involve

a journal of mathematics

msp

2018 vol. 11, no. 4

involve

msp.org/involve

INVOLVE YOUR STUDENTS IN RESEARCH

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

Cover: Alex Scorpan

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2018 is US $190/year for the electronic version, and $250/year (+$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP.

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

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

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

INVOLVE 11:4 (2018)
dx.doi.org/10.2140/involve.2018.11.541

Modeling of breast cancer through evolutionary game theory

Ke'Yona Barton, Corbin Smith, Jan Rychtář and Tsvetanka Sendova

(Communicated by Kenneth S. Berenhaut)

We present a simple mathematical model of the development and progression of breast cancer based on evolutionary game theory. Four types of cellular populations are considered: stromal (native) cells, macrophages, benign tumor cells, and motile (malignant) tumor cells. Despite the relative simplicity of the model, it provides a way to explore the interactions between the various cell types and suggests potential approaches to managing and treating cancer.

1. Introduction

The third most common cancer in the world is breast cancer, succeeding lung and stomach cancer [Ford et al. 1998]. In women worldwide it is the leading cancer and there are more than 10^6 new cases each year. There are many genes associated with an increased probability of a person developing breast cancer, more commonly known amongst which are the BRCA1 and BRCA2 genes [Ford et al. 1998; Slamon et al. 1987].

There has been a substantial amount of research which makes use of mathematical models based on evolutionary game theory (EGT) and attempts to gain insight into the principal mechanisms that govern the development of cancer; see for example [Basanta et al. 2012; Orlando et al. 2012; Bach et al. 2001; Tomlinson and Bodmer 1997]. EGT, introduced in the 1970s by John Maynard Smith, was first used to analyze contests between rival species, competing for an important resource (e.g., food, territory, etc.). If one takes the view of tumor and stromal (native) cells as species, the same type of mathematical techniques, previously used in an ecological context, can be applied to study the progression of cancer. In recent years this approach has been applied to study various aspects of cancer. For example, [Basanta et al. 2012] uses a three cell species model to investigate prostate cancer tumor-stroma interaction; [Bach et al. 2001] and [Liu and Liu 2012] develop respectively two and three species models to study the synergistic effects of

MSC2010: primary 91A22; secondary 91A40.
Keywords: evolutionary game theory, mathematical modeling, breast cancer.

interactions between stromal cells and tumor cells, which often result in malignancy. Gatenby and Vincent [2003] conducted a study on tumor cells and used game theory to improve an existing linear model. In other studies, Mansury et al. [2006] and Basanta et al. [2008] employed game theory to model tumor growth in the brain.

Our proposed game-theoretical model of breast cancer builds on the model of [Liu and Liu 2012]. As in that paper, our model incorporates the growth-factor secreting stromal cells (native cells), motile tumor cells and proliferative cells (benign tumor cells). However, our model also incorporates macrophages, which play an important role in the development of breast cancer [Qian and Pollard 2012; Lamagna et al. 2006; Qian et al. 2009; Chen et al. 2011]. Macrophages have been shown to have a complex interaction with tumor cells and act in a dual role — in the beginning stages of cancer, they act as a defense mechanism against cancer by attacking tumor cells; however, they also produce growth factor, which in later stages can actually promote tumor growth [Lamagna et al. 2006; Chen et al. 2011].

Macrophages are large blood cells, produced as a result of the differentiation of monocytes. Monocytes travel through the blood stream and are produced in bone marrow. Once monocytes leave the blood stream, they turn into macrophages. These cells travel the body ingesting and destroying bacteria, cleaning up cellular debris, other harmful particles, dead cells and microbes [Børresen-Dale 2003]. Macrophages play an important role in the development of tumor cells. They ingest and destroy the cells. After they ingest the tumor cells, they use some of the materials in the cell for survival. They produce a growth factor that the macrophages and the tumor cells both benefit from [Mansury et al. 2006].

2. Model

We will assume there are four different types of cells in the body:

(a) the native cells (NC), which are the healthy stromal cells;

(b) the macrophages (MΦ), which are part of the immune system;

(c) the benign tumor cells (BTC), lump-forming cancer cells that lack the ability to metastasize;

(d) the motile tumor cells (MTC), metastatic cancer cells that can invade neighboring tissues.

The concentrations of the various types of cells are denoted by ϱ_{NC}, $\varrho_{\mathrm{M\Phi}}$, ϱ_{BTC} and ϱ_{MTC} respectively. The concentrations are between 0 and 1 and satisfy $\varrho_{\mathrm{NC}} + \varrho_{\mathrm{M\Phi}} + \varrho_{\mathrm{BTC}} + \varrho_{\mathrm{MTC}} = 1$.

We will now set up costs and benefits for each type of cell. Both the native cells and macrophages produce growth factor, which benefits all types of cells. As in [Archetti 2013], the cost of producing the growth factor, c_{G}, and the benefits of the

symbol	meaning
ϱ_{NC}	concentration of native cells
$\varrho_{\mathrm{M\Phi}}$	concentration of macrophages
ϱ_{BTC}	concentration of benign tumor cells
ϱ_{MTC}	concentration of motile tumor cells
c_{G}	cost of producing the growth factor
b_{G}	benefits of receiving the growth factor
c_{S}	cost of sharing the spaces
$c_{\mathrm{M,M\Phi}}$	cost of the ability to move for MΦ
$c_{\mathrm{M,MTC}}$	cost of the ability to move for MTC
b_{R}	benefits of reproducing quickly
c_{D}	cost of being destroyed by macrophages
W_X	net benefit for a given type of cells $X \in \{\mathrm{NC, M\Phi, BTC, MTC}\}$

Table 1. Model parameters and notation.

growth factor, b_{G}, will be assumed to be the same for all types of the cells. The macrophages and motile tumor cells can move and we will assume that the ability comes at the costs $c_{\mathrm{M,M\Phi}}$, and $c_{\mathrm{M,MTC}}$ respectively. The native cells and benign tumor cells stay in place and thus have to share the resources with other native and benign tumor cells, which comes at the cost c_{S}. The cancer cells can reproduce faster than native cells or macrophages, which we model by additional benefit b_{R} to the cancer cells, but the cancer cells can be destroyed by macrophages, which we model by additional cost c_{D} to the cancer cells. Overall, when the concentrations of the cells are ϱ_{NC}, $\varrho_{\mathrm{M\Phi}}$, ϱ_{BTC} and ϱ_{MTC}, the net benefits (benefits minus the costs) to each type of the cells are

$$W_{\mathrm{NC}} = b_{\mathrm{G}}(\varrho_{\mathrm{NC}} + \varrho_{\mathrm{M\Phi}}) - c_{\mathrm{G}} - c_{\mathrm{S}}(\varrho_{\mathrm{NC}} + \varrho_{\mathrm{BTC}}), \tag{1}$$

$$W_{\mathrm{M\Phi}} = b_{\mathrm{G}}(\varrho_{\mathrm{NC}} + \varrho_{\mathrm{M\Phi}}) - c_{\mathrm{G}} - c_{\mathrm{M,M\Phi}}, \tag{2}$$

$$W_{\mathrm{BTC}} = b_{\mathrm{R}} + b_{\mathrm{G}}(\varrho_{\mathrm{NC}} + \varrho_{\mathrm{M\Phi}}) - c_{\mathrm{S}}(\varrho_{\mathrm{NC}} + \varrho_{\mathrm{BTC}}) - c_{\mathrm{D}}\varrho_{\mathrm{M\Phi}}, \tag{3}$$

$$W_{\mathrm{MTC}} = b_{\mathrm{R}} + b_{\mathrm{G}}(\varrho_{\mathrm{NC}} + \varrho_{\mathrm{M\Phi}}) - c_{\mathrm{M,MTC}} - c_{\mathrm{D}}\varrho_{\mathrm{M\Phi}}. \tag{4}$$

For example, (1) reads that a native cell (a) benefits from the growth factor produced by (other) native cells and the macrophages, shown by the term $b_{\mathrm{G}}(\varrho_{\mathrm{NC}} + \varrho_{\mathrm{M\Phi}})$, (b) pays the cost of producing the growth factor itself, shown by the term c_{G}, and (c) pays the cost of sharing the space with other native cells and benign tumor cells, shown by the term $c_{\mathrm{S}}(\varrho_{\mathrm{NC}} + \varrho_{\mathrm{BTC}})$.

The notation and model parameters are summarized in Table 1.

Similarly to the models presented in [Basanta et al. 2008; Liu and Liu 2012; Bach et al. 2001], the situation described by (1)–(4) could be modeled as a matrix game when the interactions between individual cells are assumed to be pairwise and the payoff matrix is given by

$$\begin{array}{c|cccc}
\text{payoff to}\downarrow & \text{encounter with}\rightarrow\ \text{MTC} & \text{M}\Phi & \text{NC} & \text{BTC} \\
\hline
\text{MTC} & b_{\text{R}}-c_{\text{M,MTC}} & b_{\text{R}}-c_{\text{M,MTC}}-c_{\text{D}}+b_{\text{G}} & b_{\text{R}}-c_{\text{M,MTC}}+b_{\text{G}} & b_{\text{R}}-c_{\text{M,MTC}} \\
\text{M}\Phi & -c_{\text{G}}-c_{\text{M,M}\Phi} & b_{\text{G}}-c_{\text{G}}-c_{\text{M,M}\Phi} & b_{\text{G}}-c_{\text{G}}-c_{\text{M,M}\Phi} & -c_{\text{G}}-c_{\text{M,M}\Phi} \\
\text{NC} & -c_{\text{G}} & b_{\text{G}}-c_{\text{G}} & b_{\text{G}}-c_{\text{G}}-c_{\text{S}} & b_{\text{G}}-c_{\text{G}}-c_{\text{S}} \\
\text{BTC} & b_{\text{R}} & b_{\text{R}}+b_{\text{G}}-c_{\text{D}} & b_{\text{R}}+b_{\text{G}}-c_{\text{S}} & b_{\text{R}}-c_{\text{S}}
\end{array} \tag{5}$$

To make sure that the entries of matrix (5) are nonnegative, it is customary to add a fixed number (for example 1) to all of them.

3. Results

We are interested in deriving conditions which ensure that the cancer cells (or at least the metastatic tumor cells) eventually die out.

3.1. ***Coexistence of native cells and macrophages.*** We first derive conditions on the parameters which ensure a healthy organism; i.e., the coexistence of native cells and macrophages (with no tumor cells) is an evolutionarily stable state (ESS). The assumption that there are only native cells and macrophages requires that $\varrho_{\text{BTC}} = 0$ and $\varrho_{\text{MTC}} = 0$ and consequently $\varrho_{\text{NC}} + \varrho_{\text{M}\Phi} = 1$. Subtracting (2) from (1) yields

$$W_{\text{NC}} - W_{\text{M}\Phi} = c_{\text{M,M}\Phi} - c_{\text{S}}\varrho_{\text{NC}}. \tag{6}$$

Recall that the net benefit from interaction (fitness) for the native cells is denoted by W_{NC} and for the macrophages, by $W_{\text{M}\Phi}$. It follows from (6) that

$$W_{\text{NC}} \gtreqless W_{\text{M}\Phi} \quad \text{if and only if} \quad \varrho_{\text{NC}} \lesseqgtr \frac{c_{\text{M,M}\Phi}}{c_{\text{S}}}$$

(in other words, native cells do better than macrophages if there are too many macrophages, and vice versa). Consequently, the only candidates for the stable healthy proportion of the cells are $\varrho_{\text{NC}} = c_{\text{M,M}\Phi}/c_{\text{S}}$ and $\varrho_{\text{M}\Phi} = (c_{\text{S}} - c_{\text{M,M}\Phi})/c_{\text{S}}$.

Since, in this scenario, we would like for the ESS to include no tumor cells, we need to derive the conditions which ensure that tumor cells (in tiny amounts) still do worse than the native cells. Subtracting (1) from (3) yields

$$W_{\text{BTC}} - W_{\text{NC}} = b_{\text{R}} + c_{\text{G}} - c_{\text{D}}\varrho_{\text{M}\Phi}, \tag{7}$$

while subtracting (2) from (4) yields

$$W_{\text{MTC}} - W_{\text{M}\Phi} = b_{\text{R}} + c_{\text{G}} + (c_{\text{M,M}\Phi} - c_{\text{M,MTC}}) - c_{\text{D}}\varrho_{\text{M}\Phi}. \tag{8}$$

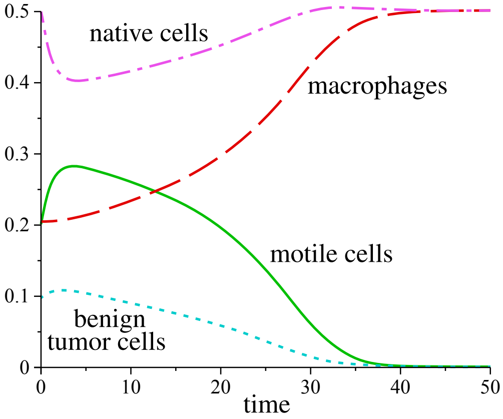


Figure 1. If (9) is satisfied, then the tumor cells eventually extinct. In this figure the values of the parameters are as follows: $b_{\mathrm{R}} = 1$, $c_{\mathrm{G}} = 2$, $b_{\mathrm{G}} = 4$, $c_{\mathrm{D}} = 7$, $c_{\mathrm{M,MTC}} = c_{\mathrm{M,M\Phi}} = 1$, $c_{\mathrm{S}} = 2$.

It follows that, in a healthy body where $\varrho_{\mathrm{M\Phi}} = (c_{\mathrm{S}} - c_{\mathrm{M,M\Phi}})/c_{\mathrm{S}}$, both the benign tumor cells and the motile tumor cells do worse than healthy cells if and only if

$$b_{\mathrm{R}} + c_{\mathrm{G}} + \max\{0, c_{\mathrm{M,M\Phi}} - c_{\mathrm{M,MTC}}\} < c_{\mathrm{D}} \frac{c_{\mathrm{S}} - c_{\mathrm{M,M\Phi}}}{c_{\mathrm{S}}}. \tag{9}$$

In particular, increasing the value of c_{D} (or the ability of macrophages to destroy tumor cells) or decreasing the value of b_{R} (the reproductive advantage of the tumor cells) ensures that the fitness of both types of tumor cells is smaller than the fitness of the native cells and the macrophages and that the body will stay healthy.

Moreover, when condition (9) is satisfied, and the initial state of the system involves relatively small amounts of tumor cells, the tumor cells eventually go extinct; see for example Figure 1, which shows the evolution of the four cell types under the replicator dynamics [Hofbauer and Sigmund 1998]

$$\frac{d}{dt}\varrho_{\text{cell type}} = \varrho_{\text{cell type}}(W_{\text{cell type}} - \overline{W}), \tag{10}$$

where $\overline{W}$ is the average fitness, given by

$$\overline{W} = \sum_i \varrho_i W_i.$$

The summation index i varies over all four cell types.

3.2. ***Coexistence of native cells, macrophages, and benign tumor cells.*** We note that if $c_{\mathrm{M,M\Phi}} \leq c_{\mathrm{M,MTC}}$, then by (7) and (8), motile tumor cells do worse than benign tumor cells in a healthy body. It is thus possible that the body will be able to

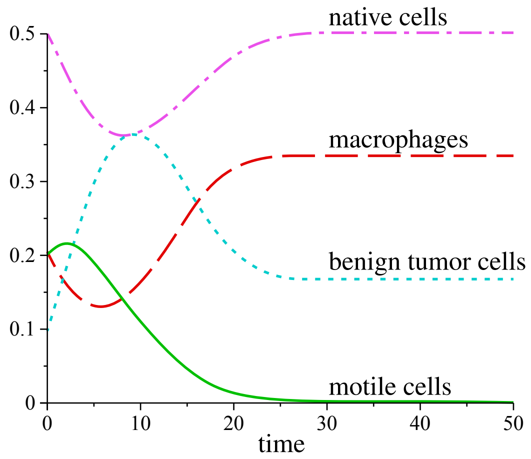


Figure 2. If (13) holds, then the motile tumor cells eventually extinct even when the benign tumor cells can stay in the body. The parameters are as follows: $b_{\mathrm{R}} = 1$, $c_{\mathrm{G}} = 2$, $b_{\mathrm{G}} = 4$, $c_{\mathrm{D}} = 7$, $c_{\mathrm{M,MTC}} = 1$, $c_{\mathrm{M,M\Phi}} = 0.8$, $c_{\mathrm{S}} = 1.2$

get rid of the dangerous motile tumor cells even if it is not able to get rid of the less dangerous benign tumor cells. This is the situation that we will investigate now.

More precisely, we will want to see under what conditions it is possible to have $\varrho_{\mathrm{MTC}} = 0$ as a stable condition. As in Section 3.1, subtracting (2) from (1) yields

$$W_{\mathrm{NC}} - W_{\mathrm{M\Phi}} = c_{\mathrm{M,M\Phi}} - c_{\mathrm{S}}(\varrho_{\mathrm{NC}} + \varrho_{\mathrm{BTC}}). \tag{11}$$

An ESS requires that the fitnesses of each of the coexisting types of cells be equal to each other. In particular, $W_{\mathrm{NC}} = W_{\mathrm{M\Phi}}$ and since $\varrho_{\mathrm{MTC}} = 0$, we also get $\varrho_{\mathrm{NC}} + \varrho_{\mathrm{BTC}} = 1 - \varrho_{\mathrm{M\Phi}}$. Thus, it follows from (11) that, as in Section 3.1,

$$\varrho_{\mathrm{M\Phi}} = \frac{c_{\mathrm{S}} - c_{\mathrm{M,M\Phi}}}{c_{\mathrm{S}}}. \tag{12}$$

Since subtracting (2) from (4) still yields (8), we get that no motile tumor cells are possible only if

$$b_{\mathrm{R}} + c_{\mathrm{G}} + c_{\mathrm{M,M\Phi}} - c_{\mathrm{M,MTC}} < c_{\mathrm{D}} \frac{c_{\mathrm{S}} - c_{\mathrm{M,M\Phi}}}{c_{\mathrm{S}}}. \tag{13}$$

Thus, if it is difficult to ensure that condition (9) is satisfied for a patient, one can still attempt to satisfy (13), for example by increasing the value of $c_{\mathrm{M,MTC}}$ (the cost of movement for the tumor cells) or decreasing the value of $c_{\mathrm{M,M\Phi}}$ (the cost of movement for the macrophages), and thus prevent the development of metastatic cancer.

Figure 2 shows the evolution of the concentrations of the four cell types as a function of time under the replicator dynamics (10) when (9) is not satisfied but

(13) still holds. We can see that the benign tumor cells stay in the body but the motile tumor cells die out.

Note that in the case when $c_{\mathrm{M,MTC}} < c_{\mathrm{M,M\Phi}}$, the motile tumor cells can thrive in the body whenever benign tumor cells can.

4. Conclusions and discussion

In this paper we presented and analyzed a game-theoretical model of breast cancer. We have extended the model of [Liu and Liu 2012] by explicitly incorporating the macrophages. As observed in [Qian and Pollard 2012; Lamagna et al. 2006; Qian et al. 2009; Chen et al. 2011] and confirmed by the analysis of our model, the macrophages indeed play a crucial role in the development and prevention of cancer.

Our model suggests at least three possible ways of cancer treatment. One is to increase the damage to the tumor cells caused by macrophages (or in a similar fashion, increase the ability of macrophages to destroy tumor cells). Another way is to decrease the reproductive advantage of the tumor cells, i.e., their ability to reproduce much more quickly than healthy cells. And a third way is to increase the cost of mobility for the tumor cells. The last scenario may not completely prevent the cancer from developing in the body, but it may prevent dangerous metastatic tumors.

Acknowledgment

This research was supported by the MAA's NREUP program funded by the NSA (grant H98230-13-1-0270) and the NSF (grant DMS-1156582).

References

[Archetti 2013] M. Archetti, "Dynamics of growth factor production in monolayers of cancer cells and evolution of resistance to anticancer therapies", *Evol. App.* **6**:8 (2013), 1146–1159.

[Bach et al. 2001] L. A. Bach, S. M. Bentzen, J. Alsner, and F. B. Christiansen, "An evolutionary-game model of tumour-cell interactions: possible relevance to gene therapy", *Eur. J. Cancer* **37**:16 (2001), 2116–2120.

[Basanta et al. 2008] D. Basanta, M. Simon, H. Hatzikirou, and A. Deutsch, "Evolutionary game theory elucidates the role of glycolysis in glioma progression and invasion", *Cell Prolif.* **41**:6 (2008), 980–987.

[Basanta et al. 2012] D. Basanta, J. G. Scott, M. N. Fishman, G. Ayala, S. W. Hayward, and A. R. A. Anderson, "Investigating prostate cancer tumour-stroma interactions: clinical and biological insights from an evolutionary game", *Brit. J. Cancer* **106**:1 (2012), 174–181.

[Børresen-Dale 2003] A.-L. Børresen-Dale, "Tp53 and breast cancer", *Human Mutation* **21** (2003), 292–300.

[Chen et al. 2011] J. Chen, Y. Yao, C. Gong, F. Yu, S. Su, J. Chen, B. Liu, H. Deng, F. Wang, L. Lin, H. Yao, F. Su, K. S. Anderson, Q. Liu, M. E. Ewen, X. Yao, and E. Song, "Ccl18 from tumor-associated macrophages promotes breast cancer metastasis via pitpnm3", *Cancer Cell* **19**:4 (2011), 541–555.

[Ford et al. 1998] D. Ford, D. Easton, M. Stratton, S. Narod, D. Goldgar, P. Devilee, D. Bishop, B. Weber, G. Lenoir, J. Chang-Claude, H. Sobol, M. Teare, J. Struewing, A. Arason, S. Scherneck, J. Peto, T. Rebbeck, P. Tonin, S. Neuhausen, R. Barkardottir, J. Eyfjord, H. Lynch, B. Ponder, S. Gayther, J. Birch, A. Lindblom, D. Stoppa-Lyonnet, Y. Bignon, A. Borg, U. Hamann, N. Haites, R. Scott, C. Maugard, H. Vasen, S. Seitz, L. Cannon-Albright, A. Schofield, and M. Zelada-Hedman, "Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families", *Amer. J. Human Genetics* **62**:3 (1998), 676–689.

[Gatenby and Vincent 2003] R. A. Gatenby and T. L. Vincent, "Application of quantitative models from population biology and evolutionary game theory to tumor therapeutic strategies", *Molecular Cancer Therapeutics* **2**:9 (2003), 919–927.

[Hofbauer and Sigmund 1998] J. Hofbauer and K. Sigmund, *Evolutionary games and population dynamics*, Cambridge University Press, 1998. MR Zbl

[Lamagna et al. 2006] C. Lamagna, M. Aurrand-Lions, and B. A. Imhof, "Dual role of macrophages in tumor growth and angiogenesis", *J. Leukocyte Biol.* **80**:4 (2006), 705–713.

[Liu and Liu 2012] Q. Liu and Z. Liu, "Malignancy through cooperation: an evolutionary game theory approach", *Cell Proliferation* **45**:4 (2012), 365–377.

[Mansury et al. 2006] Y. Mansury, M. Diggory, and T. S. Deisboeck, "Evolutionary game theory in an agent-based brain tumor model: exploring the 'genotype-phenotype' link", *J. Theoret. Biol.* **238**:1 (2006), 146–156. MR

[Orlando et al. 2012] P. A. Orlando, R. A. Gattenby, and J. S. Brown, "Cancer treatment as a game: integrating evolutionary game theory into the optimal control of chemotherapy", *Phys. Biol.* **9**:6 (2012), art. id. 065007.

[Qian and Pollard 2012] B.-Z. Qian and J. W. Pollard, "New tricks for metastasis-associated macrophages", *Breast Cancer Research* **14** (2012), art. id. 316.

[Qian et al. 2009] B. Qian, Y. Deng, J. H. Im, R. J. Muschel, Y. Zou, J. Li, R. A. Lang, and J. W. Pollard, "A distinct macrophage population mediates metastatic breast cancer cell extravasation, establishment and growth", *PLoS One* **4**:8 (2009), art. id. e6562.

[Slamon et al. 1987] D. J. Slamon, G. M. Clark, S. G. Wong, W. J. Levin, A. Ullrich, and W. L. McGuire, "Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene", *Science* **235**:4785 (1987), 177–182.

[Tomlinson and Bodmer 1997] I. P. M. Tomlinson and W. F. Bodmer, "Modelling the consequences of interactions between tumour cells", *Brit. J. Cancer* **75**:2 (1997), 157–160.

Received: 2015-01-13 Revised: 2017-07-08 Accepted: 2017-07-20

keyona.barton@bennett.edu *Bennett College, Greensboro, NC, United States*

corbin.smith@bennett.edu *Bennett College, Greensboro, NC, United States*

rychtar@uncg.edu *Department of Mathematics and Statistics, The University of North Carolina at Greensboro, Greensboro, NC, United States*

tsendova@math.msu.edu *Department of Mathematics, Michigan State University, East Lansing, MI, United States*

INVOLVE 11:4 (2018)
dx.doi.org/10.2140/involve.2018.11.549

The isoperimetric problem in the plane with the sum of two Gaussian densities

John Berry, Matthew Dannenberg, Jason Liang and Yingyi Zeng

(Communicated by Gaven Martin)

We consider the isoperimetric problem for the sum of two Gaussian densities in the line and the plane. We prove that the double Gaussian isoperimetric regions in the line are rays and that if the double Gaussian isoperimetric regions in the plane are half-spaces, then they must be bounded by vertical lines.

1. Introduction

Sudakov and Tsirelson, and independently Borell, see [Morgan 2009, 18.2], proved that for $\mathbb{R}^n$ endowed with a Gaussian measure, half-spaces bounded by hyperplanes are isoperimetric, i.e., minimize weighted perimeter for given weighted volume. Cañete et al. [2010, Question 6], in response to a question of Brancolini, conjectured that for $\mathbb{R}^n$ endowed with a finite sum of Gaussian measures centered on the x-axis, half-spaces bounded by vertical hyperplanes are isoperimetric. We consider the case of two such Gaussians in $\mathbb{R}^1$ and $\mathbb{R}^2$. Our Theorem 3.16 proves that on the double Gaussian line, rays are isoperimetric. Section 4 provides evidence that on the double Gaussian plane, half-spaces are isoperimetric.

1.1. ***The double Gaussian line.*** Theorem 3.16 states that the isoperimetric regions in the double Gaussian line are rays. We may assume that the two Gaussians have centers at 1 and -1. For small variances, the theorem follows by comparison with the single Gaussian. For larger variances, additional quantitative and stability arguments are needed to rule out certain nonray cases.

1.2. ***The double Gaussian plane.*** A conjecture of Cañete et al. [2010, Question 6], appearing in this paper as Conjecture 4.1, states that isoperimetric regions in the double Gaussian plane are half-planes bounded by vertical lines. We use variational arguments to show that horizontal and vertical lines are the only lines that are candidates, and that vertical lines always beat horizontal lines.

MSC2010: 49Q10, 53A10.
Keywords: isoperimetric, Gaussian.

2. First and second variations

Formulas 2.3 and 2.6 state standard first and second variation formulas, analogous to the first and second derivative conditions for local minima of twice-differentiable real functions.

Definition 2.1. A *density* e^{ψ} on $\mathbb{R}^n$ is a positive, continuous function used to weight volume and hypersurface area. Given a density e^{ψ}, the (weighted) *volume* of a region R is given by

$$\int_R e^{\psi}\, dV_0.$$

The (weighted) *hypersurface area* of its boundary ∂R is given by

$$\int_{\partial R} e^{\psi}\, dA_0.$$

R is called *isoperimetric* if no other region of the same weighted volume has a boundary with smaller hypersurface area.

We now assume that the density e^{ψ} is smooth. The existence and regularity of isoperimetric regions for densities of finite total volume is standard.

Existence and Regularity 2.2 [Morgan 2009, 5.5, 9.1, 8.5]. *Suppose that e^{ψ} is a density in the line or plane such that the line or plane has finite measure A_0. Then for any $0 < A < A_0$, an isoperimetric region R of weighted volume A exists and is a finite union of intervals bounded by finitely many points in the line or a finite union of regions with smooth boundaries in the plane.*

Let e^{ψ} be a smooth density on $\mathbb{R}^{n+1}$. Let R be a smooth region in $\mathbb{R}^{n+1}$. Let φ_t be a smooth, one-parameter family of deformations on $\mathbb{R}^{n+1}$ such that φ_0 is the identity. For a given $x \in \partial R$, the deformation $\varphi_t(x)$ traces out a small path in $\mathbb{R}^{n+1}$ beginning at x and $\varphi_t(\partial R)$ is a curve for each t. Therefore $\{\varphi_t\}$, where $|t| < \epsilon$, describes a perturbation of ∂R. Define

$$V(t) = \int_{\varphi_t(R)} e^{\psi}\, dV_0 \quad \text{and} \quad P(t) = \int_{\varphi_t(\partial R)} e^{\psi}\, dA_0.$$

First Variation Formulas 2.3 [Rosales et al. 2008, Lemma 3.1]. *Suppose that $\boldsymbol{n}$ and H are the inward unit normal and mean curvature of ∂R. Let X be the vector field $d\varphi_t/dt$ and $u = \langle X, \boldsymbol{n}\rangle$. Then we have that*

$$V'(0) = -\int_{\partial R} e^{\psi} u\, dA_0 \quad \textit{and} \quad P'(0) = -\int_{\partial R} (nH - \langle \nabla\psi, \boldsymbol{n}\rangle) e^{\psi} u\, dA_0.$$

Since any isoperimetric curve is a local minimum among all curves enclosing a certain volume A, it satisfies $P'(0) = 0$ for any φ_t such that $V(t) = A$ for small t.

Corollary 2.4. *If a curve ∂R is isoperimetric, then $(nH - \langle \nabla\psi, \boldsymbol{n}\rangle)$ is constant on ∂R.*

Proof. If a curve ∂R is isoperimetric, then it satisfies $P'(0) = 0$. By Formula 2.3, this occurs if and only if $(nH - \langle \nabla\psi, \boldsymbol{n}\rangle)$ is constant on ∂R. □

Definition 2.5. Let C be a boundary in the line or plane with unit inward normal $\boldsymbol{n}$ and let κ denote the standard curvature. For a density e^{ψ}, we call $\kappa_{\psi} = \kappa - d\psi/d\boldsymbol{n}$ the *generalized curvature* of C.

By Corollary 2.4, all isoperimetric curves have constant generalized curvature. In the real line, $n = 0$, so isoperimetric curves have $\langle \nabla\psi, \boldsymbol{n}\rangle$ constant. For the interval $[a, b]$, the generalized curvature evaluated at b is equal to $\psi'(b)$, while the generalized curvature evaluated at a is equal to $-\psi'(a)$.

Second Variation Formula 2.6 [Rosales et al. 2008, Proposition 3.6]. *Let the real line be with smooth density e^{ψ}. If a one-dimensional boundary $l = \partial R$ satisfies $P'(0) = 0$ for any volume-preserving $\{\varphi_t\}$, then*

$$(P - \kappa_{\psi} V)''(0) = \int_l f u^2 \left(\frac{d^2\psi}{dx^2} \right) da.$$

Proof. This formula comes from Proposition 3.6 in [Rosales et al. 2008], where the second variation is stated for arbitrary dimensions. Some terms from the general formula cancel in the one-dimensional case. □

Corollary 2.7. *Let S be a subset of the real line such that $\psi''(x) \le 0$ for all $x \in S$ with equality holding at no more than one point. If B is an isoperimetric boundary contained in S, then B is connected and thus a single point.*

Proof. If B has at least two connected components, then since by Existence and Regularity 2.2 B consists of a finite union of points, there is a nontrivial volume-preserving flow on B given by moving one component so as to increase the volume and the other so as to decrease it. By Formula 2.6, the second variation satisfies

$$(P - \kappa_{\psi} V)''(0) = \int_B f u^2 (\psi''(x))\, da < 0.$$

This contradicts that B is isoperimetric. □

3. Isoperimetric regions on the double Gaussian line

Theorem 3.16 states that for the real line with density given by the sum of two Gaussians with the same variance a^2, isoperimetric regions are rays bounded by single points. This theorem is a necessary condition for Conjecture 4.1, which states that isoperimetric regions in the double Gaussian plane are half-planes bounded by vertical lines. Propositions 3.4, 3.14, and 3.15 treat the cases $a^2 \geq 1$, $1 > a^2 > \frac{1}{2}$, and $\frac{1}{2} \geq a^2 > 0$.

Lemma 3.5 shows that if the Gaussians have the same variance, we can reduce the problem to ruling out a few noninterval, but still symmetrical, cases. When the Gaussians have different variances, the problem is harder and not treated by our results.

Let $g_{c,a}$ denote the Gaussian density with mean c and variance a^2, and let

$$f_{c,a}(x) = \frac{1}{2}\left(\frac{e^{-(x-c)^2/(2a^2)} + e^{-(x+c)^2/(2a^2)}}{a\sqrt{2\pi}}\right) = \tfrac{1}{2}(g_{c,a}(x) + g_{-c,a}(x)).$$

Let

$$f(x) = \tfrac{1}{2}(f_1(x) + f_2(x)) = \tfrac{1}{2}(g_{1,a}(x) + g_{-1,a}(x)).$$

In one dimension, the regions are unions of intervals and their boundaries are points. Since the total measure is finite, isoperimetric regions exist by Existence and Regularity 2.2. For a given weighted length A, we seek to find the set of points with the smallest total density which bounds a region of weighted length A. Since the complement of a region of weighted length A has weighted length $1 - A$, we can assume that our regions have weighted length $0 \leq A \leq \frac{1}{2}$.

The following proposition shows that it suffices to consider the density f.

Proposition 3.1. *Suppose B is an isoperimetric boundary enclosing a region L of weighted length A for the density $f_{1,a}(x)$. Then for any $b > 0$, we have bB is an isoperimetric boundary enclosing region bL of weighted length A for the density $f_{b,ab}(x)$.*

Proof. Let g denote the standard Gaussian density.

First, we show that for any boundary P enclosing a region Q, the weighted length of bQ for the density $f_{b,ab}(x)$ is the same as the weighted length of Q for the density $f_{1,a}(x)$. We have that

$$\begin{aligned}|Q| = \int_Q f_{1,a}(x)\,dx &= \frac{1}{2}\int_Q g_{1,a}(x)\,dx + \frac{1}{2}\int_Q g_{-1,a}(x)\,dx \\ &= \frac{1}{2}\int_{(Q-1)/a} g(x)\,dx + \frac{1}{2}\int_{(Q+1)/a} g(x)\,dx \\ &= \frac{1}{2}\int_{(bQ-b)/(ab)} g(x)\,dx + \frac{1}{2}\int_{(bQ+b)/(ab)} g(x)\,dx \\ &= \frac{1}{2}\int_Q g_{b,ab}(x)\,dx + \frac{1}{2}\int_Q g_{-b,ab}(x)\,dx = |bQ|,\end{aligned}$$

where $|\cdot|$ denotes the weighted length in the appropriate densities.

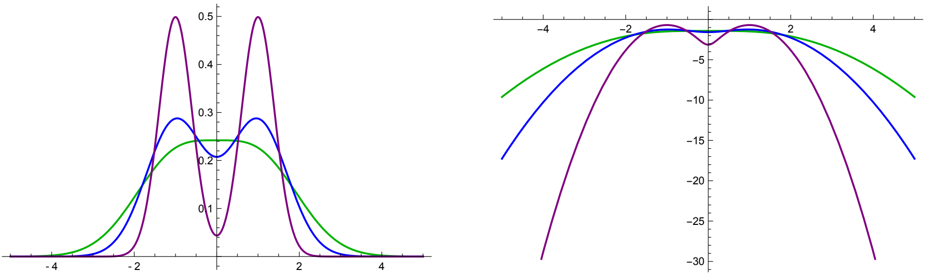


Figure 1. Plots of f (left) and ψ (right). The purple curves are for $a^2 = 0.16$, the blue curves for $a^2 = \frac{1}{2}$, and the green curves for $a^2 = 1$.

Second, for any two boundaries P_1 and P_2, we have $f_{b,ab}(bx) = (1/b) f_{1,a}(x)$ for $x \in P_i$. Thus, $|P_1| \geq |P_2|$ in the density $f_{1,a}(x)$ exactly when $|bP_1| \geq |bP_2|$ in the density $f_{b,ab}(x)$.

Therefore $|bL| = A$ in the density $f_{b,ab}(x)$, and if any other boundary P enclosing region Q satisfies $|Q| = A$ in the density $f_{b,ab}(x)$, then since B is isoperimetric, we have $|B| \leq |P/b|$ in the density $f_{1,a}(x)$. Therefore $|bB| \leq |P|$ in the density $f_{b,ab}(x)$, so bP is isoperimetric. □

As a result of Proposition 3.1, it suffices to consider the density

$$f = \tfrac{1}{2}(f_1 + f_2) = \tfrac{1}{2}(g_{1,a} + g_{-1,a}).$$

Proposition 3.2. *Let X be the disjoint union of two real-lines X_1 and X_2, each with a standard Gaussian density scaled so that it has weighted length $\frac{1}{2}$. For any given length $0 < A < \frac{1}{2}$, the isoperimetric region in X of length A is a ray contained entirely in X_1 or X_2.*

Proof. Let B be an isoperimetric boundary and B_i its intersection with X_i. If B_1 and B_2 are nonempty, then they each must be a single point since the isoperimetric boundaries for the single Gaussian are always single points. Assume, in contradiction to the proposition, that $B_i = \{b_i\}$ for $i = 1, 2$ is the i-th component on the i-th Gaussian bounding a ray L_i of weighted length A_i. Since $A_1 + A_2 < \frac{1}{2}$, it is possible to put a point b_1' on the first Gaussian at the same height as that of b_2 bounding a ray L_1' disjoint from L_1 and with weighted length A_2. Consider the boundary $B' = \{b_1, b_1'\}$, which has the same weighted perimeter as that of B. There exists a single point on B_1 bounding a ray of area A and with weighted density smaller than $|B'| = |B|$. This contradicts the fact that B is isoperimetric. □

Proposition 3.3. *For the double Gaussian density f, the log derivative ψ' is given by*

$$\psi'(x) = a^{-2}\Big(-x + \tanh\frac{x}{a^2}\Big).$$

Proof. We have

$$\psi'(x) = \frac{\dfrac{-e^{-(-1+x)^2/(2a^2)}(-1+x)}{a^2} + \dfrac{-e^{-(1+x)^2/(2a^2)}(1+x)}{a^2}}{e^{-(-1+x)^2/(2a^2)} + e^{-(1+x)^2/(2a^2)}}.$$

By using the substitution

$$\tanh\Big(\frac{x}{a^2}\Big) = \frac{e^{x/a^2} - e^{-x/a^2}}{e^{x/a^2} + e^{-x/a^2}},$$

we get

$$\psi'(x) = a^{-2}\Big(-x + \tanh\frac{x}{a^2}\Big). \qquad \square$$

Proposition 3.4. *For the double Gaussian density f, if $a \geq 1$, isoperimetric boundaries are single points.*

Proof. For any given a, we have

$$\psi'(x) = a^{-2}\Big(-x + \tanh\frac{x}{a^2}\Big),$$

$$\psi''(x) = a^{-4}\Big(-a^2 + \operatorname{sech}^2\frac{x}{a^2}\Big),$$

$$\psi'''(x) = -2a^{-6}\operatorname{sech}^2\frac{x}{a^2}\tanh\frac{x}{a^2}.$$

As shown in Figure 2, $\psi'''(x)$ is positive for any $x < 0$ and negative for $x > 0$, so $\psi''(x)$ achieves its unique maximum at $x = 0$ for any given a. We have $\psi''(0) = (1 - a^2)/a^4$, so $\psi''(0)$ is greater than 0 for $a < 1$, and less than or equal to 0 for $a \geq 1$. If $a \geq 1$, by Corollary 2.7, isoperimetric boundaries are always connected. Since isoperimetric boundaries consist of finite unions of points, they must be single points. □

Lemma 3.5. *Let p and q be two real functions with $p(0) = q(0)$. Suppose p and q satisfy*

(1) $p'(0) = q'(0) \geq 0$,

(2) $q''(0) \geq p''(0)$,

(3) $q''(0) \geq 0$, *and*

(4) $p''' < 0$ *and* $q''' > 0$ *on* $(0, \infty)$.

For any $a, b > 0$, if $p(a) = q(b)$, then $q'(b) > p'(a)$.

Proof. As in Figure 3, for all $x > 0$, by (2) and (4) we have $q''(x) > p''(x)$ and by (3) and (4) we have $q''(x) > 0$. If we choose a' so that $q'(a') = p'(a)$, we will have $a' < a$.

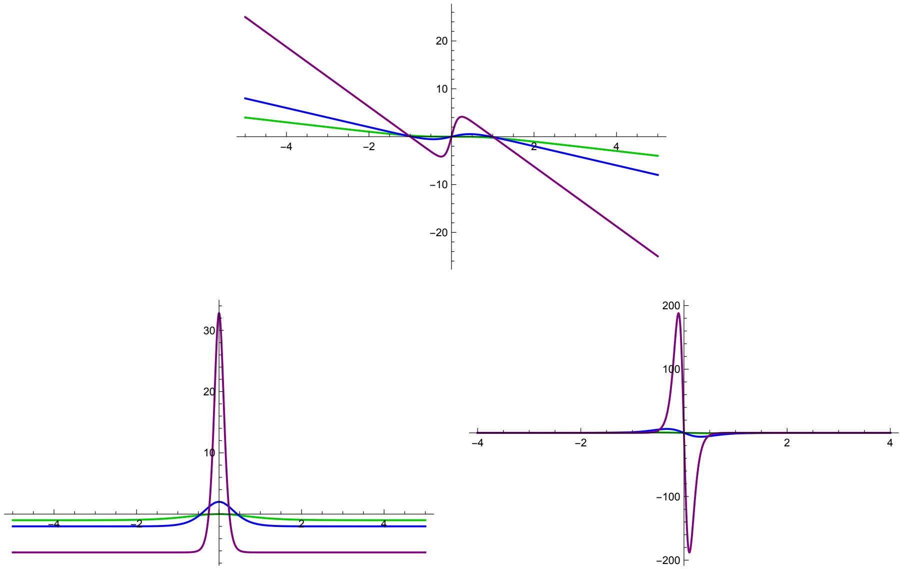


Figure 2. Plots of ψ' (top), ψ'' (bottom left), and ψ''' (bottom right). The purple curves are for $a^2 = 0.16$, the blue curves for $a^2 = \frac{1}{2}$, and the green curves for $a^2 = 1$.

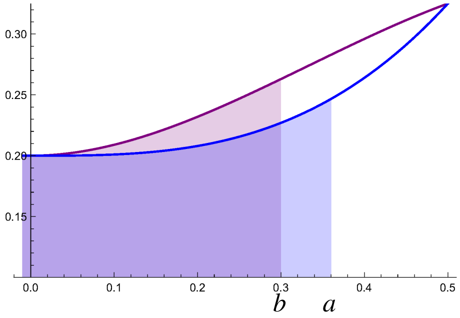


Figure 3. The purple curve is q', and the blue curve p'. When the areas are equal, as in the picture, q' is higher.

Since by (4) p' is concave and q' is convex,

$$q(a') = \int_0^{a'} q'(t)\,dt \le \tfrac{1}{2}(a' * q'(a')) < \tfrac{1}{2}(a * p'(a)) \le \int_0^{a'} q'(t)\,dt = p(a) = q(b).$$

Therefore $b > a'$, so $q'(b) > p'(a)$, as asserted. □

Proposition 3.6. *Suppose $[a, b]$ is an interval of f-weighted length $0 < A < \frac{1}{2}$ with $-1 < a < b < 1$. Then there exists a union of rays $B = (-\infty, c] \cup [d, \infty]$ of f_1-weighted length A such that $f_1(c) < f_1(b) < f(b)$ and $f_1(d) < f_2(a) < f(a)$.*

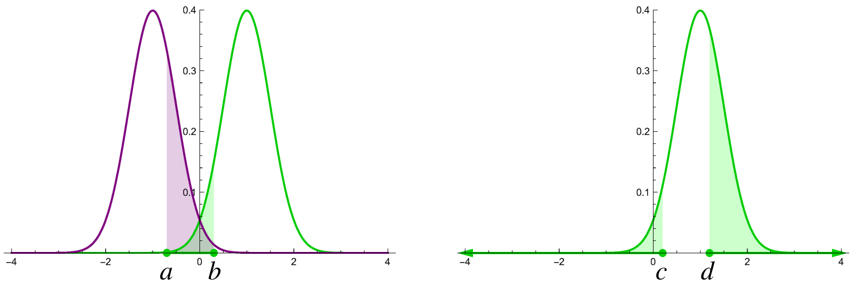


Figure 4. Left: an interval in the double Gaussian. Right: two rays in the single Gaussian. The total areas are the same, but the heights in the right graph are slightly lower.

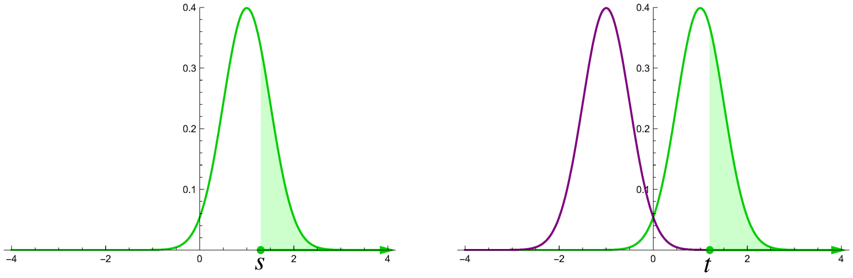


Figure 5. Left: ray in the single Gaussian. Right: ray in the double Gaussian. The total areas are the same, but the height in the right graph is slightly lower.

Proof. Since $2+a = 1+(a-(-1))$, we have $f_1(2+a) = f_2(a)$. The union of rays $(-\infty, b] \cup [2+a, \infty)$ has greater f_1-weighted length than the f-weighted length of $[a, b]$. Therefore there exists $c < t$ and $d > 2+a$ such that $(-\infty, c] \cup [d, \infty)$ has f_1-weighted length A, and

$$f_1(c) + f_1(d) < f_1(b) + f_2(a) < f(b) + f(a).$$

See Figure 4. □

Proposition 3.7. *If* $[s, \infty)$ *has* $\left(\frac{1}{2}\right) f_1$*-weighted length* $0 < A \leq \frac{1}{4}$*, then there exists* $t > s$ *such that* $[t, \infty)$ *has* f*-weighted length* A*.*

Proof. If $[s, \infty)$ has $\left(\frac{1}{2}\right) f_1$-weighted length $0 < A \leq \frac{1}{4}$, then $s \geq 1$. The interval $[s, \infty)$ has f-weighted length greater than A. Therefore there exists $t > s$ such that $[t, \infty)$ has f-weighted length A. See Figure 5. □

Now we begin analyzing the case where the variance satisfies $0 < a^2 < 1$.

Proposition 3.8. *If* a^2 *satisfies* $0 < a^2 \leq 1$*, then* $\psi''(x) = 0$ *exactly when* $x = \pm a^2 \operatorname{arccosh}(1/a)$.

Proof. This follows from the formula for $\psi''(x)$ given in Proposition 3.4. □

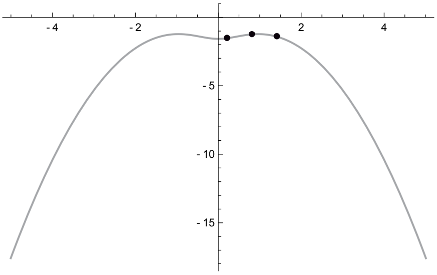


Figure 6. On the graph of $\psi = \log f$, there are at most three points with $x > 0$ with the same value for $|\psi'(x)|$.

Suppose that a^2 is a variance. In the proof of the following proposition, we will use the quantity

$$c_a = a^2 \operatorname{arccosh}(1/a).$$

Proposition 3.9. *Suppose* $0 < a^2 \leq 1$ *and* B *is an isoperimetric boundary with at least one point* s *in* $[0, c]$, *where* $c = c_a$, *enclosing a region of weighted-length* $0 < A < \frac{1}{2}$. *Then the boundary* B *is one of the following*:

(1) *a single point* s *enclosing the ray* $[s, \infty)$,

(2) $\{s, t\}$, *where* $t > s$, *enclosing the interval* $[s, t]$,

(3) $\{s, t\}$, *where* $s > 0 > t$, *enclosing the interval* $[t, s]$,

(4) $\{s, -s, t\}$ *enclosing* $[-s, s] \cup (-\infty, t]$, $[-s, s] \cup [t, \infty)$ *or* $[s, t] \cup (-\infty, -s]$.

The analogous claims apply if $s \in [-c, 0]$.

Proof. Since B is isoperimetric, it can contain at most one point x at which $\psi''(x) < 0$. If it contained two such points, then by slightly shifting the two points we could create a new region with the same weighted length. By Formula 2.6, the boundary of this region would have a smaller total density. Therefore B can contain at most one point outside of $[-c, c]$.

In addition, B has constant curvature, so $|\psi'|$ is constant on B (see Figure 6). Since $\psi''(s)$ is positive on $[0, c)$ and negative on $(c, \infty]$, there exists one point $t > s > 0$ such that $\psi'(t) = \psi'(s)$ and one point $u > t > s > 0$ such that $-\psi'(u) = \psi'(s)$. Therefore B is a subset of $\{s, t, u, -s, -t, -u\}$. Suppose B is not (1). If B contains no points outside of $[-c, c]$, then B is (3). Suppose B contains one point y outside of $[-c, c]$. If $t > 0$, then the only possibilities are (2) or (4). If $t < 0$, then the only possibilities are (3) or (4). The regions enclosed follow from the fact that we assume $0 < A < \frac{1}{2}$. □

Proposition 3.10. *Suppose* B *is an isoperimetric boundary with at least one point* $s \in [-c, c]$. *If* B *is of type (3) in Proposition 3.9 and* $0 < a^2 \leq \frac{1}{2}$, *then the region* R *enclosed by* B *has* f*-weighted length no more than* $\frac{1}{4}$.

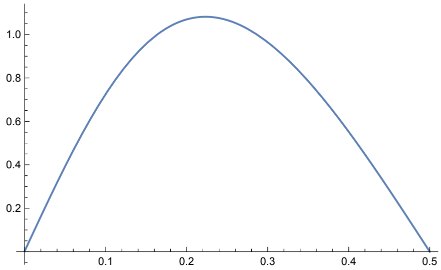


Figure 7. $\psi'(s) - \psi'(1-s)$.

Proof. We have

$$\frac{d}{dx}(x - \operatorname{arccosh}(x)) = 1 - \frac{1}{\sqrt{x-1}\sqrt{1+x}} > 0$$

for $x > \sqrt{2}$, so $x - \operatorname{arccosh}(x)$ is increasing on $(\sqrt{2}, \infty)$. If $y = 1/x$, then the function $y - \operatorname{arccosh}(y) - \frac{1}{2}$ decreases on $(0, 1/\sqrt{2})$. Since $\sqrt{2} - \operatorname{arccosh}(\sqrt{2}) - \frac{1}{2} > 0$, we have $\operatorname{arccosh}(y) < y - \frac{1}{2}$ on $(0, 1/\sqrt{2})$. Therefore

$$c < a - \frac{a^2}{2} \le \frac{1}{\sqrt{2}} - \frac{1}{4} < \frac{1}{2}.$$

Consider the function

$$I(x) = \int_{x-1}^{x} f_1(x)\,dx + \int_{x-1}^{x} f_2(x)\,dx$$

which sends x to the weighted length of $[x-1, x]$. Then

$$\begin{aligned} I'(x) &= f_1(x) - f_1(x-1) + f_2(x) - f_2(x-1) \\ &= [f_2(x) - f_1(x-1)] + [f_1(x) - f_2(x-1)]. \end{aligned}$$

For $|x| < \frac{1}{2}$, both the bracketed quantities are negative, so I is decreasing on $[0, c]$. We have

$$I(0) = \int_{-1}^{0} f_2(x)\,dx + \int_{-1}^{x} f_1(x)\,dx = \int_{-1}^{1} f_2(x)\,dx < \int_{-1}^{\infty} f_2(x)\,dx = \tfrac{1}{4}.$$

Therefore if we can show that $s - t \le 1$, we will have that the f-weighted length of $[s, t]$ is less than $I(s) \le \frac{1}{4}$ and be done. This follows immediately when $t = -s$, since $s \le c < \frac{1}{2}$. When $t \ne -s$, we observe that $s - 1$ is to the left of $-c$, so it suffices to show that $\psi'(s-1) \ge \psi'(t) = -\psi'(s)$. Thus we want to show that

$$\begin{aligned} \psi'(s-1) + \psi'(s) &= \psi'(s) - \psi'(1-s) \\ &= ([(1-s) - s] + [\tanh(s/a^2) - \tanh((1-s)/a^2)])/(a^2) \ge 0 \end{aligned}$$

on $\left[0, \frac{1}{2}\right]$; see Figure 7.

This is equivalent to showing that

$$\gamma(s) := \left([(1-s)-s] + \left[\tanh\left(\frac{s}{a^2}\right) - \tanh\left(\frac{1-s}{a^2}\right)\right]\right) \geq 0$$

on $[0, c]$. Since $|\tanh| < 1$, we have $\gamma(0) > 0$. In addition, $\gamma\left(\frac{1}{2}\right) = 0$. Therefore it suffices to show that γ achieves its minimum value on $\left[0, \frac{1}{2}\right]$ at $s = \frac{1}{2}$. We will do this by using the first derivative test to show that there is only one other local extremum in the interval and further demonstrating that this local extremum is not the minimum point.

We have

$$\gamma'(s) = \frac{\operatorname{sech}^2(s/a^2)}{a^2} + \frac{\operatorname{sech}^2((1-s)/a^2)}{a^2} - 2.$$

Since $1/a^2 \geq 2$, we have $\gamma'(0) > 0$. In addition,

$$\gamma'\left(\tfrac{1}{2}\right) = \frac{2\operatorname{sech}^2(1/(2a^2))}{a^2} - 2.$$

By using the substitution

$$\operatorname{sech}^2(x) = \frac{4}{e^{2x} + e^{-2x} + 2},$$

we get

$$\operatorname{sech}^2\left(\frac{x}{2}\right) = \frac{4}{e^{1/x} + e^{-1/x} + 2} \leq \frac{4}{e^{1/x} + 2}.$$

Therefore

$$\operatorname{sech}^2\left(\frac{x}{2}\right)\left(\frac{1}{x}\right) \leq \frac{4}{xe^{1/x} + 2x}.$$

We have

$$\alpha(x) := (xe^{1/x} + 2x)' = (2 + e^{1/x} - e^{1/x}/x).$$

When $0 < x \leq \frac{1}{2}$, we have

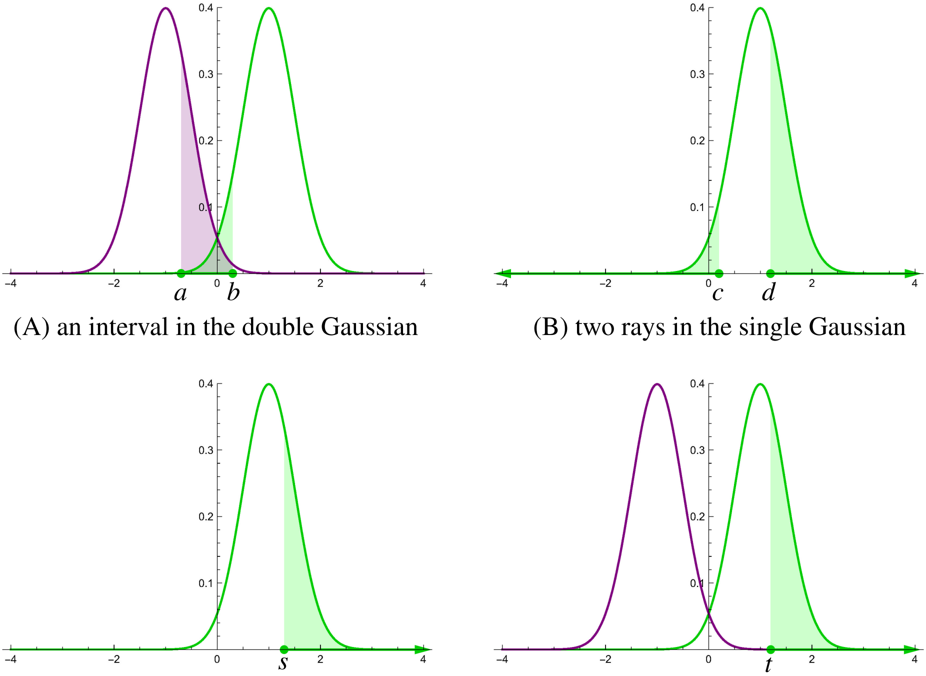
$$\alpha(x) \leq 2 + e^{1/x} - 2e^{1/x} = 2 - e^{1/x} \leq 2 - e^2 < 0.$$

Therefore $\alpha(x)$ attains a minimum value of $\frac{1}{2}e^2 + 1 > 4$ on $\left(0, \frac{1}{2}\right]$. This shows that

$$\operatorname{sech}^2\left(\frac{x}{2}\right)\left(\frac{1}{x}\right) \leq \frac{4}{xe^{1/x} + 2x} < 1$$

on $\left(0, \frac{1}{2}\right]$, so $\gamma'\left(\frac{1}{2}\right) < 0$.

By the intermediate value theorem, there exists $z_1 \in \left(0, \frac{1}{2}\right)$ such that $\gamma'(z_1) = 0$. It follows that $z_2 = 1 - z_1 > \frac{1}{2}$ is also a zero of γ'. Now $\operatorname{sech}^2(x) = \operatorname{sech}^2(-x)$ tends to 0 as x tends to ∞, so $\gamma' < 0$ for some $s \ll 0$. Therefore there exists z_3 in $(-\infty, 0)$ such that $\gamma'(z_3) = 0$, and $z_4 = 1 - z_3 > 1$ is also a zero of γ'.



(A) an interval in the double Gaussian

(B) two rays in the single Gaussian

(C) a ray in the single Gaussian

(D) a ray in the double Gaussian

Figure 8. When all the areas are the same, we have (A) > (B) > (C) and (D) > (C).

Again using the substitution

$$\operatorname{sech}^2(x) = \frac{4}{e^{2x} + e^{-2x} + 2},$$

we see that $\gamma'(s)$ is a rational function of e^{2s/a^2} whose numerator is quartic. Therefore γ' has at most four zeros, so z_1 is the only zero of γ' in $\left(0, \frac{1}{2}\right)$. Since $\gamma'(0) > 0$,

$$\gamma(z_1) > \gamma(0) > \gamma\left(\tfrac{1}{2}\right),$$

so $\gamma(s) \geq \gamma\left(\frac{1}{2}\right) = 0$ for $s \in \left[0, \frac{1}{2}\right]$. □

Proposition 3.11. *If the variance satisfies* $0 < a^2 \leq \frac{1}{2}$, *then the isoperimetric boundaries B with one point b in* $[0, c]$ *cannot be of type* (3) *in Proposition 3.9.*

Proof. Let A be the weighted length of B. If, in contradiction to the proposition, B is of type (3) in Proposition 3.9, then B is of the form $[a, b]$, where $-1 < a < b < 1$, as shown in Figure 8(A). By Proposition 3.6, there exists a union of rays $(-\infty, c] \cup [d, \infty)$ with f_1-weighted length A such that $f_1(c) + f_1(d) < f(a) + f(b)$. This is shown in Figure 8(B). By the solution to the single Gaussian isoperimetric

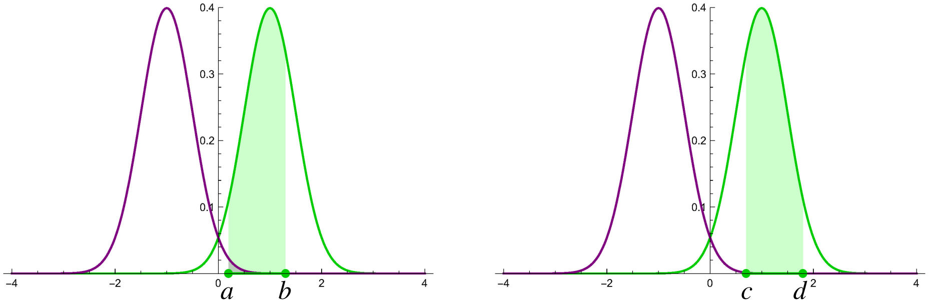


Figure 9. Left: original ray. Right: reflected ray.

problem, there exists a ray $[s, \infty)$, as shown in Figure 8(C), with f_1-weighted length A such that $f_1(t) < f_1(c) + f_1(d)$. By Proposition 3.10, $A \leq \frac{1}{4}$, so $s \geq 1$. By Proposition 3.7, there exists a ray $[t, \infty)$, as shown in Figure 8(D), with f-weighted length A such that $t > s$.

To get a contradiction to the fact that B is isoperimetric, we show $(f(a)+f(b)) - f(t) > 0$. Write

$$(f(a)+f(b)) - f(t) = [(f(a)+f(b)) - (f_1(c)+f_1(d))] \\ + [(f_1(c)+f_1(d)) - f_1(s)] + [f_1(s) - f(t)].$$

Since $[(f_1(c)+f_1(d)) - f_1(s)] > 0$, it suffices to show that $[(f(a)+f(b)) - (f_1(c)+f_1(d))] > [f(t) - f_1(s)]$. Since $f(a) > f_1(d)$, we have

$$[(f(a)+f(b)) - (f_1(c)+f_1(d))] > f(b) - f_1(c) > f(b) - f_1(b) = f_2(b).$$

Since $f(t) < f(s)$, we have

$$[f(t) - f_1(s)] < f(s) - f_1(s) = f_2(s).$$

Since $-1 < b < 1 < s$, we have $f_2(s) < f_2(b)$, and this proves the claim. □

Proposition 3.12. *If the variance satisfies $a^2 \leq \frac{1}{2}$, then the isoperimetric boundaries B with one point $b > 0$ in $[-c, c]$ cannot be of type (2) in Proposition 3.9.*

Proof. We know $f(a) < f(b)$ (recall the concavity/convexity argument), and since $f_2(b) < f_2(a)$, we must have $f_1(a) < f_1(b)$.

Pick $d > c$ such that $f_1(c) = f_1(b)$ and $f_1(d) = f_1(a)$. In other words, we get $[c, d]$ by reflecting $[a, b]$ over the line $x = 1$. See Figure 9. Since $a < 1$, we either have $c < 1 < d$ or $1 < d < c$.

In the first case, we have that $[c, d]$ has the same f_1-length as $[a, b]$, and since $c > a$ and $d > b$, we have $f_2(c) < f_2(a)$ and $f_2(d) < f_2(b)$. Therefore $f(c)+f(d) < f(a)+f(b)$. At the same time, the f_2-length of $[c, d]$ is less than that of $[a, b]$. This difference is at most the f_2-length of $[a, \infty)$. Since $f_1(d) = f_1(a) > f_2(a)$, we can find $e > d$ such that $[c, e]$ has f-length A. In addition, $f(c) + f(e) < f(c) + f(d) < f(a) + f(b)$, so $[a, b]$ is not isoperimetric.

In the second case, we have that $[d, c]$ has the same f_1-length as $[a, b]$, and since $d, c > a, b$, we have $f_2(d) < f_2(a)$ and $f_2(c) < f_2(b)$. Therefore $f(c) + f(d) < f(a) + f(b)$. At the same time, the f_2-length of $[d, c]$ is less than that of $[a, b]$. This difference is at most the f_2-length of $[a, \infty)$. Since $f_1(c) = f_1(a) > f_2(a)$, we can find $e > c$ such that $[d, e]$ has the f-length A. In addition, $f(c) + f(e) < f(c) + f(d) < f(a) + f(b)$, so $[a, b]$ is not isoperimetric. □

Proposition 3.13. *If the variance satisfies $a^2 \leq \frac{1}{2}$, then the isoperimetric boundaries B with one point b in $[-c, c]$ cannot be of type (4) in Proposition 3.9.*

Proof. We may assume without loss of generality that $b \geq 0$. Suppose B is of type (4) in Proposition 3.9. Then the region L enclosed by B consists of the union of an interval of type (2) or (3) in Proposition 3.9 and a ray. Apply Propositions 3.11 and 3.12 to get a new region L' that beats the interval. Since $A < \frac{1}{2}$, L' may be chosen to not intersect the ray. Then the union of L' and the ray beats L. □

Proposition 3.14. *If B is an isoperimetric boundary and the variance satisfies $a^2 \leq \frac{1}{2}$, then B is a single point.*

Proof. If B does not contain a point $s \in [-c, c]$, then by Corollary 2.7, then B is a single point. Otherwise, apply Propositions 3.11–3.13 to complete the proof. □

Proposition 3.15. *For the line endowed with density $f(x)$, if the variance a^2 is such that $\frac{1}{2} \leq a^2 < 1$, then isoperimetric regions R are always rays with boundary B consisting of a single point.*

Proof. By Proposition 3.8, we have that $\psi''(x) = 0$ exactly when x is $c = \pm a^2 \operatorname{arccosh}(1/a)$. Since $\psi'''(x) > 0$ for $x < 0$ and $\psi'''(x) < 0$ for $x > 0$, we have that ψ'' is negative outside of $[-c, c]$ and is positive in $(-c, c)$.

Suppose that B is an isoperimetric boundary containing more than two points. By Corollary 2.7, B does not lie entirely outside $[-c, c]$. Since $\psi''(x) > 0$ on $(-c, c)$ and $\psi''(\pm c) = 0$, the maximum and minimum of $\psi'(x)$ on $[-c, c]$ are achieved at c and $-c$ with $\psi'(-c)$ negative and $\psi'(c)$ positive. Since $\psi'(x)$ tends to $-\infty$ as x approaches ∞, there exists a unique point $b > c$ such that $f(\pm b) = f(\pm c)$. Since $b > c$, we have $\psi''(x) < 0$ outside of $[-b, b]$.

We claim that B must lie in $[-b, b]$. Since $|\psi'(x)|$ is constant on B, to show that $B \subset [-b, b]$ it suffices to show that the maximum and minimum of $\psi'(x)$ on $[-b, b]$ are achieved at $-b$ and b. Since 0 is a local minimum for $f(x)$, it suffices to show that $|\psi'(b)| > |\psi'(c)|$. Since $\psi'(c)$ is postive and $\psi''(x) < 0$ for $x > c$, there exists a unique point $d > c$ where $\psi'(d) = 0$ and ψ' changes from positive to negative at d. To apply Lemma 3.5, consider functions p and q denoting the

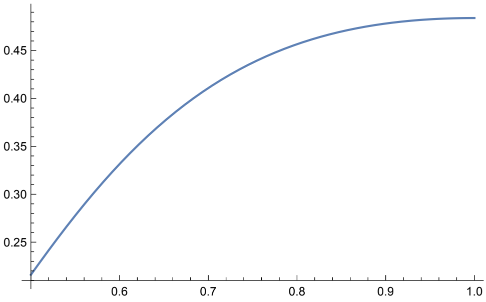


Figure 10. $2f(0, a)$ for various values of a.

increase in ψ moving left of d and the decrease in ψ moving right of d:

$$p(x) = \psi(d) - \psi(d - x),$$
$$q(x) = g(x) = \psi(d) - \psi(d + x),$$

which satisfy the hypotheses of Lemma 3.5. Since $\psi(c) = \psi(b)$, we have

$$|\psi'(c)| = \psi'(c) = p'(d - c) < g'(b - d) - \psi'(b) = |\psi'(b)|.$$

There are five candidates for the minimum points of $f(x)$ on $[-b, b]$: $\pm b$, 0, and $\pm d$. Since $d > c$, we have $\psi''(d) < 0$, so $\pm d$ is not a candidate. Since, also by the preceding paragraph, $\psi'(x)$ is positive between 0 and c, we have $f(b) = f(c) > f(0)$. Therefore the minimum on this interval is $f(0)$. We have

$$\frac{d}{da}(f(0, a)) = -\frac{\sqrt{1/a^2}(-1 + a^2)e^{-1/(2a^2)}}{a^3\sqrt{2\pi}} > 0$$

for all $a \in [-1/\sqrt{2}, 1)$. Therefore we have

$$2f(0, a) \geq 2f(0, 1/\sqrt{2}) \approx 0.415107\ldots.$$

See Figure 10.

To finish the proof, we must show that $f(x, a) < 0.415107\ldots$ for all x and all $a \in [1/\sqrt{2}, 1)$. Consider the numerator n of f given by

$$n(x) = e^{-(x-1)^2/(2a^2)} + e^{-(x+1)^2/(2a^2)}.$$

For a given x, we have n increases when a increases, so

$$n(x) \leq m(x) = e^{-(x-1)^2/2} + e^{-(x+1)^2/2}.$$

Since

$$\frac{d}{dx}(\log m(x)) = \tanh(x) - x,$$

which has the same sign as $-x$, we see that $m(x)$ is maximized at 0. Therefore $n(x) \leq m(0) < 1.22$, so

$$f(x) < \frac{m(0)}{2\sqrt{2\pi}a} \leq \frac{1.22}{2\sqrt{\pi}} \approx 0.345.$$

This means that there is a ray which beats B, contradicting the fact that B is isoperimetric. □

Theorem 3.16. *The isoperimetric boundaries for the double Gaussian density f are always single points enclosing rays.*

Proof. To cover the three cases, apply Propositions 3.4, 3.14, and 3.15. □

4. Isoperimetric regions on the double Gaussian plane

This section describes evidence for the conjecture of Cañete et al., given here as Conjecture 4.1, which states that double Gaussian isoperimetric boundaries in the plane are vertical lines. Proposition 4.4 proves that horizontal and vertical lines are the only stationary lines. Proposition 4.5 proves that vertical lines are better than horizontal lines. First we prove some incidental symmetry results (Propositions 4.2 and 4.3).

Conjecture 4.1 [Cañete et al. 2010, Question 6]. *Let $f(x, y) = e^{\psi(x,y)}$ be the normalized sum of two Gaussian densities with the same variance and different centers. Isoperimetric regions are half-planes enclosed by lines perpendicular to the line connecting the two centers.*

By the planar analogue of Proposition 3.1, it suffices to prove this conjecture in the case where the centers are $c_1 = (1, 0)$ and $c_2 = (-1, 0)$.

Then we have

$$f(x, y) = e^{\psi(x,y)} = \frac{1}{4\pi a^2} e^{-y^2/(2a^2)} \left(e^{-(x-1)^2/(2a^2)} + e^{-(x+1)^2/(2a^2)}\right).$$

The next two propositions describe some symmetry properties of isoperimetric curves. For a curve C, let A_C denote the weighted area enclosed by C.

Proposition 4.2. *Consider a density g symmetric about the x-axis. If a closed, embedded curve C encloses the same weighted area above and below the x-axis, then there is a curve C' which is symmetric about the x-axis, encloses the same weighted area, and has weighted perimeter no greater than that of C.*

Proof. Let C_1 and C_2 be the parts of C in the open upper and lower half-planes chosen so that the weighted perimeter of C_1 is no bigger than that of C_2. Consider the curve C' formed by joining C_1 with its reflection over the x-axis and taking the closure. Let w denote the part of C on the x-axis and w_1 denote the part of C' on the x-axis. Since g is symmetric about the x-axis, $A_C = A_{C'}$. In addition,

$$|C'| - |C| = (2|C_1| + |w_1|) - (|C_1| + |C_2| + |w|) = (|C_1| - |C_2|) + (|w_1| - |w|).$$

We have $|C_1| - |C_2| \leq 0$ by assumption, and since the part of C which intersects the x-axis must include w_1, we know $|w_1| - |w| < 0$. Therefore $|C'| - |C| \leq 0$. □

Proposition 4.3. *Consider a density symmetric about the x-axis. If C is a closed embedded planar curve symmetric about the x-axis, then the part C' of C in the open upper half-plane encloses half as much weighted area with half the weighted length.*

Proof. Suppose that C is a curve that is symmetric about the x-axis and encloses area A. Since C is symmetric about the x-axis, C cannot have nonzero perimeter on the x-axis. Then C' encloses area $\frac{1}{2}A_C$ in the upper half-plane and has weighted perimeter $\frac{1}{2}|C|$. □

Proposition 4.4. *If the plane is endowed with density f, then horizontal and vertical lines have generalized curvature* 0 *and are the only lines which have constant generalized curvature.*

Proof. Let $\psi = \ln f$. Then

$$\nabla\psi(x, y) = \left(\frac{-x + \tanh(x/a^2)}{a^2}, \frac{-y}{a^2}\right).$$

In addition, the normal to the line $y = cx + b$ is $(-c, 1)/\sqrt{c^2+1}$ at all points of the line. Therefore the generalized curvature of such a line evaluated at $(0, b)$ is

$$0 - \nabla\psi(0, b) \cdot \frac{(-c, 1)}{\sqrt{c^2+1}} = \frac{b}{a^2\sqrt{1+c^2}},$$

and by an analogous computation the generalized curvature evaluated at $(1, c+b)$ is

$$\frac{c+b}{a^2\sqrt{1+c^2}} + \frac{c(-1+\tanh(1/a^2))}{a^2\sqrt{1+c^2}}.$$

Thus the generalized curvatures at $(0, b)$ and $(1, c+b)$ are equal exactly when $c = 0$. This shows that only nonvertical lines that could possibly have constant curvature are the horizontal lines $y = b$. Such lines have normal $(0, 1)$, and this, combined with our formula with the gradient, shows that horizontal lines have constant curvature b/a^2.

An explicit computation of the same variety shows that the vertical line $x = b$ has constant curvature

$$\frac{b - \tanh(b/a^2)}{a^2}.$$

□

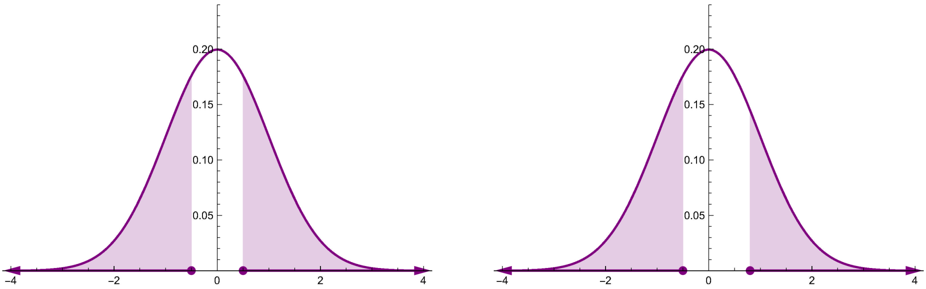


Figure 11. Left: symmetric rays. Right: nonsymmetric rays. When the purple areas are equal, the two nonsymmetric rays are more efficient than the two symmetric rays. The efficiency increases as the disparity between the rays increases, and the limiting case is a single ray, which is the isoperimetric region.

Proposition 4.5. *In the plane with double Gaussian density f, vertical lines enclose a given area with less perimeter than horizontal lines.*

Proof. We now compare the perimeters of and areas enclosed by the horizontal line $x = b$ and the vertical line $y = c$. By symmetry and the fact that we may assume the areas are less than $\frac{1}{2}$, we can assume that b and c are positive and consider the areas of the regions $x > b$ and $y > c$.

The area enclosed by the vertical line is

$$\int_b^\infty \int_{-\infty}^\infty f(x, y)\, dy\, dx = \int_b^\infty \frac{e^{-(x-1)^2/(2a^2)} + e^{-(x+1)^2/(2a^2)}}{2a\sqrt{2\pi}},$$

which is the same as the weighted length of the ray $R_b = [b, \infty)$ on the double Gaussian line. The perimeter of the vertical line is

$$\int_{-\infty}^\infty f(b, y)\, dy = \frac{e^{-(b-1)^2/(2a^2)} + e^{-(b+1)^2/(2a^2)}}{2a\sqrt{2\pi}},$$

which is exactly the cost of R_b on the double Gaussian line.

The area enclosed by the horizontal line is

$$\int_c^\infty \int_{-\infty}^\infty f(x, y)\, dx\, dy = \int_c^\infty \frac{e^{-y^2/(2a^2)}}{2a\sqrt{2\pi}}\, dy,$$

which is the same as the weighted length of the ray $R_c = [c, \infty)$ on the single Gaussian (of total weighted-length 1) line. The perimeter of the horizontal line is

$$\int_{-\infty}^\infty f(x, c)\, dx = \frac{e^{-c^2/(2a^2)}}{2a\sqrt{2\pi}},$$

which is exactly the cost of R_c on the single Gaussian line.

Therefore it suffices to show that a ray on the double Gaussian line of length A costs less than a ray on the single Gaussian line of the same weighted length. Consider the line with density g given by a single Gaussian of total length $\frac{1}{2}$. The ray on the single Gaussian is equivalent to the union of two disjoint, symmetric rays on the g-line. The ray on the double Gaussian is equivalent to the union of two disjoint, nonsymmetric rays on the g-line. By applying the first and second variation arguments to a single Gaussian density, we see that two nonsymmetric rays are always better than two symmetric rays of the same total weighted-length. See Figure 11. □

Therefore if the isoperimetric curve corresponding to area A is a line, then it is a vertical line.

Acknowledgements

This paper is the work of the Williams College NSF "SMALL" 2015 Geometry Group. We thank our advisor Professor Morgan for his support. We would like to thank the National Science Foundation, Williams College, the University of Chicago, and the MAA for supporting the "SMALL" REU and our travel to MathFest 2015.

References

[Cañete et al. 2010] A. Cañete, M. Miranda, Jr., and D. Vittone, "Some isoperimetric problems in planes with density", *J. Geom. Anal.* **20**:2 (2010), 243–290. MR Zbl

[Morgan 2009] F. Morgan, *Geometric measure theory: a beginner's guide*, 4th ed., Academic Press, 2009. MR Zbl

[Rosales et al. 2008] C. Rosales, A. Cañete, V. Bayle, and F. Morgan, "On the isoperimetric problem in Euclidean space with density", *Calc. Var. Partial Differential Equations* **31**:1 (2008), 27–46. MR Zbl

Received: 2016-08-01 Accepted: 2016-08-21

jtb1@williams.edu
*Department of Mathematics and Statistics, Williams College,
Williamstown, MA, United States*

mdannenberg@g.hmc.edu
*Department of Mathematics, Harvey Mudd College,
Claremont, CA, United States*

liangj@uchicago.edu
*Department of Mathematics, University of Chicago,
Chicago, IL, United States*

yzeng@smcm.edu
*Department of Mathematics and Computer Science, St. Mary's
College of Maryland, St. Mary's City, MD, United States*

INVOLVE 11:4 (2018)
dx.doi.org/10.2140/involve.2018.11.569

Finiteness of homological filling functions

Joshua W. Fleming and Eduardo Martínez-Pedroza

(Communicated by Kenneth S. Berenhaut)

Let G be a group. For any $\mathbb{Z}G$-module M and any integer $d > 0$, we define a function $\mathrm{FV}_M^{d+1} : \mathbb{N} \to \mathbb{N} \cup \{\infty\}$ generalizing the notion of $(d+1)$-dimensional filling function of a group. We prove that this function takes only finite values if M is of type FP_{d+1} and $d > 0$, and remark that the asymptotic growth class of this function is an invariant of M. In the particular case that G is a group of type FP_{d+1}, our main result implies that its $(d+1)$-dimensional homological filling function takes only finite values.

1. Introduction

For a contractible cellular complex X and an integer $d > 0$, the homological filling function $\mathrm{FV}_X^{d+1} : \mathbb{N} \to \mathbb{N}$ measures the difficulty of filling cellular d-cycles with $(d+1)$-chains; a precise definition is below. They are higher-dimensional homological generalizations of isoperimetric functions. For a group G admitting a compact classifying space $K(G, 1)$ with universal cover X, the equivalence growth rate of the function FV_X^{d+1} provides an invariant of the group. The initial motivation of this work was to provide a direct argument that FV_X^{d+1} takes only finite values for such complex X, addressing what the authors perceived as a gap in the literature. In this article we provide a self-contained proof based on the algebraic approach to define the homological filling functions from [Hanlon and Martínez-Pedroza 2016], and on our way, we prove a more general result that defines a new collection of invariants for $\mathbb{Z}G$-modules.

The topological perspective. We assume all spaces are combinatorial complexes and all maps are combinatorial; see for example [Bridson and Haefliger 1999, Part I, Chapter 8, Appendix]. A G-action on a complex X is *proper* if for all compact subcomplexes K of X the collection $\{g \in G \mid K \cap g(K) \neq \varnothing\}$ is finite. The G-action is *cocompact* if there is a compact subcomplex K of X such that the collection $\{gK \mid g \in G\}$ covers X. For a complex X, the cellular d-dimensional chain group

MSC2010: primary 20F65, 20J05; secondary 16P99, 28A75, 57M07.
Keywords: Dehn functions, homological filling function, isoperimetric inequalities, finiteness properties of groups.

$C_d(X, \mathbb{Z})$ is a free $\mathbb{Z}$-module with a natural ℓ_1-norm induced by a basis formed by the collection of all d-dimensional cells of X, each cell with a chosen orientation from each pair of opposite orientations. This norm, denoted by $\|\cdot\|_1$, is the sum of the absolute value of the coefficients in the unique representation of the chain as a linear combination of elements of the basis. Let $Z_d(X, \mathbb{Z})$ denote the $\mathbb{Z}$-module of integral d-cycles, and $\partial_{d+1} : C_{d+1}(X, \mathbb{Z}) \to Z_d(X, \mathbb{Z})$ be the boundary map. The *$(d+1)$-dimensional filling function of X* is the function $\mathrm{FV}_X^{d+1} : \mathbb{N} \to \mathbb{N} \cup \{\infty\}$ defined as

$$\mathrm{FV}_X^{d+1}(k) = \sup\{\|\gamma\|_\partial \mid \gamma \in Z_d(X, \mathbb{Z}),\ \|\gamma\|_1 \le k\},$$

where

$$\|\gamma\|_\partial = \inf\{\|\mu\|_1 \mid \mu \in C_{d+1}(X, \mathbb{Z}),\ \partial(\mu) = \gamma\},$$

where the supremum and infimum of the empty set are defined as zero and ∞ respectively. In words, $\mathrm{FV}_X^{d+1}(k)$ is the most efficient upper bound on the size of fillings by $(d+1)$-chains of d-cycles of norm at most k. A complex X is *d-acyclic* if the reduced homology groups $\overline{H}_i(X, \mathbb{Z})$ are trivial for $0 \le i \le d$. As mentioned above, the initial motivation of this work was to provide a proof of Theorem 1.1, which the authors perceived as a gap in the literature. The main contribution of this note is a generalization to an algebraic framework of the following statement; see Theorem 1.3.

Theorem 1.1. *Let d be a positive integer and let G be a group acting properly and cocompactly by cellular automorphisms on a d-acyclic complex X. Then $FV_X^{d+1}(m)$ is finite for all $m \in \mathbb{N}$.*

Theorem 1.1 was known to hold in the following cases:

- For $d = 1$, it is a result of [Gersten 1999, Proposition 2.4].
- For $d \ge 1$ and under the extra assumption that G admits a combing, it follows from [Epstein et al. 1992, Theorem 10.3.6]; see also [Behrstock and Druţu 2015, Lemma 3.7].
- For $d \ge 3$, Hanlon and the second author observed in [Hanlon and Martínez-Pedroza 2016, Section 3.3] that Theorem 1.1 holds using results of Alonso, Pride and Wang [Alonso et al. 1999] in conjunction with an argument from Abrams, Brady, Dani and Young [Abrams et al. 2013]. The results in [Alonso et al. 1999] rely on nontrivial machinery from homotopy theory. The failure of the argument for $d = 2$ relies on an application of the Hurewicz theorem; for details see [Hanlon and Martínez-Pedroza 2016, Section 3.3].

Current results in the literature leave open the statement of Theorem 1.1 for the case $d = 2$. Our argument in this note proving Theorem 1.1 does not rely on previous results, it is valid for all $d > 0$, and it is elementary. The argument might be known to the experts, but to our knowledge does not appear in the literature, and this note fills this gap. Let us sketch the argument from a topological perspective; for an algebraic proof see Section 2.

Sketch of the proof of Theorem 1.1, from a topological perspective. Consider the combinatorial path metric on the 1-skeleton of X, and for any d-cycle σ (which is a formal finite sum of d-cells) define its diameter $\operatorname{diam}(\sigma)$ as the diameter of the set consisting of vertices (0-cells) which are in the closure of at least one d-cell defining σ. A d-cycle σ is called *connected* if the subcomplex of X formed by taking the closure of the union of d-cells defining σ is connected (and has no cut-points).

Let $m > 0$. Since G acts properly and cocompactly on X, there is an integer $C \geq 0$ that bounds the diameter of any d-cell of X, and hence for any connected d-cycle σ,

$$\operatorname{diam}(\sigma) \leq C\|\sigma\|_1.$$

From here, it follows that the induced G-action on the set of connected d-cycles with ℓ_1-norm $\leq m$ has finitely many G-orbits. Since X is d-acyclic, $\|\sigma\|_\partial < \infty$ for each d-cycle σ. Therefore, there is an integer $M = M(m)$ such that

$$\sigma \text{ is connected and } \|\sigma\|_1 \leq m \quad \Longrightarrow \quad \|\sigma\|_\partial \leq M.$$

Let σ be an arbitrary d-cycle with ℓ_1-norm $\leq m$. Then one shows that σ can be decomposed as a sum of connected d-cycles $\sum_{i=1}^k \sigma_i$, where $k \leq \|\sigma\|_1 = \sum_{i=1}^k \|\sigma_i\|_1$. Hence

$$\|\sigma\|_\partial \leq \sum_{i=1}^{k} \|\sigma_i\|_\partial \leq k \cdot M \leq m \cdot M.$$

Therefore $\mathrm{FV}_X^{d+1}(m) \leq m \cdot M < \infty$. □

Remark 1.2. Under the assumptions of Theorem 1.1, it is known that the growth rate of the function FV_X^{d+1} is a quasi-isometry invariant of the group G. This was first addressed by Fletcher [1998, Theorem 2.1] under the assumption that X is the universal cover of $K(G, 1)$. Young [2011, Lemma 1] provided a proof of the quasi-isometry invariance in the general context of Theorem 1.1. Notably, these works do not address that these functions are finite.

The algebraic perspective, and our main result. Our main result is an algebraic analog of Theorem 1.1. Recall that for a group G, a $\mathbb{Z}G$-module M is of type FP_n if there exists a partial resolution of $\mathbb{Z}G$-modules

$$P_n \xrightarrow{\varphi_n} P_{n-1} \xrightarrow{\varphi_{n-1}} \cdots \xrightarrow{\varphi_2} P_1 \xrightarrow{\varphi_1} P_0 \to M \to 0$$

such that each P_i is a finitely generated projective $\mathbb{Z}G$-module. For a $\mathbb{Z}G$-module M of type FP_{d+1} we define the $(d+1)$-filling function FV_M^{d+1} of M, see Definition 2.5, and prove the following result.

Recall that the *growth rate class* of a function $\mathbb{N} \to \mathbb{N}$ is defined as follows. Given two functions $f, g : \mathbb{N} \to \mathbb{N}$, define the relation $f \preceq g$ if there is $C > 0$ such that $f(n) \le Cg(Cn + C) + Cn + C$ for all $n \in \mathbb{N}$, and let $f \sim g$ if both $f \preceq g$ and $g \preceq f$. This yields an equivalence relation where the equivalence class of a function f is called the *growth rate class of* f.

Theorem 1.3. *Let M be a $\mathbb{Z}G$-module of type FP_{d+1}:*

(1) *For all positive integers d and k, we have* $\mathrm{FV}_M^{d+1}(k) < \infty$.

(2) *The growth rate of the function* $\mathrm{FV}_M^{d+1} : \mathbb{N} \to \mathbb{N}$ *only depends on M.*

This result provides a new collection of invariants for $\mathbb{Z}G$-modules that remains to be studied. The invariant is interesting even in the case that $M = \mathbb{Z}$ and G is suitable. In this case, the filling functions $\mathrm{FV}_{\mathbb{Z}}^{d+1}$ correspond to the filling invariants of the group G, usually denoted by FV_G^{d+1}, in the context of Theorem 1.1 and Remark 1.2. There are computations by Young [2016] in the case that G corresponds to a discrete Heisenberg group answering a conjecture of Gromov [1993, Chapter 5], estimations in the case that G is the special linear group $\mathrm{SL}(n, \mathbb{Z})$ by Epstein and Thurston [1992, Chapter 10], and general results in the case that G is a hyperbolic group by Gersten [1996] and Mineyev [2000], among others. In [Hanlon and Martínez-Pedroza 2016, Remark 3.4], it was observed that there was no proof in the literature that if G is of type FP_3 (i.e., $\mathbb{Z}$ is of type FP_3 as a module over $\mathbb{Z}G$) then FV_G^3 is finite-valued; observe that this is a consequence of Theorem 1.3.

This note contains a proof of the first statement of Theorem 1.3. The proof of the second statement appears in [Hanlon and Martínez-Pedroza 2016, Theorem 3.5] for the case that $M = \mathbb{Z}$, but the argument works verbatim for the general case.

Organization. The rest of the paper is organized as follows: Section 2 contains some preliminary definitions including the definition of FV_M^{d+1}, the statement of the main technical result of the article, Proposition 2.4, and arguments implying Theorems 1.1 and 1.3. Section 3 is devoted to the proof of Proposition 2.4. Section 4 discusses some geometric examples illustrating some matters about Theorem 1.1.

2. Main technical result and proofs of the main theorems

Let G be a group and let S be a G-set. The set of all orbits of S under the G-action is denoted by S/G. The free abelian group $\mathbb{Z}[S]$ with S as a free generating set can be made into a $\mathbb{Z}G$-module that we shall call the permutation module on S. The $\mathbb{Z}$-basis S induces a G-equivariant norm, called the ℓ_1-*norm*, given by $\left\| \sum_{s \in S} n_s s \right\|_S = \sum_{s \in S} |n_s|$, where $n_s \in \mathbb{Z}$.

If the G-action on S is free, then $\mathbb{Z}[S]$ is a free module over $\mathbb{Z}S$. Conversely, if F is a free $\mathbb{Z}G$-module with a chosen $\mathbb{Z}G$-basis $\{\alpha_i \mid i \in I\}$, then F is isomorphic to the permutation module $\mathbb{Z}[S]$, where $S = \{g\alpha_i \mid g \in G,\ i \in I\}$ with the natural G-action. In this case the $\mathbb{Z}G$-basis $\{\alpha_i \mid i \in I\}$ of F induces an ℓ_1-norm as before.

Definition 2.1 (Gersten's filling norms). Let $\eta : F \to M$ be a surjective morphism of $\mathbb{Z}G$-modules where F is finitely generated and free with a chosen finite $\mathbb{Z}G$-basis, and the induced *filling norm* on M is defined by

$$\|m\|_\eta = \min\{\|x\|_F \mid x \in F, \eta(x) = m\},$$

where $\|\cdot\|_F$ denotes the induced ℓ_1-norm on F.

Remark 2.2 (induced ℓ_1-norms are filling norms). Let $\mathbb{Z}[S]$ be a permutation $\mathbb{Z}G$-module such that G acts freely on S and the quotient S/G is finite. Then $\mathbb{Z}[S]$ is a finitely generated free $\mathbb{Z}G$-module and the ℓ_1-norm $\|\cdot\|_S$ is a filling norm. This statement holds without the assumption that G acts freely on S. Since we do not use this fact, we leave its verification to the reader.

Definition 2.3. Let $\rho : \mathbb{Z}[S] \to \mathbb{Z}[T]$ be a morphism of permutation $\mathbb{Z}G$-modules such that the kernel $K = \ker\rho$ is finitely generated. Let $\|\cdot\|_K$ denote a filling norm on K and let $\|\cdot\|_S$ denote the ℓ_1-norm on $\mathbb{Z}[S]$ induced by S. Define the function $\mathrm{FV}_\rho : \mathbb{N} \to \mathbb{N} \cup \{\infty\}$ as

$$\mathrm{FV}_\rho(n) = \sup\{\|x\|_K \mid x \in K,\ \|x\|_S \leq n\}.$$

Proposition 2.4. *Let $\rho : \mathbb{Z}[S] \to \mathbb{Z}[T]$ be a morphism. Suppose that S/G and T/G are finite, T has finite G-stabilizers for all $t \in T$, and* $\ker\rho$ *is finitely generated. Then* $\mathrm{FV}_\rho(n) < \infty$ *for all $n \in \mathbb{N}$.*

In the rest of this section, we deduce Theorems 1.1 and 1.3 from Proposition 2.4.

Proof of Theorem 1.1. Let G be a group acting properly and compactly by cellular automorphisms on a d-connected complex X. The free abelian groups $C_d(X)$ and $C_{d+1}(X)$ are permutation $\mathbb{Z}G$-modules over the G-sets of d-cells and $(d+1)$-cells of X, respectively. Observe that the definition of FV_X^{d+1} coincides with Definition 2.3 of FV_{∂_d} for the boundary map $C_d(X) \xrightarrow{\partial_d} C_{d-1}(X)$. The proof concludes by verifying the hypothesis of Proposition 2.4 for this morphism.

Since the G-action on X is cocompact, there are finitely many G-orbits of d-cells and $(d+1)$-cells; in particular $C_{d+1}(X)$ is a finitely generated $\mathbb{Z}G$-module. Since X is d-acyclic, the sequence

$$C_{d+1}(X) \xrightarrow{\partial_{d+1}} C_d(X) \xrightarrow{\partial_{d+1}} C_{d-1}(X)$$

is exact and hence $\ker(\partial_d)$ is a finitely generated $\mathbb{Z}G$-module. Since the G-action is proper, the stabilizer of each d-cell of X is finite. □

Definition 2.5. Let M be a $\mathbb{Z}G$-module of type FP_{d+1}. The $(d+1)$*-filling function of* M is the function

$$\mathrm{FV}_M^{d+1} : \mathbb{N} \to \mathbb{N} \cup \{\infty\}$$

defined as follows. Let

$$P_{d+1} \xrightarrow{\varphi_{d+1}} P_d \xrightarrow{\varphi_d} \cdots \xrightarrow{\varphi_2} P_1 \xrightarrow{\varphi_1} P_0 \to M \to 0$$

be a FP_{d+1}-resolution for M. Chose filling norms on P_{d+1} and P_d denoted by $\|\cdot\|_{P_{d+1}}$ and $\|\cdot\|_{P_d}$ respectively. Then

$$\mathrm{FV}_M^{d+1}(k) = \sup\{\|x\|_{\varphi_{d+1}} \mid x \in \ker \varphi_d,\ \|x\|_{P_d} \leq k\},$$

where

$$\|x\|_{\varphi_{d+1}} = \min\{\|y\|_{P_{d+1}} \mid y \in P_{d+1},\ \varphi_{d+1}(y) = x\}.$$

The proof of Theorem 1.3 uses the following lemma.

Lemma 2.6 [Brown 1982, Chapter VIII, Proposition 4.3]. *A $\mathbb{Z}G$-module M is of type FP_d if and only if M admits a partial resolution of free finitely generated $\mathbb{Z}G$-modules of the form*

$$F_{d+1} \to F_d \to \cdots \to F_1 \to F_0 \to M \to 0.$$

Proof of Theorem 1.3. Since M is of type FP_{d+1}, by Lemma 2.6, there exists a partial resolution of free and finitely generated $\mathbb{Z}G$-modules

$$F_{d+1} \xrightarrow{\varphi_{d+1}} F_d \xrightarrow{\varphi_d} \cdots \xrightarrow{\varphi_2} F_1 \xrightarrow{\varphi_1} F_0 \to M \to 0$$

such that $\ker \varphi_n$ is finitely generated for n such that $d \geq n \geq 0$. Consider the finitely generated free modules F_d and F_{d-1} as permutation modules $\mathbb{Z}[S]$ and $\mathbb{Z}[T]$ respectively. Finite generation and freeness implies that we can assume that G acts freely and with finitely many orbits on both S and T. Since the induced ℓ_1-norms on $\mathbb{Z}[S]$ and $\mathbb{Z}[T]$ are in particular filling norms, the definition of FV_M^{d+1} coincides with Definition 2.3 of $\mathrm{FV}_{\varphi_{d+1}}$. Then the first statement of the theorem on the finiteness of FV_M^{d+1} follows by applying Proposition 2.4 to $\mathrm{FV}_{\varphi_{d+1}}$.

The proof of the second statement that the growth rate of FV_M^{d+1} is independent of the choice of partial resolution and filling norms appears in [Hanlon and Martínez-Pedroza 2016, Theorem 3.5] for the case that $M = \mathbb{Z}$ and G is a group of type FP_{d+1}. The argument for arbitrary M follows verbatim by replacing each occurrence of $\mathbb{Z}$ by M. Let us remark that the heart of the argument is the fact that any two projective resolutions of M are chain homotopy equivalent [Brown 1982, p. 24, Theorem 7]. □

3. Finiteness

This section contains the proof Proposition 2.4. Let S and T be G-sets. For $x \in \mathbb{Z}[S]$ with $x = \sum_{s\in S} n_s s$, we denote by $\langle x, s\rangle$ the integer n_s. For $x \in \mathbb{Z}[T]$ and $t \in T$ we define analogously $\langle x, t\rangle$.

Definition 3.1 (x is a part of y). Let $x, y \in \mathbb{Z}[S]$. We say x *is a part of* y, denoted by $x \preceq_S y$, to mean that for each $s \in S$ if $\langle x, s\rangle > 0$ then $\langle x, s\rangle \leq \langle y, s\rangle$, and if $\langle x, s\rangle < 0$ then $\langle y, s\rangle \leq \langle x, s\rangle$. Note that this is equivalent to $\langle x, s\rangle\langle y, s\rangle \geq \langle x, s\rangle^2$ for all $s \in S$.

Definition 3.2 (S-intersect). For $x, y \in \mathbb{Z}[S]$, the S-intersection of x and y is defined as $x \cap_S y = \{s \in S \mid \langle x, s\rangle\langle y, s\rangle < 0\}$.

Remark 3.3. Let $x, y \in \mathbb{Z}[S]$. Then $\|x + y\|_S = \|x\|_S + \|y\|_S$ if and only if $x \cap_s y = \varnothing$. Indeed,

$$\|x+y\|_S = \sum_{s\in S} |\langle x, s\rangle + \langle y, s\rangle| \leq \sum_{s\in S} |\langle x, s\rangle| + \sum_{s\in S} |\langle y, s\rangle| = \|x\|_S + \|y\|_S$$

with equality if and only if $\langle x, s\rangle$ and $\langle y, s\rangle$ have the same sign for all $s \in S$.

Throughout the rest of this section, let

$$\mathcal{D}_1 = S \cup \{-s \mid s \in S\}.$$

Furthermore, let $\rho : \mathbb{Z}[S] \to \mathbb{Z}[T]$ denote a morphism of $\mathbb{Z}G$-modules.

Definition 3.4 (ρ-intersect). A pair of elements $x, y \in \mathbb{Z}[S]$ have *nontrivial ρ-intersection*, denoted by $x \cap_\rho y \neq \varnothing$, if there exists $x_1, y_1 \in \mathcal{D}_1$ such that $\rho(x_1) \cap_T \rho(y_1) \neq \varnothing$ where $x_1 \preceq_S x$ and $y_1 \preceq_S y$.

Definition 3.5 (ρ-connected). For each integer $n \geq 1$, let $\mathcal{D}_n$ be the collection of elements of $\mathbb{Z}[S]$ of the form $x = \sum_{i=1}^n x_i$, where each $x_i \in \mathcal{D}_1$ and for every $k < n$ the elements $\sum_{i=1}^k x_i$ and x_{k+1} have trivial S-intersection and nontrivial ρ-intersection. An element $x \in \mathbb{Z}[S]$ is *ρ-connected* if $x \in \mathcal{D}_n$ for some $n \geq 1$.

Remark 3.6. For $x \in \mathbb{Z}[S]$, we have $x \in \mathcal{D}_n$ if and only if x is ρ-connected and $\|x\| = n$.

Lemma 3.7. *If* $0 \neq z \in \ker \rho$, *then there exists x such that*

(1) $x \preceq_S z$, *in particular*, $\|z - x\|_S < \|z\|_S$,

(2) $x \in \ker \rho$, *and*

(3) *x is ρ-connected.*

Proof. Let $0 \neq z \in \ker \rho$ be an arbitrary element. Consider the set

$$\Omega = \{x \preceq_S z \mid x \neq 0,\ x \text{ is } \rho\text{-connected}\};$$

this is a nonempty finite set partially ordered by $\preceq_S$. Let $x \in \Omega$ be a maximal element. We claim that $x \in \ker\rho$. Suppose that $x \notin \ker\rho$. We have $\rho(x)$ and $\rho(z-x)$ are nonzero and satisfy

$$\rho(x)+\rho(z-x)=0.$$

Since $\rho(x)\neq 0$ there exists $t \in T$ such that $\langle \rho(x), t\rangle \neq 0$. Therefore

$$\langle \rho(z-x), t\rangle = -\langle \rho(x), t\rangle.$$

Since $\rho(z-x)\neq 0$, there exists $s \in S$ for which

$$\langle z-x, s\rangle\langle \rho(s), t\rangle\langle \rho(z-x), t\rangle > 0.$$

This implies

$$\langle z-x, s\rangle\langle \rho(s), t\rangle\langle \rho(x), t\rangle < 0.$$

Now define $\lambda = \langle z-x, s\rangle / |\langle z-x, s\rangle|$. We show $x+\lambda s$ is ρ-connected. First observe that $x \cap_S \lambda s = \varnothing$ since $x \preceq_S z$ and $\lambda s \preceq_S z$. Moreover, note that $x \cap_\rho \lambda s \neq \varnothing$ since

$$\langle \rho(x), t\rangle\langle \rho(\lambda s), t\rangle = \langle \rho(x), t\rangle\langle \rho(s), t\rangle\lambda < 0.$$

Therefore $x+\lambda s$ is ρ-connected and $x \precneqq_S x+\lambda s \preceq_S z$. This contradicts the maximality of x and therefore $x \in \ker\rho$. □

Proposition 3.8. *For all nonzero $z \in \ker\rho$, there exist ρ-connected elements $x_1 \ldots, x_n \in \ker\rho$ such that*

(1) $z = x_1 + \cdots + x_n$,

(2) $x_i \preceq_S z$ *for each* i.

Proof. Applying Lemma 3.7 to $z \in \ker\rho$, there exists a ρ-connected element $x_1 \in \ker\rho$ such that $x_1 \preceq_S z$. If $z - x_1 \neq 0$ then there exists a ρ-connected element $x_2 \in \ker\rho$ such that $x_2 \preceq_S z - x_1 \prec_S z$. If $z - x_1 - x_2 \neq 0$ then there exists a ρ-connected element $x_3 \in \ker\rho$ such that $x_3 \preceq_S z - x_1 - x_2 \prec z - x_1 \prec z$. This process must terminate for some positive integer n since

$$\|z - x_1 - \cdots - x_k\| > \|z - x_1 - \cdots - x_k - x_{k+1}\| \geq 0$$

if $z - x_1 - \cdots - x_k \neq 0$. Hence we obtain ρ-connected elements $x_1, \ldots, x_n \in \ker\rho$ such that $x_i \preceq_S z$ for each i, and $z = x_1 + \cdots + x_n$. □

Remark 3.9. For $x, y \in \mathbb{Z}[S]$, the relations $x \preceq_S y$, $x \cap_S y \neq \varnothing$, and $x \cap_\rho y \neq \varnothing$ are preserved by the G-action on $\mathbb{Z}[S]$. Thus, if $x \in \mathcal{D}_n$ and $g \in G$ then $gx \in \mathcal{D}_n$. It follows that $\mathcal{D}_n$ is a G-set.

Proposition 3.10. *Suppose that S and T have finitely many G-orbits and each element of T has finite G-stabilizer. Then for every $n \geq 1$, the set $\mathcal{D}_n$ has finitely many G-orbits.*

Before the proof of the Proposition 3.10, we introduce the following lemmas.

Lemma 3.11. *Suppose S has finitely many G-orbits, and each element of T has finite G-stabilizer. Then for every $t \in T$, the set $S(t) = \{s \in S \mid \langle \rho(s), t\rangle \neq 0\}$ is finite.*

Proof. For any $t \in T$, $s \in S$, and $g \in G$, we have $\langle \rho(gs), gt\rangle = \langle \rho(s), t\rangle$. For each $s \in S$, let $T(s) = \{t \in T \mid \langle \rho(s), t\rangle \neq 0\}$. As ρ is a morphism, $T(s)$ is a finite set for all $s \in S$. Now, fix $t \in T$ and let $s_1, \dots, s_m$ be representatives of G-orbits of S. Then

$$S(t) = \bigcup_{i=1}^{m} \bigl\{gs_i \bigm| g \in G,\ \langle \rho(s_i), g^{-1}t\rangle \neq 0\bigr\} = \bigcup_{i=1}^{m} \bigcup_{r \in T(s_i)} \{gs_i \mid g \in G, g^{-1}t = r\}.$$

Observe that the set $\{g \in G \mid g^{-1}t = r\}$ is in one-to-one correspondence with $G_t = \{g \in G \mid gt = t\}$. By assumption, G_t is finite and thus for each $i \in \{1, \dots, m\}$ and $r \in T(s_i)$ the set $\{gs_i \mid g \in G, g^{-1}t = r\}$ is finite. Therefore, the set $S(t)$ is finite. □

Lemma 3.12. *Suppose S has finitely many G-orbits and that T has finite G-stabilizers for each $t \in T$. Then for all $n \in \mathbb{Z}_+$ and for all $y \in \mathcal{D}_n$ the set $\{x \in \mathcal{D}_1 \mid x \cap_\rho y \neq \varnothing\}$ is finite.*

Proof. For $y \in \mathbb{Z}[S]$ denote by $\mathcal{D}_1(y)$ the set $\{x \in \mathcal{D}_1 \mid x \cap_\rho y \neq \varnothing\}$. Let $y \in \mathcal{D}_n$. By definition, $y = \sum_{i=1}^n x_i$, where each $x_i \in \mathcal{D}_1$ and for each $k < n$, the elements $\sum_{i=1}^k x_i$ and x_{k+1} have trivial S-intersection and nontrivial ρ-intersection. It follows from the definition of ρ-intersect that

$$\mathcal{D}_1(y) = \{x \in \mathcal{D}_1 \mid x \cap_\rho y \neq \varnothing\} = \bigcup_{i=1}^{n} \{x \in \mathcal{D}_1 \mid x \cap_\rho x_i \neq \varnothing\} = \bigcup_{i=1}^{n} \mathcal{D}_1(x_i).$$

Therefore, to conclude it is enough to show that $\mathcal{D}_1(s)$ is finite for every $s \in \mathcal{D}_1$.

Let $s \in \mathcal{D}_1$. Observe that

$$\mathcal{D}_1(s) = \bigcup_{t \in T} \bigl\{x \in \mathcal{D}_1 \bigm| \langle \rho(x), t\rangle \langle \rho(s), t\rangle < 0\bigr\} \subset \bigcup_{t \in T} \bigl\{x \in \mathcal{D}_1 \bigm| \langle \rho(x), t\rangle \langle \rho(s), t\rangle \neq 0\bigr\}.$$

It is immediate that $\{t \in T \mid \langle \rho(s), t\rangle \neq 0\}$ is finite. Hence the union on the right is over a collection with finitely many nonempty sets. By Lemma 3.11, for any $t \in T$ the set $\{x \in \mathcal{D}_1 \mid \langle \rho(x), t\rangle \neq 0\}$ is finite, and hence $\{x \in \mathcal{D}_1 \mid \langle \rho(x), t\rangle \langle \rho(s), t\rangle \neq 0\}$ is finite. Therefore the expression on the right is the union of a finite collection of finite sets, and we conclude that $\mathcal{D}_1(s)$ is finite. □

Proof of Proposition 3.10. We prove by induction on n. For $n = 1$ the result follows from the assumption that S has finitely many G-orbits and the definition of $\mathcal{D}_1$.

Suppose $\mathcal{D}_n$ has finitely many G-orbits with representatives $y_1, \dots, y_\ell$. For each $1 \leq k \leq \ell$, let A_k be the collection of elements A_k of $\mathcal{D}_1$ such that

$$y_k \cap_S z = \varnothing \quad \text{and} \quad y_k \cap_\rho z \neq \varnothing.$$

By Lemma 3.12, the collection A_k is finite. The proof concludes with the verification of the following claim.

Claim. *The set*

$$\{y_k + z \mid 1 \leq k \leq \ell \textit{ and } z \in A_k\}$$

is a collection of representatives of G-orbits of $\mathcal{D}_{n+1}$.

Let $x \in \mathcal{D}_{n+1}$. Then $x = \sum_{i=1}^{n+1} x_i$, where each $x_i \in \mathcal{D}_1$ and for every $k < n$ the elements $\sum_{i=1}^{k} x_i$ and x_{k+1} have trivial S-intersection and nontrivial ρ-intersection. By definition, $\sum_{i=1}^{n} x_i$ is in $\mathcal{D}_n$. Hence $\sum_{i=1}^{n} x_i = g y_j$ for some $g \in G$ and some $1 \leq j \leq \ell$. It follows that $x = g y_j + x_{n+1}$ and therefore $g^{-1}x = y_j + g^{-1}x_{n+1}$. By Remark 3.9, we have that $z = g^{-1}x_{n+1}$ is an element of A_j. Therefore $x = g y_i + gz = g(y_i + z)$. This proves the claim. □

Proof of Proposition 2.4. Let K denote $\ker \rho$, and let $\|\cdot\|_K$ denote a chosen filling norm on K. By Proposition 3.10, for each positive integer n, the G-set $\bigcup_{i=1}^{n} \mathcal{D}_i$ has finitely many G-orbits. Therefore, for each $n \in \mathbb{Z}_+$ there is an integer B_n such that for every $x \in \bigcup_{i=1}^{n} \mathcal{D}_i$, we have $\|x\|_K \leq B_n$.

Let $0 \neq z \in K$ such that $\|z\|_S \leq n$. By Proposition 3.8, there exist ρ-connected elements $x_1, \dots, x_m \in K$ such that $m \leq n$, $z = x_1 + \cdots + x_m$, and $x_i \prec z$, $i = 1, \dots, m$. By Remark 3.6, each $x_i \in \mathcal{D}_n$. Therefore, by the triangle inequality,

$$\|z\|_K \leq \sum_{i=1}^{m} \|x_i\|_K \leq m \cdot B_n \leq n \cdot B_n.$$

This shows that $\mathrm{FV}_\rho(n) \leq n \cdot B_n < \infty$. □

Remark 3.13. Observe that Proposition 2.4 can be generalized as follows. Consider the sequence of modules $\ker \rho \to \mathbb{Z}[S] \xrightarrow{\rho} \mathbb{Z}[T]$, where $|S/G|, |T/G| < \infty$ and T has finite G-stabilizers for all $t \in T$. Let $\|\cdot\|_K$ be a G-invariant norm on K; then for all $n \in \mathbb{N}$,

$$\sup\{\|x\|_K \mid x \in K,\ \|x\|_S \leq n\} < \infty.$$

In particular, K being finitely generated induces a filling norm which is G-invariant.

4. Examples

A graph Γ is called *fine* if for every edge e and each integer $n > 0$, the number of circuits of length at most n which contain e is finite. By a circuit we mean a closed edge path that does not pass through a vertex more than once. The length of a circuit is defined as the number of edges.

Theorem 4.1 [Martínez-Pedroza 2016, Theorem 1.3]. *Let X be a cocompact G-cell complex with finite stabilizers of* 1*-cells. The following two statements are equivalent*:

(1) *X has fine* 1*-skeleton and the homology group $H_1(X, Z)$ is trivial.*

(2) $\mathrm{FV}_X(k) < \infty$ *for any integer k.*

This result allows us to exhibit examples that contrast with Theorem 1.1 as follows:

- There is a group G acting cocompactly, not properly, and by cellular automorphisms on a simply connected complex X for which $\mathrm{FV}^2_X(m)$ is finite for all $m \in \mathbb{N}$.

  In particular, the converse of Theorem 1.1 does not hold.

- There is a group G acting cocompactly by cellular automorphisms on a simply connected complex X for which $FV^2_X(m)$ is infinite for some $m \in \mathbb{N}$. In particular, the properness assumption in Theorem 1.1 cannot be removed.

The two examples are based on the notion of coned-off Cayley complex. We use the version from [Groves and Manning 2008], which we briefly recall below; for another version see [Martínez-Pedroza 2017, Section 3].

Let G be a group and let P be a subgroup. The group G is *finitely generated relative to P* if there is a finite subset $S \subset G$ such that the natural map $F(S) * P \to G$ is surjective, where $F(S)$ denotes the free group on S, and $F(S) * P$ denotes the free product of $F(S)$ and P. In this case S is called a *finite relative generating set of G with respect to P*.

Suppose that S is a finite relative generating set of G with respect to P. Without loss of generality assume that S is closed under inverses. The *coned-off Cayley graph* $\widehat{\Gamma} = \widehat{\Gamma}(G, P, S)$ is the graph with vertex set consisting of all elements of G and all left cosets of P; the edge set is the collection of pairs $(g, gs) \in G \times G$ for $g \in G$ and $s \in S$, and pairs (g, gP) for $g \in G$. Observe that the left action of G on itself extends to a left action on $\widehat{\Gamma}$. Vertices of $\widehat{\Gamma}$ of the form gP are called *cone-vertices*. Observe that the G-stabilizers of cone-vertices correspond to conjugates of P; in particular, if P is infinite, the action is not proper. Moreover, the G-stabilizers of 1-cells of $\widehat{\Gamma}$ are trivial. It is well known that the assumption that S is a relative generating set implies that $\widehat{\Gamma}$ is path-connected as a combinatorial complex; in fact, this is an equivalence, as remarked in [Hruska 2010].

Under the assumptions, the group G is *finitely presented relative to P* if there is a finite subset $R \subset F(S) * P$ such that the kernel of the map $F(S) * P \to G$ is the

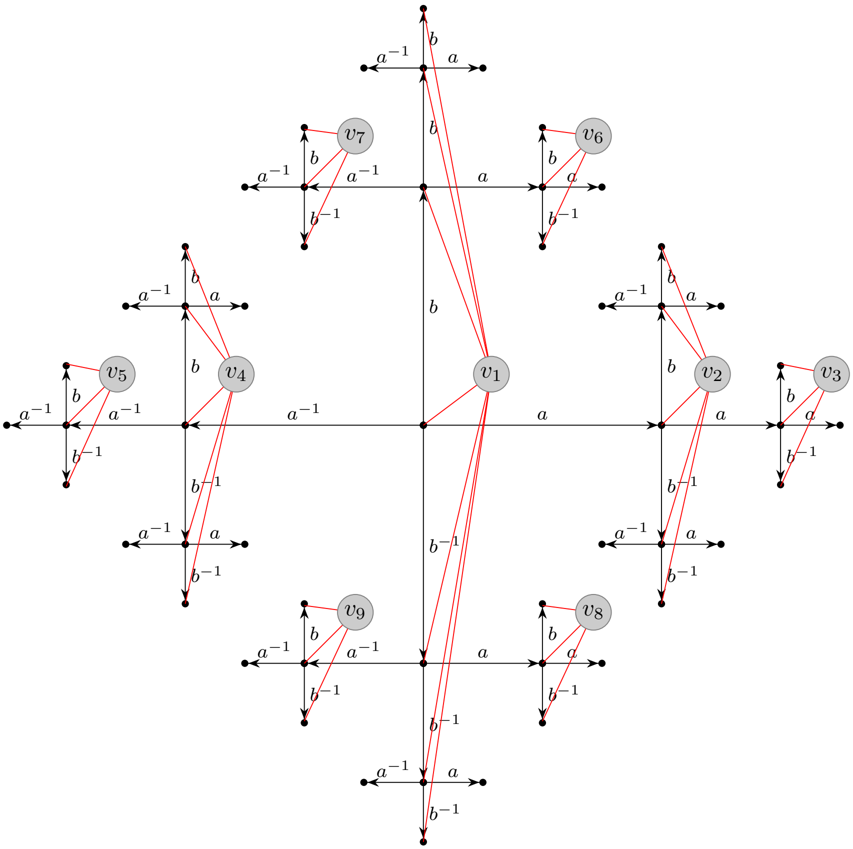


Figure 1. The coned-off Cayley graph $\widehat{\Gamma}(G, P, S)$, where G is the free group in two letters $S = \{a, b\}$ and P is the cyclic subgroup $\langle b \rangle$.

smallest normal subgroup containing R. In this case, we say that

$$\langle S, P | R \rangle \tag{1}$$

is a finite relative presentation of G with respect to P. It is an exercise to show that if G is finitely presented and P is finitely generated, then G is finitely presented relative to P. We refer the reader to [Osin 2006] for an exposition on finite relative presentations.

Assume that P is finitely generated, that (1) is a finite relative presentation of G with respect to P, and that $S \cap P$ is a generating set of P. The *coned-off Cayley complex* $\widehat{C} = \widehat{C}(G, P, S, R)$ is the 2-dimensional complex with 1-skeleton the coned-off Cayley graph $\widehat{\Gamma}(G, P, S)$ obtained by equivariantly attaching 2-cells as follows. For each word $r \in R$ correspond a loop in $\widehat{\Gamma}$. Attach a 2-cell with trivial stabilizer to each such loop, and extend in a manner equivariant under the G-action

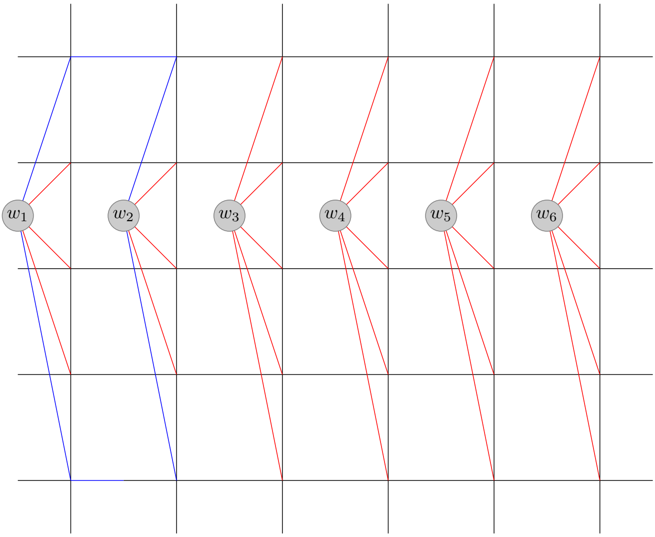


Figure 2. The coned-off Cayley graph $\widehat{\Gamma}(G, P, S)$, where G is the free abelian group in two letters $S = \{a, b\}$ and P is the cyclic subgroup $\langle b \rangle$.

on $\widehat{\Gamma}$. Similarly, for each $P \in \mathcal{P}$, for each generator in $s \in S \cap P$ and each $g \in G$, we have a corresponding loop in $\widehat{\Gamma}$ of length 3 passing through the vertices g, gs, gP. Attach a 2-cell with trivial stabilizer to each such loop, equivariantly under the G-action. The resulting G-complex $\widehat{C}$ is simply connected [Groves and Manning 2008, Lemma 2.48], the G-action is cocompact by construction, and if P is infinite, the G-action is not proper. Now we consider the 2-dimensional filling function $FV^2_{\widehat{C}}$ of $\widehat{C}$.

Example 4.2. Let G be the free group of rank 2, let $S = \{a, b\}$ be a free generating set, and let P be the cyclic subgroup generated by b. It is an observation that the coned-off Cayley graph $\widehat{\Gamma}(G, P, S)$, see Figure 1, is a fine graph and hence Theorem 4.1 implies that $\mathrm{FV}^2_{\widehat{C}}(m) < \infty$ for every $m \in \mathbb{N}$. Similar examples can be constructed by considering relatively hyperbolic groups.

Example 4.3. Let G be the free abelian group of rank 2, let $S = \{a, b\}$ be a generating set, and let P be the cyclic subgroup generated by b. The coned-off Cayley graph $\widehat{\Gamma}(G, P, S)$, see Figure 2, is not fine since there are infinitely many circuits of length 6 passing through the edge from b to P. By Theorem 4.1, we have that $\mathrm{FV}^2_{\widehat{C}}(m) = \infty$ for some $m \in \mathbb{N}$. In fact, one can verify that $\mathrm{FV}^2_{\widehat{C}}(6) = \infty$.

Remark 4.4. Theorem 1.1 does not hold for $d = 0$ in the natural setting of defining FV^1_X by taking $Z_0(X, \mathbb{Z})$ to be the kernel of the augmentation map. Consider a finitely generated infinite group G acting properly and cocompactly on a connected graph X; for example, take X to be the Cayley graph of G with respect to a finite generating set. Then X is infinite, and the formal difference $\gamma = b - a$ between two distinct vertices a and b of X is a 0-cycle for which $|\gamma|_\partial$ can be made arbitrarily large by taking a and b sufficiently far apart; roughly speaking, a 1-chain μ such that $\partial\mu = b - a$ contains a combinatorial edge path from a to b and hence $\|\mu\|_1$ is at least the length of the shortest edge path from a to b. Hence $\mathrm{FV}^1_X(2) = \infty$ in this case.

Acknowledgments

Fleming was funded by a Science Undergraduate Research Award (SURA) of Memorial University during a part of this project. Martínez-Pedroza is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC). We thank the referee for comments on the article, and Lana Martínez-Aoki for assistance during this work.

References

[Abrams et al. 2013] A. Abrams, N. Brady, P. Dani, and R. Young, "Homological and homotopical Dehn functions are different", *Proc. Natl. Acad. Sci. USA* **110**:48 (2013), 19206–19212. MR Zbl

[Alonso et al. 1999] J. M. Alonso, X. Wang, and S. J. Pride, "Higher-dimensional isoperimetric (or Dehn) functions of groups", *J. Group Theory* **2**:1 (1999), 81–112. MR Zbl

[Behrstock and Druţu 2015] J. Behrstock and C. Druţu, "Combinatorial higher dimensional isoperimetry and divergence", preprint, 2015. arXiv

[Bridson and Haefliger 1999] M. R. Bridson and A. Haefliger, *Metric spaces of non-positive curvature*, Grundlehren der Math. Wissenschaften **319**, Springer, 1999. MR Zbl

[Brown 1982] K. S. Brown, *Cohomology of groups*, Graduate Texts in Mathematics **87**, Springer, 1982. MR Zbl

[Epstein et al. 1992] D. B. A. Epstein, J. W. Cannon, D. F. Holt, S. V. F. Levy, M. S. Paterson, and W. P. Thurston, *Word processing in groups*, Jones and Bartlett, Boston, 1992. MR Zbl

[Fletcher 1998] J. L. Fletcher, *Homological group invariants*, Ph.D. thesis, University of Utah, 1998, available at https://search.proquest.com/docview/304455841.

[Gersten 1996] S. M. Gersten, "Subgroups of word hyperbolic groups in dimension 2", *J. London Math. Soc.* (2) **54**:2 (1996), 261–283. MR Zbl

[Gersten 1999] S. M. Gersten, "Homological Dehn functions and the word problem", preprint, 1999, available at http://www.math.utah.edu/~sg/Papers/df9.pdf.

[Gromov 1993] M. Gromov, "Asymptotic invariants of infinite groups", pp. 1–295 in *Geometric group theory, II* (Sussex, 1991), edited by G. A. Niblo and M. A. Roller, London Math. Soc. Lecture Note Ser. **182**, Cambridge Univ. Press, 1993. MR Zbl

[Groves and Manning 2008] D. Groves and J. F. Manning, "Dehn filling in relatively hyperbolic groups", *Israel J. Math.* **168** (2008), 317–429. MR Zbl

[Hanlon and Martínez-Pedroza 2016] R. G. Hanlon and E. Martínez-Pedroza, "A subgroup theorem for homological filling functions", *Groups Geom. Dyn.* **10**:3 (2016), 867–883. MR Zbl

[Hruska 2010] G. C. Hruska, "Relative hyperbolicity and relative quasiconvexity for countable groups", *Algebr. Geom. Topol.* **10**:3 (2010), 1807–1856. MR Zbl

[Martínez-Pedroza 2016] E. Martínez-Pedroza, "A note on fine graphs and homological isoperimetric inequalities", *Canad. Math. Bull.* **59**:1 (2016), 170–181. MR Zbl

[Martínez-Pedroza 2017] E. Martínez-Pedroza, "Subgroups of relatively hyperbolic groups of Bredon cohomological dimension 2", *J. Group Theory* (online publication July 2017).

[Mineyev 2000] I. Mineyev, "Higher dimensional isoperimetric functions in hyperbolic groups", *Math. Z.* **233**:2 (2000), 327–345. MR Zbl

[Osin 2006] D. V. Osin, *Relatively hyperbolic groups: intrinsic geometry, algebraic properties, and algorithmic problems*, Mem. Amer. Math. Soc. **843**, 2006. MR Zbl

[Young 2011] R. Young, "Homological and homotopical higher-order filling functions", *Groups Geom. Dyn.* **5**:3 (2011), 683–690. MR Zbl

[Young 2016] R. Young, "High-dimensional fillings in Heisenberg groups", *J. Geom. Anal.* **26**:2 (2016), 1596–1616. MR Zbl

Received: 2016-09-15 Accepted: 2017-07-22

jwf572@mun.ca *Memorial University of Newfoundland, St. John's, NL, Canada*

emartinezped@mun.ca *Memorial University of Newfoundland, St. John's, NL, Canada*

INVOLVE 11:4 (2018)
dx.doi.org/10.2140/involve.2018.11.585

Explicit representations of 3-dimensional Sklyanin algebras associated to a point of order 2

Daniel J. Reich and Chelsea Walton

(Communicated by Michael E. Zieve)

The representation theory of a 3-dimensional Sklyanin algebra S depends on its (noncommutative projective algebro-) geometric data: an elliptic curve E in $\mathbb{P}^2$, and an automorphism σ of E given by translation by a point. Indeed, by a result of Artin, Tate, and van den Bergh, we have that S is module-finite over its center if and only if σ has finite order. In this case, all irreducible representations of S are finite-dimensional and of at most dimension $|\sigma|$.

In this work, we provide an algorithm in Maple to directly compute all irreducible representations of S associated to σ of order 2, up to equivalence. Using this algorithm, we compute and list these representations. To illustrate how the algorithm developed in this paper can be applied to other algebras, we use it to recover well-known results about irreducible representations of the skew polynomial ring $\mathbb{C}_{-1}[x, y]$.

1. Introduction

We work over the ground field $\mathbb{C}$. The motivation of this work is to study, up to equivalence, irreducible finite-dimensional representations (irreps) of *Sklyanin algebras* S of global dimension 3 (Definition 1.2). Past work on this problem includes results on bounds on the dimension of irreps of S [Walton 2012], and on a geometric parametrization of (trace-preserving) irreps of S [De Laet and Le Bruyn 2015]. The focus of this paper is to determine, for a class of Sklyanin algebras, all *explicit* irreps up to equivalence. Namely, we compute *irreducible matrix solutions* to the defining equations of S, up to an action of a general linear group. A geometric parametrization of the set of irreps of S is also presented, as this is the typical approach to understanding aspects of Sklyanin algebras.

MSC2010: 16S38, 16G99, 16Z05.
Keywords: Azumaya locus, irreducible representation, Maple algorithm, 3-dimensional Sklyanin algebra.

Remark 1.1. We directly compute the irreps via a Maple algorithm. A more conceptual technique, using noncommutative projective algebraic geometry (and Clifford theory for these particular Sklyanin algebras), can be used to solve this problem. We nevertheless hold to the computational approach because it can be adapted (much more easily in some cases) to other algebras; for further discussion of the complexity of this approach, see Remarks 1.10 and 1.11.

To begin, let us define the algebra under investigation.

Definition 1.2 [Artin et al. 1990]. The *3-dimensional Sklyanin algebra* $S := S(a, b, c)$ over $\mathbb{C}$ is generated by three noncommuting variables x, y, z subject to the relations

$$ayz + bzy + cx^2 \;=\; azx + bxz + cy^2 \;=\; axy + byx + cz^2 \;=\; 0. \tag{1.3}$$

Here, $[a : b : c] \in \mathbb{P}^2_{\mathbb{C}}$, with $abc \neq 0$ and $(3abc)^3 \neq (a^3 + b^3 + c^3)^3$.

This algebra is rather resistant to noncommutative Gröbner basis methods; that is, it is difficult to write down a $\mathbb{C}$-vector space basis of S (consisting of monomials in x, y, z). See, for instance, [Bellamy et al. 2016, Exercise 1.7]. (The reader may also be interested in [Iyudu and Shkarin 2017].) In fact, it is common practice to consider the geometric data of S in the context of noncommutative projective algebraic geometry [Artin et al. 1990; Bellamy et al. 2016; Stafford and Van den Bergh 2001] to analyze its ring-theoretic behavior. By [Artin et al. 1990, Equations 1.6 and 1.7], the geometric data of $S(a, b, c)$ consists of an elliptic curve $E := E_{a,b,c} \subset \mathbb{P}^2_{\mathbb{C}}$ defined the equation

$$E_{a,b,c} :\; (a^3 + b^3 + c^3)(uvw) - (abc)(u^3 + v^3 + w^3) = 0, \tag{1.4}$$

and an automorphism of this elliptic curve $\sigma := \sigma_{a,b,c}$ given by

$$\sigma_{a,b,c}([u : v : w]) = [acv^2 - b^2uw \;:\; bcu^2 - a^2vw \;:\; abw^2 - c^2uv]. \tag{1.5}$$

Here, the automorphism is given by translation of the point $[a : b : c] \in E_{a,b,c}$, where $[1 : -1 : 0]$ is the origin of $E_{a,b,c}$. The *order* of σ, denoted by $|\sigma|$, is the smallest $n \in \mathbb{N}$ such that $\sigma^n = \mathrm{id}_E$. If no such n exists, then $|\sigma| = \infty$. Consider the following terminology.

Definition 1.6. We say that a Sklyanin algebra $S(a, b, c)$ is *associated to a point* ($[a : b : c] \in E_{a,b,c}$) *of order* n if the automorphism $\sigma_{a,b,c}$ has order n.

The role of this geometric data for our work will be explained towards the end of this section.

Now let us recall some basic representation theory terminology. Take n to be a positive integer. An *n-dimensional representation* of $S := S(a, b, c)$ is an algebra homomorphism $\psi : S \to \mathrm{End}(V)$, where V is a $\mathbb{C}$-vector space of dimension n. Since

$\mathrm{End}(V)$ is isomorphic to $\mathrm{Mat}_n(\mathbb{C})$, there is a one-to-one correspondence between the n-dimensional representations of $S(a, b, c)$ and the $n \times n$ *matrix solutions* (X, Y, Z) to the system of equations (1.3). Here, $X = \psi(x)$, $Y = \psi(y)$, and $Z = \psi(z)$.

Next, we discuss irreducibility. Given a representation $\psi : S \to \mathrm{End}(V)$, a subspace W of V is called S-*stable* if $\psi(s)(w) \in W$ for all $s \in S$, $w \in W$. Such a subspace W yields a *subrepresentation* of S, given as $\psi' : S \to \mathrm{End}(W)$. We say that ψ is *irreducible* if the only S-stable subspaces of V are $\{0\}$ and itself, that is, if there are no proper subrepresentations ψ' of ψ. Similarly, there is a notion of irreducibility for a matrix solution (X, Y, Z) to (1.3); see Lemma 2.1.

Now we recall when two representations/matrix solutions of S are equivalent. We say that n-dimensional representations $\psi, \phi : S \to \mathrm{End}(V)$ are *equivalent* if there exists a matrix $Q \in \mathrm{GL}_n(\mathbb{C})$ such that $\psi(s) = Q\phi(s)Q^{-1}$ for all $s \in S$. Likewise, two matrix solutions (X_0, Y_0, Z_0) and (X_1, Y_1, Z_1) to (1.3) are *equivalent* if there exists $Q \in \mathrm{GL}_n(\mathbb{C})$ such that $Q^{-1}X_0Q = X_1$, $Q^{-1}Y_0Q = Y_1$, and $Q^{-1}Z_0Q = Z_1$. Note that two equivalent representations/matrix solutions are either both irreducible or both reducible.

As the reader can imagine, studying explicit finite-dimensional representations of the algebras $S(a, b, c)$ is difficult computationally. Now by [Walton 2012, Theorem 1.3], we only have nontrivial finite-dimensional representations of S when the automorphism σ of (1.5) has finite order. So, we refine our goal: we study the irreps of $S(a, b, c)$ associated to a point $[a : b : c] \in E_{a,b,c}$ of order 2. Note that the order-1 case is precisely the case when S is commutative (Lemma 2.4).

Lemma 1.7 (Lemma 2.5). *A Sklyanin algebra $S(a, b, c)$ is associated to a point $[a : b : c] \in E_{a,b,c}$ of order 2 if and only if $a = b$.*

In this case, we assume that $a = b = 1$ by rescaling. Therefore, our goal is to study the representation theory of the 3-dimensional Sklyanin algebra $S(1, 1, c)$, where by Definition 1.2, $c \neq 0$, $c^3 \neq 1, -8$. By Lemma 2.6, all 1-dimensional irreps of $S(1, 1, c)$ are trivial, and all irreps of $S(1, 1, c)$ are finite-dimensional, of at most dimension 2. Thus, we only need to compute the irreps of dimension 2; we achieve this as follows.

Theorem 1.8. *The nontrivial explicit irreps (or matrix solutions) of the 3-dimensional Sklyanin algebra $S(1, 1, c)$ are of dimension 2. They are classified up to equivalence; the representatives of equivalence classes of irreps of $S(1, 1, c)$ are provided in (5.1)–(5.2) and (6.1)–(6.5) in Sections 5 and 6, respectively.*

In Section 2, we provide background material and some preliminary results. In Section 3, we give an outline (Steps 0–2, 3a, 3b) of our algorithm to prove Theorem 1.8. The algorithm then begins in Section 4, where we determine all of the 2-dimensional representations of $S(1, 1, c)$, and exclude “families” of reducible representations; this is Steps 0–2 of the algorithm. In Sections 5 and 6, we determine representatives of equivalence classes of 2-dimensional irreps of $S(1, 1, c)$; this is Steps 3a and 3b of the algorithm.

The study of the irreps of $S(1, 1, c)$ ends in Section 7, where for completion, we discuss a geometric parametrization of equivalence classes of irreps of $S(1, 1, c)$; e.g., we illustrate the *Azumaya locus* of $S(1, 1, c)$ over the center of $S(1, 1, c)$. Namely we have the result below.

Theorem 1.9 (Theorem 7.1). *The set of equivalence classes of irreps of $S(1, 1, c)$ is in bijective correspondence with the points of the 3-dimensional affine variety*:

$$X_c := \mathbb{V}\big(g^2 - c^2(u_1^3 + u_2^3 + u_3^3) - (c^3 - 4)u_1u_2u_3\big) \subseteq \mathbb{C}^4_{\{u_1,u_2,u_3,g\}}.$$

In particular, $X_c \setminus \{\underline{0}\}$ is the Azumaya locus of S over its center (i.e., points of $X_c \setminus \{\underline{0}\}$ correspond to 2-dimensional irreps of S), and the origin of X_c corresponds to the trivial representation of S.

Remark 1.10. We would like to point out that one can adjust our algorithm to prove Theorem 1.8 to examine equivalence classes of irreps of other algebras with generators and relations, especially those that are module-finite over their center. Although, the run-time and complexity of the output of the algorithm is in direct correlation with the number of generators and relations of the algebra, along with the algebra's *polynomial identity degree* (PI degree), if applicable.

We illustrate the remark above in Section 8, where we tailor our algorithm to examine irreps of the skew polynomial ring

$$\mathbb{C}_{-1}[x, y] := \mathbb{C}\langle x, y\rangle/(xy + yx).$$

Like $S(1, 1, c)$, it is well known that all irreps of $\mathbb{C}_{-1}[x, y]$ are finite-dimensional, of dimension at most 2 (Lemma 8.1(c)). See Proposition 8.3 and Corollary 8.5 for the results on the representation theory of $\mathbb{C}_{-1}[x, y]$.

Remark 1.11. Part of the novelty of this work is that we obtain noncommutative algebraic/representation-theoretic results with Maple, which is a computer algebra system that is used typically for commutative computations. We hope that in the future the task of determining equivalence classes of irreps of noncommutative algebras (presented by generators and relations) can be achieved easily using a computer algebra system that handles noncommutative Gröbner bases, such as GAP [Cohen and Knopper 2016].

Remark 1.12. Unless stated otherwise, computational results in this work are performed with the computer algebra system Maple (version 16). All code (including comments) is available on the authors' professional websites, and in the preprint version of this work available on the ArXiv: http://arxiv.org/abs/1512.09167.

2. Preliminaries

We begin with a result on the irreducibility of a representation/matrix solution of a Sklyanin algebra $S = S(a, b, c)$. This result is well known, and we will use it often without mention.

Lemma 2.1. *Let $\psi : S \to \mathrm{End}(V)$ be an n-dimensional representation of S, with corresponding matrix solution (X, Y, Z) to the system of equations (1.3). Then, the following are equivalent:*

(a) *ψ is irreducible.*

(b) *The corresponding S-module V (where S acts on V via ψ) is simple.*

(c) *ψ is surjective.*

(d) *$\psi(S)$ generates $\mathrm{End}(V) \cong \mathrm{Mat}_n(\mathbb{C})$ as a $\mathbb{C}$-algebra.*

(e) *Every matrix in $\mathrm{Mat}_n(\mathbb{C})$ can be expressed as a noncommutative polynomial in (X, Y, Z) over $\mathbb{C}$.* □

If any of the above conditions hold, we say that the matrix solution (X, Y, Z) is *irreducible*.

On the other hand, we can determine when a matrix solution of S is reducible by using Lemma 2.1.

Corollary 2.2. *An $n \times n$ matrix solution (X, Y, Z) to (1.3) (corresponding to a representation ψ of S) is reducible if and only if there exists a subspace W of V of dimension $m < n$ with $X \cdot w, Y \cdot w, Z \cdot w \in W$ for all $w \in W$. Here, we embed W into V so that $\cdot$ is given by matrix multiplication.* □

If S is a Sklyanin algebra associated to a point of infinite order, then by [Walton 2012, Theorem 1.3(i)], we have that all finite-dimensional irreps of S are trivial. On the other hand, Sklyanin algebras associated to points of finite order have an interesting representation theory, due to the following result.

Proposition 2.3. *Let $S(a, b, c)$ be a Sklyanin algebra associated to a point of finite order. Then, all irreducible representations of $S(a, b, c)$ are finite-dimensional, of at most dimension $|\sigma_{a,b,c}|$.*

Proof. In this case, we have that the Sklyanin algebra $S(a, b, c)$ is module-finite over its center by [Artin et al. 1991, Theorem 7.1]. Further, $S(a, b, c)$ has PI degree $|\sigma_{a,b,c}|$ by [Walton 2012, Proposition 1.6]. Hence, the irreducible representations of $S(a, b, c)$ are all finite-dimensional by [McConnell and Robson 2001, Theorem 13.10.3(a)], of dimension at most $|\sigma_{a,b,c}|$ by [Brown and Goodearl 1997, Proposition 3.1]. □

Now we analyze parameters $(a, b, c) \in \mathbb{C}^3$ so that the automorphism $\sigma_{a,b,c}$ from (1.5) has finite order. Recall that two projective points $[m_1 : m_2 : m_3]$, $[n_1 : n_2 : n_3] \in \mathbb{P}^2_{\mathbb{C}}$

are equal if and only if $m_1n_2 - m_2n_1 = m_1n_3 - m_3n_1 = m_2n_3 - m_3n_2 = 0$ if and only if $n_i = \lambda m_i$ for all $i = 1, 2, 3$, for some nonzero $\lambda \in \mathbb{C}$. Omitting the conditions on parameters a, b, c for now, it is worth noting the following the result.

Lemma 2.4. *The automorphism $\sigma_{a,b,c}$ from (1.5) has order 1 if and only if $a = 1$, $b = -1$, $c = 0$. In this case, $S(1, -1, 0)$ is the commutative polynomial ring $\mathbb{C}[x, y, z]$.*

Proof. If σ has order 1, then we obtain $[acv^2-b^2uw : bcu^2-a^2vw : abw^2-c^2uv] = [u : v : w]$. Therefore, $bcu^2w - (a^2 + ab)vw^2 + c^2uv^2 = 0$, which (by taking the coefficient of uv^2) implies $c = 0$. Without loss of generality, take $a = 1$. Now, $[-b^2uw : -vw : bw^2] = [u : v : w]$, and we must have that $b = -1$ since $-vw^2 = bvw^2$. Therefore, the forward direction holds. For the converse, note that $\sigma_{1,-1,0}([u : v : w]) = [-uw : -vw : -w^2] = [u : v : w]$, so $\sigma_{1,-1,0}$ has order 1. The last statement is clear. □

Consider the following preliminary results about Sklyanin algebras associated to a point of order 2.

Lemma 2.5. *Take $S = S(a, b, c)$ to be a 3-dimensional Sklyanin algebra associated to the automorphism $\sigma_{a,b,c}$ of (1.5). Then, $|\sigma_{a,b,c}| = 2$ if and only if $a = b$.*

Proof. Without loss of generality, take $a = 1$. The code for this result (see Remark 1.12) implies that $b = 1$ and there are no conditions on c (other than those in Definition 1.2).

The converse is clear by the computation above, but we can verify this directly. If $a = b = 1$, then $\sigma_{1,1,c}([u : v : w]) = [cv^2 - uw : cu^2 - vw : w^2 - c^2uv]$. So,

$$\begin{aligned}\sigma^2_{1,1,c}([u : v : w]) &= \big[u(c^3u^3+c^3v^3+w^3-3c^2uvw) : v(c^3u^3+c^3v^3+w^3-3c^2uvw)\\ &\qquad : w(c^3u^3+c^3v^3+w^3-3c^2uvw)\big]\\ &= [u : v : w],\end{aligned}$$

as desired. □

Hence, to work with Sklyanin algebras $S(a, b, c)$ associated to a point of order 2, we take $a = b = 1$.

Lemma 2.6. *We have the following statements for the Sklyanin algebra $S(1, 1, c)$.*

(a) *The only 1-dimensional representation of $S(1, 1, c)$ is the trivial representation.*

(b) *All irreducible representations of $S(1, 1, c)$ are finite-dimensional, of at most dimension equal to 2.*

Proof. (a) One can compute this directly, or by using a short routine; see Remark 1.12.

(b) This follows from Proposition 2.3 and Lemma 2.5. □

3. Methodology and terminology

In this section, we provide an outline of the algorithm used to prove Theorem 1.8; see Sections 4–6 for the full details. The goal is to obtain *irreducible representative families* of $S(1, 1, c)$ as defined below.

Definition 3.1. We say that a set of matrix solutions of the defining equations of $S(1, 1, c)$ (or of equations (1.3) with $a = b = 1$) is a *representative family of matrix solutions*, if no two members within the set are equivalent. Further, we call this set an *irreducible representative family* if all of its members are irreducible matrix solutions of $S(1, 1, c)$.

Note that we aim to have the parameter c of $S(1, 1, c)$ free. So due to Maple's default alpha ordering, *we refer to c as* `zc` *in the code below*.

First, we make the following simplification.

Step 0: assume the matrix X is in Jordan form. Due to Lemma 2.6 we know that all nontrivial irreps of $S(1, 1, c)$ are of dimension 2. Hence, we only study 2×2 matrix solutions (X, Y, Z) of (1.3) with $(a, b, c) = (1, 1, c)$. Initially, the entries of X, Y, Z are x_ℓ, y_ℓ, z_ℓ for $\ell = 1, 2, 3, 4$. We further simplify the problem by assuming that X is in Jordan form. This simplification is made because we wish to classify the irreps up to equivalence, and equivalence is determined by simultaneous conjugation by an invertible matrix. So, we take X to be either a single 2×2 Jordan block or diagonal so that we have 3 or 2 less unknowns, respectively. We consider these cases separately.

Step 1: find all families of matrix solutions. Now, we solve (1.3) with $(a, b, c) = (1, 1, c)$ for 2×2 matrices (X, Y, Z). The output consists of 2-dimensional (*matrix solution*) *families* of $S(1, 1, c)$. The solutions are grouped according to the default behavior of Maple. We refer to these groups as `Families`.

Step 2: eliminate reducible matrix solutions. We run this step now to cut down on the run-time of the algorithm and the complexity of its output. Given a family of matrix solutions, we use Corollary 2.2 to determine if all members of this family are reducible. Namely, we let `w =«p,q»` be a basis of a 1-dimensional subspace W of $\mathbb{C}^2$. Note that if $p = p_1 + p_2 i$ and $q = q_1 + q_2 i$ for $i := \sqrt{-1}$ and $p_1, p_2, q_1, q_2 \in \mathbb{R}$, then $(p, q) \neq (0, 0)$ precisely when $p\bar{p} + q\bar{q} \neq 0$. We examine when W is stable under the action of $S(1, 1, c)$; namely, we need each of Xw, Yw, Zw to be a scalar multiple of w. So, we solve for p, q subject to the conditions

- W is not the zero subspace ...... `p*conjugate(p)+q*conjugate(q)<>0`,
- $XW \subset W$ .... `p*Xw[2][1]-q*Xw[1][1]=0`,
- $YW \subset W$ .... `p*Yw[2][1]-q*Yw[1][1]=0`,
- $ZW \subset W$ .... `p*Zw[2][1]-q*Zw[1][1]=0`,
- conditions on c.

If there is a solution, then this implies that all members of the specified family are reducible. We remove such families from further computations by forming a list `NonRedFams` consisting of families for which there is no p, q satisfying the conditions above.

Steps 3a and 3b are independent of each other, and either can be run after Step 2.

Step 3a: account for equivalence between families. For the remaining families of matrix solutions, we determine conditions when members of one family `NonRedFams[i]` is equivalent to members of another family `NonRedFams[j]`. These conditions are collected in the list `BetweenFams`.

We do so as follows. First, we force variables of `NonRedFams[i]` to be in terms of u_ℓ, v_ℓ, w_ℓ instead of x_ℓ, y_ℓ, z_ℓ for $\ell = 1, 2, 3, 4$; this is executed with

```
eval(NonRedFams[...],ChangeVars).
```

Next, we conjugate the relabeled matrices simultaneously by a 2×2 matrix `Q` to form `Xconj`, `Yconj`, `Zconj`. Then, we solve for variables $u_\ell, v_\ell, w_\ell, x_\ell, y_\ell, z_\ell$ subject to the conditions

- `Xconj` is equal to the X-matrix `Xj` of `NonRedFams[j]` .... `Equiv1=0`,
- `Yconj` is equal to the Y-matrix `Yj` of `NonRedFams[j]` .... `Equiv2=0`,
- `Zconj` is equal to the Z-matrix `Zj` of `NonRedFams[j]` .... `Equiv3=0`,
- conditions on c and invertibility of `Q`.

The output is `[i,j,{`conditions on $u_\ell, v_\ell, w_\ell, x_\ell, y_\ell, z_\ell$`}]`, which we interpret as follows.

Interpretation: We can eliminate `NonRedFams[i]` from our consideration if all of its members are equivalent to members of `NonRedFams[j]` for some $j \neq i$. This occurs if we get an output

[`i`, `j`, ...{each of u_ℓ, v_ℓ, w_ℓ is free }...] for $i < j$, or

[`j`, `i`, ...{each of x_ℓ, y_ℓ, z_ℓ is free }...] for $j < i$.

We obtain that `NonRedFams[i]` forms a representative family if we get output

`[i,i,...{`restrictions on $u_\ell, v_\ell, w_\ell, x_\ell, y_\ell, z_\ell$ `}...]`

under one of the following conditions:

- (i) each of x_ℓ, y_ℓ, z_ℓ is free and (ii) each of u_ℓ, v_ℓ, w_ℓ is free, or depends only on x_ℓ, y_ℓ, z_ℓ; or
- (i) each of u_ℓ, v_ℓ, w_ℓ is free and (ii) each of x_ℓ, y_ℓ, z_ℓ is free, or depends only on u_ℓ, v_ℓ, w_ℓ.

In either case above, we set the free variables in (ii) equal to 1 to obtain representative families. Otherwise, a careful examination is needed.

Conditions $u_\ell, v_\ell, w_\ell, x_\ell, y_\ell, z_\ell$ may depend on entries of the matrix `Q`. In this case, we can conclude that such variables are free as long as this does not violate invertibility of `Q`.

Step 3b: check for full irreducibility conditions. Here, we run the same code as in Step 2 except that we solve for p, q along with all variables x_ℓ, y_ℓ, z_ℓ. The conditions are collected in a list called `IrConditions`. If the output for `NonRedFams[i]` is `[i]` (or empty), then all members of `NonRedFams[i]` are irreducible.

4. Families of nonreducible representations of $S(1, 1, c)$

Here, we execute Steps 0–2 of the algorithm discussed in the previous section. Namely, we find all 2-dimensional representations of $S(1, 1, c)$ by determining 2×2 matrix solutions (X, Y, Z) to (1.3) with $a = b = 1$. Here, X is in Jordan form, either one Jordan block or two Jordan blocks (diagonal). Moreover, we eliminate the families of solutions for which all of its members are reducible. See Remark 1.12 and we obtain the results below.

We start with the output of Steps 0–2 for `NonRedFams` when X is one Jordan block. For all matrix solutions, we have

$$X = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}. \tag{4.1}$$

The rows below show the matrices Y, Z for the five matrix solutions:

$$\begin{gathered}
\begin{pmatrix} -y_4 & \big(y_4^2+(y_4^4-8y_4z_4^3)^{1/2}\big)/(2cz_4^2) \\ -cz_4^2 & y_4 \end{pmatrix}, \quad \begin{pmatrix} -z_4 & 0 \\ \frac{1}{2}c\big(y_4^2+(y_4^4-8y_4z_4^3)^{1/2}\big)-cy_4^2 & z_4 \end{pmatrix}; \\
\begin{pmatrix} -y_4 & -\big(-y_4^2+(y_4^4-8y_4z_4^3)^{1/2}\big)/(2cz_4^2) \\ -cz_4^2 & y_4 \end{pmatrix}, \quad \begin{pmatrix} -z_4 & 0 \\ -\frac{1}{2}c\big(-y_4^2+(y_4^4-8y_4z_4^3)^{1/2}\big)-cy_4^2 & z_4 \end{pmatrix}; \\
\begin{pmatrix} \alpha/(cz_3) & -\big(-cz_2^2z_3-cz_2z_4^2+2\alpha z_4/(cz_3)\big)/z_3 \\ -cz_2z_3-cz_4^2 & -\alpha/(cz_3) \end{pmatrix}, \quad \begin{pmatrix} -z_4 & z_2 \\ z_3 & z_4 \end{pmatrix}; \\
\begin{pmatrix} -\beta/(cz_3) & -\big(-cz_2^2z_3-cz_2z_4^2+2\beta z_4/(cz_3)\big)/z_3 \\ -cz_2z_3-cz_4^2 & \beta/(cz_3) \end{pmatrix}, \quad \begin{pmatrix} -z_4 & z_2 \\ z_3 & z_4 \end{pmatrix}; \\
\begin{pmatrix} -y_4 & y_4^2/(cz_4^2) \\ -cz_4^2 & y_4 \end{pmatrix}, \quad \begin{pmatrix} -z_4 & 2y_4/(cz_4) \\ 0 & z_4 \end{pmatrix};
\end{gathered} \tag{4.2}$$

where,

$$\begin{aligned}
\alpha &= c^2z_2z_3z_4 + c^2z_4^3 + \big(3c^4z_2^2z_3^2z_4^2 + 3c^4z_2z_3z_4^4 + c^4z_4^6 - cz_3^3 + c^4z_3^3z_2^3\big)^{1/2}, \\
\beta &= -c^2z_2z_3z_4 - c^2z_4^3 + \big(3c^4z_2^2z_3^2z_4^2 + 3c^4z_2z_3z_4^4 + c^4z_4^6 - cz_3^3 + c^4z_3^3z_2^3\big)^{1/2}.
\end{aligned} \tag{4.3}$$

When X is two Jordan blocks, `NonRedFams` gives six matrix solutions:

$$X=\begin{pmatrix} cz_4^2/(2y_4) & 0 \\ 0 & -cz_4^2/(2y_4)\end{pmatrix},\quad Y=\begin{pmatrix} -y_4 & -(y_4^3-z_4^3)/(y_4y_3) \\ y_3 & y_4\end{pmatrix},$$
$$Z=\begin{pmatrix} -z_4 & -z_4(8y_4^3+c^3z_4^3)/(4y_4^2y_3) \\ 0 & z_4\end{pmatrix};$$

$$X=\begin{pmatrix} -x_4 & 0 \\ 0 & x_4\end{pmatrix},\quad Y=\begin{pmatrix} 0 & 0 \\ y_3 & 0\end{pmatrix},\quad Z=\begin{pmatrix} 0 & -cx_4^2/y_3 \\ 0 & 0\end{pmatrix};$$

$$X=\begin{pmatrix} cy_4^2/(2z_4) & 0 \\ 0 & -cy_4^2/(2z_4)\end{pmatrix},\quad Y=\begin{pmatrix} -y_4 & -y_4(8z_4^3+c^3y_4^3)/(4z_3z_4^2) \\ 0 & y_4\end{pmatrix},$$
$$Z=\begin{pmatrix} -z_4 & (y_4^3-z_4^3)/(z_4z_3) \\ z_3 & z_4\end{pmatrix};$$

$$X=\begin{pmatrix} -x_4 & 0 \\ 0 & x_4\end{pmatrix},\quad Y=\begin{pmatrix} 0 & -cx_4^2/z_3 \\ 0 & 0\end{pmatrix},\quad Z=\begin{pmatrix} 0 & 0 \\ z_3 & 0\end{pmatrix}; \tag{4.4}$$

$$X=\begin{pmatrix} \gamma/(c^2y_3z_3) & 0 \\ 0 & -\gamma/(c^2y_3z_3)\end{pmatrix},\quad Y=\begin{pmatrix} -y_4 & -\bigl(2z_4\gamma/(c^2y_3z_3)+cy_4^2\bigr)/(cy_3) \\ y_3 & y_4\end{pmatrix},$$
$$Z=\begin{pmatrix} -z_4 & -\bigl(-2\gamma y_4/(c^2y_3z_3)+cz_4^2\bigr)/(cz_3) \\ z_3 & z_4\end{pmatrix};$$

$$X=\begin{pmatrix} -\delta/(c^2y_3z_3) & 0 \\ 0 & \delta/(c^2y_3z_3)\end{pmatrix},\quad Y=\begin{pmatrix} -y_4 & -\bigl(2z_4\delta/(c^2y_3z_3)+cy_4^2\bigr)/(cy_3) \\ y_3 & y_4\end{pmatrix},$$
$$Z=\begin{pmatrix} -z_4 & -\bigl(2\delta y_4/(c^2y_3z_3)+cz_4^2\bigr)/(cz_3) \\ z_3 & z_4\end{pmatrix};$$

where

$$\begin{aligned}\gamma&=-z_3^2z_4-y_3^2y_4+\bigl(z_3^4z_4^2+2z_3^2z_4y_3^2y_4+y_3^4y_4^2+c^3y_3z_3^3y_4^2+c^3y_3^3z_3z_4^2-2c^3y_3^2z_3^2y_4z_4\bigr)^{1/2},\\ \delta&=z_3^2z_4+y_3^2y_4+\bigl(z_3^4z_4^2+2z_3^2z_4y_3^2y_4+y_3^4y_4^2+c^3y_3z_3^3y_4^2+c^3y_3^3z_3z_4^2-2c^3y_3^2z_3^2y_4z_4\bigr)^{1/2}.\end{aligned} \tag{4.5}$$

5. Equivalence and irreducibility: one-Jordan-block case

We wish to classify the matrix solutions from Steps 0–2 (in the previous section) up to equivalence and extract the irreducible equivalence classes. So, we would like to know under what conditions is a matrix solution equivalent to a member of the same/different solution family. We then specify conditions for which the representative of an equivalence class of matrix solutions is irreducible. This achieved with Steps 3a and 3b, respectively, as described in Section 3. In this section, we continue the algorithm of Section 4 in the case when X is one Jordan block; see Remark 1.12.

The output of Steps 0–3a can be viewed by entering the following:

```
for i from 1 to nops(BetweenFams) do     print(BetweenFams[i]):     end do:
```

For interpretation, consider the snippets of output

```
[1, 2, {q1 = q1, q2 = q2, q3 = 0, q4 = q1, v4 = v4, w4 = w4,

                 3     3        1/2
            -4 w4   + v4   - v4 %1                          4 q1 w4
       y4 = -----------------------, z4 = -w4,  zc = - -----------------}],
                   2     1/2                                 2     1/2
                 v4   - %1                             q2 (v4   - %1   )

          4       3
  %1 := v4   - 8 w4   v4
                                                                         2
                                                                 zc q2 w4
[1, 5, {q1 = q1, q2 = q2, q3 = 0, q4 = q1, v4 = 0, w4 = w4, y4 = ---------,
                                                                    q1
        z4 = w4, zc = zc}]
```

In the first snippet, one sees that with a choice of q_1 and q_2, the parameter c can be considered free without violating the invertibility of Q. We can also conclude that any member of NonRedFams[1] is equivalent to a member of NonRedFams[2], except when $v_4^2-(v_4^4-8w_4^3v_4)^{1/2}=0$, or equivalently when v_4 or $w_4=0$. From the second snippet of output, we see that any member of NonRedFams[1] is equivalent to a member of NonRedFams[5] when $v_4 = 0$. Moreover by (4.1)–(4.3), we have that in NonRedFams[1] w_4 (identified with z_4) cannot be 0. So, we exclude NonRedFams[1] from further computation.

Now consider another two snippets of output:

```
[2, 4, {q1 = q1, q2 = q2, q3 = 0, q4 = q1, v4 = v4, w4 = w4, z2 = z2,

                       2
           (2 RootOf(_Z   + 1 + _Z) w4 q2 - q1 z2) q1
      z3 = -----------------------------------------,
                                    2
                                  q2

                      2
             RootOf(_Z   + 1 + _Z) w4 q2 - q1 z2
      z4 = - -----------------------------------,
                            q2

                           2
             2 (2 RootOf(_Z   + 1 + _Z) w4 q2 - q1 z2) q1
      zc = - -------------------------------------------}]
                      2      4        3    1/2    2
                   (v4   + (v4   - 8 w4   v4)    ) q2
                                                                         2
                                                                 zc q2 w4
[2, 5, {q1 = q1, q2 = q2, q3 = 0, q4 = q1, v4 = 0, w4 = w4, y4 = ---------,
                                                                    q1
  z4 = w4,  zc = zc}]
```

Through a choice of q_1 and q_2, we consider c to be free in [2,4,...]. We conclude that any member of NonRedFams[2] is equivalent to a member of

NonRedFams[4] for all values of v_4 and w_4 except when $v_4^2+(v_4^4-8w_4^3v_4)^{1/2}=0$, or equivalently when v_4 or $w_4=0$. From the second snippet of output, we see that if $v_4=0$, any member of NonRedFams[2] is equivalent to a member of NonRedFams[5]. From (4.1)–(4.3), we see that w_4 (identified with z_4) in NonRedFams[2] cannot be 0. So, we exclude NonRedFams[2] from further computation.

Now take into account the following snippets of output:

```
[3, 4, {q1 = q1, q2 = q2, q3 = 0, q4 = q1, w2 = w2, w3 = w3, w4 = w4,

                         2        2
       2 q1 w4 q2 + w2 q1  - w3 q2                 -w3 q2 + q1 w4
  z2 = ----------------------------, z3 = w3, z4 = --------------, zc = zc}]
                      2                                   q1
                    q1

[4, 5]
```

This implies NonRedFams[3] is equivalent to NonRedFams[4]. So, we exclude NonRedFams[3] from further computation. Further, no member of NonRedFams[4] is equivalent to a member of NonRedFams[5].

Finally, we determine when the remaining families are representative families. Consider

```
[4, 4, {q1 = q1, q2 = q2, q3 = 0, q4 = q1, w2 = w2, w3 = w3, w4 = w4,

                                  2        2
       2 q1 w4 q2 + w2 q1  - w3 q2                 -w3 q2 + q1 w4
  z2 = ----------------------------, z3 = w3, z4 = --------------, zc = zc}]
                      2                                   q1
                    q1

                        q1 (-y4 + v4)
[5, 5, {q1 = q1, q2 = - -------------, q3 = 0, q4 = q1, v4 = v4, w4 = w4,
                                2
                           zc w4

  y4 = y4, z4 = w4, zc = zc}]
```

We get that a member of NonRedFams[5] is equivalent to another member of this family for any value of y_4. Without loss of generality, set $y_4=1$. So, NonRedFams[5] is a representative family with $y_4=1$.

In NonRedFams[4], we obtain any value for z_4, say a, by setting

$$q_2=(w_4-a)q_1/w_3.$$

(Note that by (4.1)–(4.3), z_3, identified by w_3, is not equal to 0.) This choice of q_2 does not violate the invertibility of Q. Further, it is easy to check that in this case, $z_2=w_2$. Thus, without loss of generality, set $z_4=1$. So, NonRedFams[4] is a representative family with $z_4=1$.

Given the results above, we only need to execute Step 3b for NonRedFams[4] and NonRedFams[5], but we complete this for the whole list NonRedFams as follows:

```
IrConditions:=[]:
for i from 1 to nops(NonRedFams) do
Xw:=Multiply(NonRedFams[i][1][1],w): Yw:=Multiply(NonRedFams[i][1][2],w):
Zw:=Multiply(NonRedFams[i][1][3],w):
Ir:=solve([p*conjugate(p)+q*conjugate(q)<>0,
          p*Xw[2][1]-q*Xw[1][1],p*Yw[2][1]-q*Yw[1][1],p*Zw[2][1]-q*Zw[1][1],
          zc<>0,zc^3<>1,zc^3<>-8]):
IrConditions:=[op(IrConditions),[i,Ir]]:
end do:
```

To see the output, enter

```
for i from 1 to nops(IrConditions) do    print(IrConditions[i]):    end do:
```

One gets that, for each i, all members of NonRedFams[i] are irreducible matrix solutions of $S(1,1,c)$.

Now by entering

```
eval(NonRedFams[4],[z4=1]); eval(NonRedFams[5],[y4=1]);
```

one obtains the representatives of equivalence classes of irreducible matrix solutions (X,Y,Z) of equations (1.3), where X is assumed to be one Jordan block. The output is as follows:

$$X=\begin{pmatrix}0&1\\0&0\end{pmatrix},\quad Y=\begin{pmatrix}-\beta/(cz_3)&-\bigl(-cz_2^2z_3-cz_2+2\beta/(cz_3)\bigr)/z_3\\-cz_2z_3-c&\beta/(cz_3)\end{pmatrix},\quad Z=\begin{pmatrix}-1&z_2\\z_3&1\end{pmatrix};$$
$$X=\begin{pmatrix}0&1\\0&0\end{pmatrix},\quad Y=\begin{pmatrix}-1&1/(cz_4^2)\\-cz_4^2&1\end{pmatrix},\quad Z=\begin{pmatrix}-z_4&2/(cz_4)\\0&z_4\end{pmatrix}; \tag{5.1}$$

where

$$\beta=-c^2z_2z_3-c^2+\bigl(3c^4z_2^2z_3^2+3c^4z_2z_3+c^4-cz_3^3+c^4z_3^3z_2^3\bigr)^{1/2}. \tag{5.2}$$

6. Equivalence and irreducibility: two-Jordan-block case

As in the one-Jordan-block case, we wish to classify the matrix solutions from Steps 0–2 (in Section 4) up to equivalence and extract the irreducible equivalence classes. So, we would like to know under what conditions a matrix solution is equivalent to a member of the same/different solution family. We then specify conditions for which the representative of an equivalence class of matrix solutions is irreducible. This achieved with Steps 3a and 3b, respectively, as described in Section 3. In this section, we continue the algorithm of Section 4 in the case when X is two Jordan blocks.

To execute Step 3a, as described in Section 3, enter the code for Step 3a used in Section 5 (see Remark 1.12). (The memory and time for this operation was

27068.0 MB and 523.78 seconds, respectively.) The output of Steps 0–3a can be viewed by entering the following:

```
for i from 1 to nops(BetweenFams) do    print(BetweenFams[i]):    end do:
```

Consider the following snippet of output:

```
                  v3 q4
[1, 1, {q1 = -----, q2 = 0, q3 = 0, q4 = q4, v3 = v3, v4 = y4, w4 = z4,
                   y3

  y3 = y3, y4 = y4, z4 = z4, zc = zc}]
```

Note that $y_3 \neq 0$ in NonRedFams[1] by (4.4)–(4.5). So, NonRedFams[1] is a representative family with y_3 (identified with v_3) equal to 1 without loss of generality.

Now take

```
                   v3 q4
[2, 2,  {q1 = -----, q2 = 0, q3 = 0, q4 = q4, u4 = x4, v3 = v3, x4 = x4,
                    y3

  y3 = y3, zc = zc}]
```

Note that $y_3 \neq 0$ in NonRedFams[2] by (4.4)–(4.5). So, NonRedFams[2] is a representative family with y_3 (identified with v_3) equal to 1 without loss of generality.

Consider the output

```
                   w3 q4
[3, 3,  {q1 = -----, q2 = 0, q3 = 0, q4 = q4, v4 = y4, w3 = w3, w4 = z4,
                    z3

  y4 = y4, z3 = z3, z4 = z4, zc = zc}]
```

Note that $z_3 \neq 0$ in NonRedFams[3] by (4.4)–(4.5). So, NonRedFams[3] is a representative family with z_3 (identified with w_3) equal to 1 without loss of generality.

Next, consider the snippet of output below:

```
                                   2
                             zc x4  q3
[2, 4, {q1 = 0, q2 = - ---------, q3 = q3, q4 = 0, u4 = -x4, v3 = v3,
                               z3 v3

  x4 = x4, z3 = z3, zc = zc}]
```

By (4.4)–(4.5), we have that $z_3 \neq 0$ for NonRedFams[4]. So by the output above, we get that any member of NonRedFams[4] is equivalent to a member NonRedFams[2]. We exclude NonRedFams[4] from further computation.

Consider the output:

```
                  v3 q4                                                  z3 v3
[5, 5, {q1 = -----, q2 = 0, q3 = 0, q4 = q4, v3 = v3, v4 = y4, w3 = -----,
                   y3                                                     y3

  w4 = z4, y3 = y3, y4 = y4, z3 = z3, z4 = z4, zc = zc}]
```

We have that $y_3 \neq 0$ in `NonRedFams[5]` by (4.4)–(4.5). Without loss of generality, we can take y_3 (identified with v_3) to be 1. In this case, $w_3 = z_3$. So, `NonRedFams[5]` is a representative family with $y_3 = 1$.

Now let us take

```
             v3 q4                                                 z3 v3
[5, 6, {q1 = -----, q2 = 0, q3 = 0, q4 = q4, v3 = v3, v4 = y4, w3 = -----,
              y3                                                    y3

  w4 = z4, y3 = y3, y4 = y4, z3 = z3, z4 = z4, zc = zc}]
```

Note that by (4.4)–(4.5), we have $y_3 \neq 0$ for `NonRedFams[6]`. So by the output above, we get that any member of `NonRedFams[6]` is equivalent to a member `NonRedFams[5]`. We exclude `NonRedFams[6]` from further computation.

We still need to analyze the equivalence between members of `NonRedFams[1]`, `NonRedFams[2]`, `NonRedFams[3]`, and `NonRedFams[5]`. In this case, the output is easier to interpret if we run Step 3b before Step 3a again.

Given the results above, we only need to execute Step 3b for `NonRedFams[1]`, `NonRedFams[2]`, `NonRedFams[3]`, and `NonRedFams[5]`, but we complete this for the whole list `NonRedFams` by entering the code for Step 3b (see Remark 1.12). Consider the snippets

```
           y4 q
[1, {p = - ----, q = q, y3 = y3, y4 = y4, z4 = 0, zc = zc}]
            y3

[2, {p = 0, q = q, x4 = 0, y3 = y3, zc = zc}]

           z4 q
[3, {p = - ----, q = q, y4 = 0, z3 = z3, z4 = z4, zc = zc}]
            z3

                                         2
[5, {p = 0, q = q, y3 = 0, y4 = RootOf(_Z  + 1 + _Z) z4, z3 = z3, z4 =
       z4, zc = zc},

    {p = - ----, q = q, y3 = y3, y4 = -----, z3 = z3, z4 = z4, zc = zc}]
           z4 q                       z4 y3
            z3                         z3
```

We obtain that

- members of `NonRedFams[1]`, `NonRedFams[2]`, and `NonRedFams[3]` are irreducible precisely when $z_4 \neq 0$, $x_4 \neq 0$, and $y_4 \neq 0$, respectively, and
- members of `NonRedFams[5]` are irreducible precisely when $\{y_3 \neq 0, y_4 \neq e^{\pm 2\pi i/3} z_4\}$ or $\{y_3 z_4 \neq y_4 z_3\}$.

We execute Step 3a again for the families highlighted above; we refer to the code in Remark 1.12. From the output, we obtain that $z_4 = 0$ in `NewNonRedFams[1]` precisely when any member of `NewNonRedFams[1]` is equivalent to a member of `NewNonRedFams[2]`. On the other hand, we have that $x_4 = 0$ in `NewNonRedFams[2]` precisely when any member of `NewNonRedFams[2]` is equivalent to a member of

`NewNonRedFams[1]`. However, we know members of `NewNonRedFams[1]` and `NewNonRedFams[2]` are reducible when $z_4 = 0$ and $x_4 = 0$, respectively.

Now by a choice of q_2, q_3, we can consider c to be free in `[1,3,...]`. So, we get that $z_4 = \zeta y_4$ for $\zeta^3 = 1$ in `NewNonRedFams[3]` precisely when any member of `NewNonRedFams[3]` is equivalent to a member of `NewNonRedFams[1]`.

Putting this together we conclude that:

- `NewNonRedFams[1]=eval(NonRedFams[1],[y3=1])` is an irreducible representative family when $z_4 \neq 0$;
- `NewNonRedFams[2]=eval(NonRedFams[2],[y3=1])` is an irreducible representative family when $x_4 \neq 0$;
- `NewNonRedFams[3]=eval(NonRedFams[3],[z3=1])` is an irreducible representative family when $y_4 \neq 0$, and there is no overlap with `NewNonRedFams[1]` when $z_4 \neq \zeta y_4$ for $\zeta^3 = 1$;
- `NewNonRedFams[4]=eval(NonRedFams[5],[y3=1])` is an irreducible representative family when $y_4 \neq e^{\pm 2\pi i/3} z_4$ and $z_4 \neq y_4 z_3$.

We obtain the following representatives of equivalence classes of irreducible matrix solutions (X, Y, Z) of equations (1.3), where X is assumed to be two Jordan blocks:

$$X = \begin{pmatrix} cz_4^2/(2y_4) & 0 \\ 0 & -cz_4^2/(2y_4) \end{pmatrix}, \quad Y = \begin{pmatrix} -y_4 & -(y_4^3 - z_4^3)/y_4 \\ 1 & y_4 \end{pmatrix}, \\ Z = \begin{pmatrix} -z_4 & -z_4(8y_4^3 + c^3 z_4^3)/(4y_4^2) \\ 0 & z_4 \end{pmatrix} \tag{6.1}$$

for $z_4 \neq 0$,

$$X = \begin{pmatrix} -x_4 & 0 \\ 0 & x_4 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad Z = \begin{pmatrix} 0 & -c^2 x_4^2 \\ 0 & 0 \end{pmatrix} \tag{6.2}$$

for $x_4 \neq 0$,

$$X = \begin{pmatrix} cy_4^2/(2z_4) & 0 \\ 0 & -cy_4^2/(2z_4) \end{pmatrix}, \quad Y = \begin{pmatrix} -y_4 & -y_4(8z_4^3 + c^3 y_4^3)/(4z_4^2) \\ 0 & y_4 \end{pmatrix}, \\ Z = \begin{pmatrix} -z_4 & (y_4^3 - z_4^3)/z_4 \\ 1 & z_4 \end{pmatrix} \tag{6.3}$$

for $y_4 \neq 0,\ z_4 \neq \zeta y_4,\ \zeta^3 = 1$, and

$$X = \begin{pmatrix} \gamma/(c^2 z_3) & 0 \\ 0 & -\gamma/(c^2 z_3) \end{pmatrix}, \quad Y = \begin{pmatrix} -y_4 & -\big(2z_4\gamma/(c^2 z_3) + cy_4^2\big)/c \\ 1 & y_4 \end{pmatrix}, \\ Z = \begin{pmatrix} -z_4 & -\big(-2\gamma y_4/(c^2 z_3) + cz_4^2\big)/cz_3 \\ z_3 & z_4 \end{pmatrix} \tag{6.4}$$

for $y_4 \neq e^{\pm 2\pi i/3} z_4$, $z_4 \neq y_4 z_3$, where

$$\gamma = -z_3^2 z_4 - y_4 + \left(z_3^4 z_4^2 + 2z_3^2 z_4 y_4 + y_4^2 + c^3 z_3^3 y_4^2 + c^3 z_3 z_4^2 - 2c^3 z_3^2 y_4 z_4\right)^{1/2}. \quad (6.5)$$

7. Geometric parametrization of irreducible representations of $S(1, 1, c)$

Since the Sklyanin algebra $S = S(1, 1, c)$ is module finite over its center, we can use the center Z of S to provide a geometric parametrization of the set of equivalence classes of irreducible representations of S. (Recall by Definition 1.2, $c \neq 0$, $c^3 \neq 1, -8$.) Namely, we depict the *Azumaya locus* of $S(1, 1, c)$ over its center [Brown and Goodearl 2002, III.1.7]. We refer the reader to [Smith et al. 2000] for an introduction to affine varieties.

Theorem 7.1. *Let Z be the center of the Sklyanin algebra $S = S(1, 1, c)$.*

(a) *We have that Z is generated by $u_1 = x^2$, $u_2 = y^2$, $u_3 = z^2$,*

$$g = cy^3 + yxz - xyz - cx^3,$$

subject to the degree-6 relation

$$F := g^2 - c^2(u_1^3 + u_2^3 + u_3^3) - (c^3 - 4)u_1 u_2 u_3 = 0.$$

(b) *The set of equivalence classes of irreducible representations of S is in bijective correspondence with the set of maximal ideals of the center Z of S. Here, a representative ψ of an equivalence class of an irrep of S corresponds to $(\ker \psi) \cap Z$, a maximal ideal of Z.*

(b) *The geometric parametrization of the set of equivalence classes of irreducible representations of S is the 3-dimensional affine variety (3-fold)*

$$X_c := \mathbb{V}(F) \in \mathbb{C}^4_{\{u_1, u_2, u_3, g\}}.$$

In particular, $X_c \setminus \{\underline{0}\}$ is the Azumaya locus of S over Z. Indeed, points of $X_c \setminus \{\underline{0}\}$ (the smooth locus of X_c) correspond to irreducible 2-dimensional representations of S, and the origin of X_c corresponds to the trivial representation of S.

Taking a value of c, say 5, we can visualize the 3-fold X_c by taking 2-dimensional slices at various values of u_1. See Figure 1 below.

Proof of Theorem 7.1. (a) We have that Z is generated by three algebraically independent elements u_1, u_2, u_3 of degree 2 and one element g of degree 3, subject to one relation F of degree 6, by [Smith and Tate 1994, Theorems 3.7, 4.6, and 4.7]. Now part (a) follows by direct computation in the algebra $S(1, 1, c)$. One can do this by hand, but we execute this with the computer algebra software GAP using the GBNP package for noncommutative Gröbner bases [Cohen and Knopper 2016]. We check that u_1, u_2, u_3, g commute with each of $x_1 := x$, $y_1 := y$, $z_1 := z$; the

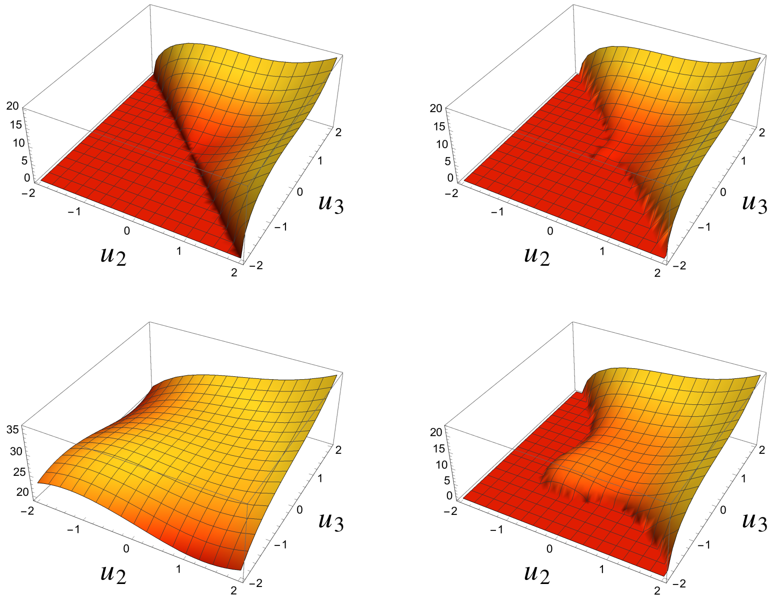


Figure 1. Real part of $\sqrt{25(u_1^3+u_2^3+u_3^3)+21u_1u_2u_3}$ at $u_1 = 0, 0.3, 1, 3$ (clockwise from the top left).

code for this is publicly available, see Remark 1.12. To view g, for instance, enter `PrintNP(g);`. The output of the last twelve lines are all 0. Thus, u_1, u_2, u_3, g are all central elements of $S(1, 1, 5)$. One can replace $c = 5$ with various values of $c \neq 0, 1, -8$, and this yields the same output.

Now to see that F is the relation of Z, more care is needed. Enter

```
PrintNP(MulQA(g,g,GB));
PrintNP(MulQA(u1,MulQA(u1,u1,GB),GB));
PrintNP(MulQA(u2,MulQA(u2,u2,GB),GB));
PrintNP(MulQA(u3,MulQA(u3,u3,GB),GB));
PrintNP(MulQA(u1,MulQA(u2,u3,GB),GB));
```

and compare terms to derive the coefficients of F as claimed.

(b) The arguments below are standard in ring theory and in representation theory, but we provide details for the reader's convenience. Recall from Lemma 2.6 that all nontrivial irreducible representations of S are of dimension 2. Let maxSpec(A) denote the set of maximal ideals of an algebra A. Moreover, a *primitive* ideal of A is an ideal that arises as the kernel of an irreducible representation of A; denote the set of such ideals by prim(A). Take [Irrep(A)] to be the set of equivalence classes of irreducible representations of A.

Since S is PI, we see that there is a bijective correspondence between [Irrep(S)] and prim(S) as follows. Equivalent representations of S have the same kernel, so we get a surjective map ϕ: [Irrep(S)]$\to$ prim(S), given by $\psi \mapsto \ker \psi$. On the other hand, take $P \in \mathrm{prim}(S)$, that is, the kernel of an irreducible representation ψ of S. Then, ψ is also an irreducible representation of S/P. Now $S/P \cong \mathrm{Mat}_t(\mathbb{C})$ for $t = 1$ or 2 by [Brown and Goodearl 2002, Theorem I.13.5(1)], and all irreducible representations of matrix algebras are equivalent to the identity representation by the Skolem–Noether theorem. So, $P \in \mathrm{prim}(S)$ has a unique preimage $\phi^{-1}(P)$ in [Irrep(S)].

Moreover, we see that there is a bijective correspondence between [Irrep(S)] and maxSpec(S) as follows. Maximal ideals are primitive. On the other hand, take P to be a nonzero primitive ideal of S. Again, by [Brown and Goodearl 2002, Theorem I.13.5(1)], S/P is isomorphic to a matrix ring, which is simple. Thus, P is a maximal ideal of S. So it suffices to show that the ideals of maxSpec(S) and of maxSpec(Z) are in bijective correspondence.

Consider the map

$$\eta : \mathrm{maxSpec}(S) \to \mathrm{maxSpec}(Z), \quad M \mapsto M \cap Z.$$

The map η is well-defined and surjective by [Brown and Goodearl 2002, Proposition III.1.1(5)]. Now by Lemma 2.6, the trivial representation of S corresponds to the augmentation (maximal) ideal $S_+ := (x, y, z)$ of S, and the set of equivalence classes of nontrivial irreducible representations of S corresponds to the maximal ideals M of S not equal to S_+. Thus, $\eta(S_+) = Z_+$, and it suffices to show that the ideals of $\mathrm{maxSpec}(S)\backslash S_+$ and of $\mathrm{maxSpec}(Z)\backslash Z_+$ are in bijective correspondence.

Take Az(S) to be the set of maximal ideals $\mathfrak{m}$ of Z so that (i) $\mathfrak{m} = M \cap Z$ for $M \in \mathrm{maxSpec}(S)$, and (ii) M is the kernel of a 2-dimensional irreducible representation of S. Namely, Az(S) is the *Azumaya locus* of S over Z. Consider the map

$$\rho : \mathrm{Az(S)} \to \mathrm{maxSpec}(S), \quad \mathfrak{m} \mapsto \mathfrak{m}S.$$

We get that $\eta\rho(\mathfrak{m}) = \eta(\mathfrak{m}S) = (\mathfrak{m}S) \cap Z = \mathfrak{m}$; the last equality holds by [Brown and Goodearl 2002, Theorem III.1.6(3)]. So, η is bijective on $\rho(\mathrm{Az}(S))$. Since $\mathrm{Az}(S) = \mathrm{maxSpec}(Z)\backslash Z_+$ by Lemma 2.6, and since ρ is injective, we conclude that η is bijective on $\mathrm{maxSpec}(S)\backslash S_+$, as desired.

(c) To see that the claim follows from parts (a) and (b), we have to show that the smooth locus of X_c consists of all nonzero points. This is achieved by using [Smith et al. 2000, Theorem 6.2]; namely, we verify that the common zero set of the vanishing of all partial derivatives of F is the origin of X_c:

```
F:=g^2-c^2*(u1^3+u2^3+u3^3) - (c^3-4)*u1*u2*u3;
solve([diff(F,g),diff(F,u1),diff(F,u2),diff(F,u3)],[g,u1,u2,u3]);
>                 [[g = 0, u1 = 0, u2 = 0, u3 = 0]]
```

This completes the proof. □

Remark 7.2. One may push the result above further and study the *moduli space* (or *GIT quotient*) that parametrizes the set of equivalence classes of irreducible representations of S. But this is not the focus of this work here. On the other hand, if one wants to understand irreducible representations of S topologically, then one could consider the *Jacobson topology* (or *hull-kernel topology*) on the set $\mathrm{prim}(S)$.

Remark 7.3. The code available via Remark 1.12 verifies that the irreps produced in (5.1)–(5.2) and (6.1)–(6.5) indeed correspond to points on X_c. One must first run the algorithm in the previous sections: Sections 4 and 5 for the one-Jordan-block case, and Sections 4 and 6 for the two-Jordan-block case.

By evaluating `simplify(U1);`, `simplify(U2);`, `simplify(U3);`, and `simplify(G);` for each of the six irreducible representative families above, we obtain the corresponding points on the 3-fold $X_c = \mathbb{V}(F) \subset \mathbb{C}^4_{\{u_1,u_2,u_3,g\}}$.

8. Irreducible representations of $\mathbb{C}_{-1}[x, y] := \mathbb{C}\langle x, y\rangle/(xy + yx)$

The purpose of this section is to illustrate our algorithm of Sections 3-6 (Steps 0–2, 3a, 3b) by replacing the Sklyanin algebra $S(1, 1, c)$ with a class of algebras that are much better understood. Here, we study irreducible representations of the skew polynomial ring

$$\mathbb{C}_{-1}[x, y] := \mathbb{C}\langle x, y\rangle/(xy + yx),$$

up to equivalence; these results are well known. At the end of the section, we provide a geometric parametrization of these irreps, akin to Theorem 7.1 for $S(1, 1, c)$. Now we remind the reader of a few preliminary results.

Lemma 8.1. (a) *The* 1*-dimensional irreps of* $\mathbb{C}_{-1}[x, y]$ *are, up to equivalence, of the form*

$$\begin{aligned} \rho_\alpha &: \mathbb{C}_{-1}[x, y] \to \mathbb{C}, \quad x \mapsto \alpha, \quad y \mapsto 0 \quad \textit{for } \alpha \in \mathbb{C}, \\ \rho_\beta &: \mathbb{C}_{-1}[x, y] \to \mathbb{C}, \quad x \mapsto 0, \quad y \mapsto \beta \quad \textit{for } \beta \in \mathbb{C}. \end{aligned} \tag{8.2}$$

(b) *All irreducible representations of* $\mathbb{C}_{-1}[x, y]$ *are finite-dimensional, of at most dimension* 2.

Proof. (a) This follows by an easy computation.

(b) By [Brown and Goodearl 1997, Proposition 3.1; 2002, Example I.14.3(1)], an irrep of $\mathbb{C}_{-1}[x, y]$ is of at most dimension 2. □

With the lemma above, we see that to classify irreps of $\mathbb{C}_{-1}[x, y]$, we just need to compute the 2-dimensional irreps

$$\psi : \mathbb{C}_{-1}[x, y] \to \mathrm{Mat}_2(\mathbb{C}), \quad x \mapsto X, \quad y \mapsto Y,$$

up to equivalence.

Without loss of generality, we can assume that X is in Jordan form; that is, either one Jordan block or diagonal. Now the code for this part (available publicly, see Remark 1.12) was adapted from Sections 3–6 by removing all lines and conditions involving the generator z, and by changing the defining relations of the algebra.

We obtain the result below.

Proposition 8.3. *All irreducible representations ϕ of $\mathbb{C}_{-1}[x, y]$ are of dimensions 1 or 2. In dimension 1, irreps are of the form (8.2). In dimension 2, all irreps, up to equivalence, take the form*

$$\psi_{\alpha,\beta} : \mathbb{C}_{-1}[x, y] \longrightarrow \mathrm{Mat}_2(\mathbb{C}), \quad x \mapsto \begin{pmatrix} -\alpha & 0 \\ 0 & \alpha \end{pmatrix}, \quad y \mapsto \begin{pmatrix} 0 & 1 \\ \beta & 0 \end{pmatrix} \tag{8.4}$$

for $\alpha, \beta \in \mathbb{C}$ with $\alpha\beta \neq 0$.

Proof. The first two statements follow from Lemma 8.1. To get the last statement, we run the adapted algorithm above. We only obtain reducible representations in the one-Jordan-block case; just enter `NonRedFams;` and `IrConditions;` to see this.

On the other hand, in the two-Jordan-block case, we first print off `NonRedFams` (we've converted the output to standard format for readability):

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} y_1 & y_2 \\ y_3 & y_4 \end{bmatrix}; \qquad \begin{bmatrix} x_1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & y_4 \end{bmatrix};$$

$$\begin{bmatrix} 0 & 0 \\ 0 & x_4 \end{bmatrix}, \begin{bmatrix} y_1 & 0 \\ 0 & 0 \end{bmatrix}; \qquad \begin{bmatrix} x_1 & 0 \\ 0 & x_4 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix};$$

$$\begin{bmatrix} -x_4 & 0 \\ 0 & x_4 \end{bmatrix}, \begin{bmatrix} 0 & y_2 \\ y_3 & 0 \end{bmatrix}.$$

Consider the following snippets of output from `BetweenFams`:

```
[2, 3, {q1 = 0, q2 = q2, q3 = q3, q4 = 0, u1 = x4, v4 = y1, x4 = x4,
  y1 = y1}]
                                                         y3 q2       y2 q3
[5, 5, {q1 = 0, q2 = q2, q3 = q3, q4 = 0, u4 = -x4, v2 = -----, v3 = -----,
                                                          q3          q2

  x4 = x4, y2 = y2, y3 = y3}]
```

So, any member of `NonRedFams[3]` is equivalent to a member of `NonRedFams[2]`, and therefore `NonRedFams[3]` is removed from our consideration.

Moreover, `NonRedFams[5]` forms an equivalence family as x_4, y_2, y_3 are free. Take into consideration the output from `IrConditions`:

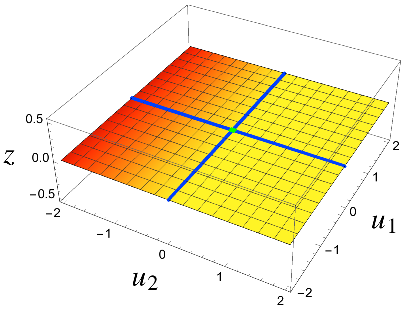


Figure 2. Affine 2-space parametrizing irreps of $\mathbb{C}_{-1}[x, y]$; axes parametrize 1-dimensional irreps.

```
                            2                2
                        y3 p  + p y4 q - y2 q
[1, {p = p, q = q, y1 = ----------------------, y2 = y2, y3 = y3, y4 = y4}]
                                 p q

[2, {p = p, q = 0, x1 = x1, y4 = y4}]

[4, {p = p, q = 0, x1 = x1, x4 = x4}]

[5, {p = p, q = 0, x4 = x4, y2 = y2, y3 = 0},

    {p = 0, q = q, x4 = x4, y2 = 0, y3 = y3},

                                    2
                                y3 p
    {p = p, q = q, x4 = 0, y2 = -----, y3 = y3}]
                                  2
                                 q
```

Now, we can conclude that NonRedFams[1], NonRedFams[2], NonRedFams[4] consist of reducible representations, so these families are eliminated from our consideration. Further, NonRedFams[5] forms an irreducible representative family with $y_2 = 1$; we can see this by adapting and running the algorithm for Step 3b in Section 6 in this case. □

The geometric parametrization of the equivalence classes of irreducible representations of $\mathbb{C}_{-1}[x, y]$ is given as follows; see also Figure 2.

Corollary 8.5. *We have the following statements.*

(a) *We have that the center Z of $\mathbb{C}_{-1}[x, y]$ is the commutative polynomial ring generated by $u_1 := x^2$ and $u_2 := y^2$.*

(b) *The set of equivalence classes of irreducible representations of S are in bijective correspondence with the set of maximal ideals of $\mathbb{C}[x^2, y^2]$.*

(c) *The geometric parametrization of the set of equivalence classes of irreducible representations of* $\mathbb{C}_{-1}[x, y]$ *is the* 2*-dimensional affine space* $\mathbb{C}^2_{\{u_1,u_2\}}$. *In particular*

- *points of* $\mathbb{C}^2 \setminus \mathbb{V}(u_1u_2)$ *correspond to irreducible* 2*-dimensional representations of* $\mathbb{C}_{-1}[x, y]$,
- *points on the axes* $\mathbb{V}(u_1u_2)$ *not equal to the origin correspond to nontrivial* 1*-dimensional representations of* $\mathbb{C}_{-1}[x, y]$, *and*
- *the origin corresponds to the trivial representation of* $\mathbb{C}_{-1}[x, y]$.

Proof. (a) The algebra $\mathbb{C}_{-1}[x, y]$ has a $\mathbb{C}$-vector space basis given by $\{x^i y^j \mid i, j \in \mathbb{N}\}$. Since $(x^i y^j)x = (-1)^j x^{i+1} y^j = x^{i+1} y^j$ and $y(x^i y^j) = (-1)^i x^i y^{j+1} = x^i y^{j+1}$ implies that i, j are even, the result is clear.

(b) This follows by the proof of Theorem 7.1(b).

(c) The first statement follows, as $\operatorname{Spec}(Z) = \mathbb{C}^2_{\{u_1,u_2\}}$. Now the remaining statements hold by (8.4) and (8.2), where $u_1 = \alpha^2$ and $u_2 = \beta$. □

Acknowledgments

This project was inspired by Walton's conversation with Andrew Morrison and Balázs Szendrői at the Interactions between Algebraic Geometry and Noncommutative Algebra workshop at the Mathematisches Forschungsinstitut Oberwolfach in May 2014. The authors are grateful to the referees of our of submission, particularly to a referee who discussed with us a approach to this problem using Clifford algebras (as mentioned in Remark 1.1). Travel for Reich was supported by an Undergraduate Research Supplement of Walton's National Science Foundation grant DMS-1550306.

References

[Artin et al. 1990] M. Artin, J. Tate, and M. Van den Bergh, "Some algebras associated to automorphisms of elliptic curves", pp. 33–85 in *The Grothendieck Festschrift, I*, edited by P. Cartier et al., Progr. Math. **86**, Birkhäuser, Boston, 1990. MR Zbl

[Artin et al. 1991] M. Artin, J. Tate, and M. Van den Bergh, "Modules over regular algebras of dimension 3", *Invent. Math.* **106**:2 (1991), 335–388. MR Zbl

[Bellamy et al. 2016] G. Bellamy, D. Rogalski, T. Schedler, J. T. Stafford, and M. Wemyss, *Noncommutative algebraic geometry*, Mathematical Sciences Research Institute Publications **64**, Cambridge Univ. Press, 2016. MR Zbl

[Brown and Goodearl 1997] K. A. Brown and K. R. Goodearl, "Homological aspects of Noetherian PI Hopf algebras of irreducible modules and maximal dimension", *J. Algebra* **198**:1 (1997), 240–265. MR Zbl

[Brown and Goodearl 2002] K. A. Brown and K. R. Goodearl, *Lectures on algebraic quantum groups*, Birkhäuser, Basel, 2002. MR Zbl

[Cohen and Knopper 2016] A. M. Cohen and J. W. Knopper, "GAP package GBNP: computing Gröbner bases of noncommutative polynomials", software package, 2016, available at http://www.gap-system.org/Packages/gbnp.html.

[De Laet and Le Bruyn 2015] K. De Laet and L. Le Bruyn, "The geometry of representations of 3-dimensional Sklyanin algebras", *Algebr. Represent. Theory* **18**:3 (2015), 761–776. MR Zbl

[Iyudu and Shkarin 2017] N. Iyudu and S. Shkarin, "Three dimensional Sklyanin algebras and Gröbner bases", *J. Algebra* **470** (2017), 379–419. MR Zbl

[McConnell and Robson 2001] J. C. McConnell and J. C. Robson, *Noncommutative Noetherian rings*, revised ed., Graduate Studies in Mathematics **30**, American Mathematical Society, Providence, RI, 2001. MR Zbl

[Smith and Tate 1994] S. P. Smith and J. Tate, "The center of the 3-dimensional and 4-dimensional Sklyanin algebras", *K-Theory* **8**:1 (1994), 19–63. MR Zbl

[Smith et al. 2000] K. E. Smith, L. Kahanpää, P. Kekäläinen, and W. Traves, *An invitation to algebraic geometry*, Springer, New York, 2000. MR Zbl

[Stafford and Van den Bergh 2001] J. T. Stafford and M. Van den Bergh, "Noncommutative curves and noncommutative surfaces", *Bull. Amer. Math. Soc.* (*N.S.*) **38**:2 (2001), 171–216. MR Zbl

[Walton 2012] C. Walton, "Representation theory of three-dimensional Sklyanin algebras", *Nuclear Phys. B* **860**:1 (2012), 167–185. MR Zbl

Received: 2016-10-14 Revised: 2017-02-08 Accepted: 2017-02-22

djreich@ncsu.edu *Department of Mathematics, North Carolina State University,*
Raleigh, NC, United States

notlaw@temple.edu *Department of Mathematics, Temple University,*
Philadelphia, PA, United States

INVOLVE 11:4 (2018)
dx.doi.org/10.2140/involve.2018.11.609

A classification of Klein links as torus links

Steven Beres, Vesta Coufal, Kaia Hlavacek, M. Kate Kearney,
Ryan Lattanzi, Hayley Olson, Joel Pereira and Bryan Strub

(Communicated by Kenneth S. Berenhaut)

We classify Klein links. In particular, we calculate the number and types of components in a $K_{p,q}$ Klein link. We completely determine which Klein links are equivalent to a torus link, and which are not.

1. Introduction

When we began thinking about Klein knots, we were told that they were uninteresting since all Klein knots are torus knots. We decided to see if we could prove that statement using elementary methods, and to consider whether it was also true about Klein links. In our first paper [Alvarado et al. 2016], we presented our constructions and results leading up to our discovery of a class of Klein links that are not equivalent to any torus links.

In this paper, we show exactly which Klein links are torus links, and which are not. We begin in Section 2 with defining our notation for Klein links, which is based on the standard notation for torus links. In Section 3 we define two functions, the wrapping function and the hitting function, which help us to describe components of our links as they traverse a standard link diagram. We introduce several preliminary results in Section 4. We compute the number of components in a link $K_{p,q}$. Each of these components is itself a Klein knot, and we also describe the knot type of these components. Section 5 includes our main result, Theorem 12, which gives a complete classification of which Klein links are equivalent to torus links and which are *knot*.

Some of our results are identical or similar to results proved by another group using braids. However, our methods are different. Explicitly, our Lemma 2 is [Bush et al. 2014, Proposition 6.1], our Theorem 3 is [Bush et al. 2014, Proposition 6.2], and our Lemma 7 is [Catalano et al. 2010, Theorem 2].

MSC2010: 57M25.
Keywords: knot theory, torus links, Klein links.
The authors received partial support from McDonald Work Awards .

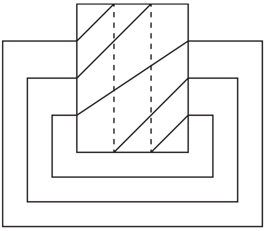


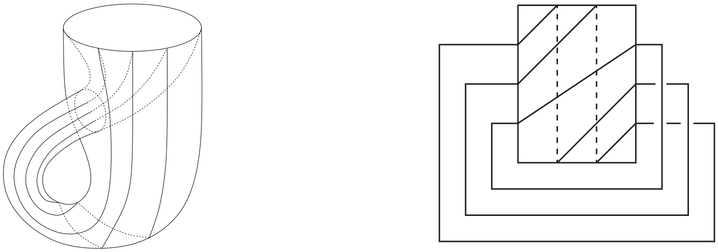
Figure 1. Planar diagram for the torus knot $T_{2,3}$.

2. Constructions

We begin with a brief description of the standard construction of torus links [Adams 1994; Murasugi 1996] and our analogous construction of Klein links. For nonnegative integers p and q, the torus link $T_{p,q}$ is the link on the torus which crosses the longitude p times and crosses the meridian q times, with no crossing on the torus itself. We illustrate the construction of $T_{2,3}$ on a planar diagram in Figure 1. The rectangle in the figure is a planar diagram for the torus, with the gluings (left side to right side, and top to bottom) understood.

We will construct Klein links in a similar way, being careful of certain issues. Klein bottles do not exist in three-dimensional space, and knots are trivial in four-dimensional space. To avoid this, we will work with punctured Klein bottles in three-dimensional space. The puncture occurs where the Klein bottle appears to (but does not) intersect itself. Warning: the notation of the knots and links we work with will be dependent on the relative position of the puncture. Mimicking the construction of $T_{p,q}$, the Klein knot $K_{2,3}$ is illustrated in Figure 2.

The corresponding planar diagram representation of $K_{2,3}$ is modeled after the torus version, except that we need to account for the Möbius-band twist and be mindful of the puncture. We deform the Klein bottle so that the twist produces a pattern of additional crossings as in Figure 2, with the puncture occurring in the lower left corner.



$K_{2,3}$ on a Klein bottle

planar diagram for $K_{2,3}$

Figure 2. Klein link $K_{2,3}$.

Note that $K_{p,0}$ is the p-component unlink.

We emphasize that the class of links that we are denoting by $K_{p,q}$ and the results in this paper are dependent on placing the puncture in the lower left corner. We do not consider Klein links with the puncture placed in different positions in this paper. Furthermore, deformations of our links are as links in space, not on the Klein bottle, and so the puncture does not affect deformations. For this reason, and since our puncture is always in the lower left corner, we do not include it in our illustrations.

It is worth noting that, while the diagrams are configured a bit differently, our $K_{p,q}$ Klein links are the same as the $K(p,q)$ Klein links found in [Bush et al. 2014; Freund and Smith-Polderman 2013; Shepherd et al. 2012]. Additionally, some of the same authors of the previously cited papers have done preliminary work in which they found explicit relationships between Klein links with different choices of puncture. There are certainly more questions to be answered in this regard.

3. The wrapping and hitting functions

We start with some definitions.

Definition 1. A *component* is a maximal connected subset of the link. A *horizontal node* is a position on the top of the planar diagram that a component passes through. A *vertical node* is a position on the left of the planar diagram that a component passes through. A *strand* is a subset of a component that passes exactly once horizontally through the planar diagram. Typically we denote the strand by the vertical node the strand passes through on the left side of the planar diagram.

The underlying keys to many of our results are our "wrapping" and "hitting" functions. Given a component entering the left side of the rectangle in the planar diagram construction of $K_{p,q}$ (see Figure 3), the wrapping function describes where that particular component reenters the left side of the rectangle. For $1 \le x \le q$, let x be the node in $K_{p,q}$ as in Figure 3. Then the wrapping function is given by

$$W_{p,q}(x) = 1 - x + p \pmod q.$$

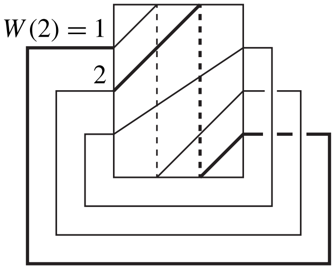


Figure 3. For $K_{2,3}$, $W(2) = 1$.

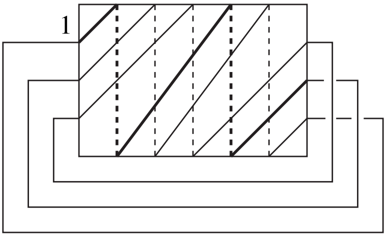


Figure 4. The hitting function $H_{5,3}(1) = 2$.

For an in-depth exploration of the wrapping function, as well as a proof of the next lemma, see [Alvarado et al. 2016].

Lemma 2. *For any $p, q \geq 0$, we have $W^2_{p,q}(x) = x$. Therefore, every component of $K_{p,q}$ wraps at most twice.*

While the wrapping function describes the horizontal movement of a strand, the hitting function addresses the vertical travel. Given a particular strand x starting at node x in a planar diagram of a $K_{p,q}$, we can determine how many times x hits the top of the planar diagram before reaching the right edge of the planar diagram. This is denoted by $H_{p,q}(x)$. Given p, q and x, where $1 \leq x \leq q$, we can use the following formula to find $H_{p,q}(x)$:

$$H_{p,q}(x) = \left\lfloor \frac{p-x}{q} \right\rfloor + 1, \tag{1}$$

where $\lfloor t \rfloor$ is the greatest integer function.

Note that the hitting function depends on the strand. To see how many vertical nodes a component passes through, we apply the hitting function to each strand in the component and add.

Applying the hitting function to the first strand of $K_{5,3}$ gives $H_{5,3}(1) = 2$, which is illustrated in Figure 4.

To see that the hitting function is defined correctly, notice that by construction a strand passes through the $(k+1)$ horizontal nodes $x,\ x+q,\ x+2q, \ldots, x+kq$, where $x + kq \leq p < x + (k+1)q$. So we have

$$\begin{aligned} x + kq \leq p &< x + (k+1)q, \\ kq \leq p - x &< (k+1)q, \\ k \leq \frac{p-x}{q} &< k+1. \end{aligned}$$

It follows that $k+1 = \lfloor (p-x)/q \rfloor + 1$. Thus the hitting function is correctly defined in (1).

4. Preliminary results

Our primary goal in this paper is to describe which Klein links are equivalent to torus links. In the interest of doing so, we will build up several results that break down the link $K_{p,q}$ into components and describe those components. Our first result gives the number of components in the link $K_{p,q}$.

Theorem 3 (number of components). *For a Klein link $K_{tq+n,q}$ with $q > 0$, $t \geq 0$ and $0 \leq n < q$:*

- *For q odd there are $\frac{1}{2}(q+1)$ components.*
- *For q even, n even there are $\frac{1}{2}q$ components.*
- *For q even, n odd there are $\frac{1}{2}q+1$ components.*

Moreover, in the case that $\frac{1}{2}(n+1)$ or $\frac{1}{2}(q+n+1)$ are integers, the strands at these nodes wrap only once. All other strands wrap twice.

Proof. It is enough the count the number of strands that wrap once. Then we divide the number of remaining vertical nodes by 2 to find how many components wrap twice, then add these two values.

To find the single-wrapping components, consider the equation $W_{tq+n,q}(x) = x$ (that is, the strand x wraps to itself). Then,

$$\begin{aligned} x = W_{tq+n,q}(x) &\equiv 1 - x + tq + n \pmod{q}, \\ x &\equiv 1 - x + n \pmod{q}, \\ 2x &\equiv n + 1 \pmod{q}. \end{aligned}$$

We will make use of this last modular equation in the following cases.

Case 1 (q *odd*): Since q is odd, 2 has a multiplicative inverse modulo q, which is $\frac{1}{2}(q+1)$. Solving the modular equation above, we have

$$\begin{aligned} 2x &\equiv n + 1 \pmod{q}, \\ x &\equiv \tfrac{1}{2}(q+1)(n+1) \pmod{q}. \end{aligned}$$

Thus $x = \frac{1}{2}(q+1)(n+1) + kq$ for some integer k. Since $1 \leq x \leq q$, we have $1 \leq \frac{1}{2}(q+1)(n+1) + kq \leq q$. The length of this interval is $q-1$; thus there can be at most one k-value solution. Since q is odd, there is at least one strand that wraps only once, which means there is at least one k-value solution. It follows that there is exactly one k-value solution, and thus exactly one component that wraps once and $\frac{1}{2}(q-1)$ components that wrap twice. Therefore, we have $\frac{1}{2}(q-1)+1 = \frac{1}{2}(q+1)$ components.

Case 2 (q *even, n even*): In this case, rewriting the modular equation we get that $2x = n + 1 + kq$ for some integer k. We have that $2x$ and kq are even integers, and

$n+1$ is an odd integer. Then the equation $2x = n+1+kq$ has no solutions, and thus every component wraps twice. Therefore, there are $\frac{1}{2}q$ components.

Case 3 (*q even, n odd*): In this case, we again have $2x = n+1+kq$. Solving the equation for x gives $x = \frac{1}{2}(n+1)+\frac{1}{2}kq$. Recall that $1 \le x \le q$ and $0 \le n < q$. Thus, we have two solutions: one when $k = 0$ and the other when $k = 1$. Thus there are two components that wrap once and $\frac{1}{2}(q-2)$ components that wrap twice. Therefore, there are $\frac{1}{2}(q-2)+2 = \frac{1}{2}q+1$ components. □

Now that we have determined the number of components in a $K_{tq+n,q}$, we would like to know how many times each of these components wraps around the meridian and longitude of the Klein bottle, as well as their knot type. We will denote by $L = a \cdot P \cup b \cdot Q$ a link which is composed of a copies of a knot (or link) P, and b copies of knot (or link) Q. The copies of P and Q may be linked.

Theorem 4 (types of components). *Consider $K_{tq+n,q}$ with $q > 0$, $t \ge 0$ and $0 \le n < q$. Then*:

(1) *If q even and n odd*, *then*

$$K_{tq+n,q} \equiv \tfrac{1}{2}(n-1) \cdot K_{2t+2,2} \cup \tfrac{1}{2}(q-n-1) \cdot K_{2t,2} \cup K_{t+1,1} \cup K_{t,1}.$$

(2) *If q, n odd*, *then*

$$K_{tq+n,q} \equiv \tfrac{1}{2}(n-1) \cdot K_{2t+2,2} \cup \tfrac{1}{2}(q-n) \cdot K_{2t,2} \cup K_{t+1,1}.$$

(3) *If q odd and n even*, *then*

$$K_{tq+n,q} \equiv \tfrac{1}{2}n \cdot K_{2t+2,2} \cup \tfrac{1}{2}(q-n-1) \cdot K_{2t,2} \cup K_{t,1}.$$

(4) *If q, n even*, *then*

$$K_{tq+n,q} \equiv \tfrac{1}{2}n \cdot K_{2t+2,2} \cup \tfrac{1}{2}(q-n) \cdot K_{2t,2}.$$

Proof. According to Theorem 3, the only components that wrap once are the components through $x_1^* = \frac{1}{2}(n+1)$ and $x_2^* = \frac{1}{2}(q+n+1)$ when these values are integers (one or both), and all other components wrap twice.

It is advantageous to inspect the wrapping function $W(x)$ for a number of specific values:

$$\begin{aligned} W(1) &= n, & W(n+1) &= q,\\ W(2) &= n-1, & W(n+2) &= q-1,\\ W(3) &= n-2, & W(n+3) &= q-2. \end{aligned}$$

In general, we have

$$\begin{aligned} W(x) &= \ n+1-x && \text{for } x < x_1^*,\\ W(x) &= \ q+n+1-x && \text{for } n < x < x_2^*. \end{aligned}$$

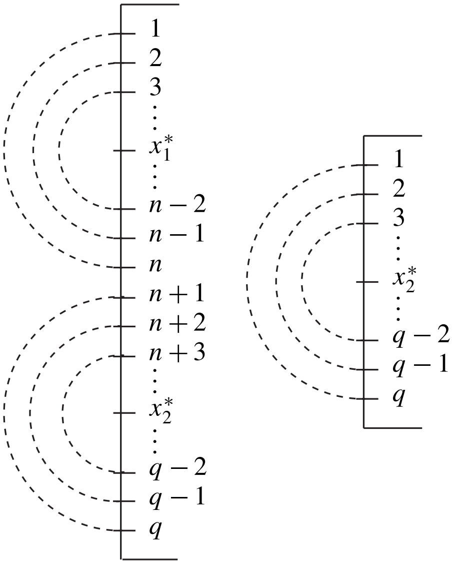


Figure 5. The wrapping of $K_{tq+n,q}$ (on the left), and the wrapping of $K_{tq,q}$ (on the right).

We see that there are now two symmetry points, $x_1^* = \frac{1}{2}(n+1)$ and $x_2^* = \frac{1}{2}(q+n+1)$, regardless of whether these are integers or not, and the wrapping of $K_{tq+n,q}$ can be pictured as in the left side of Figure 5.

If $n = 0$, however, we see that $x_1^* = \frac{1}{2}$ and there are no nodes $x < x_1^*$. In that case, we have only one symmetry point as in the right side of Figure 5.

Next, recalling that $0 \le n < q$ and $1 \le x \le q$, we simplify the hitting function as follows:

$$H(x) = \left\lfloor \frac{tq+n-x}{q} \right\rfloor + 1 = \left\lfloor \frac{n-x}{q} \right\rfloor + t + 1 = \begin{cases} t+1 & \text{if } x \le n, \\ t & \text{if } x > n. \end{cases}$$

Notice that the components symmetric about (but not on) x_1^* wrap twice and hit $t+1$ times on each wrap, so they are all of the form $K_{2(t+1),2} = K_{2t+2,2}$. When n is odd, there is a component passing through x_1^* and it wraps once and hits $t+1$ times, making it a $K_{t+1,1}$. Components symmetric about (but not on) x_2^* wrap twice and hit t times on each wrap, so they are all of the form $K_{2t,2}$. When $q+n$ is odd, there is a component passing through x_2^* and it wraps once and hits t times, making it a $K_{t,1}$.

All that is left is to count the number of components of each type, depending on the parity of q and n, using Theorem 3. For example, if q is even and $n > 0$ is even, then there are a total of $\frac{1}{2}q$ components, with $\frac{1}{2}n$ of them symmetric about x_1^* and $\frac{1}{2}q - \frac{1}{2}n = \frac{1}{2}(q-n)$ of them about x_2^*. Thus, in this case, $K_{tq+n,q} \equiv \frac{1}{2}n \cdot K_{2t+2,2} \cup \frac{1}{2}(q-n) \cdot K_{2t,2}$. We leave it to the reader to finish counting for the remaining three cases. □

We now have a complete characterization of the types of components for any Klein link. To establish an equivalence to a torus link, we need to establish an equivalence of the components. We present a collection of lemmas about the components of torus and Klein links that we will use to prove the classification theorem.

In the next lemma, and many of the subsequent results, we make use of the linking number of a pair of components in a link.

Definition 5. To define the *linking number* of two components C_1 and C_2 of a link, we first orient the link (choose a direction of travel for each component). Next, assign $+1$ to a crossing between if the undergoing strand goes from the right side to the left side of the overgoing strand (right-handed crossing). If the undergoing strand moves from left to right (left-handed crossing) it is assigned a -1. Considering all crossings involving a strand from C_1 and a strand from C_2, add all of the signed crossing numbers (the $+1$s and -1s), take the absolute value of this sum, and divide by two. The resulting value is called the linking number of the two components, and is denoted by $\mathrm{lk}(C_1, C_2)$.

Lemma 6. *All components of a torus link have the same knot type. Additionally, every pair of components in a torus link have the same linking number.*

Proof. As discussed in Section 2, the torus link $T_{p,q}$ is given by identifying the top and bottom, and left and right sides of the square together with the knot that hits the top p times and the side q times, that is, the line with slope p/q, and appropriate translation; see [Flapan 2016]. We can identify components by examining each strand along the left-hand side, just as we have for Klein links. In contrast to the picture with Klein links, we can make the observation here that a vertical translation by $1/q$ produces the same link, but with the ordering of strands (and hence components) shifted by 1. Since each component is a translation of the others, all components must have the same knot type.

Next we consider the linking of pairs of components. As we saw above, each component is a translation of the others. Furthermore, considering the strands along the left-hand side, if we have n components they must be represented by the first n strands from the top. If we translate a strand vertically by a/q and find that we have reached another strand of the same component, then every translation by $(c * a)/q$ will also return the same component. Hence, to have n components, we must find our first repeated component in the translation by n/q (so the first strand shifts to the $(n+1)$-st strand), and so the first n strands each represent different components. Now, we see that if we consider components x_i, x_j, and x_k, we know that x_k is a translation of x_j, and in particular it is a translation by less than n/q, and hence does not cross any strand of x_i in the process of translating. This is enough to guarantee that the linking number of x_i with x_j is equal to the linking number of x_i with x_k. Finally, we see that any pair of components in the torus link have the same linking number. □

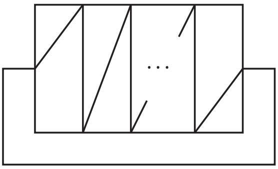


Figure 6. $K_{p,1}$ is an unknot.

The proof of Lemma 7 follows directly from the construction; see Figure 6.

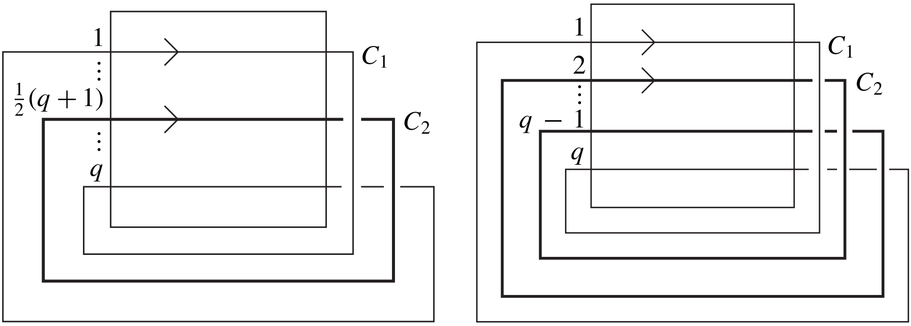
Lemma 7. *For all* p, $K_{p,1}$ *is an unknot.*

The next two lemmas address the linking numbers of certain components of $K_{p,q}$ in special cases.

Lemma 8. *If* $q \geq 3$ *is odd, then* $K_{0,q}$ *contains a pair of components with linking number* 1. *If* $q \geq 4$ (*even or odd*), *then* $K_{0,q}$ *contains a pair of components with linking number* 2.

Proof. First note that $K_{0,q}$ has crossings only outside of the rectangle, and all crossings are of the same type (with all strands oriented to point into the right-hand side of the rectangle, and all crossings are right-hand crossings).

For $q \geq 3$ and odd, let C_1, C_2 be the components passing through nodes 1 and $\frac{1}{2}(q+1)$, respectively. Using the wrapping function, we have that $W(1) = 1-1+0 \equiv q \pmod{q}$ and $W\big(\frac{1}{2}(q+1)\big) = 1 - \frac{1}{2}(q+1) + 0 \equiv q + \frac{1}{2}(1-q) \equiv \frac{1}{2}(q+1) \pmod{q}$. Thus, component C_1 passes through nodes 1 and q, wrapping twice, and C_2 passes through node $\frac{1}{2}(q+1)$ and wraps only once. See Figure 7(a). Since C_1 wraps twice, while C_2 wraps only once, they cross each other exactly twice. Hence C_1 and C_2 have exactly two crossings, both outside of the rectangle, and the linking number is $\operatorname{lk}(C_1, C_2) = \frac{2}{2} = 1$.



(a) $K_{0,q}$ with q odd

(b) $K_{0,q}$ with q even

Figure 7. $K_{0,q}$.

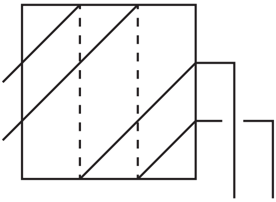


Figure 8. Two components of $K_{n,n}$ on a single wrap.

For $q \geq 4$, let C_1 again be the component passing through nodes 1 and q. Using the wrapping function, we denote by C_2 the component that passes through 2 and $q-1$. See Figure 7(b). In particular, they both wrap twice. It follows that they cross each other exactly four times, and the linking number is $\mathrm{lk}(C_1, C_2) = \frac{4}{2} = 2$. □

Lemma 9. *For $n \geq 3$, $K_{n,n}$ has a pair of components with nonzero linking number.*

Proof. Considering the planar diagram, all crossings inside of the rectangle are left-hand crossings, with our choice of orientation, and every crossing outside of the rectangle is right-handed. Let C_1, C_2 be the components passing through nodes 1 and 2, respectively. We will calculate the linking number for the pair C_1, C_2 by counting the number of crossings inside of the rectangle and the number of crossings outside.

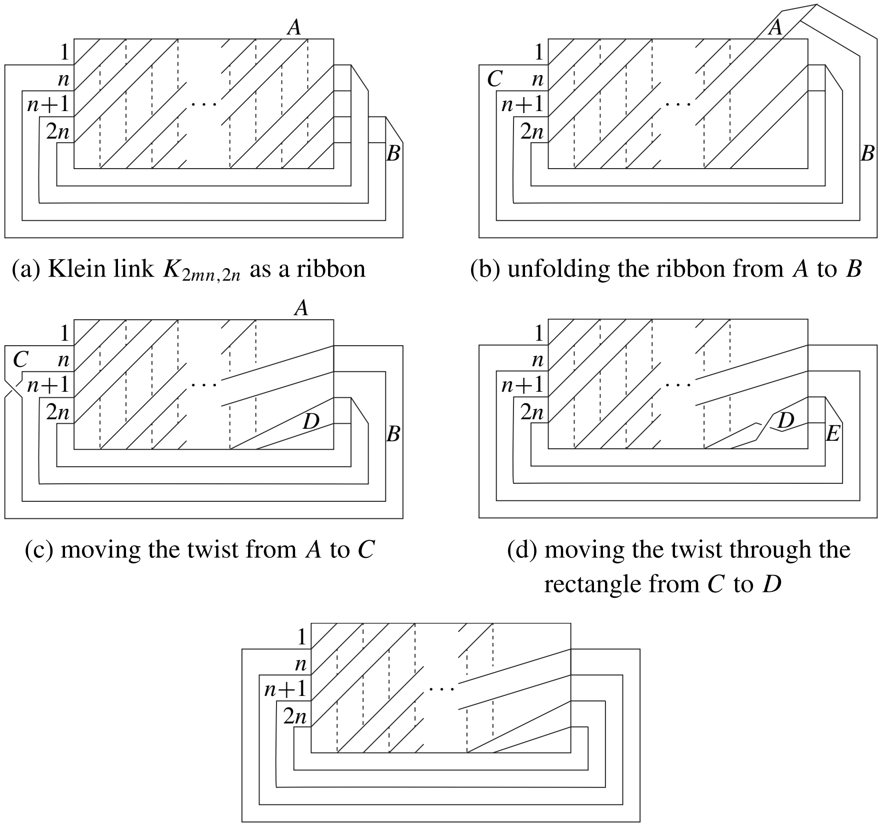
First, we have that $W(1) = 1 - 1 + n = n \neq 1 \pmod{n}$ and so C_1 wraps twice for $n \geq 3$. If $n = 3$, then $W(2) = 1 - 2 + 3 = 2 \pmod{3}$, and thus C_2 wraps once. If $n \geq 4$, then $W(2) = n - 1 \neq 2 \pmod{n}$ and so C_2 wraps twice. For all $1 \leq x \leq n$, $H(x) = \lfloor (n-x)/n \rfloor + 1 = 1$. Thus each component hits the top of the rectangle exactly once each time it wraps. It follows that, on each wrap, the two components cross twice in the rectangle and once outside of the rectangle, as shown in Figure 8.

For $n = 3$, C_1 wraps twice and C_2 wraps once, so they cross a total of $2(2) = 4$ times inside the rectangle and $2(1) = 2$ times outside the rectangle. The linking number is $|(2-4)/2| = 1$. For $n \geq 3$, both C_1 and C_2 wrap twice, so they cross a total of $4(2) = 8$ times inside the rectangle and $2(2) = 4$ times outside the rectangle, giving a linking number of $\left|\frac{1}{2}(4-8)\right| = 2$. In both cases, the pair has nonzero linking number. □

The next lemma is a generalization of [Alvarado et al. 2016, Theorem 6] and was proved by one of the authors of that paper, Enrique Alvarado.

Lemma 10. *For all m and n, we have $K_{2mn,2n} \equiv T_{2mn-n,2n}$.*

Proof. A $K_{2mn,2n}$ has $2n$ strands entering or leaving each side of the rectangle in the planar diagram. We collect together the first n strands (strands 1 through n) to form a single ribbon. Notice that since $W(1) \equiv 2n \pmod{2n}$ and $W(n) \equiv n+1 \pmod{2n}$, the ribbon exits the right side and wraps around to reenter the left side through



(a) Klein link $K_{2mn,2n}$ as a ribbon

(b) unfolding the ribbon from A to B

(c) moving the twist from A to C

(d) moving the twist through the rectangle from C to D

(e) canceling the twists at D and E

Figure 9. Manipulating the ribbon form of $K_{2mn,2n}$ into $T_{2mn-n,2n}$.

strands $n+1$ through $2n$. Thus the entire link $K_{2mn,2n}$ consists of just one ribbon that wraps twice from left to right, as in Figure 9(a).

The transformation to $T_{2mn-n,2n}$ is illustrated in Figure 9. First, unfold the ribbon between the points labeled A and B, as in Figure 9(b), then move the remaining twist at A through B to C, as in Figure 9(c). We also slide the ribbon at point A down from the top of the rectangle to the right side, leaving $2mn-n$ strands through the top and $2n$ strands through the right side of the rectangle. Next, move the twist through the rectangle to point D. To do this, we are doing a series of moves as shown in Figure 10.

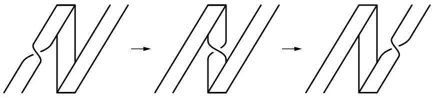


Figure 10. Moving the twist through the rectangle.

We end up with a twist at D, as in Figure 9(d). The twist at D cancels the twist at E, resulting in the ribbon form of the torus link $T_{2mn-n,2n}$, as in Figure 9(e). □

Lemma 11. *For $t \geq 2$, we have $K_{2t+2,2} \not\equiv K_{2t,2}$, and neither are unknots.*

Proof. By Lemma 10, $K_{2t+2,2} \equiv T_{2t+1,2}$ and $K_{2t,2} \equiv T_{2t-1,2}$. The torus links are nontrivial and not equivalent since they have different determinants [Livingston 1993]. Hence, $K_{2t+2,2} \equiv T_{2t+1,2} \not\equiv T_{2t-1,2} \equiv K_{2t,2}$. □

5. The classification theorem

Having built our preliminary results, we are now ready to state and prove our main result, which describes exactly which Klein links are equivalent to torus links, and which are not. Without further ado…

Theorem 12 (the classification theorem). *Let $p = tq + n$ with $t \geq 0$ and $0 \leq n < q$. All Klein links $K_{p,q}$ which are equivalent to torus links are listed in the following table*:

q	$0, 1, 2$	3	4	*even*	*odd*
p	$0 \leq p$	$0 \leq p \leq 4$	2	$p = tq$	$p = q + 1$

All other Klein links have no torus equivalent.

We present an immediate (but important) corollary before the proof of Theorem 12.

Corollary 13. *Every Klein knot is equivalent to some torus knot.*

Proof. A Klein knot is a Klein link with one component. By Theorem 3, the only possible q values for a Klein knot are 1 and 2. Thus, the only Klein knots are of the forms $K_{p,1}$ and $K_{p,2}$. By Theorem 12, all such knots have a torus equivalent. □

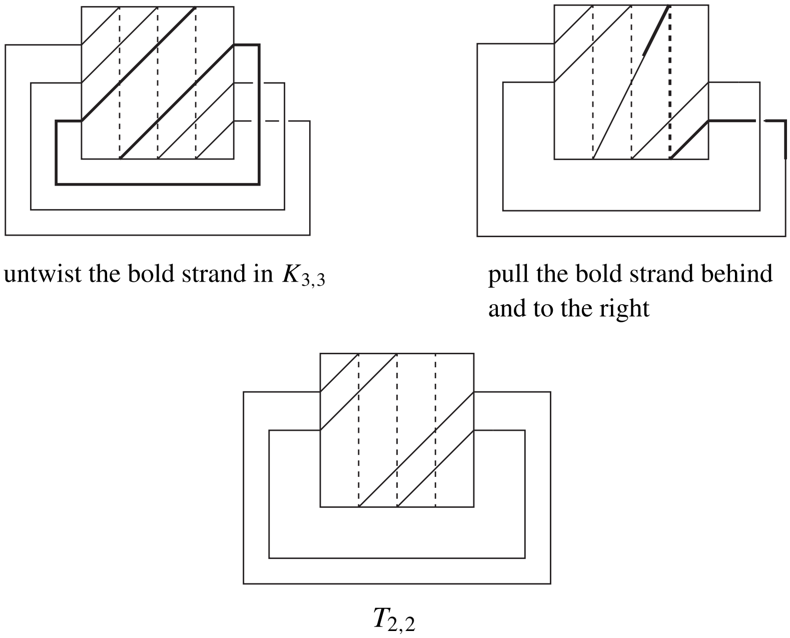
We emphasize that the corollary is a result about knots, not links. It is well-known and can be found in [Alvarado et al. 2016; Catalano et al. 2010; Freund and Smith-Polderman 2013].

Proof of Theorem 12. We first show that the Klein links listed in the table are, indeed, equivalent to some torus link.

Case 1 ($q = 0$): For each $p \geq 0$, by the way we construct Klein links, $K_{p,0}$ is a p-component unlink, hence equivalent to a torus link.

Case 2 ($q = 1$): For each $p \geq 0$, $K_{p,1}$ is an unknot by Lemma 7, hence equivalent to a torus link.

Case 3 ($q = 2$): One can see that $K_{0,2}$ is an unknot, and by [Alvarado et al. 2016, Theorem 5], for each $p \geq 1$, we know $K_{p,2} \equiv T_{p-1,2}$.



untwist the bold strand in $K_{3,3}$

pull the bold strand behind and to the right

$T_{2,2}$

Figure 11. $K_{3,3} \equiv T_{2,2}$.

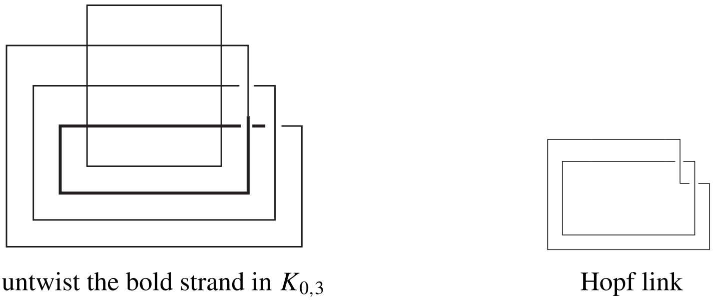
Case 4 ($q = 3$): Both $K_{1,3}$ and $K_{2,3}$ are 2-component unlinks, and thus are torus links. For $p = 3$, we can see in Figure 11 that $K_{3,3}$ is equivalent to $T_{2,2}$, which is a Hopf link.

With a little untwisting as shown in Figure 12 we see that $K_{0,3}$ is also equivalent to a Hopf link, and hence $T_{2,2}$.

Case 5 ($q = 4,\ p = 2$): By inspection, $K_{2,4}$ is a 2-component unlink.

Case 6 (q *even,* $p = tq,\ t \neq 0$): By Lemma 10, $K_{tq,q} \equiv T_{tq-q/2,q}$.

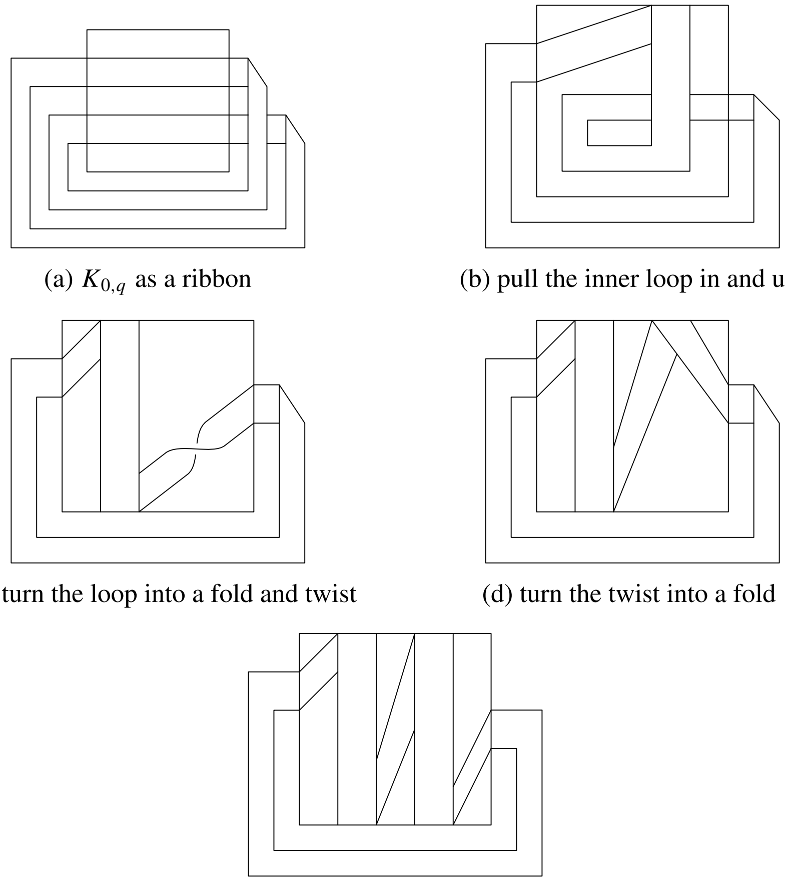
Case 7 ($t = n = 0,\ q \geq 4$ *and even*): Similar to the proof of Lemma 10, we collect together the strands through the first $\frac{1}{2}q$ nodes to form a ribbon. Using the wrapping



untwist the bold strand in $K_{0,3}$

Hopf link

Figure 12. $K_{0,3}$ is equivalent to a Hopf link.



(a) $K_{0,q}$ as a ribbon

(b) pull the inner loop in and up

(c) turn the loop into a fold and twist

(d) turn the twist into a fold

(e) reposition the right fold

Figure 13. $K_{0,q}$ with $q \geq 4$ even is a torus link.

function, these $\frac{1}{2}q$ strands wrap to the strands through nodes $\frac{1}{2}q+1$ through q so that we have a single ribbon that wraps twice. See Figure 13(a). Manipulate the ribbon as in Figure 13(b)–(e). Notice that the resulting ribbon in Figure 13(e) represents a torus link, though with the vertical wrapping opposite to the way we usually wrap.

Case 8 (*q odd, $p = q+1$*): Using [Alvarado et al. 2016, Theorem 4] and Lemma 10, we have $K_{q+1,q} \equiv K_{q+1,q+1} \equiv T_{q+1,(q+1)/2}$.

Our next step is to show that all other Klein links, those not listed in the table in Theorem 12, have no torus equivalence.

Case 9 (*$t = n = 0$, $q \geq 5$ and odd*): By Lemma 8, $K_{0,q}$ with $q \geq 5$ odd has pairs of components with different linking numbers, one pair with linking number 1 and another pair with linking number 2. Thus it cannot be equivalent to a torus link by Lemma 6.

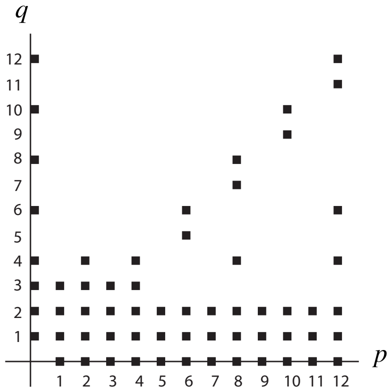


Figure 14. Graph showing which $K_{p,q}$ are torus links.

Case 10 ($t = 0$ *and either* $n = 1,\ q \geq 4$*; or* $n = 2,\ q \geq 5$*; or* $n \geq 3$): Since $t = 0$ and $n < q$, we can use [Alvarado et al. 2016, Theorem 3] to write $K_{p,q} = K_{n,q} \equiv K_{n,n} \cup K_{0,q-n}$, where $K_{n,n}$ and $K_{0,q-n}$ are unlinked. It follows that $K_{n,q}$ must have two components, one from $K_{n,n}$ and one from $K_{0,q-n}$, whose linking number is zero. Now, if $n = 1,\ q \geq 4$ or $n = 2,\ q \geq 5$, then $K_{0,q-n}$ has components with nonzero linking number by Lemma 8. On the other hand, if $n \geq 3$, then $K_{n,n}$ has components with nonzero linking number by Lemma 9. In both cases, $K_{n,q}$ must have components with nonzero linking number. Since it also has components with linking number zero, $K_{n,q}$ cannot be equivalent to a torus link by Lemma 6.

Case 11 ($t = 1$ *and either* $n = 0,\ q \geq 5$ *and odd; or* $n = 1,\ q \geq 4$ *and even*): We are looking at either $K_{q,q}$ with $q \geq 5$ and odd, or $K_{q+1,q} \equiv K_{q+1,q+1}$ with $q + 1 \geq 5$ and odd by [Alvarado et al. 2016, Theorem 4]. Thus, by [Alvarado et al. 2016, Theorem 7], neither can be equivalent to a torus link.

Case 12 (*either* $t = 1,\ n \geq 2$*; or* $t \geq 2,\ n = 0,\ q \geq 3$ *and odd; or* $t \geq 2,\ n = 1,\ q \geq 3$*; or* $t \geq 2,\ n \geq 2$*, n and q not both even; or* $t \geq 2,\ n \geq 2$*, n and q both even*): In each of these cases, by Theorem 4, $K_{p,q}$ contains either:

(1) $K_{2t+2,2}$ and at least one of $K_{t+1,1}$ or $K_{t,1}$ (with $t \geq 1$), or

(2) $K_{2t,2}$ and at least one of $K_{t+1,1}$ or $K_{t,1}$ (with $t \geq 2$), or

(3) $K_{2t+2,2}$ and $K_{2t,2}$ (with $t \geq 2$).

For each situation, $K_{p,q}$ contains components that are nonequivalent knots by Lemmas 7 and 11. Thus, $K_{p,q}$ has no torus equivalence by Lemma 6.

We leave it to the reader to check that all possible cases have been addressed. Figure 14, showing which Klein links have a torus equivalence, might help. □

Recall that every Klein knot is equivalent to a torus knot. From the sparseness of the graph in Figure 14, it is interesting to note that relatively few Klein links are equivalent to torus links. Thus, they warrant further study. For example, we plan to finish calculating the linking numbers for all Klein links (some further work has been done in [Bush et al. 2014]). Other link invariants could also be calculated. As noted in our construction, our Klein links are dependent on the relative position of the puncture on the Klein bottle. We need to investigate the effects on our results if we choose a different position for the puncture. On a more ambitious scale, we would like to determine a complete classification of Klein links, not just in terms of their relation to torus links.

References

[Adams 1994] C. C. Adams, *The knot book*, W. H. Freeman and Co., New York, 1994. MR Zbl

[Alvarado et al. 2016] E. Alvarado, S. Beres, V. Coufal, K. Hlavacek, J. Pereira, and B. Reeves, "Klein links and related torus links", *Involve* **9**:2 (2016), 347–359. MR Zbl

[Bush et al. 2014] M. A. Bush, K. R. French, and J. R. H. Smith, "Total linking numbers of torus links and Klein links", *Rose-Hulman Undergrad. Math J.* **15**:1 (2014), 73–92. MR

[Catalano et al. 2010] L. Catalano, D. Freund, R. Ruzvidzo, J. Bowen, and J. Ramsay, "A preliminary study of Klein knots", pp. 10–17 in *Proceedings of the Midstates Conference for Undergraduate Research in Computer Science and Mathematics at Wittenberg University*, 2010.

[Flapan 2016] E. Flapan, *Knots, molecules, and the universe: an introduction to topology*, American Mathematical Society, Providence, RI, 2016. MR Zbl

[Freund and Smith-Polderman 2013] D. Freund and S. Smith-Polderman, "Klein links and braids", *Rose-Hulman Undergrad. Math J.* **14**:1 (2013), 71–84. MR

[Livingston 1993] C. Livingston, *Knot theory*, Carus Mathematical Monographs **24**, Mathematical Association of America, Washington, DC, 1993. MR Zbl

[Murasugi 1996] K. Murasugi, *Knot theory and its applications*, Birkhäuser, Boston, 1996. MR Zbl

[Shepherd et al. 2012] D. Shepherd, J. Smith, S. Smith-Polderman, J. Bowen, and J. Ramsay, "The classification of a subset of Klein links", pp. 38–47 in *Proceedings of the Midstates Conference for Undergraduate Research in Computer Science and Mathematics at Ohio Wesleyan University*, 2012.

Received: 2016-11-01 Revised: 2017-08-09 Accepted: 2017-08-16

sberes@zagmail.gonzaga.edu *Gonzaga University, Spokane, WA, United States*

coufal@gonzaga.edu *Department of Mathematics, Gonzaga University, Spokane, WA, United States*

khlavacek@zagmail.gonzaga.edu *Department of Mathematics, Gonzaga University, Spokane, WA, United States*

kearney@gonzaga.edu *Department of Mathematics, Gonzaga University, Spokane, WA, United States*

rlattanzi@zagmail.gonzaga.edu *Gonzaga University, Spokane, WA, United States*

holson3@zagmail.gonzaga.edu *Gonzaga University, Spokane, WA, United States*

jp3465@drexel.edu *Department of Mathematics, Drexel University, Philadelphia, PA, United States*

bstrub@zagmail.gonzaga.edu *Gonzaga University, Spokane, WA, United States*

INVOLVE 11:4 (2018)
dx.doi.org/10.2140/involve.2018.11.625

Interpolation on Gauss hypergeometric functions with an application

Hina Manoj Arora and Swadesh Kumar Sahoo

(Communicated by Kenneth S. Berenhaut)

We use some standard numerical techniques to approximate the hypergeometric function

$$ {}_2F_1[a,b;c;x] = 1 + \frac{ab}{c}x + \frac{a(a+1)b(b+1)}{c(c+1)}\frac{x^2}{2!} + \cdots $$

for a range of parameter triples (a,b,c) on the interval $0 < x < 1$. Some of the familiar hypergeometric functional identities and asymptotic behavior of the hypergeometric function at $x = 1$ play crucial roles in deriving the formula for such approximations. We also focus on error analysis of the numerical approximations leading to monotone properties of quotients of gamma functions in parameter triples (a,b,c). Finally, an application to continued fractions of Gauss is discussed followed by concluding remarks consisting of recent works on related problems.

1. Introduction and preliminaries

For a complex number z and $c \neq 0, -1, -2, -3, \ldots$, the *hypergeometric series* is defined by

$$ 1 + \sum_{n=1}^{\infty} \frac{(a)_n(b)_n}{(c)_n(1)_n} z^n. $$

Here $(a)_n$ denotes the shifted factorial notation defined, in terms of the gamma function, by

$$ (a)_n = \frac{\Gamma(a+n)}{\Gamma(a)} = \begin{cases} a(a+1)\cdots(a+n-1) & \text{if } n \geq 1, \\ 1 & \text{if } n = 0,\ a \neq 0. \end{cases} $$

Note that the hypergeometric series defines an analytic function, denoted by the symbol ${}_2F_1[a,b;c;z]$, in $|z| < 1$. As quoted in the historical remarks in [Anderson et al. 1997, 1.55, p. 24], the concept of hypergeometric series was first introduced

MSC2010: primary 65D05; secondary 33B15, 33B20, 33C05, 33F05.
Keywords: interpolation, hypergeometric function, gamma function, error estimate.

by J. Wallis in 1656 to refer to a generalization of the geometric series. Less than a century later, Euler extensively studied the analytic properties of the hypergeometric function and found, for instance, its integral representation; see [Anderson et al. 1997, Theorem 1.19(2)]. Gauss made his first contribution to the subject in 1812. Due to the outstanding contribution made by Gauss to the field, the hypergeometric function is also sometimes known as the *Gauss hypergeometric function*. Most elementary functions which are solutions to certain differential equations can be written in terms of the Gauss hypergeometric functions. One can easily verify by using the Frobenius technique that the function ${}_2F_1[a,b;c;z]$ is one of the solutions of the *hypergeometric differential equation* [Andrews et al. 1999; Beals and Wong 2010; Rainville 1960]

$$z(1-z)w''+(c-(a+b+1)z)w'-abw=0.$$

We refer to [Rainville 1943; 1960] for Kummer's 24 solutions to the hypergeometric differential equation, and to [Beals and Wong 2010] for related applications. The asymptotic behavior of ${}_2F_1[a,b;c;z]$ near $z=1$ reveals that

$${}_2F_1[a,b;c;1]=\frac{\Gamma(c-a-b)\Gamma(c)}{\Gamma(c-a)\Gamma(c-b)}<\infty, \quad \text{valid for } \operatorname{Re}(c-a-b)>0. \tag{1-1}$$

Interpolating polynomials for elementary real functions such as trigonometric functions, logarithmic functions, exponential functions, etc. have already been derived in undergraduate texts in numerical analysis; see for instance [Atkinson 1978]. These elementary functions are in fact hypergeometric functions with specific parameters a,b,c; see for instance [Andrews et al. 1999; Rainville 1960]. Most of such polynomial approximations are computed when the functional values at the given boundary points are possible. Hence the asymptotic behavior (1-1) of the hypergeometric function near $z=1$ motivates us to construct interpolating polynomials for real hypergeometric functions ${}_2F_1[a,b;c;x]$, $a,b,c\in\mathbb{R}$, $c\notin\{0,-1,-2,-3,\dots\}$, of a real variable x using several numerical techniques in the interval $[0,1]$; however, the interval may be extended to $[-1,1]$ as the hypergeometric series in x is convergent for $|x|<1$ and it has a certain asymptotic behavior near -1 as well, with suitable choices of the parameters a,b,c; see for instance [Rainville 1960, Theorem 26]. More precisely, when we compute an interpolating polynomial $p_n(x)$ of a hypergeometric function ${}_2F_1[a,b;c;x]$ on $[0,1]$ we take the value ${}_2F_1[a,b;c;1]$ in the sense that the hypergeometric function defined at $x=1$ by means of its asymptotic behavior at $x=1$; see (1-1). Several hypergeometric functional identities also play a crucial role in determining functional values at the interpolating points.

The following lemmas are useful in describing the error analysis for the interpolating polynomials that we obtained in this paper. Our subsequent paper(s) in this series will cover the study of interpolating polynomials using other techniques.

Lemma 1.1 [Anderson et al. 1997, Lemma 1.33(1), p. 13; see also Lemma 1.35(2)]. *If* $a, b, c \in (0, \infty)$, *then* ${}_2F_1[a, b; c; x]$ *is strictly increasing on* $[0, 1)$. *In particular, if* $c > a + b$ *then for* $x \in [0, 1]$ *we have*

$$ {}_2F_1[a, b; c; x] \leq \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)}. $$

Lemma 1.2 [Anderson et al. 1997, Lemma 2.16(2), p. 36]. *The gamma function* $\Gamma(x)$ *is a log-convex function on* $(0, \infty)$. *In other words, the logarithmic derivative,* $\Gamma'(x)/\Gamma(x)$, *of the gamma function is increasing on* $(0, \infty)$.

Note that in all the plots in this paper, graphs drawn in blue represent the original functions and graphs drawn in red represent interpolating polynomials.

2. Linear interpolation on ${}_2F_1[a, b; c; x]$

For performing linear interpolation of the function ${}_2F_1[a, b; c; x] = f(x)$, we consider the end points $x_0 = 0$ and $x_1 = 1$ of the interval [0, 1]. The functional values at these points are respectively $f(0) = 1$ and $f(1)$, described in (1-1). Hence, the equation of the segment of the straight line joining 0 and 1 is

$$ P_l(x) = f(x_0) + \frac{x-x_0}{x_1-x_0}(f(x_1)-f(x_0)) = \frac{\Gamma(c)\Gamma(c-a-b)-\Gamma(c-a)\Gamma(c-b)}{\Gamma(c-a)\Gamma(c-b)}x+1, $$

when $c - a - b > 0$ and $c \neq 0, -1, -2, -3, \ldots$. The polynomial $P_l(x)$ represents the linear interpolation of ${}_2F_1[a, b; c; x]$ interpolating at 0 and 1.

Using Lemma 1.1, we obtain the following error estimate:

Lemma 2.1. *Let* $a, b, c \in (-2, \infty)$ *with* $c - a - b > 2$. *The deviation of the given function* $f(x) = {}_2F_1[a, b; c; x]$ *from the approximating function* $P_l(x)$ *for all values of* $x \in [0, 1]$ *is estimated by*

$$ |E_l(f, x)| = |f(x) - P_l(x)| \leq \frac{|a(a+1)b(b+1)|}{8} \frac{\Gamma(c)\Gamma(c-a-b-2)}{\Gamma(c-a)\Gamma(c-b)}. $$

Proof. Maximizing

$$ |E_l(f, x)| = \tfrac{1}{2}x(1-x)|f''(x)| $$

in [0, 1] yields

$$ \tfrac{1}{8}(1-0)^2 \max_{0 \leq x \leq 1} |f''(x)|, $$

where $f(x) = {}_2F_1[a, b; c; x]$. The following well-known derivative formula is useful:

$$ \frac{d}{dx}\,{}_2F_1[a, b; c; x] = \frac{ab}{c}\,{}_2F_1[a+1, b+1; c+1; x]. \tag{2-1} $$

The proof follows from (1-1), Lemma 1.1, (2-1), and the fact that

$$ \Gamma(x+1) = x\Gamma(x). $$ □

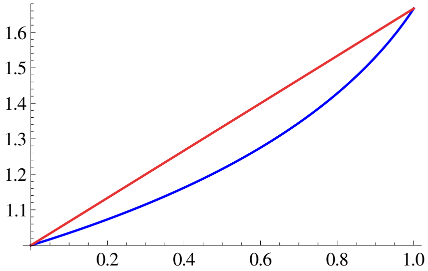


Figure 1. Linear interpolation of ${}_2F_1[1, 2; 6; x]$ at 0 and 1.

nodes x_i	0	0.25	0.5	0.75	1
actual values ${}_2F_1[1, 2; 6; x_i]$	1	1.0936	1.2149	1.3843	1.6667
polynomial approx. by $P_l(x_i)$	1	1.1667	1.3333	1.5000	1.6667
validity of error bounds by $E_l(f, x_i)$	0	$0.0731 < 1.25$	$0.1184 < 1.25$	$0.1157 < 1.25$	0

Table 1. Comparison of the functional and linear polynomial values.

Remark 2.2. It follows from Lemma 2.1 that there is no error for the choices $a = 0$, $a = -1$, $b = 0$, $b = -1$. In other words, for these choices $E_l(f, x)$ vanishes.

Figure 1 shows linear interpolation of the hypergeometric function at 0 and 1, whereas Table 1 compares the values of the hypergeometric function up to four decimal places with its interpolating polynomial values in the interval [0, 1] for the choice of parameters $a = 1$, $b = 2$ and $c = 6$. Figure 1 and Table 1 also indicate errors at various points within the unit interval except at the end points.

3. Quadratic interpolation on ${}_2F_1[a, b; c; x]$

Let the three points in consideration for quadratic interpolation be $x_0 = 0$, $x_1 = 0.5$ and $x_2 = 1$. The functional values at $x_0 = 0$ and $x_2 = 1$ can be found easily in terms of the parameters but the functional value at $x_1 = 0.5$ can be obtained through different identities involving hypergeometric functions ${}_2F_1[a, b; c; x]$ dealing with various constraints on the parameters a, b, c. This section consists of two subsections and in each subsection the method to obtain the functional value of ${}_2F_1[a, b; c; x]$ at $x = 0.5$ uses three different identities. Finally, we compare the resultant interpolations. In fact we observe that the interpolating polynomial remains unchanged in two cases, although the approaches are different (see the subsection on page 630 for more details).

Quadratic interpolation on ${}_2F_1[a, 1-a; c; x]$. This section deals with the value ${}_2F_1\left[a, b; c; \frac{1}{2}\right]$, where $a + b = 1$ due to the following identity of Bailey [1935,

p. 11] (see also [Rainville 1960, p. 69]):

$$
\begin{aligned}
&{}_2F_1\bigl[a,\,1-a;\,c;\,\tfrac{1}{2}\bigr]\\
&\qquad=\frac{2^{1-c}\,\Gamma(c)\,\Gamma\bigl(\frac{1}{2}\bigr)}{\Gamma\bigl(\frac{1}{2}(c+a)\bigr)\,\Gamma\bigl(\frac{1}{2}(1+c-a)\bigr)}=\frac{\Gamma\bigl(\frac{1}{2}c\bigr),\,\Gamma\bigl(\frac{1}{2}(1+c)\bigr)}{\Gamma\bigl(\frac{1}{2}(c+a)\bigr)\,\Gamma\bigl(\frac{1}{2}(1+c-a)\bigr)},
\end{aligned}
\tag{3-1}
$$

where c is a positive integer. It follows from (3-1) that

$$
\Gamma\bigl(\tfrac{1}{2}c\bigr)\Gamma\bigl(\tfrac{1}{2}(1+c)\bigr)=2^{1-c}\sqrt{\pi}\,\Gamma(c),
\tag{3-2}
$$

since $\Gamma\bigl(\frac{1}{2}\bigr)=\sqrt{\pi}$. In this case, we obtain

$$
\begin{aligned}
f(x_0)&=f(0)={}_2F_1[a,\,1-a;\,c;\,0]=1,\\
f(x_1)&=f(0.5)={}_2F_1\bigl[a,\,1-a;\,c;\,\tfrac{1}{2}\bigr]=\frac{\Gamma\bigl(\frac{1}{2}c\bigr)\,\Gamma\bigl(\frac{1}{2}(1+c)\bigr)}{\Gamma\bigl(\frac{1}{2}(c+a)\bigr)\,\Gamma\bigl(\frac{1}{2}(1+c-a)\bigr)},\\
f(x_2)&=f(1)={}_2F_1[a,\,1-a;\,c;\,1]=\frac{\Gamma(c)\Gamma(c-1)}{\Gamma(c-a)\Gamma(c+a-1)}\quad(c>1).
\end{aligned}
$$

Consider the well-known Lagrange fundamental polynomials

$$
\begin{aligned}
L_0(x)&=\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)},\\
L_1(x)&=\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)},\\
L_2(x)&=\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}.
\end{aligned}
$$

Then the quadratic interpolation of $f(x)={}_2F_1[a,\,1-a;\,c;\,x]$ becomes

$$
\begin{aligned}
P_{q_3}(x)&=f(x_0)L_0(x)+f(x_1)L_1(x)+f(x_2)L_2(x)\\
&=(2x^2-3x+1)+(-4x^2+4x)\frac{\Gamma\bigl(\frac{1}{2}c\bigr)\,\Gamma\bigl(\frac{1}{2}(1+c)\bigr)}{\Gamma\bigl(\frac{1}{2}(c+a)\bigr)\,\Gamma\bigl(\frac{1}{2}(1+c-a)\bigr)}\\
&\qquad\qquad+(2x^2-x)\frac{\Gamma(c)\Gamma(c-1)}{\Gamma(c-a)\Gamma(c+a-1)}.
\end{aligned}
$$

This leads to the following result.

Theorem 3.1. *Let $a, b, c\in\mathbb{R}$ be such that $c>1$. Then*

$$
\begin{aligned}
P_{q_1}(x)&=\biggl(2-\frac{4\Gamma\bigl(\frac{1}{2}c\bigr)\Gamma\bigl(\frac{1}{2}(1+c)\bigr)}{\Gamma\bigl(\frac{1}{2}(c+a)\bigr)\Gamma\bigl(\frac{1}{2}(1+c-a)\bigr)}+\frac{2\Gamma(c)\Gamma(c-1)}{\Gamma(c-a)\Gamma(c+a-1)}\biggr)x^2\\
&\qquad+\biggl(\frac{4\Gamma\bigl(\frac{1}{2}c\bigr)\Gamma\bigl(\frac{1}{2}(1+c)\bigr)}{\Gamma\bigl(\frac{1}{2}(c+a)\bigr)\Gamma\bigl(\frac{1}{2}(1+c-a)\bigr)}-\frac{\Gamma(c)\Gamma(c-1)}{\Gamma(c-a)\Gamma(c+a-1)}-3\biggr)x+1.
\end{aligned}
$$

is a quadratic interpolation of ${}_2F_1[a,\,1-a;\,c;\,x]$ in $[0, 1]$.

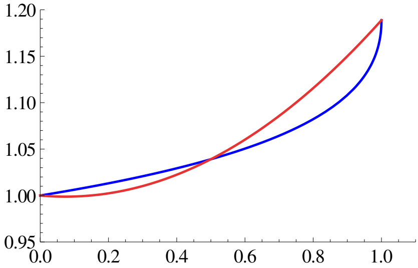


Figure 2. The quadratic interpolation of ${}_2F_1[0.9, 0.1; 1.5; x]$ at $0, 0.5$, and 1.

Remark 3.2. It is evident that when $a = 0, 1$, then $P_{q_1}(x) = {}_2F_1[a, 1-a; c; x] = 1$ for all $x \in [0, 1]$ and for all $c > 1$. Moreover, for all $c > 1$, we have the following three natural observations:

(i) If $-1 < a < 0$, then $P_{q_1}(x)$ and ${}_2F_1[a, 1-a; c; x]$ decrease together in $[0, 1]$.

(ii) If $0 < a < 1$, then $P_{q_1}(x)$ and ${}_2F_1[a, 1-a; c; x]$ increase together in $[0, 1]$.

(iii) If $1 < a < 2$, then $P_{q_1}(x)$ and ${}_2F_1[a, 1-a; c; x]$ decrease together in $[0, 1]$.

Indeed, these follow from derivative test. More observations are stated later while estimating the error (see Remark 3.10).

An interpolating polynomial $P_{q_1}(x)$ of ${}_2F_1[a, 1-a; c; x]$ for certain choices of parameters a and c is as shown in Figure 2.

Remark 3.3. Note that in Theorem 3.1, the parameter c cannot be chosen such that $c \le \frac{1}{2}(a+b+1)$ since the choice $b = 1-a$ results in $c \le 1$, which is a contradiction to the assumption that $c > 1$. In particular, $c \ne \frac{1}{2}(a+b+1)$ in Theorem 3.1, which is the negation of a constraint that will be considered in the next subsection.

Quadratic interpolation on ${}_2F_1\big[a, b; \frac{1}{2}(a+b+1); x\big]$. In this section, $f(x) = {}_2F_1[a, b; c; x]$, $c = \frac{1}{2}(a+b+1)$, is first interpolated using the following quadratic transformation obtained from [Andrews et al. 1999, (3.1.3)]; see also [Rainville 1960, Theorem 2.5].

Lemma 3.4. *If $\frac{1}{2}(a+b+1)$ is a positive integer, and if $|x| < 1$ and $|4x(1-x)| < 1$, then*

$$ {}_2F_1\big[a, b; \tfrac{1}{2}(a+b+1); x\big] = {}_2F_1\big[\tfrac{1}{2}a, \tfrac{1}{2}b; \tfrac{1}{2}(a+b+1); 4x(1-x)\big]. \tag{3-3}$$

If we choose $x = 0.5$ then the right-hand side of (3-3) computes the asymptotic behavior of the hypergeometric function at 1. Hence the functional value at $x = 0.5$ of the function $f(x) = {}_2F_1\big[a, b; \frac{1}{2}(a+b+1); x\big]$ can be obtained with the help of (1-1). Due to Lemma 3.4 and (1-1), in this case, the constraints on the parameters are computed as

- $a+b<1$;
- $a+b\neq-(2n+1)$ for $n\in\mathbb{N}\cup\{0\}$.

One can easily obtain that

$$f(x_0)={}_2F_1\big[a,b;\tfrac{1}{2}(a+b+1);0\big]=1;$$

$$f(x_1)={}_2F_1\big[a,b;\tfrac{1}{2}(a+b+1);\tfrac{1}{2}\big]=\frac{\sqrt{\pi}\,\Gamma\big(\frac{1}{2}(a+b+1)\big)}{\Gamma\big(\frac{1}{2}(a+1)\big)\Gamma\big(\frac{1}{2}(b+1)\big)},$$

$$f(x_2)={}_2F_1\big[a,b;\tfrac{1}{2}(a+b+1);1\big]=\frac{\Gamma\big(\frac{1}{2}(1-a-b)\big)\Gamma\big(\frac{1}{2}(a+b+1)\big)}{\Gamma\big(\frac{1}{2}(a+1-b)\big)\Gamma\big(\frac{1}{2}(b+1-a)\big)}=\frac{\cos\big(\frac{\pi}{2}(b-a)\big)}{\cos\big(\frac{\pi}{2}(b+a)\big)},$$

where $f(x_2)$ is obtained by the well-known Euler's reflection formula (in nonintegral variable x)

$$\Gamma(x)\Gamma(1-x)=\frac{\pi}{\sin(\pi x)}.$$

This leads to the additional constraints on the parameters

$$\begin{aligned} &a+b\neq 1\pm 2n \quad \text{and} \quad a-b\neq -1\pm 2n, \quad n\in\mathbb{Z} \quad \text{or}\\ &a+b\neq -1\pm 2n \quad \text{and} \quad a-b\neq 1\pm 2n, \quad n\in\mathbb{Z}. \end{aligned} \tag{3-4}$$

(These constraints may be relaxed when one does not use Euler's reflection formula!)

Thus, the first quadratic interpolation of $f(x)={}_2F_1\big[a,b;\frac{1}{2}(a+b+1);x\big]$ becomes

$$\begin{aligned} P_{q2}(x)&=f(x_0)L_0(x)+f(x_1)L_1(x)+f(x_2)L_2(x)\\ &=(2x^2-3x+1)+(-4x^2+4x)\frac{\sqrt{\pi}\,\Gamma\big(\frac{1}{2}(a+b+1)\big)}{\Gamma\big(\frac{1}{2}(a+1)\big)\Gamma\big(\frac{1}{2}(b+1)\big)}\\ &\qquad\qquad+(2x^2-x)\frac{\cos\big(\frac{\pi}{2}(b-a)\big)}{\cos\big(\frac{\pi}{2}(b+a)\big)}. \end{aligned}$$

This leads to the following result.

Theorem 3.5. *Let $a,b\in\mathbb{R}$ and $n\in\mathbb{N}\cup\{0\}$ be such that $a+b\neq-(2n+1)$ and $a+b<1$. If either $a+b\neq 1\pm 2n$ and $a-b\neq -1\pm 2n$, or $a+b\neq -1\pm 2n$ and $a-b\neq 1\pm 2n$ hold, then*

$$\begin{aligned} P_{q2}(x)=&\left(2-\frac{4\sqrt{\pi}\,\Gamma\big(\frac{1}{2}(a+b+1)\big)}{\Gamma\big(\frac{1}{2}(a+1)\big)\Gamma\big(\frac{1}{2}(b+1)\big)}+\frac{2\cos\big(\frac{\pi}{2}(b-a)\big)}{\cos\big(\frac{\pi}{2}(b+a)\big)}\right)x^2\\ &+\left(\frac{4\sqrt{\pi}\,\Gamma\big(\frac{1}{2}(a+b+1)\big)}{\Gamma\big(\frac{1}{2}(a+1)\big)\Gamma\big(\frac{1}{2}(b+1)\big)}-\frac{\cos\big(\frac{\pi}{2}(b-a)\big)}{\cos\big(\frac{\pi}{2}(b+a)\big)}-3\right)x+1 \end{aligned}$$

is a quadratic interpolation of ${}_2F_1\big[a,b;\frac{1}{2}(a+b+1);x\big]$ in $[0,1]$.

Secondly, we discuss quadratic interpolation of the same function ${}_2F_1[a, b; c; x]$, $c = \frac{1}{2}(a+b+1)$, in [0, 1], but using a different hypergeometric identity. Finally, we observe that both the interpolations are same except at a minor difference in one of the constraints.

Recall the transformation formula [Rainville 1960, Theorem 20, p. 60]:

Lemma 3.6. *If* $|x| < 1$ *and* $|x/(1-x)| < 1$, *then we have*

$$ {}_2F_1[a, b; c; x] = (1-x)^{-a}{}_2F_1\left[a, c-b; c; \frac{-x}{1-x}\right]. $$

Note that $-x/(1-x) = -1$ for $x = 0.5$. This suggests that to find the value $f(0.5) = 2^a{}_2F_1[a, c-b; c; -1]$ we can use the following identity [Rainville 1960, Theorem 26, p. 68]; see also [Beals and Wong 2010].

Lemma 3.7. *Let* $a', b' \in \mathbb{R}$. *If* $1 + a' - b' \neq \{0, -1, -2, -3, \dots\}$ *and* $b' < 1$, *then we have*

$$ {}_2F_1[a', b'; a'-b'+1; -1] = \frac{\Gamma(a'-b'+1)\Gamma\left(\frac{1}{2}a'+1\right)}{\Gamma(a'+1)\Gamma\left(\frac{1}{2}a'-b'+1\right)}. $$

Comparison of the parameters $a' = a,\ b' = c - b$ and $a' - b' + 1 = c$ leads to

$$ {}_2F_1[a, c-b; c; -1] = \frac{\Gamma(a-c+b+1)\Gamma\left(\frac{1}{2}a+1\right)}{\Gamma(a+1)\Gamma\left(\frac{1}{2}a-c+b+1\right)} \tag{3-5} $$

with the constraints

- $2c = a + b + 1$;
- $c \neq \{0, -1, -2, -3, \dots\} \iff a + b \neq -(2n+1),\ n \in \mathbb{N} \cup \{0\}$;
- $c - b < 1 \iff a - b < 1$.

Under these conditions, (3-5) leads to

$$ f(x_1) = f(0.5) = {}_2F_1\left[a, b; \tfrac{1}{2}(a+b+1); \tfrac{1}{2}\right] = 2^a \frac{\Gamma\left(\frac{1}{2}(a+b+1)\right)\Gamma\left(\frac{1}{2}a+1\right)}{\Gamma(a+1)\Gamma\left(\frac{1}{2}(b+1)\right)} $$
$$ = \frac{2^{a-1}\,\Gamma\left(\frac{1}{2}(a+b+1)\right)\,\Gamma\left(\frac{1}{2}a\right)}{\Gamma(a)\,\Gamma\left(\frac{1}{2}(b+1)\right)} = \frac{\sqrt{\pi}\,\Gamma\left(\frac{1}{2}(a+b+1)\right)}{\Gamma\left(\frac{1}{2}(a+1)\right)\Gamma\left(\frac{1}{2}(b+1)\right)}, $$

where the last equality holds by (3-2). Also as discussed at the beginning of this subsection, we have

$$ f(x_0) = f(0) = {}_2F_1\left[a, b; \tfrac{1}{2}(a+b+1); 0\right] = 1, $$
$$ f(x_2) = f(1) = {}_2F_1\left[a, b; \tfrac{1}{2}(a+b+1); 1\right] = \frac{\cos\left(\frac{\pi}{2}(b-a)\right)}{\cos\left(\frac{\pi}{2}(b+a)\right)}, \quad a+b<1, $$

with additional constraints obtained in (3-4) (here also (3-4) may be relaxed!).

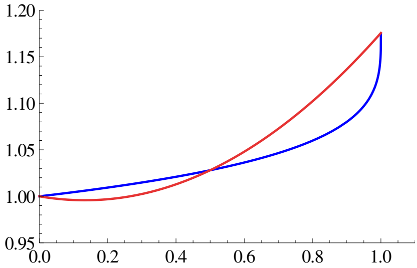


Figure 3. The quadratic interpolation of ${}_2F_1[0.1, 0.3; 0.7; x]$ at 0, 0.5, and 1.

Thus, the second quadratic interpolation of $f(x) = {}_2F_1\big[a, b; \frac{1}{2}(a+b+1); x\big]$ remains same as the first quadratic interpolation obtained in Theorem 3.5 but with an additional constraint $a - b < 1$. This shows that the quadratic interpolation obtained by Theorem 3.5 is stronger than what was discussed so far using Lemmas 3.6 and 3.7. A quadratic interpolation of ${}_2F_1\big[a, b; \frac{1}{2}(a+b+1); x\big]$ is shown in Figure 3.

Error estimates. The error estimate in quadratic interpolation of ${}_2F_1[a, b; c; x]$ interpolating at 0, 0.5, 1 in [0, 1] is formulated as below:

Lemma 3.8. *Let $P_q(x)$ be a quadratic interpolation of $f(x) = {}_2F_1[a, b; c; x]$ interpolating at* 0, 0.5, 1 *in* [0, 1]. *If $a, b, c \in (-3, \infty)$ with $c - a - b > 3$, then the deviation of $f(x)$ from $P_q(x)$ is estimated by*

$$|E_q(f, x)| = |f(x) - P_q(x)| \\ \le \tfrac{1}{6}M\big|a(a+1)(a+2)b(b+1)(b+2)\big|\,\frac{\Gamma(c)\Gamma(c-a-b-3)}{\Gamma(c-a)\Gamma(c-b)}$$

for all values of $x \in [0, 1]$, where M is defined by

$$M := \begin{cases} \frac{1}{12}(3-\sqrt{3})\big(-1+\frac{1}{6}(3-\sqrt{3})\big)\big(-1+\frac{1}{3}(3-\sqrt{3})\big), & x < \frac{1}{2}, \\ -\frac{1}{12}(3+\sqrt{3})\big(-1+\frac{1}{6}(3+\sqrt{3})\big)\big(-1+\frac{1}{3}(3+\sqrt{3})\big), & x > \frac{1}{2}. \end{cases} \tag{3-6}$$

Proof. We need to estimate

$$\max_{0\le x\le 1} \tfrac{1}{6}\big|x(x-0.5)(x-1)\big| \max_{0\le x\le 1} |f'''(x)|,$$

where $f(x) = {}_2F_1[a, b; c; x]$. Note that

$$\max_{0\le x\le 1} |x(x-0.5)(x-1)| = M \ (\approx 0.0481125\ldots)$$

by (3-6). We apply the well known derivative formula (2-1) to maximize $|f'''(x)|$, $0 \le x \le 1$. The proof follows from (1-1), Lemma 1.1, (2-1), and the fact that

$$\Gamma(x+1) = x\Gamma(x). \qquad \square$$

The following result is an immediate consequence of Lemma 3.8 which estimates the difference $E_{q_1}(f,x) = {}_2F_1[a, 1-a; c; x] - P_{q_1}(x)$ in $[0, 1]$.

Corollary 3.9. *Let $a, c \in \mathbb{R}$ be such that $-3 < a < 4$ and $c > 4$. Then the deviation of ${}_2F_1[a, 1-a; c; x]$ from $P_{q_1}(x)$ is estimated by*

$$\begin{aligned}|E_{q_1}(f,x)| &= |f(x) - P_{q_1}(x)| \\ &\le \tfrac{1}{6}M\big|a(a+1)(a+2)(1-a)(2-a)(3-a)\big|\,\frac{\Gamma(c)\Gamma(c-4)}{\Gamma(c-a)\Gamma(c+a-1)}\end{aligned}$$

for all values of $x \in [0, 1]$, where M is obtained by (3-6).

Remark 3.10. It follows from Corollary 3.9 that there is no error for any of the choices $a = -2, -1, 0, 1, 2, 3$. In other words, for any of these choices, $E_{q_1}(f,x)$ vanishes.

Similarly, as a consequence of Lemma 3.8, we obtain:

Corollary 3.11. *Let $a, b \in \mathbb{R}$ be such that $-7 < a + b < -5$. Then the deviation of ${}_2F_1\big[a, b; \tfrac{1}{2}(a+b+1); x\big]$ from $P_{q_2}(x)$ is estimated by*

$$\begin{aligned}|E_{q_2}(f,x)| &= |f(x) - P_{q_2}(x)| \\ &\le \tfrac{1}{6}M\big|a(a+1)(a+2)b(b+1)(b+2)\big|\,\frac{\Gamma\big(\frac{1}{2}(a+b+1)\big)\Gamma\big(\frac{1}{2}(-a-b-5)\big)}{\Gamma\big(\frac{1}{2}(b-a+1)\big)\Gamma\big(\frac{1}{2}(a-b+1)\big)}\end{aligned}$$

for all values of $x \in [0, 1]$, where M is obtained by (3-6).

Remark 3.12. It follows from Corollary 3.11 that since $E_{q_2}(f,x)$ vanishes for the choices $a = -2, -1, 0$ and $b = -2, -1, 0$, there is no error for these choices of the parameters a and b.

Now we describe a slightly deeper analysis on the error obtained in Corollary 3.9 through the following lemma, which is a consequence of Lemma 1.2. A similar analysis can be described for Corollary 3.11.

Lemma 3.13. *Let $a, c \in \mathbb{R}$ be such that $c > 4$. If either $1 < a < 4$ or $-3 < a < 0$ holds, then the quotient*

$$\frac{\Gamma(c)\Gamma(c-4)}{\Gamma(c-a)\Gamma(c+a-1)}$$

decreases when c increases.

Proof. We use Lemma 1.2. Since $c - a > c - 4 > 0$, on one hand we have

$$\frac{\Gamma'(c-4)}{\Gamma(c-4)} - \frac{\Gamma'(c-a)}{\Gamma(c-a)} < 0.$$

On the other hand, since $c < c + a - 1$, we have

$$\frac{\Gamma'(c)}{\Gamma(c)} - \frac{\Gamma'(c+a-1)}{\Gamma(c+a-1)} < 0.$$

Thus, if

$$g(c)=\frac{\Gamma(c)\Gamma(c-4)}{\Gamma(c-a)\Gamma(c+a-1)},$$

it follows that

$$\begin{aligned}\frac{g'(c)}{g(c)}&=\frac{\Gamma'(c)}{\Gamma(c)}+\frac{\Gamma'(c-4)}{\Gamma(c-4)}-\frac{\Gamma'(c-a)}{\Gamma(c-a)}-\frac{\Gamma'(c+a-1)}{\Gamma(c+a-1)}\\&=\left(\frac{\Gamma'(c-4)}{\Gamma(c-4)}-\frac{\Gamma'(c-a)}{\Gamma(c-a)}\right)+\left(\frac{\Gamma'(c)}{\Gamma(c)}-\frac{\Gamma'(c+a-1)}{\Gamma(c+a-1)}\right)<0.\end{aligned}$$

By the definition of the gamma function, obviously, one can see that $\Gamma(x)>0$ for $x>0$. This shows that $g(c)>0$ and hence $g'(c)<0$. Thus, $g(c)$ decreases for $1<a<4<c$.

For $c>4$, if $-3<a<0$ holds then we consider the rearrangement

$$\frac{g'(c)}{g(c)}=\left(\frac{\Gamma'(c)}{\Gamma(c)}-\frac{\Gamma'(c-a)}{\Gamma(c-a)}\right)+\left(\frac{\Gamma'(c-4)}{\Gamma(c-4)}-\frac{\Gamma'(c+a-1)}{\Gamma(c+a-1)}\right)$$

and show that $g'(c)/g(c)<0$. □

Using Mathematica or other similar tools, one can see that Lemma 3.13 even holds true for the remaining range $0\le a\le 1$. This suggests the following conjecture.

Conjecture 3.14. *Let $a, c\in\mathbb{R}$ be such that $0\le a\le 1$ and $c>4$. Then the quotient*

$$\frac{\Gamma(c)\Gamma(c-4)}{\Gamma(c-a)\Gamma(c+a-1)}$$

decreases when c increases.

Thus, we observe that when $c>4$ increases then the error $E_{q_1}(f,x)$ estimated in Corollary 3.9 decreases (see also Figures 4 and 5).

Figures 4 and 5 describe the quadratic interpolation of the hypergeometric functions ${}_2F_1[a,1-a,c,x]$ at 0, 0.5 and 1, whereas Tables 2 and 3 compare the values of the hypergeometric function up to four decimal places with its interpolating

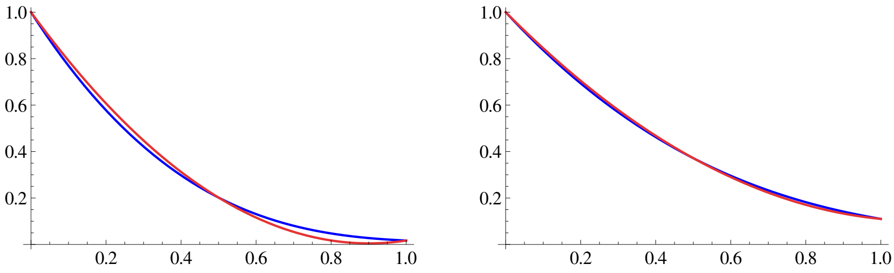


Figure 4. The error estimate $E_{q_1}(f,x)$ when $a=3.9$ and c increases from 4.5 to 6.5.

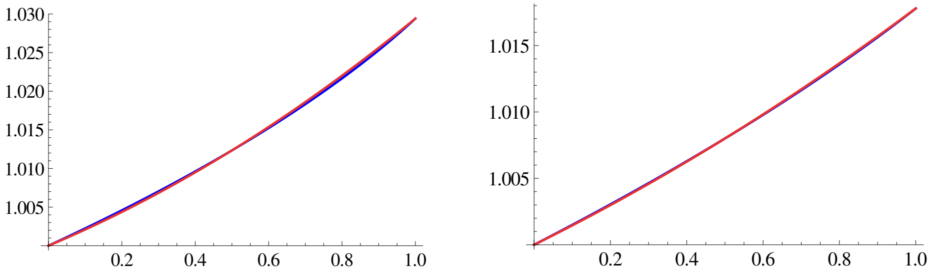


Figure 5. The error estimate $E_{q_1}(f,x)$ when $a = 0.9$ and c increases from 4.1 to 6.1.

nodes x_i	0	0.25	0.5	0.75	1
actual values ${}_2F_1[3.9,-2.9;5;x_i]$	1	0.5372	0.2516	0.0998	0.0367
polynomial approx. by $P_{q_1}(x_i)$	1	0.5591	0.2516	0.0775	0.0367
validity of error bounds by $E_{q_1}(f,x_i)$	0	$0.0219 < 0.0274$	0	$0.0223 < 0.0274$	0
actual values ${}_2F_1[3.9,-2.9;6;x_i]$	1	0.6027	0.3358	0.1724	0.0845
polynomial approx. by $P_{q_1}(x_i)$	1	0.6163	0.3358	0.1585	0.0845
validity of error bounds by $E_{q_1}(f,x_i)$	0	$0.0136 < 0.0158$	0	$0.0139 < 0.0158$	0

Table 2. Comparison of the functional and quadratic polynomial values.

nodes x_i	0	0.25	0.5	0.75	1
actual values ${}_2F_1[0.9,0.1;5;x_i]$	1	1.0047	1.0099	1.0158	1.0227
polynomial approx. by $P_{q_1}(x_i)$	1	1.0046	1.0099	1.0160	1.0227
validity of error bounds by $E_{q_1}(f,x_i)$	0	$0.0001 < 0.0016$	0	$0.0002 < 0.0016$	0
actual values ${}_2F_1[0.9,0.1;6;x_i]$	1	1.0039	1.0082	1.0128	1.0182
polynomial approx. by $P_{q_1}(x_i)$	1	1.0038	1.0082	1.0129	1.0182
validity of error bounds by $E_{q_1}(f,x_i)$	0	$0.0001 < 0.0004$	0	$0.0001 < 0.0004$	0

Table 3. Comparison of the functional and quadratic polynomial values.

polynomial values in the interval [0, 1] for the choice of parameters $a = 3.9$, $c = 5$ and $a = 0.9$, $c = 6$ respectively. Figures 4 and 5 and Tables 2 and 3 also indicate errors at various points within the unit interval except at the interpolating points at $x = 0, 0.5, 1$.

The error estimate $|E_{q_2}(f, x)|$ for the function ${}_2F_1\big[a, b; \frac{1}{2}(a+b+1); x\big]$ can be analyzed in a similar way, and hence we omit the proof.

4. An application

In this section, we briefly consider interpolation of a continued fraction that converges to a quotient of two hypergeometric functions. Gauss used the contiguous relations to give several ways to write a quotient of two hypergeometric functions as a continued fraction. For instance, it is well known that

$$\frac{{}_2F_1[a+1, b; c+1; x]}{{}_2F_1[a, b; c; x]} = \cfrac{1}{1+\cfrac{\frac{(a-c)b}{c(c+1)}x}{1+\cfrac{\frac{(b-c-1)(a+1)}{(c+1)(c+2)}x}{1+\cfrac{\frac{(a-c-1)(b+1)}{(c+2)(c+3)}x}{1+\cfrac{\frac{(b-c-2)(a+2)}{(c+3)(c+4)}x}{1+\ddots}}}}}, \quad |x| < 1. \tag{4-1}$$

On one hand, if we adopt the basic linear interpolation method that we discussed in Section 2 (that is, linear interpolation directly) to the function

$$g(x) = \frac{{}_2F_1[a+1, b; c+1; x]}{{}_2F_1[a, b; c; x]}$$

at $x_0 = 0$ and $x_1 = 1$, we obtain the linear interpolation of the above continued fraction in the form

$$R_l(x) = g(x_0) + \frac{x - x_0}{x_1 - x_0}(g(x) - g(x_0)) = 1 + \left(\frac{b}{c-b}\right)x, \quad c - b > a,$$

since $g(x_0) = 1$ and $g(x_1) = c/(c-b)$. For the choice $a = 1,\ b = 2,\ c = 6$, this approximation is also shown in Figure 6.

On the other hand, an application of linear interpolation of ${}_2F_1[a, b; c; x]$ obtained in Section 2 leads to the following approximation of the above continued

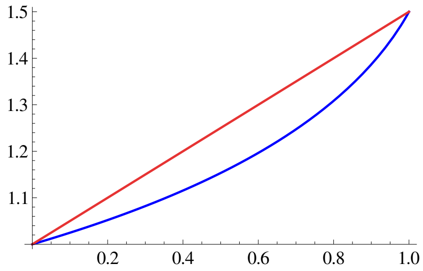


Figure 6. Approximation of ${}_2F_1[a{+}1,b;c{+}1;x]/{}_2F_1[a,b;c;x]$ through $R_l(x)$.

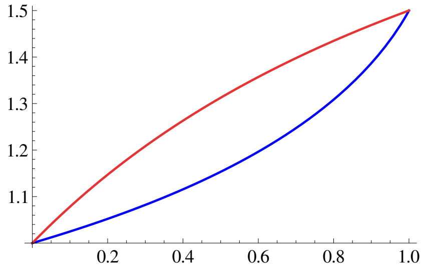


Figure 7. Approximation of ${}_2F_1[a{+}1,b;c{+}1;x]/{}_2F_1[a,b;c;x]$ through $R_r(x)$.

fraction in terms of ratio of polynomial approximation (we call this *rational interpolation*):

$$\begin{aligned}R_r(x) &= \frac{1}{P_l(x)}\left(\frac{\Gamma(c+1)\Gamma(c-a-b)-\Gamma(c-a)\Gamma(c-b+1)}{\Gamma(c-a)\Gamma(c-b+1)}x+1\right)\\ &= \frac{\big[c\Gamma(c)\Gamma(c-a-b)/(c-b)-\Gamma(c-a)\Gamma(c-b)\big]x+\Gamma(c-a)\Gamma(c-b)}{\big[\Gamma(c)\Gamma(c-a-b)-\Gamma(c-a)\Gamma(c-b)\big]x+\Gamma(c-a)\Gamma(c-b)}\\ &= 1+\frac{b}{c-b}\left[\frac{\Gamma(c-a-b)\Gamma(c)\,x}{\big[\Gamma(c)\Gamma(c-a-b)-\Gamma(c-a)\Gamma(c-b)\big]x+\Gamma(c-a)\Gamma(c-b)}\right],\end{aligned}$$

where $c-a-b>0$. For the choice $a=1,\ b=2,\ c=6$, this approximation is also shown in Figure 7.

Observe that

$$R_r(x_0)=1=R_l(x_0) \quad \text{and} \quad R_r(x_1)=\frac{c}{c-b}=R_l(x_1)$$

and hence R_r also interpolates the continued fraction under consideration at 0 and 1. Further we observe that both the approximations $R_l(x)$ and $R_r(x)$ of the continued fraction are easy to obtain and the first approximation, i.e., $R_l(x)$, is in a simpler form than $R_r(x)$, as expected. Now, it would be interesting to know which one would give the best approximation to the continued fraction under consideration. With the special choice $a=1,\ b=2,\ c=6$, we see from Figures 6 and 7 that $R_l(x)$ is a better approximation than $R_r(x)$. One may ask: does it happen for arbitrary parameters a, b, c? Since $R_l(x) = R_r(x)$ if and only if $\Gamma(c)\Gamma(c-a-b) = \Gamma(c-a)\Gamma(c-b)$, the answer to this question is yes except when $\Gamma(c)\Gamma(c-a-b) = \Gamma(c-a)\Gamma(c-b)$.

This leads to the following result:

Theorem 4.1. *Let $R_l(x)$ and $R_r(x)$ be respectively the linear interpolation and the rational interpolation of the quotient ${}_2F_1[a+1, b; c+1; x]/{}_2F_1[a, b; c; x]$ (equivalently, of the continued fraction (4-1)). Then $R_l(x)$ and $R_r(x)$ coincide with each other if and only if $\Gamma(c)\Gamma(c-a-b) = \Gamma(c-a)\Gamma(c-b)$ holds for $c-a-b > 0$.*

5. Concluding remarks and future scope

Recall that, in this paper, we use some standard interpolation techniques to approximate the hypergeometric function

$$
{}_2F_1[a, b; c; x] = 1 + \frac{ab}{c}x + \frac{a(a+1)b(b+1)}{c(c+1)}\frac{x^2}{2!} + \cdots
$$

for a range of parameter triples (a, b, c) on the interval $0 < x < 1$. Some of the familiar hypergeometric functional identities and asymptotic behavior of the hypergeometric function at $x = 1$ played crucial roles in deriving the formula for such approximations. One can expect similar formulae using other well-known interpolations and obtain better approximations for the hypergeometric function; however, we discuss such results in an upcoming manuscript(s). Different numerical methods for the computation of the confluent and Gauss hypergeometric functions were studied recently in [Pearson et al. 2017]. Such investigation may be extended to the q-analog of the hypergeometric functions, namely, Heine's basic hypergeometric functions; for instance refer to [Chen and Fu 2011] for similar discussions.

We also focus on error analysis of the numerical approximations leading to monotone properties of quotients of gamma functions in parameter triples (a, b, c). Monotone properties of the gamma function and its quotients in different forms are of recent interest to many researchers; see for instance [Alzer 1993; Anderson and Qiu 1997; Bustoz and Ismail 1986; Chen and Zhou 2014; Giordano and Laforgia 2001; Gautschi 1959; Luo et al. 2017; Mortici and Dumitrescu 2017]. In this paper, we also studied and stated a conjecture (see Conjecture 3.14) related to monotone properties of quotients of gamma functions to analyze the error estimate of the numerical approximations under consideration.

Finally, an application to continued fractions of Gauss is also discussed. Approximations of continued fractions in different forms are also attractive to many researchers; see [Lu et al. 2017; 2016].

Acknowledgements

This work was carried out when Arora was in internship at IIT Indore during the summer of 2014. The authors would like to thank the referee and the editor for their valuable remarks on this paper.

References

[Alzer 1993] H. Alzer, "Some gamma function inequalities", *Math. Comp.* **60**:201 (1993), 337–346. MR Zbl

[Anderson and Qiu 1997] G. D. Anderson and S.-L. Qiu, "A monotoneity property of the gamma function", *Proc. Amer. Math. Soc.* **125**:11 (1997), 3355–3362. MR Zbl

[Anderson et al. 1997] G. D. Anderson, M. K. Vamanamurthy, and M. K. Vuorinen, *Conformal invariants, inequalities, and quasiconformal maps*, Wiley, New York, 1997. MR Zbl

[Andrews et al. 1999] G. E. Andrews, R. Askey, and R. Roy, *Special functions*, Encyclopedia of Mathematics and its Applications **71**, Cambridge Univ. Press, 1999. MR Zbl

[Atkinson 1978] K. E. Atkinson, *An introduction to numerical analysis*, Wiley, New York, 1978. MR Zbl

[Bailey 1935] W. N. Bailey, *Generalized hypergeometric series*, Cambridge Tracts in Math. and Math. Phys. **32**, Cambridge Univ. Press, 1935. Zbl

[Beals and Wong 2010] R. Beals and R. Wong, *Special functions*, Cambridge Studies in Advanced Mathematics **126**, Cambridge Univ. Press, 2010. MR Zbl

[Bustoz and Ismail 1986] J. Bustoz and M. E. H. Ismail, "On gamma function inequalities", *Math. Comp.* **47**:176 (1986), 659–667. MR Zbl

[Chen and Fu 2011] S. H. L. Chen and A. M. Fu, "A $2n$-point interpolation formula with its applications to q-identities", *Discrete Math.* **311**:16 (2011), 1793–1802. MR Zbl

[Chen and Zhou 2014] B. Chen and H. Zhou, "On completely monotone of an arbitrary real parameter function involving the gamma function", *Appl. Math. Comput.* **242** (2014), 658–663. MR Zbl

[Gautschi 1959] W. Gautschi, "Some elementary inequalities relating to the gamma and incomplete gamma function", *J. Math. and Phys.* **38**:1-4 (1959), 77–81. MR Zbl

[Giordano and Laforgia 2001] C. Giordano and A. Laforgia, "Inequalities and monotonicity properties for the gamma function", *J. Comput. Appl. Math.* **133**:1-2 (2001), 387–396. MR Zbl

[Lu et al. 2016] D. Lu, L. Song, and C. Ma, "A quicker continued fraction approximation of the gamma function related to the Windschitl's formula", *Numer. Algorithms* **72**:4 (2016), 865–874. MR Zbl

[Lu et al. 2017] D. Lu, X. Liu, and T. Qu, "Continued fraction approximations and inequalities for the gamma function by Burnside", *Ramanujan J.* **42**:2 (2017), 491–500. MR Zbl

[Luo et al. 2017] S. Luo, J. Wei, and W. Zou, "On a transcendental equation involving quotients of gamma functions", *Proc. Amer. Math. Soc.* **145**:6 (2017), 2623–2637. MR Zbl

[Mortici and Dumitrescu 2017] C. Mortici and S. Dumitrescu, "Efficient approximations of the gamma function and further properties", *Comput. Appl. Math.* **36**:1 (2017), 677–691. MR Zbl

[Pearson et al. 2017] J. W. Pearson, S. Olver, and M. A. Porter, "Numerical methods for the computation of the confluent and Gauss hypergeometric functions", *Numer. Algorithms* **74**:3 (2017), 821–866. MR Zbl

[Rainville 1943] E. D. Rainville, *Intermediate course in differential equations*, Wiley, New York, 1943.

[Rainville 1960] E. D. Rainville, *Special functions*, Macmillan, New York, 1960. MR Zbl

Received: 2016-11-21 Revised: 2017-07-07 Accepted: 2017-07-21

hina.arora.256@gmail.com *Discipline of Electrical Engineering,*
Indian Institute of Technology, Indore, India

hina.arora@stonybrook.edu *Department of Applied Mathematics & Statistics,*
Stony Brook University, Stony Brook, NY, United States

swadesh@iiti.ac.in *Discipline of Mathematics, Indian Institute of Technology,*
Indore, India

INVOLVE 11:4 (2018)
dx.doi.org/10.2140/involve.2018.11.643

Properties of sets of nontransitive dice with few sides

Levi Angel and Matt Davis

(Communicated by Kenneth S. Berenhaut)

We define and investigate several properties that sets of nontransitive dice might have. We prove several implications between these properties, which hold in general or for dice with few sides. We also investigate some algorithms for creating sets of 3-sided dice that realize certain tournaments.

1. Nontransitive dice

Consider a set of three 3-sided dice, A, B, and C, numbered in the following way:

A	9	5	1
B	8	4	3
C	7	6	2

(1)

In this example, if we rolled each die one time, die A would beat die B $\frac{5}{9}$ of the time, die B would beat die C $\frac{5}{9}$ of the time, and die C would beat die A $\frac{5}{9}$ of the time. We say that die A "beats" or "wins against" die B if the probability that A rolls higher than B is greater than $\frac{1}{2}$. (Of course, in this case we could also say that B loses against A.) We use the notation $\succeq$ for the relation "beats", so that in this example $A \succeq B$, $B \succeq C$, and $C \succeq A$. This is an example of *nontransitivity*, since the relation $\succeq$ on $\{A, B, C\}$ is nontransitive. The study of such sets of dice dates back to [Steinhaus and Trybuła 1959; Trybuła 1961], although [Gardner 1970] was highly influential in raising interest in them. Numerous examples of nontransitive dice have since been constructed. This paper will examine a number of questions related to the construction of such sets of dice, particularly focusing on those with a small number of sides.

In what follows, we will always have a set of n k-sided dice, with the faces of each die labeled with a number from $\{1, 2, \ldots, kn\}$. We will assume each number from this set is used exactly once.

MSC2010: primary 05B30; secondary 60C05.
Keywords: nontransitive dice, probability, tournaments.

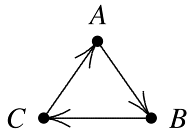


Figure 1. A nontransitive tournament on three vertices.

The relation $\succeq$ on a set of dice can be visualized as a directed graph. A *tournament* on n vertices is a directed realization of the complete graph K_n. In other words, it is a directed graph on the vertices $\{1, 2, \dots, n\}$ where for any pair of vertices i and j, either there is an edge from i to j or from j to i, but not both. We can interpret this as a definition of a relation on a set of dice — we say that a set of dice $\{X_1, X_2, \dots, X_n\}$ *realizes* a tournament T if $X_i \succeq X_j$ if and only if there is an edge from i to j in T. So the set of dice given in (1) realizes the tournament in Figure 1.

Previous work has shown that for any tournament T, it is possible to construct a set of dice that realizes this tournament. See [Angel and Davis 2017; Schaefer 2017; Bednay and Bozóki 2013] for some examples of such algorithms.

2. Properties of sets of dice

There are a number of properties that a set of dice might have that we will work with. By abuse of notation, for a given die X, we will use the same letter to represent the random variable giving the value when the die is rolled.

Uniform. We say a set of dice is *uniform* if there is a constant p so that, whenever $X \succeq Y$, we have $P(X>Y)=p$. Note that this is similar to the notion of *balanced dice* in [Schaefer and Schweig 2017], but uniformity is slightly stronger for nontransitive dice. The set of dice in (1) is uniform since $P(A>B) = P(B>C) = P(C>A) = \frac{5}{9}$.

Columned. We say a set of n dice with k sides is *columned* if the j-th smallest side on each die is chosen from the numbers $(j-1)n+1, \dots, jn$. That is, the smallest side from every die contains a number 1 through n, the second-smallest side of every die contains a number from $n+1$ through $2n$ and so on until the largest side of every die contains a number from $(k-1)n+1$ through kn. Put another way, the sides of each die are a transversal of the collection $\{\{1, 2, \dots, n\}, \{n+1, n+2, \dots, 2n\}, \dots, \{(k-1)n+1, \dots, kn\}\}$. The set of dice in (1) is columned since each die contains one number from $\{1, 2, 3\}$, one from $\{4, 5, 6\}$, and one from $\{7, 8, 9\}$.

Regular. If we have an odd number n of dice, we say the set of dice is *regular* if each die wins against $\frac{1}{2}(n-1)$ of the dice and loses against $\frac{1}{2}(n-1)$ dice. A set of dice is regular exactly if the tournament it realizes is a regular graph. The set of dice in (1) is regular since each die beats exactly one other die.

For our final property, we need one other notion. Given a die X in a set of dice, the *total number of face wins* for X is the number of ordered pairs (a, b) where a is a number on die X, b is a number on a different die in the set, and $a > b$. We similarly define the number of face wins for a die X over a die Y to be the number of ordered pairs (a, b) where a is on X, b is on Y, and $a > b$. In the example in (1), A has five face wins over B, corresponding to the pairs $(9, 8)$, $(9, 4)$, $(9, 3)$, $(5, 4)$, $(5, 3)$. Also, A has four face wins over C for a total of nine face wins.

This notion counts the total number of ways for die X to beat another die when X is rolled against another die. In other words, if we have a set S of k-sided dice containing X, and we sum $P(X{>}Y)$ for all $Y \neq X$ in S, the (unreduced) result will be a fraction with k^2 in the denominator. The total number of face wins is the numerator of that fraction. (Notice that in (1), $P(A{>}B) = \frac{5}{9}$ and $P(A{>}C) = \frac{4}{9}$, corresponding to its face wins.)

Equitable. We say a set of dice is *equitable* if each die has the same total number of face wins. The set of dice in (1) is equitable since each die has exactly nine total face wins.

We observe that, for a die X in a set of n k-sided dice,

$$\sum_{j \text{ is a face of } X} j = \text{total number of face wins for } X + \binom{k+1}{2}. \tag{2}$$

To see this, note that since our dice are numbered from 1 to nk, a face labeled j will be at least as large as the j faces labeled $1, 2, \dots, j$. However, when counting total face wins, we do not count the wins a die's face would earn over faces on the same die (including the tie against itself). There are always exactly $1+2+\cdots+k = \binom{k+1}{2}$ of these, accounting for the extra term in (2). Thus, for a set of n k-sided dice that use the numbers $1, 2, \dots, kn$ once each, equitability is equivalent to the condition that the total of the faces of each die is the same. This means that equitability is not always possible for a given number of sides and number of dice — specifically, an even number of dice each with an odd number of sides cannot be equitable.

We also explain here one way of thinking about sets of dice that is sometimes useful, which we call the *face rankings* of a die. For an ordered list of k-sided dice $X_1, X_2, \dots, X_n$, we can associate to each die a list of numbers that encodes the number of face wins for each die over the next die in the list (or for X_n over X_1), one face at a time. Specifically, for each die, we give a list of k numbers. The first number corresponds to the highest face on the die and tells us how many faces of the *next* die it is higher than, i.e., how many face wins the given die has as a result of that face. The second number similarly corresponds to the second-highest face of the die in the same way, etc. So in the example in (1), the corresponding list would be

A	320
B	311
C	221

since, for example, die A's highest face beats all of B's faces, its middle face beats two of B's faces, and its lowest face beats none of B's faces. Notice that these lists give a number of relations between the faces, which in this case (but not in all cases) are enough to reconstruct the entire set of dice. We can see A's highest face is larger than B's highest, which is larger than C's highest two faces, which are larger than A's second-highest face, etc.

For example, the set of face rankings

A	320
B	221
C	221
D	311

would describe the set of dice

A	11	8	1
B	9	7	4/5
C	10	6	3
D	12	4/5	2

(3)

where the two spaces marked 4/5 contain the faces 4 and 5 in some order. These faces are not uniquely determined by the face rankings.

With the notion of face wins and (2), we can establish some general implications between the properties described above.

Theorem 1. *Given a regular tournament on an odd number n of dice, any uniform set of dice that realize that tournament is equitable.*

Proof. Assume the dice have k sides, and that if $X \succeq Y$, then $P(X{>}Y) = j/k^2$. Then a die X wins against $\frac{1}{2}(n-1)$ dice with j face wins each and loses against $\frac{1}{2}(n-1)$ dice with $k^2 - j$ face wins each. Thus the total number of face wins for X must be $\frac{1}{2}k^2(n-1)$, and so the set is equitable. □

Theorem 2. *A uniform equitable set of an odd number n of k-sided dice must be regular.*

Proof. Note that a uniform equitable set of dice cannot be transitive, since the die that beats all others would have a greater total number of face wins than the other dice. Assume that if $X \succeq Y$, then $P(X{>}Y) = j/k^2$. Given a die X, there are $k^2(n-1)$ pairs consisting of a face of X and a face of another die. By equitability, the face from X is the higher value in exactly half of those pairs. So X has a total number of face wins equal to $\frac{1}{2}k^2(n-1)$. However, adding up the total number of face wins by comparing X to each other die means we must write $\frac{1}{2}k^2(n-1)$ as a sum of $n-1$ numbers, each of which is either j or k^2-j. This is only possible with exactly $\frac{1}{2}(n-1)$ of each. □

3. Implications between properties for 3-sided dice

For sets of small dice, there are some additional implications between these properties. In what follows, we will use the following theorem; see [Savage 1994] or [Trybuła 1961].

Theorem 3. *Suppose the numbers* $1, 2, \dots, kn$ *are arranged on a set of three* k*-sided dice, labeled* A_1, A_2, A_3. *Then at least one of the probabilities* $P(A_1 > A_2)$, $P(A_2 > A_3)$ *is less than* $\frac{1}{2}(\sqrt{5}-1)$.

For a set of three dice, at least one of the given probabilities must be less than or equal to $\frac{5}{9}$, since $\frac{6}{9} > \frac{1}{2}(\sqrt{5}-1)$. However, for a set of four dice, it is possible to arrange the dice in a cycle so each one beats the next with probability $\frac{2}{3}$. The dice described by Gardner [1970], now known as Efron dice for their discoverer, are an example of such dice.

This theorem also inspires the following theorem, which is particular to sets of 3-sided dice of any size.

Theorem 4. *Suppose the numbers* $1, 2, \dots, kn$ *are arranged on a set of* k 3*-sided dice, labeled* A_1, A_2, $A_3, \dots, A_k$. *Then at least one of the probabilities* $P(A_1 > A_2)$, $P(A_2 > A_3), \dots, P(A_k > A_1)$ *is less than* $\frac{2}{3}$.

Proof. Assume that the dice are numbered so that $A_1 \succeq A_2, \dots, A_{k-1} \succeq A_k$, $A_k \succeq A_1$. If no such k-cycle can be formed, then one of the given probabilities is in fact less than $\frac{1}{2}$. Also, assume each winning probability is at least $\frac{2}{3}$. This means each die has at least six face wins over the next die in the cycle. For 3-sided dice, this implies the middle face of a die is larger than at least two faces on the next die. (The only possible lists of face rankings with six face wins are 330 or 222.) Thus each middle face of a die is greater than the middle face of the next die in the cycle. But, this implies (by going all the way around the cycle) that each middle face is larger than itself, a contradiction. □

Theorem 5. *A nontransitive uniform set of* 3*-sided dice is columned.*

Proof. In the case that any die X has two numbers from $1, \dots, n$ or $2n+1, \dots, 3n$, there would be another die Y that had no numbers from that set. This would lead to one of X or Y beating the other with probability at least $\frac{2}{3}$. Unless the set of dice is transitive, this would force a cycle of at least three dice, each of which beats the next with probability at least $\frac{2}{3}$ (by uniformity), contradicting Theorem 4. □

Of course, one can easily make a transitive uniform set of 3-sided dice that is not columned merely by making each die in the list strictly better than the next.

Theorem 6. *A set of an odd number of equitable columned* 3*-sided dice must be uniform.*

Proof. Recall that an odd number of dice are necessary in this case for equitability to be possible. By the columned property, every die must have at least three face wins against any other die, since the largest number on each die is guaranteed to be higher than the smaller two numbers on the other dice, etc. Thus for any dice X and Y (assuming without loss of generality that $X \succeq Y$), we have $P(X{>}Y) \leq \frac{2}{3}$. However, if $P(X{>}Y) = \frac{2}{3}$, this would imply that X's largest face is greater than Y's largest face, X's second-largest face is greater than Y's, etc. This contradicts equitability. Thus if $X \succeq Y$, then $P(X{>}Y) = \frac{5}{9}$. □

Theorem 7. *A set of an odd number of regular, columned* 3*-sided dice has to be uniform.*

Proof. Since the set of dice is columned, the only way for a die X to have six face wins over a die Y is if X's largest face is greater than Y's largest face, X's second-largest face is greater than Y's, etc. This, however, implies that X beats every die that Y beats, as well as Y, so the set could not be regular. So the only possible numbers of face wins for one die over another are 5 and 4. Thus if $X \succeq Y$, $P(X{>}Y) = \frac{5}{9}$, the definition of uniformity. □

Theorem 8. *A regular equitable set of an odd number n of* 3*-sided dice must be uniform.*

Proof. By regularity, each die wins against exactly $\frac{1}{2}(n-1)$ other dice. By equitability, the total number of face wins for any die is $\frac{9}{2}(n-1)$. So the average number of face wins for any die against another is $\frac{9}{2}$. Thus if a die X has six face wins against a die Y, then X must have three face wins or fewer against some other die Z, or else its average number of face wins would be greater than $\frac{9}{2}$. Similarly, Z would have only three face wins against some die, and this will eventually create a cycle of 3-sided dice where each die beats the next with probability $\frac{2}{3}$. This contradicts Theorem 4. Thus, if $X \succeq Y$, then $P(X{>}Y) \neq \frac{2}{3}$, and so $P(X{>}Y) = \frac{5}{9}$. □

The previous few theorems, along with Theorems 1 and 2, imply the following corollary.

Corollary 9. *If a set of an odd number of* 3*-sided dice has any three of the properties equitable, columned, uniform, and regular, then it must have the fourth property. If the set of dice has two of these properties, at least one of which is regular or equitable, then it has all four properties.*

Note that it is possible for a set of dice to be only uniform and columned.

Example 10.

A	9	6	1
B	8	5	2
C	7	4	3

In this case A beats both B and C $\frac{5}{9}$ of the time and B beats C $\frac{5}{9}$ of the time, making this set of dice uniform. While this set of dice is columned, C doesn't win against any die, and A's face sum is 16 whereas C's is 14, so the example is not equitable or regular.

The theorems at the end of Section 2 suggest that generally, uniformity is the strongest condition, but the others become slightly more powerful with small dice. Generally, any one of the properties can exist alone, although a set of 3-sided dice which is uniform but not columned must be transitive.

Example 11.

A	15	5	4
B	12	11	1
C	14	7	3
D	13	9	2
E	10	8	6

This set of dice is equitable (since each die has 18 face wins) but has none of the other properties.

Example 12.

A	15	7	2
B	14	5	4
C	13	11	1
D	12	9	3
E	10	8	6

This set of dice is regular (since each die beats 2 other dice) but has none of the other properties.

Example 13.

A	14	7	1
B	11	8	2
C	15	9	3
D	12	10	4
E	13	6	5

This set of dice is columned (since each die contains one face each from the sets $\{1, 2, 3, 4, 5\}$, $\{6, 7, 8, 9, 10\}$, and $\{11, 12, 13, 14, 15\}$) but has none of the other properties.

Note that for an even number of 3-sided dice, regularity and equitability are impossible. However, we can replace these notions with weak versions. For a set of an even number n of k-sided dice (where k is odd), we say the set is *weakly regular* if every die beats either $\frac{n}{2}$ or $\frac{n}{2}-1$ other dice. We say the set is *weakly equitable* if the number of face wins for each die is within $\frac{1}{2}$ of the average number of face wins. For our dice, this is equivalent to the sum of the labels on each die being either $\left\lceil \frac{1}{2}k(kn+1) \right\rceil$ or $\left\lfloor \frac{1}{2}k(kn+1) \right\rfloor$.

Some of the theorems above generalize to the weaker versions, with the same proof. Theorems 1, 2, and 6 directly generalize to the weaker notions of regularity and equitability. However, Theorems 7 and 8 do not generalize.

First, a weakly regular columned set of dice need not be uniform.

Example 14.

A	10	8	3
B	9	7	2
C	12	6	1
D	11	5	4

Here, A and D beat two dice each, while B and C each beat one die, so these dice are weakly regular. However, $P(A>B)=\frac{2}{3}$, while the other winning probabilities are $\frac{5}{9}$, so the dice are not uniform.

Also, a weakly regular and weakly equitable set of dice need not be uniform.

Example 15.

A	8	7	5
B	12	4	3
C	11	6	2
D	10	9	1

The face sums of these dice are all 19 or 20, so they are weakly equitable, and each die beats 1 or 2 other dice, so they are weakly regular. However, $P(A>B)=\frac{2}{3}$ and $P(D>A)=\frac{2}{3}$, while the other winning probabilities are $\frac{5}{9}$, so these dice are not uniform.

Returning to our strong versions of the properties, we note that the statement that any three of these properties implies the fourth is specific to 3-sided dice. For sets of dice with more sides, it is possible to create sets of dice which have three of these properties, but not the fourth, if the missing property is either columned or uniform.

Example 16.

A	15	13	7	3	2
B	14	12	9	4	1
C	11	10	8	6	5

This is an example of three 5-sided dice which are equitable, regular, and uniform, but not columned. Note also that Algorithm 4.2 in [Schaefer and Schweig 2017] gives a way of constructing more examples which are equitable, regular, and uniform, but not columned.

The following theorem gives a large class of counterexamples.

Theorem 17. *For an odd number $n > 3$, there exists a set of n n-sided dice which is regular, equitable and columned but not uniform. In fact, each die beats the dice that it wins against with a different probability. The set of winning probabilities for a given die are the $\frac{1}{2}(n-1)$ possible winning probabilities closest to $\frac{1}{2}$.*

Proof. We begin by constructing the dice so that each die beats the next one in the list (cyclically) with $\frac{1}{2}n(n+1)-1$ face wins. To do so, take the numbers $1, n+1, 2n+1, \ldots, n^2-n+1$ (all congruent to 1 mod n) and place them on different dice. This can be done arbitrarily, but we assume without loss of generality that they are placed as shown here:

A	1					
B		$n+1$				
C			$2n+1$			
D				$3n+1$		
E					$4n+1$	
$\vdots$						$\ddots$

Then, place the number that is congruent to 2 mod n in each column above the number congruent to 1 mod n, cycling around to the bottom row when necessary.

A	1	$n+2$				
B		$n+1$	$2n+2$			
C			$2n+1$	$3n+2$		
D				$3n+1$	$4n+2$	
E					$4n+1$	$\ddots$
$\vdots$						$\ddots$

Then we repeat the process, placing the number congruent to 3 mod n in each column above the number congruent to 2 mod n, etc. This process creates a columned set of dice, shown here for $n = 5$.

A	1	7	13	19	25
B	5	6	12	18	24
C	4	10	11	17	23
D	3	9	15	16	22
E	2	8	14	20	21

Now, each die will contain exactly one number from each congruence class mod n, so the total on the die will be $\frac{1}{2}n(n^2+1)$. By construction, a die X earns $\frac{1}{2}n(n+1)-1$ face wins over the die Y after it, since each face of X is larger than the corresponding face of Y except the face of X congruent to 1 mod n. But X earns $\frac{1}{2}n(n+1)-2$ face wins over the die Z after Y, since each face of X is greater than the corresponding face of Z except the faces of X congruent to 1 or 2 mod n. This pattern repeats, and X earns one fewer face win against every successive die after it in the list. Thus X wins against exactly $\frac{1}{2}(n-1)$ other dice, but with different winning probabilities. □

Note that using face rankings, it possible to show that the construction above is the only way to create a columned n-cycle of n-sided dice where each die has $\frac{1}{2}n(n+1)-1$ face wins against the next one in the cycle. For a columned set of dice, there are exactly n ways for one die to have $\frac{1}{2}n(n+1)-1$ face wins over another. If die X beats die Y with exactly $\frac{1}{2}n(n+1)$ face wins, each face of X would be greater than the face of Y in the same column, so for X to get one fewer face win, exactly one of its faces must be smaller than the corresponding face of Y. But, no such list of face rankings can repeat in a set of n n-sided dice, or else we would be missing one such pattern, which would create a cycle within a single column, which is impossible.

This section gives a relatively complete picture of the possibilities for 3-sided dice. We attempt to generalize to sets of 4-sided dice with some success.

4. Implications between properties for 4-sided dice

Note that for 4-sided dice, it is no longer necessarily the case that a uniform set of nontransitive dice must be columned.

Example 18.

A	16	8	6	3
B	15	13	5	2
C	14	11	9	1
D	12	10	7	4

This set of dice is not columned since D has no face from the set {13, 14, 15, 16}. However, every winning probability is $\frac{9}{16}$, and the set contains the cycle $A \succeq B$, $B \succeq C$, $C \succeq A$.

In fact, we have the following:

Theorem 19. *A uniform columned set of three 4-sided dice is transitive.*

Proof. Assume that $A \succeq B$, $B \succeq C$, and $C \succeq A$. Then by uniformity and Theorem 3, $P(A>B) = P(B>C) = P(C>A) = \frac{9}{16}$. Thus, since the dice are columned, the only face rankings that are possible are 4320, 4311, 4221, or 3321. But, choosing any three of those will give us one column where each face ranking has the same number, implying that each face in that column would have to be larger than the corresponding face on the next die, even cyclically, which is impossible. □

Corollary 20. *A uniform columned set of* 4*-sided dice is transitive.*

Proof. Given a uniform columned set of 4-sided dice, if it is not transitive, then it contains some 3-cycle. Call the dice in that cycle A, B, and C. Then we can convert A, B, and C into a set of columned dice labeled by $1, \dots, 12$ by "compressing" the numbers in each column. So the smallest number on each die is changed to 1, 2, or 3, but keeping the numbers in the same relative order as in the original set of dice. Repeating this process for each column gives us a uniform columned set of three 4-sided dice, which must be transitive, a contradiction. □

This theorem gives us two more corollaries.

Corollary 21. *A set of columned equitable* 4*-sided dice must contain some evenly matched dice — dice with equal probability of beating each other.*

Proof. By the columned property, every die must have at least six face wins against any other die, since the largest number on each die is guaranteed to be higher than the smaller three numbers on the other dice, etc. Thus for any dice X and Y (assuming $X \succeq Y$ without loss of generality), $P(X{>}Y) \leq \frac{10}{16}$. However, if $P(X{>}Y) = \frac{10}{16}$, this would imply that X's largest face is greater than Y's largest face, X's second-largest face is greater than Y's, etc. This contradicts equitability. Thus for every pair of dice X and Y where $X \succeq Y$, we have $P(X{>}Y) = \frac{9}{16}$. Thus, if there were no evenly matched dice, the set of dice would be uniform, a contradiction. □

Corollary 22. *There are no sets of regular columned* 4*-sided dice.*

Proof. For a set of an odd number of 4-sided dice that is regular and columned, if $P(X{>}Y) = \frac{10}{16}$, then each face of X is higher than the corresponding face of Y. Thus X beats any die that Y beats, contradicting the regularity assumption. So if $X > Y$, then $P(X{>}Y) = \frac{9}{16}$. Then, since regularity implies that there are no evenly matched dice, the set of dice must be uniform, a contradiction. □

Note that Theorem 19 and Corollary 20 can in some sense theoretically be generalized to larger sizes of dice. However, the theorem is not as powerful, since it applies only to uniform sets of dice where the winning probability is $\bigl(\frac{1}{2}n(n+1)-1\bigr)/n^2$. So, for example, a columned set of 5-sided dice with uniform winning probability $\frac{14}{25}$ is impossible (see the note after Theorem 17), but if we want probability $\frac{13}{25}$, such a set of dice is possible:

A	21	20	12	7	5
B	22	19	11	9	4
C	23	18	15	6	3
D	24	17	14	8	2
E	25	16	13	10	1

But, if we ignore the columned property for the moment, we could ask whether any two of the other properties will imply the last remaining one for a set of 4-sided dice. Two of these implications are special cases of the theorems of Section 2. However, the question of whether a regular and equitable set of 4-sided dice must be uniform is unclear. To this point, we have yet to even find an example of a regular and equitable set of 4-sided dice to test the implication. We suspect that an equitable set of 4-sided dice must include at least two evenly matched dice somewhere, but have not been able to prove this conjecture. (This would be a strengthening of Corollary 21.)

We note here for completeness that our theorems on 4-sided dice are based on the fact that relatively few winning probabilities are possible. For larger sizes of dice, there are multiple possible winning probabilities, and generalizations of Theorem 19 (and its implications) tend not to hold.

The following example gives a set of three 6-sided dice which are columned, uniform, equitable and regular, showing that for sets of dice with a larger even number of sides, the columned property can coexist with the others.

Example 23.

A	18	14	12	7	4	2
B	17	13	11	9	6	1
C	16	15	10	8	5	3

Here, each die beats one other, with probability $\frac{19}{36}$, so the dice are regular and uniform. Moreover, the face sums all equal 57, so the dice are equitable.

However, uniformity does not imply columned for larger sets of dice. The following set of dice is adapted from [Savage 1994]. It is regular, uniform, and equitable, but not columned.

Example 24.

A	18	10	9	8	7	5
B	17	16	15	4	3	2
C	14	13	12	11	6	1

Note that each die still beats one other, and all winning probabilities are still $\frac{19}{36}$, so regularity and uniformity still hold. Also, each die has face sum 57, so the dice are still equitable.

So generally, for larger dice, the columned property is independent of the others.

5. Some tournaments achievable on 3-sided dice

One area of interest in the study of nontransitive dice is finding sets of dice with relatively few sides that realize a given tournament; see [Bozóki 2014] for one example. Given our focus on properties of sets of 3-sided dice, it is interesting to investigate which tournaments are actually realizable on 3-sided dice.

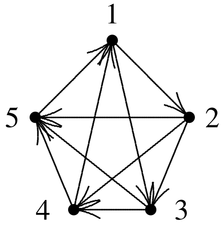


Figure 2. The 2-almost transitive tournament on five vertices.

First, for $1 < j < n$ we define the j-almost transitive tournament on n dice to be the tournament on n dice $X_1, \dots, X_n$, where $X_i \succeq X_k$ if $i < k$, or if $k = 1$ and $i \in \{n - j + 1, n - j + 2, \dots, n\}$. Intuitively, this tournament is almost transitive since each die except X_1 beats the dice after it in the list. However, the last j dice beat X_1. Figure 2 shows the 2-almost transitive tournament on five vertices.

Theorem 25. *Given integers $j < n$, there exists a columned set of n 3-sided nontransitive dice which realize the j-almost transitive tournament.*

Proof. We construct a table as follows. The third column contains $3n$ through $2n + 1$, in order, from top to bottom. In the second column, die X_{n-j+1}, which is the lowest-numbered die which beats X_1, has face $2n$. The remaining numbers are added downward from it in order, wrapping around to the top after placing $2n - j + 1$ on die X_n. In the first column, X_1 receives the face 1, then dice X_{n-j+1} through X_n contain the numbers 2 through $j + 1$, in order, and dice X_2 through X_{n-j} receive the numbers $j + 2$ through n, in order.

For example, the 2-almost transitive tournament on seven dice is realized by

X_1	1	12	21
X_2	4	11	20
X_3	5	10	19
X_4	6	9	18
X_5	7	8	17
X_6	2	14	16
X_7	3	13	15

Then one can check easily that the dice from X_2 through X_{n-j} defeat each other transitively since the last two columns are in descending order. However, X_1 loses to X_{n-j+1} through X_n since its smallest two faces are smaller. But dice X_2 through X_{n-j} all beat X_{n-j+1} through X_n because of the first and third columns. Lastly, the dice X_{n-j+1} through X_n beat each other transitively because of the second and third columns. □

For a tournament T, let T' denote the *opposite tournament*, the tournament on the same set of dice with all edges reversed. Note that if we have a set of n k-sided dice labeled with the numbers 1 through nk that realizes a tournament T, we can replace each face label j with the label $nk+1-j$ to get a set of dice that realize T'. In the case of the j-almost transitive tournament T, we can see that T' is the tournament where $X_i \succeq X_k$ if $i > k$, or if $i = 1$ and $k \in \{n-j+1, n-j+2, \dots, n\}$. We call this the j-upsetter tournament on n vertices, since it can be obtained from a transitive tournament by making the "last-place" die beat the j dice that won against the most dice in the transitive tournament. Thus as a corollary of this theorem, it is always possible to construct a set of 3-sided dice that realize the j-upsetter tournament on n vertices.

We also define the cyclic tournament on $2n+1$ vertices to be the regular tournament on the dice $X_1, \dots, X_{2n+1}$ where each die beats the next n dice in the list, wrapping around to the beginning as necessary. (This name is given to it because it can be constructed from the data of a cyclic group.)

Theorem 26. *The cyclic tournament on* $2n+1$ *vertices is realizable with* 3*-sided dice.*

Proof. We construct a table as follows. In the first column, we add the numbers in the order $2, 4, 6, \dots, 2n, 1, 3, \dots, 2n+1$, i.e., counting by twos mod $2n+1$. In the second column, place $4n+2$ on the same die as the entry 1 and add the remaining numbers in the second column in order downward from that die. Then in the third column we add the numbers $6n+3$ through $4n+3$ starting with $6n+3$ on the first die and moving downward in order.

For example, the cyclic tournament on seven dice is realized by

X_1	2	10	21
X_2	4	9	20
X_3	6	8	19
X_4	1	14	18
X_5	3	13	17
X_6	5	12	16
X_7	7	11	15

To see that this realizes the cyclic tournament, notice that X_{n+1} has the entries $1, 4n+2, 5n+3$. So it loses to the n dice above it because of the first and third columns, but it beats the n dice below it because of the last two columns. Then for any die X_k where $k < n+1$, we can see that X_k will beat X_{k+1} through X_n because of the last two columns. It will beat X_{n+1} through X_{n+k} because of the first and last columns. However, X_{n+k+1}'s first column contains the entry that is one more than X_k's first column, so X_k loses to X_{n+k+1} through X_{2n+1}. The dice after X_{n+1} can be examined similarly. □

Another construction that is very helpful is the "blow-up" of a tournament. (The terminology is borrowed from a vaguely similar concept in algebraic geometry.) Say we have a tournament S on the vertices $Y_1, \dots, Y_m$, and a tournament T on the vertices $X_1, \dots, X_n$. We can form a new tournament U on the vertices $Y_1, \dots, Y_m, X_2, X_n$, where in U, the relation $\succeq$ between X_i and X_j or Y_i and Y_j is the same as in T or S respectively, and $Y_i \succeq X_j$ exactly if $X_1 \succeq X_j$ in T. Intuitively, the vertex X_1 in T has been "blown up" into an entire copy of S, which has the same relationship to the other X_j as X_1 did. We call U the blow-up of T at X_1 with by S.

Theorem 27. *If there is a columned set of k-sided dice that realize S and a set of k-sided dice that realize T, then there is a set of k-sided dice that realize the blowup of T at any vertex X by S.*

Proof. Let X be the die representing the vertex at which we blow up T, and assume it has faces $a_1, a_2, \dots, a_k$, where $a_1 < a_2 < \cdots < a_k$. We choose a small $\epsilon > 0$. Then for each die Y_i in our realization of S, we replace its smallest label y_{i1} by $a_1 + y_{i1}\epsilon$, its second-lowest face y_{i2} by $a_2 + y_{i2}\epsilon$, etc. (This will of course create a set of dice labeled with numbers other than the usual integers, but we will adjust accordingly at the end of the algorithm.) We claim that the new dice Y_i that we have just constructed will have the same relationships to each other as the original dice realizing S. To see this, note that for faces in the same "column", we have $y_{ij} < y_{kj}$ if and only if $a_j + y_{ij}\epsilon < a_j + y_{kj}\epsilon$. For numbers in different columns, $y_{ij} < y_{km}$ whenever $j < m$. But since $a_j < a_m$ in this case, we will also have $a_j + y_{ij}\epsilon < a_m + y_{km}\epsilon$. Moreover, we can choose ϵ small enough that $a_j + y_{ij}\epsilon < a_j + 1$ always holds, so that every entry $a_j + y_{ij}\epsilon$ is in the same position as a_j relative to the faces of the other dice that realized T. That is, the new die Y_i will beat (or lose to) those other dice in the same way that X did. So, if we remove X from the set of dice and include the altered Y_i's, the resulting set will realize the blowup of T at X by S. And finally, we can alter the actual numbers on the resulting dice set by replacing the lowest number on all the faces by a 1, the second-lowest number by 2, etc., without changing the structure of $\succeq$. □

Example 28. As an example of this theorem, we can take both S and T to be the dice set of (1), and let X be die C. For clarity, we will call the dice in S lowercase a, b, and c. The algorithm (using $\epsilon = .1$) originally gives

A	1	5	9
B	3	4	8
a	2.1	6.5	7.9
b	2.3	6.4	7.8
c	2.2	6.6	7.7

Converting these to the numbers 1 through 15, we obtain

A	1	7	15
B	5	6	14
a	2	9	13
b	4	8	12
c	3	10	11

As another example, the algorithm above for constructing a set of 3-sided dice realizing the 1-almost transitive tournament implicitly makes use of the blow-up algorithm. If we start with the dice set of (1) as T, and S as the dice set

X_1	1	$2n$	$3n$
X_2	2	$2n-1$	$3n-1$
X_3	3	$2n-2$	$3n-2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

realizing the transitive tournament, then performing the algorithm to blow up T at B by S will give the same construction of a dice set realizing the 1-almost transitive tournament as Theorem 25.

Notice that these theorems allow us to construct a wide range of 3-sided realizations of tournaments. But in general, not all tournaments are realizable with 3-sided dice. An exhaustive computer search found that all tournaments with up to seven vertices could be realized on 3-sided dice, but that approximately 95 tournaments on eight vertices (out of 6880) could not be realized on 3-sided dice. On nine vertices, there are even some regular tournaments that cannot be realized with 3-sided dice. The question of exactly which tournaments can be realized on 3-sided dice seems difficult but interesting.

References

[Angel and Davis 2017] L. Angel and M. Davis, "A direct construction of nontransitive dice sets", *J. Combin. Des.* (online publication June 2017).

[Bednay and Bozóki 2013] D. Bednay and S. Bozóki, "Constructions for nontransitive dice sets", pp. 15–23 in *Proceedings of the 8th Japanese-Hungarian Symposium on Discrete Mathematics and Its Applications* (Veszprém, Hungary, 2013), edited by A. Frank et al., Budapest Univ. Technology and Economics, 2013.

[Bozóki 2014] S. Bozóki, "Nontransitive dice sets realizing the Paley tournaments for solving Schütte's tournament problem", *Miskolc Math. Notes* **15**:1 (2014), 39–50. MR Zbl

[Gardner 1970] M. Gardner, "The paradox of the nontransitive dice and the elusive principle of indifference", *Sci. Amer.* **223**:6 (1970), 110–114.

[Savage 1994] R. P. Savage, Jr., "The paradox of nontransitive dice", *Amer. Math. Monthly* **101**:5 (1994), 429–436. MR Zbl

[Schaefer 2017] A. Schaefer, “Balanced non-transitive dice, II: Tournaments”, preprint, 2017. arXiv

[Schaefer and Schweig 2017] A. Schaefer and J. Schweig, “Balanced nontransitive dice”, *College Math. J.* **48**:1 (2017), 10–16. MR

[Steinhaus and Trybuła 1959] H. Steinhaus and S. Trybuła, “On a paradox in applied probabilities”, *Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys.* **7** (1959), 67–69. Zbl

[Trybuła 1961] S. Trybuła, “On the paradox of three random variables”, *Zastos. Mat.* **5**:4 (1961), 321–332. MR Zbl

Received: 2016-12-15 Revised: 2017-05-22 Accepted: 2017-08-14

langel@muskingum.edu *Department of Mathematics and Computer Science,*
Muskingum University, New Concord, OH, United States

mattd@muskingum.edu *Department of Mathematics and Computer Science,*
Muskingum University, New Concord, OH, United States

INVOLVE 11:4 (2018)
dx.doi.org/10.2140/involve.2018.11.661

Numerical studies of serendipity and tensor product elements for eigenvalue problems

Andrew Gillette, Craig Gross and Ken Plackowski

(Communicated by Antonia Vecchio)

While the use of finite element methods for the numerical approximation of eigenvalues is a well-studied problem, the use of serendipity elements for this purpose has received little attention in the literature. We show by numerical experiments that serendipity elements, which are defined on a square reference geometry, can attain the same order of accuracy as their tensor product counterparts while using dramatically fewer degrees of freedom. In some cases, the serendipity method uses only 50% as many basis functions as the tensor product method while still producing the same numerical approximation of an eigenvalue. To encourage the further use and study of serendipity elements, we provide a table of serendipity basis functions for low-order cases and a Mathematica file that can be used to generate the basis functions for higher-order cases.

1. Introduction

Computational approximation of eigenvalues is a topic of ongoing interest across a broad spectrum of the applied mathematics community, due in part to the wide variety of application areas where it is required. In this work, we compare two finite element methods for the computation of eigenvalues of the Laplacian: tensor product and serendipity. While tensor product finite element methods have been used for decades to compute eigenvalues, the lesser known serendipity elements have been employed rarely, if ever, in this context, despite the fact that they are expected to require fewer computations to achieve the same order of accuracy.

The potential benefits of a serendipity element eigenvalue solver are obvious from a rough estimate of the degrees of freedom required for a method with $\mathcal{O}(h^p)$ error decay. Here, h indicates the maximum diameter of an affinely mapped square mesh element and $p \geq 1$ indicates the maximum exponent of any variable appearing in a basis for the element. The tensor-product finite element method for H^1-conforming problems in $\mathbb{R}^n$ uses $(p+1)^n$ basis functions per element, while the

MSC2010: 35P15, 41A25, 65H17, 65N30.
Keywords: serendipity finite elements, eigenvalue approximation, h-refinement, p-refinement.

serendipity method uses roughly $p^n/n!$ for large p. Thus, for domains in $\mathbb{R}^2$, an $\mathcal{O}(h^p)$ serendipity method has about 50% of the number of basis functions as its tensor product counterpart, while for domains in $\mathbb{R}^3$, an $\mathcal{O}(h^p)$ serendipity method has only 17% of the number of basis functions as a tensor product method. As we show by numerical evidence, these computational savings are not restricted to an asymptotic regime but can be realized even in domains in $\mathbb{R}^2$ and for values $p \leq 6$.

The body of prior work studying finite element methods for eigenvalue approximation dates back to the 1970s [Hackbusch 1979] and is quite large due to the many options available when designing finite element schemes and the many kinds of inquiries that could be made. An excellent survey of the research in this area was given by Boffi [2010]. While many works are concerned with approximation of the spectrum of the Laplacian (e.g., concerns about pollution and completeness of the computed spectrum), here we focus on the accurate computation of individual eigenvalues to a high order of accuracy with the goal of minimizing the number of global degrees of freedom. A similar kind of study by Wang, Monk, and Szabó [Wang et al. 1996] compared h- and p-refinement schemes on tetrahedra for computing resonant modes in a cavity using tetrahedral elements. This work focuses on square elements, which offer greater ability to reduce the number of global degrees of freedom than simplicial elements.

In this paper, we carry out a series of numerical experiments to compare the accuracy of serendipity and tensor product finite element methods in the context of eigenvalue computation. We compare square and L-shaped domains, Dirichlet and Neumann boundary conditions, and h- and p-refinement strategies. To ensure a fair comparison, we implement basis functions for both tensor product and serendipity elements using the construction process described in the work of Floater and Gillette [2017], which uses interpolation conditions based on partial derivative data at edge and cell midpoints. To the best of our knowledge, this is the first time such functions have been tested numerically.

Our results show that a p-refinement strategy with serendipity elements is preferable to the same strategy with tensor product elements in a variety of domain and boundary condition scenarios. In particular, we find many specific instances where the serendipity elements achieve the same order of accuracy as the corresponding tensor product element with only 50% the number of degrees of freedom. The results also show that an h-refinement strategy does not always favor serendipity elements, meaning application context is essential when deciding between the use of tensor product and serendipity elements.

The remainder of the paper is organized as follows. In Section 2, we review the eigenvalue problem for the Laplace equation with Neumann and Dirichlet boundary conditions, as well as the derivation of a Galerkin finite element method. Following this is a discussion of the two families of finite elements studied in this paper:

tensor product and serendipity. In Section 3, we state interpolation conditions that involve both values and derivative values and compute the basis functions for both the tensor product and the serendipity finite elements. We also discuss the relevant components for implementation via Mathematica and MATLAB. In Section 4, we provide a description of our results and a discussion of the comparison between the tensor product and serendipity elements. This includes comparisons of the aforementioned scenarios. In Section 5, we summarize our conclusions and give some directions for future research. Finally, in the Appendix, we give tables of the serendipity basis functions that we use and provide a link to a Mathematica code that can be used for further studies.

2. Finite element methods for eigenvalue problems

Our focus in this work is the scalar-valued Laplace eigenvalue problem. With Dirichlet boundary conditions, the problem is to find $\lambda \in \mathbb{R}$ and $u \in H^2(\Omega)$ such that

$$\begin{cases} -\Delta u = \lambda u & \text{in } \Omega, \\ \quad u = 0 & \text{on } \partial\Omega. \end{cases} \tag{1}$$

With Neumann boundary conditions, the problem is to find $\lambda \in \mathbb{R}$ and $u \in H^2(\Omega)$ such that

$$\begin{cases} -\Delta u = \lambda u & \text{in } \Omega, \\ \mathrm{d}u/\mathrm{d}\boldsymbol{n} = 0 & \text{on } \partial\Omega, \end{cases} \tag{2}$$

where $\boldsymbol{n}$ is the unit vector normal to the boundary of Ω.

We consider two subsets of $\mathbb{R}^2$ for the domain Ω: the unit square $[0, 1]^2$ and the L-shaped domain, $[0, 2]^2 - (1, 2]^2$. On $[0, 1]^2$, the eigenvalues for the Dirichlet problem (1) are

$$(m^2 + n^2)\pi^2 \quad \text{for } m, n \in \{1, 2, \ldots\} \text{ and } \Omega = [0, 1]^2.$$

For the Neumann problem (2) on $[0, 1]^2$, the eigenvalues are

$$(m^2 + n^2)\pi^2 \quad \text{for } m, n \in \{0, 1, 2, \ldots\} \text{ and } \Omega = [0, 1]^2;$$

the only difference being that m and n are allowed to have value 0. For (m, n) pairs with $m \neq n$, the corresponding eigenvalue has multiplicity at least 2, while those with $m = n$ have multiplicity 1 and are called "simple".

On the L-shaped domain, Dauge [2003] has given benchmark computations with at least eight digits of accuracy for the lowest nonzero eigenvalues for the Neumann problem. The first four of these are

$$\lambda^{(1)} = 1.4756218450, \quad \lambda^{(2)} = 3.5340313683,$$
$$\lambda^{(3)} = 9.8696044011, \quad \lambda^{(4)} = 11.389479398.$$

Note that $\lambda^{(3)} = 2\pi^2$, which is also an eigenvalue for the Dirchlet problem. In our experiments, we look at approximating $2\pi^2$ for each kind of boundary condition, as well as the approximation of $\lambda^{(1)}$ for the Neumann case.

Discretization of (1) for numerical approximation begins with the weak form of (1). Set $V := H_0^1(\Omega)$ and find $\lambda \in \mathbb{R}$ and $u \in V,\ u \neq 0$, such that

$$\int_\Omega \nabla u \cdot \nabla v = \lambda \int_\Omega uv \quad \text{for all } v \in V. \tag{3}$$

A Galerkin finite element method seeks a solution to (3) that holds over a finite-dimensional subspace $V_{h,p} \subset V$: find $\lambda_{h,p} \in \mathbb{R}$ and $u_{h,p} \in V_{h,p},\ u_{h,p} \neq 0$, such that

$$\int_\Omega \nabla u_{h,p} \cdot \nabla v_{h,p} = \lambda_{h,p} \int_\Omega u_{h,p}\, v_{h,p} \quad \text{for all } v_{h,p} \in V_{h,p}. \tag{4}$$

The dimension of $V_{h,p}$ is determined by the type of element used (tensor product or serendipity, in our case) in addition to the parameters h and p. Here, h indicates the maximum diameter of an element in the mesh and p indicates the maximum exponent of any variable appearing in the monomial basis for the element. Hence, as $h \to 0$ or $p \to \infty$, we have $\dim V_{h,p} \to \infty$.

We consider two possible choices for $V_{h,p}$ that are subsets of $H^1(\Omega)$ and are associated to a partition of Ω into a mesh of squares. We will follow notational conventions from the periodic table of the finite elements [Arnold and Logg 2014a; 2014b] to describe the two choices in terms of the local spaces on each square element. The first choice for a local space is $\mathcal{Q}_p^-\Lambda^0(\square_2)$, more commonly known as the tensor product element of order p on a square [Arnold et al. 2015]. This element has 1 degree of freedom per vertex, $(p-1)$ degrees of freedom per edge, and $(p-1)^2$ degrees of freedom associated to the interior, for a total of $(p+1)^2$ degrees of freedom per square element. The second choice for a local space is $\mathcal{S}_p\Lambda^0(\square_2)$, known as the serendipity element of order p on a square [Arnold and Awanou 2011]. The serendipity element has the same degrees of freedom associated to vertices and edges of the square, but only $\frac{1}{2}(p-3)(p-2)$ degrees of freedom[1] associated to the interior of the square. It has a total of $\frac{1}{2}(p^2+3p+6)$ degrees of freedom per element.

In addition to the type of domain Ω (square or L-shaped), the family of element ($\mathcal{Q}^-$ or $\mathcal{S}$), and the order of p selected, the dimension of $V_{h,p}$ depends on the maximum diameter of a mesh element. We only consider meshes where all elements are squares of the same side length h, so that the maximum diameter of a mesh element is $\sqrt{2}h$. By this convention, if $h = 1/N$ for an integer $N \geq 1$, the square domain will have N^2 elements and the L-shaped domain will have $3N^2$ elements.

[1] For $p = 1$, there are no interior degrees of freedom; the formula applies for any $p \geq 2$.

By counting the total number of vertices, edges, and elements in the mesh, we have the formula

$$\dim V_{h,p} = (\#\text{ vertices}) + (p-1)\cdot(\#\text{ edges}) + \left(\frac{\#\text{ DoF}}{\text{interior}}\right)\cdot(\#\text{ elements}),$$

where the number of degrees of freedom (DoF) per interior depends on the choice of $\mathcal{Q}^-$ or $\mathcal{S}$, as described above. Note that when Dirichlet boundary conditions are used, the values of degrees of freedom associated to the boundary of the domain are set to zero, which decreases the dimension of $V_{h,p}$.

The goal of the numerical experiments in this paper is to study the following question: given a domain, a set of boundary conditions, a rough guess for an eigenvalue λ, an h-refinement or p-refinement strategy, and a desire to attain a precise estimate of λ while avoiding fruitless growth in $\dim V_{h,p}$, is it better to use $\mathcal{Q}^-$ or $\mathcal{S}$ elements? Since the $\mathcal{S}_p\Lambda^0$ and $\mathcal{Q}_p^-\Lambda^0$ elements each contain polynomials of total degree at most p and $\dim \mathcal{S}_p\Lambda^0 < \dim \mathcal{Q}_p^-\Lambda^0$ for $p \geq 2$, we might expect that the serendipity elements would be preferable in every case. On the other hand, perhaps the "extra" approximation power afforded by the larger basis in the tensor product element provides better eigenvalue estimation overall. To make a fair comparison, we implement serendipity and tensor product elements by the same methodology, and then report their results when used in a series of computational experiments.

3. Implementation of serendipity elements

Here, for the first time, we compute and employ the basis functions for $\mathcal{S}_p\Lambda^0(\square_2)$ with Hermite-like interpolation conditions at edge midpoints, as described in [Floater and Gillette 2017]. We review the degrees of freedom for these elements here and explain how the process outlined in that paper was used to derive the basis functions employed in our numerical experiments.

Serendipity degrees of freedom. The term "serendipity element" has appeared in various mathematical and engineering texts since the 1970s [Brenner and Scott 1994; Ciarlet 1978; Hughes 1987; Mandel 1990; Szabó and Babuška 1991; Strang and Fix 1973], referring to the fact that these elements seemed to achieve $O(h^p)$ accuracy with fewer degrees of freedom than their tensor product counterparts. Arnold and Awanou [2011] provided degrees of freedom in the classical finite element sense for the H^1-conforming version of these spaces: for a d-dimensional face $\square_d$ of an n-cube $\square_n$, the order-p serendipity degrees of freedom for a scalar function u are

$$u \mapsto \int_{\square_d} uq \quad \text{for all } q \in \mathcal{P}_{p-2d}(\square_d), \tag{5}$$

where $\mathcal{P}_{p-2d}(\square_d)$ denotes the space of polynomials in n variables of degree $\leq p-2d$ on face $\square_d$. For $n=2$ and $p \geq 2$, the space of polynomials associated to the degrees of freedom (5) is denoted by $\mathcal{S}_p\Lambda^0(\square_2)$ and given by

$$\mathcal{S}_p\Lambda^0(\square_2) = \mathcal{P}_p(\square_2) \oplus \operatorname{span}\{x^p y, xy^p\}. \tag{6}$$

It is shown in [Arnold and Awanou 2011] that the degrees of freedom (5) are unisolvent for (6), but a consideration of how to construct suitable basis functions for the implementation of these elements in applications was not provided.

Basis functions for serendipity elements. We use a procedure outlined by Floater and Gillette [2017] to construct basis functions for the $\mathcal{S}_p\Lambda^0(\square_2)$ element. To the best of our knowledge, these functions have not been constructed explicitly or used in numerical experiments previously. The procedure is also used to construct bases for the $\mathcal{Q}_p^-\Lambda^0(\square_2)$ element.

Given $p \geq 1$, we first we define a set of $p+1$ functions over $[-1, 1]$, denoted by

$$\Phi_p[x] := \{\phi_1(x), \dots, \phi_{p+1}(x)\}.$$

Let D denote the endpoints and midpoint of $[-1, 1]$, i.e., $D = \{-1, 0, 1\}$, and denote the Kronecker delta function by

$$\delta_i(j) = \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases}$$

Define $\Phi_1[x] := \{(1-x)/2, (1+x)/2\}$. For $p \geq 2$, fix the interpolation properties[2]

$$\phi_1(x_0) = \delta_{-1}(x_0) \qquad \forall x_0 \in D, \tag{7}$$

$$\phi_2(x_0) = \delta_0(x_0) \qquad \forall x_0 \in D, \tag{8}$$

$$\phi_{p+1}(x_0) = \delta_1(x_0) \qquad \forall x_0 \in D, \tag{9}$$

$$\phi_i^{(k)}(0) = 0 \qquad \forall i \in \{1, 2, p+1\},\ \forall k \in \{1, \dots, p-2\}, \tag{10}$$

$$\phi_i(x_0) = 0 \qquad \forall x_0 \in D,\ \forall i \in \{3, \dots, p\}, \tag{11}$$

$$\phi_i^{(i-2)}(0) = 1 \qquad \forall i \in \{3, \dots, p\}, \tag{12}$$

$$\phi_i^{(k)}(0) = 0 \qquad \forall i \in \{3, \dots, p\},\ \forall k \in \{1, \dots, i-3\}. \tag{13}$$

For $i = 1$ to $p+1$, we find the lowest-degree polynomial ϕ_i that satisfies the above constraints. Since there are at most $p+1$ constraints for each i, this process uniquely defines a set of $p+1$ polynomials, each of degree at most p. Moreover, ϕ_1, ϕ_2, and ϕ_{p+1} are the only functions in the set that have nonzero values at -1, 0, and 1, respectively, while the functions ϕ_3 through ϕ_p have linearly independent

[2] If a set of indices on the right is empty, the property should be treated as vacuous.

p	$\phi_1(x)$	$\phi_2(x)$	$\phi_3(x)$	$\phi_4(x)$	$\phi_5(x),\ \phi_6(x)$
1	$-\frac{1}{2}(x-1)$	$\frac{1}{2}(x+1)$			
2	$\frac{1}{2}(x-1)x$	$1-x^2$	$\frac{1}{2}x(x+1)$		
3	$-\frac{1}{2}(x-1)x^2$	$1-x^2$	$x-x^3$	$\frac{1}{2}x^2(x+1)$	
4	$\frac{1}{2}(x-1)x^3$	$1-x^4$	$x-x^3$	$-\frac{1}{2}(x-1)x^2(x+1)$	$\frac{1}{2}x^3(x+1)$
5	$-\frac{1}{2}(x-1)x^4$	$1-x^4$	$x-x^5$	$-\frac{1}{2}(x-1)x^2(x+1)$	$-\frac{1}{6}(x-1)x^3(x+1),\ \frac{1}{2}x^4(x+1)$

Table 1. Basis functions for $\Phi_p[x]$ with $1 \le p \le 5$.

constraints on their derivatives at 0. Thus, for each $p \ge 1$, $\Phi_p[x]$ is a basis for $\mathcal{P}_p([-1,1])$. The sets $\Phi_1[x], \dots, \Phi_5[x]$ are listed explicitly in Table 1.

By taking tensor products of the $\Phi_p[x]$ sets, we can build out bases for tensor product and serendipity spaces over $[-1,1]^n$ for any $n \ge 1$, although we consider only $n = 2$ here. We fix the notation

$$\Phi_{pq} := \{\phi_i(x)\phi_j(y) : \phi_i(x) \in \Phi_p[x],\ \phi_j(y) \in \Phi_q[y]\},$$

where p and q need not be distinct. A basis for the tensor product space $\mathcal{Q}_p^-\Lambda^0(\square_2)$ can be computed immediately as

$$\text{basis for } \mathcal{Q}_p^-\Lambda^0(\square_2) = \Phi_{pp}.$$

A basis for the serendipity space $\mathcal{S}_p\Lambda^0(\square_2)$ is more involved to describe but only slightly more difficult to compute. First, an addition operation on sets of the type Φ_{pq} is defined as follows. To build the set $\Phi_{pq} + \Phi_{rs}$, let $M = \max\{p, q, r, s\}$ and build a square array of indices $\{1, \dots, M+1\} \times \{1, \dots, M+1\}$. Associate the function $\phi_i(x)\phi_j(y) \in \Phi_{pq}$ to index $\{k, \ell\}$ according to the rule

$$\phi_i(x)\phi_j(y) \mapsto \begin{cases} \{M+1, j\} & \text{if } i = p+1,\ j < q+1, \\ \{i, M+1\} & \text{if } i < p+1,\ j = q+1, \\ \{M+1, M+1\} & \text{if } i = p+1,\ j = q+1, \\ \{i, j\} & \text{otherwise.} \end{cases}$$

Associate the function $\phi_i(x)\phi_j(y) \in \Phi_{rs}$ to indices according to the same rule, replacing p by r and q by s. Initialize $\mathbb{A}_{pq,M}$ as an $(M+1)\times(M+1)$ array of zeros, then place the functions from Φ_{pq} into $\mathbb{A}_{pq,M}$ according to their index assignment. Define $\mathbb{A}_{rs,M}$ analogously, using functions from Φ_{rs}. The set $\Phi_{pq} + \Phi_{rs}$ is then defined to be the set of nonzero entries of $\mathbb{A}_{pq,M} + \mathbb{A}_{rs,M}$. In practice, this reindexing and summation procedure is carried out by inserting rows or columns of zeros at appropriate places into the arrays storing Φ_{pq} and Φ_{rs} and then adding the arrays together.

A basis for $\mathcal{S}_p\Lambda^0(\square_2)$ can then be written as a linear combination of this addition operation on some Φ_{rs} sets. For $p=1$ through $p=6$, these linear combinations are

$$\mathcal{S}_1\Lambda^0(\square_2)\text{ basis} = \Phi_{11}, \tag{14}$$

$$\mathcal{S}_2\Lambda^0(\square_2)\text{ basis} = \Phi_{21}+\Phi_{12}-\Phi_{11}, \tag{15}$$

$$\mathcal{S}_3\Lambda^0(\square_2)\text{ basis} = \Phi_{31}+\Phi_{13}-\Phi_{11}, \tag{16}$$

$$\mathcal{S}_4\Lambda^0(\square_2)\text{ basis} = \Phi_{41}+\Phi_{14}+\Phi_{22}-(\Phi_{21}+\Phi_{12}), \tag{17}$$

$$\mathcal{S}_5\Lambda^0(\square_2)\text{ basis} = \Phi_{51}+\Phi_{15}+\Phi_{32}+\Phi_{23}-(\Phi_{31}+\Phi_{13}+\Phi_{22}), \tag{18}$$

$$\mathcal{S}_6\Lambda^0(\square_2)\text{ basis} = \Phi_{61}+\Phi_{16}+\Phi_{42}+\Phi_{24}+\Phi_{33}-(\Phi_{41}+\Phi_{14}+\Phi_{23}+\Phi_{32}). \tag{19}$$

The derivation of these linear combinations is given in [Floater and Gillette 2017, §5] using different notation. The techniques in that paper can produce bases in this way for $\mathcal{S}_p\Lambda^0(\square_n)$ for any $p \geq 1$ and $n \geq 1$. As an example, in the Appendix, we provide the two-dimensional serendipity basis functions for $p=1$ to 4.

***Implementation via Mathematica and* MATLAB.** We use Mathematica to compute the bases for $\mathcal{Q}_p^-\Lambda^0(\square_2)$ and $\mathcal{S}_p\Lambda^0(\square_2)$ according to the procedure just described and the process of basis generation is summarized below. The Mathematica function `InterpolatingPolynomial` is used to produce the sets $\Phi_p[x]$ based on conditions (7)–(13). For example, $\phi_3(x)\in\Phi_3[x]$ should satisfy

$$\phi_3(-1)=\phi_3(0)=\phi_3(1)=0,$$

as well as $\phi_3'(0)=1$. The unique cubic polynomial satisfying these constraints is computed by the command

```
InterpolatingPolynomial[{{-1, 0}, {0, 0, 1}, {1, 0}}, x].
```

We define a function `interpolatingList[p]` that creates the required inputs to `InterpolatingPolynomial` for each $\phi_i\in\Phi_p[x]$. We also define a function `genTable2D[p,q,M]` that builds the array $\mathbb{A}_{pq,M}$. Bases for $\mathcal{S}_p\Lambda^0(\square_2)$ are constructed by simplifying linear combinations of appropriate `genTable2D[r,s,M]` arrays according to (14)–(19); the value of M is set to p for each term in the combination so that the output is a $(p+1)\times(p+1)$ array with exactly $\dim\mathcal{S}_p\Lambda^0(\square_2)$ nonzero entries. The basis for $\mathcal{Q}_p^-\Lambda^0(\square_2)$ is built by the command `genTable2D[p,p,p]`, which generates a $(p+1)\times(p+1)$ array with all entries nonzero.

Once the basis functions are created, we pass them to a finite element solver in `MATLAB` in order to compute approximate eigenvalues. The resulting finite element problem is given by

$$\lambda M v = L v,$$

where M is the mass matrix and L is the stiffness matrix with

$$M = [M_{i,j}] \quad \text{such that } M_{i,j} = \int_\Omega \psi_i \psi_j \,\mathrm{d}A,$$
$$L = [L_{i,j}] \quad \text{such that } L_{i,j} = \int_\Omega \langle \nabla \psi_i, \nabla \psi_j \rangle \,\mathrm{d}A,$$

where ψ_i, ψ_j range over a basis for $\mathcal{Q}_p^- \Lambda^0(\square_2)$ or $\mathcal{S}_p^- \Lambda^0(\square_2)$. The finite element solver takes a local approach, making use of the basis functions defined over a reference element as above (specifically $[-1, 1]^2$). By calculating the desired entries of the mass and stiffness matrices over the reference element, scaling, and assembling on the global square or L-shaped domain, we produce global mass and stiffness matrices.

Furthermore, in the derivation of the variational form of the problem, the imposition of the Neumann conditions is encoded by the vanishing of any integrals over the boundary of the domain. To impose the Dirichlet conditions, it is necessary to manipulate the equations in the discrete problem that solve for the coefficients corresponding to boundary nodes. Traditionally, this is realized by setting each of the coefficients corresponding to value interpolating nodes on the boundary equal to zero. As the tensor product and serendipity basis functions that we use include interpolation of some partial derivative values along the boundary, we also set to zero the coefficients of the basis functions corresponding to those conditions.

4. Numerical experiments and results

Our numerical experiments are characterized by four choices: domain (square or L-shaped), boundary conditions (Dirichlet or Neumann), eigenvalue λ being approximated, and refinement strategy (p-refine with h fixed or h-refine with p fixed). For each choice, we report the error in the numerical approximation of λ as a function of the number of global degrees of freedom, i.e., the dimension of $V_{h,p}$. Two data series are generated in this fashion: one for tensor product elements and one for serendipity elements, using $p = 1$ through 6 for a fixed h value (p-refinement), or for $h = 1, \frac{1}{2}, \ldots, \frac{1}{5}$ for a fixed p value (h-refinement). The results of these experiments are shown in Figures 1–10.

Square domain. Our first comparison of the tensor product and serendipity elements is on a square domain with Neumann boundary conditions. Figure 1 shows the error in approximating the eigenvalue $2\pi^2$ when we fix h and allow p to vary. Ignoring for now the outlier corresponding to one of the tensor product solutions, we see that in nearly every case, using serendipity elements can match the accuracy of the eigenvalue obtained by tensor product elements with much fewer degrees of freedom. For example, in Figure 1, $h = \frac{1}{4}$, we see that we can obtain an approximate

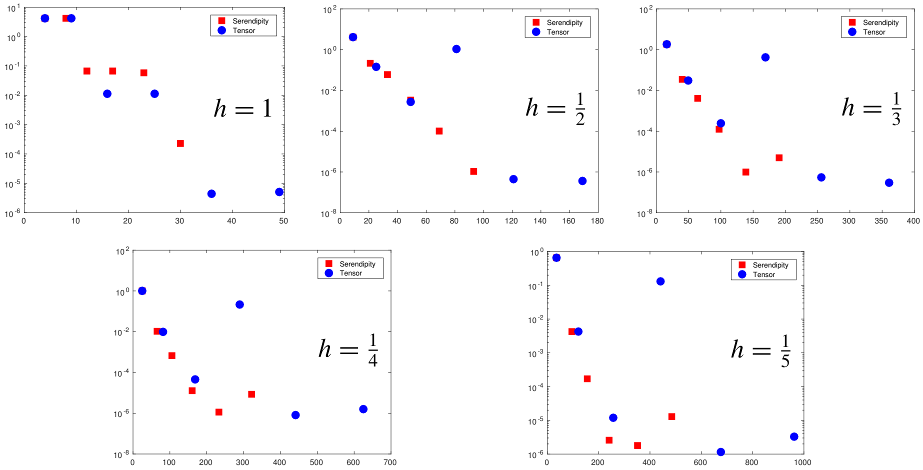


Figure 1. Square domain, Neumann conditions, $\lambda = 2\pi^2$, p-refinement experiments.

eigenvalue which differs by about 10^{-6} from the theoretical using both serendipity and tensor elements. However, when using the serendipity elements, we see a reduction in the number of degrees of freedom by approximately half compared to the tensor product element.

We see similar behavior in the Dirichlet problem, depicted in Figure 2, with the obvious difference of an overall reduction in the number of degrees of freedom. Note that since we remove the degrees of freedom corresponding to the boundary, not discretizing the mesh at all (i.e., when $h = 1$) results in having too few equations to properly solve for a nonzero eigenvalue for small p.

When we consider the Neumann problem with p fixed and h varied, we see the results depicted in Figure 3. In contrast to the previously discussed results, we see that, in nearly every case, the tensor product elements achieve better accuracy than serendipity while using fewer degrees of freedom. The only exception is when $p = 4$, also depicted in Figure 3. Here, we note a large increase in error when using tensor product elements. This effect can be seen in nearly every plot for h-refinements and accounts for the large jumps in the tensor product results where h is fixed. The reason for this error was undetermined in our experiments, but will be revisited when the L-shaped results are discussed. We see the same behavior for the Dirichlet problem in Figure 4. In results not displayed here, we analyze p- and h-refinements in approximating the nonsimple Neumann eigenvalue $5\pi^2$. The results are qualitatively similar to the previously discussed results.

We also note strange behavior when using elements of order 5 and 6. Exhibited in the Neumann case on the square in Figure 3, $p = 5$ and $p = 6$, we see that as we

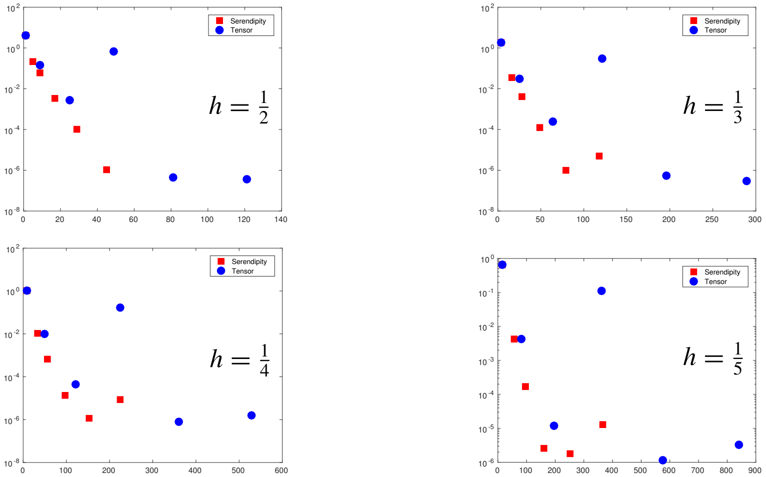


Figure 2. Square domain, Dirichlet conditions, $\lambda = 2\pi^2$, p-refinement experiments.

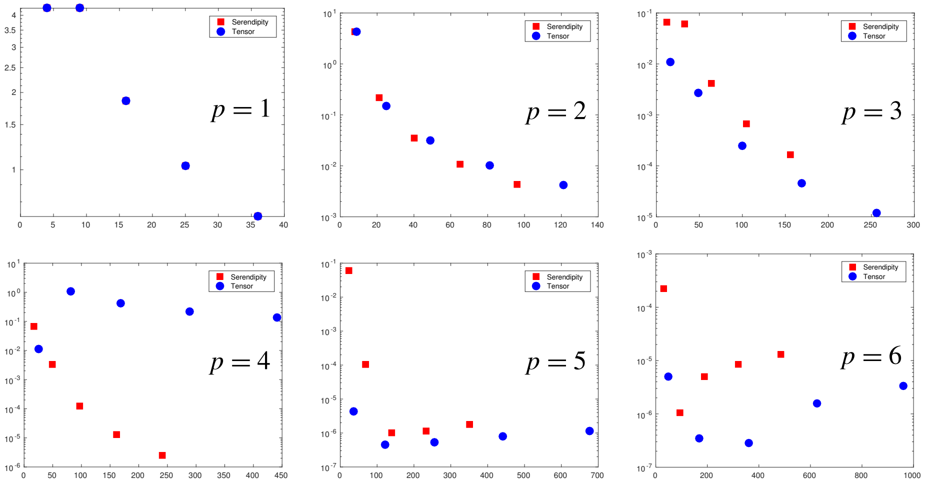


Figure 3. Square domain, Neumann conditions, $\lambda = 2\pi^2$, h-refinement experiments.

refine our mesh further, the error increases. The error sometimes increases higher than lower-order elements solving the same problem, as seen in many of the plots when h is fixed; the trend in error seems to “flair up” towards the end. The reason for this behavior is likely due to numerical roundoff errors.

L-shaped domain. On the L-shaped domain, we see in Figures 5–8 nearly the same patterns described above when approximating the eigenvalue $2\pi^2$. We note that

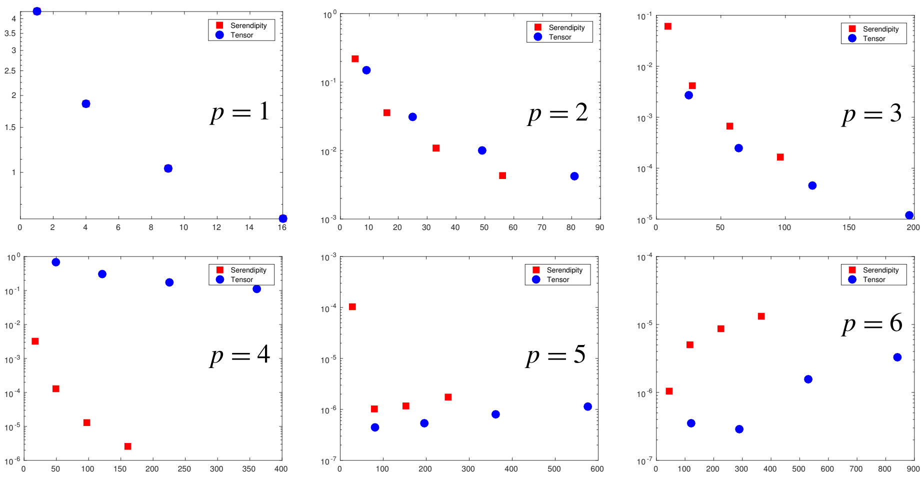


Figure 4. Square domain, Dirichlet conditions, $\lambda = 2\pi^2$, h-refinement experiments.

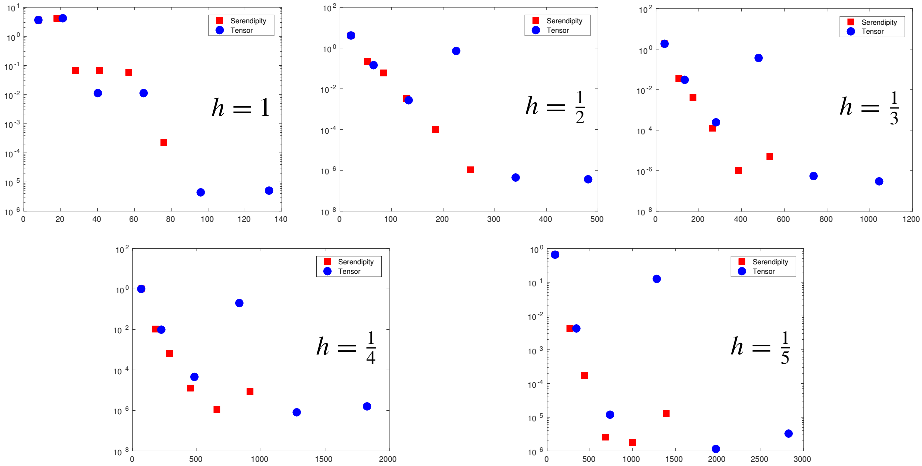


Figure 5. L-shaped domain, Neumann conditions, $\lambda = 2\pi^2$, p-refinement experiments.

when h is fixed, the savings achieved by serendipity elements is increased even further. For example, with Neumann boundary conditions and $h = \frac{1}{4}$ (Figure 5), for the $p = 5$ case, both the serendipity and tensor product elements exhibit an error of about 10^{-6}. The number of degrees of freedom used in the serendipity case however is less than half of that of the tensor case. With Dirichlet boundary conditions as seen in Figure 6, $h = \frac{1}{4}$, this savings is further increased, with serendipity elements using nearly a third of the degrees of freedom used by tensor product elements.

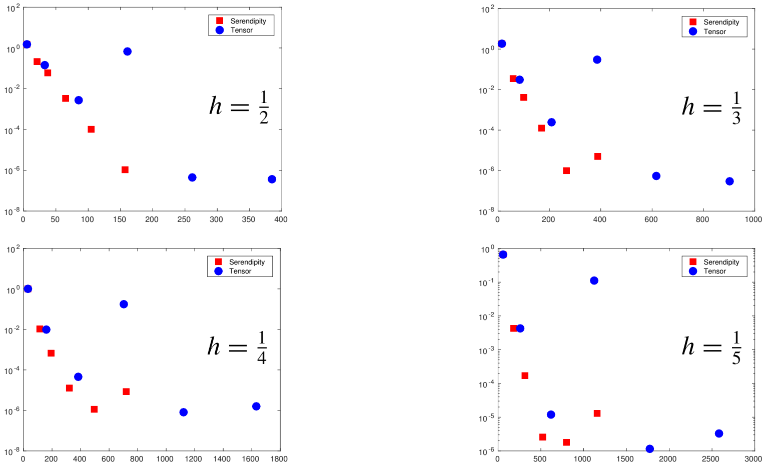


Figure 6. L-shaped domain, Dirichlet conditions, $\lambda = 2\pi^2$, p-refinement experiments.

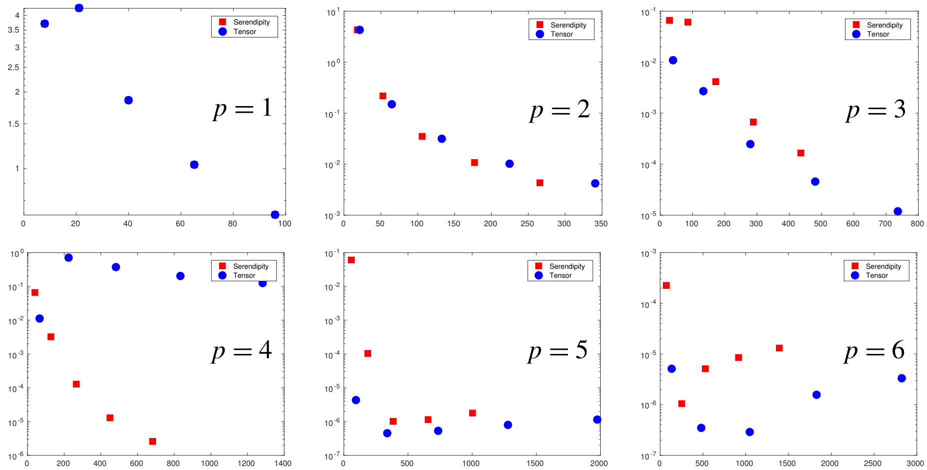


Figure 7. L-shaped domain, Neumann conditions, $\lambda = 2\pi^2$, h-refinement experiments.

In addition to the plots described above, we have also added plots depicting the results of approximating the Neumann eigenvalue numerically approximated as 1.4756218450. Figure 9 and Figure 10 show that these results mostly correspond to the previously exhibited behavior with the exception that in Figure 9, the tensor product elements also achieve better approximations when refining p. We also note that in Figure 10, $p = 4$, the order-4 tensor product elements have a large decrease

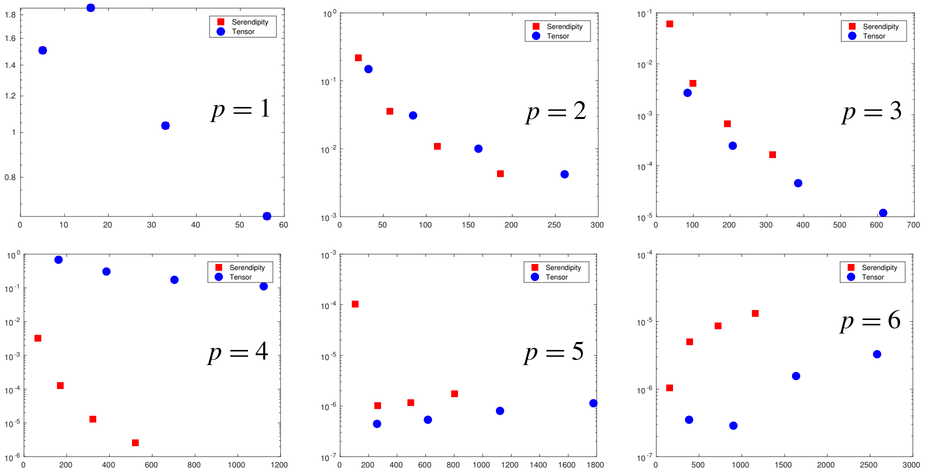


Figure 8. L-shaped domain, Dirichlet conditions, $\lambda = 2\pi^2$, h-refinement experiments.

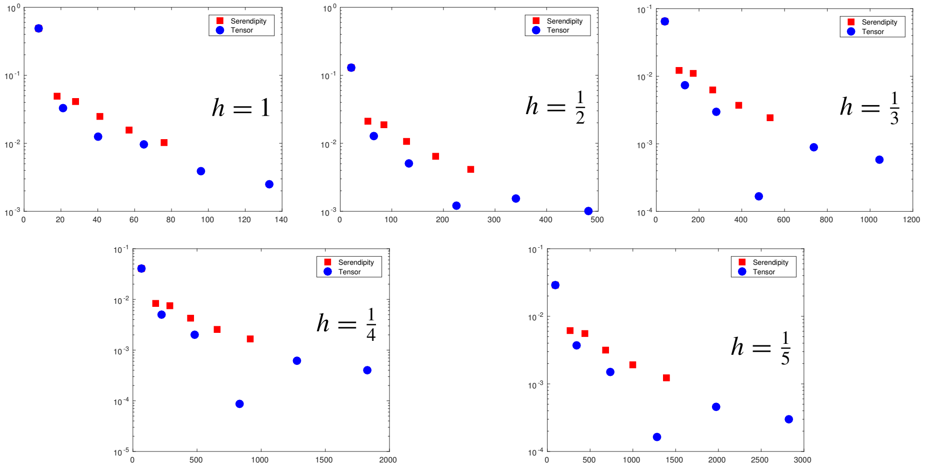


Figure 9. L-shaped domain, Neumann conditions, $\lambda = 1.4756218450$, p-refinement experiments.

in error. This behavior contrasts the increase we saw when approximating $2\pi^2$ with order-4 tensor product elements over the square and is, again, unexplained.

Spectrum comparison. We also compare the spectrum of eigenvalues that are computed by the tensor product and serendipity elements on the square versus the theoretical spectrum. The results are shown in Figure 11. We see that the eigenvalues calculated by the tensor product and serendipity elements are nearly

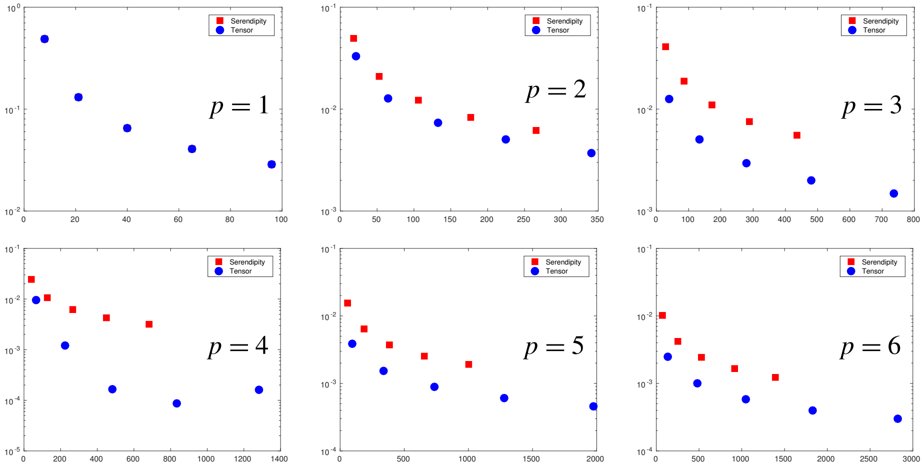


Figure 10. L-shaped domain, Neumann conditions, $\lambda = 1.4756218450$, h-refinement experiments.

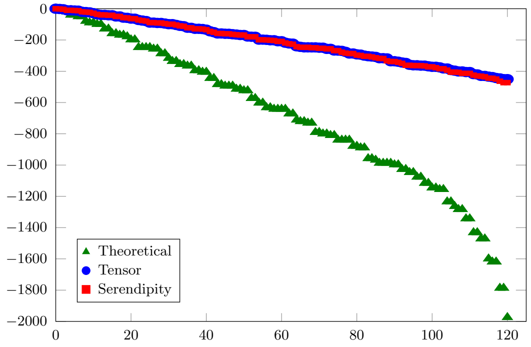


Figure 11. Spectra for $p = 3$, $h = \frac{1}{5}$, tensor, and serendipity bases along with theoretical spectrum over square domain.

the same, and, as expected, as we attempt to approximate larger eigenvalues, the results become less accurate.

5. Conclusion and future directions

A key takeaway message from our numerical experiments is that when seeking eigenvalue estimates on a fixed mesh of squares, serendipity elements do appear to fulfill their promise of producing as accurate a result as tensor product elements, despite having roughly 50% the number of degrees of freedom. Since many application contexts require a fixed domain mesh, it would be advantageous computationally to use serendipity elements in such circumstances.

Various additional experiments are planned. First, there are questions in regards to differing behavior on the square versus the L-shaped domain, and the Neumann versus Dirichlet boundary conditions. A study of serendipity elements for the Poisson equation (i.e., with nonzero boundary conditions) or for more general eigenvalue problems might help explain our results. A second issue is to resolve the dramatic aberrations in the results for the case of tensor product basis functions for the case $p = 4$. Further investigation into the pattern observed in the convergence behavior depending on mesh discretization is in progress.

We also plan to investigate the observation that mesh discretization for high degree polynomial basis functions sometimes results in less accurate approximations. We suspect that this arises from numerical roundoff errors, as the results became worse only after reaching a threshold on the order of 10^{-8}.

As discussed in the Mathematica code accompanying this paper, similar constructions for serendipity basis functions in three dimensions were also determined. In future work, we plan to extend the implementation of our finite element solver to allow for three-dimensional domains, and implement these three-dimensional serendipity basis functions in order to produce similar analysis and comparisons as those that we have found for two dimensions.

Appendix: Serendipity basis functions

The following are the serendipity element basis functions in two-dimensions from order 1 to 4. The basis functions are organized as they are calculated in Mathematica, i.e., as the sum of reindexed arrays of basis functions as discussed in Section 3. The Mathematica code that was used to generate these functions is available in the online supplement:

$$\mathcal{S}_1\Lambda^0(\square_2)\text{ basis} = \begin{pmatrix} \frac{1}{4}(1-x)(1-y) & \frac{1}{4}(1-x)(y+1) \\ \frac{1}{4}(x+1)(1-y) & \frac{1}{4}(x+1)(y+1) \end{pmatrix},$$

$\mathcal{S}_2\Lambda^0(\square_2)$ basis $=$

$$\begin{pmatrix} -\frac{1}{4}(x-1)(y-1)(x+y+1) & \frac{1}{2}(x-1)(y^2-1) & \frac{1}{4}(x-1)(x-y+1)(y+1) \\ \frac{1}{2}(x^2-1)(y-1) & 0 & -\frac{1}{2}(x^2-1)(y+1) \\ \frac{1}{4}(y-1)(-x^2+yx+y+1) & -\frac{1}{2}(x+1)(y^2-1) & \frac{1}{4}(x+1)(y+1)(x+y-1) \end{pmatrix},$$

$\mathcal{S}_3\Lambda^0(\square_2)$ basis $=$

$$\begin{pmatrix} \frac{1}{4}(x-1)(y-1)A_3 & \frac{1}{2}(x-1)(y^2-1) & \frac{1}{2}(x-1)y(y^2-1) & -\frac{1}{4}(x-1)(y+1)A_3 \\ \frac{1}{2}(x^2-1)(y-1) & 0 & 0 & -\frac{1}{2}(x^2-1)(y+1) \\ \frac{1}{2}x(x^2-1)(y-1) & 0 & 0 & \frac{1}{2}(x-x^3)(y+1) \\ -\frac{1}{4}(x+1)(y-1)A_3 & -\frac{1}{2}(x+1)(y^2-1) & \frac{1}{2}(x+1)(y-y^3) & \frac{1}{4}(x+1)(y+1)A_3 \end{pmatrix},$$

where

$$A_3 = x^2 + y^2 - 1;$$

$\mathcal{S}_4\Lambda^0(\square_2)$ basis =

$$\begin{pmatrix} B_{11} & B_{12} & B_{13} & B_{14} & B_{15} \\ \frac{1}{2}(x^2-1)(x^2-y)(y-1) & (x^2-1)(y^2-1) & 0 & 0 & -\frac{1}{2}(x^2-1)(y+1)(x^2+y) \\ \frac{1}{2}x(x^2-1)(y-1) & 0 & 0 & 0 & \frac{1}{2}(x-x^3)(y+1) \\ \frac{1}{4}(x-1)x^2(x+1)(y-1) & 0 & 0 & 0 & -\frac{1}{4}(x-1)x^2(x+1)(y+1) \\ B_{51} & B_{52} & B_{53} & B_{54} & B_{55} \end{pmatrix},$$

where

$$\begin{aligned} B_{11} &= -\tfrac{1}{4}(x-1)(y-1)(x^3-(y+1)x+y(y^2-1)), \\ B_{12} &= \tfrac{1}{2}(y^2-1)(-x^2+y^2x+x-y^2), \\ B_{13} &= \tfrac{1}{2}(x-1)y(y^2-1), \\ B_{14} &= \tfrac{1}{4}(x-1)(y-1)y^2(y+1), \\ B_{15} &= \tfrac{1}{4}(x-1)(y+1)(x^3+(y-1)x-y^3+y), \\ B_{51} &= \tfrac{1}{4}(x+1)(y-1)(-x^3+yx+x+y^3-y), \\ B_{52} &= -\tfrac{1}{2}(y^2-1)(x^2+y^2x+x+y^2), \\ B_{53} &= \tfrac{1}{2}(x+1)(y-y^3), \\ B_{54} &= -\tfrac{1}{4}(x+1)(y-1)y^2(y+1), \\ B_{55} &= \tfrac{1}{4}(x+1)(y+1)(x^3+(y-1)x+y(y^2-1)). \end{aligned}$$

Acknowledgements

All authors were supported in part by NSF Award 1522289.

References

[Arnold and Awanou 2011] D. Arnold and G. Awanou, "The serendipity family of finite elements", *Found. Comput. Math.* **11**:3 (2011), 337–344. MR Zbl

[Arnold and Logg 2014a] D. Arnold and A. Logg, "Periodic table of the finite elements", electronic reference, 2014, available at http://femtable.org.

[Arnold and Logg 2014b] D. Arnold and A. Logg, "Periodic table of the finite elements", *SIAM News* **47**:9 (2014), 9 pp.

[Arnold et al. 2015] D. Arnold, D. Boffi, and F. Bonizzoni, "Finite element differential forms on curvilinear cubic meshes and their approximation properties", *Numer. Math.* **129**:1 (2015), 1–20. MR Zbl

[Boffi 2010] D. Boffi, "Finite element approximation of eigenvalue problems", *Acta Numer.* **19** (2010), 1–120. MR Zbl

[Brenner and Scott 1994] S. C. Brenner and L. R. Scott, *The mathematical theory of finite element methods*, Texts in Applied Mathematics **15**, Springer, New York, 1994. MR Zbl

[Ciarlet 1978] P. G. Ciarlet, *The finite element method for elliptic problems*, Studies in Mathematics and its Applications **40**, North-Holland, Amsterdam, 1978. MR Zbl

[Dauge 2003] M. Dauge, "Benchmark computations for Maxwell equations for the approximation of highly singular solutions", electronic reference, 2003, available at http://tinyurl.com/daugebench.

[Floater and Gillette 2017] M. S. Floater and A. Gillette, "Nodal bases for the serendipity family of finite elements", *Found. Comput. Math.* **17**:4 (2017), 879–893. MR

[Hackbusch 1979] W. Hackbusch, "On the computation of approximate eigenvalues and eigenfunctions of elliptic operators by means of a multi-grid method", *SIAM J. Numer. Anal.* **16**:2 (1979), 201–215. MR Zbl

[Hughes 1987] T. J. R. Hughes, *The finite element method*, Prentice Hall, Englewood Cliffs, NJ, 1987. MR Zbl

[Mandel 1990] J. Mandel, "Iterative solvers by substructuring for the p-version finite element method", *Comput. Methods Appl. Mech. Engrg.* **80**:1-3 (1990), 117–128. MR Zbl

[Strang and Fix 1973] G. Strang and G. J. Fix, *An analysis of the finite element method*, Prentice Hall, Englewood Cliffs, NJ, 1973. MR Zbl

[Szabó and Babuška 1991] B. Szabó and I. Babuška, *Finite element analysis*, Wiley, New York, 1991. MR Zbl

[Wang et al. 1996] Y. Wang, P. Monk, and B. Szabó, "Computing cavity modes using the p-version of the finite element method", *IEEE Trans. Magnetics* **32**:3 (1996), 1934–1940.

Received: 2017-04-17 Accepted: 2017-07-22

agillette@math.arizona.edu *Department of Mathematics, University of Arizona,*
Tucson, AZ, United States

grosscra@msu.edu *Department of Mathematics, University of Arizona,*
Tucson, AZ, United States

plackow1@math.arizona.edu *Department of Mathematics, University of Arizona,*
Tucson, AZ, United States

INVOLVE 11:4 (2018)
dx.doi.org/10.2140/involve.2018.11.679

Connectedness of two-sided group digraphs and graphs

Patreck Chikwanda, Cathy Kriloff, Yun Teck Lee,
Taylor Sandow, Garrett Smith and Dmytro Yeroshkin

(Communicated by Ann N. Trenk)

Two-sided group digraphs and graphs, introduced by Iradmusa and Praeger, provide a generalization of Cayley digraphs and graphs in which arcs are determined by left and right multiplying by elements of two subsets of the group. We characterize when two-sided group digraphs and graphs are weakly and strongly connected and count connected components, using both an explicit elementary perspective and group actions. Our results and examples address four open problems posed by Iradmusa and Praeger that concern connectedness and valency. We pose five new open problems.

1. Introduction

Two-sided group digraphs were introduced as a generalization of Cayley digraphs by Iradmusa and Praeger [2016] and independently in [Anil Kumar 2012]; see [Iradmusa and Praeger 2016, Remark 1.6]. Given a group G and a subset S of G, the *Cayley digraph* $\mathrm{Cay}(G, S)$ has the elements of G as vertices and a directed arc from g to h when $gh^{-1} \in S$. Several authors have generalized this idea by relaxing the group conditions or the nature of the multiplication; see [Annexstein et al. 1990; Marušič et al. 1992; Gauyacq 1997; Kelarev and Praeger 2003]. The *two-sided group digraph* $2\mathrm{S}(G; L, R)$ also has elements of a group G as vertices, but two nonempty subsets, L and R, of G are used to define an arc from vertex g to vertex h in G when $h = l^{-1}gr$ for some $l \in L$ and $r \in R$. As with Cayley digraphs, by definition $2\mathrm{S}(G; L, R)$ does not have multiple arcs between two vertices, even though it is possible that $l_1^{-1}gr_1 = l_2^{-1}gr_2$ for $l_1 \neq l_2$ and $r_1 \neq r_2$ (see Section 2). A Cayley digraph is undirected when $S = S^{-1}$ and the digraph $2\mathrm{S}(G; L, R)$ is undirected when $L^{-1}gR = LgR^{-1}$ for all $g \in G$, but we do not assume this.

MSC2010: primary 05C25; secondary 05C20, 05C40.

Keywords: two-sided group digraph, Cayley graph, group, connectivity.

Lee, Sandow and Smith were supported by Idaho State University Career Path Internship funds.

It is worth noting that a continuous version of a two-sided group digraph has previously appeared in the context of Riemannian geometry as the study of biquotients. Introduced in [Gromoll and Meyer 1974], biquotients are viewed as the quotient space of a two-sided Lie group action and have been studied systematically as a source of manifolds with positive and nonnegative curvature since the work of Eschenburg [1982; 1984]. We refer to [DeVito 2011] for a broader overview of the topic.

Iradmusa and Praeger explored several properties of two-sided group digraphs and posed eight open problems. Here we address the first four problems, which concern valency and connectedness. It would also be of interest to know whether there exist vertex-transitive two-sided group digraphs that are not isomorphic to Cayley digraphs since these would have potential applications to routing and communication schemes in interconnection networks. Indeed, the remaining unresolved questions in [Iradmusa and Praeger 2016] primarily address understanding when two-sided group digraphs are vertex-transitive and when they are isomorphic to Cayley digraphs. In addition, we propose five new problems related to our results below.

Our main focus is to generalize [Iradmusa and Praeger 2016, Theorem 1.8], which gives necessary and sufficient conditions for a two-sided group digraph $2S(G; L, R)$ to be connected, assuming that L and R are inverse-closed. Theorem 2.4 solves Problem 4 in [loc. cit.] by characterizing when $2S(G; L, R)$ is connected without the inverse-closed assumption on L and R. Examples 2.5 through 2.8 in Section 2 both illustrate Theorem 2.4 and address Problems 1 and 2 in [loc. cit.] by showing that it is possible for $2S(G; L, R)$ to have constant out-valency but not constant in-valency and to be regular of valency strictly less than $|L| \cdot |R|$.

In Section 3B, building on results in Section 3A, we use elementary methods similar to those in our proof of Theorem 2.4 to generalize further. In Theorems 3.13 and 3.16, under the assumption that elements in G can be factored appropriately, we count weakly and strongly connected components, show such components must all be of the same size, and characterize their vertices. The result that all components have the same size addresses Problem 3 of [loc. cit.]. We also show that the connected components are in fact isomorphic under a condition on the normalizers of L and R. To illustrate we provide Corollaries 3.15 and 3.17 that give simple characterizations of weak and strong connectedness and give Example 3.18 in which components are isomorphic and Example 3.19 in which they are not.

In Section 4 we drop the factorization assumptions and note that connected components are contained within double cosets. Results analogous to those in Section 3B apply within a given double coset and examples demonstrate that in different double cosets the sizes of the connected components can differ.

A less explicit but more natural approach to counting strongly connected components is to view the components as orbits under a group action and to use a standard result that counts orbits. This is done in Section 5.

In Section 6 we prove that when G is a semidirect product, $G = H \rtimes K$, it is possible to determine whether $2\mathrm{S}(G; L, R)$ is connected by analyzing connectedness properties related to H in $2\mathrm{S}(G; L, R)$ and a two-sided group digraph on K. We also generalize this to the case where K is G/H for H a normal subgroup of G.

2. Preliminaries

Following some definitions, we begin with an initial result that characterizes when a two-sided group digraph is strongly connected. After some examples we compare Theorem 2.4 to [Iradmusa and Praeger 2016, Theorem 1.8].

Recall the following definition from [loc. cit.].

Definition 2.1. For nonempty subsets L and R of a group G, a *two-sided group digraph* $2\mathrm{S}(G; L, R)$ has vertex set G and a directed arc (g, h) from g to h if and only if $h = l^{-1}gr$ for some $l \in L$ and $r \in R$.

The digraph $2\mathrm{S}(G; L, R)$ is undirected when $L^{-1}gR = LgR^{-1}$ for all $g \in G$, but we work in the generality of directed graphs and consider this situation to be a special case.

Definition 2.2. Let S be a nonempty subset of a group G. A *word in S of (finite) length* $n > 0$ is a string $s_1 s_2 \cdots s_n$, where $s_1, s_2, \ldots, s_n \in S$. In general, we denote a word in S of length n by $w_{S,n}$ and write $\mathcal{W}(S)$ for the set containing all finite-length words in S.

Note that the factors in a word need not be distinct, a single group element will have numerous different representations as a word in S, and different words will be denoted by varying subscripts for the set or length on the letter w.

Definition 2.3. If g and h are vertices in a digraph, then g is *strongly connected* to h if there exists a directed path from g to h and a directed path from h to g. A digraph is *strongly connected* if every pair of vertices is strongly connected.

Theorem 2.4. *The two-sided group digraph* $2\mathrm{S}(G; L, R)$ *is strongly connected if and only if* $G = \mathcal{W}(L^{-1})\mathcal{W}(R) = \mathcal{W}(L)\mathcal{W}(R^{-1})$ *and the identity element e satisfies* $e = w_{L^{-1},i+1} w_{R,i} = w_{L^{-1},j} w_{R,j+1}$ *for some* $i, j \in \mathbb{N}$.

Proof. Assume that the two-sided group digraph $2\mathrm{S}(G; L, R)$ is strongly connected. Then given any $g \in G$, there exists a directed path from the identity element e to g, meaning $g = w_{L^{-1},n} e w_{R,n} = w_{L^{-1},n} w_{R,n}$. Hence $g \in \mathcal{W}(L^{-1})\mathcal{W}(R)$ and $G = \mathcal{W}(L^{-1})\mathcal{W}(R)$. Since there also exists a directed path from g to e, we know $e = w_{L^{-1},m} g w_{R,m}$, which implies that $g = w_{L^{-1},m}^{-1} w_{R,m}^{-1} = w_{L,m} w_{R^{-1},m}$, and hence $G = \mathcal{W}(L)\mathcal{W}(R^{-1})$. In particular there exists a directed path from l^{-1} to e, where $l \in L$, and hence $e = w_{L^{-1},i} l^{-1} w_{R,i} = w_{L^{-1},i+1} w_{R,i}$ for some i. Similarly, $e = w_{L^{-1},j} w_{R,j+1}$ since there is a directed path from r to e, where $r \in R$.

Conversely, suppose that $G = \mathcal{W}(L^{-1})\mathcal{W}(R) = \mathcal{W}(L)\mathcal{W}(R^{-1})$ and the identity element e satisfies $e = w_{L^{-1},\,i+1}w_{R,i} = w_{L^{-1},\,j}w_{R,j+1}$ for some $i, j \in \mathbb{N}$. It suffices to show that there is a directed path from e to g and from g to e for all $g \in G$; i.e., $g = w_{L^{-1},\,m}w_{R,m} = w_{L,n}w_{R^{-1},n}$ for some $m, n \in \mathbb{N}$.

Since $G = \mathcal{W}(L^{-1})\mathcal{W}(R)$, we know g has an $L^{-1}R$ factorization; i.e., $g = w_{L^{-1},\,a}w_{R,b}$ for some $a, b \in \mathbb{N}$. If $a \neq b$, it is possible to adjust the $L^{-1}R$ factorization of g so that both words have the same length by inserting the appropriate factorization of e between the words from L^{-1} and R. For example, if $a > b$, then insert $e = w_{L^{-1},\,j}w_{R,j+1}$ to obtain

$$g = w_{L^{-1},\,a}w_{R,b} = w_{L^{-1},\,a}(w_{L^{-1},\,j}w_{R,j+1})w_{R,b} = w_{L^{-1},\,a+j}w_{R,b+j+1}.$$

Repeating this process yields $g = w_{L^{-1},\,m}w_{R,m}$, where $m = a + (a-b)j$.

To see that g also has an LR^{-1} factorization with words of the same length, note that left and right multiplying by inverses of the words from L^{-1} and R respectively converts $e = w_{L^{-1},\,i+1}w_{R,i} = w_{L^{-1},\,j}w_{R,j+1}$ into $e = w_{L,i+1}w_{R^{-1},i} = w_{L,j}w_{R^{-1},j+1}$. Repeatedly inserting the appropriate LR^{-1} factorization of e into an LR^{-1} factorization of g shows $g = w_{L,n}w_{R^{-1},n}$ for any $g \in G$. Hence $2\mathrm{S}(G; L, R)$ is strongly connected. □

The following examples illustrate Theorem 2.4 and also address the first two problems posed in [Iradmusa and Praeger 2016].

Example 2.5. Consider $\Gamma = 2\mathrm{S}(A_4; L, R)$, where A_4 is the alternating group on four elements, $L = \{e, (243)\}$, and $R = \{(234), (12)(34), (132), (14)(23)\}$, as shown in Figure 1. Since G is generated by words in R or R^{-1}, we have $G = \mathcal{W}(L^{-1})\mathcal{W}(R) = \mathcal{W}(L)\mathcal{W}(R^{-1})$. Also $e = e^3 \cdot [(12)(34)]^2 = e \cdot [(12)(34)]^2$, so the hypotheses of Theorem 2.4 hold and thus Γ is strongly connected.

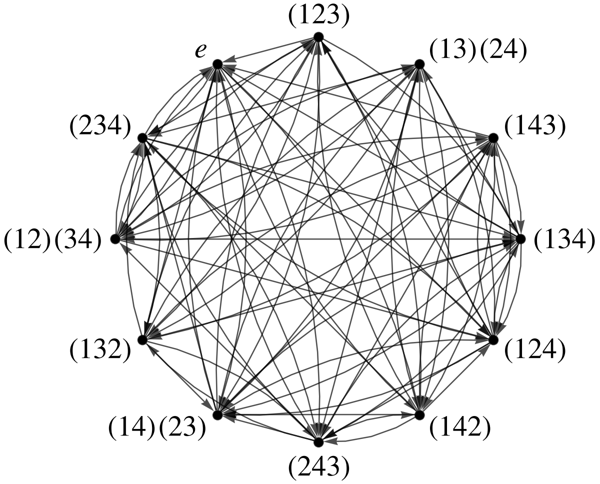


Figure 1. $2\mathrm{S}(A_4; \{e, (243)\}, \{(234), (12)(34), (132), (14)(23)\})$.

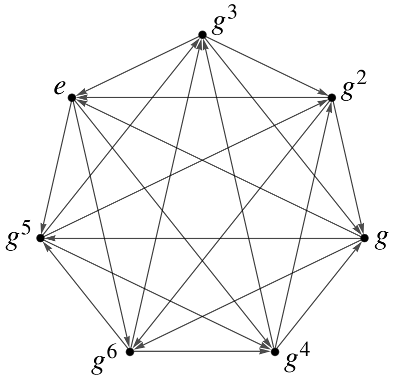


Figure 2. $2\mathrm{S}(C_7; \{g^2, g^3\}, \{e, g\})$.

This example addresses Problem 1 of [Iradmusa and Praeger 2016], which asks whether or not $2\mathrm{S}(G; L, R)$ can have constant out-valency but not constant in-valency. The digraph Γ has constant out-valency of 7; however the vertices $\{(123), (132), (142), (143), (12)(34), (13)(24)\}$ have in-valency 6 and the vertices $\{e, (234), (243), (134), (124), (14)(23)\}$ have in-valency 8. Furthermore $2\mathrm{S}(A_4; L^{-1}, R^{-1})$ will have constant in-valency of 7 and out-valency of either 6 or 8 for the same sets as in $2\mathrm{S}(G; L, R)$ because inverting L and R changes the direction of each edge.

Problem 1. For $2\mathrm{S}(G; L, R)$ with constant out-valency, what are the possible sets of in-valencies? In particular, how large can they be and how much can they differ from the out-valency?

Example 2.6. The two-sided group digraph $2\mathrm{S}(C_7; \{g^2, g^3\}, \{e, g\})$, where C_7 is the cyclic group of order 7 generated by g, satisfies the hypotheses of Theorem 2.4 and is connected. This example also addresses Problem 2 of [Iradmusa and Praeger 2016], which asks whether or not $2\mathrm{S}(G; L, R)$ can be a regular graph of valency strictly less than $|L| \cdot |R|$. Here $|L| \cdot |R| = 4$, but as seen in Figure 2, $2\mathrm{S}(C_7; \{g^2, g^3\}, \{e, g\})$ is regular with valency 3. In fact $2\mathrm{S}(C_7; \{g^2, g^3\}, \{e, g\}) \cong \mathrm{Cay}(C_7, \{g^4, g^5, g^6\})$, with g^5 arising in two different ways from the sets L^{-1} and R, explaining the valency of 3.

Example 2.7. Consider the dihedral group D_6 of order 12, generated by the reflection τ and the rotation σ of order 6. The undirected graph $2\mathrm{S}(D_6; \{\tau, \tau\sigma^5\}, \{\tau\sigma, \tau\sigma^2\})$ is regular of valency 3 and $|L| \cdot |R| = 4$ as in Example 2.6, but for a nonabelian group. For any $g \in D_6$, the set $(LL^{-1})^g \cap (RR^{-1}) = \{e, \sigma, \sigma^{-1}\}$ is of size 3 and there is a reduction in valency by 1. See Figure 3 in Section 3B.

Example 2.8. The two-sided group digraph $2\mathrm{S}(A_4; A_4, \{(243), (12)(34)\})$ has $|L| \cdot |R| = 24$, but is in fact regular with valency 12 and forms a complete undirected graph with loops. Here $(LL^{-1})^g \cap (RR^{-1}) = \{e, (124), (142)\}$ is of size 3 for each g, but the reduction in valency is much larger than in the previous examples because $L^{-1}gR$, viewed as a multiset, consists of 12 distinct elements, each with multiplicity 2.

The set $(LL^{-1})^g \cap (RR^{-1})$, where $(LL^{-1})^g = g^{-1}LL^{-1}g$, is introduced in [Iradmusa and Praeger 2016, Definition 1.4(3) and Theorem 1.5], and the condition $(LL^{-1})^g \cap (RR^{-1}) = \{e\}$ is shown in Lemma 3.1 of that paper to guarantee the valency of $2\mathrm{S}(G; L, R)$ is exactly $|L| \cdot |R|$. In Examples 2.6, 2.7, and 2.8 it is the failure of this condition which causes a drop in valency. In general, if for $h \neq e$, we have $h = g^{-1}l_1l_2^{-1}g = r_1r_2^{-1}$ for some $l_1, l_2 \in L$ and some $r_1, r_2 \in R$, then $l_1^{-1}gr_1 = l_2^{-1}gr_2$, which causes a multiplicity greater than 1 in $L^{-1}gR$ considered as a multiset. Since the elements and their multiplicities in the multiset $L^{-1}gR$ depend on g, we did not search for necessary and sufficient conditions on L and R for a two-sided group digraph to have valency strictly less than $|L| \cdot |R|$, as requested in Problem 2 of [Iradmusa and Praeger 2016]. Thus this aspect of their Problem 2 remains unresolved.

Theorem 2.4 is a generalization of the first part of the following result in [Iradmusa and Praeger 2016] and also addresses Problem 4 of that paper.

Theorem 2.9 [Iradmusa and Praeger 2016, Theorem 1.8]. *Let L and R be nonempty inverse-closed subsets of a group G, and let $\Gamma = 2\mathrm{S}(G; L, R)$:*

(1) *The graph Γ is connected if and only if $G = \langle L\rangle\langle R\rangle$ and there exist words in L and R with lengths of opposite parity whose product is e.*

(2) *If $G = \langle L\rangle\langle R\rangle$ and there do not exist words in L and R with lengths of opposite parity whose product is e then Γ is disconnected with exactly two connected components.*

Theorems 3.13 and 3.16 further generalize [Iradmusa and Praeger 2016, Theorem 1.8] by providing more general counts and characterizations of connected components. Theorem 3.13 also answers Problem 3 of that paper, by showing there cannot exist G, L, and R satisfying the hypotheses of Theorem 2.9 such that $G = \langle L\rangle\langle R\rangle$ but $2\mathrm{S}(G; L, R)$ has connected components of different sizes. We show more generally that if $G = \mathcal{W}(L \cup L^{-1})\mathcal{W}(R \cup R^{-1})$ then all connected components of $2\mathrm{S}(G; L, R)$ have the same size.

3. General connectedness results

3A. ***Connection length.*** In this section we lay the foundation for studying both weakly and strongly connected components of $2\mathrm{S}(G; L, R)$ in Section 3B.

Definition 3.1. In a digraph, a vertex g is *weakly connected* to vertex h if there is a path $g_0, g_1, \ldots, g_n$ such that $g = g_0$, $h = g_n$, and either (g_{i-1}, g_i) or (g_i, g_{i-1})

is an arc of the digraph. A digraph is *weakly connected* if each pair of its vertices is weakly connected.

If L and R are nonempty subsets of a group G, we let $\bar{L} = L \cup L^{-1}$ and $\bar{R} = R \cup R^{-1}$ and use $w_{\bar{L},m,a}$ to denote a word that contains m factors from L and a factors from L^{-1} in any order. The notation $g \sim h$ will mean g is weakly connected to h in $2\mathrm{S}(G; L, R)$, or equivalently $h = W_{\bar{L},m,a} g W_{\bar{R},a,m}$, where the capital W indicates that the corresponding factors on either side of g have opposite signs; i.e., one factor is from L^{-1} and one is from R, or alternatively, one is from L and one is from R^{-1}. If computations lead to factorizations that may not involve opposite signs on corresponding factors then W is changed to w.

We begin with two key results which will allow us to define minimum weak connection length in Definition 3.4 and which will also be used in the proof of Theorem 3.13.

Lemma 3.2. *In $2\mathrm{S}(G; L, R)$ if $g = w_{\bar{L},m,a} w_{\bar{R},n,b}$, then $g \sim l^d$ and $g \sim r^d$, where l is any element of L, r is any element of R, and $d = m + n - (a + b)$.*

Proof. Let $g = w_{\bar{L},m,a} w_{\bar{R},n,b}$ for $a, b, m, n \in \mathbb{N}$. Then we have for any $r \in R$ and $l \in L$,

$$\begin{aligned} g = w_{\bar{L},m,a} w_{\bar{R},n,b} &= w_{\bar{L},m,a} w_{\bar{R},n,b} r^{m-a} r^{-m+a} \\ &= W_{\bar{L},m,a} w_{\bar{R},m+n,a+b} W_{\bar{R},a,m} \\ &= W_{\bar{L},m,a} l^{a+b-(m+n)} l^{m+n-(a+b)} w_{\bar{R},m+n,a+b} W_{\bar{R},a,m} \\ &= W_{\bar{L},m,a} W_{\bar{L},a+b,m+n} l^{m+n-(a+b)} W_{\bar{R},m+n,a+b} W_{\bar{R},a,m}. \end{aligned}$$

Corresponding factors can be adjusted to have opposite signs because the repeated r and r^{-1} and l and l^{-1} can be rearranged as needed. A similar construction yields $g \sim r^d$. □

The following corollary is stated in terms of L, but an analogous statement in terms of R also holds.

Corollary 3.3. *Let L and R be nonempty subsets of a group G:*

(1) *In $2\mathrm{S}(G; L, R)$ there exist two words in L of different lengths that are weakly connected if and only if there is a word in L that is weakly connected to e.*

(2) *In $2\mathrm{S}(G; L, R)$ there exists a word $w_{L,n}$ weakly connected to e if and only if there exists a word $w_{L^{-1},n}$ weakly connected to e.*

Proof. For (1), if $w_{L,m} \sim w_{L,n}$, assume without loss of generality that $m < n$, left multiply by $w_{L,m}^{-1}$, and apply Lemma 3.2 to obtain $e \sim l^{n-m}$ for $l \in L$. Conversely if $e \sim w_{L,m}$, left multiply by some $l \in L$ and apply Lemma 3.2.

For (2), if $e = W_{\bar{L},m,a} w_{L,n} W_{\bar{R},a,m}$ then $e = w_{L,n}^{-1} w_{\bar{L},a,m} w_{\bar{R},m,a}$. Now apply Lemma 3.2 to obtain $e \sim (l^{-1})^n$ for $l \in L$. The converse is achieved analogously. □

These results yield that the following notion is well-defined.

Definition 3.4. The *minimum weak connection length in G relative to* (L, R) is the minimum length k of a word purely in L, L^{-1}, R, or R^{-1} that is weakly connected to e, and is infinite if there is no such minimum. Algebraically this is equivalent to the minimum length of a word w purely in L, L^{-1}, R, or R^{-1} such that $e = W_{\bar{L},m,a} w W_{\bar{R},a,m}$ for some $a, m \in \mathbb{N}$.

Here and in the next section we impose the additional assumption that the set of words in L and the set of words in R are subgroups of G in order to adapt weak connectedness results to the case of strong connectedness using Proposition 3.7 and Corollary 3.8. The following proposition provides two further means of verifying that sets of words are subgroups.

Proposition 3.5. *Given any nonempty subset S of a group G, the following are equivalent*:

(1) *$\mathcal{W}(S)$ is a subgroup of G.*

(2) $\mathcal{W}(S) = \mathcal{W}(S^{-1})$.

(3) $\mathcal{W}(S) = \langle S \rangle$.

Proof. Clearly (2) implies (3) and (3) implies (1) so it remains to show that (1) implies (2). Assume $\mathcal{W}(S)$ is a subgroup of G and let $w \in \mathcal{W}(S)$. Then $w^{-1} \in \mathcal{W}(S)$ by assumption and $w^{-1} = s_1 s_2 \cdots s_k$, where $k > 0$. Hence $w = (s_1 s_2 \cdots s_k)^{-1} = s_k^{-1} s_{k-1}^{-1} \cdots s_1^{-1} \in \mathcal{W}(S^{-1})$.

Now suppose that $w \in \mathcal{W}(S^{-1})$. Then $w = s_1^{-1} s_2^{-1} \cdots s_k^{-1} = (s_k s_{k-1} \cdots s_1)^{-1} \in \mathcal{W}(S)$ because $\mathcal{W}(S)$ is a subgroup of G. Hence $\mathcal{W}(S) = \mathcal{W}(S^{-1})$ and the result follows. □

Remark 3.6. Notice that if G is a finite group, any subset S of G will satisfy the statements in Proposition 3.5. The statements will also hold in any group if the subset S is inverse-closed, as is assumed in places in [Iradmusa and Praeger 2016], or if all elements of S have finite order.

Proposition 3.7. *In the two-sided group digraph* $2\mathrm{S}(G; L, R)$, *if $\mathcal{W}(L)$ and $\mathcal{W}(R)$ are subgroups of G, there is a directed path from g to h if and only if there is a directed path from h to g.*

Proof. Suppose that there is a directed path from g to h in $2\mathrm{S}(G; L, R)$. Then we have $h = w_{L^{-1},n} g w_{R,n}$ for some $n \in \mathbb{N}$ which implies $g = w_{L^{-1},n}^{-1} h w_{R,n}^{-1}$.

Since $\mathcal{W}(L^{-1})$ and $\mathcal{W}(R)$ are both subgroups of G, we know $w_{L^{-1},n}^{-1} \in \mathcal{W}(L^{-1})$ and $w_{R,n}^{-1} \in \mathcal{W}(R)$; i.e., inverses can be expressed as words in the original set. It will be sufficient to show that both of the inverses can be expressed as words in their respective sets with the same length.

First suppose that $w_{L^{-1},n}^{-1} = w_{L^{-1},a}$ and $w_{R,n}^{-1} = w_{R,b}$ for some $a, b \in \mathbb{N}$. Then we have $e = w_{L^{-1},n} w_{L^{-1},a}$ and similarly $e = w_{R,n} w_{R,b}$ of total lengths at least 1. Using that e is the identity, we have

$$\begin{aligned} e &= w_{L^{-1},n} w_{L^{-1},a} (w_{L^{-1},n} w_{L^{-1},a})^{n+b-1} \\ &= w_{L^{-1},n} w_{L^{-1},a} w_{L^{-1},(n+a)(n+b-1)} \\ &= w_{L^{-1},n} w_{L^{-1},a+(n+a)(n+b-1)}. \end{aligned}$$

This shows that $w_{L^{-1},n}^{-1}$ can be expressed as a word in L^{-1} of length $(n+a)(n+b)-n$. Similarly, we can express $w_{R,n}^{-1}$ as a word in R of the same length. Therefore there is a directed path from h to g in $2S(G; L, R)$. □

Corollary 3.8. *In the two-sided group digraph* $2S(G; L, R)$, *if* $\mathcal{W}(L)$ *and* $\mathcal{W}(R)$ *are subgroups of* G, *then* $g \in G$ *is weakly connected to* $h \in G$ *if and only if* g *is strongly connected to* h, *and hence weakly connected components are identical to strongly connected components.*

Proof. Assume that g is weakly connected to h in $2S(G; L, R)$. Then there exists a path $g_0, g_1, \dots, g_n$ with $g = g_0$ and $h = g_n$ such that either (g_{i-1}, g_i) or (g_i, g_{i-1}) is an arc for $1 \le i \le n$. For every arc of the form (g_i, g_{i-1}), apply Proposition 3.7. This generates a new directed path $g_0', g_1', \dots, g_m'$ with $g = g_0'$ and $h = g_m'$ such that (g_{i-1}', g_i') is an arc for $1 \le i \le m$. Applying Proposition 3.7 again yields that g is strongly connected to h. □

Under the hypothesis that $\mathcal{W}(L)$ and $\mathcal{W}(R)$ are subgroups of G, Proposition 3.7 and Corollary 3.8 allow us to convert any statement about weak connectedness into a corresponding statement about strong connectedness. This leads to the following results analogous to Lemma 3.2 and Corollary 3.3 and consequently a well-defined notion of minimum strong connection length.

Lemma 3.9. *In* $2S(G; L, R)$ *if* $\mathcal{W}(L)$ *and* $\mathcal{W}(R)$ *are subgroups of* G *and* $g = w_{L^{-1},a} w_{R,n}$, *then* g *is strongly connected to* l^d *and to* r^d, *where* l *is any element of* L, r *is any element of* R, *and* $d = n - a$.

Corollary 3.10. *Let* $\mathcal{W}(L)$ *and* $\mathcal{W}(R)$ *be subgroups of* G:

(1) *In* $2S(G; L, R)$ *there exist two words in* L *of different lengths that are strongly connected if and only if there is a word in* L *that is strongly connected to* e.

(2) *In* $2S(G; L, R)$ *there exists a word* $w_{L,n}$ *strongly connected to* e *if and only if there exists a word* $w_{L^{-1},n}$ *strongly connected to* e.

Definition 3.11. Assuming that $\mathcal{W}(L)$ and $\mathcal{W}(R)$ are subgroups of G, the *minimum strong connection length in* G *relative to* (L, R) is the minimum length k of a word purely in L, L^{-1}, R, or R^{-1} that is strongly connected to e, and is infinite if there is no such minimum. Algebraically this is equivalent to the minimum length of a word v purely in L, L^{-1}, R, or R^{-1} such that $e = w_{L^{-1},n} v w_{R,n}$ for some $n \in \mathbb{N}$.

Lemma 3.9 and Corollary 4.10 also lead to the following version of Theorem 2.4.

Corollary 3.12. *Let $\mathcal{W}(L)$ and $\mathcal{W}(R)$ be subgroups of G. The two-sided group digraph $2\mathrm{S}(G; L, R)$ is strongly connected if and only if $G = \langle L\rangle\langle R\rangle$ and $e = w_{L^{-1},i} w_{R,j}$, where $|i - j| = 1$.*

3B. ***Connected components.*** In this section we count numbers of connected components and characterize their vertices, assuming that elements of G factor as a word in $\bar{L} = L \cup L^{-1}$ times a word in $\bar{R} = R \cup R^{-1}$.

Theorem 3.13. *Let L and R be nonempty subsets of a group G. If $G = \mathcal{W}(\bar{L})\mathcal{W}(\bar{R})$ and k is the minimum weak connection length for G relative to (L, R), then the two-sided group digraph $2\mathrm{S}(G; L, R)$ has exactly k weakly connected components all of the same size. Moreover, if $L \cap N_G(L) \neq \varnothing$ or $R \cap N_G(R) \neq \varnothing$, then all components are isomorphic.*

Proof. Assume $G = \mathcal{W}(\bar{L})\mathcal{W}(\bar{R})$ and let k be the minimum weak connection length for G relative to (L, R). If k is infinite, then by Corollary 3.3, any two words in L of different lengths are not weakly connected to each other and it follows that $2\mathrm{S}(G; L, R)$ will have infinitely many connected components. Otherwise $k \in \mathbb{N}$ and by Corollary 3.3, we may assume $e = W_{\bar{L},m,a} l^k W_{\bar{R},a,m}$. For $0 \le i < j < k$ we claim that $l^i \neq l^j$ and there is no path between l^i and l^j.

If $l^i = l^j$ for some $0 \le i < j < k$, then $e = l^{j-i}$, contradicting the minimality of k as the weak connection length for G relative to (L, R). Similarly, if $l^i = W_{\bar{L},m,a} l^j W_{\bar{R},a,m}$ then $e = l^{-i} W_{\bar{L},m,a} l^j W_{\bar{R},a,m}$ and Lemma 3.2 yields $e \sim l^{j-i}$, which again contradicts the minimality of k. This shows that $2\mathrm{S}(G; L, R)$ has at least k weakly connected components.

To show that $2\mathrm{S}(G; L, R)$ has exactly k weakly connected components, we first notice that since $G = \mathcal{W}(\bar{L})\mathcal{W}(\bar{R})$, Lemma 3.2 means that for every $g \in G$, we have $g \sim l^d$ for some integer d. Hence it suffices to show that for all $d \in \mathbb{Z}$, we have $l^d \sim l^i$ for some $0 \le i < k$. This statement is true since by Lemma 3.2, $e \sim l^{-k}$ and $e \sim l^k$, which allow d to be reduced modulo k.

Fix $l \in L$ and let Γ_i for $0 \le i < k$ be the weakly connected component of $2\mathrm{S}(G; L, R)$ containing l^i. Then the Γ_i are distinct and the union of $\Gamma_0, \dots, \Gamma_{k-1}$ is $2\mathrm{S}(G; L, R)$. To see that all of the connected components have the same size, consider the injective maps $\phi_i : \Gamma_0 \to \Gamma_i$ for $1 \le i < k$ defined by $\phi_i(h) = l^i h$. The map sending h to $l^{-i}h$ is also injective and is an inverse to ϕ_i, showing that ϕ_i is bijective and all connected components have the same size.

Now assume $l \in L \cap N_G(L)$ and let Γ_i and ϕ_i be defined as above. The maps ϕ_i will preserve arcs because if (x, y) is an arc in Γ_0 then $y = l_1^{-1} x r_1$ for some $l_1 \in L$

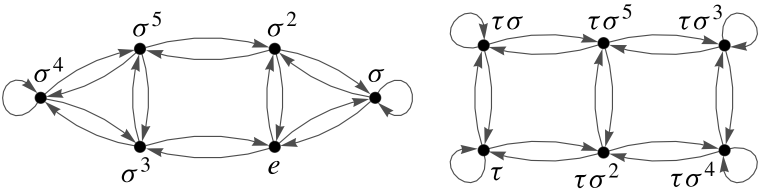


Figure 3. $2\mathrm{S}(D_6; \{\tau, \tau\sigma^5\}, \{\tau\sigma, \tau\sigma^2\})$.

and $r_1 \in R$, and, since $l \in N_G(L)$, for some $l_2 \in L$

$$\phi_i(y) = l^i y = l^i l_1^{-1} x r_1 = l_2^{-1} l^i x r_1 = l_2^{-1} \phi_i(x) r_1.$$

Similarly, if $(\phi_i(x), \phi_i(y))$ is an arc in Γ_i, then (x, y) is an arc in Γ_0. Thus the disjoint connected components are isomorphic to each other.

For the case when $R \cap N_G(R) \neq \varnothing$, note that the above proof can be modified using the set $\{r^i\}_{i=0}^{k-1}$ to describe the Γ_i and defining $\phi_i(h) = hr^i$ instead. □

Remark 3.14. In Theorem 3.13 if in fact $L \cap N_G(L) = L$ and $R \cap N_G(R) = R$, then $2\mathrm{S}(G; L, R)$ is also vertex-transitive by [Iradmusa and Praeger 2016, Theorem 1.13].

Corollary 3.15. *The two-sided group digraph* $2\mathrm{S}(G; L, R)$ *is weakly connected if and only if* $G = \mathcal{W}(\bar{L})\mathcal{W}(\bar{R})$ *and there exists some element of* $\bar{L}$ *or* $\bar{R}$ *that is weakly connected to* e.

Using Proposition 3.7 and Corollary 3.8 as described before Lemma 3.9 yields the following.

Theorem 3.16. *Let* $\mathcal{W}(L)$ *and* $\mathcal{W}(R)$ *be subgroups of* G. *If* $G = \mathcal{W}(L)\mathcal{W}(R)$ *and* k *is the minimum strong connection length for* G *relative to* (L, R), *then the two-sided group digraph* $2\mathrm{S}(G; L, R)$ *has exactly* k *strongly connected components all of the same size. Moreover, if* $L \cap N_G(L) \neq \varnothing$ *or* $R \cap N_G(R) \neq \varnothing$, *then all components are isomorphic.*

Corollary 3.17. *Let* $\mathcal{W}(L)$ *and* $\mathcal{W}(R)$ *be subgroups of* G. *Then the two-sided group digraph* $2\mathrm{S}(G; L, R)$ *is strongly connected if and only if* $G = \mathcal{W}(L)\mathcal{W}(R)$ *and there exists some element of* $\bar{L}$ *or* $\bar{R}$ *that is strongly connected to* e.

Example 3.18. Consider $2\mathrm{S}(D_6; \{\tau, \tau\sigma^5\}, \{\tau\sigma, \tau\sigma^2\})$ as in Example 2.7. Since $\tau \in L$ and $\sigma = (\tau\sigma^5)\tau \in \mathcal{W}(L)$, we know $D_6 = \mathcal{W}(L) = \mathcal{W}(L)\mathcal{W}(R)$. Since $e \not\sim \tau$ but $e \sim \tau^2 = e$, the graph, as seen in Figure 3, has two strongly connected components of the same size, as shown in Theorem 3.16. Notice that $N_{D_6}(L) = N_{D_6}(R) = \{e, \sigma^3\}$ does not intersect L or R so the fact that the components are not isomorphic does not violate Theorem 3.16.

Example 3.19. Consider $2\mathrm{S}(D_{10}; \{\sigma\}, \{\tau, \sigma^3\})$. It is clear that $D_{10} = \mathcal{W}(R) = \mathcal{W}(L)\mathcal{W}(R)$. Since $e \not\sim \tau$ but $e \sim \tau^2 = e$, the graph, as seen in Figure 4, has two

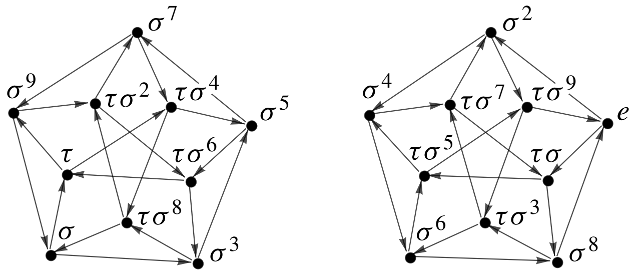


Figure 4. $2S(D_{10}; \{\sigma\}, \{\tau, \sigma^3\})$.

strongly connected components of the same size as shown in Theorem 3.16. Notice that $N_{D_{10}}(L) = \langle\sigma\rangle$ so $\sigma \in L \cap N_{D_{10}}(L)$ and the components are isomorphic by Theorem 3.16.

4. Double cosets

Recall that for a Cayley digraph $\mathrm{Cay}(G, S)$ the coset $\langle S\rangle g$ is the weakly connected component of the digraph containing $g \in G$ and that if H and K are subgroups of a group G then the double cosets HgK for $g \in G$ partition G into (possibly different sized) subsets. In the two-sided group digraph $2S(G; L, R)$ the component containing $g \in G$ need only be contained in the double coset $\langle L\rangle g\langle R\rangle$.

Proposition 4.1. *The weakly or strongly connected component of* $2S(G; L, R)$ *containing g is a subset of the double coset* $\langle L\rangle g\langle R\rangle$.

Proof. Let h be weakly connected to g; that is, h is of the form $W_{\bar{L},m,a}gW_{\bar{R},a,m}$ for some $W_{\bar{L},m,a} \in \mathcal{W}(\bar{L}) = \langle L\rangle$ and $W_{\bar{R},a,m} \in \mathcal{W}(\bar{R}) = \langle R\rangle$. Then $h \in \langle L\rangle g\langle R\rangle$ and the weakly or strongly connected component containing g lies in $\langle L\rangle g\langle R\rangle$. □

In Theorem 4.5, without the assumption that $G = \mathcal{W}(\bar{L})\mathcal{W}(\bar{R})$, we count connected components within double cosets analogously to Theorem 3.13. Connected components in a given double coset have the same size, but between different double cosets the sizes of components can differ. This is illustrated in Figure 5 for Example 4.8, Figure 7 for Example 4.14, and Figure 8 for Example 4.15.

Let L and R be nonempty subsets of G and fix a set S of double coset representatives for $\langle L\rangle$ and $\langle R\rangle$. Each g in G lies in a double coset $\langle L\rangle s\langle R\rangle$ for some $s \in S$, and s will play the role in $\langle L\rangle s\langle R\rangle$ that the identity element played in Sections 3A and 3B.

Lemma 4.2. *In* $2S(G; L, R)$ *if* $g = w_{\bar{L},m,a}sw_{\bar{R},n,b}$ *for* $s \in G$, *then* $g \sim l^d s$ *and* $g \sim sr^d$, *where* l *is any element of* L, r *is any element of* R, *and* $d = m+n-(a+b)$.

Proof. This proof is identical to the proof of Lemma 3.2 with s inserted between the words from $\bar{L}$ and words from $\bar{R}$. □

Corollary 4.3. *In* $2\mathrm{S}(G; L, R)$ *the following hold with* $s \in G$:

(1) *There exist words* $w_{L,m}$ *and* $w_{L,n}$ *with* $m \neq n$ *such that* $w_{L,m}s \sim w_{L,n}s$ *if and only if there exists* $w_{L,k}$ *such that* $w_{L,k}s \sim s$. *One can take* $k = |m - n|$.

(2) *There exists a word* $w_{L,n}$ *such that* $w_{L,n}s \sim s$ *if and only if there exists a word* $w_{L^{-1},n}$ *such that* $w_{L^{-1},n}s \sim s$.

(3) *If* g *is in* $\langle L\rangle s\langle R\rangle$ *then* $w_{L,k}g \sim g$ *for some* $w_{L,k}$ *in* $\mathcal{W}(L)$ *if and only if* $w'_{L,k}s \sim s$ *for some* $w'_{L,k}$ *in* $\mathcal{W}(L)$.

Proof. The first two parts follow similarly to their analogues in Corollary 3.3.

For (3), by symmetry it is enough to prove one direction. Let $g = w_{\bar{L},m,a}sw_{\bar{R},n,b}$ and $w_{L,k}g \sim g$. Rewriting $w_{L,k}g \sim g$ in terms of s yields $w_{L,k}w_{\bar{L},m,a}sw_{\bar{R},n,b} \sim w_{\bar{L},m,a}sw_{\bar{R},n,b}$. Applying Lemma 4.2 to both sides yields $l^{k+d}s \sim l^d s$. Hence $l^k s \sim s$ by (1). □

By the first two parts of Corollary 4.3, if there exists a minimum length k_s of a word w in L such that $ws \sim s$, then it is also the minimum length of such a word in L^{-1}, and by Corollary 4.3(3) the minimum such length is independent of the representative of a double coset. Inserting $r^{k_s}r^{-k_s}$ to the right of s shows that k_s is also the minimum length of a word w in R such that $sw \sim s$, and hence, by an R version of Corollary 4.3, k_s is also the minimum such length of a word in R^{-1}. Thus the following definition for the minimum weak connection length in $\langle L\rangle s\langle R\rangle$ is well-defined.

Definition 4.4. The *minimum weak connection length in* $\langle L\rangle s\langle R\rangle$ is the minimum length k_s of a word w purely in L or L^{-1} such that $ws \sim s$ in $2\mathrm{S}(G; L, R)$, or the minimum length k_s of a word w purely in R or R^{-1} such that $sw \sim s$ in $2\mathrm{S}(G; L, R)$. Take k_s to be infinite if there is no such minimum. Algebraically this is equivalent to the minimum length of a word w purely in L or L^{-1} such that $s = W_{\bar{L},m,a}wsW_{\bar{R},a,m}$ for some $a, m \in \mathbb{N}$, or the minimum length of a word w purely in R or R^{-1} such that $s = W_{\bar{L},m,a}swW_{\bar{R},a,m}$ for some $a, m \in \mathbb{N}$.

Theorem 4.5. *Let* L *and* R *be nonempty subsets of a group* G. *If* k_s *is the minimum weak connection length for* $\langle L\rangle s\langle R\rangle$, *then the double coset* $\langle L\rangle s\langle R\rangle$ *within* $2\mathrm{S}(G; L, R)$ *consists of exactly* k_s *weakly connected components all of the same size. Moreover, if* $L \cap N_G(L) \neq \varnothing$ *or* $R \cap N_G(R) \neq \varnothing$, *then all components within the same double coset are isomorphic.*

Proof. This follows from Lemma 4.2 and Corollary 4.3 exactly as in the proof of Theorem 3.13. □

Corollary 4.6. *In the two-sided group digraph* $2\mathrm{S}(G; L, R)$ *there are* $\sum_{s\in S} k_s$ *weakly connected components, where* S *is a set of double coset representatives for* G *modulo* $\langle L\rangle$ *and* $\langle R\rangle$ *and* k_s *is the minimum weak connection length for* $\langle L\rangle s\langle R\rangle$.

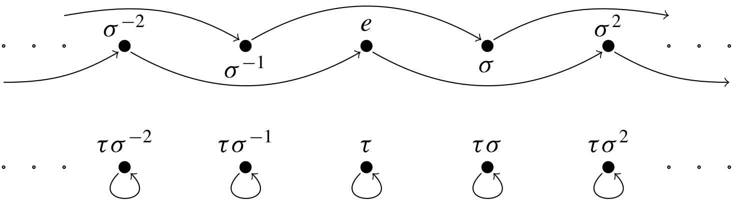


Figure 5. $2\mathrm{S}(D_\infty; \{\sigma^a\}, \{\sigma^b\})$, where $a=-1,\ b=1$.

Remark 4.7. Note that in Theorem 4.5 if in fact $L \cap N_G(L) = L$ and $R \cap N_G(R) = R$, then by an argument similar to that in [Iradmusa and Praeger 2016, Theorem 1.13] the subgraph $\langle L\rangle s \langle R\rangle$ is vertex-transitive.

Example 4.8. Consider $2\mathrm{S}(D_\infty; \{\sigma^a\}, \{\sigma^b\})$ with $\gcd(a, b) = 1$ and where D_∞ is the group of isometries of $\mathbb{Z}$ with the presentation

$$D_\infty = \langle \sigma, \tau \mid \tau^2 = e,\ \sigma\tau = \tau\sigma^{-1}\rangle.$$

We think of σ as right translation and τ as negation. Since $\langle L\rangle = \{\sigma^{an}\}$ and $\langle R\rangle = \{\sigma^{bn}\}$ with a, b relatively prime, D_∞ has two double cosets, namely $\langle L\rangle\langle R\rangle = \langle\sigma\rangle$ and $\langle L\rangle\tau\langle R\rangle = \tau\langle\sigma\rangle$.

It is easy to see that each $g \in D_\infty$ has exactly one out-neighbor and one in-neighbor (possibly the same). If $g = \sigma^n \in \langle L\rangle\langle R\rangle$, then g lies on the arcs $(\sigma^n, \sigma^{n+(b-a)})$ and $(\sigma^{n-(b-a)}, \sigma^n)$. If instead $g = \tau\sigma^n \in \langle L\rangle\tau\langle R\rangle$, then g lies on the arcs $(\tau\sigma^n, \tau\sigma^{n+a+b})$ and $(\tau\sigma^{n-(a+b)}, \tau\sigma^n)$. Therefore, the structure of the graph depends on $b-a$ and $a+b$.

If $b-a \neq 0$, then the double coset $\langle L\rangle\langle R\rangle$ consists of $|b-a|$ weakly connected components each consisting of σ^n with n fixed modulo $|b-a|$. If $b-a=0$, then the arcs are of the form (σ^n, σ^n), and the double coset consists of isolated points linked only to themselves, and so has infinitely many connected components. Both of these cases illustrate the results of Theorem 4.5.

The value of $a+b$ plays the same role for the structure of the double coset $\langle L\rangle\tau\langle R\rangle$. Two example graphs are provided, Figure 5 for $a=-1, b=1$ and Figure 6 for $a=1, b=2$.

Proposition 3.7 and Corollary 3.8 again yield corresponding strongly connected results.

Lemma 4.9. *In $2\mathrm{S}(G; L, R)$ if $\mathcal{W}(L)$ and $\mathcal{W}(R)$ are subgroups of G and $g = w_{L^{-1},a}sw_{R,n}$ for $s \in G$, then g is strongly connected to $l^d s$ and to sr^d, where l is any element of L, r is any element of R, and $d = n-a$.*

Corollary 4.10. *In $2\mathrm{S}(G; L, R)$ if $\mathcal{W}(L)$ and $\mathcal{W}(R)$ are subgroups of G, then the following three properties hold for any $s \in G$:*

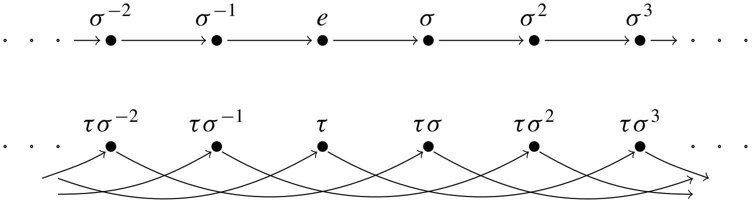


Figure 6. $2\mathrm{S}(D_\infty; \{\sigma^a\}, \{\sigma^b\})$, where $a = 1,\ b = 2$.

(1) *There exist words $w_{L^{-1},m}$ and $w_{L^{-1},n}$ with $m \neq n$ such that $w_{L^{-1},m}s$ is strongly connected to $w_{L^{-1},n}s$ if and only if there exists $w_{L^{-1},k}$ such that $w_{L^{-1},k}s$ is strongly connected to s. In practice, $k = |m - n|$.*

(2) *There exists a word $w_{L,n}$ such that $w_{L,n}s$ is strongly connected to s if and only if there exists a word $w_{L^{-1},n}$ such that $w_{L^{-1},n}s$ is strongly connected to s.*

(3) *If g is in $\langle L\rangle s\langle R\rangle$ then $w_{L,k}g$ is strongly connected to g for some $w_{L,k}$ in $\mathcal{W}(L)$ if and only if $w'_{L,k}s$ is strongly connected to s for some $w'_{L,k}$ in $\mathcal{W}(L)$.*

Definition 4.11. Assuming that $\mathcal{W}(L)$ and $\mathcal{W}(R)$ are subgroups of G, the *minimum strong connection length in $\langle L\rangle s\langle R\rangle$* is the minimum length k_s of a word w purely in L or L^{-1} such that ws is strongly connected to s in $2\mathrm{S}(G; L, R)$, or the minimum length k_s of a word w purely in R or R^{-1} such that sw is strongly connected to s in the two-sided group digraph $2\mathrm{S}(G; L, R)$. Take k_s to be infinite if there is no such minimum. Algebraically this is equivalent to the minimum length of a word v purely in L or L^{-1} such that $s = w_{L^{-1},n}vsw_{R,n}$ for some $n \in \mathbb{N}$, or the minimum length of a word v purely in R or R^{-1} such that $s = w_{L^{-1},n}svw_{R,n}$ for some $n \in \mathbb{N}$.

Theorem 4.12. *Let $\mathcal{W}(L)$ and $\mathcal{W}(R)$ be subgroups of G. If k_s is the minimum strong connection length for $\langle L\rangle s\langle R\rangle$ in $2\mathrm{S}(G; L, R)$, then the double coset $\langle L\rangle s\langle R\rangle$ consists of exactly k_s strongly connected components all of the same size. Moreover, if $L \cap N_G(L) \neq \varnothing$ or $R \cap N_G(R) \neq \varnothing$, then all components within the same double coset are isomorphic.*

Problem 2. Theorems 3.13, 3.16, 4.5, and 4.12 provide sufficient conditions for connected components to be isomorphic. Find necessary and sufficient conditions for this to occur.

Corollary 4.13. *Let $\mathcal{W}(L)$ and $\mathcal{W}(R)$ be subgroups of G. The two-sided group digraph $2\mathrm{S}(G; L, R)$ consists of $\sum_{s\in S} k_s$ strongly connected components, where S is a set of double coset representatives for G modulo $\langle L\rangle$ and $\langle R\rangle$ and k_s is the minimum strong connection length for $\langle L\rangle s\langle R\rangle$.*

Example 4.14. The digraph $2\mathrm{S}(D_3 \times C_3; \{(\tau\sigma^2, g^2)\}, \{(e, g^2), (\tau, g^2)\})$ shown in Figure 7 is an example of Theorem 4.12. The two double cosets in $G = D_3 \times C_3$ are

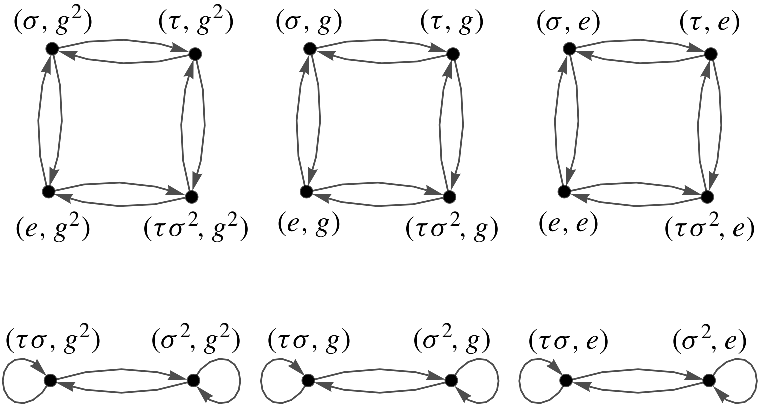


Figure 7. $2\mathrm{S}(D_3 \times C_3; \{(\tau\sigma^2, g^2)\}, \{(e, g^2), (\tau, g^2)\})$.

$\langle L\rangle\langle R\rangle$ and $\langle L\rangle(\sigma^2, e)\langle R\rangle$, both of which have minimum strong connection length of 3. Since L consists of a single element, $L \cap N_G(L) \neq \varnothing$ and all components within each double coset are isomorphic.

Example 4.15. Another example is provided by $2\mathrm{S}(A_5; \{(235)\}, \{(243), (254)\})$, shown in Figure 8. There are three double cosets in A_5 modulo $\langle L\rangle$ and $\langle R\rangle$, whose representatives are the identity, (123), and (145). The minimum strong connection length is 3 in the first two double cosets and 1 in the third. The connected components of $\langle L\rangle\langle R\rangle$ have four vertices. All the connected components in the other two double cosets contain 12 vertices and are isomorphic. That the components within $\langle L\rangle(123)\langle R\rangle$ are isomorphic follows from the fact that L consists of a single element, so $L \cap N_G(L) \neq \varnothing$.

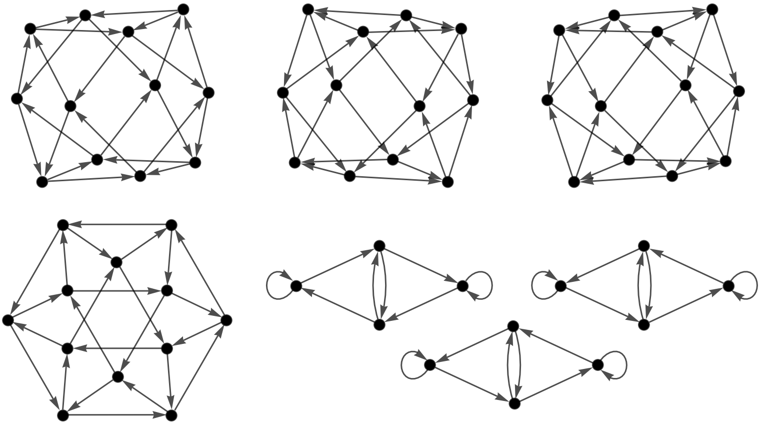


Figure 8. $2\mathrm{S}(A_5; \{(235)\}, \{(243), (254)\})$.

5. Orbit counting

Another way to count strongly connected components is to use group actions. We briefly review necessary background material.

A group G *acts* (on the right) on a set X if there exists a function $\alpha : X \times G \to X$, where $(x, g) \mapsto x.g$ such that $x.e = x$ and for all $g_1, g_2 \in G$ and all $x \in X$, we have $x \cdot (g_1 g_2) = (x \cdot g_1).g_2$. If G acts on a set X, then for any $x \in X$, the set $x.G = \{x.g \mid g \in G\}$ is the *orbit* of x under G. It can be shown that X is the disjoint union of its orbits. If G acts on X, the *stabilizer* of $x \in X$ is the subgroup $G_x = \{g \mid x.g = x\}$ of G and the set fixed by $g \in G$ is $X^g = \{x \mid x.g = x\}$. The following well known results are used to prove Theorem 5.3.

Lemma 5.1. *Suppose that a group G acts on a set X. If $x \in X$, then the mapping $\phi : G_x \backslash G \to x.G$ defined by $\phi(G_x g) = x.g$ is well-defined and bijective. Thus, $|G| = |x.G||G_x|$.*

Lemma 5.2. *Suppose that a group G acts on a set X:*

(1) *If $x \in X$ and $g \in G$, the stabilizer of $x.g$ is $G_{x.g} = g^{-1} G_x g$.*

(2) *If x and y are in the same orbit under G, then $|G_x| = |G_y|$.*

Theorem 5.3. *Suppose that a group G acts on a set X. The number N of distinct orbits of G on X satisfies*

$$N \cdot |G| = \sum_{g \in G} |X^g|.$$

Proof. The case where X or G is infinite is trivial so let X and G be finite. Consider the set $Y = \{(x, g) \mid g \in G,\ x \in X,\ x.g = x\} \subset X \times G$. We may count elements of Y as $|Y| = \sum_{g \in G} |X^g| = \sum_{x \in X} |G_x|$. Alternatively, consider representatives $x_1, x_2, x_3, \ldots, x_N$ from each orbit of X. If x is in the same orbit as x_i, then $x.G = x_i.G$ and hence, by Lemma 5.2, $|G_x| = |G_{x_i}|$. We therefore have, by Lemma 5.1,

$$\sum_{g \in G} |X^g| = \sum_{i=1}^{N} \sum_{x \in x_i.G} |G_x| = \sum_{i=1}^{N} |x_i.G||G_{x_i}| = \sum_{i=1}^{N} |G| = N \cdot |G|. \qquad \square$$

We apply this result to $2\mathrm{S}(G; L, R)$. Define

$$U = \{(w_{L,n}, w_{R,n}) \mid w_{L,n} \in \mathcal{W}(L),\ w_{R,n} \in \mathcal{W}(R)\} \subseteq G \times G.$$

We show that if $\mathcal{W}(L)$ and $\mathcal{W}(R)$ are subgroups of G then U is a subgroup of $G \times G$. The set U is clearly closed under multiplication. The fact that U is closed under inverses follows from the proof of Proposition 3.7. Since U is not empty it contains an identity and U is a group under composition.

The action of U on G is induced by the standard action of $G \times G$ on G by $g \cdot (g_1, g_2) = g_1^{-1} g g_2$; that is, $g \cdot (w_{L,n}, w_{R,n}) = w_{L,n}^{-1} g w_{R,n}$. One can check that this is in fact a right action. For each element g in G, the orbit $g \cdot U$ is the strongly connected component of $2\mathrm{S}(G; L, R)$ containing g.

Corollary 5.4. *Let $2\mathrm{S}(G; L, R)$ be a two-sided group digraph where $\mathcal{W}(L)$ and $\mathcal{W}(R)$ are subgroups of G and with the group U acting on G as defined above. The number N of strongly connected components in $2\mathrm{S}(G; L, R)$ satisfies $N \cdot |U| = \sum_{u \in U} |G^u|$.*

Example 5.5. Let $2\mathrm{S}(G; L, R)$ be a connected digraph and let H_N be any group of order N. Then $2\mathrm{S}(G \times H_N; L \times \{e\}, R \times \{e\})$ has N connected components. This shows that the number N of connected components may be arbitrarily large.

Problem 3. For a given group G, how many connected components can $2\mathrm{S}(G; L, R)$ have?

Note that if $G = \langle L \rangle \langle R \rangle$ then by Theorem 3.13 or Theorem 3.16 the number of connected components will divide $|G|$, but Example 4.15 shows this need not hold in general.

Based on the group action perspective and our observation about the connection between two-sided group digraphs and biquotients, we pose a question motivated by a common construction in the biquotient setting. We first define a generalization of $2\mathrm{S}(G; L, R)$.

Definition 5.6. Let G be a group and U be a nonempty subset of $G \times G$. Define the digraph $2\mathrm{S}(G; U)$ to have vertex set G and a directed arc (g, h) from g to h if and only if $h = u_l^{-1} g u_r$ for some $(u_l, u_r) \in U$.

Remark 5.7. Observe that if $U = L \times R$, then $2\mathrm{S}(G; U) = 2\mathrm{S}(G; L, R)$.

Motivated by the biquotient literature, we note a correspondence between the digraphs $2\mathrm{S}(G; U)$ and $2\mathrm{S}(G \times G; \Delta G, U)$, where $\Delta G = \{(g, g) \mid g \in G\}$ is the diagonal of $G \times G$. This correspondence is given by the map $\phi : G \times G \to G$, $\phi(g_1, g_2) = g_1^{-1} g_2$. Direct computation shows that the map ϕ takes arcs of $2\mathrm{S}(G \times G; \Delta G, U)$ to arcs of $2\mathrm{S}(G; U)$. Additionally, it produces a bijection between the connected components of $2\mathrm{S}(G \times G; \Delta G, U)$ and the connected components of $2\mathrm{S}(G; U)$. This allows the number of connected components of $2\mathrm{S}(G; U)$ to be counted using our preceding results, especially when one notes that by Theorems 4.5 and 4.12 the connected components of $2\mathrm{S}(G \times G; \Delta G, U)$ are precisely the double cosets.

Problem 4. Under what conditions on U do the connected components of $2\mathrm{S}(G; U)$ have the same size? Under what conditions are they isomorphic?

6. Reduction results

In this section we prove Proposition 6.2, which relates connectedness of a two-sided digraph for a semidirect product group to connectedness properties for the factors, and pose a final general problem. Remark 6.1 will be useful in the proof of Proposition 6.2.

Remark 6.1. A group G is said to be a *semidirect product* of its subgroups H and K, written $G = H \rtimes K$, if H is a normal subgroup of G, $G = HK$, and $H \cap K = \{e\}$. A subgroup K of a group G is a *retract* of G if there exists a homomorphism $\phi : G \to G$ such that $\phi(g) \in K$ for all $g \in G$ and $\phi(k) = k$ for all $k \in K$.

If K is a retract of a group G with retraction map ϕ, then it is easy to verify that $G = H \rtimes K$ for $H = \ker\phi$. Conversely, if $G = H \rtimes K$ then the map ϕ defined by $\phi(hk) = k$ is well-defined because $H \cap K = \{e\}$ and is a group homomorphism because $H \trianglelefteq G$. Hence $G = H \rtimes K$ if and only if K is a retract of G with retraction ϕ and $H = \ker\phi$. Denote $\phi(L)$ by L^{ϕ}.

Proposition 6.2. *Let K be a retract of a group G under the retraction ϕ. Then* $2\mathrm{S}(G; L, R)$ *is weakly connected if and only if* $2\mathrm{S}(K; L^{\phi}, R^{\phi})$ *is weakly connected and* $\ker\phi$ *is weakly connected within* $2\mathrm{S}(G; L, R)$.

Proof. Assume that $2\mathrm{S}(G; L, R)$ is weakly connected. Then certainly $H = \ker\phi$ is weakly connected within $2\mathrm{S}(G; L, R)$. Observe that $2\mathrm{S}(K; L^{\phi}, R^{\phi})$ is also weakly connected because the retraction $\phi : G \to K$ sends the arc $(g, l^{-1}gr)$ in $2\mathrm{S}(G; L, R)$ to the arc

$$(\phi(g), \phi(l^{-1}gr)) = (\phi(g), \phi(l)^{-1}\phi(g)\phi(r))$$

in $2\mathrm{S}(K; L^{\phi}, R^{\phi})$; i.e., ϕ induces a retraction from $2\mathrm{S}(G, L, R)$ to $2\mathrm{S}(K; L^{\phi}, R^{\phi})$.

Conversely, assume that $2\mathrm{S}(K; L^{\phi}, R^{\phi})$ is weakly connected and $H = \ker\phi$ is weakly connected within $2\mathrm{S}(G; L, R)$. We show that for every $g \in G$ there is a path in $2\mathrm{S}(G; L, R)$ from the identity to g. Write $g = hk$ for $h \in H$ and $k \in K$. Using that there is a path from e to k in $2\mathrm{S}(K, L^{\phi}, R^{\phi})$, write $k = w_{\bar{L}^{\phi},m,a} w_{\bar{R}^{\phi},a,m}$ and then

$$g = h w_{\bar{L}^{\phi},m,a} w_{\bar{R}^{\phi},a,m}.$$

For each factor $k_i \in \bar{R}^{\phi}$ in $w_{\bar{R}^{\phi},a,m}$, find $h_i \in H$ so that $h_i k_i \in \bar{R}$ and insert $h_i^{-1}h_i$ before k_i in $w_{\bar{R}^{\phi},a,m}$. Insert similarly appropriate expressions for the identity before each factor from $\bar{L}^{\phi}$ in $w_{\bar{L}^{\phi},m,a}$. Then use $H \trianglelefteq G$ to rewrite g as

$$g = w_{\bar{L},m,a} h' w_{\bar{R},a,m},$$

where $h' \in H$, exhibiting a path from h' to g. Since there is a path from e to h' in $2\mathrm{S}(G; L, R)$, there is also a path from e to g in $2\mathrm{S}(G; L, R)$. This proves $2\mathrm{S}(G; L, R)$ is weakly connected. □

Problem 5. Develop analogues of earlier results about numbers of connected components and isomorphisms between them in the setting of semidirect products.

Example 6.3. Consider the digraph $2\mathrm{S}(D_6; \{\sigma\}, \{\sigma^2, \tau\})$, where $D_6 = \langle\sigma\rangle \rtimes \langle\tau\rangle$. Given $\sigma^n \in D_6$, the arc $(\sigma^n, \sigma^{-1}\sigma^n\sigma^2) = (\sigma^n, \sigma^{n+1})$ shows that $\langle\sigma\rangle$ is weakly connected in $2\mathrm{S}(D_6; \{\sigma\}, \{\sigma^2, \tau\})$. Furthermore, $2\mathrm{S}(\langle\tau\rangle; L^\phi, R^\phi) = 2\mathrm{S}(\langle\tau\rangle; \{e\}, \{e, \tau\})$ is connected since the graph consists of two vertices e and τ with arcs between e and τ and loops at each. Therefore by Proposition 6.2, $2\mathrm{S}(D_6; \{\sigma\}, \{\sigma^2, \tau\})$ is weakly connected.

Example 6.4. The two-sided group digraph $2\mathrm{S}(D_6; \{\tau, \tau\sigma^5\}, \{\tau\sigma, \tau\sigma^2\})$ from Example 3.19 is disconnected. Here $2\mathrm{S}(K; L^\phi, R^\phi)$ consists of isolated vertices e and τ with a loop at each and H is weakly connected within $2\mathrm{S}(D_6; \{\tau, \tau\sigma^5\}, \{\tau\sigma, \tau\sigma^2\})$.

Using an argument similar to the one in the proof of Proposition 6.2, one can prove the following.

Corollary 6.5. *Given a group G and a normal subgroup N let $\phi : G \to G/N$ be the canonical projection. Then $2\mathrm{S}(G; L, R)$ is weakly connected if and only if $2\mathrm{S}(G/N; L^\phi, R^\phi)$ is weakly connected and N is weakly connected within $2\mathrm{S}(G; L, R)$.*

In both Proposition 6.2 and Corollary 6.5 under the further assumption that $\mathcal{W}(L)$ and $\mathcal{W}(R)$ are subgroups of G similar conclusions hold for strong connectedness.

Acknowledgments

We thank Briana Foster-Greenwood for helpful discussions, particularly during the initial phases of this project, and Yu Chen for the observation behind Corollary 5.4. We thank the referee for the encouragement to add further problems and for particular suggestions leading to Problems 1 and 3.

References

[Anil Kumar 2012] V. Anil Kumar, "Generalized Cayley digraphs", *Pure Math. Sci.* **1**:1 (2012), 1–12. Zbl

[Annexstein et al. 1990] F. Annexstein, M. Baumslag, and A. L. Rosenberg, "Group action graphs and parallel architectures", *SIAM J. Comput.* **19**:3 (1990), 544–569. MR Zbl

[DeVito 2011] J. DeVito, *The classification of simply connected biquotients of dimension at most 7 and 3 new examples of almost positively curved manifolds*, Ph.D. thesis, University of Pennsylvania, 2011, available at https://search.proquest.com/docview/878684574/.

[Eschenburg 1982] J.-H. Eschenburg, "New examples of manifolds with strictly positive curvature", *Invent. Math.* **66**:3 (1982), 469–480. MR Zbl

[Eschenburg 1984] J.-H. Eschenburg, *Freie isometrische Aktionen auf kompakten Lie-Gruppen mit positiv gekrümmten Orbiträumen*, Schriftenreihe Math. Inst. Univ. Münster (2) **32**, Univ. Münster, 1984. MR Zbl

[Gauyacq 1997] G. Gauyacq, "On quasi-Cayley graphs", *Discrete Appl. Math.* **77**:1 (1997), 43–58. MR Zbl

[Gromoll and Meyer 1974] D. Gromoll and W. Meyer, "An exotic sphere with nonnegative sectional curvature", *Ann. of Math.* (2) **100**:2 (1974), 401–406. MR Zbl

[Iradmusa and Praeger 2016] M. N. Iradmusa and C. E. Praeger, "Two-sided group digraphs and graphs", *J. Graph Theory* **82**:3 (2016), 279–295. MR Zbl

[Kelarev and Praeger 2003] A. V. Kelarev and C. E. Praeger, "On transitive Cayley graphs of groups and semigroups", *European J. Combin.* **24**:1 (2003), 59–72. MR Zbl

[Marušič et al. 1992] D. Marušič, R. Scapellato, and N. Zagaglia Salvi, "Generalized Cayley graphs", *Discrete Math.* **102**:3 (1992), 279–285. MR Zbl

Received: 2017-05-26 Revised: 2017-07-26 Accepted: 2017-08-01

pchikwanda@gsu.edu *Department of Mathematics and Statistics,*
Georgia State University, Atlanta, GA, United States

krilcath@isu.edu *Department of Mathematics and Statistics,*
Idaho State University, Pocatello, ID, United States

leeyunt@isu.edu *Idaho State University, Pocatello, ID, United States*

sandtayl@isu.edu *Idaho State University, Pocatello, ID, United States*

smitgarr@isu.edu *Idaho State University, Pocatello, ID, United States*

yerodmyt@isu.edu *Idaho State University, Pocatello, ID, United States*

INVOLVE 11:4 (2018)
dx.doi.org/10.2140/involve.2018.11.701

Nonunique factorization over quotients of PIDs

Nicholas R. Baeth, Brandon J. Burns,
Joshua M. Covey and James R. Mixco

(Communicated by Vadim Ponomarenko)

We study factorizations of elements in quotients of commutative principal ideal domains that are endowed with an alternative multiplication. This study generalizes the study of factorizations both in quotients of PIDs and in rings of single-valued matrices. We are able to completely describe the sets of factorization lengths of elements in these rings, as well as compute other finer arithmetical invariants. In addition, we provide the first example of a finite bifurcus ring.

1. Introduction

Of course every commutative principal ideal domain (PID) is a unique factorization domain and every nonzero nonunit factors uniquely as a product of irreducible (prime) elements. It is not surprising that this property of unique factorization passes, in some sense, to any quotient ring of a PID. However, if D is a PID and n is the product of two or more primes in D, then $D/(n)$ contains nonzero zerodivisors that make factorization more interesting. For example, in $\mathbb{Z}/(900)$, $\overline{30}$ factors only as $\overline{30}=\overline{2}\cdot\overline{3}\cdot\overline{5}$, while $\overline{100}$ factors as $\overline{2^2}\cdot\overline{5^2}\cdot\overline{46^a}\cdot\overline{55^b}$ for any $a,b\in\mathbb{N}_0$. In fact, if D is a PID and n is the product of at least two primes of D, there are elements in $D/(n)$ that have unique factorization and others that have infinitely many factorizations — and of arbitrarily long lengths. A complete characterization of how elements factor over quotients of PIDs is given in [Baeth et al. 2017] and is summarized here in Proposition 3.1. The goal of this note is to study factorizations in quotients of PIDs endowed with an alternative multiplicative structure. The purpose is threefold: First, by introducing a more general multiplication in a principal ideal ring, we generalize both the results of [Baeth et al. 2017] (factorization in quotients of PIDs) and of [Baeth et al. 2011; Jacobson 1965] (factorization in rings of single-valued matrices). Secondly, we give examples of finite bifurcus rings, thus giving an affirmative answer to Open Problem 2.1.3 of [Adams et al. 2009]. Finally, we provide an even larger class of examples of commutative rings R such that every element of R is

MSC2010: 13A05, 13F15.
Keywords: factorizations, zerodivisors, bifurcus.

a zerodivisor and such that the set of factorization lengths of each element is a discrete interval, with many of these intervals being infinite.

We begin by defining, for any commutative ring R, an alternate multiplicative structure. Let R be a commutative ring and fix an element $k \in R$. We now define multiplication in $S_k(R)$ which, as an additive abelian group, is equal to R. For each pair of elements $r, s \in R$, we define the product of the corresponding elements $[r], [s] \in S_k(R)$ to be $[r][s] = [krs]$. The notation is convenient when distinguishing multiplication in R and in $S_k(R)$ and is motivated by the following (though less general) formulation of $S_k(R)$. With k a positive integer, we denote by $[r]$ the $k \times k$ single-valued matrix whose k^2 entries all equal r. With $S_k(R)$ the set of all such matrices over R and viewing R as a $\mathbb{Z}$-algebra so that

$$k \cdot r = \underbrace{r + \cdots + r}_{k} = kr,$$

we see that if $[r], [s] \in S_k(R)$, then $[r][s] = [krs]$ as in the original definition. With $R = \mathbb{Z}$, the ring of integers, and $k = 2$, this structure was introduced in [Jacobson 1965] to give examples of nonunique factorization of integers. This study was generalized in [Baeth et al. 2011] to $k \geq 2$ where more precise information about factorizations was gathered. Over the past several decades, factorization theory, and in particular the study of lengths of factorizations of elements in rings and semigroups, has become a major area of algebraic and combinatorial research. See, for example, the recent expository article [Geroldinger 2016] or the comprehensive text [Geroldinger and Halter-Koch 2006]. We will illustrate, using the structure of $S_k(R)$ where R is either a PID or the quotient of a PID, the existence of rings for which the factorization length set of every element is a discrete interval.

If R is a commutative ring, $R^\times$ denotes the set of *units* — elements with multiplicative inverses. Of course if R does not have a multiplicative identity, then $R^\times = \varnothing$. We say that an element $[r] \in S_k(R)$ is *irreducible* if it is impossible to write $[r] = [x][y]$ for any $[x], [y] \in S_k(R)$. In the cases of interest (see Setup 3.2) $S_k(R)$ has no units and this definition coincides with the usual definition of irreducibility in integral domains and cancellative semigroups and to the definition of *very strong irreducibles* as in [Ağargün et al. 2001; Anderson and Valdes-Leon 1996; 1997] in rings with zerodivisors. In this note we will first determine the set of irreducible elements of $S_k(R)$. Then, for each nonirreducible element $[r] \in S_k(R)$, we will compute its *length set*

$$\mathsf{L}([r]) = \{t : [r] = [x_1] \cdots [x_t] \text{ with each } [x_i] \text{ irreducible}\}.$$

This invariant is well-studied in the realm of cancellative commutative semigroups, see [Geroldinger and Halter-Koch 2006; Geroldinger 2016], and was computed for $S_k(\mathbb{Z})$ in [Baeth et al. 2011]. When R is either a principal ideal domain or a quotient of a principal ideal domain, we will show that $\mathsf{L}([r])$ is always either a singleton set or an interval of integers. When $a, b \in \mathbb{Z}$ with $a < b$, we denote by $[a, b]$ the discrete interval $\{a, a+1, \dots, b\}$. Similarly, $[a, \infty) = \{a, a+1, \dots\}$. Throughout, if D is PID, then for elements $x, y \in D$, we denote by $(x, y) = \{rx + sy : r, s \in R\}$ the ideal generated by x and y. A greatest common divisor d of x and y is an element r such that $(x, y) = (r)$. Note that with $D^\times$ denoting the set of units of D, $(x, y) = (r) = (s)$ if and only if $s = ru$ for some $u \in D^\times$.

In the remainder of this section, before turning our attention to proper quotients of PIDs, we generalize the results of [Baeth et al. 2011]. In Section 2 we give some preliminary results about the structure of $S_k(R)$ where R is the quotient of a PID. Our main results are contained in Section 3, where we describe factorizations of elements in $S_k(R)$ where R is a quotient of a PID.

The following lemma and theorem describe factorization in $S_k(D)$ where D is a PID. It should not be surprising that the results obtained here are essentially the same as those obtained in [Baeth et al. 2011], where $R = \mathbb{Z}$ (and k is a positive integer). In fact, the proofs of these results are only slightly modified from those in that paper and thus we do not include them here.

Lemma 1.1. *Let D be a PID, let $k \in D\backslash(D^\times \cup \{0\})$, and let $[a] \in S_k(D)$. Then $[a]$ is irreducible in $S_k(D)$ if and only if $k \nmid a$.*

For $a, b \in D$, we define $\nu_b(a)$ to be the largest integer m such that a is divisible by b^m. Then we have the following classification of length sets in $S_k(D)$ when D is a PID.

Theorem 1.2. *Let D be a PID, let $k \in D\backslash(D^\times \cup \{0\})$, and let $[a] \in S_k(D)$.*

(1) *If k is prime, then $|\mathsf{L}([a])| = 1$.*

(2) *If $k = p^m$ for some prime p, then*

$$\mathsf{L}([a]) = \left[\left\lceil \frac{\nu_p(a) + m}{2m - 1} \right\rceil, \nu_m(a) + 1\right].$$

(3) *If k is not the power of a prime, then $\mathsf{L}([a]) = [2, \nu_m(a) + 1]$.*

We note that if k is prime, then $S_k(D)$ is *half-factorial*; that is, the length set of any factorization is a singleton set. When k is not prime, each element has either a singleton length set or its length set is a discrete interval. When k is not the power of a prime, $S_k(D)$ is *bifurcus*; that is, every nonirreducible element can be represented as the product of two irreducible elements.

2. The structure of $S_k(D/(n))$

Throughout the next two sections, $R = D/(n)$, where D is a commutative principal ideal domain and n is a nonzero nonunit nonprime of D. For convenience we use the notation $\bar{x}$ to denote the coset $x+(n)$ in $D/(n)$. Before investigating factorization in $S_k(R)$ in Section 3, we give some preliminary results and make a few basic observations about $S_k(R)$. We begin by showing that $S_k(R)$ has no multiplicative identity except for in the trivial case, where $S_k(R) \cong R$.

Proposition 2.1. *Let $R = D/(n)$, where D is a PID and $n \in D \setminus (D^\times \cup \{0\})$. The following statements are equivalent*:

(1) 1 *is a greatest common divisor of k and n.*

(2) *$S_k(R)$ has a multiplicative identity.*

(3) *$S_k(R) \cong R$.*

Proof. If 1 is a greatest common divisor of k and n, there exist $x, y \in D$ with $kx + ny = 1$. Then, in R, $\bar{k}\bar{x} = \bar{1}$. For any $[\bar{a}] \in S_k(R)$, $[\bar{a}][\bar{x}] = [\overline{axk}] = [\bar{a}]$ and $[\bar{x}]$ is the multiplicative identity of $S_k(R)$. Conversely, suppose $S_k(n)$ has a multiplicative identity $[\bar{u}]$. Then $[\bar{1}][\bar{u}] = [\bar{1}]$ and so $\bar{u}\bar{k} = \bar{1}$ in $D/(n)$. But then $ku + nv = 1$ for some $v \in D$, and so 1 is a greatest common divisor of k and n. Therefore (1) and (2) are equivalent. The fact that (3) implies (2) is trivial since $R = D/(n)$ has a multiplicative identity. We now show that (1) implies (3). Since 1 is a greatest common divisor of k and n, we have $\overline{k^{-1}}\bar{k} = \bar{1}$ for some $k^{-1} \in D$. It is then trivial to check that the map $\varphi : D/(n) \to S_k(R)$ defined by $\varphi(\bar{a}) = [\overline{k^{-1}a}]$ is a ring isomorphism. □

Before investigating the multiplicative structure of $S_k(R)$, we note that k need only be considered modulo n. If $k \equiv k' \bmod n$ with $k, k' \in D$, then $\bar{k} = \overline{k'}$ in R and the following result is immediate.

Proposition 2.2. *Let $k \equiv k' \bmod n$.*

(1) *If $k' = 0$, then all nonzero elements of $S_k(R)$ are irreducible.*

(2) *If $k' \neq 0$, then $S_k(R) \cong S_{k'}(R)$.*

Suppose that $S_k(R) \ncong R$. Clearly $[\bar{0}]$ is a zerodivisor of $S_k(R)$. If $d \neq 1$ is a greatest common divisor of k and n, then $k = dy$ and $n = dz$ for some $y, z \in D$. Consider $[\overline{az}] \in S_k(R)$ with $a \in D$. Then

$$[\overline{az}][\bar{x}] = [\overline{kazx}] = [\overline{(dy)azx}] = [\overline{(dz)ayx}] = [\overline{(n)ayx}] = [\overline{(0)ayx}] = [\bar{0}]$$

for every $[\bar{x}] \in S_k(R)$. Thus we have the following result.

Proposition 2.3. *Let D be a PID and let $R = D/(n)$ for some nonzero nonunit n of D. If* 1 *is not a greatest common divisor of k and n, then all elements of $S_k(R)$ are zerodivisors.*

Note that what the argument preceding Proposition 2.3 really shows is that for each $a \in D$, with $z = n/d$ for some greatest common divisor d of k and n, the element $[\overline{az}] \in S_k(R)$ annihilates all elements of $S_k(R)$. Moreover, if $d \neq 1$ is a greatest common divisor of k and n, then $[\overline{az}] \neq [\bar{0}]$ for some $a \in D$. That is, an element of the form $[\overline{az}]$ is a sort of *psuedozero* as it annihilates all other elements of $S_k(R)$. This element $z \in D$ has an additional interesting property in terms of factorizations. Suppose $\bar{x} = \overline{az + c}$ and $\bar{y} = \overline{bz + c}$ for some $a, b, c \in D$. Then for all $[\bar{w}] \in S_k(R)$, we have $[\bar{x}][\bar{w}] = [\bar{c}][\bar{w}] = [\bar{y}][\bar{w}]$.

3. Length sets in $S_k(R)$

The goal of this section is to compute the length set $\mathsf{L}([\bar{x}])$ for each $[\bar{x}] \in S_k(D/(n))$. We will obtain results similar to those in Theorem 1.2 but find that for some $[\bar{x}]$, $\mathsf{L}([\bar{x}])$ is unbounded, much as is the case for some elements in $D/(n)$. We begin by recalling the following proposition, [Baeth et al. 2017, Theorem 3.4], that describes factorization in $D/(n)$ with the usual multiplication.

Proposition 3.1. *Let n be a nonzero nonprime element of a PID D and let $\bar{x} \in D/(n)$ with $\gcd(x, n) = d$. If $p \mid (n/d)$ for every prime divisor p of n, then $\bar{x}$ factors uniquely in $D/(n)$ and $\mathsf{L}_{D/(n)}(\bar{x}) = \{t\} = \mathsf{L}_D(d)$. Otherwise, $\bar{x}$ has infinitely many distinct factorizations in $D/(n)$ and $\mathsf{L}_{D/(n)}(\bar{x}) = [t, \infty)$, where $\mathsf{L}_D(d) = \{t\}$.*

Since factorization in $D/(n)$ is already understood, we focus on the case when $S_k(R) \ncong R$. Based on Propositions 2.1 and 2.2 we set some blanket hypotheses for the remainder of this manuscript.

Setup 3.2. Let D be a PID, let n be a nonzero nonunit of D and let $R = D/(n)$. Also let $k \in D$ be a nonzero nonunit in D with $n \nmid k$ and $(n, k) = (d) \neq D$.

First we classify the *irreducible elements* — elements that cannot be represented as a product of two nonzero elements of $S_k(R)$.

Proposition 3.3. *Let the notation be as in Setup 3.2. Then $[\bar{a}] \in S_k(R)$ is irreducible if and only if $d \nmid a$ in D.*

Proof. Suppose that $d \mid a$. Then $a \in (d) = (k, n)$ in D and so $a = kx + ny$ for some $x, y \in D$. But then $[\bar{a}] = [\overline{kx + ny}] = [\overline{kx}] = [\bar{1}][\bar{x}]$ is not irreducible in $S_k(R)$. Conversely, suppose that $[\bar{a}]$ is not irreducible in $S_k(R)$. Then $[\bar{a}] = [\bar{x}][\bar{y}] = [\overline{kxy}]$ for some $x, y \in D$. Then $\bar{a} = \overline{kxy}$ in $D/(n)$ and so $a = kxy + nz$ for some $z \in D$. Then, since $d \mid k$ and $d \mid n$, we know $d \mid a$. □

Now that we have classified the irreducible elements of $S_k(R)$, we work to compute the length sets of nonzero elements in $S_k(R)$. Throughout we will need the following definition. For $a \in D$, define $\nu_{(n,k)}(a)$, if it exists, to be the smallest positive integer m such that $\gcd(k^m, n) \nmid a$. This gives an analog to the valuation $\nu_b(a)$ which was used in the description of lower bounds of length sets in Theorem 1.2.

Remark 3.4. Note that if $R = D/(n)$ is the quotient of a PID D and $n = p_1^{t_1} \cdots p_s^{t_s}$ with $p_1, \dots, p_s$ distinct primes in D and $t_1, \dots, t_s$ positive integers, then the decomposition of R by the Chinese remainder theorem immediately gives a decomposition on $S_k(R)$ as $S_k(R) \cong S_k(D/(p_1^{t_1})) \times \cdots \times S_k(D/(p_s^{t_s}))$. One could then study factorization in $S_k(R)$ by piecing together information about factorization in each $S_k(D/(p_i^{t_i}))$. Though this simplifies some calculations, it obfuscates exactly how elements factor in $S_k(R)$. However, this decomposition does clarify the definition of $\nu_{(n,k)}(a)$ since

$$\nu_{(p^t,k)}(a) = \min_{m \geq 1}\{m : \min\{m\nu_p(k), t\} > \nu_p(a)\} = \left\lfloor \frac{\nu_p(a)}{\nu_p(k)} + 1 \right\rfloor$$

if p is a prime in D and k is a positive integer.

In the next proposition we investigate upper bounds on $\mathsf{L}([\bar{a}])$.

Proposition 3.5. *Let the notation be as in Setup 3.2. Let* $[\bar{a}] \in S_k(n)$:

(1) *If* $\nu_{(n,k)}(a)$ *exists*, *then* $\max \mathsf{L}([\bar{a}]) \leq \nu_{(n,k)}(a)$.

(2) *If* $\nu_{(n,k)}(a)$ *does not exist*, *then* $\mathsf{L}([\bar{a}])$ *is unbounded.*

Proof. Let $[\bar{a}] \in S_k(n)$ and assume that $\nu_{(n,k)}(a)$ exists. Suppose that $[\bar{a}] = \prod_{j=1}^{l}[\bar{b}_j]$, where each $[\bar{b}_j]$ is irreducible. Then $a \equiv k^{l-1}b_1 \cdots b_l \mod n$ and so $\gcd(k^{l-1}, n) \mid a$. Thus $l - 1 < \nu_{(n,k)}(a)$ and so $l \leq \nu_{(n,k)}(a)$. Now assume that $\nu_{(n,k)}(a)$ does not exist. That is, $\gcd(k^m, n) \mid a$ for all $m \geq 1$. For $m \geq 1$, set d_m to be a greatest common divisor of k^m and n. Then $d_m = k^m x + ny$ for some $x, y \in D$. Since $d_m \mid a$, we know $a = d_m b = k^m xb + nyb$ for some $b \in D$. Then $[\bar{a}] = [\bar{1}]^m[\overline{xb}]$. Since $[\bar{1}]$ is irreducible and since $[\overline{xb}]$ is either irreducible or can be factored as the product of irreducibles, $[\bar{a}]$ has a factorization of length at least $m + 1$. Since m was arbitrarily chosen, $\mathsf{L}([\bar{a}])$ is unbounded. □

We now show that if $\nu_{(n,k)}(a)$ exists, then $[\bar{a}]$ has a factorization of length $\nu_{(n,k)}(a)$. First we observe the following fact, which is immediate using the ideal inclusion $(a, b)(a^{m-1}, b) \subseteq (a^m, b)$.

Lemma 3.6. *Let D be a PID and let $a, b \in D$. If m is a positive integer*, *then* $\gcd(a^m, b) \mid \gcd(a, b)\gcd(a^{m-1}, b)$.

Proposition 3.7. *Let the notation be as in Setup 3.2. Let* $[\bar{a}] \in S_k(R)$ *and assume that* $\nu_{(n,k)}(a)$ *exists. Then* $\nu_{(n,k)}(a) \in \mathsf{L}([\bar{a}])$.

Proof. Clearly $[\bar{1}]$ is irreducible. We will show that there is $[\bar{b}] \in S_k(n)$ such that $[\bar{a}] = [\bar{b}][\bar{1}]^{\nu_{(n,k)}(a)-1}$ with $[\bar{b}]$ irreducible. Let $d' = \gcd(k^{\nu_{(n,k)}(a)-1}, n)$, $k' = k^{\nu_{(n,k)}(a)-1}/d'$, $a' = a/d'$, and $n' = n/d'$. Then $\gcd(k', n') = 1$ and so there exist $x, y \in D$ such that $n'x + k'y = 1$. Let $b = a'y$. Then

$$k^{\nu_{(n,k)}(a)-1}b = d'k'a'y = ak'y = a - xan' = a - a'xn,$$

whence $k^{v_{(n,k)}(a)-1}b \equiv a \bmod n$. We now show that $[\bar{b}]$ is irreducible. If $d \mid b$, then since $d' \mid k^{v_{(n,k)}(a)-1}$, we have $dd' = \gcd(k, n)\gcd(k^{v_{(n,k)}(a)-1}, n) \mid bk^{v_{(n,k)}(a)-1}$. Then, by Lemma 3.6, $\gcd(k^{v_{(n,k)}(a)}, n) \mid \gcd(k, n)\gcd(k^{v_{(n,k)}(a)-1}, n)$. This would imply $\gcd(k^{v_{(n,k)}(a)}, n) \mid bk^{v_{(n,k)}(a)-1}$ and $\gcd(k^{v_{(n,k)}(a)}, n) \mid n$. But $\gcd(k^{v_{(n,k)}(a)}, n) \nmid a$, contradicting $a \equiv k^{v_{(n,k)}(a)-1}b \bmod n$. Thus $[\bar{b}]$ is irreducible and $v_{(n,k)}(a) \in \mathsf{L}([\bar{a}])$. □

Now (1) of Proposition 3.5 becomes: if $v_{(n,k)}(a)$ exists, then $\max \mathsf{L}([\bar{a}]) = v_{(n,k)}(a)$.

For the remainder of this section we consider two cases. Let d be a greatest common divisor of k and n. First we suppose that d is not the power of a prime. In this case we show that $S_k(R)$ is bifurcus and hence $\mathsf{L}([\bar{a}]) = [2, \sup \mathsf{L}([\bar{a}])]$ for all nonirreducibles $[\bar{a}] \in S_k(R)$. We then consider when d is the power of some prime in D. In this case we compute the minimum value in $\mathsf{L}([\bar{a}])$ and again show that $\mathsf{L}([\bar{a}]) \subseteq [\min \mathsf{L}([\bar{a}]), \sup \mathsf{L}([\bar{a}])]$ with equality if k is also a prime power. In each case we explicitly give factorizations of $[\bar{a}]$ of each possible length. We begin with the simpler case when d is not a prime power.

Proposition 3.8. *Let the notation be as in Setup 3.2. Suppose that $d = st$ for some relatively prime $s, t \in D$. Then $2 \in \mathsf{L}([\bar{a}])$ for all nonzero nonirreducible $[\bar{a}] \in S_k(R)$.*

Proof. If $[\bar{a}]$ is not irreducible, then $d \mid a$. Then $a \in (d) = (n, k)$ and so $a = kx + ny$ for some $x, y \in D$. Write $x = d^r z$ with $r \geq 0$ and $d \nmid z$. Then, without loss of generality, $s \nmid z$. Now

$$[\bar{a}] = [\overline{kx}] = [\overline{kd^r z}] = [\overline{ks^r t^r z}] = [\overline{s^r}][\overline{t^r z}].$$

Since $d \nmid s^r$ and $d \nmid t^r z$, we have $[\overline{s^r}]$ and $[\overline{t^r z}]$ are irreducible. □

Since $2 \in \mathsf{L}([\bar{a}])$ for all nonzero nonirreducible $[\bar{a}] \in S_k(R)$, we know $S_k(R)$ is a finite bifurcus ring. This provides an affirmative answer to Open Problem 2.1.3 of [Adams et al. 2009].

Note that if $l \in \mathsf{L}([\bar{a}])$ with $l > 2$, then $[\bar{a}] = [\bar{b}_1] \cdots [\bar{b}_l]$ with each $[\bar{b}_i]$ irreducible. Since $S_k(R)$ is bifurcus, $[\bar{b}_1][\bar{b}_2][\bar{b}_3] = [\bar{c}_1][\bar{c}_2]$ for some $[\bar{c}_1]$, $[\bar{c}_2]$ irreducible. Then $[\bar{a}] = [\bar{c}_1][\bar{c}_2][\bar{b}_4] \cdots [\bar{b}_l]$ is a factorization of $[\bar{a}]$ of length $l - 1$. Therefore we have the following corollary.

Corollary 3.9. *Let the notation be as in Setup 3.2. Let $[\bar{a}] \in S_k(n)$. Let d be a greatest common divisor of k and n and suppose that d is not a prime power in D:*

(1) *If $v_{(n,k)}(a)$ exists, then $\mathsf{L}([\bar{a}]) = [2, v_{(n,k)}(a)]$.*

(2) *If $v_{(n,k)}(a)$ does not exist, then $\mathsf{L}([\bar{a}]) = [2, \infty)$.*

In addition to a complete description of the length sets of elements in $S_k(D/(n))$, if $\gcd(k, n)$ is not a prime power, then the ring is bifurcus and [Adams et al. 2009, Theorem 1.1] tells us also the *catenary degree* is $\mathsf{c}(S_k(D/(n))) = 3$ and the *tame* *degree* is $\mathsf{t}(S_k(D/(n))) = \infty$; see [Geroldinger and Halter-Koch 2006, Chapter 1.6] for definitions.

We now consider when a greatest common divisor of k and n is a prime power and set some notation for the remainder of this section. Let $n = xp^r$, $k = yp^s$, and $d = p^t$, where p is a prime in D, $p \nmid x, y$, and $r, s \geq 1$. Then $t = \min\{r, s\} \geq 1$. Moreover, since $\bar{y} \in D/(n)^{\times}$, there is $w \in D$ with $yw \equiv 1 \bmod n$. We will consider factorizations of $[\bar{a}] \in S_k(n)$ where $a = zp^u$ with $p \nmid z$. Note that in this setting, similar to Remark 3.4,

$$\nu_{(n,k)}(a) = \min_{m \geq 1}\bigl\{m : \min\{ms, r\} > u\bigr\}.$$

Therefore $\nu_{(n,k)}(a)$ exists if and only if $r > u$. When it does exist, $\nu_{(n,k)}(a) = \lfloor u/s + 1\rfloor$. Thus we consider two cases: $r > u$ and $r \leq u$. In each case we suppose that $l \in \mathsf{L}([\bar{a}])$; i.e., $[\bar{a}] = [\bar{a}_1] \cdots [\bar{a}_l]$ with each $[\bar{a}_i]$ irreducible so that $a \equiv k^{l-1} a_1 \cdots a_l \bmod n$ and hence $p^{s(l-1)} \mid a$.

First, suppose that $u < r$. We then consider two subcases determined by the relation of $(l-1)s$ to u and r. If $u < (l-1)s$, then $p^{s(l-1)} \nmid a$ and so $[\bar{a}]$ has no factorization of length l. Alternatively, $(l-1)s \leq u < r$. Since $p^{s(l-1)} \mid a$ and $a = zp^u$, we know $p^{u-(l-1)s} \mid a_1 \cdots a_l$. As each $[\bar{a}_i]$ is irreducible, $p^t \nmid a_i$ for each i. By the pigeonhole principle, $\lceil (u-(l-1)s)/(t-1)\rceil \leq l$. Conversely, suppose $j = \lceil (u-(l-1)s)/(t-1)\rceil \leq l$. Then

$$[\bar{a}] = [\overline{p^{u-(l-1)s-(t-1)(j-1)} w^{l-1} z}][\overline{p^{t-1}}]^{j-1}[\bar{1}]^{l-j}$$

is a factorization of $[\bar{a}]$ of length l. Thus, when $u < r$, we know $[\bar{a}]$ has a factorization of length l if and only if $\lceil (u-(l-1)s)/(t-1)\rceil \leq l$, equivalently $\lceil (u+s)/(t+s-1)\rceil \leq l \leq \lfloor u/s + 1\rfloor$.

Now suppose that $r \leq u$ and consider three subcases. First, suppose that $(l-1)s \leq r \leq u$. Then $p^{r-(l-1)s} \mid a_1 \cdots a_l$ and as in the case above, $\lceil (r-(l-1)s)/(t-1)\rceil \leq l$. Conversely, if $j = \lceil (r-(l-1)s)/(t-1)\rceil \leq l$, then

$$[\bar{a}][\overline{p^{r-(l-1)s-(t-1)(j-1)w^{l-1}} z(p^{u-r}+x)}][\overline{p^{t-1}}]^{j-1}[\bar{1}]^{l-j}$$

is a factorization of $[\bar{a}]$ of length l. Now suppose that $r \leq (l-1)s < u$. Note that if $p \mid (p^{u-(l-1)s} + x + mxp^r)$ for some m, then $p \mid x$. Thus $p \nmid (p^{u-(l-1)s} + x + mxp^r)$ for all $m \in D$ and so $[\overline{p^{u-(l-1)s} + x}]$ is irreducible and

$$[\bar{a}] = [\overline{p^{u-(l-1)s} + x}][\overline{w^{l-1}}][\bar{1}]^{l-2}$$

is a factorization of $[\bar{a}]$ of length l. Finally, suppose that $r \leq u \leq (l-1)s$. Since $(p, x) = 1$, there is $v \in D$ with $vp \equiv 1 \bmod x$. That is, $vp = 1 + xb$ for some $b \in D$ and so $vp \cdot p^r = (1+xb)p^r = p^r + nb \equiv p^r \bmod n$. In fact, $v^j p^{j+r} \equiv p^r \bmod n$

for all $j \geq 0$. Now, choosing $j > (l-1)s + r - u$,

$$[\overline{v^j p^{r+j+(u-r)-(l-1)s} + x}][\overline{w^{l-1}z}][\bar{1}]^{l-2}$$

is a factorization of $[\bar{a}]$ of length l. Thus, when $r \leq u$, we know $[\bar{a}]$ has a factorization of length l if and only if $l \geq \lceil (r+s)/(t+s-1) \rceil$. In summary, we have the following proposition.

Proposition 3.10. *Let the notation be as in Setup 3.2. Let $[\bar{a}] \in S_k(n)$. Let $n = xp^r$, $k = yp^s$, $d = p^t$, and $a = zp^u$, where p is a prime in D, $p \nmid x, y, z$, and $r, s \geq 1$:*

(1) *If $\nu_{(n,k)}(a)$ exists, then $\mathsf{L}([\bar{a}]) = [\lceil (u+s)/(t+s-1) \rceil, \nu_{(n,k)}(a)]$.*

(2) *If $\nu_{(n,k)}(a)$ does not exist, then $\mathsf{L}([\bar{a}]) = [\lceil (r+s)/(t+s-1) \rceil, \infty)$.*

Even though $S_k(D/(n))$ is not bifurcus if $\gcd(k, n)$ is a prime power, we can still bound the catenary degree and compute the tame degree. Since for any $[\bar{a}] \in S_k(D/(n))$, we have $\min \mathsf{L}([\bar{a}]) \leq \lceil (r+s)/(t+s-1) \rceil$, an argument analogous to that of [Adams et al. 2009, Theorem 1.1] gives that $\mathsf{c}(S_k(D/(n))) \leq \lceil (r+s)/(t+s-1) \rceil$. Since there exist elements with arbitrarily long factorization lengths, [Geroldinger and Halter-Koch 2006, Theorem 1.6.6] gives that $\mathsf{t}(S_k(D/(n))) \geq \rho(S_k(D/(n)) = \infty$.

In conclusion, whenever $[\bar{a}] \in S_k(R)$ with $(k, n) \neq D$, we have $\mathsf{L}([\bar{a}]) = [\min \mathsf{L}([\bar{a}]), \sup \mathsf{L}([\bar{a}])]$, with $\sup \mathsf{L}([\bar{a}]) = \infty$, if and only if $\nu_{(n,k)}(a)$ does not exist. Together, Corollary 3.9 and Proposition 3.10 completely describe the length sets of elements in the ring $S_k(R)$ subject to the conditions laid out in Setup 3.2. The remaining cases are either trivial or are dealt with in Theorems 1.2 and 3.1. Moreover, the catenary degree is always bounded and the tame degree is always infinite.

References

[Adams et al. 2009] D. Adams, R. Ardila, D. Hannasch, A. Kosh, H. McCarthy, V. Ponomarenko, and R. Rosenbaum, "Bifurcus semigroups and rings", *Involve* **2**:3 (2009), 351–356. MR Zbl

[Ağargün et al. 2001] A. G. Ağargün, D. D. Anderson, and S. Valdes-Leon, "Factorization in commutative rings with zero divisors, III", *Rocky Mountain J. Math.* **31**:1 (2001), 1–21. MR Zbl

[Anderson and Valdes-Leon 1996] D. D. Anderson and S. Valdes-Leon, "Factorization in commutative rings with zero divisors", *Rocky Mountain J. Math.* **26**:2 (1996), 439–480. MR Zbl

[Anderson and Valdes-Leon 1997] D. D. Anderson and S. Valdes-Leon, "Factorization in commutative rings with zero divisors, II", pp. 197–219 in *Factorization in integral domains* (Iowa City, 1996), edited by D. D. Anderson, Lecture Notes in Pure and Appl. Math. **189**, Dekker, New York, 1997. MR Zbl

[Baeth et al. 2011] N. Baeth, V. Ponomarenko, D. Adams, R. Ardila, D. Hannasch, A. Kosh, H. McCarthy, and R. Rosenbaum, "Number theory of matrix semigroups", *Linear Algebra Appl.* **434**:3 (2011), 694–711. MR Zbl

[Baeth et al. 2017] N. Baeth, B. Burns, and J. Mixco, "A fundamental theorem of modular arithmetic", *Period. Math. Hungar.* **75** (2017), 356–367.

[Geroldinger 2016] A. Geroldinger, "Sets of lengths", *Amer. Math. Monthly* **123**:10 (2016), 960–988. MR Zbl

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

[Jacobson 1965] B. Jacobson, "Matrix number theory: an example of nonunique factorization", *Amer. Math. Monthly* **72**:4 (1965), 399–402. MR Zbl

Received: 2017-05-31 Accepted: 2017-08-14

baeth@ucmo.edu *School of Computer Science and Mathematics, University of Central Missouri, Warrensburg, MO, United States*

branjburns@gmail.com *Americo Financial Life and Annuities, Kansas City, MO, United States*

jcovey94@gmail.com *Department of Mathematics, Washington University in Saint Louis, Saint Louis, MO, United States*

james.mixco@slu.edu *Department of Mathematics and Computer Science, Saint Louis University, Saint Louis, MO, United States*

INVOLVE 11:4 (2018)
dx.doi.org/10.2140/involve.2018.11.711

Locating trinomial zeros

Russell Howell and David Kyle

(Communicated by Michael Dorff)

We derive formulas for the number of interior roots (i.e., zeros with modulus less than 1) and exterior roots (i.e., zeros with modulus greater than 1) for trinomials of the form $z^n + z^k - 1$, where $1 \le k \le n-1$. Combined with earlier work by Brilleslyper and Schaubroeck, who focus on unimodular roots (i.e., zeros that lie on the unit circle), we give a complete count of the location of zeros of these trinomials.

1. Introduction

The investigation of zeros of analytic functions has a long and rich history, with many important results focusing on specialized cases. Indeed, the study of zeros of trinomials dates to the 19th century, and a recent paper by Melman [2012] gives historical references in addition to providing information on the location of zeros. Even more recently, [Brilleslyper and Schaubroeck 2014], which won a Pólya award, investigated trinomials of the form

$$p(z) = z^n + z^k - 1 \quad (n \ge 2,\ 1 \le k \le n-1). \tag{1}$$

Their main result characterizes the *unimodular roots* (i.e., zeros that lie on the unit circle) of $p(z)$:

Theorem 1. *Let $p(z) = z^n + z^k - 1$ and let $g = \gcd(n,k)$. If 6 divides $n/g + k/g$, then p has exactly $2g$ unimodular roots, occurring in conjugate pairs z_m and $\bar{z}_m$, determined by $z_m = \exp[i(\pi/(3g) + 2\pi m/g)]$, where $0 \le m \le g-1$.*

In that paper, they called for the discovery of a formula (involving n and k) that would calculate the number of *interior roots* (i.e., zeros with modulus less than 1) of these trinomials. In [Brilleslyper and Schaubroeck $\ge$ 2018] they developed a conjecture,

$$\text{number of interior roots} = 2g\left\lfloor \frac{n+k-g}{6g} \right\rfloor + g,$$

and proved it for the special case when $k = 1$.

MSC2010: primary 30C15, 97I80; secondary 11-02.
Keywords: trinomials, complex analysis, Diophantine equations, zeros of functions.

Here we show that their conjecture is correct in general. Specifically we prove, for $1 \le k \le n-1$, the equivalent formula

$$\text{number of interior roots} = 2g\left\lceil \frac{n+k}{6g} \right\rceil - g. \tag{2}$$

Our proof proceeds in three steps.

First, we show that any interior root must lie in what we call an *interior region*, that any such region contains at most one root, and that the maximum number of these regions matches (2). Next, we show that a similar situation holds for *exterior roots* (i.e., zeros with modulus greater than 1) with respect to *exterior regions*, where the maximum number of these regions matches (3), given by

$$\text{number of exterior roots} = n - 2g\left\lfloor \frac{n+k}{6g} \right\rfloor - g. \tag{3}$$

Finally, we show that adding together the number of unimodular roots (if any), the maximum number of interior regions, and the maximum number of exterior regions results in n, the degree of the trinomial, so that these regions contain exactly one root.

We begin by analyzing where interior roots must be located. To do so, we generally follow the approach in [Brilleslyper and Schaubroeck ≥ 2018], but with some modifications. Throughout, the term *trinomial* and the notation $p(z)$ designate a function as defined in (1).

2. The location of interior roots

In what follows we suppose $p(z_0) = 0$ for some z_0 with $|z_0| < 1$.

2.1. ***Native zones for interior roots.*** The assumption that $p(z_0) = 0$ leads to the equation $z_0^k(z_0^{n-k}+1) = 1$. Using the additional assumption that $|z_0| < 1$ and taking the modulus of both sides reveal that $|z_0^k| < 1$ and $|z_0^{n-k}+1| > 1$. Thus, z_0^{n-k} must lie outside the circle $|z+1| = 1$, and z_0^k must lie inside the circle $|z| = 1$. But if $|z_0^k| < 1$, then $|z_0^{n-k}| < 1$ as well, so z_0^{n-k} must also lie inside the circle $|z| = 1$. The two circles intersect at points whose arguments are $\pm\frac{2}{3}\pi$, so $\operatorname{Arg}(z_0^{n-k}) \in \left(-\frac{2}{3}\pi, \frac{2}{3}\pi\right)$. It follows that the point z_0 itself must lie inside one of $n-k$ possible disjoint regions, which we dub *native zones*:

$$N_m = \left\{ re^{i\theta} : \theta \in \left(-\frac{2\pi}{3(n-k)} + m\frac{2\pi}{(n-k)}, \frac{2\pi}{3(n-k)} + m\frac{2\pi}{(n-k)} \right) \right\}, \tag{4}$$

where $0 < r < 1$ and $m \in \mathbb{Z}$.

Although there are only $n-k$ distinct native zones N_m, we allow the index m to range over the integers. Doing so will assist us later in counting the number of these zones satisfying certain restrictions.

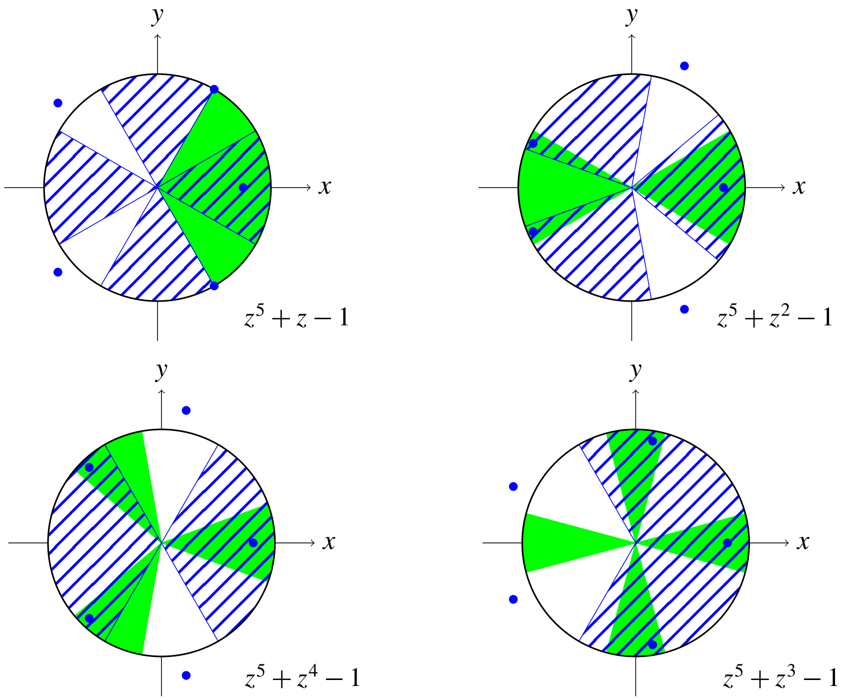


Figure 1. The unit disk with native zones (hatched), echo zones (shaded), and roots (large dots) for the trinomials z^5+z^k-1, where $1 \le k \le 4$.

2.2. ***Echo zones for interior roots.*** We can get further information on the location of z_0 by considering a related polynomial $q(z)$ defined by

$$q(z) = -z^n p(1/z) = z^n - z^{n-k} - 1.$$

A straightforward calculation reveals that $p(z) = 0$ if and only if $q(1/\bar{z}) = 0$.

Let $w_0 = 1/\bar{z}_0$, and note that $\mathrm{Arg}(z_0) = \mathrm{Arg}(w_0)$. Thus, z_0 and w_0 are echos of each other across the unit circle, and are zeros, respectively, of $p(z)$ and $q(z)$.

Write $q(w_0) = {w_0}^n - {w_0}^{n-k} - 1 = 0$ as $w_0^{n-k}({w_0}^k - 1) = 1$. Taking the modulus of both sides reveals that $|w_0^{n-k}| > 1$ (because $|z_0| < 1$) and $|{w_0}^k - 1| < 1$. Using an analysis similar to that which led to the definition of native zones enables us to conclude that $\mathrm{Arg}(w_0^k) = \mathrm{Arg}(z_0^k) \in \left(-\frac{1}{3}\pi, \frac{1}{3}\pi\right)$. It follows that the point z_0 itself must lie inside one of k possible disjoint regions E_j, which we call *echo zones*:

$$E_j = \left\{ re^{i\theta} : \theta \in \left(-\frac{\pi}{3k} + j\frac{2\pi}{k}, \frac{\pi}{3k} + j\frac{2\pi}{k} \right) \right\}, \tag{5}$$

where $0 < r < 1$ and $j \in \mathbb{Z}$.

As with the native zones, we allow the index j for the echo zones E_j to range over the integers.

2.3. ***Interior regions for interior roots.*** The preceding analysis shows that any interior root must lie in a nonempty intersection of a native zone and an echo zone, which we call an *interior region*. Figure 1 depicts this result for the trinomials $z^n + z^k - 1$, where $n = 5$ and $1 \le k \le 4$. Note that every interior root is in an interior region, and, in the case of Figure 1 (upper-left), there are $2g = 2$ unimodular roots as guaranteed by Theorem 1. Further, extending the radii of native and echo zones indicates that every exterior root is in neither a native nor an echo zone. The next section shows more precisely that these roots must be located in what we call *exterior regions*.

3. The location of exterior roots

Under the hypothesis that $p(z_0) = 0$, where $|z_0| > 1$, the same process for analyzing interior roots can be used to show that all exterior roots belong to an intersection of an *exterior native zone* and an *exterior echo zone*, defined respectively as

$$EN_m = \left\{ re^{i\theta} : \theta \in \left(\frac{2\pi}{3(n-k)} + m\frac{2\pi}{(n-k)},\ \frac{4\pi}{3(n-k)} + m\frac{2\pi}{(n-k)} \right) \right\}, \tag{6}$$

where $1 < r < \infty$, $m \in \mathbb{Z}$; and

$$EE_j = \left\{ re^{i\theta} : \theta \in \left(\frac{\pi}{3k} + j\frac{2\pi}{k},\ \frac{5\pi}{3k} + j\frac{2\pi}{k} \right) \right\}, \tag{7}$$

where $1 < r < \infty$, $j \in \mathbb{Z}$. As with the corresponding native and echo zones, we allow m and j to range over the integers.

We call any nonempty intersection of (6) and (7) an *exterior region*.

4. Upper bounds for roots

The last two sections collectively show that every interior root must belong to an interior region, and every exterior root must belong to an exterior region. In this section we establish that each such region contains at most one root.

In proving (2) for the case when $k = 1$, Brilleslyper and Schaubroeck demonstrated that exactly one root of $p(z)$ resides in each of the disjoint angular regions

$$R_a = \left\{ re^{i\theta} : \theta \in \left(\frac{2a\pi}{n} - \frac{\pi}{2n},\ \frac{2a\pi}{n} + \frac{\pi}{2n} \right) \right\}, \tag{8}$$

where $0 < r < 2$ and $0 \le a \le n-1$.

They called these regions *Rouché sectors* [Brilleslyper and Schaubroeck ≥ 2018], an appropriate choice because their demonstration makes creative use of Rouché's theorem, which can be found in almost any standard text for a first course in complex analysis [Mathews and Howell 2012, pp. 340–341]. For completeness we state the theorem here.

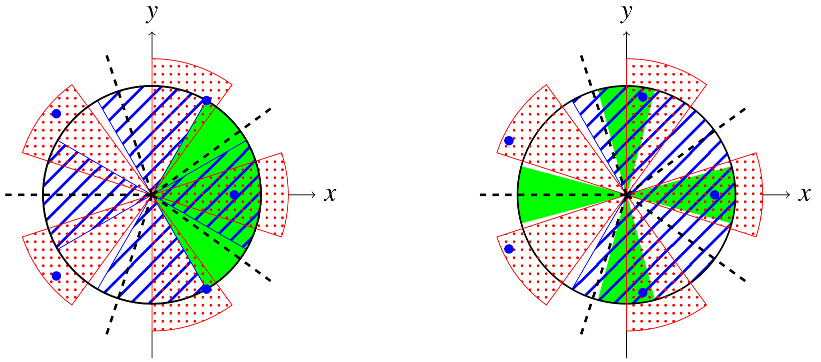


Figure 2. Native zones (hatched), echo zones (shaded), Rouché sectors (dotted), and roots (large dots) for $z^5 + z - 1$ (left) and $z^5 + z^4 - 1$ (right). The dashed lines are midway between the Rouché sectors.

Theorem 2 (Rouché's theorem). *Let Γ be a simple closed positively oriented contour in $\mathbb{C}$, and let f and g be analytic functions in a simply connected domain that contains Γ. If $|f(z) - g(z)| < |g(z)|$ for all $z \in \Gamma$, then f and g have the same number of zeros inside Γ.*

The demonstration that $p(z)$ has a zero (henceforth root) in any sector R_a comes from applying Rouché's theorem to the functions $f(z) = p(z)$ and $g(z) = z^n - 1$ evaluated on the boundary of the sector defined in (8). Each sector is centered around only one n-th root of unity, so $g(z)$ has exactly one root in each. Therefore, $p(z)$ has exactly one root in each Rouché sector.

Figure 2 illustrates this situation for the trinomials $z^5 + z - 1$ and $z^5 + z^4 - 1$, where all interior roots lie in the intersection of an interior region and a Rouché sector, and all exterior roots lie in the intersection of an exterior region and a Rouché sector. In each case the number of interior and exterior regions match, respectively, (2) and (3).

Now, if an interior region contained more than one root, then that region would have to intersect at least two Rouché sectors, and for some integer a contain one of the rays $\{z = re^{i\theta_a} : 0 < r < 1\}$, where $\theta_a = \pi/n + 2\pi a/n$, which is midway between the respective Rouché sectors (see Figure 2).

Suppose that some ray $z = re^{i\theta_a}$ were in an interior region. Then, for some integers m and j, we have $re^{i\theta_a} \in N_m$ and $re^{i\theta_a} \in E_j$ for $0 < r < 1$. According to the definitions of N_m and E_j, see (4) and (5), we thus get the inequalities

$$-\frac{2\pi}{3(n-k)} + m\frac{2\pi}{(n-k)} < \frac{\pi}{n} + a\frac{2\pi}{n} < \frac{2\pi}{3(n-k)} + m\frac{2\pi}{(n-k)},$$

that is,

$$-\frac{5}{3} < -\frac{k}{n} + 2a - a\frac{2k}{n} - 2m < -\frac{1}{3}, \tag{9}$$

if θ_a were in a native zone, and

$$-\frac{\pi}{3k} + j\frac{2\pi}{k} < \frac{\pi}{n} + a\frac{2\pi}{n} < \frac{\pi}{3k} + j\frac{2\pi}{k},$$

that is,

$$-\frac{1}{3} < \frac{k}{n} + a\frac{2k}{n} - 2j < \frac{1}{3}, \tag{10}$$

if θ_a were in an echo zone. Combining (9) and (10) gives

$$-2 < 2a - 2n - 2j < 0 \quad \text{or} \quad -1 < a - n - j < 0,$$

which is impossible because j, m, and a are integers.

By the same process we can determine that no ray $z = re^{i\theta_a}$ is in an exterior region, so that each exterior region has at most one root.

Thus, an upper bound for the number of interior and exterior roots is, respectively, the number of interior and exterior regions. The next few sections establish that the maximum number of these regions matches (2) and (3).

5. Counting interior regions

Each native and echo zone has the general form $\{re^{i\theta} : \alpha < \theta < \beta,\ 0 < r < 1\}$. To simplify language we will call the ray $z = re^{i\beta}$, where $0 < r < 1$, the *right border* of the given zone. (For exterior zones, of course, $1 < r < \infty$.) With this understanding, we proceed to count how many interior regions there are for a given trinomial $p(z)$, where a working assumption will be $\gcd(n, k) = 1$. In a subsequent section we show how to extend this assumption to the case when $\gcd(n, k) = g > 1$.

Recall that an interior region consists of a nonempty intersection $N_m \cap E_j$ of a native and echo zone. Figure 3 illustrates that there are three cases to consider for such an intersection: the right border of an echo zone belongs to a native zone (Figure 3, left), the right border of a native zone belongs to an echo zone (Figure 3, center) or their right borders coalign (Figure 3, right). Our task is to count the interior regions in each case.

Case 1: The right border of an echo zone belongs to a native zone (Figure 3, left).

Then, by (4) and (5), for some $j,\ m \in \mathbb{Z}$,

$$\begin{gathered} -\frac{2\pi}{3(n-k)} + m\frac{2\pi}{(n-k)} < \frac{\pi}{3k} + j\frac{2\pi}{k} < \frac{2\pi}{3(n-k)} + m\frac{2\pi}{(n-k)} \quad \text{or} \\ -\frac{n+k}{6} < j(n-k) - mk < \frac{3k-n}{6}. \end{gathered} \tag{11}$$

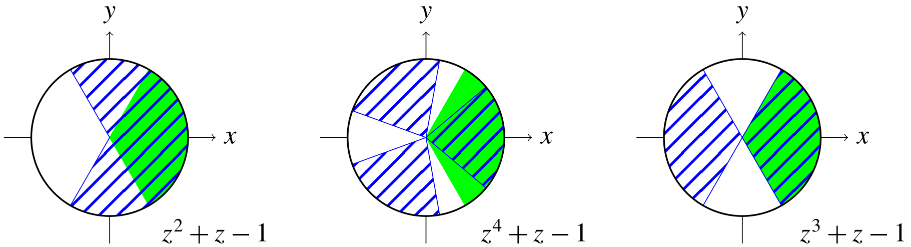


Figure 3. Trinomials illustrating that either the right border of an echo zone belongs to a native zone (left), the right border of a native zone belongs to an echo zone (center), or their right borders coalign (right).

To count the interior regions in this category, we first determine all values of m and j satisfying (11). By a standard result in number theory (see, for example, [Uspensky and Heaslet 1939, pp. 54–57]) we know that, because $\gcd(n-k,k)=1$, the Diophantine equation $j(n-k)-mk=c$ has a solution j_c, m_c for any integer $c \in \left(-\frac{1}{6}(n+k), \frac{1}{6}(3k-n)\right)$. Furthermore, the set of all solutions is given by

$$j = j_c + kt \quad \text{and} \quad m = m_c + (n-k)t \quad \text{for} \quad t \in \mathbb{Z}. \tag{12}$$

According to (4) and (5), $E_{j_c} = E_{j_c+kt}$ and $N_{m_c} = N_{m_c+(n-k)t}$ for all $t \in \mathbb{Z}$. Hence, from solution set (12), we see that to every integer $c \in \left(-\frac{1}{6}(n+k), \frac{1}{6}(n-3k)\right)$ there corresponds exactly one interior region $N_{m_c} \cap E_{j_c}$. In other words, the maximum number of interior regions in this category — and thus the maximum number of interior roots — is the number of integers between $-\frac{1}{6}(n+k)$ and $\frac{1}{6}(3k-n)$. The number of integers in an open interval (a,b) for $a,b \in \mathbb{R}$ is the difference between the last integer and first integer plus 1, that is, $(\lceil b \rceil - 1) - (\lfloor a \rfloor + 1) + 1$. Combining that fact with the result that, for $x \in \mathbb{R}$, $\lfloor -x \rfloor = -\lceil x \rceil$, yields a formula for the number of integers in the interval $\left(-\frac{1}{6}(n+k), \frac{1}{6}(3k-n)\right)$, and thus the maximum number of interior regions for Case 1:

$$\left(\left\lceil \frac{3k-n}{6} \right\rceil - 1\right) - \left(\left\lfloor -\frac{n+k}{6} \right\rfloor + 1\right) + 1 = \left\lceil \frac{3k-n}{6} \right\rceil + \left\lceil \frac{n+k}{6} \right\rceil - 1. \tag{13}$$

Cases 2 *and* 3: The right border of a native zone belongs to an echo zone, or their right borders coalign (Figure 3, center and right, respectively).

Then, for some $j, m \in \mathbb{Z}$,

$$\begin{gathered} -\frac{\pi}{3k} + j\frac{2\pi}{k} < \frac{2\pi}{3(n-k)} + m\frac{2\pi}{(n-k)} \le \frac{\pi}{3k} + j\frac{2\pi}{k} \quad \text{or} \\ -\frac{n+k}{6} < mk - j(n-k) \le \frac{n-3k}{6}. \end{gathered} \tag{14}$$

By using (14) and the same analysis as in Case 1, we find that there is exactly one interior region for each integer in the interval $\bigl(-\frac{1}{6}(n+k), \frac{1}{6}(n-3k)\bigr]$. The last integer in this interval is $\bigl\lfloor \frac{1}{6}(n-3k)\bigr\rfloor$ and the first integer is $\bigl\lfloor -\frac{1}{6}(n+k)\bigr\rfloor+1$. Therefore, the number of integers in the interval $\bigl(-\frac{1}{6}(n+k), \frac{1}{6}(n-3k)\bigr]$, and thus the maximum number of interior regions in this category, is

$$\left\lfloor \frac{n-3k}{6}\right\rfloor - \left(\left\lfloor -\frac{n+k}{6}\right\rfloor + 1\right) + 1 = -\left\lceil \frac{3k-n}{6}\right\rceil + \left\lceil \frac{n+k}{6}\right\rceil. \tag{15}$$

Combining the cases: Adding together (13) and (15) gives the desired formula for the maximum number of interior regions, and therefore the maximum number of interior roots when $\gcd(n,k)=1$:

$$2\left\lceil \frac{n+k}{6}\right\rceil - 1. \tag{16}$$

6. Counting exterior regions

Again using the assumption that $\gcd(n,k)=1$, we now obtain counts for exterior regions. As with interior regions, we have three cases to consider: the right border of an exterior echo zone (7) belongs to an exterior native zone (6), the right border of an exterior native zone belongs to an exterior echo zone, or their right borders coalign.

With the same techniques used in the previous section, we find that, in the first case, we must count the integers in the interval

$$\left(\frac{n+k}{6} - n + k, \frac{n+3k}{6} - n + k\right).$$

The identities $\lfloor x+n\rfloor = \lfloor x\rfloor + n$ and $\lceil x+n\rceil = \lceil x\rceil + n$ (valid for $n \in \mathbb{Z}$ and $x \in \mathbb{R}$) assist in obtaining the following count:

$$\left(\left\lceil \frac{n+3k}{6} - n + k\right\rceil - 1\right) - \left(\left\lfloor \frac{n+k}{6} - n + k\right\rfloor + 1\right) + 1 = \left\lceil \frac{n+3k}{6}\right\rceil - \left\lfloor \frac{n+k}{6}\right\rfloor - 1. \tag{17}$$

For the last two cases combined we must count the integers in the interval

$$\left(\frac{n+k}{6} - k, -\frac{n+3k}{6} + n - k\right].$$

Floor and ceiling function identities then assist in yielding the following amount:

$$\left\lfloor -\frac{n+3k}{6} + n - k\right\rfloor - \left(\left\lfloor \frac{n+k}{6} - k\right\rfloor + 1\right) + 1 = -\left\lceil \frac{n+3k}{6}\right\rceil + n - \left\lfloor \frac{n+k}{6}\right\rfloor. \tag{18}$$

Adding together the counts described in (17) and (18) reveals that the maximum number of exterior regions is

$$n - 2\left\lfloor \frac{n+k}{6}\right\rfloor - 1, \tag{19}$$

which is thus an upper bound for the maximum number of exterior roots when $\gcd(n,k)=1$.

7. Verifying the general formulas

For the interior roots of $p(z)=z^n+z^k-1$, where $\gcd(n,k)=g>1$, we appeal to the related polynomial $\tilde{p}(z)=z^{n/g}+z^{k/g}-1$. From [Brilleslyper and Schaubroeck 2014, Lemma 2], we know that the roots of $p(z)$ are in g-to-one correspondence with the roots of $\tilde{p}(z)$, and this correspondence does not disrupt the classification of roots into interior, unimodular, or exterior categories. Since $\gcd(n/g,k/g)=1$, we can use n/g and n/k, respectively, in place of n and k in (16) to get the maximum number of interior roots for $\tilde{p}(z)$:

$$2\left\lceil\frac{n/g+k/g}{6}\right\rceil-1=2\left\lceil\frac{n+k}{6g}\right\rceil-1.$$

The maximum number of interior roots for $p(z)$, then, is $2g\lceil(n+k)/(6g)\rceil-g$, which is exactly (2).

Using the same procedure, it can be shown that, when $\gcd(n,k)=g>1$, (19) morphs to give $n-2g\lfloor(n+k)/(6g)\rfloor-g$ as the maximum number of exterior roots for $p(z)$, which is exactly (3).

To complete our analysis we note that, when there are no unimodular roots, (2) and (3), when added together, give the maximum number of roots for $p(z)$:

$$\left(2g\left\lceil\frac{n+k}{6g}\right\rceil-g\right)+\left(n-2g\left\lfloor\frac{n+k}{6g}\right\rfloor-g\right). \tag{20}$$

When $p(z)$ has unimodular roots, Theorem 1 guarantees that the maximum number of roots it has is

$$2g+\left(2g\left\lceil\frac{n+k}{6g}\right\rceil-g\right)+\left(n-2g\left\lfloor\frac{n+k}{6g}\right\rfloor-g\right). \tag{21}$$

But according to Theorem 1, unimodular roots occur precisely when $6g$ divides $n+k$. Thus, $\lceil(n+k)/(6g)\rceil=\lfloor(n+k)/(6g)\rfloor+1$ in (20), and $\lceil(n+k)/(6g)\rceil=\lfloor(n+k)/(6g)\rfloor$ in (21). In both cases, then, the expressions sum to n, which equals the total number of roots for $p(z)$. Because interior and exterior regions are the only possible locations for interior and exterior roots, the maximum numbers of interior and exterior roots as expressed in (2) and (3) must be attained.

The enumeration of interior, exterior, and unimodular roots of trinomials $p(z)$ is now complete. For convenience, we summarize the results.

Theorem 3. *For $n\geq 2$, $1\leq k\leq n-1$, and $g=\gcd(n,k)$, the trinomial $p(z)=z^n+z^k-1$ has $2g\lceil(n+k)/(6g)\rceil-g$ interior roots, $n-2g\lfloor(n+k)/(6g)\rfloor-g$ exterior roots, and, when $6g$ divides $n+k$, it has $2g$ unimodular roots.*

References

[Brilleslyper and Schaubroeck 2014] M. Brilleslyper and L. Schaubroeck, "Locating unimodular roots", *College Math. J.* **45**:3 (2014), 162–168. MR

[Brilleslyper and Schaubroeck $\geq$ 2018] M. Brilleslyper and L. Schaubroeck, "Counting interior roots of trinomials", to appear in *Math. Mag.*

[Mathews and Howell 2012] J. H. Mathews and R. W. Howell, *Complex analysis for mathematics and engineering*, 6th ed., Jones & Bartlett Learning, Burlington, MA, 2012. Zbl

[Melman 2012] A. Melman, "Geometry of trinomials", *Pacific J. Math.* **259**:1 (2012), 141–159. MR Zbl

[Uspensky and Heaslet 1939] J. V. Uspensky and M. A. Heaslet, *Elementary number theory*, McGraw-Hill, New York, 1939. MR Zbl

Received: 2017-07-21 Accepted: 2017-08-29

howell@westmont.edu *Department of Mathematics, Westmont College,*
Santa Barbara, CA, United States

dkyle@westmont.edu *Department of Mathematics, Westmont College,*
Santa Barbara, CA, United States

Guidelines for Authors

Submissions in all mathematical areas are encouraged. All manuscripts accepted for publication in *Involve* are considered publishable in quality journals in their respective fields, and include a minimum of one-third student authorship. Submissions should include substantial faculty input; faculty co-authorship is strongly encouraged. Authors may submit manuscripts in PDF format on-line at the Submission page at the Involve website.

Originality. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles in *Involve* are usually in English, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

Format. Authors are encouraged to use LaTeX but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@msp.org with details about how your graphics were generated.

White space. Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Proofs. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

involve

2018 vol. 11 no. 4

Modeling of breast cancer through evolutionary game theory 541
KE'YONA BARTON, CORBIN SMITH, JAN RYCHTÁŘ AND TSVETANKA SENDOVA

The isoperimetric problem in the plane with the sum of two Gaussian densities 549
JOHN BERRY, MATTHEW DANNENBERG, JASON LIANG AND YINGYI ZENG

Finiteness of homological filling functions 569
JOSHUA W. FLEMING AND EDUARDO MARTÍNEZ-PEDROZA

Explicit representations of 3-dimensional Sklyanin algebras associated to a point of order 2 585
DANIEL J. REICH AND CHELSEA WALTON

A classification of Klein links as torus links 609
STEVEN BERES, VESTA COUFAL, KAIA HLAVACEK, M. KATE KEARNEY, RYAN LATTANZI, HAYLEY OLSON, JOEL PEREIRA AND BRYAN STRUB

Interpolation on Gauss hypergeometric functions with an application 625
HINA MANOJ ARORA AND SWADESH KUMAR SAHOO

Properties of sets of nontransitive dice with few sides 643
LEVI ANGEL AND MATT DAVIS

Numerical studies of serendipity and tensor product elements for eigenvalue problems 661
ANDREW GILLETTE, CRAIG GROSS AND KEN PLACKOWSKI

Connectedness of two-sided group digraphs and graphs 679
PATRECK CHIKWANDA, CATHY KRILOFF, YUN TECK LEE, TAYLOR SANDOW, GARRETT SMITH AND DMYTRO YEROSHKIN

Nonunique factorization over quotients of PIDs 701
NICHOLAS R. BAETH, BRANDON J. BURNS, JOSHUA M. COVEY AND JAMES R. MIXCO

Locating trinomial zeros 711
RUSSELL HOWELL AND DAVID KYLE