domains”, *Math. Z.* **256**:3 (2007), 521–549. [MR](#) [Zbl](#)
- [Metcalf et al. 2020] J. Metcalf, J. Sterbenz, and D. Tataru, “Local energy decay for scalar fields on time dependent non-trapping backgrounds”, *Amer. J. Math.* **142**:3 (2020), 821–883. [MR](#) [Zbl](#)
- [Morawetz 1968] C. S. Morawetz, “Time decay for the nonlinear Klein–Gordon equations”, *Proc. Roy. Soc. London Ser. A* **306** (1968), 291–296. [MR](#) [Zbl](#)
- [Moschidis 2016] G. Moschidis, “The r^p -weighted energy method of Dafermos and Rodnianski in general asymptotically flat spacetimes and applications”, *Ann. PDE* **2**:1 (2016), art. id. 6. [MR](#) [Zbl](#)
- [Sideris 2000] T. C. Sideris, “Nonresonance and global existence of prestressed nonlinear elastic waves”, *Ann. of Math. (2)* **151**:2 (2000), 849–874. [MR](#) [Zbl](#)
- [Sideris and Tu 2001] T. C. Sideris and S.-Y. Tu, “Global existence for systems of nonlinear wave equations in 3D with multiple speeds”, *SIAM J. Math. Anal.* **33**:2 (2001), 477–488. [MR](#) [Zbl](#)
- [Sogge 2008] C. D. Sogge, *Lectures on non-linear wave equations*, 2nd ed., International Press, Boston, 2008. [MR](#) [Zbl](#)
- [Sterbenz 2005] J. Sterbenz, “Angular regularity and Strichartz estimates for the wave equation”, *Int. Math. Res. Not.* **2005**:4 (2005), 187–231. [MR](#) [Zbl](#)
- [Yang 2015a] S. Yang, “Global solutions of nonlinear wave equations with large data”, *Selecta Math. (N.S.)* **21**:4 (2015), 1405–1427. [MR](#) [Zbl](#)
- [Yang 2015b] S. Yang, “Global stability of solutions to nonlinear wave equations”, *Selecta Math. (N.S.)* **21**:3 (2015), 833–881. [MR](#) [Zbl](#)

Received: 2021-01-27

Accepted: 2023-01-07

mfacci@live.unc.edu

*Department of Mathematics, University of North Carolina,
Chapel Hill, NC, United States*

mcentarffer@unc.edu

*Department of Mathematics, University of North Carolina,
Chapel Hill, NC, United States*

metcalfe@email.unc.edu

*Department of Mathematics, University of North Carolina,
Chapel Hill, NC, United States*

Cones and ping-pong in three dimensions

Gabriel Frieden, Félix Gélinas and Étienne Soucy

(Communicated by Jim Haglund)

We study the hypergeometric group in $GL_3(\mathbb{C})$ with parameters $\alpha = (\frac{1}{4}, \frac{1}{2}, \frac{3}{4})$ and $\beta = (0, 0, 0)$. We give a new proof that this group is isomorphic to the free product $\mathbb{Z}/4\mathbb{Z} * \mathbb{Z}/2\mathbb{Z}$ by exhibiting a ping-pong table. Our table is determined by a simplicial cone in \mathbb{R}^3 , and we prove that this is the unique simplicial cone (up to sign) for which our construction produces a valid ping-pong table.

1. Introduction

Beukers and Heckman [1989] defined a *hypergeometric group* to be a subgroup of $GL_n(\mathbb{C})$ generated by three matrices R, T, U such that $U^{-1}TR = I$, U and R have no shared eigenvalues, and $T - I$ is a rank-1 matrix. The name is due to the fact that these groups arise as monodromy groups of hypergeometric differential equations.

One of the main results of [Beukers and Heckman 1989] says that the Zariski closure of a primitive hypergeometric group is either a finite subgroup of $GL_n(\mathbb{C})$ or one of the matrix groups $SL_n(\mathbb{C})$, $SO_n(\mathbb{C})$, $Sp_n(\mathbb{C})$. If H is a subgroup of $GL_n(\mathbb{Z})$ whose Zariski closure is $G(\mathbb{C})$ (where G is a matrix group GL_n , SL_n , SO_n , Sp_n , etc.), H is said to be *arithmetic* if it has finite index in $G(\mathbb{Z})$, and *thin* otherwise. Arithmetic subgroups of $GL_n(\mathbb{Z})$ have long been a central object of study in number theory, but in recent years there has been increasing interest in thin subgroups; see [Sarnak 2014].

The question of whether a given primitive hypergeometric group is arithmetic or thin has been studied in [Chen, Yang and Yui 2008; Venkataramana 2014; Singh and Venkataramana 2014; Fuchs, Meiri and Sarnak 2014; Brav and Thomas 2014; Filip and Fougeron 2021] and is rather subtle. Fuchs, Meiri, and Sarnak [2014] showed that several infinite families of hypergeometric groups with closure $SO_n(\mathbb{C})$ (n odd) are thin. On the other hand, for hypergeometric groups with closure $Sp_n(\mathbb{C})$, one infinite family is known to be arithmetic [Venkataramana 2014], but the only known thin examples are in $Sp_4(\mathbb{C})$ [Singh and Venkataramana 2014; Brav and Thomas 2014; Filip and Fougeron 2021].

MSC2020: 20E06.

Keywords: hypergeometric group, free product of groups, ping-pong lemma.

In this paper, we are interested in a particular infinite family of primitive hypergeometric groups. For $n \geq 2$, define $R_n, U_n, T_n \in \mathrm{GL}_n(\mathbb{C})$ by

$$R_n = \begin{pmatrix} 0 & 0 & 0 & & 0 & 0 & -1 \\ 1 & 0 & 0 & \cdots & 0 & 0 & -1 \\ 0 & 1 & 0 & & 0 & 0 & -1 \\ \vdots & & \ddots & & \vdots & & \\ 0 & 0 & 0 & & 0 & 0 & -1 \\ 0 & 0 & 0 & \cdots & 1 & 0 & -1 \\ 0 & 0 & 0 & & 0 & 1 & -1 \end{pmatrix}, \quad U_n = \begin{pmatrix} 0 & 0 & 0 & & 0 & 0 & \pm 1 \\ 1 & 0 & 0 & \cdots & 0 & 0 & \mp n \\ 0 & 1 & 0 & & 0 & 0 & \pm \binom{n}{n-2} \\ \vdots & & \ddots & & \vdots & & \\ 0 & 0 & 0 & & 0 & 0 & \binom{n}{3} \\ 0 & 0 & 0 & \cdots & 1 & 0 & -\binom{n}{2} \\ 0 & 0 & 0 & & 0 & 1 & n \end{pmatrix},$$

and $T_n = U_n R_n^{-1}$, where the signs in the last column of U_n alternate. Let H_n be the hypergeometric group generated by R_n, U_n, T_n . The *parameters* of H_n are

$$\alpha = \left(\frac{1}{n+1}, \frac{2}{n+1}, \dots, \frac{n}{n+1} \right), \quad \beta = (0, \dots, 0).$$

This means that the eigenvalues of R_n and U_n are $e^{2\pi i/(n+1)}, \dots, e^{2\pi i n/(n+1)}$ and $1, \dots, 1$, respectively. It follows from the criterion in [Beukers and Heckman 1989] that H_n has Zariski closure $\mathrm{Sp}_n(\mathbb{C})$ if n is even and $\mathrm{SO}_n(\mathbb{C})$ if n is odd.¹ The group H_n arises in algebraic geometry as the monodromy group of a well-studied family of degree- n hypersurfaces in \mathbb{P}^{n-1} known as the *Dwork family*; see, e.g., [Katz 2009].

The group H_n is known to be arithmetic when $n = 2, 3$ (see [Fuchs, Meiri and Sarnak 2014]) and was shown in [Brav and Thomas 2014] to be thin when $n = 4$. According to [Sarnak 2014], it “seems likely” that H_n is thin for all even $n \geq 4$. If this is true, it would provide the first examples of thin subgroups of $\mathrm{Sp}_n(\mathbb{C})$ for $n \geq 6$.

To show that H_4 is thin, Brav and Thomas used the ping-pong lemma to prove that H_4 is isomorphic to the free product $\mathbb{Z}/5\mathbb{Z} * \mathbb{Z}$. The following conjecture generalizes this result and would imply that H_n is thin for $n \geq 4$.

Conjecture 1.1. *If $n \geq 2$, then*

$$H_n = \langle R_n, T_n \rangle = \langle R_n \rangle * \langle T_n \rangle = \begin{cases} \mathbb{Z}/(n+1)\mathbb{Z} * \mathbb{Z} & \text{if } n \text{ is even,} \\ \mathbb{Z}/(n+1)\mathbb{Z} * \mathbb{Z}/2\mathbb{Z} & \text{if } n \text{ is odd.} \end{cases}$$

This paper undertakes a detailed study of [Conjecture 1.1](#) in the case $n = 3$. In [Section 2](#), we use the ping-pong lemma to give an elementary proof that $H_3 \cong \mathbb{Z}/4\mathbb{Z} * \mathbb{Z}/2\mathbb{Z}$. To apply the ping-pong lemma, one must define a “ping-pong table” in a set on which the group acts. In our case, we consider the natural action of

¹Note, for odd n , the group $H_n = H_{2k+1}$ preserves a symmetric bilinear form of signature $(k+1, k)$. By contrast, [Fuchs, Meiri and Sarnak 2014] studies the case of Lorentzian signature $(2k, 1)$.

3×3 matrices on \mathbb{R}^3 , and our ping-pong table is determined by a simplicial cone C in \mathbb{R}^3 . We prove in [Section 3](#) that C is (up to sign) the *only* simplicial cone which gives rise to a “valid” ping-pong table via our construction.

In [Section 4](#), we use a two-dimensional projection to illustrate the main ideas of the previous sections. Finally, in [Section 5](#), we compare our ping-pong table in the three-dimensional case with the (essentially unique) ping-pong table in the two-dimensional case, and with the more complicated ping-pong table of Brav and Thomas in the four-dimensional case. We hope that the juxtaposition of these three examples will inspire future work on [Conjecture 1.1](#) in higher dimensions.

Remark 1.2. The $n = 2$ and $n = 3$ cases of [Conjecture 1.1](#) can be obtained from classical results of [[Schwarz 1873](#); [Klein 1933](#); [Clausen 1828](#)]. Indeed, Schwarz and Klein determined the structure of a large class of hypergeometric groups in GL_2 (the so-called *Schwarz triangle groups*), one of which is H_2 . A result of Clausen implies that H_3 is the monodromy group of the symmetric square of one of the hypergeometric differential equations covered by the work of Schwarz and Klein (namely, the equation with parameters $\alpha = (\frac{1}{8}, \frac{3}{8})$ and $\beta = (0, 0)$). It follows that H_3 is isomorphic to the Schwarz triangle group corresponding to these parameters. We refer the reader to [[Heckman 2015](#), §2.2, 3.2] for a nice account of this story.

2. A three-dimensional ping-pong table

2A. Cones. Given vectors $v_1, \dots, v_k \in \mathbb{R}^n$, define the *open cone* C generated by v_1, \dots, v_k to be the set of strictly positive linear combinations of the v_i . That is,

$$C = \{a_1 v_1 + \dots + a_k v_k \mid a_i \in \mathbb{R}_{>0}\}.$$

We will sometimes write $C = \text{cone}(v_1, \dots, v_k)$. Note that C is unchanged if one of the generators v_i is replaced by a positive scalar multiple λv_i , $\lambda > 0$. The cone C is said to be *simplicial* if the generators v_1, \dots, v_k are linearly independent.

For a subset $S \subseteq \mathbb{R}^n$, we write \bar{S} for the closure of S (in the Euclidean topology). If $C = \text{cone}(v_1, \dots, v_k)$, then

$$\bar{C} = \{a_1 v_1 + \dots + a_k v_k \mid a_i \in \mathbb{R}_{\geq 0}\}.$$

We call \bar{C} the *closed cone* generated by v_1, \dots, v_k .

A subset $S \subseteq \mathbb{R}^n$ is *convex* if, for any two points $x, y \in S$, the line segment

$$\{\lambda x + (1 - \lambda)y \mid \lambda \in [0, 1]\}$$

connecting x and y is contained in S . It is easy to verify that cones (both open and closed) are convex.

2B. Free products and the ping-pong lemma. Let G_1, \dots, G_d be subgroups of a group G . A *word (in the elements of the G_j)* is a finite sequence (x_1, \dots, x_n) ,

such that each x_i is an element of at least one of the G_j . Each word gives rise to an element of G by multiplication; that is, (x_1, \dots, x_n) gives rise to the element $g = x_1 \cdots x_n \in G$. In this case, we say that (x_1, \dots, x_n) is an *expression* for g , or that g can be *expressed* as the word (x_1, \dots, x_n) . The *group generated by the subgroups* G_j , denoted by $\langle G_1, \dots, G_d \rangle$, is the subgroup of G consisting of all elements that can be expressed as words in the elements of the G_j . The subgroups G_j are said to *generate* G if $G = \langle G_1, \dots, G_d \rangle$.

If G_1, \dots, G_d generate G , there are in general many expressions for each element of G as a word in the elements of the G_j . We say that a word (x_1, \dots, x_n) is *reduced* if none of the x_i is the identity element, and, for $i = 1, \dots, n - 1$, the elements x_i and x_{i+1} are not both contained in a single G_j . The idea is that identity elements can be removed from a word without changing the resulting element of G , and if $x_i, x_{i+1} \in G_j$, these two elements can be replaced by the single element $x_i x_{i+1} \in G_j$. By convention, the empty word gives rise to the identity element of G , and is considered to be reduced.

Definition 2.1. Let G_1, \dots, G_d be subgroups of a group G . The group G is the *free product* of the G_j if each $g \in G$ has a unique reduced expression in the elements of the G_j . In this case, one writes

$$G = G_1 * \cdots * G_d.$$

We encourage the reader to verify that if G is the free product of G_1, \dots, G_d , then $G_i \cap G_j = \{1\}$ for $i \neq j$. It may also be instructive to find a counterexample to the converse of this statement.

The following result, which is known as the *ping-pong lemma*, is a standard tool for proving that two subgroups of a larger group generate a free product.

Lemma 2.2 [Lyndon and Schupp 1977]. *Let G, H be two nontrivial subgroups of a group K such that at least one of G and H has more than two elements. Suppose K acts on a set S and there are two nonempty subsets $X, Y \subset S$ satisfying the following properties:*

- (1) $X \cap Y = \emptyset$.
- (2) If $g \in G \setminus \{1\}$ and $x \in X$, then $gx \in Y$.
- (3) If $h \in H \setminus \{1\}$ and $y \in Y$, then $hy \in X$.

*Then the subgroup of K generated by G and H is a free product; that is, $\langle G, H \rangle = G * H$.*

We will refer to the sets X and Y as a *valid ping-pong table* (for G and H) if they satisfy the hypotheses of the ping-pong lemma.

2C. A ping-pong table in \mathbb{R}^3 . We now consider the three-dimensional case of [Conjecture 1.1](#). Writing $R = R_3$, $U = U_3$, and $T = T_3$, we have

$$R = \begin{pmatrix} 0 & 0 & -1 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}, \quad U = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & -3 \\ 0 & 1 & 3 \end{pmatrix}, \quad T = \begin{pmatrix} -1 & 0 & 0 \\ 2 & 1 & 0 \\ -4 & 0 & 1 \end{pmatrix}.$$

Note that $R^4 = T^2 = I$.

Theorem 2.3. *The subgroup of $GL_3(\mathbb{R})$ generated by R and T is the free product of $\langle R \rangle$ and $\langle T \rangle$; that is,*

$$\langle R, T \rangle = \mathbb{Z}/4\mathbb{Z} * \mathbb{Z}/2\mathbb{Z}.$$

Proof. The group $GL_3(\mathbb{R})$ acts on \mathbb{R}^3 by matrix multiplication. We will find disjoint subsets $X, Y \subset \mathbb{R}^3$ such that all elements of X are sent to Y by R, R^2 , and R^3 , and all elements of Y are sent to X by T , which will allow us to conclude that $\langle R, T \rangle \cong \langle R \rangle * \langle T \rangle$ by the ping-pong lemma.

Let C be the open cone generated by the vectors

$$u = (1, -2, 1), \quad v = (1, 0, 3), \quad w = (0, -1, 1),$$

that is, $C = \{au + bv + cw \mid a, b, c \in \mathbb{R}_{>0}\}$. Define

$$X = C \cup -C, \quad Y = RX \cup R^2X \cup R^3X.$$

It is immediately clear from this definition that each nonidentity element of $\langle R \rangle$ maps X into Y , so hypothesis (2) of the ping-pong lemma is satisfied.

Now suppose there is a point $p = au + bv + cw \in X \cap Y$. Since $p \in X$, the coefficients a, b, c are all nonzero and of the same sign. Since $p \in Y$, there exists a point $q = xu + yv + zw \in X$ (so again x, y, z are nonzero and of the same sign) such that R, R^2 or R^3 maps q to p . Explicitly, we have

$$\begin{aligned} q &= (x+y, -2x-z, x+3y+z), \\ Rq &= (-x-3y-z, -2y-z, -3x-3y-2z), \\ R^2q &= (3x+3y+2z, 2x+z, 3x+y+z), \\ R^3q &= (-3x-y-z, 2y+z, -x-y), \\ p &= (a+b, -2a-c, a+3b+c). \end{aligned}$$

This gives us three systems $p = R^i q$ which solve to

$$\begin{aligned} p = Rq &\implies a = -y, & b = -x - 2y - z, & c = 4y + z, \\ p = R^2q &\implies a = x + 2y + z, & b = 2x + y + z, & c = -4x - 4y - 3z, \\ p = R^3q &\implies a = -2x - y - z, & b = -x, & c = 4x + z. \end{aligned}$$

Remembering that the triples (x, y, z) and (a, b, c) must be nonzero and either all positive or all negative, we obtain a contradiction in each case:

- In the first case, if x, y, z are positive, then $a = -y$ is negative, but $c = 4y + z$ is positive, and vice versa in the negative case.
- In the second case, again $a = x + 2y + z$ and $c = -4x - 4y - 3z$ cannot have the same sign if x, y, z have the same sign.
- The same goes in the third case for $a = -2x - y - z$ and $c = 4x + z$.

These contradictions prove that X and Y are indeed disjoint.

We will now verify that T sends Y into X using a similar argument. As before, let $q = xu + yv + zw$ be a point in X . If we apply T to Rq, R^2q, R^3q , we get

$$\begin{aligned} TRq &= (x + 3y + z, -2x - 8y - 3z, x + 9y + 2z), \\ TR^2q &= (-3x - 3y - 2z, 8x + 6y + 5z, -9x - 11y - 7z), \\ TR^3q &= (3x + y + z, -6x - z, 11x + 3y + 4z). \end{aligned}$$

This time solving the systems $p = TR^i q$ (where $p = au + bv + cw$) nets us

$$\begin{aligned} p = TRq &\implies a = x + 2y + z, & b = y, & c = 4y + z, \\ p = TR^2q &\implies a = -2x - y - z, & b = -x - 2y - z, & c = -4x - 4y - 3z, \\ p = TR^3q &\implies a = x, & b = 2x + y + z, & c = 4x + z. \end{aligned}$$

In this case we see that the signs of a, b, c all properly match, which confirms that T does send Y into X , completing the proof. \square

2D. Matrix logarithms. At this point, the reader may be wondering how we arrived at the definition of the cone C . The explanation requires an examination of the linear maps TR and TR^{-1} , and their logarithms. In addition to motivating the choice of generators u, v , and w , the formulas derived below play an essential role in the proof of the uniqueness of C in the next section.

The matrix $U = TR$ has Jordan form

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

This means that 1 is the only eigenvalue of TR , and the corresponding eigenspace is one-dimensional. The vector $u = (1, -2, 1)$ spans this eigenspace. The matrix

$$R(TR)^{-1}R^{-1} = TR^{-1}$$

has the same Jordan form as TR , and its one-dimensional eigenspace is spanned by $v = (1, 0, 3)$.

By the hypotheses of the ping-pong lemma, any positive integer power of the linear transformations TR and TR^{-1} must map X to itself. To understand the powers of these matrices, we use the Taylor expansions of \log and \exp , which allow us to define

$$(TR)^t = \exp(t \log(TR)) \quad \text{and} \quad (TR^{-1})^t = \exp(t \log(TR^{-1}))$$

for all $t \in \mathbb{R}$. For TR , we compute

$$\begin{aligned} \log(TR) &= (TR - I) - \frac{1}{2}(TR - I)^2 + \frac{1}{3}(TR - I)^3 - \dots \\ &= \begin{pmatrix} -1 & 0 & 1 \\ 1 & -1 & -3 \\ 0 & 1 & 2 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 1 & 1 & 1 \\ -2 & -2 & -2 \\ 1 & 1 & 1 \end{pmatrix} + 0 - \dots = \begin{pmatrix} -\frac{3}{2} & -\frac{1}{2} & \frac{1}{2} \\ 2 & 0 & -2 \\ -\frac{1}{2} & \frac{1}{2} & \frac{3}{2} \end{pmatrix}, \end{aligned}$$

and then

$$\begin{aligned} (TR)^t &= \exp(t \log(TR)) = I + t \log(TR) + \frac{t^2}{2!} \log(TR)^2 + \dots \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + t \begin{pmatrix} -\frac{3}{2} & -\frac{1}{2} & \frac{1}{2} \\ 2 & 0 & -2 \\ -\frac{1}{2} & \frac{1}{2} & \frac{3}{2} \end{pmatrix} + t^2 \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -1 & -1 & -1 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{pmatrix}. \end{aligned} \tag{2-1}$$

Similarly, we compute

$$(TR^{-1})^t = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + t \begin{pmatrix} -\frac{3}{2} & -\frac{1}{2} & \frac{1}{2} \\ -3 & 1 & 1 \\ -\frac{3}{2} & -\frac{5}{2} & \frac{1}{2} \end{pmatrix} + t^2 \begin{pmatrix} \frac{3}{2} & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 0 & 0 \\ \frac{9}{2} & -\frac{3}{2} & -\frac{3}{2} \end{pmatrix}. \tag{2-2}$$

Let $P = \log(TR)$ and $Q = \log(TR^{-1})$ (these are the coefficients of t in (2-1) and (2-2), respectively). As the reader may easily verify, both P and Q have rank 2, and their column spaces intersect in the line spanned by $w = (0, -1, 1)$. It is perhaps not clear why this intersection should be useful in defining a ping-pong table. In Section 4B, we consider a two-dimensional projection that clearly illustrates the significance of this intersection.

3. Uniqueness of the cone C

Let C' be the open cone generated by three linearly independent vectors $u', v', w' \in \mathbb{R}^3$, and define

$$X = C' \cup -C', \quad Y = RX \cup R^2X \cup R^3X.$$

The goal of this section is to prove the following uniqueness theorem.

Theorem 3.1. *If X and Y form a valid ping-pong table for $\langle R \rangle$ and $\langle T \rangle$, then $C' = C$ or $C' = -C$, where $C = \text{cone}(u, v, w)$ is the cone defined in the previous section.*

The proof consists of two steps, the first of which is carried out in the following:

Lemma 3.2. *Suppose X and Y form a valid ping-pong table for $\langle R \rangle$ and $\langle T \rangle$.*

(a) *Let $M = (TR^j)^t$ for fixed $j \in \{1, 2, 3\}$ and $t \in \mathbb{Z}_{>0}$. Either*

$$M(\bar{C}') \subseteq \bar{C}' \quad \text{or} \quad M(\bar{C}') \subseteq -\bar{C}'.$$

(b) *The lines spanned by u and v are contained in \bar{X} .*

(c) *Two of the generators of C' are u and v (or $-u$ and $-v$).*

Proof. The hypotheses of the ping-pong lemma imply that M maps X into X . Since linear transformations are continuous, this implies that M maps \bar{X} into \bar{X} . Suppose there are nonzero vectors $q_1, q_2 \in \bar{C}'$ such that $M(q_1) \in \bar{C}'$ and $M(q_2) \in -\bar{C}'$. Linear transformations map line segments to line segments, so the convexity of \bar{C}' implies that the line segment from $M(q_1)$ to $M(q_2)$ is contained in $\bar{X} = \bar{C}' \cup -\bar{C}'$. This can only happen if the line segment connecting $M(q_1)$ and $M(q_2)$ passes through the origin, that is, if $M(q_1) = -\lambda M(q_2)$ for some $\lambda > 0$. Since M is invertible, this would imply that $q_1 = -\lambda q_2$, so $q_1, q_2 \in \bar{C}' \cap -\bar{C}' = \{0\}$, a contradiction. This proves (a).

To prove part (b), we will show that for any nonzero vector $q = (x, y, z) \in \mathbb{R}^3$, the vectors $(TR)^t(q)$ approach the line generated by u as t approaches infinity, and the vectors $(TR^{-1})^t(q)$ approach the line generated by v . By (2-1), we have

$$(TR)^t(q) = \begin{pmatrix} \frac{1}{2}(x+y+z)t^2 + \frac{1}{2}(-3x-y+z)t + x \\ -(x+y+z)t^2 + 2(x-z)t + y \\ \frac{1}{2}(x+y+z)t^2 + \frac{1}{2}(-x+y+3z)t + z \end{pmatrix}. \quad (3-1)$$

For a nonzero vector a , let \hat{a} denote the normalization of a (i.e., the vector a divided by its Euclidean norm). Using the fact that $\lim_{t \rightarrow \infty} \widehat{(TR)^t(q)}$ depends only on the coefficients of the highest power of t appearing in $(TR)^t(q)$, we find that

$$\lim_{t \rightarrow \infty} \widehat{(TR)^t(q)} = \begin{cases} \frac{x+y+z}{|x+y+z|} \hat{u} & \text{if } x+y+z \neq 0, \\ \frac{z-x}{|z-x|} \hat{u} & \text{if } x+y+z = 0 \text{ and } x \neq z, \\ \frac{x}{|x|} \hat{u} & \text{if } x+y+z = 0 \text{ and } x = z. \end{cases} \quad (3-2)$$

In all cases, the normalization of $(TR)^t(q)$ approaches $\pm \hat{u}$, one of the two unit eigenvectors of TR . Similarly, using (2-2), we find that

$$\lim_{t \rightarrow \infty} \widehat{(TR^{-1})^t(q)} = \begin{cases} \frac{3x-y-z}{|3x-y-z|} \hat{v} & \text{if } 3x-y-z \neq 0, \\ \frac{-y}{|y|} \hat{v} & \text{if } 3x-y-z = 0 \text{ and } y \neq 0, \\ \frac{x}{|x|} \hat{v} & \text{if } 3x-y-z = 0 \text{ and } y = 0, \end{cases} \quad (3-3)$$

so in all cases the normalization of $(TR^{-1})^t(q)$ approaches $\pm\hat{v}$, one of the two unit eigenvectors of TR^{-1} .

If $q \in X$, then as observed in the proof of part (a), $(TR)^t(q)$ and $(TR^{-1})^t(q)$ must be in X for any positive integer t . Thus, since X is closed under scalar multiplication (and nonempty), the previous calculations tell us that each point on the lines spanned by u and v is a limit point of a sequence of points in X , so these lines are in the closure of X . This proves (b).

It remains to prove (c). By part (b), we may assume that u is contained in \bar{C}' (possibly after replacing C' with $-C'$). Suppose that u is not a generator of C' . This means that u is contained in the interior of \bar{C}' , or in the interior of a face of \bar{C}' . In either case, we can find a vector $q = (x, y, z)$ which is not a scalar multiple of u such that the line segment

$$\{u + \lambda q \mid |\lambda| \leq \epsilon\}$$

is contained in \bar{C}' for sufficiently small $\epsilon > 0$. All points (a, b, c) which satisfy both $a + b + c = 0$ and $c - a = 0$ are on the line spanned by $u = (1, -2, 1)$, so we must have $x + y + z \neq 0$ or $z - x \neq 0$. We may assume that $x + y + z > 0$, or that $x + y + z = 0$ and $z - x > 0$. By (3-2), the sequence $(TR)^t(u + \lambda q)$ approaches the ray generated by u if $\lambda \geq 0$ and the ray generated by $-u$ if $\lambda < 0$. This contradicts part (a).

A similar argument using (3-3) shows that v must be a generator of C' or $-C'$. To see that v must in fact be a generator of C' , note that $(TR)^t(u) = u$ for all t , and $(TR)^t(v) = (TR)^t(1, 0, 3)$ approaches the ray generated by u by (3-2). Now part (a) guarantees that $v \notin -\bar{C}'$. \square

Remark 3.3. The proof of part (b) works for any valid ping-pong table in which X is closed under scalar multiplication.

Proof of Theorem 3.1. By Lemma 3.2(c), we may assume (possibly after replacing C' with $-C'$) that two of the generators of C' are u and v . Suppose $C' = \text{cone}(u, v, w')$, where

$$w' = \lambda u + \mu v + \eta w = \begin{pmatrix} \lambda + \mu \\ -2\lambda - \eta \\ \lambda + 3\mu + \eta \end{pmatrix}$$

for some $\lambda, \mu, \eta \in \mathbb{R}$. Since u, v, w' are assumed to be linearly independent, we must have $\eta \neq 0$. We first show that $\eta > 0$.

Applying (3-1) to $v = (1, 0, 3)$, we obtain

$$(TR)^t(v) = \begin{pmatrix} 2t^2 + 1 \\ -4t^2 - 4t \\ 2t^2 + 4t + 3 \end{pmatrix}.$$

Solving a system of linear equations, we find that $(TR)^t(v) = au + bv + cw'$, where

$$a = 2t^2 - \frac{4\lambda}{\eta}t, \quad b = 1 - \frac{4\mu}{\eta}t, \quad c = \frac{4}{\eta}t.$$

Since $v \in \bar{X}$, the hypotheses of the ping-pong lemma require that $(TR)^t(v)$ be in \bar{X} for all $t \in \mathbb{Z}_{>0}$. This means that, for such t , we must have $a, b, c \geq 0$ or $a, b, c \leq 0$. For large t , we can see that a is positive and c has the same sign as η . This shows that η must be positive, as claimed.

Scaling w' by a positive constant does not change C' , so we may assume that $w' = \lambda u + \mu v + w$. We will now show that $TR(X) \not\subseteq X$ if $\mu \neq 0$ and $TR^{-1}(X) \not\subseteq X$ if $\lambda \neq 0$. Suppose $x, y, z > 0$, so that $q = xu + yv + zw'$ is in C' . Solving a system of linear equations, we find that $TR(q) = au + bv + cw'$, where

$$\begin{aligned} a &= x + (2 - 4\lambda)y + (1 + 2\mu - 4\lambda\mu)z, \\ b &= (1 - 4\mu)y - 4\mu^2z, \\ c &= 4y + (1 + 4\mu)z. \end{aligned}$$

The crucial feature of these formulas is the presence of μ^2 in the equation for b . This means that if $\mu \neq 0$, then by choosing z sufficiently large, we can make b negative. But for any fixed choice of z , we can make a positive by choosing x sufficiently large. This shows that there is a choice of $x, y, z > 0$ such that a and b do not have the same sign, contradicting the assumption that TR maps X to itself. We conclude that $\mu = 0$.

Next, we compute $TR^{-1}(q) = a'u + b'v + c'w'$, where

$$\begin{aligned} a' &= (1 - 4\lambda)x - 4\lambda^2z, \\ b' &= y + (2 - 4\mu)x + (1 + 2\lambda - 4\lambda\mu)z, \\ c' &= 4x + (1 + 4\lambda)z. \end{aligned}$$

If $\lambda \neq 0$, we can make a' negative by choosing z sufficiently large, and then we can make b' positive by choosing y sufficiently large. This contradicts the assumption that TR^{-1} maps X to itself, so we must have $\lambda = 0$. We conclude that w' is a positive scalar multiple of w . \square

4. Two-dimensional projection

4A. Definition of the projection. In order to better understand the algebraic arguments in the previous sections, it is useful to project from \mathbb{R}^3 to a plane, where we can more easily visualize what is going on. Given a linear functional $\phi: \mathbb{R}^3 \rightarrow \mathbb{R}$, we can send a vector $s \in \mathbb{R}^3$ to $s/\phi(s)$, provided $\phi(s) \neq 0$. Since ϕ is linear, $\phi(s/\phi(s)) = \phi(s)/\phi(s) = 1$. Thus, the map $\rho: s \mapsto s/\phi(s)$ amounts to projecting s onto the plane $P = \{s \mid \phi(s) = 1\}$.

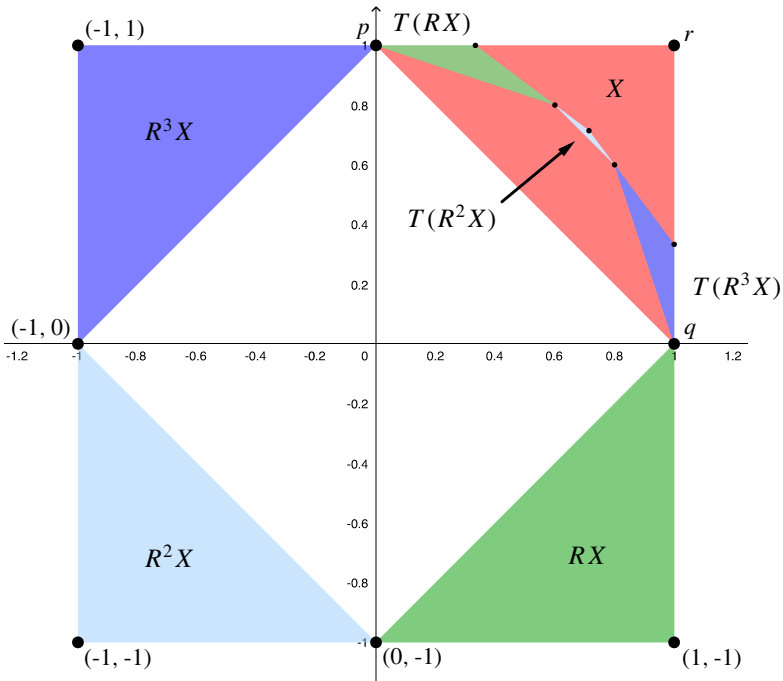


Figure 1. The large triangle in the first quadrant (colored red) is the projection of $X = \pm C$. The three large triangles in the other quadrants are the projections of RX , R^2X , and R^3X . The smaller triangles in the first quadrant are the projections of $T(RX)$, $T(R^2X)$, and $T(R^3X)$.

We will use the projection ρ determined by the linear functional

$$\phi(x, y, z) = x - y + z.$$

This choice of ϕ satisfies $\phi(u), \phi(v), \phi(w) > 0$, so the cone generated by u, v, w projects to a triangle in the plane P . We need to choose a system of coordinates on P . The fundamental theorem of affine geometry tells us that for any three points $p, q, r \in \mathbb{R}^2$ which are not collinear, there is a unique affine transformation from P to \mathbb{R}^2 sending $\rho(u), \rho(v), \rho(w)$ to p, q, r . For simplicity, we choose

$$p = (0, 1), \quad q = (1, 0), \quad r = (1, 1),$$

which leads to the projection map

$$\rho(x, y, z) = \left(\frac{-2(x - z)}{x - y + z}, \frac{-2y}{x - y + z} \right). \tag{4-1}$$

Applying ρ to X and Y , we obtain **Figure 1**, which illustrates the fact that X and Y define a valid ping-pong table.

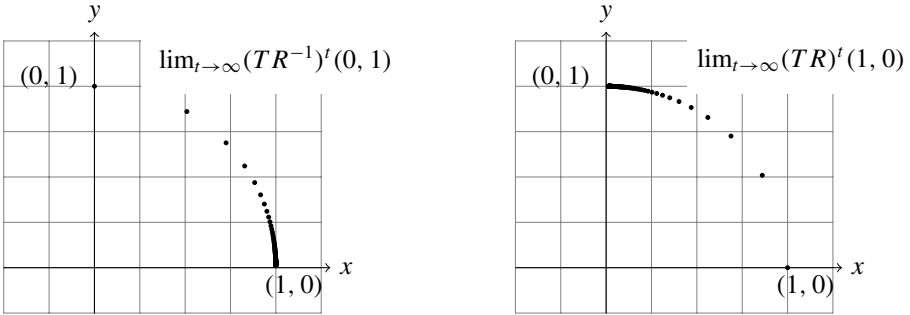


Figure 2. The points $(TR^{-1})^t(u)$ (left) and $(TR)^t(v)$ (right) in the two-dimensional projection.

We now express the maps R and T in terms of the coordinates (a, b) on \mathbb{R}^2 . The point (a, b) is the image of a line in \mathbb{R}^3 , and a straightforward computation shows that the line which maps to (a, b) is spanned by the vector (x, y, z) , where

$$x = -\frac{a}{4} - \frac{b}{4} + \frac{1}{2}, \quad y = -\frac{b}{2}, \quad z = \frac{a}{4} - \frac{b}{4} + \frac{1}{2}. \quad (4-2)$$

If we apply R and T to (x, y, z) and then apply ρ , we obtain the following formulas for the actions of R and T on \mathbb{R}^2 :

$$R(a, b) = (b, -a), \quad (4-3)$$

$$T(a, b) = \left(\frac{2a + b - 2}{2a + 2b - 3}, \frac{a + 2b - 2}{2a + 2b - 3} \right). \quad (4-4)$$

In particular, R is rotation by 90 degrees (clockwise). This explains why the projection of Y consists of the rotations of X (the red triangle) in [Figure 1](#).

Remark 4.1. The map T in (4-4) also has a geometric interpretation. In the Klein (unit disk) model of the hyperbolic plane, T is the hyperbolic rotation of angle π about the point $(\frac{1}{2}, \frac{1}{2})$. We thank Jean-Philippe Burelle for explaining this to us.

4B. Uniqueness revisited. Using the projection ρ , we can give a more visual explanation of the uniqueness of the cone C . The following argument is similar in spirit to the proof of uniqueness given in [Section 3](#), although it does not exactly correspond to the steps of that proof.

By construction, the eigenvectors u and v project to the points $p = (0, 1)$ and $q = (1, 0)$. The sequences of points $(TR)^t(1, 0)$ and $(TR^{-1})^t(0, 1)$ (for $t \in \mathbb{Z}_{>0}$) are shown in [Figure 2](#). These figures suggest that the first sequence approaches $(0, 1)$ along the unit circle, and the second sequence approaches $(1, 0)$ along the unit circle; this is verified in [Section 4C](#). The tangent line to the curve $(TR)^t(1, 0)$ becomes horizontal as $t \rightarrow \infty$, and the tangent line to $(TR^{-1})^t(0, 1)$ becomes vertical as $t \rightarrow \infty$. If the triangle formed by $(0, 1)$, $(1, 0)$, and a third point s

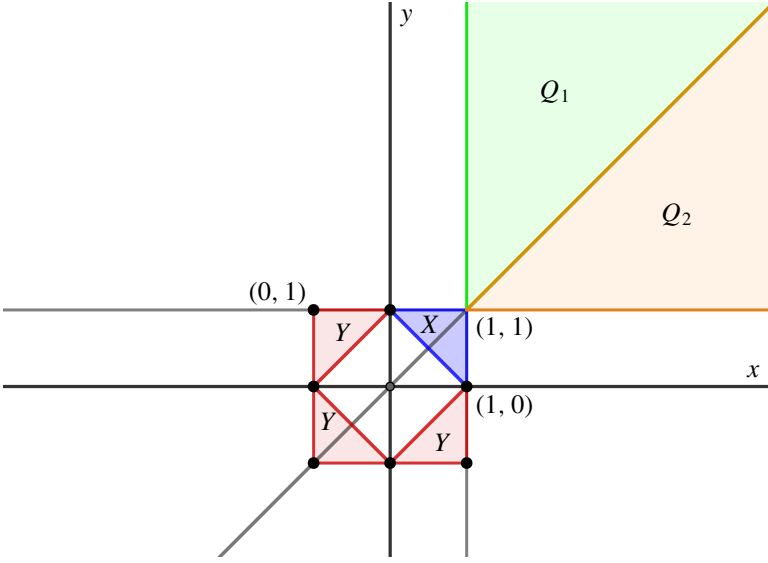


Figure 3. The cones Q_1 and Q_2 .

determines a valid ping-pong table, then the triangle must contain the intersection of these two tangent lines, which is the point $(1, 1)$. Thus, s must lie in one of the closed cones \bar{Q}_1 or \bar{Q}_2 defined by

$$\bar{Q}_1 = \mathbb{R}_{\geq 0} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + \mathbb{R}_{\geq 0} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix},$$

$$\bar{Q}_2 = \mathbb{R}_{\geq 0} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + \mathbb{R}_{\geq 0} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

These cones are shown in [Figure 3](#).

Let

$$s = \lambda_1 \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + \lambda_2 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \lambda_2 + 1 \\ \lambda_1 + \lambda_2 + 1 \\ \lambda_1 + \lambda_2 + 1 \end{pmatrix}$$

be a point in $\bar{Q}_1 \setminus (1, 1)$. This means that $\lambda_1, \lambda_2 \in \mathbb{R}_{\geq 0}$, and λ_1, λ_2 are not both zero. If X', Y' are a valid ping-pong table, then every point of Y' must be sent into X' by T . By continuity, this implies that T must send every point in \bar{Y}' to \bar{X}' . The point Rs is in \bar{Y}' , but we will show that TRs is not in \bar{X}' .

The closed triangle \bar{X}' is the intersection of the closed cones \bar{X}'_1 and \bar{X}'_2 defined by

$$\bar{X}'_1 = \mathbb{R}_{\geq 0} \left(s - \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \right) + \mathbb{R}_{\geq 0} \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix},$$

$$\bar{X}'_2 = \mathbb{R}_{\geq 0} \left(s - \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \right) + \mathbb{R}_{\geq 0} \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}.$$

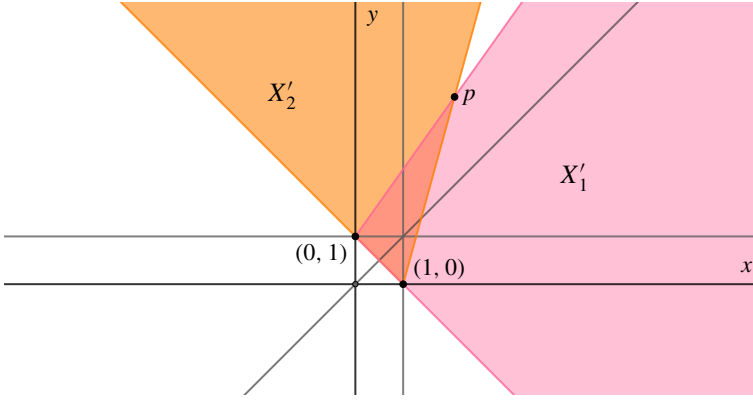


Figure 4. The cones X'_1 and X'_2 , whose intersection is X' .

These cones are illustrated in Figure 4. Suppose $TRs \in \bar{X}'_1 \cap \bar{X}'_2$. This means there are $a, b, c, d \geq 0$ such that

$$TRs = a \begin{pmatrix} \lambda_2 + 1 \\ \lambda_1 + \lambda_2 \end{pmatrix} + b \begin{pmatrix} 1 \\ -1 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} = c \begin{pmatrix} \lambda_2 \\ \lambda_1 + \lambda_2 + 1 \end{pmatrix} + d \begin{pmatrix} -1 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Using (4-3) and (4-4), we compute

$$TRs = T \begin{pmatrix} \lambda_1 + \lambda_2 + 1 \\ -\lambda_2 - 1 \end{pmatrix} = \begin{pmatrix} \frac{2\lambda_1 + \lambda_2 - 1}{2\lambda_1 - 3} \\ \frac{\lambda_1 - \lambda_2 - 3}{2\lambda_1 - 3} \end{pmatrix},$$

so a, b, c, d must be a solution to the system of linear equations

$$\begin{aligned} a(\lambda_2 + 1) + b &= \frac{2\lambda_1 + \lambda_2 - 1}{2\lambda_1 - 3}, & c\lambda_2 - d + 1 &= \frac{2\lambda_1 + \lambda_2 - 1}{2\lambda_1 - 3}, \\ a(\lambda_1 + \lambda_2) - b + 1 &= \frac{\lambda_1 - \lambda_2 - 3}{2\lambda_1 - 3}, & c(\lambda_1 + \lambda_2 + 1) + d &= \frac{\lambda_1 - \lambda_2 - 3}{2\lambda_1 - 3}. \end{aligned}$$

This system of equations has the unique solution

$$a = \frac{\lambda_1 - 1}{\theta}, \quad b = \frac{2(\lambda_1^2 + 2\lambda_1\lambda_2 + \lambda_2^2)}{\theta}, \quad c = \frac{\lambda_1 - 1}{\theta}, \quad d = \frac{-2(\lambda_2^2 + \lambda_1 + 3\lambda_2 + 1)}{\theta},$$

where

$$\theta = (2\lambda_1 - 3)(\lambda_1 + 2\lambda_2 + 1).$$

Since $\lambda_1, \lambda_2 \in \mathbb{R}_{\geq 0}$ and at least one of λ_1 and λ_2 is nonzero, b and d have opposite signs, contradicting that both are ≥ 0 . Thus, we cannot express TRs both as a nonnegative linear combination of generators of \bar{X}'_1 and as a nonnegative linear combination of generators of \bar{X}'_2 , so $TRs \notin \bar{X}'$.

Now suppose $s \in \bar{Q}_2 \setminus (1, 1)$. Let $F : (a, b) \mapsto (b, a)$ be reflection over the line $x = y$. It is clear from (4-3) and (4-4) that $FRF = R^{-1}$ and $FTF = T$. Since \bar{Q}_2 is the reflection of \bar{Q}_1 over the line $x = y$, we conclude from the previous argument that $TR^{-1}s \notin \bar{X}'$, so again X' and Y' are not a valid ping-pong table.

4C. A smaller ping-pong table. We have shown that C is the only simplicial cone that can be used to define a valid ping-pong table. If we drop the requirement that C be a simplicial cone, however, then we have additional possibilities. As Figure 1 illustrates, T maps the triangles RX , R^2X , and R^3X to three smaller triangles inside X . We can therefore obtain a smaller ping-pong table by defining X to be the union of these three triangles, and Y to be the union of the images of these triangles under R , R^2 , and R^3 . We will then be able to shrink X and Y even further. We now show that X and Y can be shrunk all the way down to the unit circle.

Lemma 4.2. *R and T map the unit circle to itself.*

Proof. Let (a, b) be a point on the unit circle. Clearly $b^2 + (-a)^2 = a^2 + b^2 = 1$, so $R(a, b)$ is on the unit circle. For $T(a, b)$, we compute

$$\left(\frac{2a + b - 2}{2a + 2b - 3}\right)^2 + \left(\frac{a + 2b - 2}{2a + 2b - 3}\right)^2 = \frac{5a^2 + 5b^2 + 8ab - 12a - 12b + 8}{4a^2 + 4b^2 + 8ab - 12a - 12b + 9}.$$

Since $a^2 + b^2 = 1$, we can simplify this to

$$\frac{5 + 8ab - 12a - 12b + 8}{4 + 8ab - 12a - 12b + 9} = 1,$$

which shows that $T(a, b)$ is on the unit circle. □

It follows from Lemma 4.2 and the discussion in Section 4A that the subsets

$$X = \{(a, b) \mid a^2 + b^2 = 1, a, b > 0\},$$

$$Y = \{(a, b) \mid a^2 + b^2 = 1, a < 0 \text{ or } b < 0\}$$

of the unit circle form a valid ping-pong table.

The projection ρ is defined by

$$\rho(x, y, z) = \left(\frac{-2(x - z)}{x - y + z}, \frac{-2y}{x - y + z}\right),$$

so the unit circle consists of the projections of vectors $(x, y, z) \in \mathbb{R}^3$ satisfying the quadratic equation

$$4(x - z)^2 + 4y^2 = (x - y + z)^2.$$

Let S be the surface in \mathbb{R}^3 defined by this equation. The maps R and T preserve this surface, so the intersection of S with the ping-pong table in \mathbb{R}^3 defined in Section 2 is a valid ping-pong table.

5. Comparison with the two-dimensional and four-dimensional cases

When $n = 2$, we have

$$R = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}, \quad U = \begin{pmatrix} 0 & -1 \\ 1 & 2 \end{pmatrix}, \quad T = \begin{pmatrix} 1 & 0 \\ -3 & 1 \end{pmatrix}.$$

As in the three-dimensional case, the matrices $U = TR$ and $RU^{-1}R^{-1} = T^{-1}R^{-1}$ have 1 as their only eigenvalue, and the corresponding eigenspace has dimension 1. The corresponding eigenvectors are $u = (-1, 1)$ and $v = (1, 2)$, and one easily verifies that the open cone C generated by u and v determines a ping-pong table by

$$X = C \cup -C, \quad Y = RX \cup R^2X.$$

One can see that C is (up to sign) the only simplicial cone with this property by an argument similar to the proof of [Lemma 3.2](#). Note that $\bar{X} \cup \bar{Y}$ is equal to all of \mathbb{R}^2 in this case.

When $n = 4$, we have

$$R = \begin{pmatrix} 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}, \quad U = \begin{pmatrix} 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 4 \\ 0 & 1 & 0 & -6 \\ 0 & 0 & 1 & 4 \end{pmatrix}, \quad T = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -5 & 1 & 0 & 0 \\ 5 & 0 & 1 & 0 \\ -5 & 0 & 0 & 1 \end{pmatrix}.$$

We describe a ping-pong table for $\langle R \rangle$ and $\langle T \rangle$, which is due to Brav and Thomas.

Theorem 5.1 [[Brav and Thomas 2014](#)]. *Let $P = \log(TR)$, and $Q = \log(T^{-1}R^{-1})$. Set $x = (0, 7, -2, 7)$, and define*

$$C^+ = \text{cone}(x, Px, P^2x, P^3x), \quad C^- = \text{cone}(x, Qx, Q^2x, Q^3x).$$

The sets

$$X = \pm C^+ \cup \pm C^-, \quad Y = RX \cup R^2X \cup R^3X \cup R^4X$$

are a ping-pong table for $\langle R \rangle$ and $\langle T \rangle$.

The proof in [[Brav and Thomas 2014](#)] shows that

$$T^k Y \subseteq \pm C^+ \quad \text{and} \quad T^{-k} Y \subseteq \pm C^-$$

for $k > 0$. (In addition, the proof shows that $TC^+ \subseteq C^+$ and $T^{-1}C^- \subseteq C^-$.)

Remark 5.2. In [[Brav and Thomas 2014](#)], the matrices R, T, U are represented in a different basis (and their T plays the role of our T^{-1}). Our matrices are obtained from theirs by conjugating by the change of basis matrix

$$S = \begin{pmatrix} 0 & 0 & 0 & 1 \\ -5 & 5 & 1 & -3 \\ 5 & -5 & -2 & 3 \\ 0 & 5 & 1 & -1 \end{pmatrix}.$$

The vector x in [Theorem 5.1](#) is a positive scalar multiple of Sv , where $v = (0, 1, -\frac{25}{12}, 0)$ is the vector defined on p. 338 of their paper (in the case $d = k = 5$).

Explicitly, the vectors defining C^+ and C^- are

$$\begin{aligned} Px &= (-5, 9, -15, 11), & P^3x &= (-1, 3, -3, 1), \\ x &= (0, 7, -2, 7), & P^2x = Q^2x &= (0, 1, -2, 1), \\ Qx &= (5, 16, -10, 14), & Q^3x &= (1, 2, -2, 4). \end{aligned}$$

We remark that P^3x is the unique (up to scalar) eigenvector of $U = TR$, and Q^3x is the unique eigenvector of $RU^{-1}R^{-1} = T^{-1}R^{-1}$. Furthermore, the matrices P^2 and Q^2 have rank 2, and their column spans intersect in the line spanned by $P^2x = Q^2x$. This vector is the analogue of w in the three-dimensional case (cf. [Section 2D](#)). In light of our results in the three-dimensional case, it seems natural to ask whether there is a vector y such that the cone

$$C = \text{cone}(P^3x, Q^3x, P^2x, y)$$

determines a ping-pong table by $X = \pm C$, $Y = RX \cup R^2X \cup R^3X \cup R^4X$. Our experiments in Sage suggest that there is no such y .

Acknowledgements

We owe a great deal of thanks to Hugh Thomas, who suggested this problem, guided our work on it during the summer of 2021, and provided valuable feedback on an earlier version of this paper. We are grateful to Benjamin Dequène for his help throughout the summer. In addition, we acknowledge the open source software package [[SageMath 2021](#)], which we used to carry out experiments and computations for this project.

Frieden was supported in part by the Canada Research Chairs program. Gélinas and Soucy were supported by NSERC Discovery Grant RGPIN-2016-04872 and Undergraduate Summer Scholarships from the Institut des Sciences Mathématiques (ISM).

References

- [Beukers and Heckman 1989] F. Beukers and G. Heckman, “Monodromy for the hypergeometric function ${}_nF_{n-1}$ ”, *Invent. Math.* **95**:2 (1989), 325–354. [MR](#) [Zbl](#)
- [Brav and Thomas 2014] C. Brav and H. Thomas, “Thin monodromy in $\text{Sp}(4)$ ”, *Compos. Math.* **150**:3 (2014), 333–343. [MR](#) [Zbl](#)
- [Chen, Yang and Yui 2008] Y.-H. Chen, Y. Yang, and N. Yui, “Monodromy of Picard–Fuchs differential equations for Calabi–Yau threefolds”, *J. Reine Angew. Math.* **616** (2008), 167–203. [MR](#) [Zbl](#)
- [Clausen 1828] T. Clausen, “Über die Fälle, wenn die Reihe von der Form $y = 1 + (\alpha/1) \cdot (\beta/\gamma)x + ((\alpha \cdot \alpha + 1)/(1 \cdot 2)) \cdot ((\beta \cdot \beta + 1)/(\gamma \cdot \gamma + 1))x^2 + \text{etc. ein Quadrat von der Form } z = 1 + (\alpha'/1) \cdot$

$(\beta'/\gamma') \cdot (\delta'/\epsilon')x + ((\alpha' \cdot \alpha' + 1)/(1 \cdot 2)) \cdot ((\beta' \cdot \beta' + 1)/(\gamma' \cdot \gamma' + 1)) \cdot ((\delta' \cdot \delta' + 1)/(\epsilon' \cdot \epsilon' + 1))x^2 +$
 etc. hat”, *J. Reine Angew. Math.* **3** (1828), 89–91. [MR](#)

[Filip and Fougeron 2021] S. Filip and C. Fougeron, “A cyclotomic family of thin hypergeometric monodromy groups in $Sp_4(\mathbb{R})$ ”, preprint, 2021. [arXiv 2106.09181](#)

[Fuchs, Meiri and Sarnak 2014] E. Fuchs, C. Meiri, and P. Sarnak, “Hyperbolic monodromy groups for the hypergeometric equation and Cartan involutions”, *J. Eur. Math. Soc. (JEMS)* **16**:8 (2014), 1617–1671. [MR](#) [Zbl](#)

[Heckman 2015] G. Heckman, “Tsinghua lectures on hypergeometric functions”, lecture notes, 2015, available at <https://www.math.ru.nl/~heckman/tsinghua.pdf>.

[Katz 2009] N. M. Katz, “Another look at the Dwork family”, pp. 89–126 in *Algebra, arithmetic, and geometry, II: in honor of Yu. I. Manin*, edited by Y. Tschinkel and Y. Zarhin, *Progr. Math.* **270**, Birkhäuser, Boston, 2009. [MR](#) [Zbl](#)

[Klein 1933] F. Klein, *Vorlesungen über die hypergeometrische Funktion*, *Grundl. Math. Wissen.* **39**, Springer, 1933. [MR](#) [Zbl](#)

[Lyndon and Schupp 1977] R. C. Lyndon and P. E. Schupp, *Combinatorial group theory*, *Ergebnisse der Math. und ihrer Grenzgebiete* **89**, Springer, 1977. [MR](#) [Zbl](#)

[SageMath 2021] The Sage Developers, *SageMath, the Sage Mathematics Software System*, 2021, available at <http://www.sagemath.org>. Version 9.3.

[Sarnak 2014] P. Sarnak, “Notes on thin matrix groups”, pp. 343–362 in *Thin groups and superstrong approximation*, edited by E. Breuillard and H. Oh, *Math. Sci. Res. Inst. Publ.* **61**, Cambridge Univ. Press, 2014. [MR](#) [Zbl](#)

[Schwarz 1873] H. A. Schwarz, “Über diejenigen fälle, in welchen die gaussische hypergeometrische reihe eine algebraische function ihres vierten elementes darstellt”, *J. Reine Angew. Math.* **75** (1873), 292–335. [MR](#)

[Singh and Venkataramana 2014] S. Singh and T. N. Venkataramana, “Arithmeticity of certain symplectic hypergeometric groups”, *Duke Math. J.* **163**:3 (2014), 591–617. [MR](#) [Zbl](#)

[Venkataramana 2014] T. N. Venkataramana, “Image of the Bureau representation at d -th roots of unity”, *Ann. of Math. (2)* **179**:3 (2014), 1041–1083. [MR](#) [Zbl](#)

Received: 2022-01-25

Revised: 2022-11-22

Accepted: 2022-12-17

gabefri@gmail.com

Université du Québec à Montréal, Montréal, QC, Canada

felixgel@yorku.ca

Université du Québec à Montréal, Montréal, QC, Canada

soucy.etienne@courrier.uqam.ca

Université du Québec à Montréal, Montréal, QC, Canada

Euclidean and affine curve reconstruction

Jose Agudelo, Brooke Dippold, Ian Klein,
Alex Kokot, Eric Geiger and Irina Kogan

(Communicated by Michael Dorff)

We consider practical aspects of reconstructing planar curves with prescribed Euclidean or affine curvatures. These curvatures are invariant under the special Euclidean group and the special affine groups, respectively, and play an important role in computer vision and shape analysis. We discuss and implement algorithms for such reconstruction, and give estimates on how close reconstructed curves are relative to the closeness of their curvatures in appropriate metrics. Several illustrative examples are provided.

1. Introduction

Rigid motions — compositions of translations, rotations and reflections — are fundamental transformations on the plane studied in a high-school geometry course. Two shapes related by these transformations are called *congruent*. The geometry studied in high school is based on the set of axioms formulated by Euclid around 300 BC and is called *Euclidean geometry*. Rigid motions make up the set of all transformations on the plane that preserve Euclidean distance between two points. A composition of two rigid motions is again a rigid motion, and the set of all rigid motions with the binary operation defined by composition satisfies the definition of a group (see [Section 2.1](#)). Naturally, this group is called the *Euclidean group* and is denoted by $E(2)$, where 2 indicates that the motions are considered in the 2-dimensional space, the plane.

To a human eye, two figures look the same if they are related by a rigid motion. However, since a reflection changes the orientation of an object, a group of orientation-preserving rigid motions, consisting of rotations and translations only, is often considered. This group is called the *special Euclidean group* and is denoted by $SE(2)$. In many applications, the congruence with respect to other groups is considered. For example, two shadows cast by the same object onto two

MSC2020: primary 53A04, 53A15, 53A55; secondary 34A45, 68T45.

Keywords: planar curves, Euclidean and affine transformations, Euclidean and affine curvatures, curve reconstruction, Picard iterations, distances.

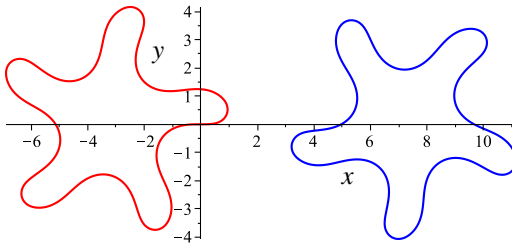


Figure 1. A special Euclidean transformation is a composition of a rotation and a translation.

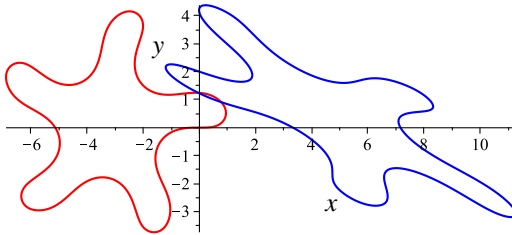


Figure 2. A special affine transformation is a composition of a unimodular linear transformation and a translation.

different planes by blocking the rays of light emitted from a lamp are related by a projective transformation. If a light source can be considered to be infinitely far away (like a sun), then the shadows are related by an affine transformation. See [Hartley and Zisserman 2004] for an excellent exposition of the roles played by projective, (special) affine, and (special) Euclidean transformations in computer vision. Starting in the 19th century, it was widely accepted that Euclidean geometry, although the most intuitive, is not the only possible consistent geometry, and that congruence can be defined relative to other transformation groups [Hawkins 1984].

In this work, we consider congruence of planar curves relative to the special Euclidean group $SE(2)$ and the *special affine group* $SA(2)$. The latter group consists of compositions of area- and orientation-preserving (i.e., *unimodular*) linear transformations and translations, and is sometimes also called the *equiaffine group*. In Figure 1, we show two curves related by a special Euclidean transformation, while in Figure 2 we show two curves related by a special affine transformation. For applications of curve matching under (special) Euclidean and affine transformations see, for instance, [Wolfson et al. 1988; Faugeras 1994; Calabi et al. 1998; Ames et al. 2002; Goldberg et al. 2004; Golubitsky et al. 2010; Flash and Handzel 2007; Hoff and Olver 2014].

It is widely known that two sufficiently smooth planar curves are $SE(2)$ -congruent if they have the same Euclidean curvature κ as a function of the Euclidean arc-length s . Somewhat less familiar, but also known from the 19th century, are the

notions of curvature and arc-length in other geometries, in particular in the special affine geometry [Guggenheimer 1977]. Similarly to the Euclidean case, one can show that two sufficiently smooth planar curves are SA(2)-congruent if they have the same affine curvature μ as a function of the affine arc-length α . Knowing that the curvature as a function of the arc-length determines a curve up to the relevant group of transformations, it is natural to ask two questions:

- (1) Is there a practical algorithm to reconstruct a curve from its curvature up to the relevant transformation group?
- (2) If two curvatures are close to each other in a certain metric, how close can the reconstructed curves be brought to each other by an element of the relevant transformation group?

In this paper, we study both of these questions, by methods and techniques that are well known. Namely, we review and implement a procedure for reconstructing a curve from its Euclidean curvature by successive integrations. The procedure for reconstructing curves from its affine curvature is more complicated and is based on Picard iterations. An implementation of these procedures can be found at https://egeig.com/research/curve_reconstruction. In [Theorem 12](#), we show how close, relative to the Hausdorff metric, two curves can be brought together by a special Euclidean transformation if their Euclidean curvatures are δ -close in the L^∞ -norm. [Theorem 19](#) addresses the same question in the special affine case.

Many of the theoretical results presented in this paper are well known and the new results presented here are hardly surprising. However, combined together and illustrated by specific examples, we believe they contribute to a better understanding of a classical, but important problem, relevant in many modern applications. This paper is the result of an REU project, which turned out to be of great pedagogical value, as it taught the students to combine the results and methods from various subjects: differential geometry, algebra, analysis and numerical analysis. In addition, this project involved theoretical work and the work of designing and implementing algorithms. The multidisciplinary nature of this project, on one hand, and its accessibility, on the other hand, allowed the undergraduate participants to truly experience the richness and challenges of mathematical research. We hope that we are able to convey to the reader the enjoyment of various aspects of the mathematical research that we experienced while working on the project.

The paper is structured as follows. [Section 2](#) contains preliminaries and is split as follows: In [Section 2.1](#), after reviewing the definitions of groups and group actions, we define the notions of congruence and symmetry of curves relative to a given group. In [Sections 2.2](#) and [2.3](#), we follow [Guggenheimer 1977] to define Euclidean and affine moving frames and invariants. In [Section 2.4](#), we introduce norms and distances, used in this paper, in the spaces of functions, matrices, matrices

of functions, and curves and prove some useful inequalities. In [Section 2.5](#), we establish some results about convergence of matrices and their norms.

[Section 3](#) contains explicit formulas for reconstructing a curve from its Euclidean curvature function and gives an upper bound on the closeness of reconstructed curves with close Euclidean curvatures. [Section 4](#) introduces a Picard iteration scheme for reconstructing a curve from its affine curvature function and gives an upper bound on the closeness of reconstructed curves with close affine curvatures. Directions of further research are indicated in [Section 5](#). In the [Appendix](#), we derive a power series representation for curves whose affine curvatures are given by a monomial.

2. Preliminaries

2.1. Congruence and symmetry of the planar curves. To keep the presentation self-contained, we remind the reader the standard definitions of groups and group-actions.

Definition 1. A *group* is a set G with a binary operation $\cdot : G \times G \rightarrow G$ that satisfies the following properties:

- (1) (associativity) $(g_1 \cdot g_2) \cdot g_3 = g_1 \cdot (g_2 \cdot g_3)$ for all $g_1, g_2, g_3 \in G$.
- (2) (identity element) There exists a unique $e \in G$ such that $e \cdot g = g \cdot e = g$ for all $g \in G$.
- (3) (inverse element) For each $g \in G$, there exists an element $h \in G$ such that $g \cdot h = h \cdot g = e$. We define $g^{-1} := h$.

Definition 2. An *action* of a group G on a set P is a map $\phi : G \times P \rightarrow P$ satisfying the following properties:

- (1) (associativity) $\phi(g_1 \cdot g_2, p) = \phi(g_1, \phi(g_2, p))$ for all $g_1, g_2 \in G$ and for all $p \in P$.
- (2) (action of the identity element) $\phi(e, p) = p$ for all $p \in P$.

We use a shorter notation $\phi(g, p) := gp$. Each element $g \in G$ determines a bijective map $g : P \rightarrow P, p \rightarrow gp$.

Groups are often defined through their actions. For example, a *rotation* in the plane by the angle $\theta > 0$ about the origin in the counterclockwise direction sends a point (x, y) in the plane to the point

$$(\bar{x}, \bar{y}) = (x \cos \theta - y \sin \theta, x \sin \theta + y \cos \theta) = (x, y)R_\theta^{-1}, \quad (1)$$

where the 2×2 matrix R_θ is given by

$$R_\theta = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}. \quad (2)$$

We multiply by the matrix on the right because we treat points (and vectors) in \mathbb{R}^2 as *row vectors*. We invert the matrix to satisfy the associativity property in the definition of the group action.¹ Rotation by $\theta = 0$ corresponds to the identity matrix and leaves all points in place, while R_θ with $\theta < 0$ corresponds to the clockwise rotation by the angle $|\theta|$. The set of matrices $\{R_\theta \mid \theta \in \mathbb{R}\}$ with the binary operation given by matrix multiplication satisfies the definition of a group. This group is called the *special orthogonal group* and is denoted by $\text{SO}(2)$. The word *special* in the name of the group indicates that $\det(R_\theta) = 1$ and so the orthonormal basis defined by its columns (or rows) is positively oriented. In fact, the group $\text{SO}(2)$ consists of all 2×2 matrices whose two columns (or two rows) form a positively oriented orthonormal basis in \mathbb{R}^2 . The map $\phi : \text{SO}(2) \times \mathbb{R}^2 \rightarrow \mathbb{R}^2$ defined by (1) satisfies the definition of a group-action. The associativity property in Definition 2 states that the action of the product of matrices $R_{\theta_1} \cdot R_{\theta_2}$ is the composition of the rotation by the angle θ_2 followed by the rotation by the angle θ_1 .

The *translation* in the plane by a vector $\mathbf{v} = (a, b)$ sends a point (x, y) to the point

$$(\bar{x}, \bar{y}) = (x, y) + (a, b) = (x + a, y + b). \quad (3)$$

The set of vectors $\mathbf{v} \in \mathbb{R}^2$ with the binary operation given by vector addition satisfies the definition of a group, with the zero vector being the identity element of this group. Formula (3) describes the action of this group on the plane. The composition of the rotation by θ followed by the translation by \mathbf{v} sends a point (x, y) to the point

$$(\bar{x}, \bar{y}) = (x \cos \theta - y \sin \theta + a, x \sin \theta + y \cos \theta + b) = (x, y)R_\theta^{-1} + \mathbf{v}. \quad (4)$$

The set of all compositions of rotations and translations also satisfies the definition of a group. It is called the *special Euclidean group* and is denoted by $\text{SE}(2)$. This is the group of all transformations in the plane that preserve distances (and, therefore, angles) in the plane, as well as the orientation. The composition of a rotation/translation pair $(R_{\theta_2}, \mathbf{v}_2)$ followed by a pair $(R_{\theta_1}, \mathbf{v}_1)$ is equivalent to the rotation $R_{\theta_1}R_{\theta_2} = R_{\theta_1+\theta_2}$ followed by the translation by the vector $\mathbf{v}_2R_{\theta_1}^{-1} + \mathbf{v}_1$. Thus we can think of the special Euclidean group as the set of pairs $\{(R_\theta, \mathbf{v}) \mid R_\theta \in \text{SO}(2), \mathbf{v} \in \mathbb{R}^2\}$ with the group operation

$$(R_{\theta_1}, \mathbf{v}_1) \cdot (R_{\theta_2}, \mathbf{v}_2) = (R_{\theta_1}R_{\theta_2}, \mathbf{v}_2R_{\theta_1}^{-1} + \mathbf{v}_1). \quad (5)$$

In other words, $\text{SE}(2) = \text{SO}(2) \ltimes \mathbb{R}^2$ is a semidirect product of the translation and rotation groups.

If in (4) and (5), we replace the rotation matrix R_θ with an arbitrary nonsingular 2×2 matrix M , we obtain an action of the *affine group*, $A(2) = \text{GL}(2) \ltimes \mathbb{R}^2$,

¹Since rotation matrices commute, the associativity property will be satisfied without the inversion, but it is essential for generalizations to other groups.

a semidirect product of the group of invertible linear transformations and translations. Restricting the matrix M to the group of *unimodular* matrices $\text{SL}(2) = \{M \mid \det(M) = 1\}$, we obtain a smaller group which is called the *special affine* or the *equiaffine* group² $\text{SA}(2) = \text{SL}(2) \ltimes \mathbb{R}^2$. A generic $\text{SA}(2)$ -transformation does not preserve distance or angles, but it preserves areas.

An action of a group on the plane induces the action on the curves in the plane. In this paper, we consider curves satisfying the following definition.

Definition 3 (planar curve). A planar curve \mathcal{C} is the image of a *continuous locally injective*³ map $\gamma : \mathbb{R} \rightarrow \mathbb{R}^2$. We call \mathcal{C} *closed* if its parametrization γ is periodic. We often restrict the domain of γ to an open or a closed interval $I \subset \mathbb{R}$.

Given a group G acting continuously on the plane, the image of a curve \mathcal{C} parametrized by γ , under a transformation $g \in G$, is the curve $g\mathcal{C} = \{gp \mid p \in \mathcal{C}\}$ parametrized by $g\gamma = g \circ \gamma$.

Definition 4. Given a group G acting on the plane, we say that two planar curves \mathcal{C}_1 and \mathcal{C}_2 are *G -congruent* ($\mathcal{C}_1 \stackrel{G}{\cong} \mathcal{C}_2$) if there exists $g \in G$, such that $\mathcal{C}_2 = g\mathcal{C}_1$.

Definition 5. An element $g \in G$ is a *G -symmetry* of \mathcal{C} if

$$g\mathcal{C} = \mathcal{C}.$$

It is easy to show that the set of such elements, denoted by $\text{sym}_G(\mathcal{C})$, is a subgroup of G , called the *G -symmetry group* of \mathcal{C} . The cardinality of $\text{sym}_G(\mathcal{C})$ is called the *symmetry index* of \mathcal{C} .

Figure 1 shows two $\text{SE}(2)$ -congruent curves, each with five $\text{SE}(2)$ -symmetries. Figure 2 shows two $\text{SA}(2)$ -congruent curves, each with five $\text{SA}(2)$ -symmetries. As a side remark, we note that the five $\text{SA}(2)$ -symmetries of the left curve in Figure 2, in fact, belong to $\text{SE}(2)$, while the five $\text{SA}(2)$ -symmetries of the right curve do not. The method of moving frames, pioneered by Bartels, Frenet, Serret, Cotton, and Darboux, and greatly extended by Cartan, allows one to solve the G -congruence problem for sufficiently smooth curves⁴ by assigning a frame of basis vectors along a curve in a way that is compatible with the G -action. We will review this classical construction of such frames for the $\text{SE}(2)$ and $\text{SA}(2)$ actions by following, for the most part, the exposition given in [Guggenheimer 1977]. For a more detailed history and generalizations to arbitrary Lie group-actions, see [Olver 2015].

²From now on we will use the term *special affine*.

³A map $\gamma : I \rightarrow \mathbb{R}^2$, where I is an open subset of \mathbb{R} , is *locally injective* if, for any $t \in I$, there exists an open neighborhood $J \subset I$ such that $\gamma|_J$ is injective.

⁴A curve is called C^k -smooth if the k -th order derivative of its parametrization γ is continuous.

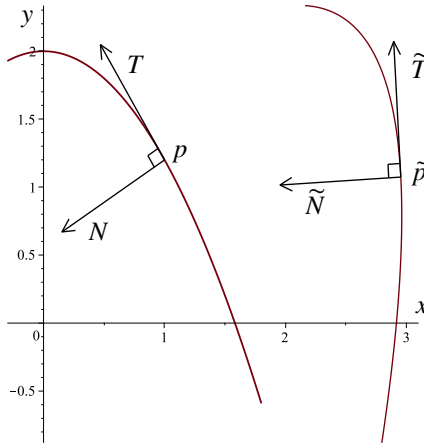


Figure 3. The SE(2)-action preserves the lengths of vectors and the angle between them.

2.2. Euclidean moving frame and invariants. The SE(2)-frame at point p of a planar curve \mathcal{C} consists of the unit tangent vector $T(p)$ and the unit normal vector $N(p)$. Orientation for $T(p)$ is defined by the parametrization γ of \mathcal{C} , while the orientation for $N(p)$ is chosen so that the pair of vectors $T(p)$ and $N(p)$ is positively oriented, i.e., the closest rotation from $T(p)$ to $N(p)$ is counterclockwise. Considering T and N to be row vectors, we combine them into an SE(2)-frame matrix

$$A_{\mathcal{C}}(p) = \begin{pmatrix} T(p) \\ N(p) \end{pmatrix}. \quad (6)$$

An important observation is that $A_{\mathcal{C}}(p)$ is an orthogonal matrix. In fact, it is precisely the rotation matrix which brings the moving frame basis consisting of $T(p)$ and $N(p)$ to the standard orthonormal basis in \mathbb{R}^2 under the action (1). An element $g \in \text{SE}(2)$ acting on \mathbb{R}^2 maps the curve \mathcal{C} to $\tilde{\mathcal{C}}$ and the point p to \tilde{p} . Since the SE(2)-action preserves tangency and length, it maps the SE(2)-frame at $p \in \mathcal{C}$ to the SE(2)-frame at $\tilde{p} \in \tilde{\mathcal{C}}$. See Figure 3 for an illustration. This compatibility property of the frame is called *equivariance* and can be expressed as

$$A_g \mathcal{C}(gp) = A_{\mathcal{C}}(p) R_g^{-1}, \quad (7)$$

where \mathcal{C} is an arbitrary curve, $p \in \mathcal{C}$, $g \in \text{SE}(2)$, R_g is the rotational part of the transformation g .

It is well known that any C^1 -smooth nondegenerate curve \mathcal{C} can be parametrized as

$$\gamma : s \rightarrow p = \gamma(s),$$

so that

$$T(p) = \gamma_s(s) \quad (8)$$

is the unit tangent vector at the point $p = \gamma(s) \in \mathcal{C}$. (Here and below, a variable in the subscript denotes the differentiation with respect to this variable.) Explicitly, if

$$\hat{\gamma} : t \rightarrow \gamma(t) = (x(t), y(t))$$

is any parametrization of \mathcal{C} , then

$$s(t) = \int_0^t |\hat{\gamma}_\tau| d\tau = \int_0^t \sqrt{x'(\tau)^2 + y'(\tau)^2} d\tau.$$

The parameter s is called the *Euclidean arc-length* parameter. Its differential

$$ds = |\hat{\gamma}_t| dt \tag{9}$$

is called the infinitesimal *Euclidean arc-length*. Clearly, the integral of ds along a curve segment produces the Euclidean length of the curve segment.

We now assume that \mathcal{C} is C^2 -smooth and note that the differentiation of the identity $|T(s)| = 1$ implies that $T_s(s)$ is orthogonal to $T(s)$, and so $T_s(s)$ is proportional to $N(s)$. Thus there is a function $\kappa(s)$, called the *Euclidean curvature function*, such that

$$T_s(s) = \kappa(s)N(s). \tag{10}$$

Explicitly, $\kappa(s) = \pm|\gamma_{ss}|$, with “+” when the rotation from γ_s to γ_{ss} is counter-clockwise and “−” otherwise. For an arbitrary parametrization $\hat{\gamma}(t)$, we have

$$\kappa(t) = \frac{\det(\hat{\gamma}_t, \hat{\gamma}_{tt})}{|\hat{\gamma}_t|^3}. \tag{11}$$

The Euclidean curvature of a circle of radius r is constant and is equal to $1/r$. The Euclidean curvature of \mathcal{C} at p equals the curvature of its osculating circle⁵ at p .

Since $|N(s)| = 1$, we know that $N_s(s)$ is proportional to $T(s)$. Furthermore, differentiating the scalar product identity $T(s) \cdot N(s) = 0$, we conclude

$$N_s(s) = -\kappa(s)T(s). \tag{12}$$

Equations (10) and (12) are called *Frenet equations* and can be written in matrix form as

$$A_s = CA,$$

where A is the Euclidean frame matrix (6), while

$$C(s) = A_s(s)A(s)^{-1} = \begin{pmatrix} 0 & \kappa(s) \\ -\kappa(s) & 0 \end{pmatrix} \tag{13}$$

⁵The osculating circle to \mathcal{C} at p passes through p , and the derivatives of the arc-length parametrizations at $s=0$ (with $s=0$ corresponding to p) of the osculating circle and \mathcal{C} coincide up to second order.

is the Euclidean *Cartan matrix*. From the equivariance property (7) and the SE(2)-invariance⁶ of C (and, therefore, of κ) it follows

$$\kappa_{gC}(gp) = \kappa_C(p),$$

where C is an arbitrary curve, $p \in C$, $g \in \text{SE}(2)$.

2.3. Affine moving frame and invariants. The action of the special affine group SA(2) preserves neither Euclidean distances nor angles. Thus the Euclidean moving frame consisting of the unit tangent and the unit normal at each point of a curve C is not compatible with the SA(2)-action. However, the SA(2)-action preserves areas, and we can use this property to define an SA(2)-equivariant frame.

It turns out that any C^2 -smooth curve C can be parametrized by

$$\gamma : \alpha \rightarrow p = \gamma(\alpha),$$

so that the area of the parallelogram defined by vectors

$$T(p) = \gamma_\alpha \quad \text{and} \quad N(p) = \gamma_{\alpha\alpha} \quad (14)$$

is 1 and the closest rotation from $T(p)$ to $N(p)$ is counterclockwise. The parameter α is called the *affine arc-length parameter*. Explicitly, if

$$\hat{\gamma} : t \rightarrow \hat{\gamma}(t) = (x(t), y(t))$$

is any parametrization of C , then

$$\alpha(t) = \int_0^t \det(\hat{\gamma}_\tau(\tau), \hat{\gamma}_{\tau\tau}(\tau))^{1/3} d\tau. \quad (15)$$

Recalling formulas (9) and (11), we rewrite (15) in terms of Euclidean curvature and arc-length:

$$\alpha(s) = \int_0^s \kappa(\tau)^{1/3} d\tau. \quad (16)$$

Vectors $T(p) = \gamma_\alpha$ and $N(p) = \gamma_{\alpha\alpha}$ are called the *affine tangent and normal* to C at p , respectively. It is important to note that although $T(p)$ is tangent to C at p , it is, in general, not of the unit length, while $N(p)$, in general, is neither perpendicular to $T(p)$ nor of the unit length. The SA(2)-*frame matrix* is then defined by

$$A_C(p) = \begin{pmatrix} T(p) \\ N(p) \end{pmatrix} = \begin{pmatrix} \gamma_\alpha \\ \gamma_{\alpha\alpha} \end{pmatrix}. \quad (17)$$

An important observation is that, by construction, $\det(A_C(p)) = 1$. In fact, this is the matrix of the unimodular linear transformation which brings the affine moving

⁶The Euclidean curvature κ changes its sign under reflections and, therefore, is not invariant under the full Euclidean group $E(2)$. Nonetheless, it is customary called the *Euclidean curvature* rather than the *special Euclidean curvature*.

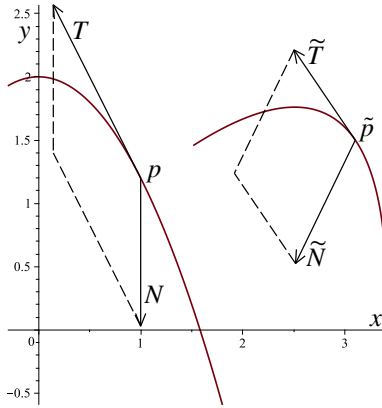


Figure 4. The SA(2)-action preserves the area of the parallelogram defined by the affine tangent and normal vectors, but not their lengths or the angle between them.

frame basis consisting of $T(p)$ and $N(p)$ to the standard orthonormal basis under the action on row vectors $\mathbf{v} \rightarrow \mathbf{v}M^{-1}$.

The affine moving frame is SA(2)-equivariant: an element $g \in \text{SA}(2)$ mapping the curve \mathcal{C} to $\tilde{\mathcal{C}}$ and the point $p \in \mathcal{C}$ to the point $\tilde{p} \in \tilde{\mathcal{C}}$ also maps the affine tangent and normal vectors at $p \in \mathcal{C}$ to the affine tangent and normal vectors at $\tilde{p} \in \tilde{\mathcal{C}}$. See [Figure 4](#) for an illustration. In the matrix form, this can be expressed as

$$A_g c(gp) = A_c(p)M_g^{-1}, \quad (18)$$

where \mathcal{C} is an arbitrary curve, $p \in \mathcal{C}$, $g \in \text{SA}(2)$, and M_g is the matrix part of g .

By definition,

$$T_\alpha(\alpha) = N(\alpha). \quad (19)$$

Using this and differentiating the identity $\det(T(\alpha), N(\alpha)) = 1$ with respect to α we obtain $\det(T(\alpha), N_\alpha(\alpha)) = 0$. This implies that N_α is proportional to T , and, therefore, there is a function $\mu(\alpha)$, called the *affine curvature function*, such that

$$N_\alpha(\alpha) = -\mu(\alpha)T(\alpha), \quad (20)$$

where

$$\mu(\alpha) = -\det(N_\alpha(\alpha), N(\alpha)) = \det(\gamma_{\alpha\alpha}(\alpha), \gamma_{\alpha\alpha\alpha}(\alpha)). \quad (21)$$

If $\hat{\gamma}(t)$ is an arbitrary parametrization of \mathcal{C} , then the formula for $\mu(t)$ is rather long (see formula (7-24) in [\[Guggenheimer 1977\]](#)), but we can get a more concise formula in terms of the Euclidean curvature and the Euclidean arc-length [\[Kogan 2003\]](#):

$$\mu = \frac{3\kappa(\kappa_{ss} + 3\kappa^3) - 5\kappa_s^2}{9\kappa^{8/3}}. \quad (22)$$

The affine curvature of a conic is constant (see [Section 4](#) for the details). The affine curvature of \mathcal{C} at p is the curvature of the osculating conic⁷ at p .

Equations (19) and (20) are the affine versions of the Frenet equations and can be written in the matrix form as

$$A_\alpha(\alpha) = C(\alpha)A(\alpha), \quad (23)$$

where A is the affine frame matrix (17), while

$$C(\alpha) = A_\alpha(\alpha)A(\alpha)^{-1} = \begin{pmatrix} 0 & 1 \\ -\mu(\alpha) & 0 \end{pmatrix} \quad (24)$$

is the *affine Cartan matrix*. From the equivariance property (18) and the SA(2)-invariance⁸ of C (and, therefore, of μ) it follows

$$\mu_{g\mathcal{C}}(gp) = \mu_{\mathcal{C}}(p),$$

where \mathcal{C} is an arbitrary curve, $p \in \mathcal{C}$, and $g \in \text{SA}(2)$.

2.4. Norms and distances. For a continuous function $f(t)$ on a closed interval $[0, L]$, let

$$\|f\|_{[0,L]} := \max_{t \in [0,L]} \{|f(t)|\}. \quad (25)$$

For a $k \times \ell$ matrix A with real entries we define

$$\langle A \rangle := \max_{\substack{i=1,\dots,k \\ j=1,\dots,\ell}} \{|a_{ij}|\}, \quad (26)$$

where a_{ij} are the entries of A and $|\cdot|$ is the usual absolute value. If $A(t)$ is a matrix whose entries are functions on a real interval $[0, L]$, we define a real-valued function

$$\langle A \rangle(t) := \langle A(t) \rangle. \quad (27)$$

If the entries of $A(t)$ are continuous functions, it is easy to show that $\langle A \rangle(t)$ is continuous on the interval $[0, L]$ and so we may define

$$\|A\|_{[0,L]} := \|\langle A \rangle(t)\|_{[0,L]} = \max_{t \in [0,L]} \langle A(t) \rangle = \max_{\substack{t \in [0,L] \\ i=1,\dots,k \\ j=1,\dots,\ell}} \{a_{ij}(t)\}, \quad (28)$$

where the first equality is the definition, and the subsequent equalities follow from (25)–(27).

⁷The osculating conic to \mathcal{C} at p passes through p , and the derivatives of the affine arc-length parametrizations at $\alpha = 0$ (with $\alpha = 0$ corresponding to p) of the osculating conic and \mathcal{C} coincide up to third order.

⁸The affine curvature μ is scaled under nonunimodular linear transformations and, therefore, is not invariant under the full affine group $A(2)$. Nonetheless, following [[Guggenheimer 1977](#)], we use the term *affine curvature* rather than the *special* or *equi*affine curvature.

We note that $\langle \cdot \rangle$ and $\| \cdot \|_{[0,L]}$ are L^∞ -norms on the vector spaces of matrices of matching sizes with real entries and functional entries, respectively, and, in particular, they satisfy the triangle inequality.

As usual, the differentiation and integration of matrices with functional entries are defined componentwise. For a matrix $A(t)$, whose entries are continuous functions on a real interval $[0, L]$, and $t \in [0, L]$ we will repeatedly use the inequalities

$$\left\langle \int_0^t A(\tau) d\tau \right\rangle \leq \int_0^t \langle A \rangle(\tau) d\tau \leq \|A\|_{[0,t]} t \leq \|A\|_{[0,L]} t \leq \|A\|_{[0,L]} L. \quad (29)$$

For a vector $\mathbf{v} \in \mathbb{R}^\ell$, its L^∞ -norm $\langle \mathbf{v} \rangle$ and its Euclidean L^2 -norm $|\mathbf{v}|$ obey the inequality

$$|\mathbf{v}| \leq \sqrt{\ell} \langle \mathbf{v} \rangle. \quad (30)$$

In this paper, the closeness of two curves is determined by the Hausdorff distance, and we recall its definition. Let P and Q be two subsets of \mathbb{R}^n . We define

$$d_{PQ} = \sup_{p \in P} \inf_{q \in Q} |p - q| \quad \text{and} \quad d_{QP} = \sup_{q \in Q} \inf_{p \in P} |p - q|.$$

Then the *Hausdorff* distance between P and Q is defined by

$$d(P, Q) = \max\{d_{PQ}, d_{QP}\}.$$

To find an upper bound for the Hausdorff distance between two *planar curves* \mathcal{C}_1 and \mathcal{C}_2 parametrized by $\gamma_1(t)$ and $\gamma_2(t)$ for $t \in [0, L]$ we note that

$$\begin{aligned} d_{\mathcal{C}_1\mathcal{C}_2} &= \sup_{\tau \in [0,L]} \inf_{t \in [0,L]} |\gamma_1(\tau) - \gamma_2(t)| \leq \sup_{\tau \in [0,L]} |\gamma_1(\tau) - \gamma_2(\tau)| \\ &\leq \sqrt{2} \sup_{\tau \in [0,L]} \langle \gamma_1(\tau) - \gamma_2(\tau) \rangle = \sqrt{2} \|\gamma_1 - \gamma_2\|_{[0,L]}. \end{aligned}$$

The same inequality holds for $d_{\mathcal{C}_2\mathcal{C}_1}$ and, therefore, for the Hausdorff distance we have

$$d(\mathcal{C}_1, \mathcal{C}_2) \leq \sqrt{2} \|\gamma_1 - \gamma_2\|_{[0,L]}. \quad (31)$$

2.5. Convergence. We recall the definition of uniform convergence:

Definition 6. Let $\{f_n\}_{n=1}^\infty$ be a sequence of real-valued functions on a set P . We say that $\{f_n\}$ converges to a function f *uniformly* on P if, for every $\varepsilon > 0$, there exists n_ε such that

$$|f_n(p) - f(p)| < \varepsilon \quad \text{for all } n > n_\varepsilon \text{ and all } p \in P. \quad (32)$$

The difference between the uniform and *pointwise* convergence is that one can choose n_ε which “works” for all $p \in P$. If P is an interval $[0, L]$, then uniform convergence of $\{f_n\}$ to f is equivalent to

$$\lim_{n \rightarrow \infty} \|f_n - f\|_{[0,L]} = 0.$$

Lemma 7. Let $\{f_n\}_{n=1}^{\infty}$ be a sequence of real-valued functions on a domain P **uniformly** convergent to a function f on P . Assume further that each of the functions f_n , and also f , achieves its maximum value on P . Then

$$\lim_{n \rightarrow \infty} \max_{p \in P} \{f_n(p)\} = \max_{p \in P} \{f(p)\}. \quad (33)$$

Proof. By assumption there exist $\{p_n\} \subset P$, $n = 0, \dots, \infty$, such that, for all $p \in P$,

$$f(p) \leq f(p_0) = m_0 \quad \text{and} \quad f_n(p) \leq f_n(p_n) = m_n, \quad n \in \mathbb{Z}_+,$$

where m_0 is the maximal value of f and m_n is the maximal value of f_n , $n \in \mathbb{Z}_+$, on P . Identity (33) can be rewritten as

$$\lim_{n \rightarrow \infty} m_n = m_0. \quad (34)$$

For an arbitrary $\varepsilon > 0$, let n_ε be such that, for all $n > n_\varepsilon$ and all $p \in P$, (32) holds, and so, for all $n > n_\varepsilon$ and all $p \in P$,

$$f(p) - \varepsilon < f_n(p) < f(p) + \varepsilon. \quad (35)$$

Substitute p_0 in the left inequality in (35) to get

$$f(p_0) - \varepsilon = m_0 - \varepsilon < f_n(p_0) \leq m_n. \quad (36)$$

Substitute p_n in the right inequality in (35) to get

$$f_n(p_n) = m_n < f(p_n) + \varepsilon \leq m_0 + \varepsilon. \quad (37)$$

Together (36) and (37) imply that for an arbitrary $\varepsilon > 0$, there exists n_ε such that, for all $n > n_\varepsilon$,

$$m_0 - \varepsilon < m_n < m_0 + \varepsilon,$$

which is equivalent to (34). \square

We say that a sequence of $k \times \ell$ matrices $\{A_n\}_{n=1}^{\infty}$ with real entries $a_{n;ij}$ converges to a $k \times \ell$ matrix A with real entries a_{ij} , if, for all $i = 1, \dots, k$, $j = 1, \dots, \ell$,

$$\lim_{n \rightarrow \infty} a_{n;ij} = a_{ij}.$$

If $\{A_n(t)\}$ is a sequence of matrices whose elements are real-valued functions on an interval $[0, L]$, then we say that $\{A_n(t)\}_{n=1}^{\infty}$ pointwise converges to $A(t)$ if, for all $t \in [0, L]$ and all $i = 1, \dots, k$, $j = 1, \dots, \ell$,

$$\lim_{n \rightarrow \infty} a_{n;ij}(t) = a_{ij}(t).$$

If the latter convergences are uniform on $[0, L]$, we say that $\{A_n(t)\}$ converges to $A(t)$ uniformly. Equivalently, the uniform convergence can be defined by

$$\lim_{n \rightarrow \infty} \|A_n - A\|_{[0, L]} = 0.$$

From [Lemma 7](#), we have the following important corollary, which we use repeatedly.

Corollary 8. (1) *Let $\{A_n\}_{n=1}^\infty$ be a sequence of matrices with real entries convergent to a matrix A , then*

$$\lim_{n \rightarrow \infty} \langle A_n \rangle = \langle A \rangle. \quad (38)$$

(2) *Let $\{A_n(t)\}_{n=1}^\infty$ be a sequence of matrices whose elements are real-valued functions on the interval $[0, L]$ pointwise convergent to a matrix of functions $A(t)$. Then, for all t ,*

$$\lim_{n \rightarrow \infty} \langle A_n \rangle(t) = \langle A \rangle(t). \quad (39)$$

(3) *If the entries of $A_n(t)$ are continuous functions and $\{A_n(t)\}_{n=1}^\infty$ converges to $A(t)$ uniformly on $[0, L]$, then*

$$\lim_{n \rightarrow \infty} \|A_n\|_{[0, L]} = \|A\|_{[0, L]}. \quad (40)$$

Proof. (1) Identity [\(38\)](#) is equivalent to

$$\lim_{n \rightarrow \infty} \max_{\substack{i=1, \dots, k \\ j=1, \dots, \ell}} \{|a_{n,ij}|\} = \max_{\substack{i=1, \dots, k \\ j=1, \dots, \ell}} \{|\lim_{n \rightarrow \infty} a_{n,ij}|\}. \quad (41)$$

Let B_n , $n \in \mathbb{Z}_+$, and B denote matrices whose elements are $|a_{n,ij}|$ and $|a_{ij}|$, respectively. Then, due to a well-known and easy-to-show fact that \lim and the absolute value are interchangeable, $\lim_{n \rightarrow \infty} B_n = B$. Note that a $k \times \ell$ matrix with real entries can be viewed as a real-valued function on a finite set of ordered pairs

$$P = \{(i, j) \mid i = 1, \dots, k, j = 1, \dots, \ell\}. \quad (42)$$

Viewed as a sequence of such functions, $\{B_n\}_{n=1}^\infty$ converges to B uniformly on P . Any function on a finite set attains its maximum and so we can apply [Lemma 7](#) to conclude that

$$\lim_{n \rightarrow \infty} \max_{p \in P} \{B_n(p)\} = \max_{p \in P} \lim_{n \rightarrow \infty} \{B_n(p)\},$$

which is equivalent to [\(41\)](#).

(2) Identity [\(39\)](#) is an immediate consequence of [\(38\)](#).

(3) Identity [\(40\)](#) is equivalent to

$$\lim_{n \rightarrow \infty} \max_{\substack{t \in [0, L] \\ i=1, \dots, k \\ j=1, \dots, \ell}} \{|a_{n,ij}(t)|\} = \max_{\substack{t \in [0, L] \\ i=1, \dots, k \\ j=1, \dots, \ell}} \{|\lim_{n \rightarrow \infty} a_{n,ij}(t)|\}. \quad (43)$$

Let $B_n(t)$ and $B(t)$ denote matrices whose elements are $|a_{n,ij}(t)|$ and $|a_{ij}(t)|$, respectively. Then $\{B_n(t)\}$ converges to $B(t)$ uniformly on $[0, L]$. Uniform convergence

implies that entries of $B(t)$ are continuous. We can view a $k \times \ell$ matrix whose entries are continuous functions on $[0, L]$ as real-valued functions on the set

$$Q = P \times [0, L],$$

where P is defined by (42). With this point of view, the sequence of functions $\{B_n\}_{n=1}^\infty$ converges to B uniformly on Q , and each of these functions attains its maximum value on Q . Thus they satisfy the assumptions of Lemma 7, and so

$$\lim_{n \rightarrow \infty} \max_{q \in Q} \{B_n(q)\} = \max_{q \in Q} \lim_{n \rightarrow \infty} \{B_n(q)\},$$

which is equivalent to (43). \square

3. Euclidean reconstruction

In this section, we review how a curve can be reconstructed from its Euclidean curvature by two successive integrations (Theorem 9). We then use these formulas to estimate how close, relative to the Hausdorff distance, two curves can be brought together by a special-Euclidean transformation, provided their Euclidean curvatures as functions of the Euclidean arc-length are δ -close in the L^∞ -norm (Theorem 12) or δ -close in the L^1 -norm (Theorem 13).

Theorem 9 (Euclidean reconstruction). *Let $\kappa(s)$ be a continuous function on an interval $[0, L]$. Then there is a unique, up to a special Euclidean transformation, curve \mathcal{C} with the Euclidean arc-length parametrization $\gamma(s) = (x(s), y(s))$, $s \in [0, L]$, such that $\kappa(s) = x'(s)y''(s) - y'(s)x''(s)$ is its Euclidean curvature.*

Proof. According to (8), (10), and (12), γ is a solution of the following system of first-order differential equations:

$$\gamma'(s) = T(s), \tag{44}$$

$$T'(s) = \kappa(s)N(s), \tag{45}$$

$$N'(s) = -\kappa(s)T(s). \tag{46}$$

Due to well-known results on the existence and uniqueness of solutions to linear ODEs [Nagle et al. 2004], there exists a unique solution of (44)–(46) with initial data

$$\gamma(0) = (0, 0), \quad T(0) = (1, 0), \quad N(0) = (0, 1). \tag{47}$$

It is easy to verify that such solution is given by

$$\gamma_0(s) = \left(\int_0^s \cos(\theta(t)) dt, \int_0^s \sin(\theta(t)) dt \right), \tag{48}$$

where

$$\theta(t) = \int_0^t \kappa(t) dt \tag{49}$$

is the tangential angle, i.e., the angle between $T = \gamma'_0(s) = (\cos(\theta(s)), \sin(\theta(s)))$ and a horizontal line. Denote a curve parametrized by γ_0 as \mathcal{C}_0 , and let \mathcal{C}_1 be another curve with Euclidean arc-length parametrization $\gamma_1(s)$, $s \in [0, L]$, such that $\kappa(s)$ is its Euclidean curvature. Let $T_1(0) = \gamma'_1(0)$ and $N_1(0) = \gamma''_1(0)$. Then there exists a unique special Euclidean transformation $g \in \text{SE}(2)$ which is a composition of translation by the vector $-\gamma_1(0)$, followed by the rotation

$$\begin{pmatrix} T_1(0) \\ N_1(0) \end{pmatrix}^{-1},$$

such that

$$g \cdot \gamma_1(0) = (0, 0), \quad g \cdot T_1 = (1, 0), \quad g \cdot N_1 = (0, 1).$$

Since κ and ds are invariant under rigid motions, it follows that the curve $g\mathcal{C}_1$ parametrized by $g\gamma_1$ satisfies (44)–(46) with the same initial data (47) and, therefore, $\mathcal{C}_0 = g\mathcal{C}_1$. \square

Formulas (48)–(49) allow us to construct a curve with prescribed Euclidean curvature. The following lemma gives a sufficient condition for a reconstructed curve to be closed. See Lemma 4 in [Musso and Nicolodi 2009] and Lemmas 1 and 2 in [Geiger and Kogan 2021].

Lemma 10. *Let $\kappa : \mathbb{R} \rightarrow \mathbb{R}$ be a periodic continuous function with minimum period ℓ if*

$$\frac{1}{2\pi} \int_0^\ell \kappa(s) ds = \frac{\xi}{m}, \quad (50)$$

where m and ξ are two relatively prime integers and $m > 1$, then, the corresponding unit speed parametrization γ , given by (48), defines a closed curve. The map γ has minimal period $m\ell$. The turning number of γ over the interval $[0, m\ell]$ is equal to ξ . If $\mathcal{C} = \text{Im}(\gamma)$ is simple, then $\xi = 1$ and m is the $\text{SE}(2)$ -symmetry index of \mathcal{C} .

Example 11. To illustrate the above lemma, consider the function

$$\kappa_1(s) = \sin(s) + \cos(s) + \frac{1}{3}. \quad (51)$$

Then $\frac{1}{2\pi} \int_0^{2\pi} \kappa_1(s) ds = \frac{1}{3}$ and the above lemma asserts that a curve with curvature function $\kappa_1(s)$ is closed with the $\text{SE}(3)$ -symmetry index of 3. Such curve, reconstructed using (48), is pictured in Figure 5, left.

On the other hand, consider

$$\kappa_2(s) = \sin(s) + \cos(s) + 1. \quad (52)$$

Then $\frac{1}{2\pi} \int_0^{2\pi} \kappa_2(s) ds = 1$ and the assumption $m > 1$ in Lemma 10 is not satisfied. Thus the lemma does not assert that a curve for which $\kappa_2(s)$ is the Euclidean curvature function is closed. In fact, the curve reconstructed using (48) is not closed, as we can see in Figure 5, right.

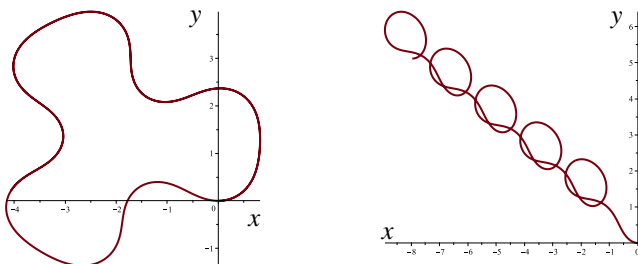


Figure 5. Left: a curve with Euclidean curvature (51). Right: a curve with Euclidean curvature (52). Lemma 10 guarantees that the left curve is closed, but does not make any assertion about the right curve.

Theorem 12 (Euclidean estimate). *Let C_1 and C_2 be two C^2 -smooth planar curves of the same Euclidean arc-length L . Assume $\kappa_1(s)$ and $\kappa_2(s)$, $s \in [0, L]$, are their respective Euclidean curvature functions. If $\|\kappa_1 - \kappa_2\|_{[0,L]} \leq \delta$, then there exists $g \in \text{SE}(2)$ such that*

$$d(C_1, g C_2) \leq \frac{\sqrt{2}}{2} \delta L^2, \tag{53}$$

where d is the Hausdorff distance.

Proof. Identifying \mathbb{R}^2 with \mathbb{C} and using Euler’s formula we may rewrite (48) as

$$\gamma(s) = \int_0^s e^{i\theta(t)} dt. \tag{54}$$

In what follows, we will use an important inequality, stating that a chord is shorter than the corresponding arc, illustrated in Figure 6:

$$|e^{i\theta_1} - e^{i\theta_2}| < |\theta_1 - \theta_2|. \tag{55}$$

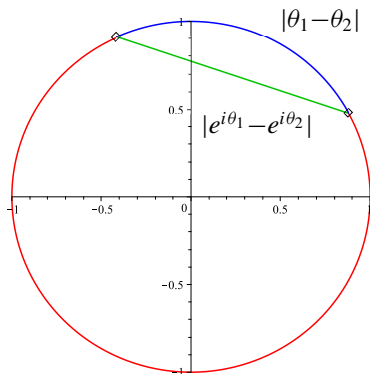


Figure 6. The length of the chord $|e^{i\theta_1} - e^{i\theta_2}|$ is shorter than the length of the arc $|\theta_1 - \theta_2|$.

For $j = 1, 2$, let $\gamma_j(s)$, $s \in [0, L]$, be the Euclidean arc length parametrization of the curve \mathcal{C}_j . Then $T_j(s) = \gamma_j'(s)$ and $N_j(s) = \gamma_j''(s)$ are the unit tangent and unit normal vectors, respectively, to \mathcal{C}_j . For $j = 1, 2$, there is a unique $g_j \in \text{SE}(2)$, such that

$$g_j \gamma_j(0) = (0, 0), \quad g_j T_j(0) = (1, 0), \quad g_j N_j(0) = (0, 1). \quad (56)$$

It follows from [Theorem 9](#) that $g_j \gamma_j(s) = \int_0^s e^{i\theta_j(t)} dt$ for $j = 1, 2$ and so

$$\begin{aligned} |g_1 \gamma_1(s) - g_2 \gamma_2(s)| &= \left| \int_0^s e^{i\theta_1(t)} dt - \int_0^s e^{i\theta_2(t)} dt \right| \stackrel{(a)}{\leq} \int_0^s |e^{i\theta_1(t)} - e^{i\theta_2(t)}| dt \\ &\stackrel{(b)}{<} \int_0^s |\theta_1(t) - \theta_2(t)| dt \stackrel{(c)}{=} \int_0^s \left| \int_0^t (\kappa_1(\tau) - \kappa_2(\tau)) d\tau \right| dt \\ &\stackrel{(*)}{\leq} \int_0^s \int_0^t |\kappa_1(\tau) - \kappa_2(\tau)| d\tau dt \\ &\stackrel{(d)}{\leq} \int_0^s \int_0^t \|\kappa_1 - \kappa_2\|_{[0, L]} d\tau dt \leq \int_0^s \int_0^t \delta d\tau dt = \frac{\delta s^2}{2}. \end{aligned} \quad (57)$$

Inequality (a) follows from properties of definite integrals. Inequality (b) follows from [\(55\)](#). Equality (c) follows from [\(49\)](#) and the properties of definite integrals. Inequality (d) follows from [\(25\)](#).

Let $g = g_1^{-1} g_2$. Then, using [\(31\)](#), [\(57\)](#) and the invariance of the Euclidean distance under the rigid motions, we have

$$\begin{aligned} d(\mathcal{C}_1, g \mathcal{C}_2) &\leq \sqrt{2} \|\gamma_1 - g \gamma_2\|_{[0, L]} = \sqrt{2} \sup_{s \in [0, L]} |\gamma_1(s) - g \gamma_2(s)| \\ &= \sqrt{2} \sup_{s \in [0, L]} |g_1 \gamma_1(s) - g_2 \gamma_2(s)| \leq \sqrt{2} \frac{\delta L^2}{2}. \quad \square \end{aligned}$$

If instead of the L^∞ -norm on the set of functions κ we use the L^1 -norm and require that $\int_0^L |\kappa_1(\tau) - \kappa_2(\tau)| d\tau \leq \delta$, then the third line of [\(57\)](#) implies the following result:

Theorem 13. *Let \mathcal{C}_1 and \mathcal{C}_2 be two C^2 -smooth planar curves of the same Euclidean arc-length L . Assume $\kappa_1(s)$ and $\kappa_2(s)$, $s \in [0, L]$, are their respective Euclidean curvature functions and*

$$\int_0^L |\kappa_1(\tau) - \kappa_2(\tau)| d\tau \leq \delta. \quad (58)$$

Then there exists $g \in \text{SE}(2)$ such that

$$d(\mathcal{C}_1, g \mathcal{C}_2) \leq \sqrt{2} \delta L.$$

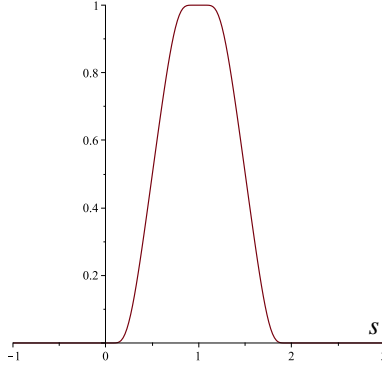


Figure 7. Bump function (59).

Proof. The proof proceeds along the same lines as the proof of [Theorem 12](#). However, inequality (*) in (57) combined with (58) implies

$$|g_1\gamma_1(s) - g_2\gamma_2(s)| < \int_0^s \delta dt = \delta s,$$

and so

$$d(\mathcal{C}_1, g\mathcal{C}_2) \leq \sqrt{2} \sup_{s \in [0, L]} |g_1\gamma_1(s) - g_2\gamma_2(s)| \leq \sqrt{2}\delta L. \quad \square$$

Example 14. To illustrate [Theorems 12](#) and [13](#), we consider a curve whose Euclidean curvature function is $\kappa(s) = \sin(s)$ and a family of curves obtained by some variations of $\kappa(s)$. To define these variations consider the smooth bump function

$$f(s) = \begin{cases} 0 & \text{if } s \leq 0, \\ e^{1/(1-s)}/(e^{1/s} + e^{1/1-s}) & \text{if } 0 < s < 1, \\ 1 & \text{if } s = 1, \\ e^{1/(s-1)}/(e^{1/(s-1)} + e^{1/(2-s)}) & \text{if } 1 < s < 2, \\ 0 & \text{if } s \geq 2, \end{cases} \quad (59)$$

shown in [Figure 7](#).

Next, for $n \in \mathbb{Z} \setminus \{0\}$, we define the functions

$$\kappa_n^*(s) = \sin(s) + \frac{2\pi}{n} f(s) \quad (60)$$

on the closed interval $[0, 2\pi]$ and let $\kappa_n(s)$ denote the periodic extension of κ_n^* to \mathbb{R} . We observe that, for any $L > 0$,

$$\|\kappa_n - \kappa\|_{[0, L]} \leq \left\| \frac{2\pi}{n} f(s) \right\|_{[0, 2\pi]} \leq \frac{2\pi}{|n|}.$$

As $|n| \rightarrow \infty$, for $n > 0$ and for $n < 0$, the sequence $\kappa_n(s)$ uniformly converges to $\sin(s)$. In [Figure 8](#), we show $\kappa_{10}(s)$, $\kappa_{20}(s)$, $\kappa_{40}(s)$, and $\kappa(s) = \sin(s)$ over their

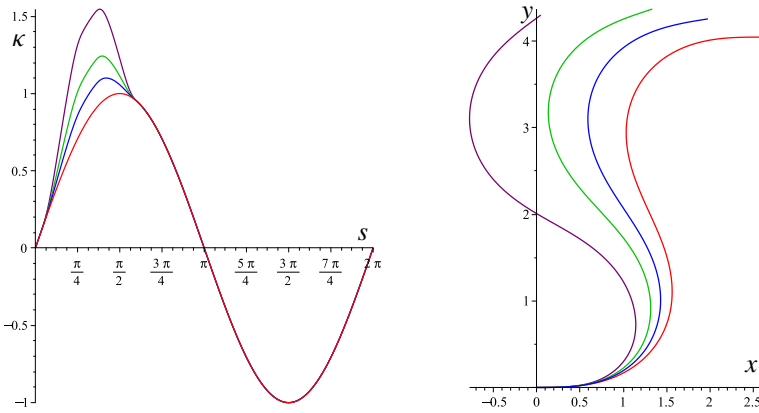


Figure 8. Left: $\kappa_{10}^*(s)$, $\kappa_{20}^*(s)$, $\kappa_{40}^*(s)$, given by (60) and $\kappa(s) = \sin(s)$, $s \in [0, 2\pi]$. Right: Curves \mathcal{C}_{10} , \mathcal{C}_{20} , \mathcal{C}_{40} , and \mathcal{C} , reconstructed from the Euclidean curvature functions shown on the left.

minimal period $[0, 2\pi]$, while in Figure 8, we show curves \mathcal{C}_{10} , \mathcal{C}_{20} , \mathcal{C}_{40} , and \mathcal{C} reconstructed from these curvatures with $s \in [0, 2\pi]$. We observe that the Hausdorff distance between \mathcal{C} and \mathcal{C}_n decreases as $|n|$ increases (and so $\delta = 2\pi/|n|$ decreases). At the same time, if we restrict s to an interval $[0, L]$, with $0 < L \leq 2\pi$, then for a fixed n , as L increases, the distance between \mathcal{C} and \mathcal{C}_n increases.

Since $\int_0^{2\pi} f(s) ds = 1$, we have $\int_0^{2\pi} \kappa_n(s) ds = 2\pi/n$. Therefore, by Lemma 10, for $n \in \mathbb{Z} \setminus \{-1, 0, 1\}$, a curve reconstructed from $\kappa_n(s)$ with $s \in [0, 2\pi n]$ is a closed curve with symmetry index $|n|$ and turning number 1. A curve reconstructed from $\kappa(s) = \sin(s)$ is, however, not closed. In Figure 9, we show the curves reconstructed from the extensions of curvatures in Figure 8 (left).

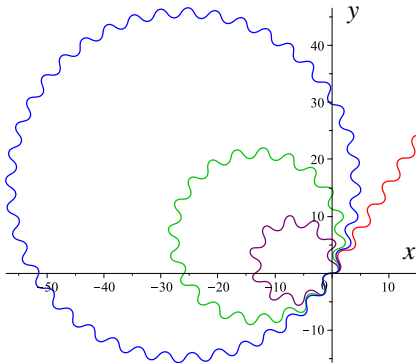
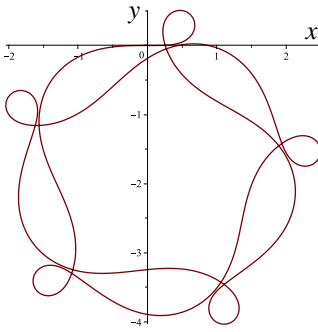
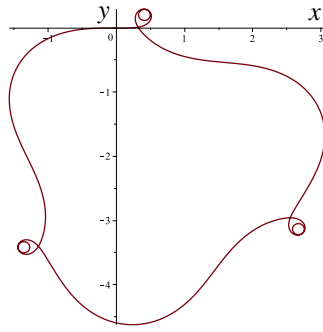


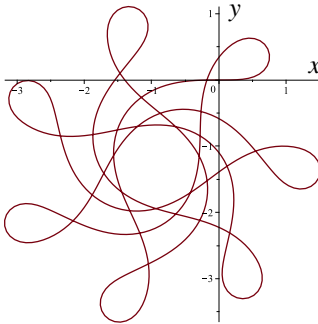
Figure 9. Closed curves reconstructed from periodic extensions $\kappa_{10}(s)$, $s \in [0, 20\pi]$, $\kappa_{20}(s)$, $s \in [0, 40\pi]$, $\kappa_{40}(s)$, $s \in [0, 80\pi]$ of κ_n^* , shown in Figure 8, and an open curve reconstructed from $\kappa(s) = \sin(s)$, $s \in [0, 12\pi]$.



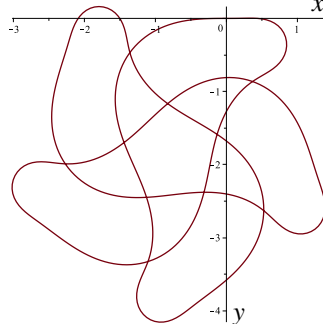
curvature $\kappa_{5/3}(s)$, $s \in [0, 10\pi]$
 turning number 3
 SE(2)-symmetry index 5.



curvature $\kappa_{3/5}$, $s \in [0, 6\pi]$
 turning number 5
 SE(2)-symmetry index 3.



curvature $\kappa_{7/3}$, $s \in [0, 14\pi]$
 turning number 3
 SE(2)-symmetry index 7.



curvature $\kappa_{-5/3}$, $s \in [0, 10\pi]$
 turning number -3
 SE(2)-symmetry index 5.

Figure 10. Closed curves reconstructed from $\kappa_r(s)$, where r is a rational number.

It is worth noting that if, in formula (60), we replace the integer n with a rational number $r = q/\xi$ such that $q \neq 1$ and ξ are relatively prime, then, by Lemma 10, a curve reconstructed from $\kappa_r(s)$, $s \in [0, 2\pi q]$, will be a closed curve with the SE(2)-symmetry index q and turning number ξ . See Figure 10 for examples.

4. Affine reconstruction

In this section, we start by showing how Picard iterations can be used to reconstruct a curve from its affine curvature. We proceed by proving some upper bounds related to Picard iterations and using them to estimate how close, relative to the Hausdorff distance, two curves can be brought together by a special affine transformation, provided the affine curvature functions of the curves are δ -close in the L^∞ -norm (Theorem 19).

Theorem 15 (affine reconstruction). *Let $\mu(\alpha)$ be a continuous function on an interval $[0, L]$. Then there is a unique, up to a special affine transformation, curve \mathcal{C} with the affine arc-length parametrization $\gamma(\alpha) = (x(\alpha), y(\alpha))$, $\alpha \in [0, L]$, such that $\mu(\alpha) = x''(\alpha)y'''(\alpha) - y''(\alpha)x'''(\alpha)$ is its affine curvature function.*

Proof. According to (14), (19) and (20), γ is a solution of the following system of first-order differential equations:

$$\gamma'(\alpha) = T(\alpha), \quad (61)$$

$$T'(\alpha) = N(\alpha), \quad (62)$$

$$N'(\alpha) = -\mu(\alpha)T(\alpha) \quad (63)$$

(equivalent to a third order ODE system of two decoupled equations $\gamma''' = -\mu\gamma'$). Due to well-known results on the existence and uniqueness of solutions to linear ODEs (see Theorems 5 and 6, Section 13.3 in [Nagle et al. 2004]), there exists a unique solution of (61)-(63) with the initial data

$$\gamma(0) = (0, 0), \quad T(0) = (1, 0), \quad N(0) = (0, 1). \quad (64)$$

Let $\gamma_0(\alpha)$ be such a solution parametrizing a curve \mathcal{C}_0 . Let \mathcal{C}_1 be another curve with the affine arc-length parametrization $\gamma_1(\alpha)$, $\alpha \in [0, L]$, such that $\mu(\alpha)$ is its affine curvature. Let $T_1 = \gamma_1'(0)$ and $N_1 = \gamma_1''(0)$. Then there exists a unique special affine transformation $g \in \text{SA}(2)$ which is a composition of translation by the vector $-\gamma_1(0)$, followed by the unimodular linear transformation

$$\begin{pmatrix} T_1(0) \\ N_1(0) \end{pmatrix}^{-1},$$

such that

$$g \cdot \gamma_1(0) = (0, 0), \quad g \cdot T_1 = (1, 0), \quad g \cdot N_1 = (0, 1).$$

Since μ and $d\alpha$ are $\text{SA}(2)$ -invariant, it follows that the curve $g\mathcal{C}_1$ parametrized by $g\gamma_1$ satisfies (61)–(63) with the same initial data (64) and, therefore, $\mathcal{C}_0 = g\mathcal{C}_1$. \square

We now consider computational aspects of reconstruction of a curve from its affine curvature. Once $T(\alpha)$ is known, γ can be reconstructed by integration, which can be done exactly or numerically depending on the complexity of $T(\alpha)$. To find $T(\alpha)$, one needs to solve the system (62)–(63).

When $\mu(\alpha)$ is a constant function, standard methods can be applied. In fact, as shown in [Guggenheimer 1977], if $\mu = 0$ then the reconstructed curve, with the initial conditions (64), is a parabola $\gamma = (\alpha, \alpha^2/2)$. When $\mu > 0$,

$$\gamma = \left(\frac{\sin(\sqrt{\mu}\alpha)}{\sqrt{\mu}}, -\frac{\cos(\sqrt{\mu}\alpha)}{\mu} \right)$$

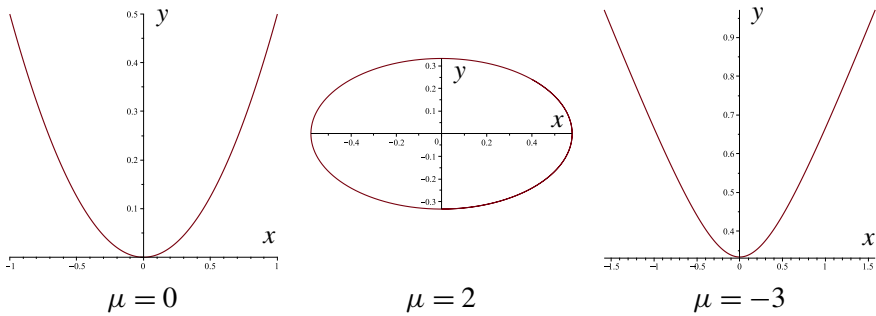


Figure 11. Examples of curves with constant special affine curvature functions.

is an ellipse. When $\mu < 0$,

$$\gamma = \left(\frac{\sinh(\sqrt{-\mu}\alpha)}{\sqrt{-\mu}}, -\frac{\cosh(\sqrt{-\mu}\alpha)}{\mu} \right)$$

is a hyperbola. See [Figure 11](#) for specific examples.

When μ is nonconstant but analytic one can use power series methods to find the solutions. The power series solutions for the case when μ is a monomial, $\mu = c\alpha^k$, are given in the [Appendix](#). For an arbitrary continuous function μ , we approximate $T(\alpha)$ by applying *Picard iterations* as follows.

As discussed in [Section 2.3](#), (62) and (63) are equivalent to the matrix equation (23), where

$$A(\alpha) = \begin{pmatrix} T(\alpha) \\ N(\alpha) \end{pmatrix}$$

is the affine frame matrix and $C(\alpha)$ is the affine Cartan matrix given by (24). The Picard iterations are defined as

$$\begin{aligned} A_0(\alpha) &= A_0, \\ A_n(\alpha) &= A_0 + \int_0^\alpha C(t)A_{n-1}(t) dt \quad \text{for } n > 0. \end{aligned} \quad (65)$$

It is well known that, on any interval, $[0, L]$ as $n \rightarrow \infty$ the sequence of $\{A_n(\alpha)\}$ uniformly converges to the unique matrix of continuous functions $A(\alpha)$ satisfying the integral equation

$$A(\alpha) = A_0 + \int_0^\alpha C(t)A(t) dt \quad (66)$$

and, therefore, the differential equation (23) with the initial value A_0 . A direct proof for the convergence of (65) to the solutions of (23) with the initial value A_0 , where C is an arbitrary continuous matrix, is given in [[Guggenheimer 1977](#), Lemma 2.12].

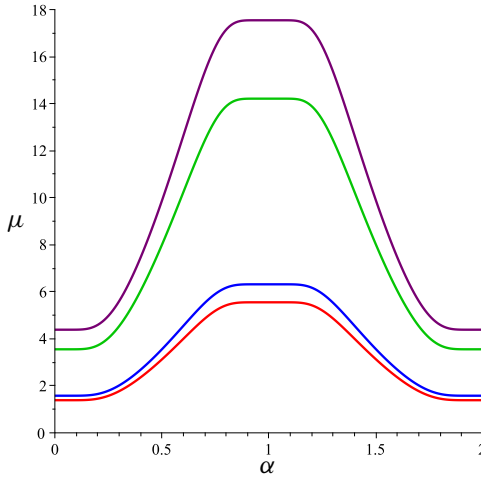


Figure 12. $\mu_{2/3}^*(\alpha)$, $\mu_{3/5}^*(\alpha)$, $\mu_{2/5}^*(\alpha)$, and $\mu_{3/8}^*(\alpha)$, $\alpha \in [0, 2]$ given by (67).

Example 16. We will briefly look at a few curves that are reconstructed from their affine curvatures. Recall the bump function $f(s)$ given by (59). Let

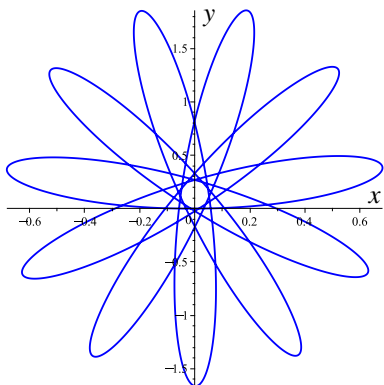
$$\mu_n^*(\alpha) = n^2 \pi^2 (f(\alpha) + 1)^2 \quad (67)$$

with domain $[0, 2]$ and let $\mu_n(\alpha)$ be the periodic extension of μ_n^* to \mathbb{R} ; see Figure 12. In Figure 13, we show approximations (using 200 Picard iterations) of curves with affine curvatures $\mu_{2/3}$, $\mu_{2/5}$, $\mu_{3/5}$, and $\mu_{3/8}$, initial conditions $\gamma(0) = (0, 0)$, and $A_0 = I$.

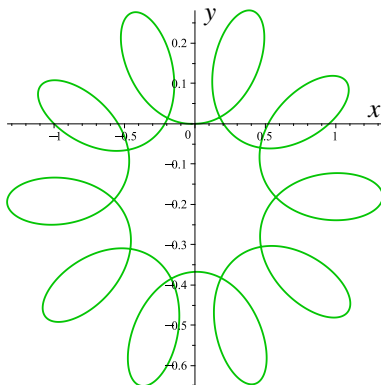
It is important to note that the affine analog to Lemma 10 is not valid. Indeed, it is shown, for instance, in Example 7.2 in [Kogan and Olver 2003], that in contrast with the Euclidean case, the total special affine curvature $\int \mu d\alpha$ of a closed curve is not topologically invariant, and thus it cannot be used to determine whether the curve is closed or open. Moreover, as remarked in [Verpoort 2011, p. 421], there does not exist a function of μ whose integral is a topological invariant. With this in mind, it is worth noting that the approximations of the curves with special affine curvatures $\mu_{2/5}$ and $\mu_{3/5}$ appear to be closed, while the curves with the affine curvature functions $\mu_{2/3}$ and $\mu_{3/8}$ show no sign that they would close if their domain was extended.

We now investigate the “closeness” of two curves reconstructed from “close” affine curvatures. We start by establishing certain upper bounds:

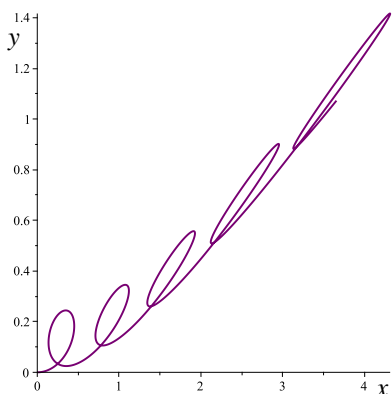
Lemma 17. *Assume that $\|C\|_{[0,L]} = \max\{1, \|\mu\|_{[0,L]}\} = c$. Let A_n be defined by the Picard iterations (65) and A be the limit of these iterations. Then for any $\alpha \in [0, L]$*



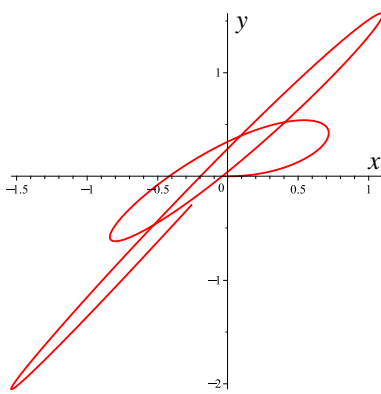
equiaffine curvature $\mu_{2/5}$ on $[0, 22]$



equiaffine curvature $\mu_{3/5}$ on $[0, 20]$



equiaffine curvature $\mu_{2/3}$ on $[0, 10]$



equiaffine curvature $\mu_{3/8}$ on $[0, 8]$

Figure 13. Approximations of curves, using 200 Picard iterations, reconstructed from periodic extensions of the affine curvature functions shown in Figure 12.

the following inequalities hold:

$$\langle A_n \rangle(\alpha) \leq \langle A_0 \rangle \sum_{i=0}^n \frac{(c\alpha)^i}{i!}, \tag{68}$$

$$\langle A \rangle(\alpha) \leq \langle A_0 \rangle e^{c\alpha}, \tag{69}$$

$$\langle A_n - A_{n-1} \rangle(\alpha) \leq \langle A_0 \rangle \frac{(c\alpha)^n}{n!}, \tag{70}$$

$$\langle A_n - A \rangle(\alpha) \leq \langle A_0 \rangle e^{c\alpha} \frac{(c\alpha)^{n+1}}{(n+1)!}. \tag{71}$$

Proof. (1) For $n = 0$, (68) states that $\langle A_0 \rangle \leq \langle A_0 \rangle$, which is trivially true. We proceed by induction. Assume that (68) holds for all $0 \leq k < n$. Then from (65),

(29) and the triangle inequality, we have

$$\langle A_n \rangle(\alpha) \leq \langle A_0 \rangle + \int_0^\alpha \langle C A_{n-1} \rangle(t) dt. \quad (72)$$

Note that, for any matrix

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix},$$

we have

$$CA = \begin{pmatrix} a_{21} & a_{22} \\ -\mu a_{11} & -\mu a_{12} \end{pmatrix},$$

and, therefore, since $c \geq \|\mu\|_{[0,L]}$ and $c \geq 1$,

$$\langle CA \rangle(t) \leq c \langle A \rangle(t). \quad (73)$$

Returning to (72) and using the inductive assumption, we then have

$$\begin{aligned} \langle A_n \rangle(\alpha) &\leq \langle A_0 \rangle + c \int_0^\alpha \langle A_{n-1} \rangle(t) dt \leq \langle A_0 \rangle + c \langle A_0 \rangle \sum_{i=0}^{n-1} \int_0^\alpha \frac{(ct)^i}{i!} dt \\ &= \langle A_0 \rangle \left(1 + c \sum_{i=0}^{n-1} \frac{c^i \alpha^{i+1}}{(i+1)!} \right) = \langle A_0 \rangle \left(1 + \sum_{i=1}^n \frac{c^i \alpha^i}{i!} \right) = \langle A_0 \rangle \sum_{i=0}^n \frac{(c\alpha)^i}{i!}. \end{aligned} \quad (74)$$

(2) To show (69), we use (39) and (68)

$$\langle A \rangle(\alpha) = \lim_{n \rightarrow \infty} \langle A_n \rangle(\alpha) \leq \langle A_0 \rangle \sum_{i=0}^{\infty} \frac{(c\alpha)^i}{i!} = \langle A_0 \rangle e^{c\alpha}. \quad (75)$$

(3) For $n = 1$, (70) states that $\langle A_1 - A_0 \rangle(\alpha) \leq \langle A_0 \rangle c\alpha$. This, indeed, holds because by (65) $A_1(\alpha) - A_0(\alpha) = \int_0^\alpha C(t)A_0 dt$, and so by (29) and (73)

$$\langle A_1 - A_0 \rangle(\alpha) \leq \int_0^\alpha \langle C(t)A_0 \rangle dt \leq \int_0^\alpha c \langle A_0 \rangle dt = \langle A_0 \rangle c\alpha.$$

We proceed by induction. Assume that (70) holds for all $1 \leq k < n$. By (65),

$$A_n(\alpha) - A_{n-1}(\alpha) = \int_0^\alpha C(t)(A_{n-1}(t) - A_{n-2}(t)) dt,$$

and then by (29), (73), and the inductive hypothesis

$$\langle A_n - A_{n-1} \rangle(\alpha) \leq \int_0^\alpha c \langle A_{n-1} - A_{n-2} \rangle(t) dt \leq \langle A_0 \rangle \int_0^\alpha c \frac{(ct)^{n-1}}{(n-1)!} dt = \langle A_0 \rangle \frac{(c\alpha)^n}{n!}.$$

(4) To show (71), we note that for any integer $j > 0$, due to the triangle inequality and (70), we have

$$\begin{aligned} \langle A_n - A \rangle(\alpha) &\leq \langle A_n - A_{n+1} \rangle(\alpha) + \langle A_{n+1} - A_{n+2} \rangle(\alpha) + \dots \\ &\quad + \langle A_{n+j-1} - A_{n+j} \rangle(\alpha) + \langle A_{n+j} - A \rangle(\alpha) \\ &\leq \langle A_0 \rangle \sum_{i=n+1}^{n+j} \frac{(c\alpha)^i}{i!} + \langle A_{n+j} - A \rangle(\alpha). \end{aligned} \quad (76)$$

Since $A_{n+j}(\alpha)$ converges to $A(\alpha)$ as $j \rightarrow \infty$, $\lim_{j \rightarrow \infty} \langle A_{n+j} - A \rangle(\alpha) = 0$, and so (76) implies

$$\langle A_n - A \rangle(\alpha) \leq \langle A_0 \rangle \sum_{i=n+1}^{\infty} \frac{(c\alpha)^i}{i!} = \langle A_0 \rangle \left(e^{c\alpha} - \sum_{i=0}^n \frac{(c\alpha)^i}{i!} \right). \quad (77)$$

Due to Taylor's remainder theorem, there exists $\alpha_0 \in [0, \alpha]$ such that

$$R_n = e^{c\alpha} - \sum_{i=0}^n \frac{(c\alpha)^i}{i!} = e^{c\alpha_0} \frac{(c\alpha)^{n+1}}{(n+1)!} \leq e^{c\alpha} \frac{(c\alpha)^{n+1}}{(n+1)!},$$

where the last inequality is true because $c > 0$ and so $e^{c\alpha}$ is an increasing function. \square

Next, we establish the bounds on the distance between two affine frames reconstructed from two δ -close (in the L^∞ norm) affine curvature functions. This result is consistent with a well-known ODE result on continuous dependence of the solutions of an ODE on its parameters (see, for instance, Theorem 10, Section 13.4 in [Nagle et al. 2004] and Theorem 3, Chapter 5 in [Birkhoff and Rota 1962]).

Proposition 18. *Let $\mu(\alpha)$ and $\tilde{\mu}(\alpha)$ be two continuous functions on the interval $[0, L]$ and let C and \tilde{C} be corresponding Cartan's matrices defined by (24). Let $\hat{c} = \max\{1, \|\mu\|_{[0, L]}, \|\tilde{\mu}\|_{[0, L]}\}$. Let A_n and \tilde{A}_n be defined by the Picard iterations (65) for the given matrices C and \tilde{C} , respectively, and A, \tilde{A} be the limits of these iterations. If $\|\mu - \tilde{\mu}\|_{[0, L]} \leq \delta$, then, for all $\alpha \in [0, L]$,*

$$\langle A_n - \tilde{A}_n \rangle(\alpha) \leq \langle A_0 \rangle \delta \alpha \sum_{i=0}^{n-1} \frac{(\hat{c}\alpha)^i}{i!} \quad \text{for } n > 0, \quad (78)$$

$$\langle A - \tilde{A} \rangle(\alpha) \leq \langle A_0 \rangle \delta \alpha e^{\hat{c}\alpha}. \quad (79)$$

Proof. (1) We first observe that, for all $\alpha \in [0, L]$, $\langle C - \tilde{C} \rangle(\alpha) = |\mu(\alpha) - \tilde{\mu}(\alpha)| < \delta$.

For $n = 1$, (78) states that $\langle A_1 - \tilde{A}_1 \rangle(\alpha) \leq \langle A_0 \rangle \delta \alpha$. This, indeed, holds because by (65), keeping in mind that $A_0(\alpha) = \tilde{A}_0(\alpha) = \langle A_0 \rangle$, we have

$$\begin{aligned} \langle A_1 - \tilde{A}_1 \rangle(\alpha) &\leq \int_0^\alpha \langle (C - \tilde{C})A_0 \rangle(t) dt \leq \int_0^\alpha \langle A_0 \rangle |\tilde{\mu}(t) - \mu(t)| dt \\ &\leq \langle A_0 \rangle \int_0^\alpha \delta dt = \langle A_0 \rangle \delta \alpha. \end{aligned}$$

We proceed by induction. Assume that (78) holds for all $1 \leq k < n$. Then

$$\begin{aligned}
\langle A_n - \tilde{A}_n \rangle(\alpha) &\stackrel{(a)}{\leq} \int_0^\alpha \langle C A_{n-1} - \tilde{C} \tilde{A}_{n-1} \rangle(t) dt \\
&= \int_0^\alpha \langle C A_{n-1} - C \tilde{A}_{n-1} + C \tilde{A}_{n-1} - \tilde{C} \tilde{A}_{n-1} \rangle(t) dt \\
&\stackrel{(b)}{\leq} \int_0^\alpha \hat{c} \langle A_{n-1} - \tilde{A}_{n-1} \rangle(t) dt + \int_0^\alpha \delta \langle \tilde{A}_{n-1} \rangle(t) dt \\
&\stackrel{(c)}{\leq} \int_0^\alpha \hat{c} \langle A_0 \rangle \delta t \sum_{i=0}^{n-2} \frac{(\hat{c}t)^i}{i!} dt + \int_0^\alpha \langle A_0 \rangle \delta \sum_{i=0}^{n-1} \frac{(\hat{c}t)^i}{i!} dt \\
&= \langle A_0 \rangle \delta \left(\hat{c} \sum_{i=0}^{n-2} \frac{\hat{c}^i \alpha^{i+2}}{i!(i+2)} + \sum_{i=0}^{n-1} \frac{\hat{c}^i \alpha^{i+1}}{(i+1)!} \right) \\
&= \langle A_0 \rangle \delta \left(\sum_{i=1}^{n-1} \frac{\hat{c}^i \alpha^{i+1}}{(i-1)!(i+1)} + \sum_{i=0}^{n-1} \frac{\hat{c}^i \alpha^{i+1}}{(i+1)!} \right) \\
&= \langle A_0 \rangle \delta \left(\alpha + \sum_{i=1}^{n-1} \hat{c}^i \alpha^{i+1} \left(\frac{1}{(i-1)!(i+1)} + \frac{1}{(i+1)!} \right) \right) \\
&= \langle A_0 \rangle \delta \alpha \left(1 + \sum_{i=1}^{n-1} \hat{c}^i \alpha^i \frac{1}{i!} \right) = \langle A_0 \rangle \delta \alpha \sum_{i=0}^{n-1} \frac{(\hat{c}\alpha)^i}{i!}, \tag{80}
\end{aligned}$$

where, for inequality (a) we used (65), (29), and the triangle inequality. In inequality (b) we use (73) and the triangle inequality, and in inequality (c) we use the inductive assumption and (68).

(2) To show (79), we note that since $A_n(\alpha)$ and $\tilde{A}_n(\alpha)$ converge to $A(\alpha)$ and $\tilde{A}(\alpha)$, respectively, as $n \rightarrow \infty$, then by (39), $\lim_{n \rightarrow \infty} \langle A_n - \tilde{A}_n \rangle(\alpha) = \langle A - \tilde{A} \rangle(\alpha)$, and so taking the limit of both sides in the inequality (78) as $n \rightarrow \infty$, we obtain (79). \square

In the next theorem, we establish an upper bound on how close (in the Hausdorff distance) two curves with δ -close (in the L^∞ -norm) affine curvature functions can be brought together by a special affine transformation.

Theorem 19 (affine estimate). *Let C_1 and C_2 be two C^3 -smooth planar curves of the same affine arc-length L . Assume $\mu_1(\alpha)$ and $\mu_2(\alpha)$, $\alpha \in [0, L]$, are their respective affine curvature functions. Assume further that C_2 satisfies the initial conditions (64).⁹ If $\|\mu_1 - \mu_2\|_{[0, L]} \leq \delta$ and $\hat{c} = \max\{1, \|\mu_1\|_{[0, L]}, \|\mu_2\|_{[0, L]}\}$, then*

⁹If we omit this assumption, then the right-hand side of (83) must be multiplied by $\langle A_2(0) \rangle$ according to (79), and so the right-hand side of (81) must be multiplied by $\langle A_2(0) \rangle$, as well.

there is $g \in \text{SA}(2)$, such that

$$d(g\mathcal{C}_1, \mathcal{C}_2) \leq \sqrt{2} \frac{\delta L}{\hat{c}} (e^{\hat{c}L} - 1), \quad (81)$$

where d is the Hausdorff distance.

Proof. For $i = 1, 2$, let $\gamma_i(\alpha)$, $\alpha \in [0, L]$, be the affine-arc length parametrization of \mathcal{C}_i , while $T_i(\alpha) = \gamma_i'(\alpha)$ and $N_i(\alpha) = \gamma_i''(\alpha)$ are the affine frame vectors along the corresponding curves. Then, there is a unique $g \in \text{SA}(2)$ such that

$$g\gamma_1(0) = \gamma_2(0) = (0, 0), \quad gT_1(0) = T_2(0) = (1, 0), \quad gN_1(0) = N_2(0) = (0, 1). \quad (82)$$

Due to the $\text{SA}(2)$ -invariance of the affine curvature function, the curve $g\mathcal{C}_1$ parametrized by $g\gamma_1(\alpha)$ has affine curvature function $\mu_1(\alpha)$. It follows from [Theorem 15](#) that $g\gamma_1(\alpha)$ is the unique solution of [\(61\)–\(63\)](#), with $\mu(\alpha) = \mu_1(\alpha)$ and $\gamma_2(\alpha)$ is the unique solution of [\(61\)–\(63\)](#), with $\mu(\alpha) = \mu_2(\alpha)$, both with initial conditions [\(82\)](#).

Denote the affine frame of $g\mathcal{C}_1$ by

$$A(\alpha) = \begin{pmatrix} gT_1(\alpha) \\ gN_1(\alpha) \end{pmatrix}$$

and the affine frame of \mathcal{C}_2 by

$$\tilde{A}(\alpha) = \begin{pmatrix} T_2(\alpha) \\ N_2(\alpha) \end{pmatrix}.$$

Then

$$\langle gT_1 - T_2 \rangle(\alpha) \leq \langle A - \tilde{A} \rangle(\alpha) \leq \delta \alpha e^{\hat{c}\alpha}, \quad (83)$$

where the first inequality is due to the definition of $\langle \cdot \rangle$ and the second inequality is due to [\(79\)](#). Since $g\gamma_1(\alpha) = \int_0^\alpha gT_1(t) dt + T_0$ and $\gamma_2(\alpha) = \int_0^\alpha T_2(t) dt + T_0$, we have, for all $\alpha \in [0, L]$,

$$\begin{aligned} \langle g\gamma_1 - \gamma_2 \rangle(\alpha) &\leq \int_0^\alpha \langle gT_1 - T_2 \rangle(t) dt \leq \int_0^\alpha \delta t e^{\hat{c}t} dt \\ &\leq \int_0^\alpha \delta L e^{\hat{c}t} dt = \frac{\delta L}{\hat{c}} (e^{\hat{c}\alpha} - 1). \end{aligned} \quad (84)$$

It then follows from [\(31\)](#) and [\(84\)](#) that

$$\begin{aligned} d(g\mathcal{C}_1, \mathcal{C}_2) &\leq \sqrt{2} \|g\gamma_1 - \gamma_2\|_{[0, L]} \\ &= \sqrt{2} \max_{\alpha \in [0, L]} \langle g\gamma_1, -\gamma_2 \rangle(\alpha) \leq \sqrt{2} \frac{\delta L}{\hat{c}} (e^{\hat{c}L} - 1). \quad \square \end{aligned}$$

5. Conclusion

We considered practical aspects of reconstructing planar curves with prescribed Euclidean or affine curvatures. An immediate extension of the current work would be the reconstruction of planar curves with prescribed projective curvatures, and obtaining distance estimates between curves, modulo a projective transformation,

compared to the distance between the projective curvatures. Indeed, the projective group, containing both the special Euclidean and the special affine groups, plays a crucial role in computer vision (see, for, instance [Faugeras and Luong 2001; Hartley and Zisserman 2004]). Extension to space curves is another direction with immediate applications.

By considering specific group actions, we take advantage of their specific structural properties and obtain results that can be immediately suitable for applications. However, the generalization of the moving frame method from [Fels and Olver 1999; Olver 2015] allows us, in principle, to generalize our approach to an action of an arbitrary Lie group G on curves (or even on higher-dimensional submanifolds) in some ambient metric space. In such a generalization, a G -equivariant moving frame map from the corresponding jet space to the group G plays the role of the G -frame matrix A , appearing in this paper, and we will seek an estimate of how close two submanifolds can be brought together by an element of G , provided the Maurer–Cartan invariants for the G -action are sufficiently close.

In this paper, we used the Hausdorff distance between curves when considering both the $SE(2)$ - and the $SA(2)$ -actions on the plane. However, while the Hausdorff distance is $SE(2)$ -invariant, it is *not* $SA(2)$ -invariant and so it does not provide a natural measure of distance between two curves in the special affine case. In a future work, it is worthwhile to explore $SA(2)$ -invariant alternatives for measuring distance between two curves, based, for instance, on the area of the region between two curves. In the generalization to other group actions, the goal would be to consider a G -invariant distance between two submanifolds.

6. Appendix

If a given special affine curvature is analytic, it is possible to reconstruct the corresponding curve by looking for power series solutions to the second-order ODE system $T_{\alpha\alpha} = -\mu(\alpha)T$. We illustrate this approach by reconstructing curves whose special affine curvatures are of the form $\mu(\alpha) = c\alpha^k$ for $c \in \mathbb{R}$ and $k \in \mathbb{N}$.

Proposition 20. *For $c \in \mathbb{R}$, $k \in \mathbb{N}$ and $T_0, N_0 \in \mathbb{R}^2$ such that $\det[T_0, N_0] = 1$, let C be the curve whose affine curvature function is $\mu(\alpha) = c\alpha^k$, the initial affine tangent vector is T_0 and the initial affine normal is N_0 . Then the affine tangent vector along C is given by the absolutely convergent power series*

$$T(\alpha) = -T_0\Gamma\left(-\frac{1}{K}\right) \sum_{i=1}^{\infty} \frac{(-c)^i \alpha^{Ki}}{i! K^{2i+1} \Gamma(-1/K + i + 1)} + N_0\Gamma\left(\frac{1}{K}\right) \sum_{i=1}^{\infty} \frac{(-c)^i \alpha^{K(i+1)}}{i! K^{2i+1} \Gamma(1/K + i + 1)}, \quad (85)$$

where $K = k + 2$ and Γ denotes the gamma function.

Proof. We first represent the tangent vector $T(\alpha)$ by

$$T = b_0 + b_1\alpha + b_2\alpha^2 + b_3\alpha^3 + \cdots + b_n\alpha^n \cdots, \quad (86)$$

where each b_i is a vector coefficient, with $b_0 = T_0$ and $b_1 = N_0$ being the initial values of the affine tangent and the affine normal, respectively.

We write out the power series representation of $T_{\alpha\alpha}$ and $-\alpha^k T$:

$$T_{\alpha\alpha} = 0b_0 + 0b_1\alpha + 2b_2 + 3 \cdot 2b_3\alpha + \cdots + n(n-1)b_n\alpha^{n-2} \cdots, \quad (87)$$

$$-\alpha^k T = -cb_0\alpha^k - cb_1\alpha^{(k+1)} - cb_2\alpha^{(k+2)} - cb_3\alpha^{(k+3)} - \cdots - cb_n\alpha^{(k+n)} - \cdots. \quad (88)$$

The equality of these two power series implies the equality of vector coefficients with the same powers of α in two series. It follows that

$$b_n = \begin{cases} 0 & \text{when } 2 \leq n \leq k+1, \\ -cb_{n-(k+2)}/(n(n-1)) & \text{when } n \geq k+2. \end{cases} \quad (89)$$

Then b_{k+2} and b_{k+3} can be written in terms of b_0 and b_1 :

$$b_{k+2} = -\frac{cb_0}{(k+2)(k+1)}, \quad b_{k+3} = -\frac{cb_1}{(k+3)(k+2)}. \quad (90)$$

Using induction, when $n \bmod (k+2) = 0$, we can express b_n in terms of b_0 , when $n \bmod (k+2) = 1$, we can express b_n in terms of b_1 , and we can show that otherwise $b_n = 0$. This gives us the power series representation for T in terms of b_0 and b_1 as

$$T(\alpha) = b_0 + b_1\alpha + \sum_{i=1}^{\infty} (-c\alpha^{k+2})^i \left(\left(\prod_{j=1}^i \frac{1}{j(k+2)(j(k+2)-1)} \right) b_0 + \left(\prod_{j=1}^i \frac{1}{j(k+2)(j(k+2)+1)} \right) b_1\alpha \right). \quad (91)$$

We can split (91) into two parts:

$$\begin{aligned} B_0 &= b_0 \sum_{i=1}^{\infty} \left(\prod_{j=1}^i \frac{1}{j(k+2)(j(k+2)-1)} \right) (-c\alpha^{k+2})^i \\ &= b_0 \sum_{i=1}^{\infty} \left(\prod_{j=1}^i \frac{1}{(j(k+2)-1)} \right) \frac{(-c\alpha^{k+2})^i}{i!(k+2)^i} = b_0 \sum_{i=1}^{\infty} \Psi_{-}(K, i) \frac{(-c\alpha^K)^i}{i!K^i} \end{aligned} \quad (92)$$

and

$$\begin{aligned} B_1 &= b_1 \sum_{i=1}^{\infty} \left(\prod_{j=1}^i \frac{1}{j(k+2)(j(k+2)+1)} \right) (-c\alpha^{k+2})^i \alpha \\ &= b_1 \sum_{i=1}^{\infty} \left(\prod_{j=1}^i \frac{1}{(j(k+2)+1)} \right) \frac{(-c\alpha^{k+2})^i \alpha}{i!(k+2)^i} = b_1 \sum_{i=1}^{\infty} \Psi_{+}(K, i) \frac{(-c\alpha^K)^i \alpha}{i!K^i}, \end{aligned} \quad (93)$$

where $K = k + 2$ and

$$\Psi_{-}(K, i) = \prod_{j=1}^i \frac{1}{(jK - 1)} = \frac{1}{K^i} \frac{1}{\prod_{j=1}^i (j - 1/K)}, \quad (94)$$

$$\Psi_{+}(K, i) = \prod_{j=1}^i \frac{1}{(jK + 1)} = \frac{1}{K^i} \frac{1}{\prod_{j=1}^i (j + 1/K)}. \quad (95)$$

These functions involve what is called *rising factorials*, defined by

$$z^{\bar{i}} := z(z + 1) \cdots (z + i - 1) = \prod_{j=0}^{i-1} (z + j).$$

Rising factorials can be expressed in terms of Γ functions, $\Gamma(z) = \int_0^{\infty} x^{z-1} e^{-x} dx$, as

$$z^{\bar{i}} = \frac{\Gamma(z + i)}{\Gamma(z)}.$$

For details see formulas (5.84), (5.85) and (5.89) on pp. 210–211 of [Graham et al. 1994]. Since

$$\prod_{j=1}^i \left(j - \frac{1}{K} \right) = -K \left(-\frac{1}{K} \right)^{\bar{i}+1} = -K \frac{\Gamma(-1/K + i + 1)}{\Gamma(-1/K)}, \quad (96)$$

$$\prod_{j=1}^i \left(j + \frac{1}{K} \right) = K \left(\frac{1}{K} \right)^{\bar{i}+1} = K \frac{\Gamma(1/K + i + 1)}{\Gamma(1/K)}, \quad (97)$$

we can rewrite (94)–(95) using Γ functions:

$$\Psi_{-}(K, i) = \prod_{j=1}^i \frac{1}{(jK - 1)} = -\frac{1}{K^{i+1}} \frac{\Gamma(-1/K)}{\Gamma(-1/K + i + 1)}, \quad (98)$$

$$\Psi_{+}(K, i) = \prod_{j=1}^i \frac{1}{(jK + 1)} = \frac{1}{K^{i+1}} \frac{\Gamma(1/K)}{\Gamma(1/K + i + 1)}. \quad (99)$$

Therefore,

$$\begin{aligned} B_0(\alpha) &= b_0 \sum_{i=1}^{\infty} \Psi_{-}(K, i) \frac{(-c\alpha^K)^i}{i! K^i} \\ &= b_0 \sum_{i=1}^{\infty} \left(-\frac{1}{K^{i+1}} \frac{\Gamma(-1/K)}{\Gamma(-1/K + i + 1)} \right) \frac{(-c\alpha^K)^i}{i! K^i} \\ &= -b_0 \Gamma\left(-\frac{1}{K}\right) \sum_{i=1}^{\infty} \frac{1}{\Gamma(-1/K + i + 1)} \frac{(-c\alpha^K)^i}{i! K^{2i+1}}, \end{aligned} \quad (100)$$

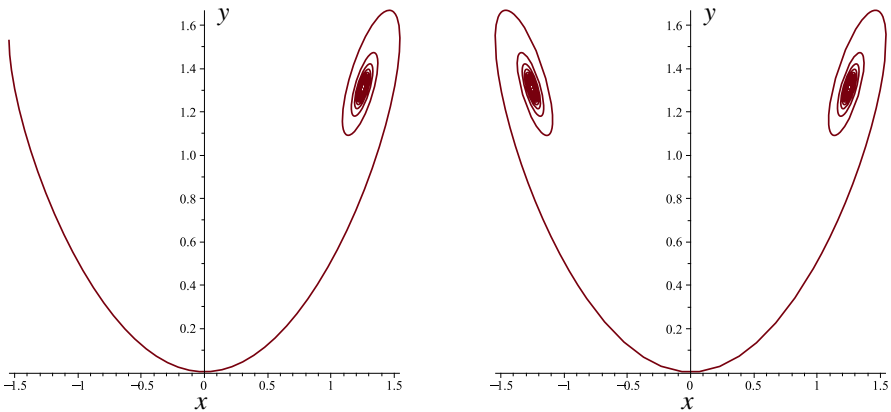


Figure 14. Curves with special affine curvature $\mu(\alpha) = \alpha$ (left) and $\mu(\alpha) = \alpha^2$ (right).

and

$$\begin{aligned}
 B_1(\alpha) &= b_1 \sum_{i=1}^{\infty} \Psi_+(K, i) \frac{(-c\alpha^K)^i \alpha}{i! K^i} \\
 &= b_1 \sum_{i=1}^{\infty} \left(\frac{1}{K^{i+1}} \frac{\Gamma(1/K)}{\Gamma(1/K + i + 1)} \right) \frac{(-c\alpha^K)^i \alpha}{i! K^i} \\
 &= b_1 \Gamma\left(\frac{1}{K}\right) \sum_{i=1}^{\infty} \frac{1}{\Gamma(1/K + i + 1)} \frac{(-c\alpha^K)^i \alpha}{i! K^{2i+1}}. \tag{101}
 \end{aligned}$$

Convergence of series (91) for all α follows from a general known result [Tenenbaum and Pollard 1963, Theorem 39.22, p. 560]. Directly, absolute convergence of subseries (92) and (93) can be verified by the ratio test, implying absolute convergence of series (91). \square

The power series for the affine arc-length parametrization $\gamma(\alpha)$ is obtained by integrating the series $T(\alpha)$. See Figure 14 for reconstructions of curves with curvatures $\mu(\alpha) = \alpha$ and $\mu(\alpha) = \alpha^2$ respectively.

Remark 21. The system $T_{\alpha\alpha} = -c\alpha^k T$ consists of two decoupled equations of the type $u''(\alpha) = -c\alpha^k u(\alpha)$, whose general solution in terms of the Bessel functions, can be found, for instance, in Section 14.1.2, subsection 7, number 3 of [Polyanin and Zaitsev 2012]. The Bessel functions can be expanded into power series involving the gamma function, recovering series (85). The advantage of formula (85) is in its explicit dependence on the initial vectors T_0 and N_0 . In addition, our direct proof illustrates how the power series approach can be applied for other analytic affine curvatures $\mu(\alpha)$.

Acknowledgements

This work was performed during the REU 2020 program at the North Carolina State University (NCSU) and was supported by the Department of Mathematics at NCSU, the NSA grant H98230-20-1-0259, and the NSF grant DMS 2051010. At the time when the project was performed, Jose Agudelo was an undergraduate student at North Dakota State University, Brooke Dippold was an undergraduate student at Longwood University, Ian Klein was an undergraduate student at Carleton College, Alex Kokot was an undergraduate student at the University of Notre Dame, and Eric Geiger was a graduate student at NCSU. Irina Kogan is a Professor of Mathematics at NCSU. The project was mentored by Eric Geiger and Irina Kogan. A poster based on this project received an honorable mention at JMM 2021.

References

- [Ames et al. 2002] A. D. Ames, J. A. Jalkio, and C. Shakiban, “Three-dimensional object recognition using invariant Euclidean signature curves”, pp. 13–23 in *Analysis, combinatorics and computing* (Dalian, China, 2000), edited by T.-X. He et al., Nova Sci., Hauppauge, NY, 2002. [MR](#) [Zbl](#)
- [Birkhoff and Rota 1962] G. Birkhoff and G.-C. Rota, *Ordinary differential equations*, Ginn and Company, Boston, 1962. [MR](#) [Zbl](#)
- [Calabi et al. 1998] E. Calabi, P. Olver, C. Shakiban, A. Tannenbaum, and S. Haker, “Differential and numerically invariant signature curves applied to object recognition”, *Int. J. Comput. Vis.* **26** (1998), 107–135.
- [Faugeras 1994] O. Faugeras, “Cartan’s moving frame method and its application to the geometry and evolution of curves in the Euclidean, affine and projective planes”, pp. 9–46 in *Applications of invariance in computer vision*, edited by J. L. Mundy et al., Lecture Notes in Computer Science **825**, Springer, 1994.
- [Faugeras and Luong 2001] O. Faugeras and Q.-T. Luong, *The geometry of multiple images: the laws that govern the formation of multiple images of a scene and some of their applications*, MIT Press, Cambridge, MA, 2001. [MR](#) [Zbl](#)
- [Fels and Olver 1999] M. Fels and P. J. Olver, “Moving coframes, II: Regularization and theoretical foundations”, *Acta Appl. Math.* **55** (1999), 127–208. [Zbl](#)
- [Flash and Handzel 2007] T. Flash and A. A. Handzel, “Affine differential geometry analysis of human arm movements”, *Biol. Cybernet.* **96**:6 (2007), 577–601. [MR](#) [Zbl](#)
- [Geiger and Kogan 2021] E. Geiger and I. A. Kogan, “Non-congruent non-degenerate curves with identical signatures”, *J. Math. Imaging Vision* **63**:5 (2021), 601–625. [MR](#) [Zbl](#)
- [Goldberg et al. 2004] D. Goldberg, C. Malon, and M. Bern, “A global approach to automatic solution of jigsaw puzzles”, *Comput. Geom.* **28**:2-3 (2004), 165–174. [MR](#) [Zbl](#)
- [Golubitsky et al. 2010] O. Golubitsky, V. Mazalov, and S. M. Watt, “Toward affine recognition of handwritten mathematical characters”, pp. 35–42 in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems* (Boston, MA, 2010), Association for Computing Machinery, New York, 2010.
- [Graham et al. 1994] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete mathematics: a foundation for computer science*, 2nd ed., Addison-Wesley, Reading, MA, 1994. [MR](#) [Zbl](#)

- [Guggenheimer 1977] H. W. Guggenheimer, *Differential geometry*, 2nd ed., Dover, New York, 1977. [MR](#) [Zbl](#)
- [Hartley and Zisserman 2004] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*, 2nd ed., Cambridge University Press, 2004. [MR](#) [Zbl](#)
- [Hawkins 1984] T. Hawkins, “The Erlanger Programm of Felix Klein: reflections on its place in the history of mathematics”, *Historia Math.* **11**:4 (1984), 442–470. [MR](#) [Zbl](#)
- [Hoff and Olver 2014] D. J. Hoff and P. J. Olver, “Automatic solution of jigsaw puzzles”, *J. Math. Imaging Vision* **49**:1 (2014), 234–250. [MR](#) [Zbl](#)
- [Kogan 2003] I. A. Kogan, “Two algorithms for a moving frame construction”, *Canad. J. Math.* **55**:2 (2003), 266–291. [MR](#) [Zbl](#)
- [Kogan and Olver 2003] I. A. Kogan and P. J. Olver, “Invariant Euler–Lagrange equations and the invariant variational bicomplex”, *Acta Appl. Math.* **76**:2 (2003), 137–193. [MR](#) [Zbl](#)
- [Musso and Nicolodi 2009] E. Musso and L. Nicolodi, “Invariant signatures of closed planar curves”, *J. Math. Imaging Vis.* **35**:1 (2009), 68–85. [MR](#) [Zbl](#)
- [Nagle et al. 2004] R. K. Nagle, E. B. Saff, and A. D. Snider, *Fundamentals of differential equations*, 6th ed., Pearson Addison-Wesley, Boston, 2004. [Zbl](#)
- [Olver 2015] P. J. Olver, “Modern developments in the theory and applications of moving frames”, pp. 14–50 in *Impact150 Stories*, edited by J. Greenlees, London Math. Soc., 2015.
- [Polyanin and Zaitsev 2012] A. D. Polyanin and V. F. Zaitsev, *Handbook of nonlinear partial differential equations*, 2nd ed., CRC Press, Boca Raton, FL, 2012. [MR](#)
- [Tenenbaum and Pollard 1963] M. Tenenbaum and H. Pollard, *Ordinary differential equations: an elementary textbook for students of mathematics, engineering, and the sciences*, Harper and Row, New York, 1963. [Zbl](#)
- [Verpoort 2011] S. Verpoort, “Curvature functionals for curves in the equi-affine plane”, *Czechoslovak Math. J.* **61**:2 (2011), 419–435. [MR](#) [Zbl](#)
- [Wolfson et al. 1988] H. Wolfson, E. Schonberg, A. Kalvin, and Y. Lamdan, “Solving jigsaw puzzles by computer”, *Ann. Oper. Res.* **12**:1-4 (1988), 51–64. [MR](#)

Received: 2022-02-13

Revised: 2023-01-28

Accepted: 2023-01-30

joseagudelo@unm.edu*University of New Mexico, Albuquerque, NM, United States*brookedippold1@gmail.com*Longwood University, Farmville, VA, United States*iklein@ncsu.edu*North Carolina State University, Raleigh, NC, United States*akokot@uw.edu*University of Washington, Seattle, WA, United States*eric.geiger@baruch.cuny.edu*Baruch College, CUNY, New York, NY, United States*iakogan@ncsu.edu*North Carolina State University, Raleigh, NC, United States*

Biological models, monotonicity methods, and solving a discrete reaction-diffusion equation

Carson Rodriguez and Stephen B. Robinson

(Communicated by Suzanne Lenhart)

The problem of interest is a discrete reaction-diffusion equation motivated by models in population biology. We consider

$$Au + \phi(u) + \lambda f(u) = 0 \quad \text{for } u \in \mathbb{R}^{n-1},$$

where $n \geq 3$, A is an $(n-1) \times (n-1)$ matrix such that $-A$ is monotone, $\phi : \mathbb{R}^{n-1} \rightarrow \mathbb{R}^{n-1}$ and $f : \mathbb{R}^{n-1} \rightarrow \mathbb{R}^{n-1}$ are smooth functions, and λ is a positive real constant. Of particular interest is the case where A is the discrete Laplacian and f is the vector-valued logistic function. The function $\phi(u)$ will encode boundary conditions. Our primary goal is to establish the existence of nonnegative solutions for several interesting choices of ϕ . For each choice we use monotonicity methods to find nonnegative solutions for appropriate ranges of λ .

1. Introduction

The problem of interest is a discrete reaction-diffusion equation motivated by models in population biology. We consider

$$Au + \phi(u) + \lambda f(u) = 0 \quad \text{for } u \in \mathbb{R}^{n-1}, \quad (1)$$

where $n \geq 3$, A is an $(n-1) \times (n-1)$ matrix such that $-A$ is monotone, $\phi : \mathbb{R}^{n-1} \rightarrow \mathbb{R}^{n-1}$ and $f : \mathbb{R}^{n-1} \rightarrow \mathbb{R}^{n-1}$ are smooth functions, and λ is a positive real constant. Of particular interest are the discrete Laplacian

$$A = \begin{bmatrix} -2 & 1 & 0 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -2 & 1 \\ 0 & \cdots & \cdots & 0 & 1 & -2 \end{bmatrix},$$

MSC2020: primary 39A27; secondary 39A12, 39A60.

Keywords: discrete nonlinear boundary value problem, reaction-diffusion equation, population model, sub- and supersolutions, density-dependent boundary conditions.

the vector-valued logistic function

$$f(u) = u(1-u) := \begin{bmatrix} u_1(1-u_1) \\ u_2(1-u_2) \\ \vdots \\ u_{n-1}(1-u_{n-1}) \end{bmatrix},$$

and the function

$$\phi(u) := \begin{bmatrix} \varphi(u) \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

which encodes boundary conditions. Our primary goal is to establish the existence of nonnegative solutions for several interesting choices of ϕ . For each choice we use monotonicity methods to find nonnegative solutions for appropriate ranges of λ .

1.1. Motivation. Begin by considering a continuous model for population growth that is often encountered in elementary ordinary differential equations. Let u represent total population measured as a percentage of carrying capacity, so $u = 1$ would mean that the population is using up all available resources, no more and no less. One way to model how the total population will change over time is the logistic equation

$$\frac{du}{dt} = \lambda u(1-u), \quad u(0) = u_0, \quad (2)$$

where λ represents a growth rate and u_0 is the initial population. Since it is impossible to have a negative population, we assume $u \geq 0$. In elementary ordinary differential equations we learn that the equilibrium solutions are found by solving

$$0 = \lambda u(1-u), \quad (3)$$

i.e., $du/dt = 0$, so the population does not change over time. We get $u \equiv 1$ or $u \equiv 0$. We also learn that if $u_0 > 0$ then $\lim_{t \rightarrow \infty} u(t) = 1$; i.e., $u \equiv 1$ is a stable equilibrium that “attracts” positive solutions.

The logistic equation above makes predictions about the total population but says nothing about how the population might be distributed over its environment. A simple way to introduce this is to model the environment as an interval $[0, 1]$ and introduce a “spatial” variable, $x \in [0, 1]$. Now we can track how a population changes over time at each point of $[0, 1]$ by solving

$$\frac{\partial u}{\partial t} = k \frac{\partial^2 u}{\partial x^2} + \lambda u(1-u). \quad (4)$$

The “reaction” term $\lambda u(1-u)$ still pushes the values of u towards 1, but the “diffusion” term, $k(\partial^2 u / \partial x^2)$, with $k > 0$, causes u to spread from higher concentrations

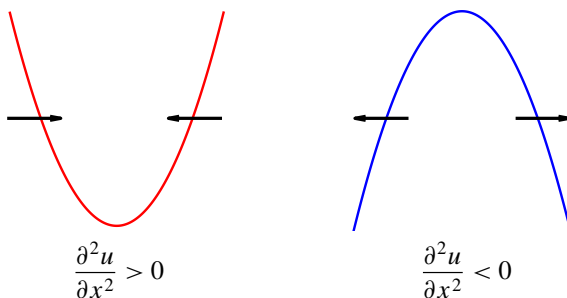


Figure 1. Diffusion effects.

to lower. **Figure 1**, left, shows the second partial derivative as positive and therefore the graph of u is concave up. The population will then diffuse in towards the point where the minimum is, and so the effect of diffusion is to increase population in that region. A similar situation occurs in **Figure 1**, right, where the population will diffuse outwards away from the maximum point and thus the effect of diffusion is to decrease population in that region. The influences of reaction and diffusion compete with each other to determine how u is changing with time.

If u reaches an equilibrium, we will have $\partial u / \partial t \equiv 0$, i.e., u is not changing over time, so our equilibrium equation is

$$0 = k \frac{\partial^2 u}{\partial x^2} + \lambda u(1 - u). \quad (5)$$

Note that we can now divide through by k and rename λ/k as λ again. If we have a stable equilibrium, then time-dependent solutions that start “near” this equilibrium move towards the equilibrium. Equilibrium solutions provide a framework for understanding the time-dependent solutions.

There is another influence that we have not yet accounted for and that is whether the population will move through the boundary in any way. The novelty of our paper comes entirely from the boundary conditions that we consider. We will impose the boundary conditions $u(1) = 0$ and $(\partial u / \partial v)(0) + g(u) = 0$. For simplicity we have imposed the Dirichlet condition at $x = 1$. The condition $u(1) = 0$ is often interpreted as the boundary at $x = 1$ being lethal. If you cross that boundary you are gone forever and cannot reenter. At the left boundary, $x = 0$, we study two different versions of $(\partial u / \partial v)(0) + g(u) = 0$. The value of $(\partial u / \partial v)(0)$ represents the outward rate of change at $x = 0$, i.e., $(\partial u / \partial v)(0) = -u'(0)$, which is often assumed to be proportional to the rate of flow of population into $[0, 1]$ across the boundary $x = 0$. The term $g(u)$ tells us how that flow depends on the population in $[0, 1]$.

In one case, we will consider $(\partial u / \partial v)(0) + g(u(0)) = 0$, so the flow is entirely determined by conditions at the boundary. Moreover we assume that g is a function modeling nonlinear “density-dependent” behavior. This interaction can come in

a lot of forms, but we consider primarily the g that models species who prefer to stay in a habitat with higher concentrations of u ; i.e., as the density increases at the boundary a smaller proportion of the population will leave through that boundary. Thus the per capita flow across the boundary, which is proportional to

$$\frac{(\partial u / \partial v)(0)}{u(0)} = -\frac{g(u(0))}{u(0)},$$

is assumed to be decreasing in absolute value. This is referred to as “negative density-dependent” behavior. A relatively simple model to consider is $g(u) = \sqrt{u+1}$. This behavior was observed in Glanville fritillary butterflies; for example, see [Cantrell and Cosner 2006]. The butterflies were observed to be less likely to leave their patch boundary if the local density of Glanville fritillary butterflies within the patch was great enough.

In the second case, we consider $(\partial u / \partial v)(0) + g(u) = 0$, where $g(u)$ is linear but “nonlocal”. Thus the flow through the boundary depends on the population density throughout the region $[0, 1]$ and not just at the boundary. Within certain restrictions our condition allows the interior population density to have either a positive or negative influence on the flow across the boundary.

In this paper we study a discretized version of (5) with boundary conditions as above. We will prove the existence of nonnegative solutions in several situations using monotonicity methods and will illustrate those results with examples and simple computations in Maple.

We were primarily motivated by [Cantrell and Cosner 2006; Bruno 2021], and subsequent papers such as [Ashley et al. 2013; Cantrell and Cosner 2007; Goddard and Shivaji 2017], where nonlinear boundary conditions are considered in combination with Allee effects. Our results provide a discrete alternative to the continuous results referenced above, and significantly generalize the boundary conditions considered in [Bruno 2021].

In order to be somewhat self-contained we will first discuss some standard theoretical background before moving on to prove the main results and then presenting some examples and computations. We end by posing a few possible research questions.

2. Theoretical background

2.1. Boundary conditions. We follow the model proposed in [Cantrell and Cosner 2006]. Consider a common formulation of the Robin boundary condition

$$\alpha \frac{\partial u}{\partial v} + (1 - \alpha)u = 0, \quad \text{where } 0 < \alpha < 1. \quad (6)$$

Note that $\partial u / \partial v$ is an outward rate of change. At the boundary $x = 0$, we have $\partial u / \partial v = -u'(0)$. At $x = 1$, we have $\partial u / \partial v = u'(1)$. The parameter α is often

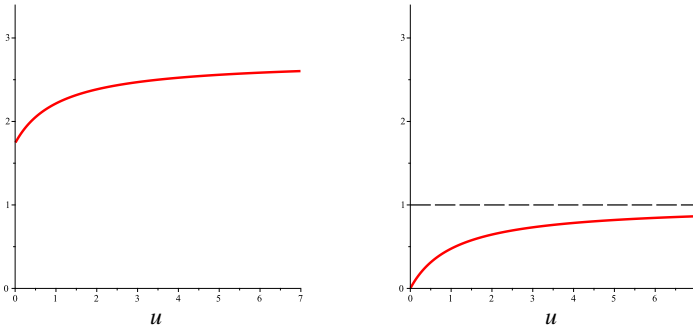


Figure 2. Graphs of $g(u)$ (left) and $\alpha(u)$ (right).

interpreted as the probability of an individual staying in the domain when reaching the boundary.

The relationship between α and u is modeled by

$$\alpha(u) = \frac{u}{u + g(u)}. \quad (7)$$

Thus, we have

$$0 = \alpha(u) \frac{\partial u}{\partial v} + (1 - \alpha(u))u = \left(\frac{u}{u + g(u)} \right) \frac{\partial u}{\partial v} + \left(\frac{u + g(u)}{u + g(u)} - \frac{u}{u + g(u)} \right) u. \quad (8)$$

Multiplying through by $u + g(u)$ we are left with $u(\partial u / \partial v + g(u)) = 0$. In this paper we will consider $\partial u / \partial v = -g(u)$. For simplicity we only impose this condition at one boundary point, and maintain a Dirichlet condition at the other. We note that several combinations of Dirichlet and density-dependent boundary conditions are considered for the continuous model in [Ashley et al. 2013].

The first case we study is when $(\partial u / \partial v)(0) = -g(u(0))$, and we assume that $g : [0, \infty) \rightarrow \mathbb{R}$ is twice continuously differentiable, with

$$g(0) = d > 0, \quad g'(u) \geq 0, \quad g''(u) \leq 0, \quad \lim_{u \rightarrow \infty} \frac{g(u)}{u} = 0. \quad (9)$$

Therefore, $\alpha \in (0, 1)$ and

$$\lim_{u \rightarrow \infty} \alpha(u) = \frac{u}{u + g(u)} = 1.$$

Thus as u increases the probability of individuals leaving through the boundary decreases; see Figure 2. These assumptions model negative density-dependent boundary behavior.

The second case that we study will consider a linear nonlocal boundary condition in the form

$$g(u) = \sum_{k=0}^n a_k u(x_k) + d,$$

where $a_k \in \mathbb{R}$, and the x_k are a collection of points in a mesh for the interval $[0, 1]$. This boundary condition models a population whose flow through the boundary depends on the full distribution of the population in $[0, 1]$. We will assume $d > 0$. Our analysis will allow the coefficients, i.e., a_k for $k = 0, \dots, n-1$, to be either positive or negative within certain bounds.

2.2. Discretization. In this subsection we discretize the boundary value problem

$$\begin{aligned} 0 &= u''(x) + \lambda u(x)(1 - u(x)), \quad x \in (0, 1), \\ u(1) &= 0, \quad \frac{\partial u}{\partial v}(0) = -g(u). \end{aligned}$$

Let $n \geq 3$ and let the step size be $h = 1/n$. For $k = 0, \dots, n$, let $x_k = k/n$ and let $u_k = u(x_k)$. We approximate the first derivative at the boundary using

$$u'(x) \approx \frac{u(x+h) - u(x)}{h} = n \left(u \left(x + \frac{1}{n} \right) - u(x) \right).$$

It follows that $(\partial u / \partial v)(0) = -u'(0) \approx -n(u_1 - u_0)$.

We approximate the second derivative in the interior using

$$\begin{aligned} u''(x) &\approx \frac{1}{h} \left(\left(\frac{u(x+h) - u(x)}{h} \right) - \left(\frac{u(x) - u(x-h)}{h} \right) \right) \\ &= n^2 \left(u \left(x + \frac{1}{n} \right) - 2u(x) + u \left(x - \frac{1}{n} \right) \right). \end{aligned} \quad (10)$$

It follows that $u''(x_k) \approx n^2(u_{k+1} - 2u_k + u_{k-1})$.

Consider the discretization of the boundary conditions. At the right endpoint the condition $u(1) = 0$ simply becomes $u_n = 0$. At the left endpoint let's first focus on the previously described nonlinear condition. The condition $(\partial u / \partial v)(0) = -g(u(0))$ becomes $-n(u_1 - u_0) = -g(u_0)$. Thus $u_1 = u_0 + (g(u_0)/n)$. However, as will be seen below, it is more useful to solve for u_0 as a function of u_1 . Let $G(u) = u + g(u)/n$, with derivatives $G'(u) = 1 + g'(u)/n$ and $G''(u) = g''(u)/n$. Using (9) we have $G(0) = d/n$, $G'(u) \geq 1$, and $G''(u) \leq 0$, so G is strictly increasing, and G is concave. It is not hard to see that $\lim_{u \rightarrow \infty} g'(u) = 0$, and thus $\lim_{u \rightarrow \infty} G'(u) = 1$. For $u < 0$ we can extend $G(u)$ as

$$\frac{d}{n} + \left(1 + \frac{g'(0)}{n} \right) u,$$

so $G(u) = 0$, where $u = -d/(n + g'(0))$.

We have $u_1 = G(u_0)$, but we want to solve for u_0 in terms of u_1 . Since G is smooth and strictly increasing, it has a smooth inverse $\varphi(u)$ by the inverse function theorem. Thus we can rewrite the boundary condition as $u_0 = \varphi(u_1)$. Using

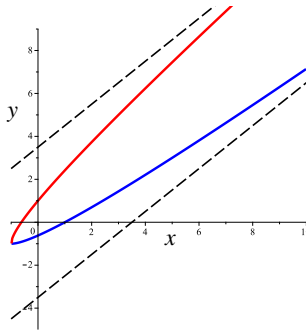


Figure 3. $G(u)$ is in red and $\varphi(u)$ is in blue. The dashed lines represent asymptotes with slope 1.

symmetry about $y = u$, we get that

$$\begin{aligned} \varphi\left(\frac{d}{n}\right) &= 0, \quad \varphi \text{ is increasing with } \varphi'(u) \leq 1, \\ \varphi \text{ is convex, } \quad \lim_{u \rightarrow \infty} \varphi'(u) &= 1; \end{aligned} \tag{11}$$

see [Figure 3](#).

The second case that we consider uses $u_n = 0$ and $-n(u_1 - u_0) = -g(u)$, where

$$g(u) = \sum_{k=0}^{n-1} a_k u_k + d = a_0 u_0 + \langle a, u \rangle + d, \tag{12}$$

where $a := (a_1, \dots, a_{n-1})$, and $\langle \cdot, \cdot \rangle$ is the standard inner product in R^{n-1} . Solving $-n(u_1 - u_0) = -g(u)$ for u_0 leads to

$$u_0 = \frac{n}{n+a_0} \left(u_1 - \frac{1}{n} \langle a, u \rangle - \frac{d}{n} \right) =: \varphi(u). \tag{13}$$

We assume

$$n + a_0 > 0, \quad \langle a, v \rangle < n, \tag{14}$$

where v is the principal eigenvector associated with the discrete Laplacian, which is normalized so that its first component is $v_1 = 1$. (See discussion of the discrete Laplacian below.)

Next we discretize the equation $u''(x) + \lambda u(x)(1 - u(x)) = 0$. Using the approximation for $u''(x)$ above, we get

$$n^2(u_{k+1} - 2u_k + u_{k-1}) + \lambda u_k(1 - u_k) = 0$$

for $k = 1, \dots, n - 1$. For $k = n - 1$ we can substitute $u_n = 0$ from the boundary condition above. For $k = 1$ we can substitute $u_0 = \varphi(u)$ from the boundary condition

above. We now have the following system of $n - 1$ equations:

$$\begin{aligned}
 n^2(u_2 - 2u_1 + 0) + n^2\varphi(u) + \lambda u_1(1 - u_1) &= 0, \\
 n^2(u_3 - 2u_2 + u_1) + 0 + \lambda u_2(1 - u_2) &= 0, \\
 &\vdots \\
 n^2(u_{n-1} - 2u_{n-2} + u_{n-3}) + 0 + \lambda u_{n-2}(1 - u_{n-2}) &= 0, \\
 n^2(0 - 2u_{n-1} + u_{n-2}) + 0 + \lambda u_{n-1}(1 - u_{n-1}) &= 0.
 \end{aligned} \tag{15}$$

Writing the system in matrix-vector form we get

$$n^2 \begin{bmatrix} -2 & 1 & 0 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -2 & 1 \\ 0 & \cdots & \cdots & 0 & 1 & -2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{n-2} \\ u_{n-1} \end{bmatrix} + n^2 \begin{bmatrix} \varphi(u) \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} + \lambda \begin{bmatrix} u_1(1-u_1) \\ u_2(1-u_2) \\ u_3(1-u_3) \\ \vdots \\ u_{n-2}(1-u_{n-2}) \\ u_{n-1}(1-u_{n-1}) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}.$$

Note that we can divide through by n^2 and absorb n^2 into the λ . Thus, we are left with (1).

2.3. Properties of the discrete Laplacian. Let $M = -A$. The main property that is required for later arguments is the fact that M is an M -matrix. This hypothesis is enough to imply the monotonicity discussed below. An excellent reference is [Berman and Plemmons 1979]. M is also symmetric and positive definite, which are more familiar properties from an introductory linear algebra class, and which imply further nice structure for the eigenvalues and eigenvectors.

For completeness we provide elementary proofs of the properties of M that are used later. In particular we show that M has positive distinct real eigenvalues, $\lambda_1 < \lambda_2 < \cdots < \lambda_{n-1}$, and that the principal eigenvalue, λ_1 , has an eigenvector v whose entries are positive and convex such that $v_k = v_{n-k}$. We will assume in all that follows that v has been normalized so that $v_1 = 1$. We refer to [Weisstein; Wikipedia] for further detail.

2.3.1. Monotonicity. A square matrix B with real entries is monotone if, for all real vectors u , $Bu \geq 0$ implies $u \geq 0$.

Theorem 1. M is monotone.

Proof. We want to show that if $Mu \geq 0$, then $u \geq 0$. Assume

$$M \begin{bmatrix} u_1 \\ \vdots \\ u_{n-1} \end{bmatrix} \geq 0.$$

Let $u_0 = 0$ and $u_n = 0$. Notice $Mu \geq 0$ implies

$$\begin{aligned} -u_0 + 2u_1 - u_2 &\geq 0, \\ -u_1 + 2u_2 - u_3 &\geq 0, \\ -u_{n-3} + 2u_{n-2} - u_{n-1} &\geq 0, \\ &\vdots \\ -u_{n-2} + 2u_{n-1} - u_n &\geq 0. \end{aligned} \tag{16}$$

So we can see the pattern $-u_{k-1} + 2u_k - u_{k+1} \geq 0$ for $k = 1, \dots, n-1$. We can further write this as

$$(u_k - u_{k-1}) + (u_k - u_{k+1}) \geq 0. \tag{17}$$

Claim 1. *If $u_k \leq u_{k-1}$ and $u_k \leq u_{k+1}$, then $u_{k-1} = u_k = u_{k+1}$.*

Proof. By assumption, we have $u_k - u_{k-1} \leq 0$ and $u_k - u_{k+1} \leq 0$ but by (17) above, these equations must strictly equal 0. Thus $u_k - u_{k-1} = 0$ and $u_k - u_{k+1} = 0$. \square

This means that u can never be concave up.

Claim 2. *Let $m = \min\{u_0, u_1, \dots, u_n\} \leq 0$. Suppose $u_k = m$ for some $1 \leq k \leq n-1$. Then $m = 0$ and $u_k = 0$ for all k .*

Proof. If $u_k = m \leq 0$, then $u_{k-1} = u_k = u_{k+1}$ by Claim 1. We can apply Claim 1 again to $u_{k-2} = u_{k-1} = u_k$ and so on until we reach the end of the boundary, $u_0 = u_1 = u_2$. Thus we have $u_0 = u_1 = \dots = u_n = 0 = m$ because we assumed $u_0 = 0$ and $u_n = 0$. Thus $m = 0$. \square

Claim 3. *If $Mu \geq 0$, then $u_k > 0$ for $k = 1, 2, \dots, n-1$.*

Proof. Let $m = \min\{u_0, u_1, \dots, u_n\} \leq 0$. If $u_k = m \leq 0$ for any $k = 1, \dots, n-1$ then $m = 0$ and $u_k = 0$ for all k based on Claim 2. Then $Mu = 0$. But this is a contradiction to our assumption. Hence $m = 0$ but only $u_0 = u_n = 0$. \square

Hence M is monotone. \square

Theorem 2. *If $\mu > 0$ then the matrix $M + \mu I$ is monotone.*

Proof. Consider, $Mu + \mu u \geq 0$. Letting $u_0 = u_n = 0$ as before this means $-u_{k-1} + 2u_k - u_{k+1} + \mu u_k \geq 0$ for $k = 1, \dots, n-1$, which can be written as $(u_k - u_{k-1}) + (u_k - u_{k+1}) + \mu u_k \geq 0$. Suppose by contradiction that there exists some u_k that is the global minimum of the sequence and is a negative value. This means $u_k \leq u_{k-1}$, $u_k \leq u_{k+1}$ and $u_k < 0$. Then when we reconsider $(u_k - u_{k-1}) + (u_k - u_{k+1}) + \mu u_k \geq 0$. We see that the values in parentheses are nonpositive, and since we assumed u_k to be negative, the third term is also negative. But this means $(u_k - u_{k-1}) + (u_k - u_{k+1}) + \mu u_k < 0$, a contradiction. Hence, the whole sequence must be greater than or equal to 0 because we know there is no negative minimum value. \square

2.3.2. Eigenvalues and eigenvectors. It turns out that one can be quite explicit about the eigenvalues and eigenvectors of M if one explores a surprising and quite interesting connection with Chebyshev polynomials of the second kind. See [Weisstein; Wikipedia] for a wealth of information about these polynomials.

To find a real eigenvalue λ of M with eigenvector v we must solve $Mv = \lambda v$. Without loss of generality we can rescale v so that $v_1 = 1$. Let $v_0 = v_n = 0$. It follows that $-v_{k-1} + 2v_k - v_{k+1} = \lambda v_k$ for $k = 1, \dots, n-1$. Rearranging we get $v_{k+1} = (2 - \lambda)v_k - v_{k-1}$. If we let $2\alpha = 2 - \lambda$, then we can express the eigenvalue problem above as a solution of the recursion relation

$$v_{k+1} = 2\alpha v_k - v_{k-1}, \quad v_0 = 0, \quad v_1 = 1. \quad (18)$$

In particular we seek solutions of (18) such that $v_n = 0$.

The given recursion relation defines polynomials in the variable α , i.e.,

$$P_k(\alpha) := v_{k+1},$$

which are known as Chebyshev polynomials of the second kind. The remaining arguments explore a few properties of these polynomials for α in the interval $[-1, 1]$ and then draw conclusions about the eigenvalue problem. We note that the zeros of $v_n = P_{n-1}(\alpha)$ correspond to solutions of the eigenvalue problem. A helpful change of variables is $\alpha = \cos \theta$ for $0 \leq \theta \leq \pi$.

Lemma 1.
$$P_k(\cos \theta) = \frac{\sin((k+1)\theta)}{\sin \theta} \quad \text{when } \sin \theta \neq 0.$$

Proof. We have $P_0(\alpha) \equiv 1$ so $P_0(\cos \theta) = \sin((0+1)\theta)/\sin \theta$. Assume

$$P_j(\cos \theta) = \frac{\sin((j+1)\theta)}{\sin \theta}, \quad j = 1, \dots, k-1,$$

when $\sin \theta \neq 0$. Then

$$\begin{aligned} P_k(\cos \theta) &= 2 \cos \theta P_{k-1}(\cos \theta) - P_{k-2}(\cos \theta), \\ &= 2 \cos \theta \frac{\sin(k\theta)}{\sin \theta} - \frac{\sin((k-1)\theta)}{\sin \theta}. \end{aligned}$$

Also

$$\begin{aligned} &\frac{\sin((k+1)\theta)}{\theta} \\ &= \frac{\sin(k\theta) \cos \theta + \sin \theta \cos(k\theta)}{\sin \theta} = \cos \theta \frac{\sin(k\theta)}{\sin \theta} + \cos(k\theta) \\ &= \cos \theta \frac{\sin(k\theta)}{\sin \theta} + \cos((k-1)\theta) \cos \theta - \sin((k-1)\theta) \sin \theta \\ &= \cos \theta \frac{\sin(k\theta)}{\sin \theta} + \frac{1}{\sin \theta} (\cos((k-1)\theta) \cos \theta \sin \theta - \sin((k-1)\theta) \sin^2 \theta) \end{aligned}$$

$$\begin{aligned}
 &= \cos \theta \frac{\sin(k\theta)}{\sin \theta} + \frac{1}{\sin \theta} (\cos((k-1)\theta) \cos \theta \sin \theta - \sin((k-1)\theta)(1 - \cos^2 \theta)) \\
 &= \cos \theta \frac{\sin(k\theta)}{\sin \theta} + \frac{\cos \theta}{\sin \theta} (\cos((k-1)\theta) \sin \theta + \sin((k-1)\theta) \cos \theta) - \frac{\sin((k-1)\theta)}{\sin \theta} \\
 &= \cos \theta \frac{\sin(k\theta)}{\sin \theta} + \frac{\cos \theta}{\sin \theta} (\sin(k\theta)) - \frac{\sin((k-1)\theta)}{\sin \theta} = 2 \cos \theta \frac{\sin(k\theta)}{\sin \theta} - \frac{\sin((k-1)\theta)}{\sin \theta}.
 \end{aligned}$$

Hence combining the results above gives

$$P_k(\cos \theta) = \frac{\sin((k+1)\theta)}{\sin \theta}. \quad \square$$

It follows from this lemma that $P_k(\alpha) = 0$ for $\alpha = \cos(j\pi/(k+1))$ with $j = 1, \dots, k$. Applying this to $k = n - 1$ leads to the eigenvalues $\lambda_j = 2(1 - \cos(j\pi/n))$ for $j = 1, \dots, n - 1$. Since M is an $(n - 1) \times (n - 1)$ matrix, this must be a complete list of the $n - 1$ simple eigenvalues for M .

The principal eigenvalue for M is given by $\lambda_1 = 2(1 - \cos(\pi/n))$. For any $k < n - 1$ we have that the entries of the principal eigenvector are

$$v_{k+1} = P_k\left(\cos\left(\frac{\pi}{n}\right)\right) = \frac{\sin((k+1)\pi/n)}{\sin(\pi/n)}.$$

The positivity, symmetry, and concavity of v all follow.

2.4. Basic existence theorem using sub- and supersolutions. A classic reference for monotonicity methods, which we adapt for discrete problems, is [Sattinger 1972]. For notational convenience in stating and proving the following theorem, we let $h(u) := \phi(u) + \lambda f(u)$. Also, in this section we use $(u^{(j)})$ to represent a sequence of vectors in R^{n-1} .

Definition. We say that $\bar{u} \in R^{n-1}$ is a supersolution of (1) if $A\bar{u} + h(\bar{u}) \leq 0$. We say that $\underline{u} \in R^{n-1}$ is a subsolution if $A\underline{u} + h(\underline{u}) \geq 0$.

Theorem 3. Let \underline{u} and \bar{u} be sub- and supersolutions, respectively, for problem (1) such that $\underline{u} \leq \bar{u}$. Then there exists a solution $u \in R^{n-1}$ for (1) such that $\underline{u} \leq u \leq \bar{u}$.

Proof. Let us examine a special case where $h : R^{n-1} \rightarrow R^{n-1}$ is nondecreasing and continuous. Recall that $-A$ is monotone, so if $-Au \geq 0$, then $u \geq 0$.

Let $u^{(0)} = \underline{u}$ and consider the recurrence formula

$$\begin{aligned}
 Au^{(1)} + h(u^{(0)}) &= 0, \\
 Au^{(2)} + h(u^{(1)}) &= 0, \\
 &\vdots \\
 Au^{(j)} + h(u^{(j-1)}) &= 0, \\
 Au^{(j+1)} + h(u^{(j)}) &= 0,
 \end{aligned}$$

and so on. We know that A is invertible so these equations are uniquely solvable.

We argue by induction that our sequence is bounded. We assumed $u^{(0)} = \underline{u} \leq \bar{u}$, so let us assume $u^{(j)} \leq \bar{u}$ and prove $u^{(j+1)} \leq \bar{u}$. We have $A(\bar{u} - u^{(j+1)}) \leq -h(\bar{u}) + h(u^{(j)})$ and notice once we multiply by a negative, the sign will switch: $-A(\bar{u} - u^{(j+1)}) \geq (h(\bar{u}) - h(u^{(j)}))$; this last relation is greater than or equal to 0 because we assumed $u^{(j)} \leq \bar{u}$ and that h is nondecreasing. Therefore, $\bar{u} - u^{(j+1)} \geq 0$, which implies $\bar{u} \geq u^{(j+1)}$.

We argue inductively that the sequence of vectors $(u^{(j)})$ is monotone. Notice that

$$A(u^{(1)} - u^{(0)}) = -h(u^{(0)}) - Au^{(0)} = -(h(\underline{u}) + A\underline{u}) \leq 0.$$

Therefore $-A(u^{(1)} - u^{(0)}) \geq 0$ and by monotonicity $u^{(1)} \geq u^{(0)}$. Assume $u^{(0)} \leq u^{(1)} \leq \dots \leq u^{(j)}$ for some $j \in \mathbb{N}$. Now, we want to examine $u^{(j+1)}$ with our recurrence formula. We have $Au^{(j+1)} + h(u^{(j)}) = 0$ and we consider

$$A(u^{(j+1)} - u^{(j)}) = -h(u^{(j)}) + h(u^{(j-1)}).$$

Thus, if we consider the negative of this equation we get

$$-A(u^{(j+1)} - u^{(j)}) = (h(u^{(j)}) - h(u^{(j-1)})) \geq 0$$

by our second assumption. Therefore, $u^{(j+1)} - u^{(j)} \geq 0$, which implies $u^{(j)} \leq u^{(j+1)}$. Hence, $u^{(0)} \leq u^{(1)} \leq \dots \leq u^{(j)} \leq u^{(j+1)}$. Thus, we have shown our sequence, $(u^{(j)})$, is monotone.

We have shown $(u^{(j)})$ is monotonically nondecreasing and is bounded above. By the monotone convergence theorem, applied componentwise, we see $u^{(j)} \rightarrow u$. So, as $j \rightarrow \infty$ in $Au^{(j)} + h(u^{(j-1)}) = 0$ we then have $Au + h(u) = 0$ with $\underline{u} \leq u \leq \bar{u}$, where we have used the continuity of A and h .

Now consider if $h(u) = \phi(u) + \lambda f(u)$, which is not monotone. Recall that ϕ is nondecreasing for both cases of interest in this paper. Let M the maximum entry of \bar{u} . Choose $\mu \geq 2M - 1$. It follows that $x(1-x) + \mu x$ is nondecreasing on $(-\infty, M]$. It follows that $f_\mu(u) := f(u) + \mu u$ is nondecreasing for all u such that $u \leq \bar{u}$. Hence $h_\mu(u) := \phi(u) + \lambda f_\mu(u)$ is nondecreasing for all u such that $u \leq \bar{u}$.

Consider $Au + h(u) = 0$. If we add and subtract μu we can manipulate our equation into $(A - \mu I)u + (h(u) + \mu u) = 0$, which becomes $A_\mu u + h_\mu(u) = 0$, where $-A_\mu := -A + \mu I$ is monotone, and $h_\mu(u)$ is monotone increasing for $u \leq \bar{u}$. Moreover it is easy to check that \underline{u} and \bar{u} are sub- and supersolutions of $A_\mu u + h_\mu(u) = 0$. By the argument above, this equation has a solution, u , such that $\underline{u} \leq u \leq \bar{u}$, which is clearly also a solution of $A(u) + h(u) = 0$. \square

3. The main results

3.1. Nonlinear boundary condition. In this section we assume the boundary conditions $u_n = 0$ and $u_0 = \varphi(u_1)$, where φ satisfies (11) as a consequence of (9).

Recall that our desired solution should be nonnegative and in particular we need $u_1 \geq d/n > 0$ so that $u_0 = \varphi(u_1) \geq 0$.

Consider $u = cv$, where $Av = -\lambda_1 v$ and $c > 0$. Given the properties of v discussed earlier we know that we can scale v to make sure the first and last terms in the vector are 1. Thus,

$$v = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix},$$

where the rest of the terms are values greater than 1. For later reference we choose $m \in \{2, \dots, n-2\}$ such that $v_m = \max\{v_1, \dots, v_{n-1}\} = \max\{v_2, \dots, v_{n-2}\}$. Then

$$Au + \phi(u) + \lambda u(1 - u) = \begin{bmatrix} -c\lambda_1 + \varphi(c) + \lambda c(1 - c) \\ -cv_2\lambda_1 + \lambda v_2 c(1 - v_2 c) \\ \vdots \\ -cv_{n-2}\lambda_1 + \lambda v_{n-2} c(1 - v_{n-2} c) \\ -c\lambda_1 + \lambda c(1 - c) \end{bmatrix}. \quad (19)$$

How can $c > 0$ be chosen to create a subsolution? We require each entry in the vector above to be nonnegative. For $k = 2, 3, \dots, n-1$ we require

$$-cv_k\lambda_1 + \lambda v_k c(1 - v_k c) \geq 0 \quad \text{or} \quad -\lambda_1 + \lambda(1 - v_k c) \geq 0. \quad (20)$$

If $c \geq 1/v_k$, then the given expression is negative, so we require $c < 1/v_k$ for each k . Thus $c < 1/v_m < 1$. From above we also know that we require $c \geq d/n$ so that $u_1 = cv_1 = c \geq d/n$. Solving for λ we get

$$\lambda \geq \frac{\lambda_1}{1 - cv_k}$$

for $k = 2, \dots, n-1$. This inequality will be true for all k if it is true just for m , thus we have the conditions

$$\lambda \geq \frac{\lambda_1}{1 - cv_m} \quad \text{and} \quad \frac{d}{n} \leq c < \frac{1}{v_m}. \quad (21)$$

For $k = 1$ we require

$$-c\lambda_1 + \varphi(c) + \lambda c(1 - c) \geq 0.$$

Since we are already requiring $c < 1$, we can solve for λ to get

$$\lambda \geq \frac{c\lambda_1 - \varphi(c)}{c(1 - c)}.$$

For $c \geq d/n$ we have $\varphi(c) \geq 0$, so

$$\frac{c\lambda_1 - \varphi(c)}{c(1 - c)} \leq \frac{\lambda_1}{1 - c} \leq \frac{\lambda_1}{1 - cv_m}.$$

Hence condition (21) will also guarantee that the first component in (19) is non-negative. The smallest value of λ that can satisfy the given inequality for the given range of c is

$$\lambda^* := \frac{\lambda_1}{1 - (d/n)v_m}.$$

For any $\lambda \geq \lambda^*$ we can solve (21) for the largest c that gives a subsolution, which is

$$c = \frac{1}{v_m} \left(1 - \frac{\lambda_1}{\lambda} \right).$$

We conclude that for all $\lambda \geq \lambda^*$ we have a subsolution

$$\underline{u} = \frac{1}{v_m} \left(1 - \frac{\lambda_1}{\lambda} \right) v.$$

How can we choose $c > 0$ to create a supersolution? In this case each entry on the right hand side of (19) must be nonpositive. For $k = 1$ we have

$$-c\lambda_1 + \varphi(c) + \lambda c(1 - c) \leq 0. \quad (22)$$

Since we eventually want the supersolution to be larger than the subsolution we know $c \geq d/n$. We also know $\varphi(d/n) = 0$ and $\varphi'(c) \leq 1$, so $\varphi(c) \leq c - d/n \leq c$. This means that (22) will be satisfied if

$$-c\lambda_1 + c + \lambda c(1 - c) \leq 0,$$

which simplifies to

$$-\lambda_1 + 1 + \lambda(1 - c) \leq 0,$$

which is equivalent to

$$\frac{1 - \lambda_1}{\lambda} + 1 \leq c. \quad (23)$$

Recall that $\lambda_1 \leq 1$ for $n \geq 3$, so the previous inequality implies $c \geq 1$. It also follows that $cv_k \geq 1$ for $k = 2, 3, \dots, n - 1$ and thus

$$-\lambda_1 cv_k + \lambda cv_k(1 - cv_k) \leq 0$$

for $k = 2, 3, \dots, n - 1$. Hence for any $\lambda \geq \lambda^*$ we have that

$$\bar{u} = \left(\frac{1 - \lambda_1}{\lambda} + 1 \right) v$$

is a supersolution.

Finally, it is clear that

$$\frac{1}{v_m} \left(1 - \frac{\lambda_1}{\lambda} \right) < 1 < \left(\frac{1 - \lambda_1}{\lambda} + 1 \right),$$

so $\underline{u} < \bar{u}$. We have proved the following.

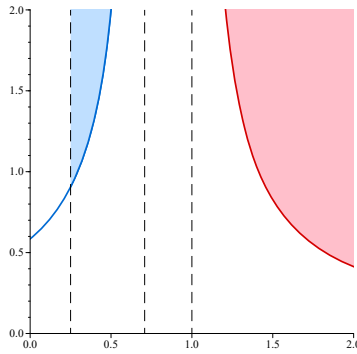


Figure 4. The blue region represents (c, λ) for subsolutions, and the pink (c, λ) for supersolutions. The blue region is bounded by $c = d/n$ and $\lambda = \lambda_1/(1 - cv_m)$ with an asymptote of $c = 1/v_m$. The pink region is bounded by $\lambda = (1 - \lambda_1)/(c - 1)$ with asymptote $c = 1$.

Theorem 4. Let A represent the $(n-1)$ -dimensional discrete Laplacian for $n \geq 3$. Let f represent the $(n-1)$ -dimensional vector valued logistic function. Let g satisfy (9), and let φ be the inverse of $G(u) = u + g(u)/n$. Let λ_1 be the principal eigenvalue of $-A$ and let v represent the positive principal eigenvector of $-A$ which is scaled so that $v_1 = v_{n-1} = 1$. Let $v_m := \max\{v_2, \dots, v_{n-2}\}$. Assume $d/n < 1/v_m$. Let

$$\lambda^* = \frac{\lambda_1}{1 - v_m(d/n)}.$$

Then for every $\lambda \geq \lambda^*$ equation (1) has a solution, u , such that

$$\frac{1}{v_m} \left(1 - \frac{\lambda_1}{\lambda} \right) v \leq u \leq \left(\frac{1 - \lambda_1}{\lambda} + 1 \right) v.$$

We illustrate the possible choices for (c, λ) to produce subsolutions and supersolutions in Figure 4.

3.2. Linear nonlocal boundary condition. In this section we assume $u_n = 0$ and $-n(u_1 - u_0) = -g(u)$, where g and φ are given by (12) and (13). We also assume (14).

Observe that

$$\varphi(cv) = \frac{n}{n+a_0} \left(c \left(1 - \frac{1}{n} \langle a, v \rangle \right) - \frac{d}{n} \right) = \frac{1}{n+a_0} (c(n - \langle a, v \rangle) - d),$$

which is a linear real-valued function of c with slope $\sigma = (n - \langle a, v \rangle)/(n + a_0) > 0$, $\varphi(0) = -d/(n + a_0) < 0$, and $\varphi(dv/(n - \langle a, v \rangle)) = 0$. It follows that the graph of $\varphi(cv)$ is an upward sloping line with negative intercept on the vertical axis, and positive intercept on the horizontal axis.

Consider first the condition $0 \leq u_0 = \varphi(cv)$. This leads to

$$c(n - \langle a, v \rangle) - d \geq 0,$$

and thus

$$c \geq \frac{d}{n - \langle a, v \rangle} > 0. \quad (24)$$

For notational convenience let $d_n := d/(n - \langle a, v \rangle)$.

How do we choose $c \geq d_n$ so that $u = cv$ is a subsolution? Once again we require every element of (19) to be nonnegative. For $k = 2, \dots, n - 1$ this is exactly as in the previous section so we require $d_n < 1/v_m$, and

$$\lambda > \frac{\lambda_1}{1 - cv_m} \quad \text{for } d_n \leq c < \frac{1}{v_m} < 1.$$

Let $\lambda^* := \lambda_1/(1 - d_n v_m)$.

For $k = 1$ we require, as before,

$$-c\lambda_1 + \varphi(cv) + \lambda c(1 - c) \geq 0.$$

Solving for λ yields

$$\lambda \geq \frac{c\lambda_1 - \varphi(cv)}{c(1 - c)}.$$

For $c \geq d_n$ we have $\varphi(cv) \geq 0$ so

$$\frac{c\lambda_1 - \varphi(cv)}{c(1 - c)} \leq \frac{c\lambda_1}{c(1 - c)} = \frac{\lambda_1}{1 - c} \leq \frac{\lambda_1}{1 - cv_m}.$$

Hence the condition

$$\lambda \geq \frac{\lambda_1}{1 - cv_m}$$

implies

$$\lambda \geq \frac{c\lambda_1 - \varphi(cv)}{c(1 - c)}.$$

As a result of the inequalities above we know that for any $\lambda \geq \lambda^*$ we can choose $c = (1/v_m)(1 - \lambda_1/\lambda)$ to get a subsolution $\underline{u} = cv$.

For $\lambda \geq \lambda^*$, as above, how can we choose c so that $u = cv$ is a supersolution? A simple first choice is $c \geq 1$, which immediately guarantees that the $k = 2, \dots, n - 1$ entries of (19) are nonpositive. Thus we are left to consider $k = 1$, i.e.,

$$-c\lambda_1 + \varphi(cv) + \lambda c(1 - c) \leq 0$$

for $c \geq 1$. We know that $\varphi(cv) \leq \sigma c$, so it suffices to choose c such that

$$-c\lambda_1 + \sigma c + \lambda c(1 - c) \leq 0.$$

Solving for positive c gives

$$c \geq 1 + \frac{\sigma - \lambda_1}{\lambda}.$$

Hence if $c = \max\{1, 1 + (\sigma - \lambda_1)/\lambda\}$, then $\bar{u} = cv$ is a supersolution. It is clear that

$$\frac{1}{v_m} \left(1 - \frac{\lambda}{\lambda_1}\right) < 1 \leq \max\left\{1, 1 + \frac{\sigma - \lambda_1}{\lambda}\right\},$$

so $\underline{u} < \bar{u}$.

Thus we have proved the following theorem.

Theorem 5. *Let A represent the $(n-1)$ -dimensional discrete Laplacian for $n \geq 3$. Let f represent the $(n-1)$ -dimensional vector-valued logistic function. Let g and φ be defined by (12) and (13), respectively, and assume (14). Let λ_1 be the principal eigenvalue of $-A$ and let v represent the positive principal eigenvector of $-A$ which is scaled so that $v_1 = v_{n-1} = 1$. Let*

$$v_m := \max\{v_2, \dots, v_{n-2}\}, \quad d_n := \frac{d}{n - \langle a, v \rangle}, \quad \sigma := \frac{n - \langle a, v \rangle}{n + a_0}, \quad \lambda^* := \frac{\lambda_1}{1 - d_n v_m}.$$

Assume $d_n < 1/v_m$. Then for every $\lambda \geq \lambda^*$ equation (1) has a solution, u , such that

$$\frac{1}{v_m} \left(1 - \frac{\lambda_1}{\lambda}\right) v \leq u \leq \max\left\{1, 1 + \frac{\sigma - \lambda_1}{\lambda}\right\} v.$$

4. Examples

In the following examples we assume $n = 4$. It follows that $\lambda_1 = 2 - \sqrt{2}$ and

$$v = \begin{bmatrix} 1 \\ \sqrt{2} \\ 1 \\ 1 \end{bmatrix}.$$

4.1. Nonlinear boundary condition. Let us consider the conditions $u_n = 0$ and $-n(u_1 - u_0) = -g(u_0)$, with $g(u) = \sqrt{u + 1}$. It is a straightforward exercise to check that g satisfies (9). We can even solve explicitly for

$$\varphi(u) = u + \frac{1}{32} - \frac{\sqrt{64u + 65}}{32}. \tag{25}$$

It follows that

$$d = 1, \quad v_m = \sqrt{2}, \quad \text{and} \quad \lambda^* = \frac{2 - \sqrt{2}}{1 - \sqrt{2}/4} \approx 0.91. \tag{26}$$

Observe that

$$\frac{d}{n} = \frac{1}{4} < \frac{1}{\sqrt{2}} = \frac{1}{v_m}.$$

Choose, for example, $\lambda = 2 > \lambda^*$. Then

$$\underline{u} = \frac{v}{2} \quad \text{and} \quad \bar{u} = \left(\frac{1 + \sqrt{2}}{\sqrt{2}} \right) v. \quad (27)$$

Thus we can apply [Theorem 4](#) to get a solution, u , of (1) such that

$$\frac{v}{2} \leq u \leq \left(\frac{1 + \sqrt{2}}{\sqrt{2}} \right) v.$$

To finish describing the solution we use the boundary conditions to determine that $u_4 = 0$ and $u_0 = \varphi(u_1)$. Since $v_1 = 1$ we know that $\frac{1}{2} \leq u_1 \leq ((1 + \sqrt{2})/\sqrt{2})$. We also know that φ is monotone increasing, so $\varphi(\frac{1}{2}) \leq u_0 \leq \varphi((1 + \sqrt{2})/\sqrt{2})$. Using the formula above we get approximately $0.22 \leq u_0 \leq 1.33$.

4.2. Nonlocal boundary condition. Now consider the conditions $u_n = 0$ and $-n(u_1 - u_0) = -g(u)$, with $g(u) = \frac{1}{2}u_0 - \frac{1}{4}u_1 - \frac{1}{8}u_2 - \frac{1}{16}u_3 + 1$. This would model a population whose probability of crossing the boundary $x = 0$ depends on the entire population distribution, but with less influence as distance from the boundary increases. It follows that

$$\begin{aligned} d &= 1, & v_m &= \sqrt{2}, & a_0 &= \frac{1}{2}, \\ a &= -\left(\frac{1}{4}, \frac{1}{8}, \frac{1}{16}\right), & \langle a, v \rangle &= -\left(\frac{5}{16} + \frac{\sqrt{2}}{8}\right) \approx -.49, \\ d_n &= \frac{16}{69 + 2\sqrt{2}} \approx .22, & \lambda^* &= \frac{-134 + 65\sqrt{2}}{14\sqrt{2} - 69} \approx .86, & \sigma &= \frac{23}{24} + \frac{\sqrt{2}}{36} \approx .998. \end{aligned} \quad (28)$$

Observe that $d_n < 1/v_m$, $n + a_0 > 0$ and $\langle a, v \rangle < n$.

Choose, for example, $\lambda = 2$. Then

$$\frac{1}{v_m} \left(1 - \frac{\lambda_1}{\lambda} \right) = \frac{1}{2}, \quad \text{and} \quad \max \left\{ 1, 1 + \frac{\sigma - \lambda_1}{\lambda} \right\} = \frac{23}{48} + \frac{37\sqrt{2}}{72} \leq 1.21.$$

Thus we can apply [Theorem 5](#) to get the solution, u , of (1) such that

$$\frac{v}{2} \leq u \leq 1.21v.$$

4.3. Computation. In this section we demonstrate how the solutions in the examples above can be approximated using a simple algorithm justified by the proof of [Theorem 3](#). The purpose is simply to illustrate previous results and not provide a best possible computational scheme. We consider once again the first example above.

The first step is to modify $f(u)$ so that it is monotone on an interval $(-\infty, M]$, where M is the maximum entry of \bar{u} . It is easy to check that $M \leq 3$. On this interval we have $f'(u) = 1 - 2u \geq -5$, so $f_5(u) := f(u) + 5u$ is monotone nondecreasing on $(-\infty, 3]$. We know that $\phi(u)$ is increasing, so $h_5(u) = \phi(u) + \lambda f_5(u)$, is increasing.

Now we apply the monotone iteration scheme to

$$(A - 10I)u + \phi(u) + 2f_5(u) = 0,$$

i.e., $u^{(0)} = \underline{u} = v/2$ and

$$(A - 10I)u^{(j+1)} + \phi(u^{(j)}) + 2f_5(u^{(j)}) = 0$$

for $n = 0, 1, 2, \dots, 20$. Note that $A - 10I = A - \lambda 5I$. The first and last iterates in this computation are

$$\begin{aligned} u^{(0)} &= \begin{bmatrix} 0.5 \\ 0.7071067810 \\ 0.5 \end{bmatrix}, \\ u^{(1)} &= \begin{bmatrix} 0.536255893875059 \\ 0.711597532800704 \\ 0.517633127733392 \end{bmatrix}, \\ u^{(2)} &= \begin{bmatrix} 0.569589664351000 \\ 0.719076398251377 \\ 0.532898818433164 \end{bmatrix}, \\ &\vdots \\ u^{(18)} &= \begin{bmatrix} 0.805305316997942 \\ 0.841375090990183 \\ 0.639281108220394 \end{bmatrix}, \\ u^{(19)} &= \begin{bmatrix} 0.809176054165159 \\ 0.844281463851946 \\ 0.641524507653473 \end{bmatrix}, \\ u^{(20)} &= \begin{bmatrix} 0.812508968309087 \\ 0.846813682192274 \\ 0.643500032182824 \end{bmatrix}. \end{aligned}$$

We used Maple to do the computations. We see that the vectors are increasing and approaching a solution $u \approx u^{(20)}$. We can add in the boundary components of u by recalling that $u_4 = 0$ and $u_0 = \phi(u_1) \approx \phi(0.81) \approx 0.50$.

5. Conclusion

Using monotonicity methods we were able to find nonnegative solutions to (1) with two different nontrivial boundary conditions of importance to applications.

For possible future research questions consider the following. Basic issues of uniqueness, multiplicity, and stability remain to be explored. In the case of multiple solutions, determining the shape and asymptotic properties of bifurcation curves

would be of interest. Establishing the convergence of discrete solutions to continuous solutions as $n \rightarrow \infty$ would be of interest. A good reference for this last question, and other interesting computational issues, is [Lewis et al. 2022]. Pairing the boundary conditions with other biologically interesting choices of $f(u)$, as in [Cantrell and Cosner 2007; Ashley et al. 2013; Goddard and Shivaji 2017] would be interesting. Finally, combining elements of the two different problems discussed above would be a good direction for further study, i.e., nonlinear nonlocal boundary conditions.

Acknowledgements

The authors are grateful to the referees for their careful and thoughtful reading which led to a significant improvement in the paper. Any remaining errors and/or oversights are still the responsibility of the authors.

References

- [Ashley et al. 2013] K. Ashley, V. Sincavage, and J. Goddard, II, “Ecological systems, nonlinear boundary conditions, and Σ -shaped bifurcation curves”, *Involve* **6**:4 (2013), 399–430. [MR](#) [Zbl](#)
- [Berman and Plemmons 1979] A. Berman and R. J. Plemmons, *Nonnegative matrices in the mathematical sciences*, Academic Press, New York, 1979. [MR](#) [Zbl](#)
- [Bruno 2021] S. Bruno, *Confirming multiple solutions in the combustion and logistic equations*, undergraduate thesis, Wake Forest University, 2021.
- [Cantrell and Cosner 2006] R. S. Cantrell and C. Cosner, “On the effects of nonlinear boundary conditions in diffusive logistic equations on bounded domains”, *J. Differential Equations* **231**:2 (2006), 768–804. [MR](#) [Zbl](#)
- [Cantrell and Cosner 2007] R. S. Cantrell and C. Cosner, “Density dependent behavior at habitat boundaries and the Allee effect”, *Bull. Math. Biol.* **69**:7 (2007), 2339–2360. [MR](#) [Zbl](#)
- [Goddard and Shivaji 2017] J. Goddard, II and R. Shivaji, “Stability analysis for positive solutions for classes of semilinear elliptic boundary-value problems with nonlinear boundary conditions”, *Proc. Roy. Soc. Edinburgh Sect. A* **147**:5 (2017), 1019–1040. [MR](#) [Zbl](#)
- [Lewis et al. 2022] T. Lewis, Q. Morris, and Y. Zhang, “Convergence, stability analysis, and solvers for approximating sublinear positive and semipositone boundary value problems using finite difference methods”, *J. Comput. Appl. Math.* **404** (2022), art. id. 113880. [MR](#) [Zbl](#)
- [Sattinger 1972] D. H. Sattinger, “Monotone methods in nonlinear elliptic and parabolic boundary value problems”, *Indiana Univ. Math. J.* **21** (1972), 979–1000. [MR](#) [Zbl](#)
- [Weisstein] E. W. Weisstein, “Chebyshev polynomial of the second kind”, available at <https://mathworld.wolfram.com/ChebyshevPolynomialoftheSecondKind.html>. From MathWorld.
- [Wikipedia] “Eigenvalues and eigenvectors of the second derivative”, Wikipedia entry, available at https://en.wikipedia.org/wiki/Eigenvalues_and_eigenvectors_of_the_second_derivative.

Received: 2022-06-28

Revised: 2023-01-14

Accepted: 2023-01-16

carson.rodriguez@alumni.wfu.edu *Department of Mathematics, Wake Forest University, Winston-Salem, NC, United States*

sbr@wfu.edu *Department of Mathematics, Wake Forest University, Winston-Salem, NC, United States*

Edge-determining sets and determining index

Sally Cockburn and Sean McAvoy

(Communicated by Anant Godbole)

A graph automorphism is a bijective mapping of the vertices that preserves adjacent vertices. A vertex-determining set of a graph is a set of vertices such that the only automorphism that fixes those vertices is the identity. The size of a smallest such set is called the determining number, denoted by $\text{Det}(G)$. The determining number is a parameter of the graph capturing its level of symmetry. We introduce the related concept of an edge-determining set and determining index, $\text{Det}'(G)$. We prove that $\text{Det}'(G) \leq \text{Det}(G) \leq 2\text{Det}'(G)$ when $\text{Det}(G) \neq 1$ and show both bounds are sharp for infinite families of graphs. Further, we investigate properties of these new concepts, as well as provide the determining index for several families of graphs, including hypercubes.

1. Introduction

We focus solely on finite, simple graphs, $G = (V, E)$, using standard terminology and notation, as can be found in [Chartrand and Zhang 2012]. A definition of particular interest for this paper is that an automorphism of a graph is a permutation of the set of vertices that respects vertex adjacency. The set of all automorphisms of a graph under the operation of composition forms a group, denoted by $\text{Aut}(G)$.

Graph theorists often try to categorize graphs by their level of symmetry. Graphs with vertices or edges that are interchangeable display one such type of symmetry. More precisely, a graph is vertex-transitive if, for all $u, v \in V(G)$, there exists $\phi \in \text{Aut}(G)$ such that $\phi(u) = v$. Edge-transitivity and arc-transitivity are defined similarly. Additionally, we will call a graph edge-flip-invariant if, for all $\{u, v\} \in E(G)$, there exists $\phi \in \text{Aut}(G)$ such that $\phi(u) = v$ and $\phi(v) = u$. In Section 2, we will show that for connected graphs, edge-flip-invariance is a stronger condition than vertex-transitivity.

The determining number of a graph is another measure of its symmetry. The motivation is to find the least number of vertices that need to be fixed to break all

MSC2020: primary 05C25; secondary 05C70.

Keywords: determining number, distinguishing index, hypercubes.

This research was supported by The Monica Odening '05 Student Internship and Research Fund in Mathematics.

of the symmetries of a graph. Intuitively, the more vertices that need to be fixed, the more symmetric the graph is.

Definition 1. A vertex subset S of a graph is a *vertex-determining set* if the only automorphism $\phi \in \text{Aut}(G)$ that satisfies $\phi(u) = u$ for all $u \in S$ is the identity automorphism. The *determining number* of G , $\text{Det}(G)$, is the size of a smallest such set: $\text{Det}(G) = \min\{|S| : S \text{ is a vertex-determining set}\}$.

Some authors use the term fixing number for this parameter (see [Erwin and Harary 2006; Gibbons and Laison 2009]), but we will continue to refer to it as the determining number. Another way to break the symmetries of a graph is to assign a color to each vertex in such a way that otherwise interchangeable vertices can be distinguished. (Notice that this need not be a proper vertex coloring, in which adjacent vertices must be assigned different colors.)

Definition 2. The *distinguishing number* of a graph G , $\text{Dist}(G)$, is the minimum number of vertex colors required so that the only automorphism $\phi \in \text{Aut}(G)$ that preserves all vertex colors is the identity automorphism.

Albertson and Boutin [2007] established a relationship between the determining number and distinguishing number.

Proposition 3 [Albertson and Boutin 2007, Theorem 3]. *We have $\text{Dist}(G) \leq \text{Det}(G) + 1$ for all graphs G .*

The idea behind the proof is that the vertices of a minimum vertex-determining set can be colored with $\text{Det}(G)$ distinct colors, and all other vertices with another color. Since the only automorphism that fixes the vertices of the vertex-determining set is the identity, this will be a distinguishing coloring with $\text{Det}(G) + 1$ colors.

Kalinowski and Piłśniak [2015] introduced the distinguishing index, based on coloring edges instead of vertices.

Definition 4. The *distinguishing index* of a graph G , $\text{Dist}'(G)$, is the minimum number of edge colors so that the only automorphism $\phi \in \text{Aut}(G)$ that preserves all edge colors is the identity automorphism.

There has been further recent research on the distinguishing index; see for example [Alikhani and Soltani 2020a; Imrich et al. 2020; Lehner et al. 2020]. Motivated by this work, in this paper we extend the concept of determining set from vertices to edges.

One complication is the following. An automorphism $\phi \in \text{Aut}(G)$ fixes an edge $e = \{u, v\} \in E(G)$ if $\phi(e) = e$, which means $\{\phi(u), \phi(v)\} = \{u, v\}$. In this case, ϕ either fixes the endvertices ($\phi(u) = u$ and $\phi(v) = v$) or switches the endvertices ($\phi(u) = v$ and $\phi(v) = u$). The following elementary observations will be useful in our investigation.

Observation 5. Let e_1, e_2 be adjacent edges in a graph G . If $\phi \in \text{Aut}(G)$ fixes both e_1 and e_2 , then ϕ fixes the endvertices of e_1 and e_2 .

Observation 6. Let $e = \{u, v\}$ be an edge in graph G such that $\deg(u) \neq \deg(v)$. If $\phi \in \text{Aut}(G)$ fixes e , then ϕ fixes the endvertices of e .

The open neighborhood $N(v)$ of a vertex v is the set of all neighbors of v ; the closed neighborhood of v is $N[v] = \{v\} \cup N(v)$. Distinct vertices u and v are called nonadjacent twins (respectively, adjacent twins) if $N(u) = N(v)$ (respectively, $N[u] = N[v]$).

Observation 7. If u, v are adjacent or nonadjacent twins in a graph G , then there exists $\phi \in \text{Aut}(G)$ such that $\phi(v) = u$, $\phi(u) = v$, and ϕ fixes all the other vertices.

Any two isolated vertices in a graph G will be nonadjacent twins, so there exists an automorphism switching them and leaving all other vertices fixed. This nontrivial automorphism fixes all edges of G . Similarly, if there is a K_2 component of G , then its endvertices are adjacent twins. The automorphism switching the endvertices of this component and leaving all other vertices fixed is a nontrivial automorphism that fixes every edge in the graph. In both of these situations, it is not possible to define an edge equivalent of the determining number. With these considerations in mind, we make the following definition.

Definition 8. Let G be a graph with no more than one isolated vertex and without K_2 as a component. An edge subset T of G is an *edge-determining set* if the only $\phi \in \text{Aut}(G)$ that satisfies $\{\phi(u), \phi(v)\} = \{u, v\}$ for all $\{u, v\} \in T$ is the identity automorphism. The *determining index*, $\text{Det}'(G)$, is the size of the smallest such set: $\text{Det}'(G) = \min\{|T| : T \text{ is an edge-determining set}\}$.

In other words, T is an edge-determining set if fixing every edge of T fixes every vertex of the graph. If G is an asymmetric graph, meaning it has no nontrivial automorphisms, then both $\text{Det}(G) = 0$ and $\text{Det}'(G) = 0$. Due to the novelty of an edge-determining set, previous authors have referred to a vertex-determining set as simply a determining set. Henceforth, we will indicate clearly whether a determining set consists of edges or vertices.

The determining index of a disconnected graph is not as simple as the sum of the determining index of each of its connected components. For example, if G is the union of two isomorphic, asymmetric, connected graphs, then the determining index is 0 for each connected component, but $\text{Det}'(G) = 1$, because there is a nontrivial automorphism switching the two components.

Erwin and Harary [2006] observed the corresponding result for determining number of disconnected graphs.

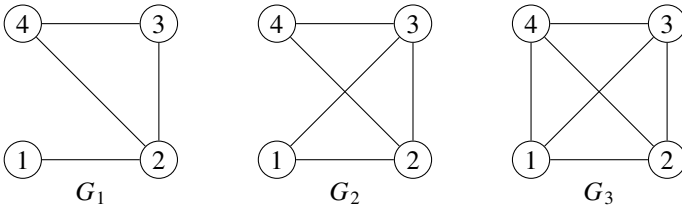


Figure 1. Exceptions to $\text{Aut}(G) \cong \text{Aut}(L(G))$.

Proposition 9 [Erwin and Harary 2006, Observation 1]. *Let $G = H_1 \cup \dots \cup H_k$, where H_i is a connected component of G . Let*

$$P = \{H_i : \text{Det}(H_i) > 0\}, \quad Q = \{H_i : \text{Det}(H_i) = 0\}$$

and let R be the set of isomorphism classes of graphs in Q . Then

$$\text{Det}(G) = |Q| - |R| + \sum_{H \in P} \text{Det}(H).$$

An analogous result holds for the determining index.

Lemma 10. *Let $G = H_1 \cup \dots \cup H_k$, where H_i is a connected component of G , at most one $H_i = K_1$, and no $H_i = K_2$. Let*

$$P = \{H_i : \text{Det}'(H_i) > 0\}, \quad Q = \{H_i : \text{Det}'(H_i) = 0\}$$

and let R be the set of isomorphism classes of graphs in Q . Then

$$\text{Det}'(G) = |Q| - |R| + \sum_{H \in P} \text{Det}'(H).$$

Proof. If $H_i \in P$, then clearly we need to fix $\text{Det}'(H_i)$ edges to fix all vertices of H_i .

If $H_i \in Q$, then, for all $H_j \in Q$ such that $i \neq j$ and $H_i \cong H_j$, there exists an automorphism of G that interchanges the vertices of these two components. Let k be the number of connected components in the isomorphism class of H_i . We need to fix at least $k - 1$ of these components. Fix one edge of each of the components in the isomorphism class of H_i except one. Then the vertices of all the components in the isomorphism class of H_i are fixed. We need to do this procedure for each distinct isomorphism class in Q . Thus, $\text{Det}'(G) = |Q| - |R| + \sum_{H \in P} \text{Det}'(H)$. \square

A useful tool for investigating connections between vertex properties and edge properties of a graph is the line graph. The line graph of a graph G is the graph $L(G)$ such that $V(L(G)) = E(G)$, with vertices $e_1 = \{u, v\}$ and $e_2 = \{x, y\}$ in $L(G)$ being adjacent if the corresponding edges are adjacent in G . It is easy to verify that an automorphism α of G induces an automorphism α' of $L(G)$ in the obvious way; for all $e = \{u, v\} \in V(L(G))$, define $\alpha'(e) = \{\alpha(u), \alpha(v)\}$. Sabidussi [1961] proved that except for the three cases shown in Figure 1, these induced automorphisms make up the entire automorphism group of $L(G)$.

Proposition 11 [Sabidussi 1961, Theorem 5.3]. *Let G be a connected graph such that $|V(G)| \geq 3$. If $G \notin \{G_1, G_2, G_3\}$, then $\text{Aut}(G) \cong \text{Aut}(L(G))$.*

Alikhani and Soltani established a relationship between the distinguishing index of a graph and the distinguishing number of the corresponding line graph.

Proposition 12 [Alikhani and Soltani 2020b, Theorem 2.3]. *If G is a connected graph such that $|V(G)| \geq 3$ and $G \neq G_2$, then $\text{Dist}'(G) = \text{Dist}(L(G))$.*

Using Proposition 11, we establish a relationship, similar to that of Proposition 12, between the determining index of a graph and the determining number of the corresponding line graph.

Theorem 13. *Let G be a connected graph such that $|V(G)| \geq 3$. Then $\text{Det}'(G) = \text{Det}(L(G))$ if and only if $G \notin \{G_1, G_2, G_3\}$.*

Proof. Assume $G \notin \{G_1, G_2, G_3\}$. We will first show $\text{Det}'(G) \geq \text{Det}(L(G))$. Let T be a minimum edge-determining set of G so that $\text{Det}'(G) = |T|$. Then T is a set of vertices in $L(G)$. Let α be an automorphism of $L(G)$ such that $\alpha(v) = v$ for all $v \in T$. By Proposition 11, there exists an isomorphism $\phi : \text{Aut}(G) \rightarrow \text{Aut}(L(G))$, so there exists $\sigma \in \text{Aut}(G)$ such that $\phi(\sigma) = \alpha$. Hence, for all edges $e = \{u, v\} \in E(G)$, $\alpha(e) = \{\sigma(u), \sigma(v)\}$. But since $\alpha(e) = e$ for all $e \in T$, $\{\sigma(u), \sigma(v)\} = \{u, v\}$ for all $\{u, v\} \in T$. Since T is an edge-determining set in G , σ is the identity in $\text{Aut}(G)$. Hence, α is the identity in $\text{Aut}(L(G))$. By definition, T is a vertex-determining set in $L(G)$. Since T is of minimum size as an edge-determining set, $\text{Det}'(G) \geq \text{Det}(L(G))$.

We will now show $\text{Det}(L(G)) \geq \text{Det}'(G)$. Let S be a minimum vertex-determining set of $L(G)$ so that $\text{Det}(L(G)) = |S|$. Then S is a set of edges in G . Let $\tau \in \text{Aut}(G)$ such that $\{\tau(u), \tau(v)\} = \{u, v\}$ for all $\{u, v\} \in S$. By Proposition 11, there is a unique $\beta \in \text{Aut}(L(G))$ such that $\phi(\tau) = \beta$. Then, for all $e = \{u, v\} \in S$, $\beta(e) = \{\tau(u), \tau(v)\} = \{u, v\} = e$. Since S is a vertex-determining set in $L(G)$, β is the identity automorphism of $L(G)$. Since $\phi : \text{Aut}(G) \cong \text{Aut}(L(G))$ is an isomorphism, the only element of $\ker(\phi)$ is the identity automorphism of G . Thus, τ is the identity and so S is an edge-determining set of G . Since S is of minimum size as a vertex-determining set of $L(G)$, $\text{Det}'(G) \leq \text{Det}(L(G))$. Hence, $\text{Det}'(G) = \text{Det}(L(G))$.

Conversely, assume $G \in \{G_1, G_2, G_3\}$. By inspection, we have $\text{Det}'(G_1) = \text{Det}'(G_2) = 1$ and $\text{Det}'(G_3) = 2$. For the determining number of their line graphs, shown in Figure 2, we have $\text{Det}(L(G_1)) = 2$, $\text{Det}(L(G_2)) = 2$, and $\text{Det}(L(G_3)) = 3$. Hence, if $G \in \{G_1, G_2, G_3\}$, then $\text{Det}'(G) \neq \text{Det}(L(G))$. \square

With the connection between G and $L(G)$, we now have a tool to find the determining index, if the determining number of the line graph is known.

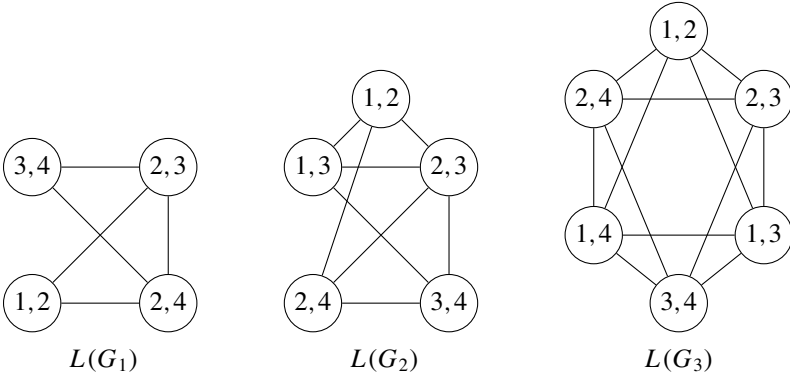


Figure 2. Line graphs to the exceptions.

Theorem 14. For all $n > 2$, $\text{Det}'(P_n) = 1$ and $\text{Det}'(C_n) = 2$, and, for all $n > 1$, $\text{Det}'(K_{1,n}) = n - 1$.

Proof. Note that $P_{n-1} = L(P_n)$, $C_n = L(C_n)$, and $K_n = L(K_{1,n})$. The determining numbers of paths, cycles and complete graphs follow easily from the definition; for $n > 2$, $\text{Det}(P_{n-1}) = 1$ and $\text{Det}(C_n) = 2$, and, for $n > 1$, $\text{Det}(K_n) = n - 1$. Now apply [Theorem 13](#). □

We can easily establish a similar relationship between the distinguishing index and the determining index to that between the distinguishing number and determining number in [Proposition 3](#).

Theorem 15. Let G be a connected graph such that $|V(G)| \geq 3$. Then $\text{Dist}'(G) \leq \text{Det}'(G) + 1$.

Proof. We first cover the special cases by inspection. If $G = G_1$, we have $\text{Dist}'(G_1) = 1 \leq 2 = \text{Det}'(G_1) + 1$. If $G = G_2$, we have $\text{Dist}'(G_2) = 1 \leq 2 = \text{Det}'(G_2) + 1$. If $G = G_3$, we have $\text{Dist}'(G_3) = 3 = \text{Det}'(G_3) + 1$. If $G \neq G_1, G_2$, or G_3 , then, by [Propositions 3](#) and [12](#),

$$\text{Dist}'(G) = \text{Dist}(L(G)) \leq \text{Det}(L(G)) + 1 = \text{Det}'(G) + 1. \quad \square$$

2. Comparing determining number and determining index

In this section, we discuss general relationships between the determining index and determining number. One important difference between these parameters is that while $\text{Det}(G) = \text{Det}(\bar{G})$, where \bar{G} is the complement of G , the same is not always true for the determining index. This can be because the determining index is undefined for the complement.

Example 16. By [Theorem 14](#), $\text{Det}'(C_3) = 2$. However, $\text{Det}'(\bar{C}_3)$ is undefined as the number of isolated vertices is 3.



Figure 3. Comparing $\text{Det}(G)$ and $\text{Det}'(G)$ on P_4 .

However, this is not always a result of $\text{Det}'(G)$ being undefined.

Example 17. By [Theorem 14](#), $\text{Det}'(K_{1,5}) = 4$. However, $\bar{K}_{1,5} = K_5 \cup K_1$, so $\text{Det}'(\bar{K}_{1,5}) = 3$. With only one isolated vertex, the edge-determining index of $\bar{K}_{1,5} = K_5 \cup K_1$ is still defined.

By [Theorem 13](#), knowing when $\text{Det}'(G) = \text{Det}'(\bar{G})$ can be viewed as a problem of knowing when $\text{Det}(L(G)) = \text{Det}(L(\bar{G}))$. Alternatively, we can ask when $\text{Det}(G) = \text{Det}'(G)$ and $\text{Det}(\bar{G}) = \text{Det}'(\bar{G})$ as $\text{Det}(G) = \text{Det}(\bar{G})$. If G is self-complementary, then the equality clearly holds.

Kalinowski and Piłśniak [\[2015\]](#) showed that $\text{Dist}'(G) \leq \text{Dist}(G) + 1$ for any graph G . It is natural to expect that there is a similar relationship between the determining index and the determining number of a graph. Before establishing a connection, we present some examples displaying the subtleties of constructing an edge-determining set from a given vertex-determining set. For a given vertex-determining set S , it seems reasonable to believe that we can construct a corresponding edge-determining set by picking one edge incident to each vertex in S . In [Figure 3](#), left, the gray vertex constitutes a determining set, but the incident bold edge fails to constitute an edge-determining set. However, [Figure 3](#), right, shows that the other incident edge does constitute an edge-determining set. In this example, $\text{Det}(P_4) = \text{Det}'(P_4) = 1$.

The example of P_4 fails to highlight all possible difficulties in selecting edges incident to vertices. While the edge we chose mattered, there was an obvious one. The endvertices of $\{3, 4\}$ have different degrees, so fixing that edge will fix its endvertices by [Observation 6](#). The choice is less obvious in a graph with more symmetry, such as the “envelope” graph H in [Figure 4](#), which is both vertex-transitive and edge-flip-invariant.

For this example, it is easily verified that $\{3, 5\}$ is a minimum vertex-determining set, so $\text{Det}(H) = 2$. In [Figure 4](#), left, the bold edges fail to constitute an edge-determining set because the reflection across a central vertical line in the drawing is an automorphism that flips edges $\{3, 4\}$ and $\{5, 6\}$ (and $\{1, 2\}$, in fact). However, [Figure 4](#), right, indicates that a different choice of one incident edge per vertex can still produce an edge-determining set. However, trying to use only the one edge between vertices 3 and 5 won't work because there is an automorphism that flips this edge (as well as the edge $\{4, 6\}$). The same is true of any singleton edge set, because H is edge-flip-invariant. Thus, $\text{Det}'(H) = 2$. With these nuances in mind, we prove the following. Recall that the distance $d(u, v)$ between vertices u and v is

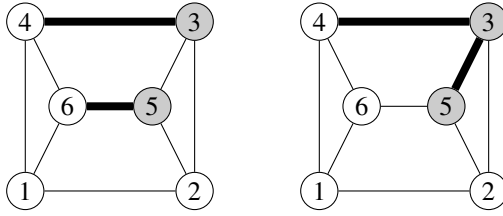


Figure 4. Comparing $\text{Det}(H)$ (left) and $\text{Det}'(H)$ (right).

the minimum number of edges in a $u - v$ path; a path of minimum length is called a $u - v$ geodesic. As is shown in [Chartrand and Zhang 2012], any subpath of a geodesic is also a geodesic.

Theorem 18. *Let G be a connected graph such that $|V(G)| \geq 3$. If $\text{Det}(G) \neq 1$, then $\text{Det}'(G) \leq \text{Det}(G)$.*

Proof. Let $S = \{s_1, \dots, s_k\}$ be a minimum vertex-determining set of G such that $k > 1$.

First suppose all the vertices of S are pairwise adjacent. Then there exists a path P such that $V(P) = S$. Let T be the edges of P , so that $|T| = k - 1$. If $k = 2$, there exists an edge e adjacent to $\{s_1, s_2\}$, since G is connected with $|V(G)| \geq 3$. In this case, we add e to T to create an edge set of size $k = 2$. Assume $\phi \in \text{Aut}(G)$ fixes the edges of T . Let $\{u, v\}, \{v, w\} \in T$ be adjacent edges. By Observation 5, if two adjacent edges are fixed, then the endvertices of those edges are fixed. Thus, $\phi(u) = u$, $\phi(v) = v$, $\phi(w) = w$. Working our way along the adjacent edges of path P , we can conclude that all the endvertices of the edges of T are fixed. Therefore, $\phi(s) = s$ for all $s \in S$ and so $\text{Det}'(G) \leq \text{Det}(G) = k$.

Now suppose not all of the vertices are pairwise adjacent. Then by renumbering the vertices in S if necessary, we can assume $d(s_1, s_2) \geq 2$. Let $P = \{s_1, u_1, \dots, u_2, s_2\}$ be an $s_1 - s_2$ geodesic. Note that it is possible for $u_1 = u_2$. We will inductively construct an edge-determining set consisting of edges incident to vertices of S . We start with the two edges $e_1 = \{s_1, u_1\}$ and $e_2 = \{s_2, u_2\}$. Let $T_2 = \{e_1, e_2\}$. Assume $\phi \in \text{Aut}(G)$ fixes these two edges. If $u_1 = u_2$, then the edges are adjacent, so, by Observation 5, $\phi(s_1) = s_1$ and $\phi(s_2) = s_2$. Otherwise, the edges are nonadjacent. In this case, suppose that ϕ flips e_1 but fixes the vertices of e_2 . Then $\phi(s_1) = u_1$, $\phi(u_1) = s_1$, and $\phi(s_2) = s_2$. Since automorphisms are distance-preserving, $d(s_1, s_2) = d(\phi(s_1), \phi(s_2)) = d(u_1, s_2)$. This contradicts our assumption that P is a geodesic. Now suppose ϕ flips both e_1 and e_2 , so that $\phi(s_1) = u_1$, $\phi(u_1) = s_1$, and $\phi(s_2) = u_2$. Again, since automorphisms are distance-preserving, $d(s_1, s_2) = d(\phi(s_1), \phi(s_2)) = d(u_1, u_2)$. This also contradicts our assumption that P is a geodesic. Thus, if ϕ fixes the edges of T_2 , then $\phi(s_1) = s_1$, and $\phi(s_2) = s_2$.

Next let $2 < i \leq k$ and let $T_{i-1} = \{e_1, e_2, \dots, e_{i-1}\}$ be a set of edges of the form $e_j = \{s_j, u_j\}$, with $d(s_1, u_j) < d(s_1, s_j)$. Assume that any automorphism

fixing every edge of T_{i-1} also fixes the vertices s_1, \dots, s_{i-1} . Since G is connected, there exists a path from s_1 to s_i . Let $e_i = \{s_i, u_i\}$ such that u_i is adjacent to s_i on an $s_1 - s_i$ geodesic. Thus $d(s_1, u_i) < d(s_1, s_i)$. Note that it is possible that $u_i = s_j$ for some $j \in \{1, \dots, i-1\}$. However, in that case it is not possible that $e_i = \{s_i, u_i\} = \{s_i, s_j\}$ is already in T_{i-1} , because this would imply $s_i = u_j$, which in turn would imply

$$d(s_1, s_j) = d(s_1, u_i) < d(s_1, s_i) = d(s_1, u_j) < d(s_1, s_j).$$

Let $T_i = T_{i-1} \cup \{e_i\}$.

Let $\phi \in \text{Aut}(G)$ be an automorphism of G that fixes the edges of T_i . Then ϕ must fix the edges of T_{i-1} and so by assumption, ϕ fixes s_1, \dots, s_i . If $u_i = s_j$ for some $j \in \{1, \dots, i-1\}$, then since $\phi(u_i) = \phi(s_j) = s_j = u_i$, it must be the case that $\phi(s_i) = s_i$ also. Otherwise, assume that ϕ flips edge e_i ; that is, assume that $\phi(s_i) = u_i$ and $\phi(u_i) = s_i$. Then since automorphisms are distance-preserving, $d(s_1, s_i) = d(\phi(s_1), \phi(s_i)) = d(s_1, u_i) < d(s_1, s_i)$, a contradiction. Thus, $\phi(s_i) = s_i$.

Finally, let $T = T_k$. If $\phi \in \text{Aut}(G)$ fixes every edge in T , then $\phi(s_i) = s_i$ for all $s_i \in S$. Since S is a determining set, ϕ is the identity automorphism. Therefore, by definition T is an edge-determining set. For each vertex in S , we added a distinct edge when constructing T . Therefore, $|T| = |S|$. Thus, $\text{Det}'(G) \leq \text{Det}(G)$. \square

Theorem 19. *Let G be a connected graph such that $|V(G)| \geq 3$. Then $\text{Det}(G) \leq 2 \text{Det}'(G)$.*

Proof. Let T be a minimum edge-determining set of G and let S be the set of endvertices of the edges in T . If the edges of T are all nonadjacent, then $|S| = 2|T|$; allowing for the possibility that some vertices are endvertices of two or more edges of T means $|S| \leq 2|T| = 2 \text{Det}'(G)$. Let $\phi \in \text{Aut}(G)$ such that ϕ fixes the vertices of S . Let $\{u, w\} \in T$. Then $\{\phi(u), \phi(w)\} = \{u, w\}$, since $u, w \in S$. Since T is an edge-determining set, ϕ is the identity automorphism. Therefore, by definition, S is a vertex-determining set. Thus, $\text{Det}(G) \leq 2 \text{Det}'(G)$. \square

If $\text{Det}(G) \neq 1$, then together Theorems 18 and 19 give us the inequality $\text{Det}'(G) \leq \text{Det}(G) \leq 2 \text{Det}'(G)$. We can algebraically rewrite this inequality to give bounds on $\text{Det}'(G)$ in terms of $\text{Det}(G)$.

Corollary 20. *Let G be a connected graph such that $|V(G)| \geq 3$. If $\text{Det}(G) \neq 1$, then $\frac{1}{2} \text{Det}(G) \leq \text{Det}'(G) \leq \text{Det}(G)$.*

We have so far avoided the case when $\text{Det}(G) = 1$. In order to make claims about $\text{Det}'(G)$ in this case, we need to establish results about edge-flip-invariant graphs. Recall that G is edge-flip-invariant if for all $\{u, v\} \in E(G)$, there exists $\phi \in \text{Aut}(G)$ such that $\phi(u) = v$ and $\phi(v) = u$. We now introduce the following definition.

Definition 21. A vertex $v \in V(G)$ has the *neighbor-swapping property* if, for all $u \in N(v)$, there exists $\phi \in \text{Aut}(G)$ such that $\phi(v) = u$ and $\phi(u) = v$.

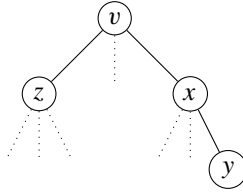


Figure 5. $x \in N(v)$ and $y \in N(x)$, $y \neq v$.

Clearly, all the vertices of K_n have this property. Many other symmetric graphs have vertices with this property such as H in Figure 4, the hypercubes Q_n and cycles C_n .

Theorem 22. *For a connected graph G , the following are equivalent:*

- (a) *There exists $v \in V(G)$ that has the neighbor-swapping property.*
- (b) *Every $v \in V(G)$ has the neighbor-swapping property.*
- (c) *G is edge-flip-invariant.*

In this case, G is vertex-transitive.

Proof. All statements in this theorem are clearly true if $G = K_1$ or K_2 . So assume $|V(G)| \geq 3$.

Assume there exists $v \in V(G)$ such that v has the neighbor-swapping property. Let $x \in N(v)$ and $y \in N(x)$ such that $y \neq v$, as shown in Figure 5. By assumption, there exists $\alpha \in \text{Aut}(G)$ such that $\alpha(x) = v$ and $\alpha(v) = x$. Then $\alpha(y) \in N(\alpha(x)) = N(v)$. Hence, there exists $z \in N(v)$ such that $\alpha(y) = z$. Since v has the neighbor-swapping property, there exists $\beta \in \text{Aut}(G)$ such that $\beta(z) = v$ and $\beta(v) = z$. Define $\sigma : V(G) \rightarrow V(G)$ by $\sigma = \alpha^{-1}\beta\alpha$. Then $\sigma(x) = y$ and $\sigma(y) = x$. Hence, for all $y \in N(x)$, there is an automorphism switching x and y . Hence, x has the neighbor-swapping property.

Now let $w \in V(G)$. Since G is connected, there exists a $v - w$ path $P = (v, u_1, \dots, u_k, w)$. By the argument above, u_1 has the neighbor-swapping property. Similarly, since $u_2 \in N(u_1)$, we know u_2 has the neighbor-swapping property. Continuing along the edges of the path, w must have the neighbor-swapping property. Thus, all the vertices of G are interchangeable with their neighbors, and so (a) implies (b).

If G satisfies (b), then by definition, G is edge-flip-invariant.

Now assume G is edge-flip-invariant. Let $v \in V(G)$ and $u \in N(v)$. By definition, there exists $\phi \in \text{Aut}(G)$ such that $\phi(u) = v$ and $\phi(v) = u$. Thus, there exists $v \in V(G)$ such that v has the neighbor-swapping property, and so (c) implies (a).

Next we show that if G is connected and edge-flip-invariant, then G is vertex-transitive. Let $v, w \in V(G)$. Since G is connected, there exists a $v - w$ path, $P = (v = u_1, u_2, \dots, u_k = w)$. Since G is edge-flip-invariant, there exists $\phi_i \in \text{Aut}(G)$

such that $\phi_i(u_i) = u_{i+1}$ and $\phi_i(u_{i+1}) = u_i$ for all $1 \leq i \leq k$. Define $\phi: V(G) \rightarrow V(G)$ by $\phi = \phi_{k-1} \circ \cdots \circ \phi_1$. Then $\phi(u_1) = u_k$. Thus, G is vertex-transitive. \square

The converse of [Theorem 22](#) does not hold; that is, not all vertex-transitive graphs are edge-flip-invariant.

Example 23. The Holt graph [1981] is the smallest graph that is vertex-transitive, edge-transitive, but not arc-transitive. Since it is edge-transitive, if there is an automorphism that flips any one edge, then it can be composed with other automorphisms to create an automorphism taking any arc to any other arc. This contradicts the fact that it is not arc-transitive.

Now we can establish the determining index when $\text{Det}(G) = 1$. In this case, it is possible that $\text{Det}'(G) > \text{Det}(G)$.

Corollary 24. *Let G be a connected graph such that $|V(G)| \geq 3$ and $\text{Det}(G) = 1$. Then*

$$\text{Det}'(G) = \begin{cases} 2 & \text{if } G \text{ is edge-flip-invariant,} \\ 1 & \text{otherwise.} \end{cases}$$

Proof. Let $S = \{v\}$ be a minimum vertex-determining set of G .

If G is not edge-flip-invariant, then by [Theorem 22](#), no vertex, including v , has the neighbor-swapping property. Thus, there exists $u \in N(v)$ such that no automorphism interchanges u and v . Let $T = \{\{u, v\}\}$. Let $\phi \in \text{Aut}(G)$ such that $\{\phi(u), \phi(v)\} = \{u, v\}$. Then $\phi(u) = u$ and $\phi(v) = v$. Since the vertex of S is fixed, $\phi(w) = w$ for all $w \in V(G)$. Hence, T is a minimum edge-determining set and $\text{Det}'(G) = 1$.

Now assume that G is edge-flip-invariant. By definition, there is a nontrivial automorphism that fixes any given edge. Then clearly $\text{Det}'(G) > 1$. Since G is connected with $|V(G)| \geq 3$, we can find two adjacent edges e_1 and e_2 such that $v \in e_1 \cup e_2$. Let $T = \{e_1, e_2\}$ and assume that $\phi \in \text{Aut}(G)$ fixes the edges of T . Then by [Observation 5](#), the endvertices of the edges are fixed. Since the vertex of S is fixed, $\phi(w) = w$ for all $w \in V(G)$. Hence, T is a minimum determining set and $\text{Det}'(G) = 2$. \square

The following examples show that there exist graphs for both cases.

Example 25. It is well known that $\text{Det}(P_4) = 1$, and clearly P_4 is not edge-flip-invariant. Hence, in accordance with [Theorem 14](#) and as seen earlier, $\text{Det}'(P_4) = 1$.

Example 26. Brooks et al. [2021] show that any automorphism fixing any vertex of the graph G_4 shown in [Figure 6](#) will fix all other vertices of the graph. Therefore, $\text{Det}(G_4) = 1$. We can use [Theorem 22](#) to show the graph is edge-flip-invariant. Looking at the edges incident to 0, we list the following automorphisms as permutations

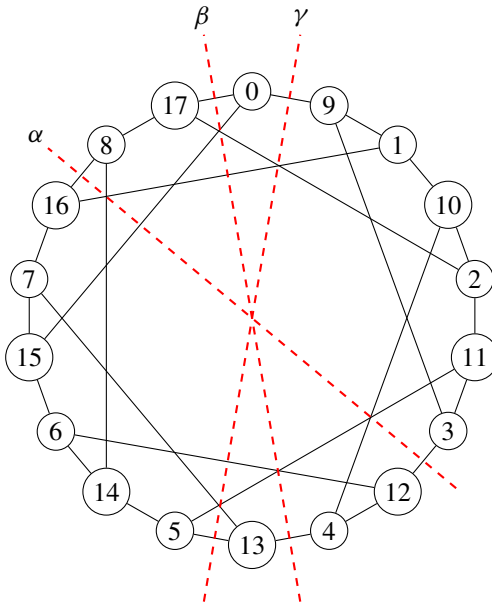


Figure 6. G_4 .

of the vertices (these are reflections across the dashed lines in [Figure 6](#)):

$$\begin{aligned} \alpha &= (0\ 15)(7\ 17)(16\ 8)(9\ 6)(1\ 14)(10\ 5)(2\ 13)(11\ 4)(3\ 12), \\ \beta &= (0\ 17)(9\ 8)(16\ 1)(7\ 10)(15\ 2)(11\ 6)(14\ 3)(12\ 5)(13\ 4), \\ \gamma &= (0\ 9)(1\ 17)(8\ 10)(16\ 2)(7\ 11)(15\ 3)(6\ 12)(14\ 4)(5\ 13). \end{aligned}$$

Thus, 0 has the neighbor-swapping property. By [Theorem 22](#), the graph is edge-flip-invariant. By [Corollary 24](#), $\text{Det}'(G_4) = 2$.

3. The determining index for some families of graphs

In this section, we find the determining index for several different families of graphs. This can give an idea of the differences between the determining index and the determining number as well as the sharpness of the bounds in [Corollary 20](#).

We first investigate complete bipartite graphs. It is easy to verify that $\text{Det}(K_{n,m}) = n + m - 2$ for $n \geq m > 1$.

Theorem 27. For $n \geq m > 1$,

$$\text{Det}'(K_{n,m}) = \begin{cases} n - 1 & \text{if } n \neq m, \\ n & \text{if } n = m. \end{cases}$$

Proof. Let $U = \{u_1, u_2, \dots, u_n\}$ and $V = \{v_1, v_2, \dots, v_m\}$ be the partite sets of $K_{n,m}$. Notice that all the vertices in U are pairwise nonadjacent twins, as are all the vertices in V .

Case 1: Assume $n > m$. By [Observation 7](#), any edge-determining set T must have at least $n - 1$ edges to cover $n - 1$ vertices in U . Let

$$T = \{\{u_i, v_i\} : 1 \leq i \leq m\} \cup \{\{u_j, v_m\} : m < j < n\}.$$

Let $\phi \in \text{Aut}(G)$ and assume ϕ fixes the edges of T . Since $n > m$, the degrees of the vertices in U and V are different. By [Observation 6](#), ϕ must fix the endvertices of every edge in T . We have fixed u_i for $1 \leq i < n$, and hence, $\phi(u_n) = u_n$. Since $|T| = n - 1$, we have $\text{Det}'(G) = n - 1$.

Case 2: Assume $n = m > 1$. Assume there exists an edge-determining set of size $n - 1$. The edges of an edge-determining set must cover $n - 1$ vertices in U and $n - 1$ vertices in V . Hence, each edge in an edge-determining set of size $n - 1$ would cover a distinct vertex in V and a distinct vertex in U . By renumbering the vertices if necessary, we can assume that $T = \{\{u_i, v_i\} : 1 \leq i \leq n - 1\}$. Now let ϕ be the nontrivial automorphism $\phi(u_i) = v_i$ and $\phi(v_i) = u_i$ for all $1 \leq i \leq n - 1$. Therefore, there does not exist an edge-determining set of size $n - 1$.

Let $T = \{\{u_i, v_i\} : 1 \leq i < n\} \cup \{u_1, v_2\}$. Assume $\phi \in \text{Aut}(G)$ fixes the edges of T . By [Observation 5](#), $\phi(u_1) = u_1$, $\phi(v_1) = v_1$, and $\phi(v_2) = v_2$. If ϕ fixes one vertex in U , then ϕ preserves the partite sets U and V setwise. Thus, if $\{\phi(u_i), \phi(v_i)\} = \{u_i, v_i\}$ for $1 \leq i < n$, then $\phi(u_i) = u_i$ and $\phi(v_i) = v_i$. We have fixed u_i for $1 \leq i < n$, and hence, $\phi(u_n) = u_n$. Similarly, we have fixed v_i for $1 \leq i < n$, and hence, $\phi(v_n) = v_n$. Since $|T| = n$, $\text{Det}'(K_{n,n}) = n$. \square

Next we look at complete graphs. Recall that $\text{Det}(K_n) = n - 1$.

Theorem 28. For $n > 2$, $\text{Det}'(K_n) = \lfloor \frac{2n}{3} \rfloor$.

Proof. Let $G = K_n$ such that $n > 2$. Note all vertices are pairwise adjacent twins. Hence, an edge-determining set of G must cover at least $n - 1$ vertices. Further, if T is an edge-determining set and $\{u, v\} \in T$, then $\{v, w\} \in T$ or $\{u, w\} \in T$ for some other $w \in V(G)$. Otherwise, there exists an automorphism that switches u and v and leaves the other vertices fixed.

Each pair of adjacent edges covers three distinct vertices. If each pair of adjacent edges in an edge-determining set covers three distinct vertices and $n \equiv 0 \pmod{3}$ or $n \equiv 1 \pmod{3}$, then these edges cover at least $n - 1$ vertices. If each pair of adjacent edges in an edge-determining set covers three distinct vertices and $n \equiv 2 \pmod{3}$, then two twin vertices would not be covered. Hence, the edge set would not be determining. Therefore, $\text{Det}'(G) \geq 2 \lfloor \frac{n}{3} \rfloor$ if $n \equiv 0 \pmod{3}$ or $n \equiv 1 \pmod{3}$. Otherwise, $\text{Det}'(G) > 2 \lfloor \frac{n}{3} \rfloor$.

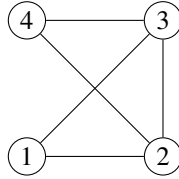


Figure 7. $K_4 - e$.

Let $V(G) = \{v_1, \dots, v_n\}$ and let

$$T = \{\{v_{i-1}, v_i\}, \{v_i, v_{i+1}\} : i \equiv 2 \pmod{3}, 0 < i < n\}.$$

Then $|T| = 2 \lfloor \frac{n}{3} \rfloor$.

Case 1: Suppose $n \equiv 0 \pmod{3}$. Then T covers all the vertices of G . Let $\phi \in \text{Aut}(G)$ and assume ϕ fixes the edges in T . By [Observation 5](#), $\phi(v_i) = v_i$, $\phi(v_{i-1}) = v_{i-1}$, and $\phi(v_{i+1}) = v_{i+1}$ for $i \equiv 2 \pmod{3}$, $0 < i < n$. Thus, T is an edge-determining set since T covers the vertices of G . Note $|T| = 2 \lfloor \frac{n}{3} \rfloor$, so T is a minimum edge-determining set. Since $n \equiv 0 \pmod{3}$, we have $\text{Det}'(G) = \lfloor \frac{2n}{3} \rfloor$.

Case 2: Suppose $n \equiv 1 \pmod{3}$. Then T covers all of the vertices of G except one. Let $\phi \in \text{Aut}(G)$ and assume ϕ fixes the edges in T . By [Observation 5](#), $\phi(v_i) = v_i$, $\phi(v_{i-1}) = v_{i-1}$ and $\phi(v_{i+1}) = v_{i+1}$ for $i \equiv 2 \pmod{3}$, $0 < i < n$. Thus, T is an edge-determining set, since T covers every vertex of G except one. Note $|T| = 2 \lfloor \frac{n}{3} \rfloor$, so T is a minimum edge-determining set. Since $n \equiv 1 \pmod{3}$,

$$\lfloor \frac{2n}{3} \rfloor = \lfloor \frac{2(n-1)}{3} + \frac{2}{3} \rfloor = 2 \lfloor \frac{n-1}{3} \rfloor = 2 \lfloor \frac{n}{3} \rfloor.$$

Case 3: Suppose $n \equiv 2 \pmod{3}$. Then there exist two vertices, v_{n-1} and v_n , not covered by an edge in T . Thus, T is not an edge-determining set. Add edge $\{v_n, v_1\}$ to T . Let $\phi \in \text{Aut}(G)$ and assume ϕ fixes the edges in T . By [Observation 5](#), $\phi(v_i) = v_i$, $\phi(v_{i-1}) = v_{i-1}$ and $\phi(v_{i+1}) = v_{i+1}$ for $i \equiv 2 \pmod{3}$, $0 < i < n$, and $\phi(v_1) = v_1$. Thus, T is an edge-determining set, since T covers every vertex of G except one. Note $|T| = 2 \lfloor \frac{n}{3} \rfloor + 1$, so T a minimum edge-determining set. Since $n \equiv 2 \pmod{3}$,

$$\lfloor \frac{2n}{3} \rfloor = \lfloor \frac{2(n-2)}{3} + \frac{4}{3} \rfloor = 2 \lfloor \frac{n-2}{3} \rfloor + 1 = 2 \lfloor \frac{n}{3} \rfloor + 1. \quad \square$$

The graphs K_n and $K_{n,m}$ illustrate that the determining index can be strictly less than the determining number, sometimes significantly. Further, the example of $K_{n,m}$ when $n > m$ indicates that a minimum edge-determining set need not simply be a cover of the vertices in a minimum vertex-determining set. Despite coming close, neither of these families show the sharpness of the upper bound $\text{Det}(G) \leq 2 \text{Det}'(G)$ in [Theorem 19](#). The following example of $K_4 - e$ indicates that the bound is sharp.

Note there are three nontrivial automorphisms of $K_4 - e$. One interchanges nonadjacent twin vertices 1 and 4, leaving 2 and 3 fixed; call it α . Another interchanges adjacent twin vertices 2 and 3, leaving 1 and 4 fixed; call it β . A third is the composition of these two automorphisms, $\alpha \circ \beta$, which geometrically is a reflection across a horizontal line through the center of the drawing. Clearly, if we fix only one vertex, then either α or β (because the composition moves all four vertices) fixes that vertex but moves others. Hence, we need to fix at least two vertices in order to fix the entire graph, say vertices 1 and 2. No nontrivial automorphism fixes both of these vertices. However, any automorphism fixing the edge $\{1, 2\}$ must fix the graph. By [Observation 6](#), no automorphism switches these two vertices as they have different degrees. Further, these vertices constitute a vertex-determining set, so this edge constitutes a minimum edge-determining set. So $\text{Det}(G) = 2 = 2 \text{Det}'(G)$. We recall a definition that will allow us to generalize this example to an infinite family of such graphs.

Definition 29. The *join* of graphs G and H is the graph $G + H$ defined by $V(G + H) = V(G) \cup V(H)$ and

$$E(G + H) = E(G) \cup E(H) \cup \{\{u, v\} : u \in G, v \in H\}.$$

Theorem 30. For all $n \in \mathbb{N}$, there exists a connected graph G such that $\text{Det}(G) = 2n$ and $\text{Det}'(G) = n$.

Proof. Let $G = N_{n+1} + K_{n+1}$ for $n \in \mathbb{N}$, where N_k is the empty graph on k vertices. Note that $K_4 - e$ is $N_2 + K_2$. Let $U = \{u_1, \dots, u_{n+1}\} = V(N_{n+1})$ and $V = \{v_1, \dots, v_{n+1}\} = V(K_{n+1})$. By construction, the vertices of U are pairwise non-adjacent twins and the vertices of V are pairwise adjacent twins. By [Observation 7](#), there exists an automorphism interchanging them and leaving all other vertices fixed. Thus, a minimum vertex-determining set must contain n vertices of U and n vertices of V . Since $\deg(u) = n + 1$ for $u \in U$ and $\deg(v) = 2n + 1$ for $v \in V$, there does not exist an automorphism interchanging the vertices of U with the vertices of V . Thus, $\text{Det}(G) = 2n$.

Similarly, an edge-determining set must cover n vertices of U and n vertices of V . Let $T = \{\{u_i, v_i\} : 1 \leq i \leq n\}$. Because there is no automorphism interchanging the vertices of U and the vertices of V , any automorphism fixing the edges of T must fix all their endvertices. Thus, $|T| = \text{Det}'(G) = n$. Therefore, $\text{Det}(G) = 2 \text{Det}'(G)$. \square

We have seen examples of infinite families such that $\text{Det}'(G) = \text{Det}(G)$, namely cycles and paths. We find a similar result for trees, where the determining number and determining index can be arbitrarily large. To demonstrate this, we require a few basic facts about trees, which can be found in any standard text on graph theory, such as [\[Chartrand and Zhang 2012\]](#). First, there is a unique path between any two distinct vertices in a tree. Second, a vertex v in a connected graph is central if

the maximum distance between v and another vertex in the graph is the minimum possible; the subgraph induced by central vertices is called the center of the graph. The center of a tree always consists either of a single vertex or a pair of adjacent vertices, and will be contained in every longest path of the tree.

Proposition 31. *The center of a tree is fixed setwise by every automorphism.*

Proof. Since path lengths are fixed by automorphisms, the image of any longest path under any automorphism ϕ is still a longest path. So the image of the center under ϕ must still be in every longest path. If the center consists of two adjacent vertices, it is possible that ϕ swaps them. \square

Erwin and Harary [2006] present many interesting results about on the determining number of trees. Our result is that determining index is always the same as the determining number.

Theorem 32. *If G is a tree such that $|V(G)| \geq 3$, then $\text{Det}(G) = \text{Det}'(G)$.*

Proof. If $\text{Det}(G) = 0$, the result is trivial. So assume $\text{Det}(G) \geq 1$.

As noted above, the center of G is either a single vertex or a pair of adjacent vertices. If the center is a pair of vertices, let v be one of them. Otherwise, let v be the center vertex.

Let T be a minimum edge-determining set of G . Because there is a unique path between v and any other vertex, for each edge in T , one endvertex of each edge must be more distant from v than the other. We let S be the set of more distant vertices. Then $|S| = |T| = \text{Det}'(G)$. To show that S is a vertex-determining set, assume $\phi \in \text{Aut}(G)$ fixes every vertex in S . Let $e = \{x, y\} \in T$, with $d(x, v) > d(y, v)$, so that $x \in S$.

Case 1: If the center is just vertex v , then by [Proposition 31](#), $\phi(v) = v$.

Case 2: Assume there are two adjacent center vertices, u and v . Then again by [Proposition 31](#), $\{\phi(u), \phi(v)\} = \{u, v\}$. By assumption $\phi(x) = x$. Suppose $\phi(v) = u$. Then $d(x, v) = d(\phi(x), \phi(v)) = d(x, u)$. However, because u and v are adjacent and trees have no cycles, $d(x, u) \neq d(x, v)$. Thus, $\phi(v) = v$.

Let $w = \phi(y)$. Since $y \in N(x)$ and $\phi(x) = x$, $w \in N(x)$. Since x is more distant from v than y , the unique $x-v$ path in the tree G must be $P = (x, y, \dots, v)$. Apply ϕ to every vertex in this path to get $P' = (\phi(x), \phi(y), \dots, \phi(v)) = (x, w, \dots, v)$. Since there exists a unique path between two vertices, so $P = P'$. Therefore, $\phi(y) = y$.

We have shown that if ϕ fixes only the endvertex of each edge in T that is more distant from v , then ϕ fixes both endvertices and hence ϕ fixes every edge in T . Since T is an edge-determining set, ϕ must be the identity automorphism. Hence, S is a vertex-determining set of G of size $\text{Det}'(G)$. Hence, $\text{Det}(G) \leq \text{Det}'(G)$. By [Corollary 20](#), $\text{Det}(G) = \text{Det}'(G)$. \square

Our last result is on hypercubes. The n -dimensional hypercube, Q_n , can be defined as the graph whose vertex set is the set of ordered n -bit strings of 0s and 1s with two vertices adjacent if they differ in exactly one bit. Hypercubes are highly symmetric graphs; not only are they vertex-transitive and edge-flip-invariant, they are also edge-transitive, arc-transitive and distance-transitive, meaning that given any four vertices u, v, x and y such that $d(u, v) = d(x, y)$, there exists an automorphism ϕ such that $\phi(u) = x$ and $\phi(v) = y$. See [Biggs 1993]. Another definition of hypercubes is based on the binary operation of Cartesian product of graphs.

Definition 33. The *Cartesian product* of graphs G and H is the graph $G \square H$ defined by $V(G \square H) = \{(u, v) : u \in G, v \in H\}$, with (u, v) and (x, y) adjacent if either $u = x$ and $\{v, y\} \in E(H)$ or $v = y$ and $\{u, x\} \in E(G)$.

A graph is prime with respect to the Cartesian product if it cannot be written as the Cartesian product of two nontrivial graphs. The prime factor decomposition of a graph is a representation of the graph as a Cartesian product of prime graphs.

We can define Q_n recursively by letting $Q_1 = K_2$ and $Q_n = Q_{n-1} \square K_2$ for $n \geq 2$. Thus, $Q_n = K_2 \square \dots \square K_2$ is a prime factor decomposition of the hypercube. A useful tool in finding the determining index of the hypercube is the characteristic matrix.

Definition 34. Let $S = (V_1, \dots, V_r)$ be an ordered set of m -tuples. The *characteristic matrix*, $M(S)$, is the $r \times m$ matrix with the ij -th entry being the j -th coordinate of V_i .

The characteristic matrix was used in [Boutin 2009] to prove the following propositions.

Proposition 35 [Boutin 2009, Lemma 1]. *Let G be a connected graph with prime factor decomposition $G = G_1 \square \dots \square G_n$. Let $S = (V_1, \dots, V_n) \subseteq V(G)$. Then S is a vertex-determining set if and only if each column of the characteristic matrix, $M(S)$, contains a vertex-determining set for the corresponding factor of G and no two columns of $M(S)$ are isomorphic images of each other.*

In this result, each vertex V_i in the ordered set S is an n -tuple $(v_{i1}, v_{i2}, \dots, v_{in})$, with $v_{ij} \in V(G_j)$. Two columns of the characteristic matrix, $[v_{1j}, \dots, v_{mj}]^T$ and $[v_{1k}, \dots, v_{mk}]^T$, are isomorphic if there exists a graph isomorphism $\psi : G_j \rightarrow G_k$ such that $\psi(v_{ij}) = v_{ik}$ for all $i \in \{1, \dots, n\}$.

Applied to $Q_n = K_2 \square \dots \square K_2$, the characteristic matrix $M(S)$ of a vertex set S will be a 0-1 matrix. Since any nonempty subset of vertices of K_2 is a vertex-determining set, any 0-1 column is guaranteed to contain a determining set for the corresponding K_2 factor of Q_n . Moreover, two 0-1 columns are isomorphic if and only if they have either all identical bits or all opposite bits.

Observation 36. Let X be a 0-1 matrix with s rows and t columns. There are 2^s different 0-1 vectors of length s , which can be partitioned into 2^{s-1} opposite

pairs. Thus, if $t > 2^{s-1}$, then X must have either two columns that are identical or two columns that are opposite.

This observation is the key to proving the result below.

Proposition 37 [Boutin 2009, Theorem 3]. *For $n \geq 1$, $\text{Det}(Q_n) = \lceil \log_2(n) \rceil + 1$.*

For $n = 1$, $Q_1 = K_2$, which does not have an determining index. For $n = 2$, $Q_2 = C_4$; by Theorem 14, $\text{Det}'(C_4) = \text{Det}(C_4) = 2$.

Theorem 38. *Let $n \geq 3$. Then*

$$\text{Det}'(Q_n) = \begin{cases} \lceil \log_2 n \rceil + 1 & \text{if } n - \lceil \log_2 n \rceil > 2^{\lceil \log_2 n \rceil - 1}, \\ \lceil \log_2 n \rceil & \text{otherwise.} \end{cases}$$

Proof. Assume T is an edge-determining set for Q_n with $|T| = \lceil \log_2(n) \rceil - 1$. Let S be the set of endvertices of the edges of T . By the proof of Theorem 19, S is a vertex-determining set. Let $M(S)$ be the corresponding characteristic matrix. By Proposition 35, the columns of $M(S)$ are nonisomorphic.

Note that the endvertices of each edge in T differ by one bit. Thus, there are at most $\lceil \log_2(n) \rceil - 1$ columns such that the two rows of $M(S)$ corresponding to the endvertices of a single edge of T differ in that column. Let X be the submatrix of $M(S)$ consisting of the other columns, namely, the columns corresponding to bits which are identical in both endvertices of every edge in T . Then X has at least $n - \lceil \log_2(n) \rceil + 1$ columns. Furthermore, X has at most $\lceil \log_2(n) \rceil - 1$ distinct rows, since the endvertices of an edge are identical except for one bit and we removed the columns in which the endvertices differ.

Let X' be the matrix obtained from X by removing any duplicate rows. Then X' has $s \leq \lceil \log_2 n \rceil - 1$ rows and $t \geq n - \lceil \log_2(n) \rceil + 1$ columns. We will need the following result, the proof of which is in the Appendix.

Lemma 39. *For all $n \geq 3$, we have $n - \lceil \log_2 n \rceil + 1 > 2^{\lceil \log_2 n \rceil - 2}$.*

Applying this inequality,

$$t \geq n - \lceil \log_2(n) \rceil + 1 > 2^{\lceil \log_2 n \rceil - 2} \geq 2^{s-1},$$

and so by Observation 36, X' must have some isomorphic columns. Hence, so does X , and so does $M(S)$. This is a contradiction. Thus, $\text{Det}'(Q_n) \geq \lceil \log_2(n) \rceil$.

By Corollary 20, $\text{Det}'(Q_n) \leq \text{Det}(Q_n) = \lceil \log_2 n \rceil + 1$. So the only two options for $\text{Det}'(Q_n)$ are $\lceil \log_2 n \rceil$ and $\lceil \log_2 n \rceil + 1$. We will show that the relative size of $n - \lceil \log_2 n \rceil$ and $2^{\lceil \log_2 n \rceil - 1}$ decides which option holds.

First assume $n - \lceil \log_2(n) \rceil > 2^{\lceil \log_2(n) \rceil - 1}$. We can repeat the argument given at the beginning of the proof to show that there can be no determining set of size $\lceil \log_2 n \rceil$. Using the same logic, we end up with a matrix X' with $s \leq \lceil \log_2 n \rceil$ rows and

$t \geq n - \lceil \log_2 n \rceil$ columns. By assumption,

$$t \geq n - \lceil \log_2(n) \rceil > 2^{\lceil \log_2 n \rceil - 1} \geq 2^{s-1},$$

so the columns of X' cannot be nonisomorphic by [Observation 36](#). Hence, in this case, $\text{Det}'(Q_n) = \lceil \log_2 n \rceil + 1$.

Now assume $n - \lceil \log_2 n \rceil \leq 2^{\lceil \log_2 n \rceil - 1}$. We will construct an edge-determining set T for Q_n of size $\lceil \log_2 n \rceil$. We begin by creating an $\lceil \log_2 n \rceil \times n$ matrix Y whose leftmost $\lceil \log_2 n \rceil$ columns are the standard basis vectors in $\mathbb{Z}_2^{\lceil \log_2 n \rceil}$ and whose remaining $n - \lceil \log_2 n \rceil$ columns are any set of nonisomorphic 0-1 columns of length $\lceil \log_2 n \rceil$. For example, when $n = 7$, one possible such matrix is

$$Y = \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 \end{array} \right].$$

We then create a corresponding $(2\lceil \log_2 n \rceil) \times n$ matrix X by creating a duplicate copy of each row, then switching only the 1 that is in the first $\lceil \log_2 n \rceil$ columns to 0. Continuing with our example, we get

$$X = \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ \hline 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ \hline 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{array} \right].$$

The first $\lceil \log_2 n \rceil$ columns of X have exactly one 1. The final $n - \lceil \log_2 n \rceil$ columns of X are still pairwise nonisomorphic, and since each has an even number of 0's and an even number of 1's, they are also nonisomorphic to any of the first $\lceil \log_2 n \rceil$ columns. Hence, by [Proposition 35](#), X is the characteristic matrix of a vertex-determining set of Q_n .

By construction, for each $i \in \{1, \dots, \lceil \log_2 n \rceil\}$, rows $2i - 1$ and $2i$ of X differ in exactly one column and therefore they correspond to a pair of adjacent vertices in Q_n . We let e_i be the edge between these vertices. We then let $T = \{e_1, e_2, \dots, e_{\lceil \log_2 n \rceil}\}$.

To show T is an edge-determining set of Q_n , suppose $\phi \in \text{Aut}(Q_n)$ fixes the edges in T . Then ϕ either fixes or switches the endvertices of each edge. To be consistent with the vertex notation in [Proposition 35](#), let V_i^0 and V_i^1 denote the two endvertices of edge e_i , with V_i^0 being the endvertex with 0 in the i -th bit. Let i and j be distinct elements of $\{1, \dots, \lceil \log_2 n \rceil\}$. (This is possible because $n \geq 3$, so $\lceil \log_2 n \rceil \geq 2$.) Suppose the last $n - \lceil \log_2 n \rceil$ columns of the i -th and j -th rows of Y differ in ℓ bits. Using the fact that the distance between two vertices of Q_n is the

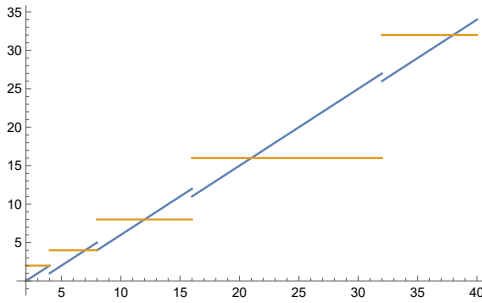


Figure 8. Comparing $x - \lceil \log_2 x \rceil$ and $2^{\lceil \log_2 x \rceil - 1}$.

number of bits in which the two corresponding bit strings differ, we get

$$d(V_i^0, V_j^0) = \ell, \quad d(V_i^0, V_j^1) = d(V_i^1, V_j^0) = \ell + 1, \quad d(V_i^1, V_j^1) = \ell + 2.$$

Since ϕ fixes the edges of T , we have $\phi(V_i^0) = V_i^0$ or V_i^1 and $\phi(V_j^0) = V_j^0$ or V_j^1 . Because automorphisms are distance-preserving, it must be the case that

$$\ell = d(V_i^0, V_j^0) = d(\phi(V_i^0), \phi(V_j^0)).$$

This is only possible if ϕ fixes the endvertices of both e_i and e_j . Thus, ϕ must fix the endvertices of every edge in T . As noted earlier, these vertices constitute a vertex-determining set of Q_n and so ϕ must be the identity. Hence, T is an edge-determining set. \square

An interesting implication of [Theorem 38](#) is that as n increases, $\text{Det}'(Q_n)$ fluctuates infinitely many times between the upper bound of $\text{Det}(Q_n)$ and the marginally smaller value of $\text{Det}(Q_n) - 1$. The graphs in [Figure 8](#) illustrate that $n - \lceil \log_2 n \rceil$ and $2^{\lceil \log_2 n \rceil - 1}$ keep alternating in relative size as n increases. An alternative formulation of this result may clarify why this happens.

Corollary 40. *Let $n \geq 3$ and let $k = \lceil \log_2 n \rceil$. Then*

$$\text{Det}'(Q_n) = \begin{cases} k & \text{if } 2^{k-1} < n \leq 2^{k-1} + k, \\ k + 1 & \text{if } 2^{k-1} + k < n \leq 2^k. \end{cases}$$

Proof. Since $k \leq 2^{k-1}$ for all $k \in \mathbb{N}$, both n and $2^{k-1} + k$ lie in the interval $(2^{k-1}, 2^k]$. Either they are equal, or one is to the left of the other. If $n \leq 2^{k-1} + k$, then $n - \lceil \log_2 n \rceil = n - k \leq 2^{k-1} = 2^{\lceil \log_2 n \rceil - 1}$, and so $\text{Det}'(Q_n) = \lceil \log_2 n \rceil$ by [Theorem 38](#). By similar reasoning, if $n > 2^{k-1} + k$, then $\text{Det}'(Q_n) = \lceil \log_2 n \rceil + 1$. \square

4. Open questions

In this paper, we have provided the determining index for several families of graphs, including paths, cycles, complete graphs, complete bipartite graphs, trees and

hypercubes. There are many more families of symmetric graphs for which this parameter could be computed, such as generalized Petersen graphs, circulant graphs, orthogonality graphs, Mycielskian graphs, Paley graphs and Praeger–Xu graphs, to name a few. In addition, there are some more conceptual open questions.

(1) We have shown that in some cases, the determining index of a graph is strictly smaller than the determining number. In these cases, the number of edges that are needed to fix in order to break all symmetries of the graph is less than the number of vertices that would be required, suggesting that fixing edges is more efficient. Are there necessary and/or sufficient conditions on G that guarantee $\text{Det}(G) = \text{Det}'(G)$?

(2) By the proof of [Theorem 38](#), the edge set

$$T = \{(1, 0, 0, 0), (0, 0, 0, 0)\}, \{(0, 1, 1, 0), (0, 0, 1, 0)\}$$

is a minimum edge-determining set for Q_4 . The set of endvertices of the edges of T is a vertex-determining set but not minimum by [Proposition 37](#). The characteristic matrices of all possible subsets with one end vertex removed are

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

Each of these characteristic matrices has two isomorphic columns, so they do not correspond to vertex-determining sets of Q_4 by [Proposition 35](#). Since $\text{Det}(Q_4) = 3$, we have found a minimum edge-determining set whose set of endvertices does not contain a minimum vertex-determining set. Do there exist conditions such that the set of endvertices of the edges in a minimum edge-determining set necessarily contains a minimum vertex-determining set?

(3) Are there expressions for $\text{Det}'(G + H)$ and $\text{Det}'(G \square H)$ in terms of $\text{Det}'(G)$ and $\text{Det}'(H)$?

Appendix: Proof of [Lemma 39](#)

Lemma 39. *For all $n \geq 3$, we have $2^{\lceil \log_2 n \rceil - 2} < \frac{n}{2} \leq n - \lceil \log_2 n \rceil + 1$.*

Proof. By definition of the ceiling function, $\lceil \log_2 n \rceil - 1 < \log_2 n \leq \lceil \log_2 n \rceil$ for any $n \in \mathbb{N}$. Then $2^{\lceil \log_2 n \rceil - 1} < 2^{\log_2 n} = n$, because exponential functions are increasing. Dividing by 2 gives $2^{\lceil \log_2 n \rceil - 2} < \frac{n}{2}$.

Next, let $f(x) = \log_2 x - \frac{x}{2}$ for $0 < x \in \mathbb{R}$. It is easily verified that the only zeroes of f occur at $x = 2$ and $x = 4$. Elementary calculus can be used to show that the graph $y = f(x)$ is concave down, with the absolute maximum value at $x = \frac{2}{\ln 2} \approx 2.885$. This means that, for all $x \geq 4$, $f(x) \leq 0$, which implies $\log_2 x \leq \frac{x}{2}$. Hence, for all $n \geq 4$

$$\lceil \log_2 n \rceil - 1 < \log_2 n \leq \frac{n}{2},$$

which can be algebraically rearranged to

$$\frac{n}{2} \leq n - \lceil \log_2 n \rceil + 1.$$

A direct calculation shows that this inequality also holds for $n = 3$. □

References

- [Albertson and Boutin 2007] M. O. Albertson and D. L. Boutin, “Using determining sets to distinguish Kneser graphs”, *Electron. J. Combin.* **14**:1 (2007), art. id. R20. [MR](#) [Zbl](#)
- [Alikhani and Soltani 2020a] S. Alikhani and S. Soltani, “The chromatic distinguishing index of certain graphs”, *AKCE Int. J. Graphs Comb.* **17**:1 (2020), 131–138. [MR](#) [Zbl](#)
- [Alikhani and Soltani 2020b] S. Alikhani and S. Soltani, “The distinguishing number and the distinguishing index of line and graphoidal graph(s)”, *AKCE Int. J. Graphs Comb.* **17**:1 (2020), 1–6. [MR](#) [Zbl](#)
- [Biggs 1993] N. Biggs, *Algebraic graph theory*, 2nd ed., Cambridge University Press, 1993. [MR](#) [Zbl](#)
- [Boutin 2009] D. L. Boutin, “The determining number of a Cartesian product”, *J. Graph Theory* **61**:2 (2009), 77–87. [MR](#) [Zbl](#)
- [Brooks et al. 2021] J. Brooks, A. Carbonero, J. Vargas, R. Flórez, B. Rooney, and D. Narayan, “Removing symmetry in circulant graphs and point-block incidence graphs”, *Mathematics* **9**:2 (2021), art. id. 116.
- [Chartrand and Zhang 2012] G. Chartrand and P. Zhang, *A first course in graph theory*, Dover, New York, 2012.
- [Erwin and Harary 2006] D. Erwin and F. Harary, “Destroying automorphisms by fixing nodes”, *Discrete Math.* **306**:24 (2006), 3244–3252. [MR](#) [Zbl](#)
- [Gibbons and Laison 2009] C. R. Gibbons and J. D. Laison, “Fixing numbers of graphs and groups”, *Electron. J. Combin.* **16**:1 (2009), art. id. R39. [MR](#) [Zbl](#)
- [Holt 1981] D. F. Holt, “A graph which is edge transitive but not arc transitive”, *J. Graph Theory* **5**:2 (1981), 201–204. [MR](#) [Zbl](#)
- [Imrich et al. 2020] W. Imrich, R. Kalinowski, M. Pilśniak, and M. Woźniak, “The distinguishing index of connected graphs without pendant edges”, *Ars Math. Contemp.* **18**:1 (2020), 117–126. [MR](#) [Zbl](#)
- [Kalinowski and Pilśniak 2015] R. Kalinowski and M. Pilśniak, “Distinguishing graphs by edge-colourings”, *European J. Combin.* **45** (2015), 124–131. [MR](#) [Zbl](#)
- [Lehner et al. 2020] F. Lehner, M. Pilśniak, and M. Stawiski, “A bound for the distinguishing index of regular graphs”, *European J. Combin.* **89** (2020), art. id. 103145. [MR](#) [Zbl](#)
- [Sabidussi 1961] G. Sabidussi, “Graph derivatives”, *Math. Z.* **76** (1961), 385–401. [MR](#) [Zbl](#)

Received: 2022-07-29

Revised: 2022-12-30

Accepted: 2022-12-30

scockbur@hamilton.edu

*Mathematics and Statistics Department, Hamilton College,
Clinton, NY, United States*

sean_mcavoy@berkeley.edu

*Department of Statistics, University of California,
Berkeley, CA, United States*

The adjacency spectra of some families of minimally connected prime graphs

Chris Florez, Jonathan Higgins, Kyle Huang,
Thomas Michael Keller and Dawei Shen

(Communicated by Vadim Ponomarenko)

In finite group theory, studying the prime graph of a group has been an important topic for almost the past half-century. Recently, prime graphs of solvable groups have been characterized in graph-theoretical terms only; this now allows the study of these graphs without any knowledge of the group-theoretical background. We approach prime graphs from a linear-algebraic angle and focus on the class of minimally connected prime graphs introduced in earlier work on the subject. As our main results, we prove new properties about the adjacency matrices of some special families of these graphs, focusing on their characteristic polynomials and spectra.

1. Introduction

This paper deals with prime graphs of finite solvable groups. The prime graph of a finite group is the graph whose vertices are the prime numbers dividing the order of the group, and two vertices are linked by an edge if and only if their product divides the order of some element of the group. Prime graphs were introduced by Gruenberg and Kegel in the 1970s and have been an object of continuous study since. They were one of the first graphs assigned to groups. This idea of representing group-theoretical data via graphs and describing them via graph-theoretical notions proved so successful that today there is a myriad of graphs (e.g., character degree graphs, conjugacy class size graphs, etc.) and a whole industry of exploring them. For this reason, today, prime graphs are often referred to as Gruenberg–Kegel graphs.

While a focus in the study of prime graphs has been on simple groups for a long time, the main result of [Gruber et al. 2015], somewhat surprisingly, is a purely graph-theoretical characterization of prime graphs of solvable groups: a (simple) graph is the prime graph of a finite solvable group if and only if its complement is triangle-free and 3-colorable. This made it possible to study simple groups whose prime graph is that of a finite solvable group; see [Gorshkov and Maslova 2018].

MSC2020: primary 05C25, 15A18; secondary 20D10.

Keywords: prime graphs, adjacency matrix, spectral graph theory.

Moreover, this characterization allowed the authors of [Gruber et al. 2015], for solvable groups, to introduce and study the idea of minimal prime graphs: connected graphs whose complement is triangle-free and 3-colorable, but removing an edge means that the complement has a triangle or is no longer 3-colorable. The groups whose prime graphs are minimal are groups which are “saturated in Frobenius actions” and have a restricted, but highly nontrivial structure, as discussed in detail in [Gruber et al. 2015]. In [Florez et al. 2020] the authors thoroughly explore some graph-theoretical properties of minimal prime graphs. Also, an alternative notion of minimal prime graphs — minimally connected prime graphs — is introduced and turns out to be closely related to minimal prime graphs.

In this paper, we study minimal prime graphs from a completely different angle, namely from a linear algebra point of view. This is one of the first studies of graphs related to groups with a linear algebra focus (the only other group-related graphs for which this has been done and that we are aware of being Cayley graphs). We will find a rich structure for some of the basic minimal and minimally connected prime graphs. We will study the determinants and the spectra, or sets of eigenvalues along with their multiplicities, of their adjacency matrices.

We now explain the specific content of the paper in some more technical detail.

A minimal prime graph (MPG) of a solvable group Γ_G is defined as a connected graph of order $n > 1$ such that $\Gamma_G \setminus \{pq\}$ is not the prime graph of a solvable group for any $pq \in E(\Gamma_G)$; see [Gruber et al. 2015]. Equivalently, an MPG is a connected graph of order $n > 1$ whose complement is triangle-free and three-colorable, but the addition of an edge to its complement induces a triangle or renders it no longer three-colorable. A minimally connected prime graph (MCPG) of a solvable group is defined similarly, but it includes that the removal of any edge may result in a disconnected graph without inducing a triangle or changing the colorability of the complement; see [Florez et al. 2020]. If \mathcal{G} is the class of minimal prime graphs and $\widehat{\mathcal{G}}$ is the class of minimally connected prime graphs, it has been shown that $\mathcal{G} \subsetneq \widehat{\mathcal{G}}$ and if $\Gamma \in \widehat{\mathcal{G}} \setminus \mathcal{G}$, then Γ is the graph of two vertex-disjoint complete graphs joined by one edge. These graphs are called complete bridge graphs, and we use the notation $B_{m,n}$ to denote the complete bridge graph of complete graphs K_m and K_n joined by one edge. For convenience, we say $m \geq n$, and in order to maintain the conditions in the definition of MCPGs, we must enforce the conditions $m \geq n > 1$ or, if $n = 1$, then $m \in \{1, 2\}$; this result is found in [Florez et al. 2020]. Another kind of MCPG (which is contained in \mathcal{G}) that we will consider in this paper are reseminant graphs. Reseminant graphs are defined as graphs generated by repeated vertex duplication on C_5 . Vertex duplication is a method of generating new graphs from old graphs, where if we have a graph G , we can produce the graph G' by introducing a new vertex v to G and an edge vv' , where $v' \in V(G)$, along with the edges vx if and only if $xv' \in E(G)$. In this paper, we are not concerned with the

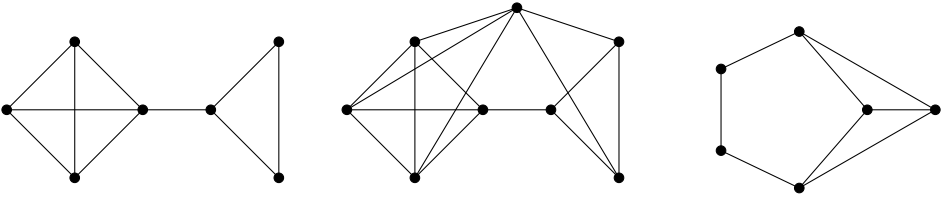


Figure 1. The complete bridge graph $B_{4,3}$ (left), the suspension graph $S(4, 3)$ (middle), and the unique reseminant graph on six vertices (right).

group-theoretic problems, so when we say “minimally connected prime graphs,” we are referring to minimally connected prime graphs of solvable groups.

In Section 2, we will study the adjacency matrices of complete bridge graphs that belong to the family of MCPGs. Specifically, this will consist of determining the characteristic polynomial for these graphs and a closer study of the relationships between the eigenvalues of $B_{m,m-1}$ for $m > 2$. In Section 3, we consider the same types of questions for reseminant graphs. Namely, we look briefly at the characteristic polynomial for the arbitrary reseminant graph and proceed to study particular families of reseminant graphs in more detail. We will focus our attention on the reseminant graphs where two nonadjacent vertices of the 5-cycle are duplicated and the case where only one vertex is duplicated. The graphs in the former case are isomorphic to the suspension graphs, which were studied in [Florez et al. 2020]. The suspension graph $S(m, n)$ is defined as the graph generated by adding a vertex to $B_{m,n}$ and connecting it with all of the nonbridge vertices. Also, for reseminant graphs in the latter category, the rigidity of this structure enables us to carefully study how the eigenvalues of these graphs are related to each other. Examples of a complete bridge graph, a suspension graph, and a reseminant graph can be found in Figure 1. We conclude the paper with a brief outlook on potential areas of future research.

Before proceeding, it will be helpful to define some additional notation that will be used. If we have a graph Γ , we will let $A(\Gamma)$ be the adjacency matrix of Γ . For a square matrix M , we will use $\phi(M, x)$ to denote its characteristic polynomial in x . We will also let $\widehat{\mathcal{G}}$ denote the set of minimally connected prime graphs, \mathcal{G} the minimal prime graphs, \mathcal{R} the reseminant graphs, and $\widetilde{\mathcal{R}}$ the reseminant graphs generated by repeatedly duplicating the same vertex.

2. Complete bridge graphs

In this section, we will determine the characteristic polynomial for complete bridge graphs as MCPGs, along with the spectra of a particular family—namely, $B_{m,m-1}$. Recall that, in order to be an MCPG, $B_{m,n}$ must be such that $m \geq n > 1$ or, if $n = 1$, then $m \in \{1, 2\}$. Also, we will restrict our attention to the complete bridge

graphs that are MCPGs with $m + n \geq 4$ (or $m > 2$ for $B_{m,m-1}$ graphs) because all these graphs share P_4 , the path graph with four vertices and three edges, as their compression graphs, while $B_{2,1}$ and $B_{1,1}$, the only remaining complete bridge graphs that are MCPGs, do not. (Compression graphs are the quotient graphs by structural equivalence classes; see [Nguyen et al. 2019] for more on structural equivalence.) The spectra for $B_{2,1}$ and $B_{1,1}$ can easily be determined with a CAS and, since they do not fit in with the general approach we will be using, we will not be concerned with them in this paper.

The following lemma will be essential for proving some of the remaining theorems in this paper. The reader unfamiliar with equitable partitions is advised to consult [Godsil and Royle 2001] (or a similar book) before proceeding. Regarding notation, Γ/S below denotes the quotient graph of Γ by S .

Lemma 2.1. *If S is an equitable partition of a graph Γ , then the characteristic polynomial of $A(\Gamma/S)$ divides the characteristic polynomial of $A(\Gamma)$.*

Now, we will determine the characteristic polynomial for any complete bridge graph that is an MCPG.

Theorem 2.2. *Let $B_{m,n}$ be an MCPG with $m + n \geq 4$ (thus, $B_{m,n}$ has P_4 as its compression graph). The characteristic polynomial of $B_{m,n}$ is*

$$\phi(A(B_{m,n}), x) = Q(x)(x + 1)^{m+n-4}, \quad (1)$$

where

$$Q(x) = x^4 + (4 - (m + n))x^3 + (mn - 3(m + n) + 5)x^2 + (2mn - 2(m + n))x + (m + n - 3). \quad (2)$$

Proof. We first show that $Q(x)$ divides the characteristic polynomial. To do so, we will first identify all structurally equivalent vertices (i.e., vertices that share the same neighbors, excluding only each other). This partitions the graph into four subgraphs, being the $m - 1$ vertices in the K_m subgraph and the $n - 1$ vertices of the K_n subgraph that are not the bridge vertices, with the two remaining equivalence classes being the two bridge vertices. Let this partition be denoted by S , which is an equitable partition. We find that the quotient matrix $A(B_{m,n}/S)$ is given by

$$A(B_{m,n}/S) = \begin{pmatrix} m-2 & 1 & 0 & 0 \\ m-1 & 0 & 1 & 0 \\ 0 & 1 & 0 & n-1 \\ 0 & 0 & 1 & n-2 \end{pmatrix}.$$

By Laplace expansion, it is easily calculated that

$$\begin{aligned} \det(Ix - A(B_{m,n}/S)) &= \phi(A(B_{m,n}/S), x) \\ &= (x^4 + (4 - (m + n))x^3 + (mn - 3(m + n) + 5)x^2 \\ &\quad + (2mn - 2(m + n))x + (m + n - 3)). \end{aligned}$$

By [Lemma 2.1](#), this must divide the characteristic polynomial of $A(B_{m,n})$.

Now, it is not too difficult to see that the remaining $m + n - 4$ eigenvalues must all be -1 . This can be seen by the fact that $B_{m,n}$ may be obtained by taking a P_4 graph (this is its compression graph) and duplicating one of its “end” vertices $m - 2$ times and the other $n - 2$ times. Duplicating a vertex always adds a -1 to the spectrum of a graph (while changing the eigenvalues that are not -1) because, when the adjacency matrix A is passed to $I + A$ (or $-I - A$), the rows (and columns) corresponding to the duplicated vertex and the new vertex are identical. Thus, it follows that $B_{m,n}$ will have -1 as an eigenvalue of multiplicity $m + n - 4$. In other words, $(x + 1)^{m+n-4}$ divides $\phi(A(B_{m,n}), x)$. The theorem follows. \square

Corollary 2.3. *The determinant of $A(B_{m,n})$ is given by*

$$\det(A(B_{m,n})) = (-1)^{m+n-1}(3 - (m + n)). \tag{3}$$

Proof. This follows easily by reading off the linear term of the characteristic polynomial of $B_{m,n}$ in [\(1\)](#). \square

Although it is nice to have the characteristic polynomial for any complete bridge graph, it is difficult to know how the roots of the polynomial $Q(x)$ from [Theorem 2.2](#) are related to each other and -1 . However, when we consider the special class $B_{m,m-1}$, we find a case where we can determine how the eigenvalues are related with relative ease. This leads us to the next result.

Corollary 2.4. *If $m > 2$, then $B_{m,m-1}$ has $m - 2$ as an eigenvalue; further, the characteristic polynomial of $B_{m,m-1}$ is*

$$\phi(A(B_{m,m-1}), x) = (x^3 + (3 - m)x^2 + (2 - 2m)x - 2)(x - (m - 2))(x + 1)^{2m-5}. \tag{4}$$

Proof. First, we let $n = m - 1$, which gives us

$$\begin{aligned} &\phi(A(B_{m,m-1}), x) \\ &= (x^4 + (5 - 2m)x^3 + (m^2 - 7m + 8)x^2 + (2m^2 - 6m + 2)x + (2m - 4))(x + 1)^{2m-5} \end{aligned}$$

by [Theorem 2.2](#). The conclusion follows by observing that

$$\begin{aligned} &x^4 + (5 - 2m)x^3 + (m^2 - 7m + 8)x^2 + (2m^2 - 6m + 2)x + (2m - 4) \\ &= (x^3 + (3 - m)x^2 + (2 - 2m)x - 2)(x - (m - 2)). \quad \square \end{aligned}$$

For the remainder of this section, we will let $\theta_1 \geq \theta_2 \geq \theta_3$ be the roots of $x^3 + (3 - m)x^2 + (2 - 2m)x - 2$.

Lemma 2.5. *If $m > 2$, then θ_1 is the unique largest eigenvalue of $B_{m,m-1}$, and it satisfies $m - 1 \leq \theta_1 \leq m$. Additionally, θ_2 and θ_3 satisfy the following inequalities:*

- (i) $-3 \leq \theta_2 + \theta_3 \leq -2$.
- (ii) $2/m \leq \theta_2\theta_3 \leq 2/(m - 1)$.

Proof. First, we note that the largest eigenvalue of any graph is bounded below by the average degree of every induced subgraph and bounded above by the maximum vertex degree (this is a standard result in spectral graph theory and can be found in [Brouwer and Haemers 2012], for example). The maximum degree of any vertex in $B_{m,m-1}$ is m and $m-1$ is the largest average degree of any induced subgraph (namely, the induced subgraph K_m). This tells us that the largest eigenvalue is bound between $m-1$ and m ; since $-1 < m-2 < m-1$, we deduce that the largest eigenvalue is θ_1 and $m-1 \leq \theta_1 \leq m$.

Now, we will utilize the fact that the sum of the eigenvalues of $A(B_{m,n})$ is the trace of the matrix (thus, 0) and the product of the eigenvalues is the determinant. Thus, we deduce

$$\sum_{i=1}^3 \theta_i = m - 3, \quad \prod_{i=1}^3 \theta_i = 2. \quad (5)$$

(This can also be determined from the coefficients of the polynomial.) The inequalities (i) and (ii) for θ_2 and θ_3 follow from these equations and the fact that $m-1 \leq \theta_1 \leq m$.

Finally, we want to show that θ_1 is the unique largest eigenvalue. By (i) of this lemma, we know that θ_3 must be negative, and by (ii), we see that θ_2 must also be negative. Thus, θ_1 and $m-2$ are the only positive eigenvalues of $B_{m,m-1}$, so the uniqueness of θ_1 as the largest eigenvalue follows since $\theta_1 > m-2$. \square

Now, we are ready to determine how the eigenvalues of $B_{m,m-1}$ are related, thereby resulting in its spectrum.

Theorem 2.6. *If $m > 2$, then $\theta_1 > m-2 > 0 > \theta_2 > -1 > \theta_3$, and*

$$\text{Spec}(B_{m,m-1}) = \left(\begin{array}{cccc} \theta_1 & m-2 & \theta_2 & -1 & \theta_3 \\ 1 & 1 & 1 & 2m-5 & 1 \end{array} \right). \quad (6)$$

Proof. By Lemma 2.5, we know $\theta_1 > m-2 > 0$ and θ_2 and θ_3 are both negative, so all that remains is to show how θ_2 and θ_3 relate to each other and -1 . Assume $\theta_2 > \theta_3 > -1$. Then $\theta_2 + \theta_3 > -2$, which contradicts Lemma 2.5(i). If $-1 > \theta_2 > \theta_3$, then $\theta_2\theta_3 > 1$, which contradicts Lemma 2.5(ii). If we assume $\theta_2 = \theta_3 = -1$, then the only $m > 2$ that satisfies Lemma 2.5 is $m = 3$, and it is easily seen that -1 is not a root of $x^3 - 4x - 2$. If $\theta_2 > \theta_3 = -1$, then Lemma 2.5(i) is not satisfied, and if $\theta_2 = -1 > \theta_3$, then Lemma 2.5(ii) is not satisfied. Thus, by this casework, we know $0 > \theta_2 > -1 > \theta_3$, which completes the proof. \square

3. Reseminant graphs

Recall that a reseminant graph is defined as a graph that can be generated by duplicating the vertices of a 5-cycle. Thus, we see that a reseminant graph may be represented by a 5-tuple, $(n_1, n_2, n_3, n_4, n_5)$ denoting how many times each vertex

of the 5-cycle is duplicated. Thus, $(0, 0, 0, 0, 0)$ corresponds to the 5-cycle itself. If only one vertex of the 5-cycle is duplicated, we will let the first number in the 5-tuple for this graph denote the number of times this vertex is duplicated. If this vertex is duplicated n times, we will use $R_{(n,0,0,0,0)}$, or \tilde{R}_n , to denote this graph. If only two vertices are duplicated, we will denote this graph by $R_{(n,m,0,0,0)}$ or $R_{(n,0,m,0,0)}$, depending on whether the two duplicated vertices are adjacent or not, where one of the vertices is duplicated n times and the other is duplicated m times, with $n \geq m$. For a general reseminant graph, we define its tuple $(n_1, n_2, n_3, n_4, n_5)$ such that the i -th vertex of the 5-cycle (i.e., the vertex represented by the n_i) is adjacent to the $(i-1)$ -th and $(i+1)$ -th vertices mod 5. We will denote this graph as $R_{(n_1,n_2,n_3,n_4,n_5)}$. Also, note that we have all isomorphisms of the form

$$R_{(n_1,n_2,n_3,n_4,n_5)} \cong \sigma \cdot R_{(n_1,n_2,n_3,n_4,n_5)} = R_{(n_{\sigma^{-1}(1)}, n_{\sigma^{-1}(2)}, n_{\sigma^{-1}(3)}, n_{\sigma^{-1}(4)}, n_{\sigma^{-1}(5)})}$$

where σ is in D_{10} , the dihedral group of order 10, consisting of the symmetries of the regular pentagon, which includes five reflections and five rotations (with one of the rotations being the trivial one). As an example, $R_{(n,0,m,0,0)} \cong R_{(n,0,0,m,0)}$, but, as stated previously, we will stick with the first notation as convention.

First, we will consider the characteristic polynomial of the arbitrary reseminant graph. Although it is quite messy, this general form enables us to study particular families of reseminant graphs more easily.

Theorem 3.1. *The characteristic polynomial of an arbitrary reseminant graph R , representing $R_{(n_1,n_2,n_3,n_4,n_5)}$, is given by $\phi(A(R), x) = P(x)(x + 1)^{n_1+n_2+n_3+n_4+n_5}$, where $P(x)$ is the determinant of*

$$\begin{pmatrix} x-n_1 & -(n_2+1) & 0 & 0 & -(n_5+1) \\ -(n_1+1) & x-n_2 & -(n_3+1) & 0 & 0 \\ 0 & -(n_2+1) & x-n_3 & -(n_4+1) & 0 \\ 0 & 0 & -(n_3+1) & x-n_4 & -(n_5+1) \\ -(n_1+1) & 0 & 0 & -(n_4+1) & x-n_5 \end{pmatrix}. \tag{7}$$

Proof. This result follows quite simply by the same reasoning we saw in the proof of [Theorem 2.2](#). We get an equitable partition of R by defining each element in the partition as the i -th vertex of the 5-cycle along with the n_i vertices obtained by duplicating it. Thus, the partition contains five elements, with sizes $n_i + 1$ for $i \in \{1, 2, 3, 4, 5\}$. If we call this partition S , we have the quotient matrix

$$A(R/S) = \begin{pmatrix} n_1 & n_2+1 & 0 & 0 & n_5+1 \\ n_1+1 & n_2 & n_3+1 & 0 & 0 \\ 0 & n_2+1 & n_3 & n_4+1 & 0 \\ 0 & 0 & n_3+1 & n_4 & n_5+1 \\ n_1+1 & 0 & 0 & n_4+1 & n_5 \end{pmatrix},$$

and by Lemma 2.1, $P(x) = \det(Ix - A(R/S)) = \phi(A(R/S), x)$ divides $\phi(A(R), x)$. The $(x + 1)^{n_1+n_2+n_3+n_4+n_5}$ term in the characteristic polynomial of $\phi(A(R), x)$ follows by reasoning similar to that used at the end of the proof of Theorem 2.2 (i.e., duplicating vertices adds a -1 to the spectrum). \square

Next, we will look briefly at suspension graphs.

Theorem 3.2. *If $B_{m,n}$ is an MCPG with $m + n > 4$, then $S(m, n) \cong R_{(m-2,0,n-2,0,0)}$.*

Proof. By hypothesis, both m and n are greater than or equal to 2. First, note that $S(m, n)$ contains an induced 5-cycle. If we choose a nonbridge vertex from the K_m and K_n subgraphs of the $B_{m,n}$ used to generate $S(m, n)$, then these, along with the bridge vertices in the $B_{m,n}$ subgraph and the vertex that was added to $B_{m,n}$ to form $S(m, n)$, yield one such example of an induced 5-cycle. The remaining $m - 2$ and $n - 2$ vertices in the K_m and K_n are in the same equivalence class (they share the same neighbors) as their respective nonbridge vertices in the induced 5-cycle. Thus, it is trivial to see that $S(m, n)$ is a reseminant graph generated by duplicating one vertex of a 5-cycle $m - 2$ times and a nonadjacent vertex of the 5-cycle $n - 2$ times. The theorem follows. \square

Given this isomorphism, we will hereby refer to reseminant graphs of type $R_{(a,0,b,0,0)}$ by their suspension graph forms.

In the following theorem, we will give the characteristic polynomial for suspension graphs (under the assumption that m and n are greater than or equal to 2). The proof is omitted, as it follows immediately from Theorem 3.1.

Theorem 3.3. *The characteristic polynomial for $S(m, n)$ is*

$$\begin{aligned} \phi(A(S(m, n)), x) &= (x^5 - (m + n - 4)x^4 + (mn - 4(m + n) + 7)x^3 + (4mn - 5(m + n) + 4)x^2 \\ &\quad + (2mn - 3)x + 4mn - 5(m + n) + 6)(x + 1)^{m+n-4}. \end{aligned} \tag{8}$$

The following corollary follows trivially from Theorem 3.3 in the same manner that Corollary 2.3 followed from Theorem 2.2.

Corollary 3.4. *The determinant of $A(S(m, n))$ is given by*

$$\det A(S(m, n)) = (-1)^{m+n} (4mn - 5(m + n) + 6). \tag{9}$$

Next, we will talk about the graphs in $\tilde{\mathcal{R}}$, i.e., the family of reseminant graphs generated by duplicating only one vertex of the 5-cycle. By Theorem 3.2, we have the following result. Figure 2 shows an example of an explicit isomorphism defined by $a_i \mapsto a'_i$.

Corollary 3.5. $\tilde{R}_n \cong S(n + 2, 2)$.

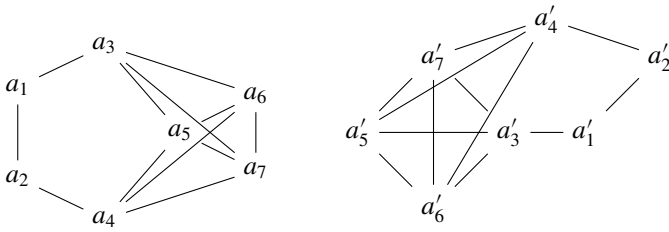


Figure 2. One of the isomorphisms between $\tilde{\mathcal{R}}_2$ and $S(4, 2)$.

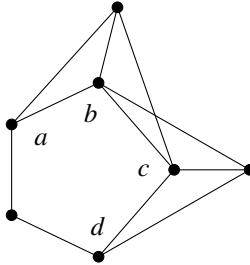


Figure 3. The graph above is a reseinant graph on seven vertices, where two distinct vertices are duplicated. Using the notation in the proof of [Theorem 3.6](#), we can let $a = v_1^*$, $b = v^* = v'_1$, $c = v' = v_2^*$, and $d = v'_2$.

Before we shift our focus to the spectral properties of $\tilde{\mathcal{R}}$, we will briefly look at these graphs from a purely graph-theoretic standpoint. First, note that K_n^- is often used to denote a complete graph with an edge removed; we shall adopt that notation. Furthermore, recall that, for a graph Γ and a subset $A \subset V(\Gamma)$, the induced subgraph $\Gamma[A]$ is the subgraph of Γ with vertex set A , with an edge between two vertices in $\Gamma[A]$ if and only if there is an edge between them in Γ .

Definition. An induced subgraph $\Gamma[\pi]$ of Γ is a *maximal K_n^- induced subgraph* of Γ if

- $\pi \subset V(\Gamma)$ with $|\pi| = n$,
- $\Gamma[\pi]$ is isomorphic to K_n^- ,
- $\Gamma[\pi \cup \{v\}]$ is not isomorphic to K_{n+1}^- for every $v \in V(\Gamma) \setminus \pi$.

Now, we are prepared to start studying $\tilde{\mathcal{R}}$, the family of reseinant graphs generated by repeatedly duplicating the same vertex of the induced 5-cycle. Below, we will use $K_{i \geq 4}^-$ to denote a maximal K_i^- induced subgraph with $i \geq 4$.

Theorem 3.6. *If $\Gamma \in \mathcal{R}$, then $\Gamma \in \tilde{\mathcal{R}}$ if and only if Γ has no more than one maximal $K_{i \geq 4}^-$ induced subgraph.*

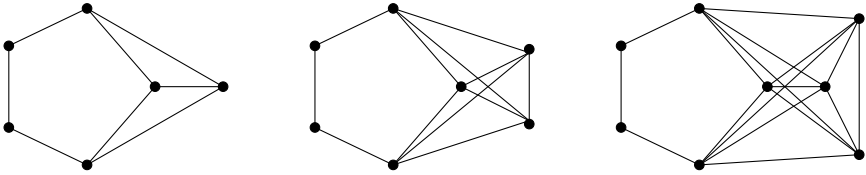


Figure 4. The graphs \tilde{R}_1 , \tilde{R}_2 , and \tilde{R}_3 , respectively. Note they all contain two vertices of degree 2 and one K_{n+3}^- induced subgraph.

Proof. Let $\Gamma \in \mathcal{R}$. The forward direction follows from the definition of $\tilde{\mathcal{R}}$. Next, assume that Γ has no more than one maximal $K_{i \geq 4}^-$ induced subgraph, but Γ is not C_5 or some other reseinant graph generated by repeatedly duplicating the same vertex in the induced subgraph isomorphic to C_5 . Thus, we know that two unique vertices v^* and v' on the 5-cycle were duplicated in the construction of Γ . Let v_1^* and v_2^* be the vertices in the induced 5-cycle of Γ that are adjacent to v^* , and similarly define v'_1 and v'_2 . Note that not all of these six vertices will be distinct because they are all on the induced 5-cycle and that v^* and v' might be adjacent. Figure 3 provides an example of how these vertex labels could work. If we let V^* be the set of vertices in Γ that are the result of duplicating v^* , and if we define V' similarly, then $\Gamma[V^* \cup \{v^*, v_1^*, v_2^*\}]$ and $\Gamma[V' \cup \{v', v'_1, v'_2\}]$ are $K_{|V^*|+3}^-$ and $K_{|V'|+3}^-$ induced subgraphs, respectively. It is also not difficult to verify that these subgraphs are maximal. Take $\Gamma[V^* \cup \{v^*, v_1^*, v_2^*\}]$, for example. Any vertex $v \in V(\Gamma) \setminus (V^* \cup \{v^*, v_1^*, v_2^*\})$ is either another vertex of the 5-cycle or a result of duplicating a vertex in the 5-cycle other than v^* . In the first case, v is not connected with an vertex in $V^* \cup \{v^*, v_1^*, v_2^*\}$, so clearly $\Gamma[V^* \cup \{v^*, v_1^*, v_2^*\} \cup \{v\}]$ is not a $K_{|V^*|+4}^-$ graph. We find the same thing in the second case, since v would not be connected with any vertex in V^* . Thus, $\Gamma[V^* \cup \{v^*, v_1^*, v_2^*\}]$ is a maximal $K_{i \geq 4}^-$ graph, as is $\Gamma[V' \cup \{v', v'_1, v'_2\}]$ by analogous reasoning. By contradiction, the desired result follows. \square

Now, we have determined an important property of $\tilde{\mathcal{R}}$ that does not apply to any other reseinant graphs. Also, we can consider the following corollary from Theorem 3.6, which is helpful for visualizing the reseinant graphs to which it applies. It will be stated without proof, as its truth should be obvious.

Corollary 3.7. *If $\Gamma \in \mathcal{R}$, then $\Gamma \in \tilde{\mathcal{R}}$ if and only if Γ has at least two adjacent vertices of degree 2.*

Observe that the properties in Theorem 3.6 and Corollary 3.7 are easily seen in small examples of graphs in \mathcal{R} , as shown in Figure 4. Now, we proceed to study the spectral properties of $\tilde{\mathcal{R}}$ graphs. As usual, φ will denote the golden ratio $(1 + \sqrt{5})/2$.

Theorem 3.8. *For all nonnegative integers n , $-\varphi$ and φ^{-1} are eigenvalues of \tilde{R}_n ; further, the characteristic polynomial of \tilde{R}_n is*

$$\phi(A(\tilde{R}_n), x) = (x^3 - (n+1)x^2 - (n+3)x + (3n+2))(x+1)^n(x^2+x-1). \quad (10)$$

Proof. The proof of this theorem will look quite similar to the proof of [Theorem 2.2](#). Partitioning \tilde{R}_n into the four vertices of the 5-cycle and the remaining vertex duplicated n times yields an equitable partition, which we will denote by S . The quotient matrix $A(\tilde{R}_n/S)$ is given by

$$A(\tilde{R}_n/S) = \begin{pmatrix} n & 1 & 0 & 0 & 1 \\ n+1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ n+1 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

By Laplace expansion, one finds

$$\begin{aligned} \det(Ix - A(\tilde{R}_n/S)) &= \phi(A(\tilde{R}_n/S), x) \\ &= x^5 - nx^4 - (2n + 5)x^3 - 3nx^2 + (4n + 5)x - 3n - 2. \end{aligned}$$

It is not difficult to verify that

$$\begin{aligned} x^5 - nx^4 - (2n + 5)x^3 - 3nx^2 + (4n + 5)x - 3n - 2 \\ = (x^3 - (n + 1)x^2 - (n + 3)x + (3n + 2))(x^2 + x - 1). \end{aligned}$$

By [Lemma 2.1](#), this must divide $\phi(A(\tilde{R}_n), x)$. Since $x^2 + x - 1$ has $-\varphi$ and φ^{-1} as roots, it follows that all graphs in $\tilde{\mathcal{R}}$ share these as eigenvalues. The fact that the remaining n eigenvalues are -1 follows by reasoning similar to that used at the end of the proof of [Theorem 2.2](#). \square

Now, we will let $\theta_1 \geq \theta_2 \geq \theta_3$ be the roots of $x^3 - (n + 1)x^2 - (n + 3)x + (3n + 2)$. The following lemma is analogous to [Lemma 2.5](#) from the previous section. In its proof, we will use the idea of the discriminant. Recall that the discriminant of a polynomial, which we denote with Δ , is a number determined by the coefficients of the polynomial (or, in our case, a new polynomial in n , since the coefficients are dependent upon n) used to determine certain properties about the roots of the original polynomial. There are well-known formulas for the discriminants of low-degree polynomials, such as $\Delta = b^2 - 4ac$ for the quadratic polynomial $ax^2 + bx + c$, but the formula grows increasingly more complicated as the degree gets larger.

Lemma 3.9. θ_1 is the unique largest eigenvalue of \tilde{R}_n and it satisfies

$$\frac{(n + 1)(n + 4)}{n + 3} \leq \theta_1 \leq n + 2.$$

Additionally, θ_2 and θ_3 satisfy the following inequalities

- (i) $-1 \leq \theta_2 + \theta_3 \leq -(n + 1)/(n + 3)$.
- (ii) $-(2 + 3n)(n + 3)/(n + 1)(n + 4) \leq \theta_2\theta_3 \leq -(2 + 3n)/(n + 2)$.

Proof. The inequalities can be determined via the same reasoning as was used in the proof of [Lemma 2.5](#). The maximum average degree of an induced subgraph is found

by looking at the (maximal) K_{n+3}^- induced subgraph. In particular, we find that the average degree of this subgraph is $(n+1)(n+4)/(n+3)$ since there are $n+3$ vertices and the sum of the degrees of the vertices is $(n+1)(n+2)+2(n+1) = (n+1)(n+4)$. If $n = 0$, clearly the maximum vertex degree is 2, and if $n > 0$, then the maximum vertex degree is realized, for example, by the vertex of the C_5 being duplicated, which has degree $n + 2$. This gives us the inequality on θ_1 . The inequalities for $\theta_2 + \theta_3$ and $\theta_2\theta_3$ follow from

$$\sum_{i=1}^3 \theta_i = n + 1, \quad \prod_{i=1}^3 \theta_i = -(2 + 3n), \tag{11}$$

which may be computed directly as in [Lemma 2.5](#) (the sum of all the eigenvalues is 0, the trace of the adjacency matrix, and the product is the determinant of the adjacency matrix). It remains to be shown that θ_1 is unique as the largest eigenvalue.

By the inequality in (i), we see that θ_3 must be negative, but θ_2 must be positive because (ii) tells us that $\theta_2\theta_3 < 0$. Thus, it does not follow immediately that θ_1 , being greater than φ^{-1} , is the unique largest eigenvalue. If $\theta_1 = \theta_2$, then the discriminant Δ of $x^3 - (n+1)x^2 - (n+3)x + (3n+2)$ must be 0. Calculating this discriminant, we find $\Delta = 13n^4 + 108n^3 + 118n^2 + 138n + 120$, which does not have any nonnegative roots, so we can conclude that there does not exist any $n \in \mathbb{Z}^+$, where \mathbb{Z}^+ is the set of positive integers $\{1, 2, 3, \dots\}$, such that $x^3 - (n+1)x^2 - (n+3)x + (3n+2)$ has a double root. The uniqueness of θ_1 as the largest eigenvalue follows. \square

Next, we present the spectrum of \tilde{R}_n .

Theorem 3.10. *If $n > 0$, then $\theta_1 > \theta_2 > \varphi^{-1} > -1 > -\varphi > \theta_3$, and*

$$\text{Spec}(\tilde{R}_n) = \begin{pmatrix} \theta_1 & \theta_2 & \varphi^{-1} & -1 & -\varphi & \theta_3 \\ 1 & 1 & 1 & n-5 & 1 & 1 \end{pmatrix}.$$

If $n = 0$, then $\theta_3 = -\varphi$, $\theta_2 = \varphi^{-1}$, and $\theta_1 = 2$, so

$$\text{Spec}(\tilde{R}_0) = \text{Spec}(C_5) = \begin{pmatrix} 2 & \varphi^{-1} & -\varphi \\ 1 & 2 & 2 \end{pmatrix}.$$

Proof. First, suppose $n = 0$. In this case, the θ_i are roots of $x^3 - x^2 - 3x + 2$, which factors into $(x - 2)(x^2 + x - 1)$, so we find $\theta_1 = 2$, $\theta_2 = \varphi^{-1}$, and $\theta_3 = -\varphi$. Along with [Theorem 3.8](#), this establishes the second part of the theorem.

Now, let $n > 0$. By [Theorem 3.8](#), we note a priori that φ^{-1} , and $-\varphi$ are eigenvalues in the spectrum with multiplicity ≥ 1 and -1 has multiplicity $\geq n - 5$. By the preceding lemma, we saw that $\theta_1 > \theta_2 > 0 > \theta_3$, so what remains to be shown is how these three eigenvalues relate to φ^{-1} , $-\varphi$, and -1 .

First, assume there is some n such that $\theta_2 < \varphi^{-1}$. By of [Lemma 3.9](#)(i), we then see that $\theta_3 > -\varphi$. However, then we find $-1 < \theta_2\theta_3 < 0$, which contradicts [Lemma 3.9](#)(ii) because $-(2+3n)/(n+2) < -1$ for all $n > 0$. Hence, we know that there is no $n > 0$

such that $\theta_2 < \varphi^{-1}$. It follows that $\theta_1 > \theta_2 \geq \varphi^{-1}$ for all $n > 0$. Now, we will show that there is no $n > 0$ such that $0 > \theta_3 > -\varphi$. Assume the contrary. Once again, by [Lemma 3.9](#), this requires that $\theta_2 < \varphi^{-1}$, which we just showed is not the case. Thus, we have another contradiction. Assume $\theta_2 = \varphi^{-1}$ for some $n > 0$. The characteristic polynomial must have all integer coefficients, and the minimal polynomial of φ^{-1} in the polynomial ring $\mathbb{Z}[x]$ is $x^2 + x - 1$, so we deduce $x^2 + x - 1$ must divide $x^3 - (n+1)x^2 - (n+3)x + (3n+2)$. This would require $\theta_3 = -\varphi$, and there must be some $k \in \mathbb{Z}$ such that $(x^2 + x - 1)(x - k) = x^3 - (n+1)x^2 - (n+3)x + (3n+2)$. This simplifies to the system of equations $-k + n = -2$ and $k - 3n = 2$, which has the unique solution $(n, k) = (0, 2)$. This contradicts our assumption $n > 0$. We get the same result if we assume first that $\theta_3 = -\varphi$, so we know that θ_1, θ_2 , and θ_3 are all distinct from each other and $\varphi^{-1}, -\varphi$, and -1 , so the theorem is proved. \square

4. Outlook

The natural question to ask after finding the spectrum of a graph is: *is this graph determined by its spectrum?* In other words, does the spectrum provide a characterization, or are there other nonisomorphic graphs that share the same spectrum. Many articles have considered this question for different families of graphs; see [[Cámara and Haemers 2014](#); [van Dam and Haemers 2003](#); [Topcu et al. 2016](#); [Wang et al. 2009](#); [Wang and Xu 2007](#)]. Given the high multiplicities of the -1 eigenvalues for both $B_{m,m-1}$ and \tilde{R}_n , it may be helpful to approach this problem by considering structural equivalence in graphs and determining which graph structures can be candidates for cospectral graphs with $B_{m,m-1}$ and \tilde{R}_n .

Another problem of interest is to determine how to account for all minimal prime graphs since the reseminant graphs only constitute a proper subset of them. Thus, it is of interest to determine whether there is a finite set of graphs that can generate all the minimal prime graphs through simple graph operations such as vertex duplication. If this were achieved, it would then be possible to study spectral properties of minimal prime graphs (and minimally connected prime graphs) in general.

Acknowledgements

This research was conducted at Texas State University under NSF-REU grant DMS-1757233 during the summer of 2020. Florez, Higgins, Huang and Keller thank the NSF and Shen gratefully acknowledges the financial support from the Office of Undergraduate Research at Washington University in St. Louis. The authors thank Texas State University for running the REU online during this difficult period of social distancing and providing a welcoming and supportive work environment. In particular, Dr. Yong Yang, the director of the REU program, is recognized for conducting an inspired and successful research program. Florez, Higgins, Huang

and Shen also thank their mentor, Keller, for his invaluable advice and guidance throughout this project. Higgins is largely responsible for the results of this paper, with guidance from the Keller. The other authors worked on other projects during the summer REU program. The authors also would like to thank the anonymous referee for helpful suggestions improving the readability of the paper.

References

- [Brouwer and Haemers 2012] A. E. Brouwer and W. H. Haemers, *Spectra of graphs*, Springer, 2012. [MR](#) [Zbl](#)
- [Cámara and Haemers 2014] M. Cámara and W. H. Haemers, “Spectral characterizations of almost complete graphs”, *Discrete Appl. Math.* **176** (2014), 19–23. [MR](#) [Zbl](#)
- [van Dam and Haemers 2003] E. R. van Dam and W. H. Haemers, “Which graphs are determined by their spectrum?”, *Linear Algebra Appl.* **373** (2003), 241–272. [MR](#) [Zbl](#)
- [Florez et al. 2020] C. Florez, J. Higgins, K. Huang, T. M. Keller, and D. Shen, “Minimal prime graphs of solvable groups”, preprint, 2020. [arXiv 2011.10403](#)
- [Godsil and Royle 2001] C. Godsil and G. Royle, *Algebraic graph theory*, Graduate Texts in Mathematics **207**, Springer, 2001. [MR](#) [Zbl](#)
- [Gorshkov and Maslova 2018] I. B. Gorshkov and N. V. Maslova, “Finite almost simple groups whose Gruenberg–Kegel graphs coincide with Gruenberg–Kegel graphs of solvable groups”, *Algebra i Logika* **57:2** (2018), 175–196. In Russian; translated in *Algebra Log.* **57:2** (2018), 115–129. [MR](#) [Zbl](#)
- [Gruber et al. 2015] A. Gruber, T. M. Keller, M. L. Lewis, K. Naughton, and B. Strasser, “A characterization of the prime graphs of solvable groups”, *J. Algebra* **442** (2015), 397–422. [MR](#) [Zbl](#)
- [Nguyen et al. 2019] T. Nguyen, D. Yang, Y. Ge, H. Li, and A. L. Bertozzi, “Applications of structural equivalence to subgraph isomorphism on multichannel multigraphs”, pp. 4913–4920 in 2019 *IEEE International Conference on Big Data* (Los Angeles, CA, 2019), IEEE, Piscataway, NJ, 2019.
- [Topcu et al. 2016] H. Topcu, S. Sorgun, and W. H. Haemers, “On the spectral characterization of pineapple graphs”, *Linear Algebra Appl.* **507** (2016), 267–273. [MR](#) [Zbl](#)
- [Wang and Xu 2007] W. Wang and C.-X. Xu, “Note: on the generalized spectral characterization of graphs having an isolated vertex”, *Linear Algebra Appl.* **425:1** (2007), 210–215. [MR](#) [Zbl](#)
- [Wang et al. 2009] J. Wang, Q. Huang, F. Belardo, and E. M. Li Marzi, “A note on the spectral characterization of dumbbell graphs”, *Linear Algebra Appl.* **431:10** (2009), 1707–1714. [MR](#) [Zbl](#)

Received: 2022-08-29

Revised: 2022-12-10

Accepted: 2022-12-17

cflorez@sas.upenn.edu

*Department of Mathematics, David Rittenhouse Lab,
University of Pennsylvania, Philadelphia, PA, United States*

jh110@illinois.edu

*Department of Mathematics, University of Illinois
Urbana-Champaign, Urbana, IL, United States*

kyle.huang@fu-berlin.de

*Department of Mathematics, Freie Universität Berlin,
Berlin, Germany*

keller@txstate.edu

*Department of Mathematics, Texas State University,
San Marcos, TX, United States*

dwshen@umich.edu

*Department of Mathematics, University of Michigan,
Ann Arbor, MI, United States*

Linear maps preserving the Lorentz spectrum of 3×3 matrices

Maria I. Bueno, Ben Faktor, Rhea Kommerell, Runze Li and Joey Veltri

(Communicated by Stephan Garcia)

For a given 3×3 real matrix A , the eigenvalue complementarity problem relative to the Lorentz cone consists of finding a real number λ and a nonzero vector $x \in \mathbb{R}^3$ such that $x^T(A - \lambda I)x = 0$ and both x and $(A - \lambda I)x$ lie in the Lorentz cone, which consists of all vectors in \mathbb{R}^3 forming a 45° or smaller angle with the positive z -axis. We refer to the set of all solutions λ to this eigenvalue complementarity problem as the Lorentz spectrum of A . Our work concerns the characterization of the linear preservers of the Lorentz spectrum on the space M_3 of 3×3 real matrices, that is, the linear maps $\phi : M_3 \rightarrow M_3$ such that the Lorentz spectra of A and $\phi(A)$ are the same for all A . We have proven that all such linear preservers take the form $\phi(A) = (Q \oplus [1])A(Q^T \oplus [1])$, where Q is an orthogonal 2×2 matrix.

1. Introduction

Let M_n denote the vector space of $n \times n$ real matrices. For a given matrix $A \in M_n$ and a closed convex cone $K \subseteq \mathbb{R}^n$, the *eigenvalue complementarity problem* consists of finding $\lambda \in \mathbb{R}$ and nonzero $x \in \mathbb{R}^n$ satisfying

$$x \in K, \quad (A - \lambda I)x \in K^* \quad \text{and} \quad x^T(A - \lambda I)x = 0, \quad (1-1)$$

where K^* denotes the dual cone of K , that is,

$$K^* = \{y \in \mathbb{R}^n : x^T y \geq 0 \text{ for all } x \in K\}.$$

This problem is a generalization of the standard eigenvalue problem for which $K = \mathbb{R}^n$ and $K^* = \{0\}$.

MSC2020: 15A18, 58C40.

Keywords: Lorentz cone, Lorentz eigenvalues, linear preservers, 3×3 matrices.

The work of Faktor and Kommerell was partially supported by the NSF grant DMS-1850663. This publication is also part of the ‘‘Proyecto de I+D+i PID2019-106362GB-I00 financiado por MCIN/AEI/10.13039/501100011033’’.

We are interested in the eigenvalue complementarity problem on the *Lorentz cone* \mathcal{K}_n , given by

$$\mathcal{K}_n = \left\{ \begin{bmatrix} \xi \\ \eta \end{bmatrix} : \xi \in \mathbb{R}^{n-1}, \eta \in \mathbb{R}, \|\xi\|_2 \leq \eta \right\}.$$

Notably, the Lorentz cone is self-dual, i.e., $\mathcal{K}_n = (\mathcal{K}_n)^*$. Thus, the eigenvalue complementarity problem for the Lorentz cone consists of finding $\lambda \in \mathbb{R}$ and nonzero $x \in \mathbb{R}^n$ satisfying

$$x \in \mathcal{K}_n, \quad (A - \lambda I)x \in \mathcal{K}_n \quad \text{and} \quad x^T(A - \lambda I)x = 0.$$

Any such solution λ is called a *Lorentz eigenvalue* of A , and any associated x is called a *Lorentz eigenvector*. The collection of all such λ is the *Lorentz spectrum* of A , denoted by $\sigma_L(A)$. If λ has an associated Lorentz eigenvector in the interior (resp. boundary) of \mathcal{K}_n , it is called an *interior* (resp. *boundary*) *Lorentz eigenvalue*. The collection of all interior (resp. boundary) Lorentz eigenvalues is called the *interior* (resp. *boundary*) *Lorentz spectrum* of A , denoted by $\sigma_{\text{int}}(A)$ (resp. $\sigma_{\text{bd}}(A)$). Note that $\sigma_L(A)$ is the (not necessarily disjoint) union of $\sigma_{\text{int}}(A)$ and $\sigma_{\text{bd}}(A)$. One distinctive property of the L-spectrum compared to the standard spectrum of a matrix is that it can be infinite. For the sake of brevity, throughout this paper we will write L-eigenvalue, L-eigenvector and L-spectrum in place of Lorentz eigenvalue, Lorentz eigenvector, and Lorentz spectrum, respectively.

The Lorentz cone is an important object of study in several areas of math, especially in optimization. The associated optimization models have applications in several fields such as engineering, finance, and control theory. The Lorentz cone is also helpful to understand the behavior of some linear maps called Z-transformations. For more on applications of the Lorentz cone, see [Alizadeh and Goldfarb 2003; Németh and Gowda 2019].

Recently, see, e.g., [Bueno et al. 2021; 2022; Seeger and Torki 2020], there has been particular interest in the characterization of the linear maps $\phi : M_n \rightarrow M_n$ which preserve the Lorentz spectrum of all matrices, that is, $\sigma_L(A) = \sigma_L(\phi(A))$ for all $A \in M_n$. We call such maps *linear preservers of the Lorentz spectrum*. In studying this problem, we assume $n \geq 3$. For $n = 2$, the Lorentz cone is a polyhedral cone, which is not the case for $n \geq 3$. The characterization of the linear preservers of the Lorentz spectrum for $n = 2$ is an immediate consequence of the characterization of the linear preservers of the Pareto spectrum since the Pareto cone, or nonnegative orthant, is a rotation of the Lorentz cone in \mathbb{R}^2 by 45° [Alizadeh and Shakeri 2017].

For $n \geq 3$, some partial results have been proven in the literature.

Theorem 1.1 [Bueno et al. 2021]. *Let $\phi : M_n \rightarrow M_n$ be a linear preserver of the L-spectrum. Then ϕ is bijective and $\phi(I_n) = I_n$.*

Theorem 1.2 [Bueno et al. 2021]. *Let Q be an orthogonal $(n-1) \times (n-1)$ matrix, and let $\phi : M_n \rightarrow M_n$ be the linear map given by*

$$\phi(A) = \begin{bmatrix} Q & 0 \\ 0 & 1 \end{bmatrix} A \begin{bmatrix} Q^T & 0 \\ 0 & 1 \end{bmatrix}. \quad (1-2)$$

Then ϕ is a linear preserver of the L-spectrum.

Conversely, if $n \geq 3$, $\phi : M_n \rightarrow M_n$ is a linear preserver of the L-spectrum, and $\phi(A) = PAQ$ for some fixed $n \times n$ matrices P and Q , then ϕ satisfies (1-2).

In [Bueno et al. 2021] it was conjectured that, for $n \geq 3$, every linear preserver of the L-spectrum must have the form (1-2). In this paper, we show that this conjecture is true for $n = 3$, providing a full characterization of the linear preservers of the L-spectrum in M_3 . Moreover, we have shown that the linear preservers of the L-spectrum on M_3 also preserve the *nature* of the L-eigenvalues, that is, $\sigma_{\text{int}}(A) = \sigma_{\text{int}}(\phi(A))$ and $\sigma_{\text{bd}}(A) = \sigma_{\text{bd}}(\phi(A))$ for all $A \in M_3$. The strategy used to prove the characterization of the linear preservers of the L-spectrum on M_3 does not seem to be easily generalizable to $n > 3$ since it crucially relies on the restrictive form of 3×3 matrices with infinitely many L-eigenvalues.

The paper is organized as follows: In Section 2, we give some properties of the L-spectrum of a matrix, as well as a full characterization of the boundary L-eigenvalues. In Section 2.1, we characterize the 3×3 real matrices with infinitely many L-eigenvalues. In Section 3, we present the main results of this paper. The proof of one of these results (Theorem 3.1) is somewhat cumbersome and is presented in Sections 4 and 5.

2. The Lorentz spectrum of a matrix

In this section, we discuss the Lorentz spectrum of a matrix in more detail. As mentioned in Section 1, we assume that $n \geq 3$. We also use the notation $\|\cdot\|$ to denote the Euclidean norm of a vector since this is the only vector norm we use in this paper.

Since \mathcal{K}_n is a cone, it is closed under positive scalar multiplication. That is, if $x \in \mathcal{K}_n$ and $\alpha \geq 0$, then $\alpha x \in \mathcal{K}_n$. Thus, if $x = [\tilde{x}^T, x_n]^T \in \mathcal{K}_n$ is an L-eigenvector of a matrix A , then x/x_n is also an L-eigenvector of A associated with the same L-eigenvalue since $x_n > 0$. Moreover, if x is in the interior (resp. boundary) of \mathcal{K}_n , so is x/x_n . Thus, from the definition of interior and boundary L-eigenvalues, we have the following results.

Theorem 2.1. *$\lambda \in \mathbb{R}$ is an interior L-eigenvalue of $A \in M_n$ if and only if there exists $\xi \in \mathbb{R}^{n-1}$ such that $\|\xi\| < 1$ and*

$$(A - \lambda I) \begin{bmatrix} \xi \\ 1 \end{bmatrix} = 0.$$

Proof. Recall that λ is an interior L-eigenvalue of A if and only if there exists an L-eigenvector $x = [\xi^T, 1]^T$ of A associated with λ in the interior of \mathcal{K}^n . This is equivalent to the conditions $\|\xi\| < 1$, $(A - \lambda I)x \in \mathcal{K}_n$, and $x^T(A - \lambda I)x = 0$. Since x and $(A - \lambda I)x$ must be orthogonal and any pair of nonzero orthogonal vectors in \mathcal{K}_n must lie on the boundary, we have $(A - \lambda I)x = 0$, which proves the result. \square

Notice that as an immediate consequence of the previous theorem, we have that the interior L-eigenvalues of a matrix A are also standard eigenvalues of A and that the corresponding L-eigenvectors are also standard eigenvectors.

Theorem 2.2. $\lambda \in \mathbb{R}$ is a boundary L-eigenvalue of $A \in M_n$ if and only if there exists $\xi \in \mathbb{R}^{n-1}$ and $s \geq 0$ such that $\|\xi\| = 1$ and

$$(A - \lambda I) \begin{bmatrix} \xi \\ 1 \end{bmatrix} = s \begin{bmatrix} -\xi \\ 1 \end{bmatrix}.$$

Proof. Recall that λ is a boundary L-eigenvalue of A if and only if there exists an L-eigenvector $x = [\xi^T, 1]^T$ of A associated with λ on the boundary of \mathcal{K}^n . This is equivalent to the conditions $\|\xi\| = 1$, $(A - \lambda I)x \in \mathcal{K}_n$, and $x^T(A - \lambda I)x = 0$. Since $(A - \lambda I)x$ must be orthogonal to x , we know $(A - \lambda I)x$ must be a nonnegative multiple of $[-\xi^T, 1]^T$, which proves the result. \square

Next we give another characterization of the boundary L-eigenvalues of a matrix. We denote the Moore–Penrose inverse of a matrix M by M^\dagger .

Theorem 2.3 [Seeger and Torki 2003]. *Let*

$$A = \begin{bmatrix} \tilde{A} & u \\ v^T & a \end{bmatrix}, \quad \text{where } \tilde{A} \in M_{n-1}, \quad u, v \in \mathbb{R}^{n-1}, \quad \text{and } a \in \mathbb{R}. \quad (2-1)$$

A real number λ is in $\sigma_{\text{bd}}(A)$ if one can write $\lambda = \mu + s$, with $\mu, s \in \mathbb{R}$ and $s \geq 0$, solving (exactly) one of the following systems:

System I:

- I.1 μ is not an eigenvalue of \tilde{A} .
- I.2 $v^T(\tilde{A} - \mu I_{n-1})^{-1}u = a - \mu - 2s$.
- I.3 $\|(\tilde{A} - \mu I_{n-1})^{-1}u\| = 1$.

System II:

- II.1 μ is an eigenvalue of \tilde{A} .
- II.2 $u \in \text{Im}(\tilde{A} - \mu I_{n-1})$.
- II.3 $v \in \text{Im}(\tilde{A}^T - \mu I_{n-1})$.
- II.4 $v^T(\tilde{A} - \mu I_{n-1})^\dagger u = a - \mu - 2s$.
- II.5 $\left\| \begin{bmatrix} \tilde{A} - \mu I_{n-1} \\ v^T \end{bmatrix}^\dagger \begin{bmatrix} u \\ a - \mu - 2s \end{bmatrix} \right\| \leq 1$.

System III:

III.1 μ is an eigenvalue of \tilde{A} with geometric multiplicity 1.

III.2 $u \in \text{Im}(\tilde{A} - \mu I_{n-1})$.

III.3 $v \notin \text{Im}(\tilde{A}^T - \mu I_{n-1})$.

III.4 $\left\| \begin{bmatrix} \tilde{A} - \mu I_{n-1} \\ v^T \end{bmatrix}^\dagger \begin{bmatrix} u \\ a - \mu - 2s \end{bmatrix} \right\| = 1$.

System IV:

IV.1 μ is an eigenvalue of \tilde{A} with geometric multiplicity at least 2.

IV.2 $u \in \text{Im}(\tilde{A} - \mu I_{n-1})$.

IV.3 $v \notin \text{Im}(\tilde{A}^T - \mu I_{n-1})$.

IV.4 $\left\| \begin{bmatrix} \tilde{A} - \mu I_{n-1} \\ v^T \end{bmatrix}^\dagger \begin{bmatrix} u \\ a - \mu - 2s \end{bmatrix} \right\| \leq 1$.

A distinguishing property of the L-spectrum compared to the standard spectrum of a matrix is that, while an $n \times n$ real matrix cannot have more than n standard eigenvalues, it may have infinitely many L-eigenvalues. The next theorem characterizes the matrices with this property.

Theorem 2.4 [Seeger and Torki 2003]. *Let $n \geq 3$ and let $A \in M_n$ be partitioned as in (2-1). The matrix A has infinitely many L-eigenvalues if and only if System IV in Theorem 2.3 is satisfied for a real eigenvalue μ of \tilde{A} and for all s in an interval $[s_1, s_2]$, where $0 \leq s_1 < s_2$.*

2.1. Matrices in M_3 with infinitely many L-eigenvalues. Here we characterize the matrices in M_3 with infinitely many L-eigenvalues. First we give a technical result, which is used in the proof of Theorem 2.6. We use the notation $\text{tr}(A)$ for the trace of a matrix A . This result can be found in [Ben-Israel and Greville 2003, Exercise 19, page 49], but we include a proof for completeness.

Lemma 2.5. *Let A be a real matrix of rank 1. Then*

$$A^\dagger = \frac{1}{\text{tr}(A^T A)} A^T.$$

Proof. Since A has rank 1, we know $A = uv^T$ for some nonzero vectors u and v . Hence

$$A^T A = vu^T uv^T = \|u\|^2 vv^T$$

also has rank 1. Thus, since $A^T A$ is symmetric, it is diagonalizable and has exactly one nonzero eigenvalue, namely, $\text{tr}(A^T A)$. Hence there exists a nonsingular matrix P such that

$$A^T A = P \begin{bmatrix} \text{tr}(A^T A) & 0 \\ 0 & 0 \end{bmatrix} P^{-1}.$$

Let $A = U\Sigma V^T$ be a singular value decomposition of A . Then

$$\Sigma = \begin{bmatrix} \sqrt{\operatorname{tr}(A^T A)} & 0 \\ 0 & 0 \end{bmatrix}$$

and

$$A^\dagger = V\Sigma^\dagger U^T = V \begin{bmatrix} 1/\sqrt{\operatorname{tr}(A^T A)} & 0 \\ 0 & 0 \end{bmatrix} U^T = \frac{1}{\operatorname{tr}(A^T A)} V\Sigma^T U^T = \frac{1}{\operatorname{tr}(A^T A)} A^T. \quad \square$$

The following theorem is the main result in this section.

Theorem 2.6. *Let $A \in M_3$. Then A has infinitely many L -eigenvalues if and only if*

$$A = \begin{bmatrix} cI_2 & 0 \\ v^T & a \end{bmatrix}, \quad \text{where } v \neq 0, a, c \in \mathbb{R}, \text{ and } c < a + \|v\|.$$

Moreover,

$$\left[\max \left\{ c, \frac{a+c-\|v\|}{2} \right\}, \frac{a+c+\|v\|}{2} \right] \subseteq \sigma_{\text{bd}}(A). \quad (2-2)$$

Proof. Assume that A is a 3×3 real matrix with infinitely many L -eigenvalues. Let us partition A as

$$A = \begin{bmatrix} \tilde{A} & u \\ v^T & a \end{bmatrix}, \quad \text{where } \tilde{A} \in M_2, u, v \in \mathbb{R}^2, \text{ and } a \in \mathbb{R}.$$

By [Theorem 2.4](#), A must have a boundary L -eigenvalue λ satisfying System IV in [Theorem 2.3](#). Thus, by [Theorem 2.2](#), $\lambda = \mu + s$ with $s \geq 0$, and there is a solution to the system of equations

$$\begin{bmatrix} \tilde{A} - \mu I_2 & u \\ v^T & a - \mu - 2s \end{bmatrix} \begin{bmatrix} \xi \\ 1 \end{bmatrix} = 0, \quad \|\xi\| = 1.$$

By condition IV.1, since μ is an eigenvalue of \tilde{A} of geometric multiplicity 2, we have $\operatorname{rank}(\tilde{A} - \mu I_2) = 0$. Thus $\tilde{A} = \mu I_2$. Moreover, by condition IV.2, $u \in \operatorname{Im}(\tilde{A} - \mu I_2)$, which means $u = 0$. By condition IV.3, we deduce that $v \neq 0$. Finally, by condition IV.4, we have

$$\left\| \begin{bmatrix} 0 \\ v^T \end{bmatrix}^\dagger \begin{bmatrix} 0 \\ a - \mu - 2s \end{bmatrix} \right\| \leq 1. \quad (2-3)$$

By [Lemma 2.5](#),

$$\begin{bmatrix} 0 \\ v^T \end{bmatrix}^\dagger = \frac{1}{\|v\|^2} [0, v].$$

Hence, (2-3) reduces to

$$\frac{|a - \mu - 2s|}{\|v\|} = \frac{1}{\|v\|^2} \left\| [0, v] \begin{bmatrix} 0 \\ a - \mu - 2s \end{bmatrix} \right\| \leq 1,$$

or equivalently,

$$\frac{a - \|v\| - \mu}{2} \leq s \leq \frac{a + \|v\| - \mu}{2}. \tag{2-4}$$

Since there are infinitely many L-eigenvalues and $s \geq 0$, we deduce $a + \|v\| - \mu > 0$. Then by taking $c = \mu$, we have an interval of L-eigenvalues $\lambda = \mu + s$ given by (2-2).

The converse holds by [Theorem 2.4](#). □

3. Main results

We now state the two main results of the paper. The first result, together with [Theorem 1.2](#), provides a full characterization of the linear maps $\phi : M_3 \rightarrow M_3$ that preserve the Lorentz spectrum.

Theorem 3.1. *Let $\phi : M_3 \rightarrow M_3$ be a linear preserver of the Lorentz spectrum. Then there exists an orthogonal matrix $Q \in M_2$ such that*

$$\phi(A) = \begin{bmatrix} Q & 0 \\ 0 & 1 \end{bmatrix} A \begin{bmatrix} Q^T & 0 \\ 0 & 1 \end{bmatrix} \quad \text{for all } A \in M_3.$$

We call Q the *orthogonal matrix associated with ϕ* .

The strategy we employ to prove [Theorem 3.1](#) is to use the linearity of the linear preservers applied to a decomposition of M_3 as a direct sum of three subspaces. More explicitly, we decompose M_3 as

$$M_3 = \left\{ \begin{bmatrix} 0 & 0 \\ v^T & a \end{bmatrix} \right\} \oplus \left\{ \begin{bmatrix} \tilde{A} & 0 \\ 0 & 0 \end{bmatrix} \right\} \oplus \left\{ \begin{bmatrix} 0 & u \\ 0 & 0 \end{bmatrix} \right\} =: S_1 \oplus S_2 \oplus S_3, \tag{3-1}$$

where $\tilde{A} \in M_2$, $u, v \in \mathbb{R}^2$, and $a \in \mathbb{R}$. The image of an arbitrary matrix $A \in M_3$ under ϕ is the sum of the images of the projections of A onto each of these subspaces.

The proof of [Theorem 3.1](#) is a direct consequence of [Theorems 4.5, 5.7, and 5.11](#), which give the images of the matrices in S_1, S_2 and S_3 , respectively, under a linear preserver of the L-spectrum. The second main result is presented next and shows that any linear preserver of the L-spectrum on M_3 must preserve the nature of the L-eigenvalues of a matrix.

Theorem 3.2. *Let $\phi : M_3 \rightarrow M_3$ be a linear preserver of the L-spectrum. Then, for all $A \in M_3$,*

$$\sigma_{\text{int}}(A) = \sigma_{\text{int}}(\phi(A)) \quad \text{and} \quad \sigma_{\text{bd}}(A) = \sigma_{\text{bd}}(\phi(A)).$$

Proof. Let ϕ be a linear preserver of the L-spectrum on M_3 , and let $A \in M_3$. Since ϕ^{-1} is also a linear preserver of the L-spectrum on M_3 , it is enough to show that $\sigma_{\text{int}}(A) \subseteq \sigma_{\text{int}}(\phi(A))$ and $\sigma_{\text{bd}}(A) \subseteq \sigma_{\text{bd}}(\phi(A))$.

Let $\lambda \in \sigma_{\text{int}}(A)$. We want to show that $\lambda \in \sigma_{\text{int}}(\phi(A))$. By [Theorem 2.1](#), there exists $\xi \in \mathbb{R}^2$ with $\|\xi\| < 1$ such that

$$(A - \lambda I_3) \begin{bmatrix} \xi \\ 1 \end{bmatrix} = 0.$$

Let $\widehat{Q} = \begin{bmatrix} Q & 0 \\ 0 & 1 \end{bmatrix}$, where Q is the orthogonal matrix associated with ϕ given by [Theorem 3.1](#). Then we have

$$(\phi(A) - \lambda I_3) \begin{bmatrix} Q\xi \\ 1 \end{bmatrix} = \widehat{Q}(A - \lambda I_3)\widehat{Q}^T \widehat{Q} \begin{bmatrix} \xi \\ 1 \end{bmatrix} = 0,$$

where $\|Q\xi\| = \|\xi\| < 1$ since Q is orthogonal. Thus, $\lambda \in \sigma_{\text{int}}(\phi(A))$ and hence $\sigma_{\text{int}}(A) \subseteq \sigma_{\text{int}}(\phi(A))$. By taking $\|\xi\| = 1$ instead of $\|\xi\| < 1$, we likewise have $\sigma_{\text{bd}}(A) \subseteq \sigma_{\text{bd}}(\phi(A))$. \square

4. Image of matrices in \mathcal{S}_1 under a linear preserver

As explained in [Section 3](#), in order to prove [Theorem 3.1](#), we determine the images of matrices in the three subspaces \mathcal{S}_1 , \mathcal{S}_2 and \mathcal{S}_3 given in (3-1) under a linear preserver of the L-spectrum. In this section, we focus on \mathcal{S}_1 .

We begin with a lemma that sheds light on the possible images under a linear preserver of the L-spectrum of matrices in \mathcal{S}_1 with infinitely many L-eigenvalues.

Lemma 4.1. *Let $\phi : M_3 \rightarrow M_3$ be a linear preserver of the L-spectrum and*

$$A = \begin{bmatrix} 0 & 0 \\ v^T & a \end{bmatrix}, \quad \text{where } v \neq 0 \text{ and } 0 < a + \|v\|.$$

Then either

$$\phi(A) = \begin{bmatrix} 0 & 0 \\ w^T & a \end{bmatrix}, \quad \text{where } \|w\| = \|v\|, \tag{4-1}$$

or

$$\phi(A) = \begin{bmatrix} 0 & 0 \\ w^T & a + \|v\| - \|w\| \end{bmatrix}, \quad \text{where } \frac{a + \|v\|}{2} \leq \|w\| \leq a + \|v\| \\ \text{and } a - \|v\| \leq 0 \leq a. \tag{4-2}$$

Proof. Since ϕ preserves the L-spectrum, by [Theorem 2.6](#), we have

$$\phi(A) = \begin{bmatrix} dI_2 & 0 \\ w^T & b \end{bmatrix}, \quad \text{where } w \neq 0, \ d, b \in \mathbb{R}, \text{ and } d < b + \|w\|.$$

We consider four cases, in which we make repeated use of [Lemma A.3](#) to determine the possibilities for $\phi(A)$ and its L-spectrum.

Case I: Assume that $0 < a - \|v\|$. In this case,

$$\sigma_L(A) = \{a\} \cup \left[\frac{a - \|v\|}{2}, \frac{a + \|v\|}{2} \right].$$

Hence $\phi(A)$ must have an isolated L-eigenvalue. We have two possible cases for $\phi(A)$:

Subcase I.1: $d < b - \|w\|$. In this case,

$$\sigma_L(\phi(A)) = \{b\} \cup \left[\frac{d+b-\|w\|}{2}, \frac{d+b+\|w\|}{2} \right].$$

Since $\sigma_L(A) = \sigma_L(\phi(A))$, we have

$$b = a, \quad d = 0, \quad \text{and} \quad \|v\| = \|w\|,$$

which leads to (4-1).

Subcase I.2: $b - \|w\| < d < b + \|w\|$ and $d > b$. In this case,

$$\sigma_L(\phi(A)) = \{b\} \cup \left[d, \frac{d+b+\|w\|}{2} \right].$$

Since $\sigma_L(A) = \sigma_L(\phi(A))$, we have

$$b = a, \quad d = \frac{a - \|v\|}{2}, \quad \text{and} \quad d + \|w\| = \|v\|.$$

However, this implies

$$a = b < d + \|w\| = \|v\|,$$

a contradiction since $\|v\| < a$ by assumption. So this subcase is impossible.

Case II: Assume that $a - \|v\| = 0$. In this case,

$$\sigma_L(A) = [0, \|v\|].$$

Hence $\phi(A)$ does not have isolated L-eigenvalues. We have two possible cases for $\phi(A)$:

Subcase II.1: $d = b - \|w\|$. In this case,

$$\sigma_L(\phi(A)) = [d, \|w\| + d].$$

Since $\sigma_L(A) = \sigma_L(\phi(A))$, we have

$$d = 0, \quad \|w\| = \|v\|, \quad \text{and} \quad b = d + \|w\| = \|v\| = a,$$

which leads to (4-1).

Subcase II.2: $b - \|w\| < d < b + \|w\|$ and $d \leq b$. In this case,

$$\sigma_L(\phi(A)) = \left[d, \frac{b + \|w\| + d}{2} \right].$$

Since $\sigma_L(A) = \sigma_L(\phi(A))$, we have

$$d = 0 \quad \text{and} \quad \frac{b + \|w\|}{2} = \|v\|,$$

or equivalently,

$$d = 0 \quad \text{and} \quad b + \|w\| = 2\|v\| = a + \|v\|.$$

Combining this with the inequalities that define this case, we see that

$$\|w\| > b - d = a + \|v\| - \|w\| = 2\|v\| - \|w\|,$$

and hence $\|w\| > \|v\|$. Similarly,

$$a - \|v\| = 0 = d \leq b = a + \|v\| - \|w\|,$$

and hence $\|w\| \leq 2\|v\|$. Altogether, Subcase II.2 gives the conditions

$$d = 0, \quad b = a + \|v\| - \|w\|, \quad \|v\| < \|w\| \leq 2\|v\|, \quad \text{and} \quad a - \|v\| = 0,$$

which leads to (4-2).

Case III: Assume that $a - \|v\| < 0 < a + \|v\|$ and $a \geq 0$. In this case,

$$\sigma_L(A) = \left[0, \frac{a + \|v\|}{2} \right].$$

Hence $\phi(A)$ does not have isolated L-eigenvalues. We have two possible cases for $\phi(A)$:

Subcase III.1: $b - \|w\| = d$. In this case,

$$\sigma_L(\phi(A)) = [d, \|w\| + d].$$

Since $\sigma_L(A) = \sigma_L(\phi(A))$, we have

$$d = 0, \quad \|w\| = \frac{a + \|v\|}{2}, \quad b = d + \|w\| = \frac{a + \|v\|}{2}.$$

Note that $b = a + \|v\| - \|w\|$ and $a - \|v\| \leq 0 \leq a$. Thus this case leads to (4-2).

Subcase III.2: $b - \|w\| < d < b + \|w\|$ and $d \leq b$. In this case,

$$\sigma_L(\phi(A)) = \left[d, \frac{b + \|w\| + d}{2} \right].$$

Since $\sigma_L(A) = \sigma_L(\phi(A))$, we have

$$d = 0 \quad \text{and} \quad b + \|w\| = a + \|v\|.$$

Combining this with the inequalities that define Subcase III.2, we see that

$$\|w\| > b - d = a + \|v\| - \|w\|,$$

and hence $\|w\| > (a + \|v\|)/2$. Similarly,

$$0 = d \leq b = a + \|v\| - \|w\|,$$

and hence $\|w\| \leq a + \|v\|$, which leads to (4-2).

Case IV: Assume that $a - \|v\| < 0 < a + \|v\|$ and $a < 0$. In this case,

$$\sigma_L(A) = \{a\} \cup \left[0, \frac{a + \|v\|}{2}\right].$$

Hence, $\phi(A)$ must have an isolated L-eigenvalue. We have two possible cases for $\phi(A)$:

Subcase IV.1: $d < b - \|w\|$. In this case,

$$\sigma_L(\phi(A)) = \{b\} \cup \left[\frac{d + b - \|w\|}{2}, \frac{d + b + \|w\|}{2}\right].$$

Since $\sigma_L(A) = \sigma_L(\phi(A))$, we have

$$b = a, \quad \frac{d + a - \|w\|}{2} = 0 \quad \text{and} \quad d + \|w\| = \|v\|.$$

However, this implies

$$a = b > d + \|w\| = \|v\|,$$

a contradiction since $a - \|v\| < 0$ by assumption. So this subcase is impossible.

Subcase IV.2: $b - \|w\| < d < b + \|w\|$ and $d > b$. In this case,

$$\sigma_L(\phi(A)) = \{b\} \cup \left[d, \frac{d + b + \|w\|}{2}\right].$$

Since $\sigma_L(A) = \sigma_L(\phi(A))$, we have

$$b = a, \quad d = 0, \quad \text{and} \quad \|w\| = \|v\|,$$

which leads to (4-1). □

We show next that the subspace \mathcal{S}_1 is invariant under linear preservers of the L-spectrum.

Lemma 4.2. *Let $\phi : M_3 \rightarrow M_3$ be a linear preserver of the L-spectrum. Then the subspace*

$$\mathcal{S}_1 = \left\{ \begin{bmatrix} 0 & 0 \\ v^T & a \end{bmatrix} : v \in \mathbb{R}^2, a \in \mathbb{R} \right\}$$

of M_3 is ϕ -invariant, that is, $\phi(\mathcal{S}_1) \subseteq \mathcal{S}_1$.

Proof. We consider three cases:

Case I: Assume $A = \begin{bmatrix} 0 & 0 \\ v^T & a \end{bmatrix}$, with $v \neq 0$ and $0 < a + \|v\|$. Then by Lemma 4.1, $\phi(A) \in \mathcal{S}_1$.

Case II: Assume $A = \begin{bmatrix} 0 & 0 \\ 0 & a \end{bmatrix}$. Let $v \in \mathbb{R}^2$ be a nonzero vector such that $0 < a + \|v\|$. We have

$$A = \begin{bmatrix} 0 & 0 \\ 0 & a \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ v^T & a \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ v^T & 0 \end{bmatrix} =: B - M.$$

Note that B and M have infinitely many L-eigenvalues by [Theorem 2.6](#). Thus, by Case I, $\phi(B), \phi(M) \in \mathcal{S}_1$. Then since ϕ is linear and \mathcal{S}_1 is a subspace,

$$\phi(A) = \phi(B) - \phi(M) \in \mathcal{S}_1.$$

Case III: Assume $A = \begin{bmatrix} 0 & 0 \\ v^T & a \end{bmatrix}$ with $v \neq 0$ and $0 \geq a + \|v\|$. Let $d \in \mathbb{R}$ be such that $0 < a + d + \|v\|$. Then we have

$$A = \begin{bmatrix} 0 & 0 \\ v^T & a \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ v^T & a + d \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & d \end{bmatrix} =: B - M.$$

Note that $\phi(B) \in \mathcal{S}_1$ by Case I and that $\phi(M) \in \mathcal{S}_1$ by Case II. Then since ϕ is linear and \mathcal{S}_1 is a subspace,

$$\phi(A) = \phi(B) - \phi(M) \in \mathcal{S}_1. \quad \square$$

We show next that we can partition the subspace \mathcal{S}_1 into three subsets which are also ϕ -invariant.

Lemma 4.3. *Let $\phi : M_3 \rightarrow M_3$ be a linear preserver of the Lorentz spectrum. Then $\phi(\mathcal{C}_1) \subseteq \mathcal{C}_1$, $\phi(\mathcal{C}_2 \cup \mathcal{C}_4) \subseteq \mathcal{C}_2 \cup \mathcal{C}_4$, and $\phi(\mathcal{C}_3 \cup \mathcal{C}_5) \subseteq \mathcal{C}_3 \cup \mathcal{C}_5$, where*

$$\begin{aligned} \mathcal{C}_1 &:= \left\{ \begin{bmatrix} 0 & 0 \\ v^T & a \end{bmatrix} : 0 \neq v \in \mathbb{R}^2, a \in \mathbb{R}, 0 < a + \|v\| \right\}, \\ \mathcal{C}_2 &:= \left\{ \begin{bmatrix} 0 & 0 \\ v^T & a \end{bmatrix} : 0 \neq v \in \mathbb{R}^2, a \in \mathbb{R}, a + \|v\| = 0 \right\}, \\ \mathcal{C}_3 &:= \left\{ \begin{bmatrix} 0 & 0 \\ v^T & a \end{bmatrix} : 0 \neq v \in \mathbb{R}^2, a \in \mathbb{R}, a + \|v\| < 0 \right\}, \\ \mathcal{C}_4 &:= \left\{ \begin{bmatrix} 0 & 0 \\ 0 & a \end{bmatrix} : a \in \mathbb{R}, a > 0 \right\}, \\ \mathcal{C}_5 &:= \left\{ \begin{bmatrix} 0 & 0 \\ 0 & a \end{bmatrix} : a \in \mathbb{R}, a \leq 0 \right\}. \end{aligned}$$

Proof. The result for \mathcal{C}_1 follows from [Lemma 4.1](#). Then by [Lemmas A.3](#) and [A.1](#), every matrix in $\mathcal{C}_2 \cup \mathcal{C}_4$ has exactly two L-eigenvalues, and every matrix in $\mathcal{C}_3 \cup \mathcal{C}_5$ has exactly one L-eigenvalue. Thus, the results for $\mathcal{C}_2 \cup \mathcal{C}_4$ and for $\mathcal{C}_3 \cup \mathcal{C}_5$ follow from [Lemma 4.2](#). \square

Next we show that any linear preserver of the L-spectrum restricted to the subspace $\mathcal{C}_4 \cup \mathcal{C}_5$ is the identity map. Henceforth we denote the matrix $e_i e_j^T$ by E_{ij} .

Lemma 4.4. *Let $\phi : M_3 \rightarrow M_3$ be a linear preserver of the L-spectrum, and let $A = \begin{bmatrix} 0 & 0 \\ 0 & a \end{bmatrix}$, where $a \in \mathbb{R}$. Then, $\phi(A) = A$.*

Proof. Assume first that $A \in \mathcal{C}_4$. Then $a > 0$ and $\sigma_L(A) = \{a, a/2\}$ by [Lemma A.1](#). We know that $\phi(A) \in \mathcal{C}_2 \cup \mathcal{C}_4$ by [Lemma 4.3](#). If $\phi(A) \in \mathcal{C}_2$, then $\sigma_L(\phi(A)) = \{a, 0\}$ by [Lemma A.3](#), a contradiction. Therefore, $\phi(A) \in \mathcal{C}_4$, which implies $\phi(A) = A$ by [Lemma A.1](#). In particular, we must have $\phi(E_{33}) = E_{33}$.

Now assume that $A \in \mathcal{C}_5$. Then $A = aE_{33}$ and $a \leq 0$, so by linearity, we have

$$\phi(A) = \phi(aE_{33}) = a\phi(E_{33}) = aE_{33} = A. \quad \square$$

We next present the main result in this section, which gives the image of matrices in the subspace \mathcal{S}_1 under linear preservers of the L-spectrum.

Theorem 4.5. *Let $\phi : M_3 \rightarrow M_3$ be a linear preserver of the L-spectrum. Then there exists an orthogonal 2×2 matrix Q such that, for any matrix $A = \begin{bmatrix} 0 & 0 \\ v^T & a \end{bmatrix} \in \mathcal{S}_1$,*

$$\phi(A) = \begin{bmatrix} 0 & 0 \\ (Qv)^T & a \end{bmatrix}.$$

We call Q the *orthogonal matrix associated with ϕ* .

Proof. By [Lemma 4.4](#) and the linearity of ϕ , it is enough to show that

$$\phi(B) := \phi\left(\begin{bmatrix} 0 & 0 \\ v^T & 0 \end{bmatrix}\right) = \begin{bmatrix} 0 & 0 \\ (Qv)^T & 0 \end{bmatrix}$$

for some orthogonal matrix Q independent of v .

This is trivially true for $v = 0$ since $\phi(0) = 0$, so assume $v \neq 0$. Since ϕ preserves the L-spectrum and $B \in \mathcal{C}_1$, by [Lemma 4.1](#), we have

$$\phi(B) = \begin{bmatrix} 0 & 0 \\ w^T & \|v\| - \|w\| \end{bmatrix}$$

for some $w \in \mathbb{R}^2$ such that $\|v\|/2 \leq \|w\| \leq \|v\|$. Let

$$H = \begin{bmatrix} 0 & 0 \\ v^T & -\|v\| \end{bmatrix} \in \mathcal{C}_2.$$

By linearity and by [Lemma 4.4](#),

$$\begin{aligned} \phi\left(\begin{bmatrix} 0 & 0 \\ v^T & -\|v\| \end{bmatrix}\right) &= \phi\left(\begin{bmatrix} 0 & 0 \\ 0 & -\|v\| \end{bmatrix}\right) + \phi\left(\begin{bmatrix} 0 & 0 \\ v^T & 0 \end{bmatrix}\right) \\ &= \begin{bmatrix} 0 & 0 \\ 0 & -\|v\| \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ w^T & \|v\| - \|w\| \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ w^T & -\|w\| \end{bmatrix}. \end{aligned}$$

Notice that $-\|w\| \in \sigma_{\text{int}}(\phi(H))$ and $\sigma_L(H) = \{-\|v\|, 0\}$ by [Lemma A.3](#). Since $\|w\| \geq \|v\|/2 > 0$, this implies $\|w\| = \|v\|$. In particular, we have

$$\phi(E_{31}) = \begin{bmatrix} 0 & 0 \\ p^T & 0 \end{bmatrix} \quad \text{and} \quad \phi(E_{32}) = \begin{bmatrix} 0 & 0 \\ q^T & 0 \end{bmatrix}, \quad \text{where } \|p\| = \|q\| = 1.$$

Let

$$p = [p_1, p_2]^T, \quad q = [q_1, q_2]^T, \quad \text{and} \quad Q = \begin{bmatrix} p_1 & q_1 \\ p_2 & q_2 \end{bmatrix}.$$

Then for any

$$A = \begin{bmatrix} 0 & 0 \\ v^T & 0 \end{bmatrix}, \quad \text{where } v = [v_1, v_2]^T,$$

we have

$$\phi(A) = v_1\phi(E_{31}) + v_2\phi(E_{32}) = \begin{bmatrix} 0 & 0 \\ v_1p^T + v_2q^T & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ (Qv)^T & 0 \end{bmatrix}.$$

Since $\|Qv\| = \|v\|$ for all v , we deduce that Q is an orthogonal matrix. \square

5. Image of matrices in \mathcal{S}_2 and \mathcal{S}_3 under a linear preserver

We begin this section with some technical lemmas that will be used in the proof of several results. Here we denote the adjugate of a matrix A by $\text{adj}(A)$ and the spectral radius of a matrix A by $\rho(A)$.

5.1. Auxiliary results.

Lemma 5.1 (Rayleigh–Ritz theorem). *Let S be a symmetric matrix and let λ_{\max} be the largest eigenvalue of S . Then*

$$\lambda_{\max} = \max_{\|x\|=1} x^T S x.$$

Lemma 5.2. *Let*

$$B_a = \begin{bmatrix} \tilde{M} & r \\ h^T & a+t \end{bmatrix}, \quad \text{where } \tilde{M} \in M_2, \ h, r \in \mathbb{R}^2, \ \text{and } t \in \mathbb{R} \text{ are fixed.}$$

Assume there is some $a_0 \in \mathbb{R}$ such that $a \in \sigma_L(B_a)$ for all $a \geq a_0$. Then a is an interior L -eigenvalue of B_a for all sufficiently large a , and we have

$$t = 0, \quad h^T r = 0, \quad \text{and} \quad h^T \text{adj}(\tilde{M})r = 0.$$

Proof. Suppose that $a \geq a_0$ is a boundary L -eigenvalue of $\phi(B_a)$. Then, by [Theorem 2.2](#), $a = \mu + s$ for some $s \geq 0$, and there exists ξ with $\|\xi\| = 1$ such that

$$0 = \begin{bmatrix} (\tilde{M} - \mu)\xi + r \\ h^T \xi + t - s \end{bmatrix} = \begin{bmatrix} \tilde{M}\xi + (s - a)\xi + r \\ h^T \xi + t - s \end{bmatrix}.$$

From the second equation, we have $s = h^T \xi + t$, and replacing it in the first equation, we get

$$0 = \tilde{M}\xi + (s - a)\xi + r = \tilde{M}\xi + (h^T \xi + t - a)\xi + r.$$

Multiplying on the left by ξ^T and taking into account that $\|\xi\| = 1$, we have

$$\begin{aligned} 0 &= \xi^T \tilde{M} \xi + h^T \xi + t - a + \xi^T r \\ &= \frac{1}{2} \xi^T \tilde{M} \xi + \frac{1}{2} \xi^T \tilde{M} \xi + h^T \xi + r^T \xi + t - a \\ &= \frac{1}{2} \xi^T (\tilde{M} + \tilde{M}^T) \xi + (h+r)^T \xi + t - a \\ &\leq \frac{1}{2} \lambda_{\max}(\tilde{M} + \tilde{M}^T) + \|h+r\| + t - a, \end{aligned} \tag{5-1}$$

where the third equality follows from the fact that $\xi^T \tilde{M} \xi$ is a number and, consequently, is equal to its transpose. The inequality follows from the Rayleigh–Ritz theorem and the Cauchy–Schwarz inequality.

Hence for $a > \frac{1}{2} \lambda_{\max}(\tilde{M} + \tilde{M}^T) + \|h+r\| + t$, condition (5-1) fails, which means that a must be an interior L-eigenvalue of $\phi(B_a)$ for all sufficiently large a . This implies there is some ξ with $\|\xi\| < 1$ such that

$$0 = \begin{bmatrix} \tilde{M} - aI & r \\ h^T & t \end{bmatrix} \begin{bmatrix} \xi \\ 1 \end{bmatrix} = \begin{bmatrix} (\tilde{M} - aI)\xi + r \\ h^T \xi + t \end{bmatrix}.$$

For $a > \rho(\tilde{M})$, we know that $\tilde{M} - aI$ must be invertible. From the first equation, we have $\xi = -(\tilde{M} - aI)^{-1}r$, and replacing it in the second equation, we get

$$\begin{aligned} 0 &= h^T \xi + t = t - h^T (\tilde{M} - aI)^{-1}r \\ &= t - \frac{h^T \operatorname{adj}(\tilde{M} - aI)r}{\det(\tilde{M} - aI)} = t - \frac{h^T (\operatorname{adj}(\tilde{M}) - aI)r}{a^2 - a \operatorname{tr}(\tilde{M}) + \det(\tilde{M})}. \end{aligned}$$

This implies

$$\begin{aligned} 0 &= t(a^2 - a \operatorname{tr}(\tilde{M}) + \det(\tilde{M})) - h^T \operatorname{adj}(\tilde{M})r + ah^T r \\ &= ta^2 + (h^T r - t \operatorname{tr}(\tilde{M}))a + t \det(\tilde{M}) - h^T \operatorname{adj}(\tilde{M})r. \end{aligned}$$

Since this holds for all sufficiently large a , we must have $t = 0$, $h^T r = h^T r - t \operatorname{tr}(\tilde{M}) = 0$, and $h^T \operatorname{adj}(\tilde{M})r = -t \det(\tilde{M}) + h^T \operatorname{adj}(\tilde{M})r = 0$. □

5.2. Image of matrices in \mathcal{S}_2 under a linear preserver. Even though our focus in this section is matrices in \mathcal{S}_2 , we start with a partial result for matrices in \mathcal{S}_3 .

Lemma 5.3. *Let $\phi : M_3 \rightarrow M_3$ be a linear preserver of the L-spectrum with associated orthogonal matrix Q , and let*

$$B = \begin{bmatrix} 0 & u \\ 0 & 0 \end{bmatrix}, \quad \text{where } u \neq 0.$$

Then

$$\phi(B) = \begin{bmatrix} \tilde{B} & Qu \\ w^T & 0 \end{bmatrix}, \quad \text{where } \det(\tilde{B}) = 0, \ w^T Qu = 0, \ \text{and } \operatorname{adj}(\tilde{B})Qu = \operatorname{tr}(\tilde{B})Qu.$$

Proof. Let

$$\phi(B) = \begin{bmatrix} \tilde{B} & v \\ w^T & b \end{bmatrix}, \quad a > \|u\|, \quad \text{and} \quad B^\perp = \begin{bmatrix} 0 & 0 \\ z^T & a \end{bmatrix},$$

where z is any vector orthogonal to u . By [Theorem 4.5](#),

$$\phi(B^\perp) = \begin{bmatrix} 0 & 0 \\ (Qz)^T & a \end{bmatrix}.$$

Thus,

$$\phi(B + B^\perp) = \phi(B) + \phi(B^\perp) = \begin{bmatrix} \tilde{B} & v \\ (w + Qz)^T & a + b \end{bmatrix}.$$

Since a is an (interior) L-eigenvalue of $B + B^\perp$ with L-eigenvector $[u^T/a, 1]^T$, we know $a \in \sigma_L(\phi(B + B^\perp))$. By [Lemma 5.2](#), we have $b = 0$, $(w + Qz)^T v = 0$, and $(w + Qz)^T \text{adj}(\tilde{B})v = 0$. By setting $z = 0$, we have, in particular, $w^T v = 0$ and $w^T \text{adj}(\tilde{B})v = 0$, which also implies $z^T Q^T v = z^T Q^T \text{adj}(\tilde{B})v = 0$ for all z orthogonal to u .

Let $c \in \mathbb{R}$, $c > 0$, and consider now the matrix $H := B + B^T + cE_{33}$. Then, by [Theorem 4.5](#),

$$\phi(H) = \begin{bmatrix} \tilde{B} & v \\ (w + Qu)^T & c \end{bmatrix}.$$

By [Lemma A.2](#), $m := (c + \sqrt{c^2 + 4\|u\|^2})/2$ is an (interior) L-eigenvalue of H , so m is also an L-eigenvalue of $\phi(H)$.

Step 1: Suppose m is a boundary L-eigenvalue of $\phi(H)$. Then, using an argument similar to that used in the first part of the proof of [Lemma 5.2](#), we get

$$\begin{aligned} 0 &= \xi^T \tilde{B} \xi + (w + Qu)^T \xi + c - 2m + \xi^T v \\ &\leq \frac{1}{2} \lambda_{\max}(\tilde{B} + \tilde{B}^T) + \|w + Qu + v\| - \sqrt{c^2 + 4\|u\|^2}. \end{aligned}$$

Hence, for

$$c > \sqrt{\left(\frac{1}{2} \lambda_{\max}(\tilde{B} + \tilde{B}^T) + \|w + Qu + v\|\right)^2 - 4\|u\|^2}$$

(or for any c if the radicand is negative), this condition fails, which means that for sufficiently large c , m must be an interior L-eigenvalue of $\phi(H)$.

Step 2: Because m is an interior L-eigenvalue of $\phi(H)$ for sufficiently large values of c , there must be some ξ with $\|\xi\| < 1$ such that

$$0 = \begin{bmatrix} (\tilde{B} - mI)\xi + v \\ (w + Qu)^T \xi + c - m \end{bmatrix}.$$

Since m approaches ∞ as c increases, we know that m will exceed the spectral radius of \tilde{B} for all sufficiently large c . Thus, from the first equation, we get

$$\xi = -(\tilde{B} - mI)^{-1}v.$$

Replacing ξ in the second equation,

$$\begin{aligned} 0 &= (w + Qu)^T \xi + c - m = c - m - (w + Qu)^T (\tilde{B} - mI)^{-1}v \\ &= c - m - \frac{(w + Qu)^T (\text{adj}(\tilde{B}) - mI)v}{m^2 - m \text{tr}(\tilde{B}) + \det(\tilde{B})}. \end{aligned}$$

Then, taking into account that $(c - m)m = -\|u\|^2$, defining $x = w + Qu$, we have

$$\begin{aligned} 0 &= (c - m)[m^2 - m \text{tr}(\tilde{B}) + \det(\tilde{B})] - x^T (\text{adj}(\tilde{B}) - mI)v \\ &= -\|u\|^2 m + \|u\|^2 \text{tr}(\tilde{B}) + (c - m) \det(\tilde{B}) - x^T \text{adj}(\tilde{B})v + mx^T v \\ &= [\|u\|^2 \text{tr}(\tilde{B}) - x^T \text{adj}(\tilde{B})v] + [x^T v - \|u\|^2 - \det(\tilde{B})]m + c \det(\tilde{B}) \\ &= [\|u\|^2 \text{tr}(\tilde{B}) - x^T \text{adj}(\tilde{B})v] + [x^T v - \|u\|^2 - \det(\tilde{B})] \left(m - \frac{c}{2}\right) \\ &\quad + [x^T v - \|u\|^2 + \det(\tilde{B})] \frac{c}{2}. \end{aligned}$$

Therefore,

$$\begin{aligned} &\left([x^T v - \|u\|^2 - \det(\tilde{B})] \left(m - \frac{c}{2}\right)\right)^2 \\ &= \left([\|u\|^2 \text{tr}(\tilde{B}) - x^T \text{adj}(\tilde{B})v] + [x^T v - \|u\|^2 + \det(\tilde{B})] \frac{c}{2}\right)^2 \end{aligned}$$

or equivalently,

$$\begin{aligned} &\left(x^T v - \|u\|^2 - \det(\tilde{B})\right)^2 \frac{c^2 + 4\|u\|^2}{4} \\ &= \left([\|u\|^2 \text{tr}(\tilde{B}) - x^T \text{adj}(\tilde{B})v] + [x^T v - \|u\|^2 + \det(\tilde{B})] \frac{c}{2}\right)^2. \end{aligned}$$

Grouping terms to rewrite this expression as a polynomial in c and using the identity $a^2 - b^2 = (a + b)(a - b)$ for all $a, b \in \mathbb{R}$, we get

$$\begin{aligned} 0 &= (x^T v - \|u\|^2)(-\det(\tilde{B}))c^2 - [\|u\|^2 \text{tr}(\tilde{B}) - x^T \text{adj}(\tilde{B})v][x^T v - \|u\|^2 + \det(\tilde{B})]c \\ &\quad + [x^T v - \|u\|^2 + \det(\tilde{B})]\|u\|^2 - [\|u\|^2 \text{tr}(\tilde{B}) - x^T \text{adj}(\tilde{B})v]^2. \end{aligned}$$

Since c can take arbitrarily large positive values, we deduce

$$\det(\tilde{B}) = 0, \quad x^T v = \|u\|^2, \quad \text{and} \quad x^T \text{adj}(\tilde{B})v = \|u\|^2 \text{tr}(\tilde{B}).$$

Recall that we showed above that $w^T v = w^T \text{adj}(\tilde{B})v = 0$. Since $x = w + Qu$, we get

$$\|u\|^2 = x^T v = u^T Q^T v \quad \text{and} \quad \|u\|^2 \text{tr}(\tilde{B}) = x^T \text{adj}(\tilde{B})v = u^T Q^T \text{adj}(\tilde{B})v.$$

Recall also that $z^T Q^T v = z^T Q^T \text{adj}(\tilde{B})v = 0$ for any z orthogonal to u . This gives us the relation

$$\begin{bmatrix} u^T Q^T \\ z^T Q^T \end{bmatrix} [v \text{adj}(\tilde{B})v] = \|u\|^2 \begin{bmatrix} 1 & \text{tr} \tilde{B} \\ 0 & 0 \end{bmatrix},$$

where the matrix on the left is nonsingular for $z \neq 0$. Hence, for each z , the equation has a unique solution $[v \text{adj}(\tilde{B})v]$. By inspection, we note that $v = Qu$ and $\text{adj}(\tilde{B})v = \text{tr}(\tilde{B})Qu$ satisfy it, so these must be the true vectors. We then obtain $\text{adj}(\tilde{B})Qu = \text{tr}(\tilde{B})Qu$, and the result follows. \square

Now we start analyzing the behavior of the matrices in S_2 under the linear preservers of the L-spectrum. As a byproduct of this work, we obtain further information about the images of matrices in S_3 under such linear preservers.

Lemma 5.4. *Let $\phi : M_3 \rightarrow M_3$ be a linear preserver of the L-spectrum with associated orthogonal matrix Q , and let $\tilde{A} \in M_2$ and $u \in \mathbb{R}^2$. Then for some $\tilde{C}, \tilde{B} \in M_2$,*

$$\phi\left(\begin{bmatrix} \tilde{A} & 0 \\ 0 & 0 \end{bmatrix}\right) = \begin{bmatrix} \tilde{C} & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \phi\left(\begin{bmatrix} 0 & u \\ 0 & 0 \end{bmatrix}\right) = \begin{bmatrix} \tilde{B} & Qu \\ 0 & 0 \end{bmatrix}.$$

Proof. Let

$$A = \begin{bmatrix} \tilde{A} & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \phi(A) = \begin{bmatrix} \tilde{C} & p \\ q^T & c \end{bmatrix}.$$

Let $a > 0$, $z \in \mathbb{R}^2$ be any nonzero vector, and

$$H := \begin{bmatrix} 0 & 0 \\ z^T & a \end{bmatrix}.$$

By [Theorem 4.5](#), we have

$$\phi(H) = \begin{bmatrix} 0 & 0 \\ (Qz)^T & a \end{bmatrix}.$$

Thus,

$$\phi(A + H) = \phi(A) + \phi(H) = \begin{bmatrix} \tilde{C} & p \\ (q + Qz)^T & c + a \end{bmatrix}.$$

Notice that a is an (interior) L-eigenvalue of $A + H$ with L-eigenvector $[0, 1]^T$. Since ϕ preserves the L-spectrum, $a \in \sigma_L(\phi(A + H))$. By [Lemma 5.2](#), we have $c = 0$ and $(q + Qz)^T p = 0$. Since Q is invertible and $z \neq 0$ is arbitrary, we have $(q + w)^T p = 0$ for all $w \neq 0$, which implies $p = 0$. Thus,

$$\phi\left(\begin{bmatrix} \tilde{A} & 0 \\ 0 & 0 \end{bmatrix}\right) = \begin{bmatrix} \tilde{C} & 0 \\ q^T & 0 \end{bmatrix}.$$

Now we show that $q = 0$. For any arbitrary nonzero vector $u \in \mathbb{R}^2$, let $a \in \mathbb{R}$ be large enough so that, by [Lemma A.4](#), a is an interior L-eigenvalue of

$$G := \begin{bmatrix} \tilde{A} & u \\ 0 & a \end{bmatrix}.$$

By the linearity of ϕ and by Lemmas 5.3 and 4.4, we have

$$\phi(G) = \begin{bmatrix} \tilde{B} + \tilde{C} & Qu \\ (w+q)^T & a \end{bmatrix}, \quad \text{where } w^T Qu = 0 \text{ and } \text{adj}(\tilde{B})Qu = \text{tr}(\tilde{B})Qu.$$

We know that for large enough a , we have $a \in \sigma_L(G)$ and hence $a \in \sigma_L(\phi(G))$. By Lemma 5.2, a is an interior L-eigenvalue of $\phi(G)$ for large enough a , $(w+q)^T Qu = 0$, and $(w+q)^T \text{adj}(\tilde{B} + \tilde{C})Qu = 0$. Since $w^T Qu = 0$, we deduce $q^T Qu = 0$. Since u is arbitrary and independent of q , we deduce that $q = 0$. Thus, the claim for matrices in S_2 follows.

Now observe that

$$\begin{aligned} 0 &= w^T \text{adj}(\tilde{B} + \tilde{C})Qu = w^T \text{adj}(\tilde{B})Qu + w^T \text{adj}(\tilde{C})Qu \\ &= \text{tr}(\tilde{B})w^T Qu + w^T \text{adj}(\tilde{C})Qu = w^T \text{adj}(\tilde{C})Qu. \end{aligned}$$

Note that we have shown that matrices $\tilde{A} \oplus [0]$ map to $\tilde{C} \oplus [0]$. Since ϕ is bijective by Theorem 1.1, any matrix $\tilde{C} \oplus [0]$ must have a preimage $\tilde{A} \oplus [0]$. In particular, we may take $\tilde{C} = \text{adj}(R)$, where $R = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$. Hence, $w^T RQu = 0$. Since $w^T Qu = 0$ and since the vectors Qu and RQu are linearly independent, we know furthermore that $w = 0$ for all nonzero u , which proves the claim for matrices in S_3 . \square

Next we show that the matrix \tilde{C} in Lemma 5.4 is closely related to \tilde{A} .

Lemma 5.5. *Let $\phi : M_3 \rightarrow M_3$ be a linear preserver of the L-spectrum with associated orthogonal matrix Q . Then there exists an invertible diagonal matrix D such that either*

$$\phi\left(\begin{bmatrix} \tilde{A} & 0 \\ 0 & 0 \end{bmatrix}\right) = \begin{bmatrix} QD\tilde{A}D^{-1}Q^T & 0 \\ 0 & 0 \end{bmatrix} \quad \text{for all } \tilde{A} \in M_2 \quad (5-2)$$

or

$$\phi\left(\begin{bmatrix} \tilde{A} & 0 \\ 0 & 0 \end{bmatrix}\right) = \begin{bmatrix} QD\tilde{A}^T D^{-1}Q^T & 0 \\ 0 & 0 \end{bmatrix} \quad \text{for all } \tilde{A} \in M_2. \quad (5-3)$$

Proof. Let $W_3 := \{\tilde{A} \oplus [a] : \tilde{A} \in M_2, a \in \mathbb{R}\}$. By Lemmas 4.4 and 5.4, the linear map $\tilde{\phi} : W_3 \rightarrow W_3$ given by $\tilde{\phi}(A) = \phi(A)$ preserves the Lorentz spectrum on W_3 . Thus, by Theorem 4.2 in [Bueno et al. 2021] with $\mathcal{M} = W_3$, there exists some invertible matrix $P \in M_2$ such that either

$$\phi\left(\begin{bmatrix} \tilde{A} & 0 \\ 0 & 0 \end{bmatrix}\right) = \tilde{\phi}\left(\begin{bmatrix} \tilde{A} & 0 \\ 0 & 0 \end{bmatrix}\right) = \begin{bmatrix} P\tilde{A}P^{-1} & 0 \\ 0 & 0 \end{bmatrix} \quad \text{for all } \tilde{A} \in M_2 \quad (5-4)$$

or

$$\phi\left(\begin{bmatrix} \tilde{A} & 0 \\ 0 & 0 \end{bmatrix}\right) = \tilde{\phi}\left(\begin{bmatrix} \tilde{A} & 0 \\ 0 & 0 \end{bmatrix}\right) = \begin{bmatrix} P\tilde{A}^T P^{-1} & 0 \\ 0 & 0 \end{bmatrix} \quad \text{for all } \tilde{A} \in M_2. \quad (5-5)$$

We will now show that $P = QD$ for some invertible diagonal matrix D . Consider matrices of the form

$$B = \begin{bmatrix} D & 0 \\ v^T & a \end{bmatrix} = \begin{bmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ v_1 & v_2 & a \end{bmatrix} \in M_3, \quad (5-6)$$

where $d_1 \neq d_2$, and $v_1 + a - d_1 \geq 0$. Then $\lambda = (v_1 + a + d_1)/2 \in \sigma_{\text{bd}}(B)$, with $\mu = d_1$, $s = (v_1 + a - d_1)/2$, and $\xi = [1, 0]^T$. Let us additionally assume that $\lambda \notin \sigma(B) = \{a, d_1, d_2\}$. Then $\lambda \notin \sigma(\phi(B))$, and by (5-4) and (5-5),

$$\phi(B) = \begin{bmatrix} PDP^{-1} & 0 \\ (Qv)^T & a \end{bmatrix}.$$

As $\sigma_L(B) = \sigma_L(\phi(B))$, we deduce that $\lambda \in \sigma_{\text{bd}}(\phi(B))$. Thus, $\lambda = \mu + s$, with $s \geq 0$, and there is some ξ such that $\|\xi\| = 1$ and

$$\begin{aligned} (PDP^{-1} - \mu I_2)\xi &= 0, \\ (Qv)^T \xi + a - \mu - 2s &= 0. \end{aligned}$$

From the first equation, we get

$$(D - \mu I_2)P^{-1}\xi = 0.$$

Let $x = P^{-1}\xi$, that is, $\xi = Px = x_1 p_1 + x_2 p_2$, where p_i denotes the i -th column of P . We consider two cases:

Case I: $\mu = d_1$. Then since $d_1 \neq d_2$, we have $x_2 = 0$ and $\xi = x_1 p_1$. As $\|\xi\| = 1$, we have $x_1 = \pm 1/\|p_1\|$. Thus,

$$\begin{aligned} 0 &= (Qv)^T \xi + a - \mu - 2s = (Qv)^T x_1 p_1 + a - \mu - 2(\lambda - \mu) \\ &= (Qv)^T x_1 p_1 + a + d_1 - (v_1 + a + d_1) = v^T (x_1 Q^T p_1) - v_1. \end{aligned}$$

Hence,

$$v^T (x_1 Q^T p_1) = v_1 = v^T \begin{bmatrix} 1 \\ 0 \end{bmatrix} = v^T e_1.$$

Since this equality holds for any v with $v_1 + a - d_1 \geq 0$ and $(v_1 + a + d_1)/2 \notin \{a, d_1, d_2\}$, we deduce that $x_1 Q^T p_1 = e_1$, or equivalently,

$$p_1 = \frac{1}{x_1} Q e_1 = \pm \|p_1\| q_1,$$

where q_1 denotes the first column of Q . By multiplying P and P^{-1} by ± 1 , we may assume without loss of generality that $p_1 = \|p_1\| q_1$.

Case II: $\mu \neq d_1$. Then since $\xi \neq 0$, we have $x_1 = 0$ and $\mu = d_2$. Thus,

$$0 = (Qv)^T \xi + a - \mu - 2(\lambda - \mu) = (Qv)^T \xi + a + d_2 - (v_1 + a + d_1) \leq \|v\| + d_2 - d_1 - v_1.$$

However, by taking $d_2 > d_1 + v_1 - \|v\|$, this condition fails, which means we can discard this case.

If we consider now the family of matrices of the form (5-6) that satisfy $v_2 + a - d_2 \geq 0$, we will have $\lambda = (v_2 + a + d_2)/2 \in \sigma_{\text{bd}}(B)$. Assuming additionally that $(v_2 + a + d_2)/2 \notin \{a, d_1, d_2\}$, that is, $\lambda \notin \sigma(B)$, a similar argument to the one above yields $p_2 = \pm \|p_2\|q_2$, so we may take the diagonal matrix D in the statement of the lemma to be $\text{diag}(\|p_1\|, \pm \|p_2\|)$, which is invertible since P is. \square

In the next lemma we narrow down the possible matrices D for which (5-2) holds and show that (5-3) cannot occur.

Lemma 5.6. *Let $\phi : M_3 \rightarrow M_3$ be a linear preserver of the L-spectrum with associated orthogonal matrix Q . Then*

$$\phi\left(\begin{bmatrix} \tilde{A} & 0 \\ 0 & 0 \end{bmatrix}\right) = \begin{bmatrix} Q & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} D\tilde{A}D^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Q^T & 0 \\ 0 & 1 \end{bmatrix} \quad \text{for all } \tilde{A} \in M_2, \quad (5-7)$$

where $D = I_2$ or $D = \text{diag}(1, -1)$.

Proof. Step 1: Let

$$A = \begin{bmatrix} 0 & 0 & 0 \\ m & 0 & 0 \\ v_1 & v_2 & a \end{bmatrix},$$

with $m \neq 0$, $a > |v_2|$, $v_1 \neq \pm v_2$, and $v_2 \notin \{0, \pm a\}$. By Lemma A.5, $\sigma_{\text{bd}}(A) = \{(\pm v_2 + a)/2\}$ and, hence $\{(\pm v_2 + a)/2\} \subseteq \sigma_L(\phi(A))$. By Lemma 5.5, $\phi(A)$ is as in (5-2) or (5-3). Assume that $\phi(A)$ is as in (5-3). Let $D = \text{diag}(d_1, d_2)$. Then,

$$\phi(A) = \begin{bmatrix} Q & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & md_1/d_2 & 0 \\ 0 & 0 & 0 \\ v_1 & v_2 & a \end{bmatrix} \begin{bmatrix} Q^T & 0 \\ 0 & 1 \end{bmatrix}$$

and $\{(\pm v_2 + a)/2\} \subseteq \sigma_{\text{bd}}(\phi(A))$. (Note that $\sigma(\phi(A)) = \{0, a\}$ and $(\pm v_2 + a)/2 \notin \{0, a\}$.) By Lemma A.5, the only potential boundary L-eigenvalues of $\phi(A)$ are $(\pm v_1 + a)/2$. However, $v_1 \neq \pm v_2$, so this is a contradiction, which implies $\phi(A)$ is as in (5-2).

Step 2: Consider matrices of the form

$$B = \left[\begin{array}{cc|c} 0 & c & 0 \\ c & 0 & 0 \\ \hline v_1 & v_2 & a \end{array} \right], \quad c > 0.$$

Then, by the conclusion of Step 1 and by Theorem 1.2,

$$\sigma_L(B) = \sigma_L(\phi(B)) = \sigma_L\left(\left[\begin{array}{cc|c} 0 & cd_1/d_2 & 0 \\ cd_2/d_1 & 0 & 0 \\ \hline v_1 & v_2 & a \end{array} \right]\right) =: \sigma_L(H).$$

Consider the family of matrices B having the boundary L-eigenvalue

$$\lambda = \frac{(v_1 + v_2)/\sqrt{2} + a + c}{2},$$

which, by Lemma A.5, happens if

$$\frac{v_1 + v_2}{\sqrt{2}} + a - c \geq 0.$$

Since $\sigma(H) = \{a, c, -c\}$, consider those matrices B for which $\lambda \notin \{a, c, -c\}$ so that $\lambda \in \sigma_{\text{bd}}(H)$. Hence, by Lemma A.5,

$$2\lambda \in \left\{ \begin{array}{l} \pm(|d_1|v_1 + |d_2|v_2)/\sqrt{d_1^2 + d_2^2} + a + |c|, \\ \pm(|d_1|v_1 - |d_2|v_2)/\sqrt{d_1^2 + d_2^2} + a - |c|. \end{array} \right.$$

This implies $d_1 = \pm d_2$, and after multiplying D by $1/d_1$ and D^{-1} by d_1 , which does not change $D\tilde{A}D^{-1}$, we get $D = I_2$ or $D = \text{diag}(1, -1)$. \square

Now we present the main result of this section, which provides a complete description of the images of matrices in \mathcal{S}_2 under linear preservers of the L-spectrum.

Theorem 5.7. *Let $\phi : M_3 \rightarrow M_3$ be a linear preserver of the L-spectrum with associated orthogonal matrix Q . Then*

$$\phi\left(\begin{bmatrix} \tilde{A} & 0 \\ 0 & 0 \end{bmatrix}\right) = \begin{bmatrix} Q\tilde{A}Q^T & 0 \\ 0 & 0 \end{bmatrix} \quad \text{for all } \tilde{A} \in M_2.$$

Proof. Let $A = \begin{bmatrix} \tilde{A} & 0 \\ 0 & 0 \end{bmatrix}$. We know that $\phi(A)$ is as in (5-7). Suppose

$$\phi\left(\begin{bmatrix} \tilde{A} & 0 \\ 0 & 0 \end{bmatrix}\right) = \begin{bmatrix} QD\tilde{A}DQ^T & 0 \\ 0 & 0 \end{bmatrix}$$

for all \tilde{A} , where $D = \text{diag}(1, -1)$. By Theorem 1.2, we know that ϕ preserves the L-spectrum if and only if the map

$$\psi(A) = \begin{bmatrix} Q^T & 0 \\ 0 & 1 \end{bmatrix} \phi(A) \begin{bmatrix} Q & 0 \\ 0 & 1 \end{bmatrix}$$

preserves the Lorentz spectrum, so we may suppose without loss of generality that $Q = I_2$.

Let

$$A = \begin{bmatrix} \tilde{A} & 0 \\ v^T & a \end{bmatrix},$$

where $v = [v_1, v_2]^T \in \mathbb{R}^2$ is such that $v_1, v_2 \neq 0$ and $\|v\| = 1$. Moreover, let $\tilde{A} = P\tilde{D}P^T$, where

$$P = \begin{bmatrix} v_1 & v_2 \\ -v_2 & v_1 \end{bmatrix}$$

is an orthogonal matrix and $\tilde{D} = \text{diag}(\lambda, v_1 v_2 + a)$, with $\lambda > v_1 v_2 + a$. Note that $v_1 v_2 + a \in \sigma_{\text{int}}(A)$ with corresponding L-eigenvector $[v_2/2, v_1/2, 1]^T$. Thus, $v_1 v_2 + a \in \sigma_L(\phi(A))$, where

$$\phi(A) = \begin{bmatrix} DP\tilde{D}P^T D & 0 \\ v^T & a \end{bmatrix}.$$

Step 1: Suppose $v_1 v_2 + a$ is a standard L-eigenvalue of $\phi(A)$. Then there is some ξ , with $\|\xi\| \leq 1$, such that

$$0 = [H - (v_1 v_2 + a)I_3] \begin{bmatrix} \xi \\ 1 \end{bmatrix} = \begin{bmatrix} DP\tilde{D}P^T D\xi - (v_1 v_2 + a)\xi \\ v^T \xi - v_1 v_2 \end{bmatrix},$$

or equivalently,

$$0 = (P\tilde{D}P^T)D\xi - (v_1 v_2 + a)D\xi, \quad (5-8)$$

$$0 = v^T \xi - v_1 v_2.$$

Since $v_1 v_2 \neq 0$, the second equation implies that $\xi \neq 0$ and

$$\xi_1 = v_2 - \frac{v_2}{v_1} \xi_2.$$

As D is invertible, $D\xi \neq 0$ and, from (5-8), we deduce that $D\xi$ is an L-eigenvector of $P\tilde{D}P^T$ associated with $v_1 v_2 + a$. This implies that $D\xi$ must be proportional to the second column of P . However, this gives a contradiction as

$$0 = \det[D\xi, p_2] = \det \begin{bmatrix} v_2(1 - \xi_2/v_1) & v_2 \\ -\xi_2 & v_1 \end{bmatrix} = v_1 v_2 \neq 0.$$

Therefore, $v_1 v_2 + a$ is not a standard L-eigenvalue of $\phi(A)$.

Step 2: Since $v_1 v_2 + a$ must be a nonstandard L-eigenvalue of $\phi(A)$, there exist $s > 0$ and ξ with $\|\xi\| = 1$ such that

$$0 = [H - (v_1 v_2 + a)I_3] \begin{bmatrix} \xi \\ 1 \end{bmatrix} + s \begin{bmatrix} \xi \\ -1 \end{bmatrix} = \begin{bmatrix} DP\tilde{D}P^T D\xi - (v_1 v_2 + a - s)\xi \\ v^T \xi - v_1 v_2 - s \end{bmatrix}.$$

Since $\xi \neq 0$, from the first equation we see that ξ is an eigenvector of $DP\tilde{D}P^T D$ corresponding to the eigenvalue $v_1 v_2 + a - s$. Because DP is orthogonal, the only eigenvalues of $DP\tilde{D}P^T D$ are λ and $v_1 v_2 + a$, which implies either $s = v_1 v_2 + a - \lambda < 0$ or $s = 0$, a contradiction.

Thus, $v_1 v_2 + a$ cannot be an L-eigenvalue of $\phi(A)$, and as a result, ϕ does not preserve the Lorentz spectrum. Hence, the claim follows. \square

5.3. Image of matrices in \mathcal{S}_3 under a linear preserver. We finish the proof of [Theorem 3.1](#) by analyzing the behavior of linear preservers of the L-spectrum on the subspace \mathcal{S}_3 . We already obtained some partial results in [Lemmas 5.3](#) and [5.4](#).

We begin by presenting an auxiliary lemma that will be used in proving some of the results in this section.

Lemma 5.8. *Let $\phi : M_3 \rightarrow M_3$ be a linear preserver of the L-spectrum with associated orthogonal matrix Q . Let*

$$\phi\left(\begin{bmatrix} 0 & u \\ v^T & 0 \end{bmatrix}\right) = \begin{bmatrix} \tilde{B} & Qu \\ (Qv)^T & 0 \end{bmatrix}.$$

Then, for any integer n , we have

$$\sigma_L\left(\begin{bmatrix} n\tilde{B} & Qu \\ (Qv)^T & 0 \end{bmatrix}\right) = \sigma_L\left(\begin{bmatrix} 0 & u \\ v^T & 0 \end{bmatrix}\right). \quad (5-9)$$

Proof. First we show that, for all integers n ,

$$\sigma_L\left(\begin{bmatrix} n\tilde{B} & Qu \\ (Qv)^T & 0 \end{bmatrix}\right) = \sigma_L\left(\begin{bmatrix} (n+1)\tilde{B} & Qu \\ (Qv)^T & 0 \end{bmatrix}\right). \quad (5-10)$$

Note that

$$\begin{bmatrix} n\tilde{B} & Qu \\ (Qv)^T & 0 \end{bmatrix} = \begin{bmatrix} Q & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} nQ^T\tilde{B}Q & u \\ v^T & 0 \end{bmatrix} \begin{bmatrix} Q^T & 0 \\ 0 & 1 \end{bmatrix},$$

so, by [Theorem 1.2](#),

$$\sigma_L\left(\begin{bmatrix} n\tilde{B} & Qu \\ (Qv)^T & 0 \end{bmatrix}\right) = \sigma_L\left(\begin{bmatrix} nQ^T\tilde{B}Q & u \\ v^T & 0 \end{bmatrix}\right).$$

Now observe that

$$\phi\left(\begin{bmatrix} nQ^T\tilde{B}Q & u \\ v^T & 0 \end{bmatrix}\right) = \begin{bmatrix} \tilde{B} + nQQ^T\tilde{B}QQ^T & Qu \\ (Qv)^T & 0 \end{bmatrix} = \begin{bmatrix} (n+1)\tilde{B} & Qu \\ (Qv)^T & 0 \end{bmatrix}.$$

Since ϕ preserves the L-spectrum, we get

$$\sigma_L\left(\begin{bmatrix} n\tilde{B} & Qu \\ (Qv)^T & 0 \end{bmatrix}\right) = \sigma_L\left(\begin{bmatrix} nQ^T\tilde{B}Q & u \\ v^T & 0 \end{bmatrix}\right) = \sigma_L\left(\begin{bmatrix} (n+1)\tilde{B} & Qu \\ (Qv)^T & 0 \end{bmatrix}\right),$$

which shows [\(5-10\)](#).

We now prove the claim by induction on n . Since ϕ preserves the L-spectrum, we know that

$$\sigma_L\left(\begin{bmatrix} 0 & u \\ v^T & 0 \end{bmatrix}\right) = \sigma_L\left(\begin{bmatrix} \tilde{B} & Qu \\ (Qv)^T & 0 \end{bmatrix}\right),$$

so [\(5-9\)](#) holds for $n = 1$. Now assume that [\(5-9\)](#) holds for some integer n . Then by [\(5-10\)](#), it holds for both $n + 1$ and $n - 1$, so the claim follows for all integers. \square

The next two lemmas analyze the image of a basis for \mathcal{S}_3 under the linear preservers. Note that the assumption on the form of the images of the two matrices in the basis is a consequence of [Lemma 5.4](#).

Lemma 5.9. *Let $\phi : M_3 \rightarrow M_3$ be a linear preserver of the L -spectrum with associated orthogonal matrix Q , and let*

$$\phi\left(\begin{bmatrix} 0 & Q^T e_1 \\ 0 & 0 \end{bmatrix}\right) = \begin{bmatrix} \tilde{B}_1 & e_1 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \phi\left(\begin{bmatrix} 0 & Q^T e_2 \\ 0 & 0 \end{bmatrix}\right) = \begin{bmatrix} \tilde{B}_2 & e_2 \\ 0 & 0 \end{bmatrix}.$$

Then there exist $x, y \in \mathbb{R}$ such that

$$\tilde{B}_1 = \begin{bmatrix} 0 & x \\ 0 & y \end{bmatrix} \quad \text{and} \quad \tilde{B}_2 = \begin{bmatrix} -x & 0 \\ -y & 0 \end{bmatrix}.$$

Proof. Let $u = [u_1, u_2]^T \in \mathbb{R}^2$ be a nonzero vector, and let

$$A = \begin{bmatrix} 0 & Q^T u \\ (Q^T u)^T & 0 \end{bmatrix}.$$

Then by [Theorem 4.5](#) and by the linearity of ϕ ,

$$\phi(A) = \begin{bmatrix} u_1 \tilde{B}_1 + u_2 \tilde{B}_2 & u \\ u^T & 0 \end{bmatrix}.$$

By [Lemma A.2](#), $\|u\| \in \sigma_{\text{bd}}(A) \subseteq \sigma_L(\phi(A))$. Suppose that $\|u\| \in \sigma_{\text{int}}(\phi(A))$. Then there exists ξ with $\|\xi\| < 1$ such that

$$0 = \begin{bmatrix} u_1 \tilde{B}_1 + u_2 \tilde{B}_2 - \|u\| I_2 & u \\ u^T & -\|u\| \end{bmatrix} \begin{bmatrix} \xi \\ 1 \end{bmatrix}.$$

Then from the second equation, we get

$$\|u\| = u^T \xi \leq \|u\| \|\xi\| < \|u\|,$$

a contradiction. Therefore, $\|u\| \in \sigma_{\text{bd}}(\phi(A))$. This means that there exist $s \geq 0$, μ , and ξ such that $\|u\| = \mu + s$, $\|\xi\| = 1$, and

$$0 = \begin{bmatrix} u_1 \tilde{B}_1 + u_2 \tilde{B}_2 - \mu I_2 & u \\ u^T & -\|u\| - s \end{bmatrix} \begin{bmatrix} \xi \\ 1 \end{bmatrix},$$

from which it follows that $s = u^T \xi - \|u\| \leq 0$. This implies $s = 0$ and hence $\mu = \|u\|$.

The only solution ξ of the equation $u^T \xi - \|u\| = 0$ on the unit circle is $\xi = u/\|u\|$. Thus, we have

$$0 = (u_1 \tilde{B}_1 + u_2 \tilde{B}_2 - \|u\| I_2) \xi + u = (u_1 \tilde{B}_1 + u_2 \tilde{B}_2) \frac{u}{\|u\|} \quad (5-11)$$

for all $u \neq 0$. Choosing $u_1 \neq u_2 = 0$ gives $\tilde{B}_1 e_1 = 0$, and choosing $u_2 \neq u_1 = 0$ gives $\tilde{B}_2 e_2 = 0$. Plugging these results in [\(5-11\)](#) gives

$$0 = u_1 u_2 (\tilde{B}_1 e_2 + \tilde{B}_2 e_1).$$

Since this must hold when $u_1, u_2 \neq 0$, we deduce $\tilde{B}_1 e_2 = -\tilde{B}_2 e_1$, and the result follows. \square

Lemma 5.10. *Let*

$$\tilde{B}_1 = \begin{bmatrix} 0 & x \\ 0 & y \end{bmatrix} \quad \text{and} \quad \tilde{B}_2 = \begin{bmatrix} -x & 0 \\ -y & 0 \end{bmatrix}$$

be as in [Lemma 5.9](#). Then $x = 0$ if and only if $y = 0$.

Proof. Suppose that $y = 0$ and $x \neq 0$, and let

$$H_n := \begin{bmatrix} n\tilde{B}_1 & e_1 \\ e_1^T & 0 \end{bmatrix}$$

for any integer n . Since

$$\phi\left(\begin{bmatrix} 0 & Q^T e_1 \\ (Q^T e_1)^T & 0 \end{bmatrix}\right) = \begin{bmatrix} \tilde{B}_1 & e_1 \\ e_1^T & 0 \end{bmatrix} = H_1,$$

[Lemma 5.8](#) implies

$$\sigma_L(H_n) = \sigma_L(H_1) = \sigma_L\left(\begin{bmatrix} 0 & Q^T e_1 \\ (Q^T e_1)^T & 0 \end{bmatrix}\right) = \sigma_L\left(\begin{bmatrix} 0 & e_1 \\ e_1^T & 0 \end{bmatrix}\right) = \{\pm 1\} \quad (5-12)$$

for all integers n , where the last equality follows from [Lemma A.2](#). Next we observe that $ny = 0$ is a standard eigenvalue of H_n with associated eigenvector

$$\xi = \begin{bmatrix} 0 & -\frac{1}{nx} & 1 \end{bmatrix}^T.$$

For n sufficiently large, we get $|1/(nx)| \leq 1$, which implies that ξ lies in the Lorentz cone and hence $0 \in \sigma_L(H_n)$. However, this is a contradiction by [\(5-12\)](#). Thus, if $y = 0$, then $x = 0$.

By applying a similar argument to

$$\begin{bmatrix} n\tilde{B}_2 & e_2 \\ e_2^T & 0 \end{bmatrix},$$

we may likewise conclude that if $x = 0$, then $y = 0$. □

We now arrive at the main result in this section and the final piece for the proof of [Theorem 3.1](#).

Theorem 5.11. *Let $\phi : M_3 \rightarrow M_3$ be a linear preserver of the L -spectrum with associated orthogonal matrix Q . Then, for all $u \in \mathbb{R}^2$,*

$$\phi\left(\begin{bmatrix} 0 & u \\ 0 & 0 \end{bmatrix}\right) = \begin{bmatrix} 0 & Qu \\ 0 & 0 \end{bmatrix}.$$

Proof. Note that it is enough to prove

$$\phi\left(\begin{bmatrix} 0 & Q^T e_1 \\ 0 & 0 \end{bmatrix}\right) = \begin{bmatrix} 0 & e_1 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \phi\left(\begin{bmatrix} 0 & Q^T e_2 \\ 0 & 0 \end{bmatrix}\right) = \begin{bmatrix} 0 & e_2 \\ 0 & 0 \end{bmatrix}$$

since $\{Q^T e_1, Q^T e_2\}$ is a basis for \mathbb{R}^2 . That is, we want to show that $\tilde{B}_1 = \tilde{B}_2 = 0$, where the matrices

$$\tilde{B}_1 = \begin{bmatrix} 0 & x \\ 0 & y \end{bmatrix} \quad \text{and} \quad \tilde{B}_2 = \begin{bmatrix} -x & 0 \\ -y & 0 \end{bmatrix}$$

are as in [Lemma 5.9](#).

Suppose for sake of contradiction that $x \neq 0$, which implies $y \neq 0$ by [Lemma 5.10](#). Let

$$H_n := \begin{bmatrix} n\tilde{B}_1 & e_1 \\ e_1^T & 0 \end{bmatrix}$$

for any integer n . By (5-12), we have $\sigma_L(H_n) = \{\pm 1\}$ for all n .

Let $n \neq 0$, $\mu = ny$, and $\lambda = \mu + s$ for some real number s . Then $\lambda \in \sigma_{\text{bd}}(H_n)$ if and only if there exists some $\xi = [\xi_1, \xi_2]^T \in \mathbb{R}^2$ such that

$$\begin{bmatrix} -ny & nx & 1 \\ 0 & 0 & 0 \\ 1 & 0 & -ny - 2s \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ 1 \end{bmatrix} = 0, \quad \|\xi\| = 1, \quad \text{and} \quad s \geq 0, \quad (5-13)$$

or equivalently,

$$-ny\xi_1 + nx\xi_2 + 1 = 0, \quad (5-14)$$

$$\xi_1 - ny - 2s = 0. \quad (5-15)$$

We now show that there exists some real number N such that either N is positive and it is possible to satisfy (5-13) for any integer $n > N$, or N is negative and it is possible to satisfy (5-13) for any integer $n < N$. Let

$$N := \begin{cases} \min(-1/\sqrt{x^2 + y^2}, -3/y) & \text{if } y > 0, \\ \max(1/\sqrt{x^2 + y^2}, -3/y) & \text{if } y < 0. \end{cases}$$

For each $n \neq 0$, (5-14) has a solution $\xi^{(n)}$ with $\|\xi^{(n)}\| = 1$ if and only if $|n| \geq 1/\sqrt{x^2 + y^2}$, which holds whenever $|n| > |N|$.

Note that (5-15) is equivalent to

$$s = \frac{\xi_1 - ny}{2}.$$

For each $\xi^{(n)}$ with $\|\xi^{(n)}\| = 1$ satisfying (5-14) and for each n in the range specified above, we know that

$$\xi_1^{(n)} \geq -1 > -3 \geq Ny > ny,$$

which guarantees that

$$s_n := \frac{\xi_1^{(n)} - ny}{2} > 0.$$

To conclude the argument, choose n in the range specified above so that $\xi^{(n)}$ and s_n satisfy (5-13). Then $\lambda = \mu + s_n \in \sigma_{\text{bd}}(H_n) = \{\pm 1\}$. However, this gives a

contradiction since

$$\lambda = \frac{\xi_1^{(n)} + ny}{2} < \frac{\xi_1^{(n)} - Ny}{2} \leq \frac{1-3}{2} = -1.$$

It follows that $x = 0$ and hence $y = 0$ by [Lemma 5.10](#). Therefore, $\tilde{B}_1 = \tilde{B}_2 = 0$. \square

6. Conclusions

In this paper, we have analyzed the linear preservers of the Lorentz spectrum of 3×3 real matrices and proven that every such linear map ϕ must be of the form

$$\phi(A) = \begin{bmatrix} Q & 0 \\ 0 & 1 \end{bmatrix} A \begin{bmatrix} Q^T & 0 \\ 0 & 1 \end{bmatrix}$$

for some orthogonal $Q \in M_2$, as conjectured in [\[Bueno et al. 2021\]](#). An immediate corollary of this result is that the linear preservers of the L-spectrum on M_3 must take interior (resp. boundary) L-eigenvalues to interior (resp. boundary) L-eigenvalues. Our proof relies on the particular form of 3×3 matrices with infinitely many L-eigenvalues, which makes it difficult to generalize this result to higher dimensions since $n \times n$ matrices with this property can be much more complicated. Thus, it is likely that different techniques will be necessary to prove the corresponding result for M_n , but we hope that some of the strategies we developed in this paper will still be applicable in this case.

Appendix: L-spectrum of some special matrices

Here we provide some results for $n \times n$ matrices, where $n \geq 3$, and one result specific to 3×3 matrices.

A1. Results for $n \times n$ matrices. The following result is an immediate consequence of Corollary 3.3 in [\[Bueno et al. 2021\]](#).

Lemma A.1. *Let*

$$A = \begin{bmatrix} cI_{n-1} & 0 \\ 0 & a \end{bmatrix} \in M_n.$$

Then

$$\sigma_L(A) = \begin{cases} \{a\} & \text{if } c > a, \\ \{a, (a+c)/2\} & \text{if } c \leq a. \end{cases}$$

The next result follows from Theorem 3.4 in [\[Bueno et al. 2021\]](#).

Lemma A.2. *Let*

$$A = \begin{bmatrix} 0 & u \\ v^T & a \end{bmatrix},$$

where $u, v \in \mathbb{R}^{n-1}$ are not both zero and $a \in \mathbb{R}$. Then:

- (1) $0 \in \sigma_{\text{int}}(A)$ (resp. $0 \in \sigma_{\text{bd}}(A)$) if and only if $u = 0$ and $|a| < \|v\|$ (resp. $|a| \leq \|v\|$).
- (2) If $\lambda \neq 0$, then $\lambda \in \sigma_{\text{int}}(A)$ (resp. $\lambda \in \sigma_{\text{bd}}(A)$) if and only if $|\lambda| > \|u\|$ (resp. $|\lambda| \geq \|u\|$) and $\lambda^2 - a\lambda - v^T u = 0$.
- (3) If $u \neq 0$, then λ is a nonstandard Lorentz eigenvalue of A if and only if one of the following holds:

- (i) $v^T u + a\|u\| - \|u\|^2 > 0$ and $\lambda = (a\|u\| + \|u\|^2 + v^T u)/(2\|u\|)$.
- (ii) $\|u\|^2 + a\|u\| - v^T u > 0$ and $\lambda = (a\|u\| - \|u\|^2 - v^T u)/(2\|u\|)$.

- (4) If $u = 0$ (and hence $v \neq 0$), then λ is a nonstandard Lorentz eigenvalue of A if and only if

$$\lambda \in \left[\frac{a - \|v\|}{2}, \frac{a + \|v\|}{2} \right] \cap (0, \infty).$$

Lemma A.3. Let

$$A = \begin{bmatrix} cI_{n-1} & 0 \\ v^T & a \end{bmatrix}, \quad \text{where } a, c \in \mathbb{R} \text{ and } 0 \neq v \in \mathbb{R}^{n-1}.$$

Then:

- (1) If $c < a - \|v\|$, then $\sigma_{\text{int}}(A) = \{a\}$ and

$$\sigma_{\text{bd}}(A) = \left[\frac{c + a - \|v\|}{2}, \frac{c + a + \|v\|}{2} \right].$$

Moreover, $\sigma_{\text{int}}(A) \cap \sigma_{\text{bd}}(A) = \emptyset$.

- (2) If $c = a - \|v\|$, then $\sigma_{\text{int}}(A) = \{a\} = \{\|v\| + c\}$, and

$$\sigma_{\text{bd}}(A) = [c, \|v\| + c].$$

Therefore, $a \in \sigma_{\text{int}}(A) \cap \sigma_{\text{bd}}(A)$.

- (3) If $a - \|v\| < c < a + \|v\|$, then $\sigma_{\text{int}}(A) = \{a, c\}$ and

$$\sigma_{\text{bd}}(A) = \left[c, \frac{c + a + \|v\|}{2} \right].$$

Therefore, $c \in \sigma_{\text{int}}(A) \cap \sigma_{\text{bd}}(A)$. Moreover, $a \in \sigma_{\text{int}}(A) \cap \sigma_{\text{bd}}(A)$ if and only if $c \leq a$.

- (4) If $c = a + \|v\|$, then

$$\sigma_{\text{int}}(A) = \{a\} \quad \text{and} \quad \sigma_{\text{bd}}(A) = \{c\}.$$

- (5) If $a + \|v\| < c$, then

$$\sigma_{\text{int}}(A) = \{a\} \quad \text{and} \quad \sigma_{\text{bd}}(A) = \emptyset.$$

Proof. This result follows from [Lemma A.2](#) and the fact that $\sigma_L(A + \gamma I) = \sigma_L(A) + \gamma$ for all $\gamma \in \mathbb{R}$. \square

Lemma A.4. *Let*

$$A = \begin{bmatrix} \tilde{A} & u \\ 0 & a \end{bmatrix},$$

where $\tilde{A} \in M_{n-1}$ and $u \in \mathbb{R}^{n-1}$ are fixed. Then $a \in \sigma_{\text{int}}(A)$ for all sufficiently large a .

Proof. We know a is an interior L-eigenvalue of A if and only if there exists ξ with $\|\xi\| < 1$ such that

$$(\tilde{A} - aI)\xi + u = 0.$$

For $a > \rho(\tilde{A})$, the matrix $\tilde{A} - aI$ is invertible, so we have

$$\xi = -(\tilde{A} - aI)^{-1}u.$$

Then since $\det(\tilde{A} - aI)$ is a polynomial in a of degree $n - 1 \geq 2$,

$$\begin{aligned} \|\xi\|^2 &= \|(\tilde{A} - aI)^{-1}u\|^2 \\ &= \left\| \frac{(\text{adj}(\tilde{A}) - aI)u}{\det(\tilde{A} - aI)} \right\|^2 = \frac{\|\text{adj}(\tilde{A})u\|^2 - 2au^T \text{adj}(\tilde{A})u + a^2\|u\|^2}{\det(\tilde{A} - aI)^2} \end{aligned}$$

approaches zero as $a \rightarrow \infty$. Thus, we likewise have $\|\xi\| < 1$ for all sufficiently large a . \square

A2. A result for 3×3 matrices.

Lemma A.5. *Let*

$$A = \begin{bmatrix} 0 & c & 0 \\ d & 0 & 0 \\ v_1 & v_2 & a \end{bmatrix}, \quad \text{where } cd \geq 0.$$

Then

$$\begin{aligned} &\sigma_{\text{bd}}(A) \\ &= \begin{cases} \frac{1}{2} \left(\pm \sqrt{\frac{c}{c+d}} v_1 \pm \sqrt{\frac{d}{c+d}} v_2 + a + \sqrt{cd} \right) & \text{if } \pm \sqrt{\frac{c}{c+d}} v_1 \pm \sqrt{\frac{d}{c+d}} v_2 + a - \sqrt{cd} \geq 0, \\ \frac{1}{2} \left(\pm \sqrt{\frac{c}{c+d}} v_1 \mp \sqrt{\frac{d}{c+d}} v_2 + a - \sqrt{cd} \right) & \text{if } \pm \sqrt{\frac{c}{c+d}} v_1 \mp \sqrt{\frac{d}{c+d}} v_2 + a + \sqrt{cd} \geq 0. \end{cases} \end{aligned}$$

Proof. Let $\lambda \in \sigma_{\text{bd}}(A)$. Then $\lambda = \mu + s$ with $s \geq 0$, and there exists $\xi = [\xi_1, \xi_2]^T$, with $\|\xi\| = 1$, such that

$$-\mu\xi_1 + c\xi_2 = 0,$$

$$d\xi_1 - \mu\xi_2 = 0,$$

$$v_1\xi_1 + v_2\xi_2 + a - \mu - 2s = 0.$$

Assume that $\mu = 0$. Then from the first equation, we get $\xi_2 = 0$ since $c \neq 0$. From the second equation, we get $\xi_1 = 0$ since $d \neq 0$, a contradiction as $\xi \neq 0$. Thus, $\mu \neq 0$, and the two first equations yield

$$(cd - \mu^2)\xi_2 = 0.$$

Since $\xi_2 = 0$ would imply $\xi_1 = 0$, contradicting $\|\xi\| = 1$, we have $\mu = \pm\sqrt{cd}$ and

$$\xi_1 = \frac{c}{\pm\sqrt{cd}}\xi_2 = \pm\sqrt{\frac{c}{d}}\xi_2.$$

Thus, since $\|\xi\| = 1$, we have

$$1 = \xi_1^2 + \xi_2^2 = \frac{c+d}{d}\xi_2^2.$$

Hence, if $\mu = \sqrt{cd}$, then

$$\xi_1 = \pm\sqrt{\frac{c}{c+d}} \quad \text{and} \quad \xi_2 = \pm\sqrt{\frac{d}{c+d}},$$

and if $\mu = -\sqrt{cd}$, then

$$\xi_1 = \pm\sqrt{\frac{c}{c+d}} \quad \text{and} \quad \xi_2 = \mp\sqrt{\frac{d}{c+d}}.$$

From the third equation, we get

$$v_1\xi_1 + v_2\xi_2 + a \mp \sqrt{cd} - 2s = 0,$$

or equivalently,

$$s = \frac{(v_1\xi_1 + v_2\xi_2) + a \mp \sqrt{cd}}{2},$$

which yields the claimed boundary L-eigenvalues $\lambda = \mu + s$ with the condition $s \geq 0$. □

References

[Alizadeh and Goldfarb 2003] F. Alizadeh and D. Goldfarb, “Second-order cone programming”, *Math. Program.* **95**:1 (2003), 3–51. [MR](#) [Zbl](#)

[Alizadeh and Shakeri 2017] R. Alizadeh and F. Shakeri, “Linear maps preserving Pareto eigenvalues”, *Linear Multilinear Algebra* **65**:5 (2017), 1053–1061. [MR](#) [Zbl](#)

[Ben-Israel and Greville 2003] A. Ben-Israel and T. N. E. Greville, *Generalized inverses: theory and applications*, 2nd ed., CMS Books Math./Ouvrages Math. SMC **15**, Springer, 2003. [MR](#) [Zbl](#)

[Bueno et al. 2021] M. I. Bueno, S. Furtado, and K. C. Sivakumar, “Linear maps preserving the Lorentz-cone spectrum in certain subspaces of M_n ”, *Banach J. Math. Anal.* **15**:3 (2021), art. id. 58. [MR](#) [Zbl](#)

[Bueno et al. 2022] M. I. Bueno, S. Furtado, A. Klausmeier, and J. Veltri, “Linear maps preserving the Lorentz spectrum: the 2×2 case”, *Electron. J. Linear Algebra* **38** (2022), 317–330. [MR](#) [Zbl](#)

[Németh and Gowda 2019] S. Z. Németh and M. S. Gowda, “The cone of \mathcal{Z} -transformations of the Lorentz cone”, *Electron. J. Linear Algebra* **35** (2019), 387–393. [MR](#) [Zbl](#)

[Seeger and Torki 2003] A. Seeger and M. Torki, “On eigenvalues induced by a cone constraint”, *Linear Algebra Appl.* **372** (2003), 181–206. [MR](#) [Zbl](#)

[Seeger and Torki 2020] A. Seeger and M. Torki, “On spectral maps induced by convex cones”, *Linear Algebra Appl.* **592** (2020), 65–92. [MR](#) [Zbl](#)

Received: 2022-08-31 Accepted: 2023-01-28

mbueno@ucsb.edu

*Department of Mathematics, University of California,
Santa Barbara, CA, United States*

benjaminfaktor@gmail.com

*Department of Mathematics, University of California,
Santa Barbara, CA, United States*

rkommerell@berkeley.edu

*Department of Mathematics, University of California,
Berkeley, CA, United States*

runzeli278@umail.ucsb.edu

*Department of Mathematics, University of California,
Santa Barbara, CA, United States*

jveltri@psu.edu

*Department of Mathematics, The Pennsylvania State
University, State College, PA, United States*

Lattice size in higher dimensions

Abdulrahman Alajmi, Sayok Chakravarty,
Zachary Kaplan and Jenya Soprunova

(Communicated by Ravi Vakil)

The lattice size of a lattice polytope is a geometric invariant which was formally introduced in the context of simplification of the defining equation of an algebraic curve, but appeared implicitly earlier in geometric combinatorics. Previous work on the lattice size was devoted to studying the lattice size in dimensions 2 and 3. We establish explicit formulas for the lattice size of a family of lattice simplices in arbitrary dimension.

1. Introduction

This paper is devoted to computing explicitly the lattice size for a family of lattice simplices in \mathbb{R}^{d+1} . We start with recalling some basic definitions related to lattice polytopes.

We say that a point $p \in \mathbb{R}^d$ is a *lattice point* if all of its coordinates are integers. A *lattice polytope* $P \subset \mathbb{R}^d$ is the convex hull of finitely many lattice points in \mathbb{Z}^d . A *lattice segment* is a segment that connects two lattice points. Such a segment is *primitive* if its only lattice points are its endpoints. The *lattice length* of a lattice segment is one less than the number of lattice points it contains (so that a primitive segment has lattice length 1). A lattice polytope is *empty* if its only lattice points are its vertices.

We say that matrix A of size d with integer entries is *unimodular* if $\det(A) = \pm 1$. The set of such matrices is denoted by $\text{GL}(d, \mathbb{Z})$. We say that a map $L: \mathbb{R}^d \rightarrow \mathbb{R}^d$ is an *affine unimodular map* if it is a composition of multiplication by a unimodular matrix and a translation by an integer vector. Such maps preserve the integer lattice $\mathbb{Z}^d \subset \mathbb{R}^d$. We say that two lattice polytopes in \mathbb{R}^d are *lattice-equivalent* if one is the image of another under an affine unimodular map.

MSC2020: 11H06, 52B20, 52C07.

Keywords: lattice size, lattice width, lattice polytopes.

Work of Chakravarty, Kaplan, and Soprunova was partially supported by NSF grant DMS-1653002.

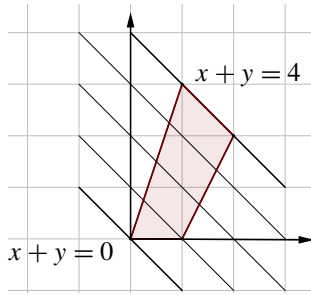


Figure 1. Lattice polygon P with $w_{(1,1)}(P) = 4$.

Let h be an integer vector in \mathbb{R}^d . For a lattice polytope $P \subset \mathbb{R}^d$, we define the *lattice width of P in the direction of h* by

$$w_h(P) = \max_{x \in P} \langle h, x \rangle - \min_{x \in P} \langle h, x \rangle,$$

where $\langle h, x \rangle$ is the standard inner product in \mathbb{R}^d . Then the *lattice width $w(P)$* of P is the minimum of $w_h(P)$ over nonzero integer vectors $h \in \mathbb{Z}^d$.

In [Figure 1](#), we illustrate the geometric meaning of this definition. Polygon P in the diagram is squeezed between the lines $x + y = 0$ and $x + y = 4$, so its width in the direction $(1, 1)$ is 4. We also have $w_{(1,0)}(P) = 2$. Since P has interior lattice points, we conclude that $w(P) = 2$.

The lattice size of a lattice polytope is an important geometric invariant of a lattice polytope that was formally introduced in [[Castricky and Cools 2015](#)], but appeared implicitly earlier in [[Arnold 1980](#); [Bárány and Pach 1992](#); [Brown and Kasprzyk 2013](#); [Lagarias and Ziegler 1991](#); [Schicho 2003](#)]. It was further studied in [[Alajmi and Soprunova 2022](#); [Harrison and Soprunova 2022](#); [Harrison et al. 2022](#); [Soprunova 2023](#)].

We next reproduce the definition of the lattice size from [[Castricky and Cools 2015](#)]. Let $0 \in \mathbb{R}^d$ be the origin and let (e_1, \dots, e_d) be the standard basis of \mathbb{R}^d . The *standard simplex* $\Delta \subset \mathbb{R}^d$ is defined by $\Delta = \text{conv}\{0, e_1, \dots, e_d\}$, where “conv” denotes the convex hull operator.

Definition 1.1. Let $P \subset \mathbb{R}^d$ be a lattice polytope. The lattice size $\text{ls}_\Delta(P)$ of P with respect to the standard simplex Δ is the smallest l such that $L(P)$ is contained in the l -dilate of Δ for some affine unimodular map $L: \mathbb{R}^d \rightarrow \mathbb{R}^d$.

Equivalently, if we let

$$l_1(P) = \max_{(x_1, \dots, x_d) \in P} (x_1 + \dots + x_d) - \min_{(x_1, \dots, x_d) \in P} x_1 - \dots - \min_{(x_1, \dots, x_d) \in P} x_d, \quad (1-1)$$

then $\text{ls}_\Delta(P)$ is the minimum of $l_1(L(P))$ over affine unimodular maps $L: \mathbb{R}^d \rightarrow \mathbb{R}^d$.

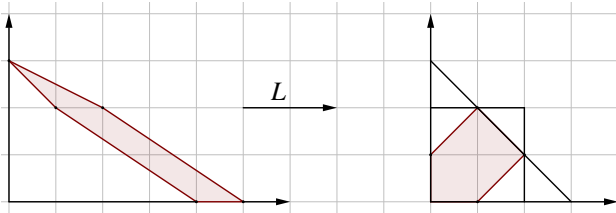


Figure 2. Illustration of [Example 1.2](#).

If in the above definition the standard simplex Δ is replaced with the unit cube $\square = [0, 1]^d$, we obtain the definition of the lattice size $\text{ls}_{\square}(P)$ with respect to the unit cube. Note that the lattice width $w(P)$ can be viewed as the lattice size with respect to the strip $\mathbb{R}^{d-1} \times [0, 1]$.

Example 1.2. Let P be the polygon with vertices $(4, 0)$, $(5, 0)$, $(2, 2)$, $(0, 3)$, and $(1, 2)$, as drawn in [Figure 2](#). Define

$$L(x, y) = \begin{bmatrix} 1 & 1 \\ -1 & -2 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} -3 \\ 6 \end{bmatrix}.$$

We get

$$L(P) = \text{conv}\{(1, 2), (2, 1), (1, 0), (0, 0), (0, 1)\},$$

so $L(P) \subset 2\square$ and $L(P) \subset 3\Delta$. Note that P has an interior lattice point, while \square and 2Δ do not, so it is impossible to unimodularly map P inside \square or 2Δ . We conclude that $\text{ls}_{\square}(P) = 2$ and $\text{ls}_{\Delta}(P) = 3$.

It was shown in [[Harrison and Soprunova 2022](#); [Harrison et al. 2022](#)] that in dimension 2 both $\text{ls}_{\Delta}(P)$ and $\text{ls}_{\square}(P)$ can be computed using basis reduction (see [[Harrison and Soprunova 2022](#); [Harrison et al. 2022](#)] for definitions and details). It is further explained in [[Harrison and Soprunova 2022](#)] that basis reduction also computes the lattice size $\text{ls}_{\square}(P)$ in dimension 3. This leads to fast algorithms for computing the lattice size in these cases. A counterexample in [[Harrison and Soprunova 2022](#)] demonstrates that a reduced basis does not necessarily compute $\text{ls}_{\Delta}(P)$ in dimension 3.

A well-known classification result of [[White 1964](#)] asserts that up to lattice equivalence empty lattice tetrahedra in \mathbb{R}^3 are of the form

$$T_{pq} = \begin{bmatrix} 1 & 0 & 0 & p \\ 0 & 1 & 0 & q \\ 0 & 0 & 1 & 1 \end{bmatrix},$$

where p and q are nonnegative relatively prime integers. (Note that in this notation T_{pq} is the convex hull of the column vectors of the matrix.) It was further shown in [[Scarf 1985](#)] that any empty lattice polytope in \mathbb{R}^3 has lattice width 1.

While it is not true that a reduced basis computes $ls_{\Delta}(P)$ for $P \subset \mathbb{R}^3$, it was shown in [Alajmi and Soprunova 2022] that this is the case for 3-dimensional empty lattice polytopes. A counterexample was provided in [Alajmi and Soprunova 2022] demonstrating that the conclusion does not generalize to all polytopes $P \subset \mathbb{R}^3$ with lattice width 1.

All the results discussed above concern the lattice size of lattice polytopes in \mathbb{R}^2 and \mathbb{R}^3 . In this paper, we consider a family of lattice simplices P in \mathbb{R}^{d+1} for arbitrary d and explicitly compute both $ls_{\Delta}(P)$ and $ls_{\square}(P)$ under some assumptions on the parameters of the family. We work with simplices of the form

$$T_{p_1 \dots p_d} = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 & p_1 \\ 0 & 1 & \dots & 0 & 0 & p_2 \\ \vdots & & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 & p_d \\ 0 & 0 & \dots & 0 & 1 & 1 \end{bmatrix},$$

where p_1, \dots, p_d are nonnegative integers. These simplices are a natural $(d+1)$ -dimensional generalization of the tetrahedra $T_{pq} \subset \mathbb{R}^3$. Each such T_{p_1, \dots, p_d} has lattice width 1. Also, it is empty if and only if $\gcd(p_1, \dots, p_d) = 1$, although we will not be making this assumption. Note that starting with dimension 4 it is no longer true that every empty lattice simplex has lattice width 1; see [Haase and Ziegler 2000].

Our main results are formulated in Theorems 3.4 and 3.5, where we provide explicit formulas for both $ls_{\Delta}(T_{p_1, \dots, p_d})$ and $ls_{\square}(T_{p_1, \dots, p_d})$ in terms of p_1, \dots, p_d under some restrictions on these parameters. Our methods are elementary and different from the ones used in earlier work in dimensions 2 and 3.

2. First lemmas

Here we provide the d -dimensional version of the introductory statements about the lattice size, which were formulated in [Alajmi and Soprunova 2022] for the case $d = 3$.

Let $P \subset \mathbb{R}^d$ be a lattice polytope and $A \in GL(d, \mathbb{Z})$. Denote the rows of A by h_1, \dots, h_d . Recall the definition of $l_1(P)$ in (1-1).

Lemma 2.1. (1) For any $h \in \mathbb{Z}^d$ we have $w_h(AP) = w_{A^T h}(P)$.

(2) For $i = 1, \dots, d$ we have $w_{e_i}(AP) = w_{h_i}(P)$.

(3) Let (e_1, \dots, e_d) be the standard basis of \mathbb{R}^d . Then $w_{e_i}(P) \leq l_1(P)$ for $i = 1, \dots, d$.

(4) For $i = 1, \dots, d$ we have $w_{h_i}(P) \leq l_1(AP)$.

(5) Let $e \in \mathbb{Z}^d$ be a vector whose entries lie in $\{0, 1\}$. Then $w_e(P) \leq l_1(P)$.

(6) Let h be the sum of any nonempty collection of rows of A . Then $w_h(P) \leq l_1(AP)$.

Proof. For (1) we have

$$\begin{aligned} w_h(AP) &= \max_{x \in P} \langle h, Ax \rangle - \min_{x \in P} \langle h, Ax \rangle \\ &= \max_{x \in P} \langle A^T h, x \rangle - \min_{x \in P} \langle A^T h, x \rangle = w_{A^T h}(P), \end{aligned}$$

and (2) is a particular case of (1).

To check (3), define $l_1 := l_1(P)$. Then $P \subset l_1 \Delta$ and hence

$$w_{e_i}(P) \leq w_{e_i}(l_1 \Delta) = l_1 = l_1(P).$$

Thus (4) follows as $w_{h_i}(P) = w_{e_i}(AP) \leq l_1(AP)$.

Next we check (5), which is similar to (3):

$$w_e(P) \leq w_e(l_1 \Delta) = l_1 = l_1(P).$$

For (6) let e be the sum of the corresponding standard basis vectors. Then by (1) and (5)

$$w_h(P) = w_e(AP) \leq l_1(AP). \quad \square$$

Lemma 2.2. *We have:*

- (1) $l_1(AP) = \max_{x \in P} \langle h_1 + \cdots + h_d, x \rangle - \min_{x \in P} \langle h_1, x \rangle - \cdots - \min_{x \in P} \langle h_d, x \rangle$.
- (2) $l_1(AP)$ does not depend on the order of rows in A .
- (3) $l_1(AP) = l_1(BP)$, where

$$B = \begin{bmatrix} h_1 \\ \vdots \\ h_{d-1} \\ -\sum_{i=1}^d h_i \end{bmatrix}.$$

Proof. (1) and (2) are clear. Let's check (3) using (1):

$$l_1(BP) = \max_{x \in P} \langle -h_d, x \rangle - \sum_{i=1}^{d-1} \min_{x \in P} \langle h_i, x \rangle - \min_{x \in P} \langle (-h_1 - \cdots - h_d), x \rangle = l_1(AP). \quad \square$$

3. Lattice size computation

Recall that

$$T_{p_1, \dots, p_d} = \text{conv}\{e_1, \dots, e_{d+1}, (p_1, \dots, p_d, 1)\} \subset \mathbb{R}^{d+1},$$

where $p_i \in \mathbb{Z}_{\geq 0}$. Define $\alpha = p_1 + \cdots + p_{d-1}$. Assume that $p_d \geq 2$ and let $k = \lfloor (p_d - 2)/(\alpha + 1) \rfloor$.

Proposition 3.1. *With p_i , α , and k defined as above, suppose that $p_d \geq \alpha^2 - \alpha$. Then $\text{ls}_\Delta(T_{p_1, \dots, p_d}) \leq k + 3$.*

Proof. Consider unimodular matrix A of size $d + 1$ defined by

$$A = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 \\ k+1 & k+1 & \dots & k+1 & -1 & p_d - \alpha(k+1) - 1 \end{bmatrix}.$$

Then

$$AT_{p_1, \dots, p_d} = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 & p_1 \\ 0 & 1 & \dots & 0 & 0 & p_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 & p_{d-1} \\ 0 & 0 & \dots & 0 & 1 & 1 \\ k+1 & k+1 & \dots & k+1 & p_d - \alpha(k+1) - 1 & -1 \end{bmatrix},$$

and hence

$$l_1(AT_{p_1, \dots, p_d}) = \max\{k + 2, p_d - \alpha(k + 1), \alpha\} - \min\{-1, p_d - \alpha(k + 1) - 1\}.$$

Since $k = \lfloor (p_d - 2)/(\alpha + 1) \rfloor$, we have $k + 1 > (p_d - 2)/(\alpha + 1)$, which implies $p_d - \alpha(k + 1) < k + 3$ and, since p_d, k and α are integers, we conclude $p_d - \alpha(k + 1) \leq k + 2$. We also have

$$k + 2 > \frac{p_d - 2}{\alpha + 1} + 1 = \frac{p_d + \alpha - 1}{\alpha + 1} \geq \frac{\alpha^2 - \alpha + \alpha - 1}{\alpha + 1} = \alpha - 1,$$

where we used the assumption $p_d \geq \alpha^2 - \alpha$. We have checked that

$$\max\{k + 2, p_d - \alpha(k + 1), \alpha\} = k + 2.$$

Next note that since $(p_d - 2)/(\alpha + 1) \geq k$ we get

$$\begin{aligned} p_d - 1 - \alpha(k + 1) &\geq p_d - 1 - \alpha\left(\frac{p_d - 2}{\alpha + 1} + 1\right) \\ &= \frac{-\alpha^2 - 1 + p_d}{\alpha + 1} \geq \frac{-\alpha^2 - 1 + \alpha^2 - \alpha}{\alpha + 1} = -1. \end{aligned}$$

Hence $\min\{-1, p_d - \alpha(k + 1) - 1\} = -1$ and $l_1(AT_{p_1, \dots, p_d}) = k + 3$, which implies $ls_\Delta(T_{p_1, \dots, p_d}) \leq k + 3$. □

Our next goal is to show that under our assumptions we have $ls_\Delta(T_{p_1, \dots, p_d}) = k + 3$. For this, we first prove two lemmas.

Lemma 3.2. *Let $h = (a_1, \dots, a_{d+1}) \in \mathbb{R}^{d+1}$ be a primitive vector with $w_h(T_{p_1, \dots, p_d}) \leq k + 2$. Then $a_d \in \{0, 1, -1\}$.*

Proof. We can assume $a_d \geq 0$. Suppose that $a_d \geq 2$. We have

$$\begin{aligned} \max\{a_1, \dots, a_{d+1}, a_1 p_1 + \dots + a_d p_d + a_{d+1}\} \\ - \min\{a_1, \dots, a_{d+1}, a_1 p_1 + \dots + a_d p_d + a_{d+1}\} \leq k + 2, \end{aligned}$$

which implies $a_1 p_1 + \dots + a_d p_d \leq k + 2$ and $a_d - a_i \leq k + 2$ for $i = 1, \dots, d - 1$. Since $a_d \geq 2$, the first inequality implies

$$a_1 p_1 + \dots + a_{d-1} p_{d-1} \leq k + 2 - 2p_d.$$

The second inequality implies $a_i \geq -k$ for $i = 1, \dots, d - 1$ and hence

$$a_1 p_1 + \dots + a_{d-1} p_{d-1} \geq -k(p_1 + \dots + p_{d-1}) = -k\alpha.$$

Combining this with what we got from the first inequality, we get $k + 2 - 2p_d \geq -k\alpha$. Hence, using the definition of k , we get

$$2p_d \leq k + 2 + k\alpha = k(\alpha + 1) + 2 \leq p_d,$$

which contradicts the assumption $p_d > 0$. □

Lemma 3.3. *Let $h = (a_1, \dots, a_{d+1}) \in \mathbb{R}^{d+1}$ be a primitive vector with $a_d = \pm 1$ and $w_h(T_{p_1, \dots, p_d}) \leq k + 2$. Then $w_h(T_{p_1, \dots, p_d}) = k + 2$.*

Proof. We can assume that $a_d = 1$. Suppose that

$$\begin{aligned} \max\{a_1, \dots, a_{d-1}, 1, a_{d+1}, a_1 p_1 + \dots + a_{d-1} p_{d-1} + p_d + a_{d+1}\} \\ - \min\{a_1, \dots, a_{d-1}, 1, a_{d+1}, a_1 p_1 + \dots + a_{d-1} p_{d-1} + p_d + a_{d+1}\} \leq k + 1. \end{aligned}$$

This implies $a_1 p_1 + \dots + a_{d-1} p_{d-1} + p_d \leq k + 1$ and $1 - a_i \leq k + 1$ for $i = 1, \dots, d - 1$. Hence for such i we have $a_i \geq -k$ and

$$-k\alpha + p_d \leq a_1 p_1 + \dots + a_{d-1} p_{d-1} + p_d \leq k + 1,$$

which implies $-k\alpha + p_d \leq k + 1$. Hence $p_d \leq k\alpha + k + 1 = k(\alpha + 1) + 1 \leq p_d - 1$, and this is impossible. □

Theorem 3.4. *Let $\alpha = p_1 + \dots + p_{d-1}$, where all p_i are positive and $p_d \geq 2$. Define $k = \lfloor (p_d - 2)/(\alpha + 1) \rfloor$. Suppose that $p_d \geq \alpha^2 - \alpha$. Then $\text{ls}_\Delta(T_{p_1, \dots, p_d}) = k + 3$.*

Proof. Suppose that there exists a unimodular map L that maps T_{p_1, \dots, p_d} inside $(k+2)\Delta$ and let A be the corresponding unimodular matrix. Then by [Lemma 2.1](#) for each of its rows h we have $w_h(T_{p_1, \dots, p_d}) \leq k + 2$. We also have the same inequality for the sum of any nonempty collection of rows of A . By [Lemma 3.2](#) each of the entries in the d -th column of A is 0, 1, or -1 , and the same applies to the sum of any collection of entries in the d -th column of A . Hence, up to permutation of rows, the d -th column of A is $(0, 0, \dots, 0, \pm 1)^T$ or $(0, 0, \dots, 0, 1, -1)^T$ and

by [Lemma 2.2](#) we can assume that it is the former. Then by [Lemma 3.3](#) we have $w_{e_{d+1}}(L(T_{p_1, \dots, p_d})) = k + 2$. Since $L(T_{p_1, \dots, p_d}) \subset (k + 2)\Delta$, we conclude that

$$(0, \dots, k + 2) \in L(T_{p_1, \dots, p_d}).$$

Similarly, we have $w_{e_1 + \dots + e_{d+1}}(L(T_{p_1, \dots, p_d})) = k + 2$ and, together with $L(T_{p_1, \dots, p_d}) \subset (k + 2)\Delta$, this implies that $L(T_{p_1, \dots, p_d})$ contains the origin. Hence $L(T_{p_1, \dots, p_d})$ and, therefore, T_{p_1, \dots, p_d} contains an edge of lattice length $k + 2$. All the edges of T_{p_1, \dots, p_d} are primitive, except, possibly, for the one connecting points $(0, \dots, 1)$ and $(p_1, \dots, p_d, 1)$, whose lattice length is $\gcd(p_1, \dots, p_d)$. Hence we conclude that $\gcd(p_1, \dots, p_d) = k + 2$. Using the assumption $p_d \geq \alpha^2 - \alpha$ we get

$$k + 2 = \gcd(p_1, \dots, p_d) \leq p_1 + \dots + p_{d-1} = \alpha \leq \frac{p_d - 2}{\alpha + 1} + 2 < k + 3.$$

Note that since all the p_i are positive, we have $\gcd(p_1, \dots, p_d) < p_1 + \dots + p_{d-1}$ unless $d = 2$ and p_2 is a multiple of p_1 . When the inequality is strict we arrive at a contradiction since the integer $p_1 + \dots + p_{d-1}$ is strictly between the consecutive integers $k + 2$ and $k + 3$.

It remains to consider the case when $d = 2$, p_2 is a multiple of p_1 , and $k + 2 = \alpha = p_1$. We have $\lfloor (p_2 - 2)/(p_1 + 1) \rfloor = k = p_1 - 2$ and hence

$$p_2 - 2 = (p_1 + 1)(p_1 - 2) + r,$$

where $0 \leq r \leq p_1$. We get $p_2 = p_1^2 - p_1 + r$ and, since p_2 is a multiple of p_1 , there are two options: $p_2 = p_1^2$ and $p_2 = p_1^2 - p_1$.

Suppose first $p_2 = p_1^2$ and let $h = (a_1, a_2, a_3)$ be a direction with $w_h(T_{p_1 p_2}) \leq k + 2 = p_1$. By [Lemma 3.2](#) we have $a_2 = 0, \pm 1$. For $a_2 = -1$ we get

$$\max\{a_1, -1, a_3, a_1 p_1 - p_1^2 + a_3\} - \min\{a_1, -1, a_3, a_1 p_1 - p_1^2 + a_3\} \leq p_1,$$

which implies $a_1 + 1 \leq p_1$ and $-a_1 p_1 + p_1^2 \leq p_1$, so we conclude that $a_1 = p_1 - 1$. Plugging in this value for a_1 we get

$$\max\{p_1 - 1, -1, a_3, a_3 - p_1\} - \min\{p_1 - 1, -1, a_3, a_3 - p_1\} \leq p_1,$$

which implies $a_3 + 1 \leq p_1$ and $p_1 - 1 - (a_3 - p_1) \leq p_1$, so $a_3 = p_1 - 1$. We conclude that for $a_2 = -1$ the only h with $w_h(T_{p_1 p_2}) \leq p_1$ is $h = (p_1 - 1, -1, p_1 - 1)$ and for such h we get $w_h(T_{p_1 p_2}) = k + 2$. Similarly, for $a_2 = 1$ such direction is $h = (1 - p_1, 1, 1 - p_1)$.

Hence in this case we can only use as rows of A vectors $\pm(p_1 - 1, -1, p_1 - 1)$ and vectors whose second component is 0. Further, we can assume that the second column of A is $(0, 0, \pm 1)^T$, and the third row is $\pm(p_1 - 1, -1, p_1 - 1)$. Then the sum of the third row with any of the first two rows will also have to be of the same form as the third row, which would imply $\det A = 0$.

We next consider the last case $p_2 = p_1^2 - p_1$. Recall that we have $k + 2 = p_1$. By Lemmas 3.2 and 3.3 if $w_h(T_{p_1 p_2}) \leq p_1$ for $h = (a_1, a_2, a_3)$ then $a_2 = 0, \pm 1$, and if $a_2 = \pm 1$ we have $w_h(T_{p_1 p_2}) = p_1$. Let's further investigate the case $a_2 = 0$. We have

$$\max\{a_1, 0, a_3, a_1 p_1 + a_3\} - \min\{a_1, 0, a_3, a_1 p_1 + a_3\} \leq p_1,$$

where we can assume $a_1 \geq 0$. Hence $a_1 p_1 \leq p_1$, which implies $a_1 = 0$ or 1 , and in the latter case the width is p_1 . We conclude that the only direction h with $w_h(T_{p_1 p_2}) < p_1$ is $(0, 0, \pm 1)$.

As before, we can assume that the second column in matrix A is $(0, 0, 1)^T$. Denote the rows of A by h_1, h_2 , and h_3 . Then out of the widths of $T_{p_1 p_2}$ in the direction of h_1 and h_2 , only one can be strictly less than p_1 . We can assume that this happens in the direction of h_1 . We also know that the width of $T_{p_1 p_2}$ in the direction of h_3 and $h_1 + h_2 + h_3$ is p_1 . We can now conclude that $L(T_{p_1 p_2})$ contains the triangle

$$\text{conv}\{(0, 0, 0), (0, p_1, 0), (0, 0, p_1)\}.$$

Note that we assumed that $p_d \geq 2$, which implies $p_1^2 - p_1 = p_2 \geq 2$ and hence $p_1 \geq 2$. Hence our conclusion implies that $T_{p_1 p_2}$ contains a lattice triangle with three nonprimitive sides, and this contradiction completes the argument. \square

We next use the work above to compute $\text{ls}_{\square}(T_{p_1, \dots, p_d})$.

Theorem 3.5. *Let $\alpha = p_1 + \dots + p_{d-1}$, where all p_i are positive and $p_d \geq 2$. Define $k = \lfloor (p_d - 2)/(\alpha + 1) \rfloor$. Suppose that $p_d \geq \alpha^2 - \alpha$. Then $\text{ls}_{\square}(T_{p_1, \dots, p_d}) = k + 2$.*

Proof. Let A be the matrix from Proposition 3.1. Then $w_{e_d}(AT_{p_1, \dots, p_d}) = 1$, $w_{e_i}(AT_{p_1, \dots, p_d}) = p_i$ for $i = 1, \dots, d - 1$, and

$$w_{e_{d+1}}(AT_{p_1, \dots, p_d}) = \max\{k+1, p_d - \alpha(k+1) - 1\} - \min\{-1, p_d - \alpha(k+1) - 1\} = k+2,$$

as shown in the proof of Proposition 3.1. It is also checked there that $\alpha \leq k + 2$, and hence after a translation by the vector $v = (0, \dots, 0, 1)^T$ we get $AT_{p_1, \dots, p_d} + v \subset [0, k + 2]^{d+1}$, which implies that $\text{ls}_{\square}(P) \leq k + 2$.

Suppose next there exists a unimodular map $L: \mathbb{R}^d \rightarrow \mathbb{R}^d$ such that $L(T_{p_1, \dots, p_d}) \subset [0, k + 1]^{d+1}$. Then the width of T_{p_1, \dots, p_d} in the direction of each of the rows of the corresponding matrix A is at most $k + 1$, but by Lemmas 3.2 and 3.3 this implies that the d -th entry of each row of A is zero, so $\det A = 0$. \square

Note that we have checked in Theorems 3.4 and 3.5 that there exists a matrix A that computes both $\text{ls}_{\Delta}(T_{p_1, \dots, p_d})$ and $\text{ls}_{\square}(T_{p_1, \dots, p_d})$. While it was shown in [Harrison and Soprunova 2022; Harrison et al. 2022] that this is always the case when $P \subset \mathbb{R}^2$, a counterexample for $P \subset \mathbb{R}^3$ was provided in [Harrison and Soprunova 2022].

Acknowledgments

We are grateful to the Kent State REU program for the hospitality. We would also like to thank the anonymous referee for useful suggestions.

References

- [Alajmi and Soprunova 2022] A. Alajmi and J. Soprunova, “Lattice size of width one lattice polytopes in \mathbb{R}^3 ”, preprint, 2022. [arXiv 2207.13124](#)
- [Arnold 1980] V. I. Arnold, “Statistics of integral convex polygons”, *Funktional. Anal. i Prilozhen.* **14**:2 (1980), 1–3. In Russian; translated in *Funct. Anal. Its Appl.* **14**:2 (1980), 79–81. [MR](#) [Zbl](#)
- [Bárány and Pach 1992] I. Bárány and J. Pach, “On the number of convex lattice polygons”, *Combin. Probab. Comput.* **1**:4 (1992), 295–302. [MR](#) [Zbl](#)
- [Brown and Kasprzyk 2013] G. Brown and A. M. Kasprzyk, “Small polygons and toric codes”, *J. Symbolic Comput.* **51** (2013), 55–62. [MR](#) [Zbl](#)
- [Castricky and Cools 2015] W. Castryck and F. Cools, “The lattice size of a lattice polygon”, *J. Combin. Theory Ser. A* **136** (2015), 64–95. [MR](#) [Zbl](#)
- [Haase and Ziegler 2000] C. Haase and G. M. Ziegler, “On the maximal width of empty lattice simplices”, *European J. Combin.* **21**:1 (2000), 111–119. [MR](#) [Zbl](#)
- [Harrison and Soprunova 2022] A. Harrison and J. Soprunova, “Lattice size and generalized basis reduction in dimension three”, *Discrete Comput. Geom.* **67**:1 (2022), 287–310. [MR](#) [Zbl](#)
- [Harrison et al. 2022] A. Harrison, J. Soprunova, and P. Tierney, “Lattice size of plane convex bodies”, *SIAM J. Discrete Math.* **36**:1 (2022), 92–102. [MR](#) [Zbl](#)
- [Lagarias and Ziegler 1991] J. C. Lagarias and G. M. Ziegler, “Bounds for lattice polytopes containing a fixed number of interior points in a sublattice”, *Canad. J. Math.* **43**:5 (1991), 1022–1035. [MR](#) [Zbl](#)
- [Scarf 1985] H. E. Scarf, “Integral polyhedra in three space”, *Math. Oper. Res.* **10**:3 (1985), 403–438. [MR](#) [Zbl](#)
- [Schicho 2003] J. Schicho, “Simplification of surface parametrizations: a lattice polygon approach”, *J. Symbolic Comput.* **36**:3-4 (2003), 535–554. [MR](#) [Zbl](#)
- [Soprunova 2023] J. Soprunova, “Bounds on area involving lattice size”, *Electron. J. Combin.* **30**:4 (2023), art. id. 4.45. [MR](#)
- [White 1964] G. K. White, “Lattice tetrahedra”, *Canadian J. Math.* **16** (1964), 389–396. [MR](#) [Zbl](#)

Received: 2022-09-09

Revised: 2022-12-23

Accepted: 2022-12-27

aalajmi@kent.edu

Department of Mathematics, Public Authority for Applied Education and Training, Ardiya, Kuwait

schakr31@uic.edu

Department of Mathematics, Statistics, and Computer Science, University of Illinois, Chicago, IL, United States

zachary.kaplan@dep.nj.gov

New Jersey Department of Environmental Protection, Trenton, NJ, United States

esopruno@kent.edu

Department of Mathematical Sciences, Kent State University, Kent, OH, United States

On the joint evolution problem for a scalar field and its singularity

Aditya Agashe, Ethan Lee and Shadi Tahvildar-Zadeh

(Communicated by Martin Bohner)

In the classical electrodynamics of point charges in vacuum, the electromagnetic field, and therefore the Lorentz force, is ill-defined at the locations of the charges. Kiessling resolved this problem by using the momentum balance between the field and the particles, extracting an equation for the force that is well-defined where the charges are located, so long as the field momentum density is locally integrable in a neighborhood of the charges.

We examine the effects of such a force by analyzing a simplified model in one space dimension. We study the joint evolution of a massless scalar field together with its singularity, which we identify with the trajectory of a particle. The static solution arises in the presence of no incoming radiation, in which case the particle remains at rest forever. We will prove the stability of the static solution for particles with positive bare mass by showing that a pulse of incoming radiation that is compactly supported away from the point charge will result in the particle eventually coming back to rest. We will also prove the nonlinear instability of the static solution for particles with negative bare mass by showing that an incoming radiation with arbitrarily small amplitude will cause the particle to reach the speed of light in finite time. We conclude by discussing modifications to this simple model that could make it more realistic.

1. Introduction and main result

Classical electromagnetism has a fundamental problem: for a charged point-particle in an electromagnetic field that is at least partly sourced by that particle, the field is not defined at the location of the particle. Because the Lorentz force that the field exerts on the particle depends on the values of the field at the particle's location, the force is also undefined where it's needed, i.e., for the particle's equations of motion to make sense. This is the infamous *radiation-reaction problem*. This problem has been the subject of intense study by some of the world's most renowned physicists

MSC2020: 35A21, 70S10, 78A35.

Keywords: radiation-reaction problem, propagation of singularities, scalar fields, point-charge sources.

and mathematicians, including Poincaré [1906] and Dirac [1938], for more than a century. An excellent account of this endeavor can be found in [Spohn 2004], where an entire chapter is devoted to recounting its history.¹

An important breakthrough came when, following up on some ideas of Poincaré [1906], Kiessling [2019] showed that if one postulated local conservation laws to hold for the total (field + particle) energy density-momentum density-stress tensor, i.e., (employing the Einstein summation convention, where repeated indices are summed over their range)

$$\partial^\mu T_{\mu\nu}^{\text{total}} = 0, \quad \text{where } T_{\mu\nu}^{\text{total}} = T_{\mu\nu}^{\text{field}} + T_{\mu\nu}^{\text{particle}}, \quad (1)$$

(once these expressions are properly defined) then the force can be determined using the law of momentum balance, provided the field momentum density is locally integrable in a neighborhood of the charge. This integrability assumption rules out the classical electromagnetic vacuum law $E = D$, $B = H$ postulated by Maxwell, but admits others, such as the Bopp–Landé–Thomas–Podolsky (BLTP for short) vacuum law [Bopp 1940; 1942; Landé 1941; Landé and Thomas 1941; Podolsky 1942].

It is of interest to study the effect of the Kiessling force on the motion of an electromagnetic point-charge. In three space dimensions using the standard electromagnetic vacuum laws, this is not possible. One can use other vacuum laws in three space dimensions, such as BLTP, that do meet Kiessling’s criterion, and for which one can prove local well-posedness of the joint field and particle dynamics [Kiessling 2019; Kiessling and Tahvildar-Zadeh \geq 2024] as well as global existence for the solution to the scattering problem of a single particle by a smooth potential [Hoang et al. 2021]. However, the expression for the force in the BLTP case is quite complicated, which makes it hard to figure out what the particle trajectories actually look like. On the other hand, by a simple scaling analysis, it is easy to see that Kiessling’s criterion may be satisfied for Maxwell’s vacuum law, so long as one works in *one* space dimension. Since there is however no viable electromagnetism in one space dimension, we instead turn to the simpler model of a *scalar* charge. Such a model has been proposed before by many authors; see, e.g., [Elskens et al. 2009]. To simplify matters even further, we will be focusing on the case of a single particle perturbed by scalar radiation. Following Weyl’s ideas [1921] on singularity theories of matter, we will take the evolving singularity of the scalar field to represent the path of the particle in space-time. This is thus a joint evolution problem for a scalar field $u(t, s)$, and the trajectory of its singularity, $s = q(t)$. The governing equations are as follows (see [Elskens et al. 2009, equations (7)–(9)]):

¹It is outside the scope of this article to mention all the various directions taken by researchers to resolve this issue. Interested readers are referred to [Spohn 2004] and its copious bibliography.

the field satisfies

$$\begin{cases} \partial_t^2 u - \partial_s^2 u = a\delta(s - q(t)), \\ u(0, s) = -\frac{a}{2}|s| + V_0(s), \\ \partial_t u(0, s) = V_1(s) \end{cases} \quad (2)$$

(δ is the Dirac delta-function), while the equations of motion for the particle are

$$\begin{cases} \dot{q} = p/(m\sqrt{1 + p^2/m^2}), \\ \dot{p} = f(t, q, \dot{q}), \\ q(0) = 0, \\ \dot{q}(0) = 0 \end{cases} \quad (3)$$

(the speed of light has been set equal to 1).

Here, (2) is the Cauchy problem for a massless wave equation sourced by the particle. We have added $V_0(s)$ and $V_1(s)$ to the initial data to represent smooth incoming radiation that is compactly supported away from the point charge. Thus $V_0, V_1 \in C_c^\infty(\mathbb{R} \setminus \{0\})$. Real constants a and m represent the charge and the (bare) rest mass of the particle. Equations (3) are simply Newton's equations of motion with Einstein's special-relativistic relation between momentum and velocity instead of Newton's. By f we denote the force exerted by the field on the particle, and its precise expression needs to be determined using another principle. Here, following Kiessling, we will use *momentum conservation* to determine f .

Remark 1. We note that the above system of equations is not fully Lorentz-covariant: the right-hand side of the wave equation in (2) does *not* transform correctly under a Lorentz transformation. This is a defect of the model (which was also pointed out in [Elskens et al. 2009]). This defect can be corrected, but the resulting system becomes harder to analyze. Some of the results in this paper have also been obtained for the fully relativistic version, and will appear elsewhere [Frolov et al. 2023].

The initial conditions in (3) can always be satisfied by going into the initial rest frame of the particle. We will use Kiessling's prescription to determine the force f on the particle. This will depend on the field u , which makes (2)–(3) a coupled system of equations for the joint evolution of the field and its singularity.

Consider first the case of no incoming radiation, i.e., $V_0 \equiv 0$ and $V_1 \equiv 0$. In that case, $u = -\frac{a}{2}|s|$, where $q(t) = 0$ for all t is clearly a time-independent solution to (2), i.e., the particle remains at rest forever. We shall see that in this case, $f \equiv 0$. In this paper we will prove:

Theorem 2. (a) *Suppose $m > 0$. For all smooth initial data (V_0, V_1) for (2) that are compactly supported away from the origin, there exists a solution $(u(t, s), q(t))$ to the joint field-particle evolution problem (2)–(3), with the property that*

- (i) the field u is Lipschitz everywhere and the particle trajectory q is C^1 ,
 - (ii) u is at least C^1 away from the particle path $s = q(t)$, and
 - (iii) for all $\epsilon > 0$, there exists $T > 0$ such that $t > T$ implies $|\dot{q}(t)| < \epsilon$.
- (b) Suppose $m < 0$. For all $\epsilon > 0$, there exists smooth, compactly supported initial data (V_0, V_1) for (2), with $\|V_0\|_{C^1(\mathbb{R})} + \|V_1\|_{C^0(\mathbb{R})} < \epsilon$, and a solution $(u(t, s), q(t))$ to the joint field-particle evolution problem (2)–(3), with the property that
- (i) the field u is Lipschitz everywhere and the particle trajectory q is C^1 ,
 - (ii) u is at least C^1 away from the particle path $s = q(t)$, and
 - (iii) the particle reaches the speed of light in finite time, i.e., there exists $T > 0$ such that $|\dot{q}(T)| = 1$.

Outline of the proof. In Section 2 we solve the field equations (2) assuming the trajectory $s = q(t)$ of the singularity is given. We do this by decomposing the field into a smooth part depending only on the incoming radiation, and a singular part sourced by the particle. In Section 3 we use this field to compute the Kiessling force f in (3), and show that it depends only on the smooth part of the field. We can thus eliminate the field from (3) and have $q(t)$ be the only unknown. In Section 4 we study (3) by turning it into a dynamical system in the plane and analyzing its phase portrait, which will allow us to prove the stability claim in Section 5 and the instability claim in Section 6.

We conclude in Section 7 by speculating on the mechanism for instability, and propose various modifications to our model that could perhaps avoid such instabilities.

2. Solving the field equations

Proposition 3. For any given trajectory $q(t)$ with $|\dot{q}| < 1$, $q(0) = 0$, and $\dot{q}(0) = 0$, the initial value problem

$$\begin{cases} \partial_t^2 u - \partial_s^2 u = a\delta(s - q(t)), \\ u(0, s) = -\frac{a}{2}|s| + V_0(s), \\ \partial_t u(0, s) = V_1(s) \end{cases} \tag{4}$$

has the unique solution

$$u(t, s) = \frac{a}{2} \begin{cases} s + V(t, s), & s < -t, \\ T_+(s + t) - t + V(t, s), & -t < s < q(t), \\ T_-(s - t) - t + V(t, s), & q(t) < s < t, \\ -s + V(t, s), & s > t, \end{cases} \tag{5}$$

where the functions T_{\pm} are defined by

$$T_{\pm}(q(x) \pm x) = x \tag{6}$$

for all x , i.e., T_{\pm} is the inverse function to $q(x) \pm x$ (which exists and is C^1 so long as $|\dot{q}| < 1$), and

$$V(t, s) = \frac{1}{2}(V_0(s-t) + V_0(s+t)) + \frac{1}{2} \int_{s-t}^{s+t} V_1(y) dy. \quad (7)$$

Furthermore, $u(t, s)$ is at least C^1 away from the path $s = q(t)$.

Proof. Define $\Psi(t, s)$ and $\Phi(t, s)$ as solving the equations

$$\begin{cases} \partial_t^2 \Phi - \partial_s^2 \Phi = 0, \\ \Phi(0, s) = -\frac{a}{2}|s| + V_0(s), \\ \partial_t \Phi(0, s) = V_1(s), \end{cases} \quad \begin{cases} \partial_t^2 \Psi - \partial_s^2 \Psi = a\delta(s - q(t)), \\ \Psi(0, s) = 0, \\ \partial_t \Psi(0, s) = 0. \end{cases} \quad (8)$$

Furthermore, define $V(t, s)$ and $U(t, s)$ as solving the equations

$$\begin{cases} \partial_t^2 V - \partial_s^2 V = 0, \\ V(0, s) = V_0(s), \\ \partial_t V(0, s) = V_1(s), \end{cases} \quad \begin{cases} \partial_t^2 U - \partial_s^2 U = 0, \\ U(0, s) = -\frac{a}{2}|s|, \\ \partial_t U(0, s) = 0. \end{cases} \quad (9)$$

We thus have that $\Phi = U + V$ and $u = \Psi + \Phi$. Note that because V_0 and V_1 are smooth functions, $V(t, s)$ is smooth as well. Hence, $V(t, s)$ contains no singularities.

We can solve for V and U (and hence Φ) by using d'Alembert's formula. We then have

$$V(t, s) = \frac{1}{2}(V_0(s-t) + V_0(s+t)) + \frac{1}{2} \int_{s-t}^{s+t} V_1(y) dy, \quad (10)$$

$$U(t, s) = -\frac{a}{4}(|s-t| + |s+t|) = \begin{cases} \frac{a}{2}s, & s \leq -t, \\ -\frac{a}{2}t, & -t < s < t, \\ -\frac{a}{2}s, & s \geq t. \end{cases} \quad (11)$$

We can solve for Ψ with Duhamel's principle. Define $W(t, s, \tau)$ as

$$\Psi(t, s) = \int_0^t W(t - \tau, s, \tau) d\tau. \quad (12)$$

It follows that

$$\begin{cases} W_{tt} - W_{ss} = 0, \\ W(0, s, \tau) = 0, \\ W_t(0, s, \tau) = a\delta(s - q(\tau)). \end{cases} \quad (13)$$

To solve for W , we apply d'Alembert's formula. We have

$$W(t, s, \tau) = \frac{1}{2} \int_{s-t}^{s+t} a\delta(y - q(\tau)) dy = \frac{a}{2} \chi_{[s-t, s+t]}(q(\tau)), \quad (14)$$

where χ is the characteristic function, i.e.,

$$\chi_{[a, b]}(x) = \begin{cases} 1, & a \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases} \quad (15)$$

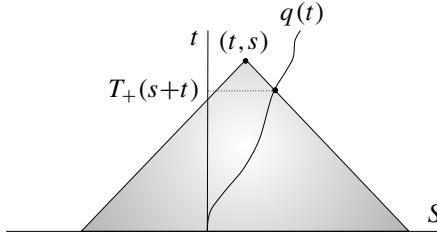


Figure 1. Retarded time T_+ .

To integrate W to get $\Psi(t, s)$, we consider the backward light cone of the event (t, s) . The τ for which $(\tau, q(\tau))$ is in this light cone will contribute $\frac{a}{2} d\tau$ to the integral. See [Figure 1](#).

Because $q(0) = 0$ and $c = 1$, we know $q(t)$ is inside the forward light cone drawn from $(0, 0)$. As a result of this, $\Psi(t, s) = 0$ when $s > t$ and $s < -t$. Inside the forward light cone of the origin, it is certainly true that the backward light cone of the event (t, s) will intersect with $s = q(t)$. Moreover, it intersects exactly once (going from time 0 to t , once $q(t)$ leaves the backward light cone of the event (t, s) , it cannot reenter due to the fact that $c = 1$). We must determine the point at which it intersects, the so-called *retarded time*. To the left of $q(t)$, this retarded time τ_2 is the solution to $q(\tau_2) + \tau_2 = s + t$, or $T_+(s + t)$ for short. The solution is hence $\frac{a}{2} T_+(s + t)$. To the right of $q(t)$, this retarded time τ_1 is the solution to $q(\tau_1) - \tau_1 = s - t$, or $T_-(s - t)$ for short. The solution is hence $\frac{a}{2} T_-(s - t)$.

We then get the following expression for Ψ :

$$\Psi(t, s) = \frac{a}{2} \begin{cases} 0, & s < -t, \\ T_+(s + t), & -t < s < q(t), \\ T_-(s - t), & q(t) < s < t, \\ 0, & s > t. \end{cases} \tag{16}$$

We see from here that Ψ , like U , is C^0 but not C^1 because it has two singularities at $s + t = 0$ and $s - t = 0$, i.e., along the light cone of the origin. We will see in [Proposition 4](#) that when we add Ψ and U , the singularities along the light cone cancel each other out.

The full solution $u(t, s)$ is thus

$$u(t, s) = \frac{a}{2} \begin{cases} s + V(t, s), & s < -t, \\ T_+(s + t) - t + V(t, s), & -t < s < q(t), \\ T_-(s - t) - t + V(t, s), & q(t) < s < t, \\ -s + V(t, s), & s > t, \end{cases} \tag{17}$$

where V is as defined in [\(10\)](#). □

We expect $u(t, s)$ to have singularities only at $s = q(t)$. We pause for a moment to show that u has no singularities along the lines $s = -t$ and $s = t$.

Proposition 4. For $u(t, s)$ given by (5), $u(t, s)$ is C^1 across $s = -t$ and $s = t$.

Proof. Because V_0 and V_1 are smooth, $V(t, s)$ is smooth. Hence, it suffices to look at the singular part of $u(t, s)$. Let

$$w(t, s) = U(t, s) + \Psi(t, s) = \frac{a}{2} \begin{cases} s, & s < -t, \\ T_+(s+t) - t, & -t < s < q(t), \\ T_-(s-t) - t, & q(t) < s < t, \\ -s, & s > t. \end{cases} \quad (18)$$

We will first show $w(t, s)$ is C^1 across $s = -t$. We have

$$\partial_s w|_{(s=-t)^-} = \partial_s \left(\frac{a}{2} s \right) = \frac{a}{2}, \quad \partial_t w|_{(s=-t)^-} = \partial_t \left(\frac{a}{2} s \right) = 0. \quad (19)$$

Recall that $T_+(s+t) = \tau_2$, where τ_2 solves

$$q(\tau_2) + \tau_2 = s + t. \quad (20)$$

Using implicit differentiation by s and t yields

$$\partial_s \tau_2 \dot{q}(\tau_2) + \partial_s \tau_2 = 1, \quad \partial_t \tau_2 \dot{q}(\tau_2) + \partial_t \tau_2 = 1 \quad (21)$$

respectively. At the line $s = -t$, we have $\tau_2 = 0$, so $\dot{q}(\tau_2) = \dot{q}(0) = 0$. Hence

$$\partial_s \tau_2 = \partial_t \tau_2 = 1. \quad (22)$$

We thus have

$$\partial_s w|_{(s=-t)^+} = \frac{a}{2} \partial_s T_+(s+t) = \frac{a}{2} \partial_s \tau_2 = \frac{a}{2}, \quad (23)$$

$$\partial_t w|_{(s=-t)^+} = \frac{a}{2} (\partial_t T_+(s+t) - 1) = \frac{a}{2} (\partial_t \tau_2 - 1) = 0. \quad (24)$$

Comparing with (19) shows that w is C^1 across $s = -t$.

The proof that $w(t, s)$ is C^1 across $s = t$ is completely analogous. \square

3. Computing the Kiessling force

We would like to combine the solution (5) for $u(t, s)$ with (3) to find a system of ODEs for $q(t)$. To do this, we need to work out the Kiessling force f in (3).

We begin by recalling that Kiessling *postulates* the local conservation of total energy-momentum for the field and particle system:

$$\partial^\mu T_{\mu\nu}^{\text{total}} = 0. \quad (25)$$

Here, $T_{\mu\nu}^{\text{total}}$ is the energy density-momentum density-stress tensor (or *energy-momentum tensor*, for short) for the field and particle system. To find the energy-momentum tensor for the field, we start with the Lagrangian. The Lagrangian for the massless scalar field is

$$\mathcal{L} = \frac{1}{2} \eta^{\mu\nu} \partial_\mu u \partial_\nu u. \quad (26)$$

Here, $\eta = \text{diag}(1, -1)$ is the Minkowski metric. The energy-momentum tensor for the field is defined as

$$T_{\mu\nu}^{\text{field}} = 2 \frac{\partial \mathcal{L}}{\partial \eta^{\mu\nu}} - \eta_{\mu\nu} \mathcal{L}, \quad (27)$$

and thus in this case

$$T_{\mu\nu}^{\text{field}} = \partial_\mu u \partial_\nu u - \frac{1}{2} \eta_{\mu\nu} \partial_\alpha u \partial^\alpha u. \quad (28)$$

Since u is expected to be singular on the worldline of the particle, the above is only well-defined away from the particle path, but our assumptions on the field are such that T^{field} can be continued into the particle path as a spacetime distribution.

The energy-momentum tensor for the particle on the other hand is defined as a distribution on spacetime that is concentrated on the worldline $x^\mu = z^\mu(\tau)$ of the particle (τ is the arclength parameter):

$$T_{\mu\nu}^{\text{particle}} := m \int \mathbf{u}_\mu \mathbf{u}_\nu \delta^{(2)}(x - z(\tau)) d\tau = \frac{m}{\mathbf{u}^0} \mathbf{u}_\mu \mathbf{u}_\nu \delta(s - q(t)), \quad (29)$$

where \mathbf{u} is the unit tangent to the worldline of the particle

$$\mathbf{u}^\mu := dz^\mu/d\tau, \quad \mathbf{u}_\mu \mathbf{u}^\mu = 1. \quad (30)$$

The definition of T^{particle} is such that

$$\partial^\mu T_{\mu\nu}^{\text{particle}} = \mathbf{f}_\nu(t) \delta(s - q(t)) \quad (31)$$

holds, where the spacetime covector \mathbf{f}_ν is the 2-force acting on the particle; see [Kiessling 2019, equation 72].

Let us take a second to consider how this relates to the $f(t, q, \dot{q})$, the force on the particle, which appears in (3). There, $f(t, q, \dot{q})$ is clearly the spatial component of a spacetime (contravariant) vector. We therefore must have

$$f(t, q(t), \dot{q}(t)) = \mathbf{f}^1(t) = -\mathbf{f}_1(t) \quad (32)$$

since we have chosen the signature $(+, -)$ for the Minkowski metric.

Hence, setting $\nu = 1$,

$$\partial^\mu T_{\mu 1}^{\text{particle}}(t, s) = -f(t, q(t), \dot{q}(t)) \delta(s - q(t)). \quad (33)$$

Going back to the energy-momentum tensor for the field, we have, for $\nu = 1$,

$$\partial^\mu T_{\mu 1}^{\text{field}} = \partial^0 T_{01}^{\text{field}} + \partial^1 T_{11}^{\text{field}} = \partial_t \pi - \partial_s \tau. \quad (34)$$

Using (25), we have

$$0 = \partial^\mu T_{\mu 1}^{\text{total}} = \partial^\mu T_{\mu 1}^{\text{field}} + \partial^\mu T_{\mu 1}^{\text{particle}} = \partial_t \pi - \partial_s \tau - f(t, q, \dot{q}) \delta(s - q(t)). \quad (35)$$

Rearranging this gives us the momentum-balance law

$$\partial_t \pi - \partial_s \tau = f(t, q, \dot{q}) \delta(s - q(t)). \quad (36)$$

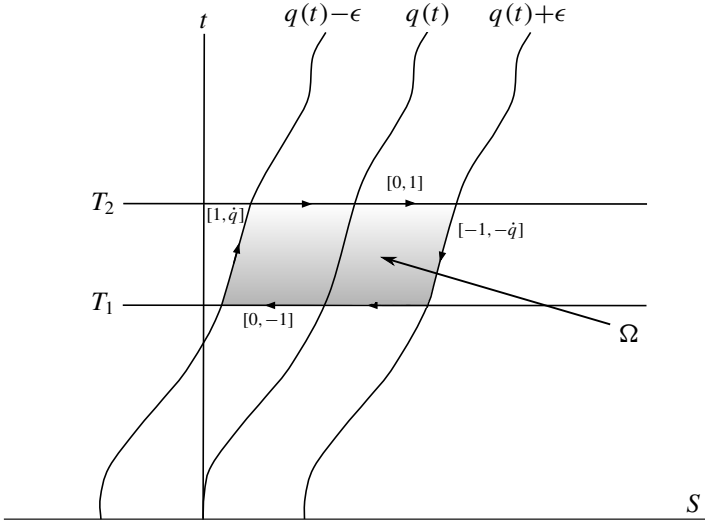


Figure 2. Region of integration for the momentum-balance equation.

From (28) we have

$$\pi(t, s) = T_{01}^{\text{field}} = u_s u_t, \quad (37)$$

$$\tau(t, s) = T_{11}^{\text{field}} = \frac{1}{2}(u_s^2 + u_t^2). \quad (38)$$

Proposition 5. *Assume that the field u is Lipschitz continuous in a tubular neighborhood of the C^1 path $(t, q(t))$ of the particle, and that u is C^1 on either side of the path. Then the force appearing in (36) is*

$$f(t, q(t), \dot{q}(t)) = -\dot{q}[\pi(t, s)]_{s=q(t)} - [\tau(t, s)]_{s=q(t)}, \quad (39)$$

where $[\cdot]_{s=q(t)}$ denotes the jump across the path.

Proof. Note that the assumptions imply that π and τ are bounded and can at most have a jump discontinuity across the path, so that the right-hand side of (39) is well-defined. Let $\epsilon > 0$ and $T_2 > T_1 \geq 0$. We will integrate (36) over the region $\Omega = \{(t, s) \mid T_1 \leq t \leq T_2, q(t) - \epsilon \leq s \leq q(t) + \epsilon\}$ and then take the limit as ϵ goes to 0. After integrating and taking the limit, the right-hand side of (36) becomes

$$\int_{T_1}^{T_2} f(t, q(t), \dot{q}(t)) dt. \quad (40)$$

After integrating and using Green's theorem, the left-hand side of (36) becomes

$$\begin{aligned} \int_{q(T_2)-\epsilon}^{q(T_2)+\epsilon} \pi(T_2, s) ds - \int_{q(T_1)-\epsilon}^{q(T_1)+\epsilon} \pi(T_1, s) ds - \int_{T_1}^{T_2} \dot{q} \pi(t, s) + \tau(t, s) \Big|_{s=q(t)+\epsilon} dt \\ + \int_{T_1}^{T_2} \dot{q} \pi(t, s) + \tau(t, s) \Big|_{s=q(t)-\epsilon} dt. \end{aligned}$$

Because π is locally integrable, the first two terms go to 0 as $\epsilon \rightarrow 0$. Taking the limit as ϵ goes to 0 in the other two terms gives us

$$-\int_{T_1}^{T_2} [\dot{q}\pi(t, s) + \tau(t, s)]_{s=q(t)} dt \quad (41)$$

for the left-hand side. Because T_1 and T_2 were arbitrary, we therefore have

$$f(t, q(t), \dot{q}(t)) = -\dot{q}[\pi(t, s)]_{s=q(t)} - [\tau(t, s)]_{s=q(t)}, \quad (42)$$

completing the proof. \square

Proposition 6. *Assume $u(t, s)$ is a solution to the joint evolution problem (2)–(3). The Kiessling force in (3) is given by*

$$f(t, q, \dot{q}) = aV_s(t, q) - \frac{a^2}{2} \frac{\dot{q}}{1 - \dot{q}^2}. \quad (43)$$

Proof. By Proposition 5, the force is

$$f(t, q(t), \dot{q}(t)) = -\dot{q}[\pi]_{s=q(t)} - [\tau]_{s=q(t)} = -\dot{q}[u_s u_t]_{s=q(t)} - \frac{1}{2}[u_s^2 + u_t^2]_{s=q(t)}, \quad (44)$$

using $u(t, s)$ as given by (5). Substituting in $u = V + w$ gives us

$$[u_t u_s]_{s=q(t)} = V_t[w_s]_{s=q(t)} + V_s[w_t]_{s=q(t)} + [w_s w_t]_{s=q(t)}, \quad (45)$$

$$\left[\frac{1}{2}u_t^2 + \frac{1}{2}u_s^2\right]_{s=q(t)} = \frac{1}{2}[w_t^2]_{s=q(t)} + \frac{1}{2}[w_s^2]_{s=q(t)} + V_s[w_s]_{s=q(t)} + V_t[w_t]_{s=q(t)}, \quad (46)$$

where we've used the fact that V is smooth. To determine the necessary values, we will first compute $w_s|_{s=q(t)^-}$, $w_s|_{s=q(t)^+}$, $w_t|_{s=q(t)^-}$, and $w_t|_{s=q(t)^+}$. We have

$$w_s|_{s=q(t)^-} = \partial_s \left(\frac{a}{2}(T_+(s+t) - t) \right) = \frac{a}{2} \frac{1}{\dot{q}(t) + 1}, \quad (47)$$

$$w_s|_{s=q(t)^+} = \partial_s \left(\frac{a}{2}(T_-(s-t) - t) \right) = \frac{a}{2} \frac{1}{\dot{q}(t) - 1}, \quad (48)$$

$$w_t|_{s=q(t)^-} = \partial_t \left(\frac{a}{2}(T_+(s+t) - t) \right) = -\frac{a}{2} \frac{\dot{q}(t)}{\dot{q}(t) + 1}, \quad (49)$$

$$w_t|_{s=q(t)^+} = \partial_t \left(\frac{a}{2}(T_-(s-t) - t) \right) = -\frac{a}{2} \frac{\dot{q}(t)}{\dot{q}(t) - 1}. \quad (50)$$

Thus, our final results for $[w_s]_{s=q(t)}$, $[w_t]_{s=q(t)}$, $[w_s w_t]_{s=q(t)}$, $[w_s^2]_{s=q(t)}$, and $[w_t^2]_{s=q(t)}$ are

$$[w_s]_{s=q(t)} = \frac{a}{2} \left(\frac{1}{\dot{q}(t) - 1} - \frac{1}{\dot{q}(t) + 1} \right) = \frac{a}{\dot{q}(t)^2 - 1}, \quad (51)$$

$$[w_t]_{s=q(t)} = -\frac{a\dot{q}(t)}{2} \left(\frac{1}{\dot{q}(t) - 1} - \frac{1}{\dot{q}(t) + 1} \right) = -\frac{a\dot{q}(t)}{\dot{q}(t)^2 - 1}, \quad (52)$$

$$[w_s w_t]_{s=q(t)} = -\frac{a^2 \dot{q}(t)}{4} \left(\frac{1}{(\dot{q}(t) - 1)^2} - \frac{1}{(\dot{q}(t) + 1)^2} \right) = -\frac{a^2 \dot{q}(t)^2}{(\dot{q}(t)^2 - 1)^2}, \quad (53)$$

$$[w_t^2]_{s=q(t)} = \frac{a^2}{4} \left(\left(\frac{\dot{q}(t)}{\dot{q}(t) - 1} \right)^2 - \left(\frac{\dot{q}(t)}{\dot{q}(t) + 1} \right)^2 \right) = \frac{a^2 \dot{q}(t)^3}{(\dot{q}(t)^2 - 1)^2}, \quad (54)$$

$$[w_s^2]_{s=q(t)} = \frac{a^2}{4} \left(\left(\frac{1}{\dot{q}(t) - 1} \right)^2 - \left(\frac{1}{\dot{q}(t) + 1} \right)^2 \right) = \frac{a^2 \dot{q}(t)}{(\dot{q}(t)^2 - 1)^2}. \quad (55)$$

Inserting these values into (44) gives us

$$f(t, q(t), \dot{q}(t)) = aV_s(t, q(t)) - \frac{a^2}{2} \frac{\dot{q}(t)}{1 - \dot{q}(t)^2}, \quad (56)$$

completing the proof. \square

Note that the first term represents the force that the external field is exerting on the particle. That is, the first term is usually taken to be the force acting on a scalar particle. The second term represents the *self-force* (the force the particle exerts on itself), which here is in the opposite direction of the motion.

4. Equations of motion as a dynamical system

We can now look at the equations of motion for the particle, which are

$$\begin{cases} \dot{q} = p/(m\sqrt{1 + p^2/m^2}), \\ \dot{p} = aV_s(t, q(t)) - \frac{1}{2}a^2\dot{q}(t)/(1 - \dot{q}(t)^2). \end{cases} \quad (57)$$

We substitute the expression for \dot{q} into the equation of \dot{p} , which results in

$$\dot{p} = aV_s(t, q(t)) - \frac{a^2}{2} \frac{p}{m} \sqrt{1 + \frac{p^2}{m^2}}. \quad (58)$$

In addition to this, let us rewrite $V_s(t, q(t))$. Recall that

$$V(t, s) = \frac{1}{2}(V_0(s+t) + V_0(s-t)) + \frac{1}{2} \int_{s-t}^{s+t} V_1(x) dx. \quad (59)$$

Hence, we have

$$V_s(t, s) = \frac{1}{2}(\dot{V}_0(s+t) + \dot{V}_0(s-t) + V_1(s+t) - V_1(s-t)). \quad (60)$$

Let us further define $F(s)$ and $G(s)$ as

$$\begin{cases} F(s) = \dot{V}_0(s) + V_1(s), \\ G(s) = \dot{V}_0(s) - V_1(s). \end{cases} \quad (61)$$

From our definitions of F and G , we can rewrite our equations of motion, specifically the expression for \dot{p} . It now becomes

$$\begin{cases} \dot{q} = p/(m\sqrt{1 + p^2/m^2}), \\ \dot{p} = \frac{1}{2}a(F(q+t) + G(q-t)) - \frac{1}{2}a^2(p/m)\sqrt{1 + p^2/m^2}. \end{cases} \quad (62)$$

We can further simplify our equations using a change of variables so that we can get rid of the square roots. We will let $p/m = \tan \theta$, so our new equations become

$$\begin{cases} \dot{q} = \sin \theta, \\ \dot{\theta} = \frac{a}{2m}(F(q+t) + G(q-t)) \cos^2 \theta - \frac{a^2}{2m} \sin \theta. \end{cases} \quad (63)$$

To get an autonomous system, we define new unknowns,

$$\begin{cases} d(t) = q(t) + t, \\ b(t) = q(t) - t, \end{cases} \quad (64)$$

and write the system of three equations as

$$\begin{cases} \dot{d} = \sin \theta + 1, \\ \dot{b} = \sin \theta - 1, \\ \dot{\theta} = \frac{a}{2m}(F(d) + G(b)) \cos^2 \theta - \frac{a^2}{2m} \sin \theta. \end{cases} \quad (65)$$

To solve this, we need to look for a solution (b, d, θ) such that

$$\begin{cases} b(0) = 0, \\ d(0) = 0, \\ \theta(0) = 0. \end{cases} \quad (66)$$

These are the consequences of our initial conditions $q(0) = 0$ and $\dot{q}(0) = 0$. Furthermore, we know the following limits for each variable: $0 < d(t) < 2t$, $-2t < b(t) < 0$, $-\frac{\pi}{2} \leq \theta(t) \leq \frac{\pi}{2}$.

We will split this into two cases: one where the bare mass, m , is positive and one where it is negative. For the first, we will prove that the solution will always be stable. For the second, we will show a case where the solution is unstable.

5. Proof of stability for positive bare mass

In the case of positive bare mass, we are concerned with (65) with $m > 0$. Recall that because $F(d)$ and $G(b)$ are defined as in (61), they must be compactly supported. Furthermore, note that it must always be the case that $\dot{d} = \sin \theta + 1 \geq 0$ and that $\dot{b} = \sin \theta - 1 \leq 0$. Hence, it suffices to look only at the region where $b \leq 0$ and $d \geq 0$. We will define $[b_L, b_R]$ such that $-\infty < b_L < b_R \leq 0$ and $G(b) = 0$ for b outside $[b_L, b_R]$. Similarly, we will define $[d_L, d_R]$ such that $0 \leq d_L < d_R < \infty$ and $F(d) = 0$ for d outside $[d_L, d_R]$. Based on the fact that $\dot{d} \geq 0$ and $\dot{b} \leq 0$, Figure 3 shows a rough sketch of the trajectory of the solution projected onto the (b, d) -plane.

Based on this, we will divide this analysis into three regions: before the radiation, during the radiation, and after the radiation. In the first region, $G(b) = F(d) = 0$, so (65) reduces to

$$\begin{cases} \dot{d} = \sin \theta + 1, \\ \dot{b} = \sin \theta - 1, \\ \dot{\theta} = -\frac{a^2}{2m} \sin \theta. \end{cases} \quad (67)$$

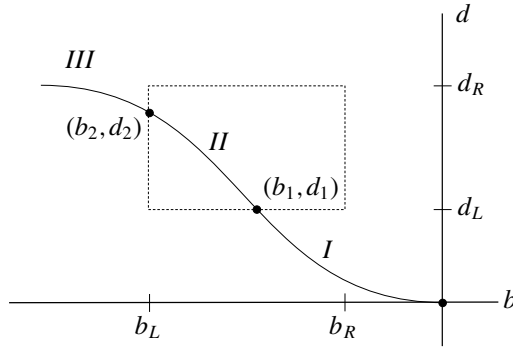


Figure 3. Projection of trajectory of the solution to (65) in the (b, d) -plane.

One can check that the unique solution to (67) with initial conditions in (66) is

$$\begin{cases} d(t) = t, \\ b(t) = -t, \\ \theta(t) = 0. \end{cases} \tag{68}$$

Hence, the particle will have the following conditions entering the second region:

$$\begin{cases} b(t_1) = b_1, \\ d(t_1) = d_1, \\ \theta(t_1) = 0, \end{cases} \tag{69}$$

where $t_1 > 0$. In the second region, note that when $\theta = \frac{\pi}{2}$, we have $\dot{\theta} = -a^2/(2m) < 0$. Similarly, when $\theta = -\frac{\pi}{2}$, we have $\dot{\theta} = a^2/(2m) > 0$. Hence, the trajectory can never cross $\theta = \frac{\pi}{2}$ and $\theta = -\frac{\pi}{2}$. The particle will thus end up with the following conditions entering the third region:

$$\begin{cases} b(t_2) = b_2, \\ d(t_2) = d_2, \\ \theta(t_2) = \theta_2, \end{cases} \tag{70}$$

where $t_2 > 0$ and $-\frac{\pi}{2} < \theta_2 < \frac{\pi}{2}$. Hence, the third interval will amount to solving the following set of ODEs with conditions specified in (70):

$$\begin{cases} \dot{b} = \sin \theta - 1, \\ \dot{d} = \sin \theta + 1, \\ \dot{\theta} = -\frac{a^2}{2m} \sin \theta. \end{cases} \tag{71}$$

We solve the third ODE explicitly in Proposition 7.

Proposition 7. Suppose we have (71) with initial conditions in (70).

- (a) If $\theta_2 = 0$, $\theta(t) = 0$ for $t > t_2$.
- (b) If $\frac{\pi}{2} > \theta_2 > -\frac{\pi}{2}$ and $\theta_2 \neq 0$, then $\lim_{t \rightarrow \infty} \theta(t) = 0$.

Proof. We know $\theta(t) = 0$ is a trivial solution which satisfies $\theta(t_2) = 0$. Noting that $-a^2 \sin \theta / (2m)$ and its derivative with respect to θ are continuous everywhere, we see that such a solution is unique. Hence, (a) follows.

For (b), assume $\theta_2 > 0$. The proof is similar for $\theta_2 < 0$. If $\theta = 0$ for some time $t_3 > t_2$, we are left with (a), and $\theta(t) = 0 < \epsilon$ for $t > t_3$. Hence, assume $\theta > 0$ at all times. Then $\sin \theta \neq 0$, and we can separate the third equation of (71):

$$\frac{1}{\sin \theta} d\theta = -\frac{a^2}{2m} dt. \quad (72)$$

Integrating, we have

$$-\ln|\csc(\theta) + \cot(\theta)| = -\frac{a^2}{2m}t + C_0 \quad (73)$$

or

$$\csc(\theta) + \cot(\theta) = C_1 e^{a^2 t / (2m)}. \quad (74)$$

We get rid of the absolute value by choosing the sign for C_1 . Here, $C_1 > 0$ since $\theta(t_2) > 0$ implies $\csc(\theta) + \cot(\theta) > 0$. For any $\epsilon > 0$, we can choose t_3 such that $\csc(\epsilon) + \cot(\epsilon) = C_1 e^{a^2 t_3 / (2m)}$. Then, for $t > t_3$,

$$\csc(\theta(t)) + \cot(\theta(t)) = C_1 e^{a^2 t / (2m)} > C_1 e^{a^2 t_3 / (2m)} = \csc(\epsilon) + \cot(\epsilon).$$

Thus, $\theta(t) < \epsilon$. □

With this, we have proved the stability of the solution for positive bare mass. In the case where $\theta_2 = 0$, we see that $\theta(t) = 0$ for $t > t_2$. Recalling that $\dot{q} = \sin \theta$, we have $\dot{q}(t) = 0$ for $t > t_2$. In the case where $\theta_2 \neq 0$, we see that since $\lim_{t \rightarrow \infty} \theta(t) = 0$ and $\dot{q} = \sin \theta$, we have $\lim_{t \rightarrow \infty} \dot{q}(t) = 0$. In either case, the speed of the particle goes to zero.

6. Proof of instability for negative bare mass

In the case of negative bare mass, we are concerned with (65) with $m < 0$. We will show that a specific case of radiation leads to an unstable solution. Assume that the radiation is purely incoming: $F \equiv 0$. Set $G_\beta(x) = \beta \sin(\pi x) \chi_{[-3, -1]}$, where $\beta \in \mathbb{R}$. We can ignore the d equation and are left with a system of two equations:

$$\begin{cases} \dot{b} = \sin \theta - 1, \\ \dot{\theta} = \frac{a}{2m} (G_\beta(b)) \cos^2 \theta - \frac{a^2}{2m} \sin \theta, \end{cases} \quad (75)$$

$$\begin{cases} b(0) = 0, \\ \theta(0) = 0. \end{cases} \quad (76)$$

We will rewrite (75) by replacing m with $-|m|$:

$$\begin{cases} \dot{b} = \sin \theta - 1, \\ \dot{\theta} = -\frac{a}{2|m|} G_\beta(b) \cos^2 \theta + \frac{a^2}{2|m|} \sin \theta, \end{cases} \quad (77)$$

$$\begin{cases} b(0) = 0, \\ \theta(0) = 0. \end{cases} \quad (78)$$

In addition, it will be useful to work with the reverse flow of the system:

$$\begin{cases} \dot{b} = -\sin \theta + 1, \\ \dot{\theta} = \frac{a \cos^2 \theta}{2|m|} (G_\beta(b)) - \frac{a^2}{2|m|} \sin \theta. \end{cases} \quad (79)$$

Note that $\beta = 0$ corresponds to the static solution. To see this, note that (77) becomes

$$\begin{cases} \dot{b} = \sin \theta - 1, \\ \dot{\theta} = \frac{a^2}{2|m|} \sin \theta. \end{cases} \quad (80)$$

The solution to this with initial conditions given in (78) is

$$\begin{cases} b(t) = -t, \\ \theta(t) = 0. \end{cases} \quad (81)$$

Since $\theta = 0$ corresponds to $\dot{q} = 0$, (81) is the static solution.

To get a better sense of the system of ODEs in (77) for nonzero β , see Figure 4. The interval $[-3, -1]$ represents the particle becoming perturbed by the incoming radiation. A particle with initial conditions given in (78) will end up with $b = -1$ and $\theta = 0$. To see this, note that outside of $[-3, -1]$, $G_\beta(b) = 0$, and (77) reduces to (80). Hence, in the interval $[-1, 0]$, the solution to (77) with initial conditions given in (78) is (81). At $t_1 = 1$, $b(t_1) = -1$ and $\theta(t_1) = 0$. At this point, we can take the system of ODEs in (78) and modify the initial conditions as follows:

$$\begin{cases} b(t_1) = -1, \\ \theta(t_1) = 0. \end{cases} \quad (82)$$

After entering the interval $[-3, -1]$, Figure 4 suggests that the particle will oscillate. The question is whether or not the particle will go back to rest ($\theta = 0$ at $b = -3$, represented by the gray line in Figure 4) or be left with some speed ($\theta \neq 0$ at $b = -3$). In the former case, the particle will remain at rest. In the latter case, the particle will go toward $\theta = \pm \frac{\pi}{2}$. Recalling that $\dot{q} = \sin \theta$, this means $\dot{q} = \pm 1$. That is, the particle will reach the speed of light in finite time. This is proved formally in the next proposition.

Proposition 8. *Suppose, for some $t_0 > 0$, we have $b(t_0) = -3$.*

- (i) *If $\theta_0 = \theta(t_0) = 0$, then $\theta(t) = 0$ for $t > t_0$.*
- (ii) *If $\frac{\pi}{2} > \theta(t_0) > 0$, then $\theta(t_1) = \frac{\pi}{2}$ for some $t_1 > t_0$.*
- (iii) *If $-\frac{\pi}{2} < \theta(t_0) < 0$, then $\theta(t_1) = -\frac{\pi}{2}$ for some $t_1 > t_0$.*

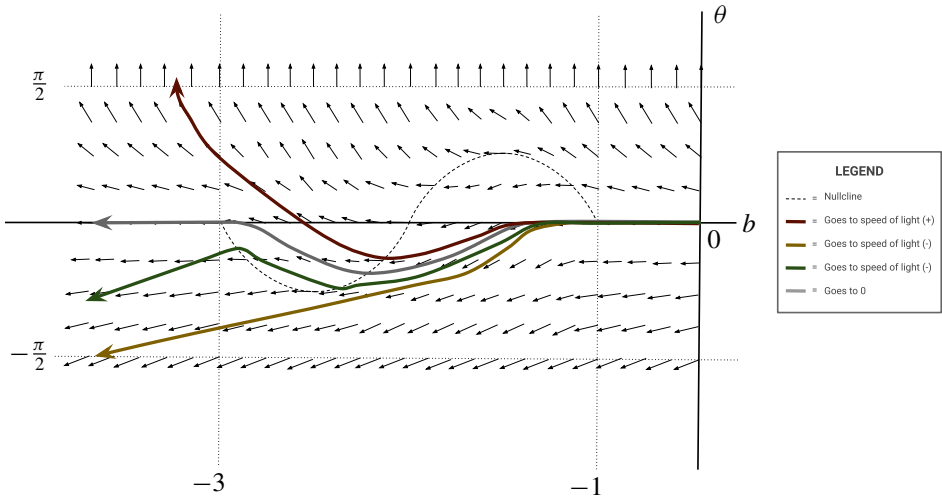


Figure 4. Hypothetical solutions to the system of ODEs.

Proof. Because $\dot{b} = \sin \theta - 1$, it is always true that $\dot{b} \leq 0$. Hence, for $t > t_0$, we have $b < -3$, and $G(b) = 0$. We can then rewrite the ODEs in (77) as (80). We know $\theta(0) = 0$ is a trivial solution and satisfies $\theta(t_0) = 0$. Since such a solution is unique, (a) follows.

We will now show (b). Because $\theta(t_0) > 0$ and $\dot{\theta} > 0$ if $\theta > 0$, it follows that $\theta(t_1) > 0$ for $t_1 > t_0$. Hence, $\sin \theta \neq 0$, and we can separate the second equation of (80):

$$\frac{1}{\sin \theta} d\theta = \frac{a^2}{2|m|} dt. \tag{83}$$

Integrating, we have

$$-\ln|\csc(\theta) + \cot(\theta)| = \frac{a^2}{2|m|} t + C_0 \tag{84}$$

or

$$\csc(\theta) + \cot(\theta) = C_1 e^{-a^2 t / (2|m|)}. \tag{85}$$

We get rid of the absolute value by choosing the sign for C_1 . Here, $C_1 > 0$ since $\theta(t_0) > 0$ implies $\csc(\theta) + \cot(\theta) > 0$. Now, for $0 \leq \theta \leq \pi$, we know $0 \leq \csc(\theta) + \cot(\theta) < \infty$ is monotonously decreasing and is invertible. Let $\Theta : [0, \pi] \rightarrow [0, \infty)$ be the inverse of $\csc(\theta) + \cot(\theta)$. We can now write the solution explicitly as

$$\theta(t) = \Theta(C_1 e^{-a^2 t / (2|m|)}). \tag{86}$$

The initial conditions tell us

$$C_1 = \frac{\csc(\theta(t_0)) + \cot(\theta(t_0))}{e^{-a^2 t_0 / (2|m|)}} = \frac{C_2}{e^{-a^2 t_0 / (2|m|)}}, \tag{87}$$

where $C_2 = \csc(\theta(t_0)) + \cot(\theta(t_0))$. Because $0 \leq \theta(t_0) < \frac{\pi}{2}$, we have $C_2 > 1$. Substituting C_1 gives us

$$\theta(t) = \Theta(C_2 e^{-a^2(t-t_0)/(2|m|)}). \quad (88)$$

Consider

$$t_1 = -\frac{2|m|}{a^2} \ln \frac{1}{C_2} + t_0 > t_0.$$

Using the fact that $\Theta(1) = \frac{\pi}{2}$ shows us that $\theta(t_1) = \frac{\pi}{2}$.

For part (c), repeat the proof for part (b), except $C_1 < 0$, and we define $\Theta : [-\pi, 0] \rightarrow (-\infty, 0]$ instead. \square

We show instability by proving the existence of a solution to (77) which satisfies the conditions of (c). To do this, we first work with the backward-flow defined in (79) and make a change of variables.

Proposition 9. *Assume $G(x) = \beta \sin(\pi x)\chi_{[-3, -1]}$. There exists an $\epsilon > 0$ such that $0 < \beta < \epsilon$ implies that the system of ODEs in (77) with initial conditions at t_1 ,*

$$\begin{cases} b(t_1) = -1, \\ \theta(t_1) = 0 \end{cases} \quad (89)$$

has a unique solution $(b(t), \theta(t))$ such that at some $t_2 > t_1$

$$\begin{cases} b(t_2) = -3, \\ \theta(t_2) < 0. \end{cases} \quad (90)$$

Proof. To start, consider the backward flow (79) with initial conditions

$$\begin{cases} b(t_1) = -3, \\ \theta(t_1) = 0 \end{cases} \quad (91)$$

and make the change of variables

$$\begin{cases} y = \theta, \\ x = b + 2. \end{cases} \quad (92)$$

Using the fact that

$$\frac{dy}{dx} = \frac{\dot{\theta}}{\dot{b}}$$

gives us the ODE

$$\begin{cases} \frac{dy_\beta}{dx} = \frac{a}{2|m|} (1 + \sin y_\beta)(\beta \sin(\pi x) - a \sec y_\beta \tan y_\beta), \\ y_\beta(-1) = 0. \end{cases} \quad (93)$$

By Lemma 10 below, there exists an $\epsilon > 0$ such that $0 < \beta < \epsilon$ implies $y_\beta(1) > 0$. Because of the uniqueness of solutions for first-order ODEs, a β satisfying the previous statement implies that the solution for the forward-flow with initial conditions specified in (89) ends up below 0. This can be intuitively seen in Figure 5. \square

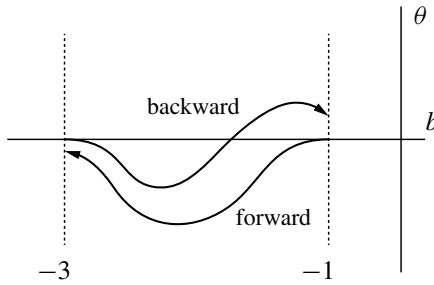


Figure 5. Relation between forward and backward solution.

Lemma 10. Assume that $a > 0$. Suppose $y_\beta : \mathbb{R} \rightarrow \mathbb{R}$ is a function satisfying

$$\begin{cases} \frac{dy_\beta}{dx} = \frac{a}{2|m|}(1 + \sin y_\beta)(\beta \sin(\pi x) - a \sec y_\beta \tan y_\beta), \\ y_\beta(-1) = 0. \end{cases} \tag{94}$$

There exists an $\epsilon > 0$ such that $0 < \beta < \epsilon$ implies $y_\beta(1) > 0$.

Proof. Consider $y(x, \beta) = y_\beta(x)$ and let $Z = \partial y / \partial \beta$. We can rewrite (94) as

$$\begin{cases} \frac{\partial y}{\partial x} = \frac{a}{2|m|}(1 + \sin y)(\beta \sin(\pi x) - a \sec y \tan y), \\ y(-1, \beta) = 0, \\ Z(-1, \beta) = 0. \end{cases} \tag{95}$$

Because $y(1, 0) = 0$, it suffices to show that $Z(1, 0) = (\partial y / \partial \beta)(1, 0) > 0$. Using

$$\frac{\partial Z}{\partial x} = \frac{\partial^2 y}{\partial x \partial \beta} = \frac{\partial^2 y}{\partial \beta \partial x},$$

and substituting $\beta = 0$ (meaning $y = 0$), we arrive at the following linear differential equation for Z :

$$\begin{cases} \frac{\partial Z}{\partial x}(x, 0) = \frac{a}{2|m|} \sin(\pi x) - \frac{a^2}{2|m|} Z, \\ Z(-1, 0) = 0. \end{cases} \tag{96}$$

The solution to this is simply

$$Z(x, 0) = \frac{a}{2|m|} e^{-a^2 x / (2|m|)} \int_{-1}^x \sin(\pi t) e^{a^2 t / (2|m|)} dt.$$

We have

$$Z(1, 0) = \frac{a}{2|m|} e^{-a^2 / (2|m|)} \int_{-1}^1 \sin(\pi t) e^{a^2 t / (2|m|)} dt = \frac{2\pi a |m| (1 - e^{-a^2 / |m|})}{4\pi^2 m^2 + a^4}. \tag{97}$$

Because a is assumed to be positive, we have $Z(1, 0) > 0$ as needed. □

7. Summary and outlook

We have shown that the static solution to this problem, where the particle remains at rest forever, is stable for particles with positive bare mass. However, for particles with negative bare mass, the static solution is highly unstable. That is, a small amount of radiation can cause the particle to accelerate to the speed of light in finite time. Though this result is not intuitive, it is also not very surprising when considering the model we used. In the initial conditions for the wave equation, we took

$$u(0, s) = -\frac{a}{2}|s| + V_0(s). \quad (98)$$

Recall that the field energy density is $\epsilon(t, s) = T_{00}^{\text{field}} = \frac{1}{2}(u_t^2 + u_s^2)$. Therefore this initial condition has an infinite amount of energy:

$$\int_{-\infty}^{\infty} \epsilon(s, 0) ds = \int_{-\infty}^{\infty} \frac{a^2}{4} + \dot{V}_0^2 ds = \infty. \quad (99)$$

Since the total energy of the system is conserved, there is an infinite amount of energy available that can be transferred to the particle, allowing it to accelerate to the speed of light. But another reason this can happen in finite time is that in this model the Kiessling force $f(t, q, \dot{q})$ itself diverges as $|\dot{q}| \rightarrow 1$.

Hence, looking forward, we would like to examine different models for the joint evolution, in which such problems are not present. In one such model, the scalar field would be governed by the Klein–Gordon equation rather than the wave equation:

$$\begin{cases} \partial_t^2 u - \partial_s^2 u + \mu^2 u = a\delta(s - q(t)), \\ u(0, s) = \frac{a}{2\mu} e^{-\mu|s|} + v_0(s), \\ \partial_t u(0, s) = v_1(s). \end{cases} \quad (100)$$

This would add a mass term to the field equations and change the part of the initial conditions that represents the static solution. In this model, the particle would start with a finite amount of energy. It would be interesting to see if a particle with negative bare mass still accelerates to the speed of light. We are currently investigating this.

Another model to consider is one in which the field equations are *fully* relativistic. The wave operator appearing in (2) is of course relativistic, but the delta source on the right-hand side of the equation is not manifestly so. It turns out that it is possible to modify this right-hand side so that the equation itself becomes fully relativistic. It is possible to show that for this modified equation for a massless scalar field, the Kiessling force will not diverge if the particle velocity reaches the speed of light, and stability of the static solution is restored. This result will appear elsewhere [Frolov et al. 2023].

Additionally, we would like to explore what would happen with two particles instead of one. Mathematically, this would involve the sum of two Dirac delta functions as the source of the wave equation. This may necessitate the use of differential-delay equations rather than simple ODEs, which would require much more intricate analysis.

Acknowledgements

We thank Vu Hoang and Maria Radosz for helping us correct the formula for the particle self-force, and Lawrence Frolov for reading through the paper and providing helpful comments. We are grateful to the anonymous referee for many helpful suggestions and comments.

References

- [Bopp 1940] F. Bopp, “Eine lineare Theorie des Elektrons”, *Ann. Physik* (5) **38** (1940), 345–384. [MR](#) [Zbl](#)
- [Bopp 1942] F. Bopp, “Lineare Theorie des Elektrons, II”, *Ann. Physik* (5) **42** (1942), 573–608. [MR](#) [Zbl](#)
- [Dirac 1938] P. A. M. Dirac, “Classical theory of radiating electrons”, *Proc. Roy. Soc. London Ser. A* **167**:929 (1938), 148–169. [MR](#) [Zbl](#)
- [Elskens et al. 2009] Y. Elkens, M. K.-H. Kiessling, and V. Ricci, “The Vlasov limit for a system of particles which interact with a wave field”, *Comm. Math. Phys.* **285**:2 (2009), 673–712. [MR](#) [Zbl](#)
- [Frolov et al. 2023] L. Frolov, S. Leigh, and A. S. Tahvildar-Zadeh, “On the joint evolution of fields and particles in one space dimension”, preprint, 2023. [arXiv 2312.06019](#)
- [Hoang et al. 2021] V. Hoang, M. Radosz, A. Harb, A. DeLeon, and A. Baza, “Radiation reaction in higher-order electrodynamics”, *J. Math. Phys.* **62**:7 (2021), art. id. 072901. [MR](#) [Zbl](#)
- [Kiessling 2019] M. K.-H. Kiessling, “Force on a point charge source of the classical electromagnetic field”, *Phys. Rev. D* **100**:6 (2019), art. id. 065012. Correction in **101**:10 (2020), art. id. 109901. [MR](#)
- [Kiessling and Tahvildar-Zadeh \geq 2024] M. K. H. Kiessling and A. S. Tahvildar-Zadeh, “Bopp–Landé–Thomas–Podolsky electrodynamics as initial value problem”, in preparation.
- [Landé 1941] A. Landé, “Finite self-energies in radiation theory, I”, *Phys. Rev.* **60**:2 (1941), 121–127.
- [Landé and Thomas 1941] A. Landé and L. H. Thomas, “Finite self-energies in radiation theory, II”, *Phys. Rev.* **60**:7 (1941), 514–523.
- [Podolsky 1942] B. Podolsky, “A generalized electrodynamics, I: Non-quantum”, *Phys. Rev.* (2) **62**:1-2 (1942), 68–71. [MR](#)
- [Poincaré 1906] H. Poincaré, “Sur la dynamique de l’électron”, *Rend. Circ. Mat. Palermo* **21** (1906), 129–175. [Zbl](#)
- [Spohn 2004] H. Spohn, *Dynamics of charged particles and their radiation field*, Cambridge University Press, 2004. [MR](#) [Zbl](#)
- [Weyl 1921] H. Weyl, “Feld und Materie”, *Ann. der Phys.* (4) **65** (1921), 541–563. [Zbl](#)

Received: 2022-10-31

Revised: 2023-01-19

Accepted: 2023-01-21

aditya_agashe@brown.edu

Brown University, Providence, RI, United States

esl75@scarletmail.rutgers.edu

Rutgers University, Piscataway, NJ, United States

shadit@math.rutgers.edu

Rutgers University, Piscataway, NJ, United States

Guidelines for Authors

Submissions in all mathematical areas are encouraged. All manuscripts accepted for publication in *Involve* are considered publishable in quality journals in their respective fields, and include a minimum of one-third student authorship. Submissions should include substantial faculty input; faculty co-authorship is strongly encouraged. Authors may submit manuscripts in PDF format on-line at the Submission page at the [Involve website](#).

Originality. Submission of a manuscript acknowledges that the manuscript is original and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles in *Involve* are usually in English, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

Format. Authors are encouraged to use L^AT_EX but submissions in other varieties of T_EX, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibT_EX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@msp.org with details about how your graphics were generated.

White space. Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Proofs. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

involve

2024

vol. 17

no. 1

An r^p -weighted local energy approach to global existence for null form semilinear wave equations	1
MICHAEL FACCI, ALEX MCENTARRFER AND JASON METCALFE	
Cones and ping-pong in three dimensions	11
GABRIEL FRIEDEN, FÉLIX GÉLINAS AND ÉTIENNE SOUCY	
Euclidean and affine curve reconstruction	29
JOSE AGUDELO, BROOKE DIPPOLD, IAN KLEIN, ALEX KOKOT, ERIC GEIGER AND IRINA KOGAN	
Biological models, monotonicity methods, and solving a discrete reaction-diffusion equation	65
CARSON RODRIGUEZ AND STEPHEN B. ROBINSON	
Edge-determining sets and determining index	85
SALLY COCKBURN AND SEAN MCAVOY	
The adjacency spectra of some families of minimally connected prime graphs	107
CHRIS FLOREZ, JONATHAN HIGGINS, KYLE HUANG, THOMAS MICHAEL KELLER AND DAWEI SHEN	
Linear maps preserving the Lorentz spectrum of 3×3 matrices	121
MARIA I. BUENO, BEN FAKTOR, RHEA KOMMERELL, RUNZE LI AND JOEY VELTRI	
Lattice size in higher dimensions	153
ABDULRAHMAN ALAJMI, SAYOK CHAKRAVARTY, ZACHARY KAPLAN AND JENYA SOPRUNOVA	
On the joint evolution problem for a scalar field and its singularity	163
ADITYA AGASHE, ETHAN LEE AND SHADI TAHVILDAR-ZADEH	