involve

a journal of mathematics

msp

2019 vol. 12, 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.

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2019 is US $195/year for the electronic version, and $260/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 12:4 (2019)
dx.doi.org/10.2140/involve.2019.12.541

Euler's formula for the zeta function at the positive even integers

Samyukta Krishnamurthy and Micah B. Milinovich

(Communicated by Filip Saidak)

We give a new proof of Euler's formula for the values of the Riemann zeta function at the positive even integers. The proof involves estimating a certain integral of elementary functions two different ways and using a recurrence relation for the Bernoulli polynomials evaluated at $\frac{1}{2}$.

1. Introduction

Let $\zeta(s)$ denote the Riemann zeta function and let $\eta(s) = (1-2^{1-s})\zeta(s)$. Then the series representations

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} \quad \text{and} \quad \eta(s) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n^s}$$

converge absolutely in the half-plane $\mathrm{Re}(s) > 1$. For $n \in \mathbb{N}$, we define the Bernoulli polynomials $B_n(x)$ via the generating function

$$\frac{ze^{xz}}{e^z - 1} = \sum_{n=0}^{\infty} B_n(x) \frac{z^n}{n!},$$

and (as usual) we call $B_n := B_n(0)$ the n-th Bernoulli number. It follows that

$$B_2 = \tfrac{1}{6}, \quad B_4 = -\tfrac{1}{30}, \quad B_6 = \tfrac{1}{42}, \quad \ldots, \quad B_{12} = -\tfrac{691}{2730}, \tag{1-1}$$

and that

$$B_{2n+1} = 0 \quad \text{for } n \in \mathbb{N}. \tag{1-2}$$

These and other standard properties of the Bernoulli numbers and Bernoulli polynomials can be found in [Montgomery and Vaughan 2007, Appendix B]. In this note we give an apparently new proof of Euler's well-known result which states that

$$\zeta(2k) = \sum_{n=1}^{\infty} \frac{1}{n^{2k}} = (-1)^{k+1} \frac{(2\pi)^{2k} B_{2k}}{2(2k)!} \quad \text{for } k \in \mathbb{N}. \tag{1-3}$$

MSC2010: primary 11M06; secondary 11B68, 11B37.
Keywords: Riemann zeta function, Euler, Basel problem, Bernoulli numbers, Bernoulli polynomials.

From (1-1) and (1-3), we see (as Euler did) that

$$\zeta(2)=\tfrac{1}{6}\pi^2,\quad \zeta(4)=\tfrac{1}{90}\pi^4,\quad \zeta(6)=\tfrac{1}{945}\pi^6,\quad \ldots,\quad \zeta(12)=\tfrac{691}{638512875}\pi^{12}.$$

In 1734, before realizing the connection to the Bernoulli numbers, Euler derived the values of $\zeta(2k)$ for $k=1,2,\ldots,6$. A few years later, in 1740, Euler discovered the formula in (1-3) relating $\zeta(2k)$ to B_{2k} for $k\in\mathbb{N}$. Some historical remarks about Euler's work on the Riemann zeta function and on other infinite series can be found in [Weil 1984, Chapter 3], see also [Ayoub 1974; Kline 1983; Varadarajan 2007], while references to numerous proofs of Euler's formula in (1-3) can be found in [de Amo et al. 2011].

Instead of evaluating $\zeta(2k)$ directly, our proof naturally evaluates the function $\eta(s)$ at the positive even integers. Since

$$B_n\big(\tfrac{1}{2}\big)=-(1-2^{1-n})B_n \quad \text{for } n\geq 0, \tag{1-4}$$

we note that Euler's result in (1-3) is equivalent to the formula

$$\eta(2k)=\sum_{n=1}^{\infty}\frac{(-1)^{n-1}}{n^{2k}}=(-1)^k\frac{(2\pi)^{2k}B_{2k}\big(\frac{1}{2}\big)}{2(2k)!} \quad \text{for } k\in\mathbb{N}. \tag{1-5}$$

We derive (1-5) in Section 3. Note that (1-1), (1-4), and (1-5) imply

$$\eta(2)=\tfrac{1}{12}\pi^2,\quad \eta(4)=\tfrac{7}{720}\pi^4,\quad \eta(6)=\tfrac{31}{30240}\pi^6,\quad \ldots,\quad \eta(12)=\tfrac{1414477}{1307674368000}\pi^{12}.$$

Since our proof of (1-5) is more straightforward in the special case $k=1$, we discuss this situation separately at the end of this article.

There is a striking resemblance between Euler's formula (1-3), relating the values of $\zeta(2k)$ to B_{2k}, and the formula (1-5), relating the values of $\eta(2k)$ to $B_{2k}\big(\frac{1}{2}\big)$. We have chosen to write the expression in (1-5) in this manner for more than simply aesthetic reasons; indeed our proof of (1-5) relies naturally on a recursive formula for the sequence $\big\{B_{2k}\big(\frac{1}{2}\big)\big\}_{k=0}^{\infty}$.

2. A recursive formula for $B_{2k}\big(\frac{1}{2}\big)$

The Bernoulli polynomials satisfy the inversion formula

$$x^n=\frac{1}{n+1}\sum_{\ell=0}^{n}\binom{n+1}{\ell}B_\ell(x)$$

for every integer $n\geq 0$. Setting $x=\frac{1}{2}$ and then observing that (1-2) and (1-4) imply $B_n\big(\frac{1}{2}\big)=0$ if n is odd, we derive the recursive formula

$$\frac{1}{2^{2k}}=\frac{1}{2k+1}\sum_{j=0}^{k}\binom{2k+1}{2j}B_{2j}\big(\tfrac{1}{2}\big) \quad \text{for } k\in\mathbb{N}. \tag{2-1}$$

3. Proof of (1-5)

We prove (1-5) by evaluating the integral

$$I_{2k} = \int_0^1 \frac{x(\log x)^{2k}}{(x^2+1)^2}\,\mathrm{d}x \quad \text{for } k \in \mathbb{N}$$

in two different ways. On one hand, we show that

$$I_{2k} = \frac{(2k)!}{2^{2k+1}}\eta(2k) \tag{3-1}$$

by expressing the integrand as a series and then integrating term-by-term. The formula (3-1) actually holds for $k = 0$ as well, since $I_0 = \frac{1}{4}$ and it can be shown that $\eta(0) = \frac{1}{2}$. On the other hand, using the residue theorem in a relatively standard way, we derive the recursive formula

$$\frac{1}{2^{2k}} = \frac{1}{2k+1}\sum_{j=0}^{k}\binom{2k+1}{2j}(-1)^j\,\frac{4I_{2j}}{\pi^{2j}}. \tag{3-2}$$

Comparing this expression to the recurrence relation for $B_{2k}\left(\frac{1}{2}\right)$ from the previous section, we can derive our desired expression for $\eta(2k)$ from (3-1) and (3-2).

Proof of (1-5). Evidently, from (2-1) and (3-2), the sequences

$$\left\{B_{2j}\left(\tfrac{1}{2}\right)\right\}_{j=0}^{\infty} \quad \text{and} \quad \left\{(-1)^j\,\frac{4I_{2j}}{\pi^{2j}}\right\}_{j=0}^{\infty}$$

satisfy the same recursion relation. Moreover, since

$$4I_0 = 4\int_0^1 \frac{x}{(x^2+1)^2}\,\mathrm{d}x = 1 = B_0\left(\tfrac{1}{2}\right),$$

the initial terms in these sequences agree and therefore these sequences are equal. Hence, from (3-1), we see that

$$B_{2k}\left(\tfrac{1}{2}\right) = (-1)^k\frac{4I_{2k}}{\pi^{2k}} = (-1)^k\frac{2(2k)!}{(2\pi)^{2k}}\eta(2k) \quad \text{for every } k \in \mathbb{N}. \qquad \square$$

It remains to establish (3-1) and (3-2).

3.1. ***Relating I_{2k} to $\eta(2k)$.*** Integrating by parts $2k$ times, we derive that

$$\int_0^1 x^{2n-1}(\log x)^{2k}\,\mathrm{d}x = \frac{(2k)!}{(2n)^{2k+1}} \tag{3-3}$$

for positive integers k and n. Alternatively, we can prove this estimate by using that the gamma function,

$$\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,\mathrm{d}x \quad \text{for } \operatorname{Re}(z) > 0,$$

satisfies the relation $\Gamma(n+1) = n!$ for $n \in \mathbb{N}$. To see this, note that the variable change $x \mapsto e^{-t/(2n)}$ implies

$$\int_0^1 x^{2n} (\log x)^{2k} \frac{\mathrm{d}x}{x} = \frac{1}{(2n)^{2k+1}} \int_0^\infty e^{-t} t^{2k} \, \mathrm{d}t = \frac{\Gamma(2k+1)}{(2n)^{2k+1}} = \frac{(2k)!}{(2n)^{2k+1}}.$$

We now express the integrand of I_{2k} as a series, interchange the sum and the integral, and then use (3-3) to integrate term-by-term. Since

$$\begin{aligned} \frac{x}{(x^2+1)^2} &= -\frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}x} \left\{ \frac{1}{1+x^2} \right\} \\ &= -\frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}x} \left\{ \sum_{n=0}^\infty (-1)^n x^{2n} \right\} = \sum_{n=1}^\infty n(-1)^{n-1} x^{2n-1} \end{aligned} \tag{3-4}$$

for $|x| < 1$, we have

$$\begin{aligned} I_{2k} = \int_0^1 \frac{x(\log x)^{2k}}{(x^2+1)^2} \, \mathrm{d}x &= \int_0^1 \sum_{n=1}^\infty n(-1)^{n-1} x^{2n-1} (\log x)^{2k} \, \mathrm{d}x \\ &= \sum_{n=1}^\infty n(-1)^{n-1} \int_0^1 x^{2n-1} (\log x)^{2k} \, \mathrm{d}x \\ &= \sum_{n=1}^\infty n(-1)^{n-1} \frac{(2k)!}{(2n)^{2k+1}} = \frac{(2k)!}{2^{2k+1}} \eta(2k) \end{aligned}$$

for every $k \in \mathbb{N}$. This proves (3-1). Note that the interchange of summation and integration is justified using Fubini's theorem since, for every $k \in \mathbb{N}$, (3-3) implies

$$\begin{aligned} \sum_{n=1}^\infty \int_0^1 |n(-1)^{n-1} x^{2n-1} \log^{2k} x| \, \mathrm{d}x &= \sum_{n=1}^\infty n \int_0^1 x^{2n-1} (\log x)^{2k} \, \mathrm{d}x \\ &= \frac{(2k)!}{2^{2k+1}} \sum_{n=1}^\infty \frac{1}{n^{2k}} < \infty. \end{aligned}$$

3.2. ***A recursive formula for I_{2k}.*** Making the variable change $x \mapsto 1/x$, it follows that

$$\int_0^1 \frac{x(\log x)^{2k}}{(x^2+1)^2} \, \mathrm{d}x = \int_1^\infty \frac{x(\log x)^{2k}}{(x^2+1)^2} \, \mathrm{d}x,$$

$$\int_0^1 \frac{x(\log x)^{2k+1}}{(x^2+1)^2} \, \mathrm{d}x = -\int_1^\infty \frac{x(\log x)^{2k+1}}{(x^2+1)^2} \, \mathrm{d}x$$

for integers $k \geq 0$. Therefore

$$I_{2k} = \frac{1}{2} \int_0^\infty \frac{x(\log x)^{2k}}{(x^2+1)^2} \, \mathrm{d}x \quad \text{and} \quad \int_0^\infty \frac{x(\log x)^{2k+1}}{(x^2+1)^2} \, \mathrm{d}x = 0. \tag{3-5}$$

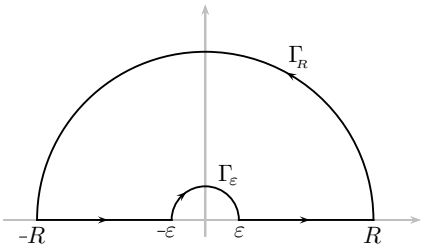


Figure 1

Now we introduce the complex-valued function

$$f(z) = \frac{z(\log z)^{2k+1}}{(1+z^2)^2},$$

where $\log z$ denotes the branch of the logarithm in $\mathbb{C}$ with $|z| > 0$ and $-\frac{\pi}{2} < \arg z < \frac{3\pi}{2}$. Note that the power of $\log z$ in the numerator of $f(z)$ is one power higher than the power of $\log x$ appearing in the integrand of I_{2k}. We integrate $f(z)$ around the positively oriented simple closed contour (shown in Figure 1) composed of the line segment $[\varepsilon, R]$ along the real-axis, the semicircle Γ_R centered at 0 of radius R starting at $z = R$ passing through $z = iR$ and ending at $z = -R$, the line segment $[-R, -\varepsilon]$ along the real-axis, and finally the semicircle Γ_ε centered at 0 of radius ε starting at $z = -\varepsilon$ passing through $z = i\varepsilon$ and ending at $z = \varepsilon$. Here ε and R denote real numbers satisfying $0 < \varepsilon < 1 < R < \infty$. The only singularity of $f(z)$ inside this contour is a double pole at $z = i$. Therefore the residue theorem implies

$$2\pi i \operatorname*{Res}_{z=i} f(z) = \int_\varepsilon^R \frac{x(\log x)^{2k+1}}{(1+x^2)^2}\,dx + \int_{\Gamma_R} f(z)\,dz + \int_{-R}^{-\varepsilon} \frac{x(\log(-x)+i\pi)^{2k+1}}{(1+x^2)^2}\,dx + \int_{\Gamma_\varepsilon} f(z)\,dz, \quad (3\text{-}6)$$

where the logarithms in the first and third integrals on the right-hand side denote the natural logarithm. Estimating trivially, we have

$$\left|\int_{\Gamma_\varepsilon} f(z)\,dz\right| \le \operatorname{length}(\Gamma_\varepsilon) \cdot \max_{z\in\Gamma_\varepsilon} |f(z)| \le (\pi\varepsilon)\left(\frac{\varepsilon(\log(-\varepsilon)+\pi)^{2k+1}}{(1-\varepsilon^2)^2}\right) \to 0$$

as $\varepsilon \to 0^+$ and

$$\left|\int_{\Gamma_R} f(z)\,dz\right| \le \operatorname{length}(\Gamma_R) \cdot \max_{z\in\Gamma_R} |f(z)| \le (\pi R)\left(\frac{R(\log R+\pi)^{2k+1}}{(1-R^2)^2}\right) \to 0$$

as $R \to +\infty$. It follows that

$$2\pi i \operatorname*{Res}_{z=i} f(z) = \int_0^\infty \frac{x(\log x)^{2k+1}}{(1+x^2)^2}\,dx + \int_{-\infty}^0 \frac{x(\log(-x)+i\pi)^{2k+1}}{(1+x^2)^2}\,dx.$$

By the second expression in (3-5), the first integral on the right-hand side equals 0. Sending $x \mapsto -x$, the second integral on the right-hand side equals

$$-\int_0^\infty \frac{x(\log x + i\pi)^{2k+1}}{(1+x^2)^2}\,\mathrm{d}x = -\sum_{\ell=0}^{2k+1}\binom{2k+1}{\ell}(i\pi)^{2k-\ell+1}\int_0^\infty \frac{x(\log x)^\ell}{(1+x^2)^2}\,\mathrm{d}x.$$

Again by (3-5), the terms in the sum with ℓ odd vanish. Hence, for even ℓ, letting $\ell = 2j$ and using the first expression in (3-5), we have

$$2\pi i \operatorname*{Res}_{z=i} f(z) = -(i\pi)^{2k+1}\sum_{j=0}^{k}\binom{2k+1}{2j}(-1)^j\frac{2I_{2j}}{\pi^{2j}}. \tag{3-7}$$

On the other hand, a straightforward calculation shows that

$$\operatorname*{Res}_{z=i} f(z) = \lim_{z\to i}\frac{\mathrm{d}}{\mathrm{d}z}\left\{\frac{z(\log z)^{2k+1}}{(z+i)^2}\right\} = -\frac{(2k+1)(i\pi)^{2k}}{2^{2k+2}}.$$

Inserting this into (3-7) and dividing by $-(2k+1)(i\pi)^{2k+1}/2$, we conclude that

$$\frac{1}{2^{2k}} = \frac{1}{2k+1}\sum_{j=0}^{k}\binom{2k+1}{2j}(-1)^j\frac{4I_{2j}}{\pi^{2j}},$$

as claimed.

3.3. ***Remarks on the case k = 1.*** Historically, the *Basel problem* asked for a closed-form evaluation of the sum

$$\zeta(2) = \sum_{n=1}^{\infty}\frac{1}{n^2}.$$

As mentioned in the Introduction, this problem was solved by Euler in 1734. Therefore, there is perhaps special interest in a direct proof of the equivalent problem of showing that

$$\eta(2) = \sum_{n=1}^{\infty}\frac{(-1)^{n-1}}{n^2} = \frac{\pi^2}{12}.$$

In this special case, our proof above can be simplified since there is no need to appeal to properties of the Bernoulli polynomials, the gamma function, or recursion relations. We sketch the details of this calculation for the interested reader.

In this case, we evaluate the integral

$$I_2 = \int_0^1 \frac{x(\log x)^2}{(x^2+1)^2}\,\mathrm{d}x$$

in two different ways. Integrating by parts twice, it can be shown that

$$\int_0^1 x^{2n-1}(\log x)^2\,\mathrm{d}x = \frac{1}{4n^3} \quad \text{for } n \in \mathbb{N}. \tag{3-8}$$

Therefore, using the series expansion in (3-4), it follows that

$$\begin{aligned} I_2 = \int_0^1 \frac{x(\log x)^2}{(x^2+1)^2}\,\mathrm{d}x &= \int_0^1 \sum_{n=1}^{\infty} n(-1)^{n-1}x^{2n-1}(\log x)^2\,\mathrm{d}x \\ &= \sum_{n=1}^{\infty} n(-1)^{n-1}\int_0^1 x^{2n-1}(\log x)^2\,\mathrm{d}x \\ &= \frac{1}{4}\sum_{n=1}^{\infty}\frac{(-1)^{n-1}}{n^2} = \frac{\eta(2)}{4}. \end{aligned}$$

As in Section 3.1, the interchange of summation and integration can be justified using Fubini's theorem. On the other hand, making the variable change $x \mapsto 1/x$, it follows that

$$I_2 = \int_0^1 \frac{x(\log x)^2}{(x^2+1)^2}\,\mathrm{d}x = \int_1^{\infty} \frac{x(\log x)^2}{(x^2+1)^2}\,\mathrm{d}x.$$

Therefore

$$I_2 = \frac{1}{2}\int_0^{\infty} \frac{x(\log x)^2}{(x^2+1)^2}\,\mathrm{d}x. \tag{3-9}$$

In order to evaluate this integral, we apply the residue theorem in a manner similar to that in the previous section. We integrate the complex-valued function

$$f(z) = \frac{z(\log z)^3}{(1+z^2)^2}$$

around the positively oriented simple closed contour shown in Figure 1. As before, $\log z$ denotes the branch of the logarithm in $\mathbb{C}$ with $|z| > 0$ and $-\frac{\pi}{2} < \arg z < \frac{3\pi}{2}$, while ε and R denote real numbers satisfying $0 < \varepsilon < 1 < R < \infty$. Then the residue theorem implies

$$\begin{aligned} 2\pi i \operatorname*{Res}_{z=i} f(z) = \int_{\varepsilon}^{R} \frac{x(\log x)^3}{(1+x^2)^2}\,\mathrm{d}x + \int_{\Gamma_R} f(z)\,\mathrm{d}z \\ + \int_{-R}^{-\varepsilon} \frac{x(\log(-x)+i\pi)^3}{(1+x^2)^2}\,\mathrm{d}x + \int_{\Gamma_\varepsilon} f(z)\,\mathrm{d}z, \end{aligned}$$

where the logarithms in the first and third integrals on the right-hand side denote the natural logarithm. As was shown in the previous section, the second and fourth integrals on the right-hand side tend to 0 as $R \to +\infty$ and $\varepsilon \to 0^+$, respectively. Since the only singularity of $f(z)$ inside this contour is a double pole at $z = i$ with

$$\operatorname*{Res}_{z=i} f(z) = \lim_{z\to i}\frac{\mathrm{d}}{\mathrm{d}z}\left\{\frac{z(\log z)^3}{(z+i)^2}\right\} = \frac{3\pi^2}{16},$$

it follows that

$$\begin{aligned}\frac{3\pi^3 i}{8} &= \int_0^\infty \frac{x(\log x)^3}{(1+x^2)^2}\,\mathrm{d}x + \int_{-\infty}^0 \frac{x(\log(-x)+i\pi)^3}{(1+x^2)^2}\,\mathrm{d}x \\ &= \int_0^\infty \frac{x(\log x)^3}{(1+x^2)^2}\,\mathrm{d}x - \int_0^\infty \frac{x(\log x+i\pi)^3}{(1+x^2)^2}\,\mathrm{d}x.\end{aligned}$$

Here we have made the variable change $x \mapsto -x$ in the second integral. Expanding the factor $(\log x + i\pi)^3$, taking imaginary parts of both sides of the equation, and then using (3-9), we deduce that

$$\frac{3\pi^3}{8} = \pi^3 \int_0^\infty \frac{x}{(1+x^2)^2}\,\mathrm{d}x - 3\pi \int_0^\infty \frac{x(\log x)^2}{(1+x^2)^2}\,\mathrm{d}x = \frac{\pi^3}{2} - 6\pi I_2.$$

This implies $I_2 = \pi^2/48$. Combining this with our previous observation that $I_2 = \eta(2)/4$, we conclude that $\eta(2) = \pi^2/12$.

Acknowledgements

Milinovich is supported in part by the NSA Young Investigator Grant H98230-16-1-0311.

References

[de Amo et al. 2011] E. de Amo, M. Díaz Carrillo, and J. Fernández-Sánchez, "Another proof of Euler's formula for $\zeta(2k)$", *Proc. Amer. Math. Soc.* **139**:4 (2011), 1441–1444. MR Zbl

[Ayoub 1974] R. Ayoub, "Euler and the zeta function", *Amer. Math. Monthly* **81** (1974), 1067–1086. MR Zbl

[Kline 1983] M. Kline, "Euler and infinite series", *Math. Mag.* **56**:5 (1983), 307–314. MR Zbl

[Montgomery and Vaughan 2007] H. L. Montgomery and R. C. Vaughan, *Multiplicative number theory, I: Classical theory*, Cambridge Studies in Advanced Mathematics **97**, Cambridge University Press, 2007. MR Zbl

[Varadarajan 2007] V. S. Varadarajan, "Euler and his work on infinite series", *Bull. Amer. Math. Soc. (N.S.)* **44**:4 (2007), 515–539. MR Zbl

[Weil 1984] A. Weil, *Number theory: an approach through history, from Hammurapi to Legendre*, Birkhäuser, Boston, MA, 1984. MR Zbl

Received: 2017-06-12 Revised: 2018-07-30 Accepted: 2018-10-28

skrishnamurt@umass.edu *Department of Physics, University of Mississippi, University, MS, United States*

Current address: *Department of Physics, University of Massachusetts, Amherst, MA, United States*

mbmilino@olemiss.edu *Department of Mathematics, University of Mississippi, University, MS, United States*

INVOLVE 12:4 (2019)
dx.doi.org/10.2140/involve.2019.12.549

Descents and des-Wilf equivalence of permutations avoiding certain nonclassical patterns

Caden Bielawa, Robert Davis, Daniel Greeson and Qinhan Zhou

(Communicated by Jim Haglund)

A frequent topic in the study of pattern avoidance is identifying when two sets of patterns Π, Π' are Wilf equivalent, that is, when $|\mathrm{Av}_n(\Pi)| = |\mathrm{Av}_n(\Pi')|$ for all n. In recent work of Dokos et al. the notion of Wilf equivalence was refined to reflect when avoidance of classical patterns preserves certain statistics. We continue their work by examining des-Wilf equivalence when avoiding certain nonclassical patterns.

1. Introduction

Let $\mathfrak{S}_n$ denote the set of permutations of $[n] := \{1, \ldots, n\}$, and let $\mathfrak{S} = \mathfrak{S}_1 \cup \mathfrak{S}_2 \cup \cdots$ be the set of all permutations of finite length. We write $\sigma \in \mathfrak{S}_n$ as $\sigma = a_1 a_2 \cdots a_n$ to indicate that $\sigma(i) = a_i$. A function $\mathrm{st} : \mathfrak{S}_n \to \mathbb{N}$ is called a *statistic*, and the systematic study of permutation statistics is generally accepted to have begun with MacMahon [1960, Volume I, Section III, Chapter V]. Four of the most well-known statistics are the *descent*, *inversion*, *major*, and *excedance* statistics, defined respectively by

$$\begin{aligned}
\mathrm{des}(\sigma) &= |\mathrm{Des}(\sigma)|,\\
\mathrm{inv}(\sigma) &= |\{(i, j) \in [n]^2 \mid i < j \text{ and } a_i > a_j\}|,\\
\mathrm{maj}(\sigma) &= \sum_{i \in \mathrm{Des}(\sigma)} i,\\
\mathrm{exc}(\sigma) &= |\{i \in [n] \mid a_i > i\}|,
\end{aligned}$$

where $\mathrm{Des}(\sigma) = \{i \in [n-1] \mid a_i > a_{i+1}\}$. Given any statistic st, one may form the generating function

$$F_n^{\mathrm{st}}(q) = \sum_{\sigma \in \mathfrak{S}_n} q^{\mathrm{st}\,\sigma}.$$

MSC2010: 05A05.
Keywords: mesh patterns, pattern avoidance, permutation statistics.
Zhou was supported in part by the Michigan State University Discovering America exchange program.

A famous result due to [MacMahon 1960] states that $F_n^{\mathrm{des}}(q) = F_n^{\mathrm{exc}}(q)$, and that both are equal to the *Eulerian polynomial* $A_n(q)$. Similarly, it is known that $F_n^{\mathrm{inv}}(q) = F_n^{\mathrm{maj}}(q) = [n]_q!$, where

$$[n]_q = 1 + q + \cdots + q^{n-1} \quad \text{and} \quad [n]_q! = [n]_q\,[n-1]_q \cdots [1]_q.$$

Let $A \subseteq [n]$, and denote by $\mathfrak{S}_A$ the set of permutations of the elements of A. The *standardization* of $\sigma = a_1 \cdots a_{|A|} \in \mathfrak{S}_A$ is the element of $\mathfrak{S}_{|A|}$ whose letters are in the same relative order as those of σ; we denote this permutation by $\mathrm{std}(\sigma)$. Now, we say that a permutation $\sigma \in \mathfrak{S}_n$ *contains the pattern* $\pi \in \mathfrak{S}_k$ if there exists a subsequence $\sigma' = a_{i_1} \cdots a_{i_k}$ of σ such that $\mathrm{std}(\sigma') = \pi$. If no such subsequence exists, then we say that σ *avoids the pattern* π. Since we will introduce additional notions of patterns, we may call such a pattern a *classical pattern* to avoid confusion. If $\Pi \subseteq \mathfrak{S}$, then we say σ *avoids* Π if σ avoids every element of Π. The set of all permutations of $\mathfrak{S}_n$ avoiding Π is denoted by $\mathrm{Av}_n(\Pi)$. In a mild abuse of notation, if $\Pi = \{\pi\}$, we will write $\mathrm{Av}_n(\pi)$. If Π, Π' are two sets of patterns and $|\mathrm{Av}_n(\Pi)| = |\mathrm{Av}_n(\Pi')|$ for all n, then we say Π and Π' are *Wilf equivalent* and write $\Pi \equiv \Pi'$.

These ideas may be combined by setting

$$F_n^{\mathrm{st}}(\Pi; q) = \sum_{\sigma \in \mathrm{Av}_n(\Pi)} q^{\mathrm{st}\,\sigma}.$$

This allows one to say that Π, Π' are st-*Wilf equivalent* if $F_n^{\mathrm{st}}(\Pi; q) = F_n^{\mathrm{st}}(\Pi'; q)$ for all n, and write this as $\Pi \overset{\mathrm{st}}{\equiv} \Pi'$. Thus, Π and Π' may be Wilf equivalent without being st-Wilf equivalent. As a concrete example, 123 and 321 are clearly not des-Wilf equivalent, even though they are Wilf equivalent. It is straightforward to check that st-Wilf equivalence is indeed an equivalence relation on $\mathfrak{S}$.

Since it is generally a difficult question to determine whether two sets are nontrivially Wilf equivalent, one should not expect it to be any easier to determine st-Wilf equivalence. However, it is certainly possible to obtain some results; see [Dokos et al. 2012] for results regarding F_n^{inv} and F_n^{maj}, and [Baxter 2014; Cameron and Killpatrick 2015] for further results, including a study of enumeration strategies for questions of this nature. In this article, we will study $F_n^{\mathrm{des}}(\Pi; q)$ for certain nonclassical patterns, called mesh patterns and barred patterns. Special cases will allow us to identify des-Wilf equivalences. We will also present several conjectural des-Wilf equivalences and provide computational evidence for these.

2. Pattern avoidance background

Classical patterns. In order to work most efficiently, it is important to recognize that certain Wilf equivalences are almost immediate to establish. For example, it is obvious that $|\mathrm{Av}_n(123)| = |\mathrm{Av}_n(321)|$, since $a_1 \cdots a_n \in \mathrm{Av}_n(123)$ if and only if $a_n a_{n-1} \cdots a_1 \in \mathrm{Av}_n(321)$. This idea can be generalized significantly.

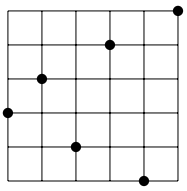


Figure 1. The plot of 342516.

The *plot* of $\sigma \in \mathfrak{S}_n$ is the set of pairs $(i, \sigma(i)) \in \mathbb{R}^2$ and will be denoted by $P(\sigma)$. The plot of 342516 is shown in Figure 1. Let

$$D_4 = \{R_0, R_{90}, R_{180}, R_{270}, r_{-1}, r_0, r_1, r_\infty\},$$

where R_θ is counterclockwise rotation of a plot by an angle of θ degrees and r_m is reflection across a line of slope m. A couple of these rigid motions have easy descriptions in terms of the one-line notation for permutations. If $\pi = a_1 a_2 \cdots a_k$ then its *reversal* is $\pi^r = a_k \cdots a_2 a_1 = r_\infty(\pi)$, and its *complement* is $\pi^c = (k+1-a_1)(k+1-a_2)\cdots(k+1-a_k) = r_0(\pi)$.

Note that $\sigma \in \mathrm{Av}_n(\pi)$ if and only if $f(\sigma) \in \mathrm{Av}_n(f(\pi))$ for any $f \in D_4$; hence $\pi \equiv f(\pi)$. For this reason, the equivalences induced by the dihedral action on a square are often referred to as the *trivial Wilf equivalences*.

Using these techniques, it is easy to show that 123 and 321 are trivially Wilf equivalent, as are all of 132, 213, 231, and 312. It is less obvious, however, whether 123 and 132 are Wilf equivalent. This question was settled by independent results due to [MacMahon 1960] and [Knuth 1969], whose combined work showed that $\mathrm{Av}_n(132)$ and $\mathrm{Av}_n(123)$ are enumerated by the *n-th Catalan number*

$$C_n = \frac{1}{n+1}\binom{2n}{n}.$$

The Catalan numbers are famous for appearing in a multitude of combinatorial situations; see [Stanley 2015] for many of them.

One of the most well-known combinatorial objects enumerated by the Catalan numbers are Dyck paths. A *Dyck path of length* $2n$ is a lattice path in $\mathbb{R}^2$ starting at $(0, 0)$ and ending at $(2n, 0)$, using steps $(1, 1)$ and $(1, -1)$, which never goes below the x-axis. See Figure 2 for an example Dyck path of length 8.

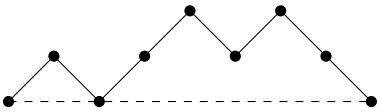
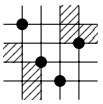


Figure 2. A Dyck path of length 8.

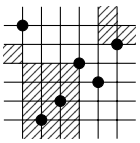
Nonclassical patterns. In this section, we will define two classes of nonclassical patterns and describe what it means for a permutation to contain or avoid them. The definitions of Wilf equivalence and des-Wilf equivalence then extend to these patterns in the same way as classical patterns, so their precise definitions will be omitted.

A *mesh pattern* is a pair (π, M), where $\pi \in \mathfrak{S}_k$ and $M \subseteq [0, k]^2$. Mesh patterns are a vast generalization of classical patterns and were first introduced by Brändén and Claesson [2011]. It is convenient to represent a mesh pattern as a grid which plots π and shades in the unit squares whose bottom-left corners are the elements of M. For example, one may represent the mesh pattern $(\pi_0, M_0) = (4213, \{(0, 2), (1, 0), (1, 1), (3, 3), (3, 4), (4, 3)\})$ as follows:

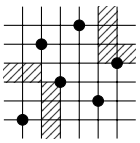
$(\pi_0, M_0) =$  .

Containment of mesh patterns is most easily understood by an informal statement and illustrative examples; the formal definition, given in [Brändén and Claesson 2011], shows that the intuition developed this way behaves as expected. We say that $\sigma \in \mathfrak{S}_n$ *contains the mesh pattern* (π, M) if σ contains an occurrence of π and the shaded regions of $P(\pi)$ corresponding to this occurrence contain no other elements of $P(\sigma)$. If σ does not contain (π, M), then we say σ *avoids* (π, M).

For the illustrative examples, first consider $\sigma = 612435$. Notice that while 6435 is an occurrence of 4213 in σ, it is not an occurrence of the mesh pattern (π_0, M_0) given above, since the shaded regions in $P(\sigma)$ dictated by M_0 yield



Now consider $\sigma' = 153624$. In this case, 5324 is an occurrence of both 4213 and (π_0, M_0) in σ', since the shading in this case is



In certain cases, determining which permutations avoid a mesh pattern (π, M) with M nonempty is equivalent to determining which permutations avoid π as a classical pattern. When this happens, we say that (π, M) has *superfluous mesh*, and Tenner [2013] identified when exactly a mesh pattern has superfluous mesh. To

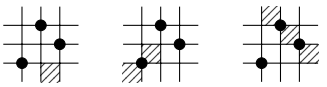
do this, we first define an *enclosed diagonal* of (π, M) to be a triple $((i, j), \varepsilon, \ell)$ where $\varepsilon \in \{-1, 1\}$, $\ell \geq 1$, and the following three properties hold:

(1) The plot of π contains the set $\{(i+d, j+\varepsilon d) \mid 1 \leq d < \ell\}$.

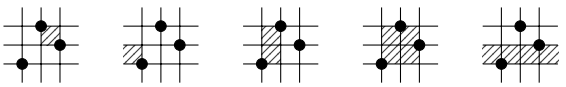
(2) The plot of π contains neither (i, j) nor $(i+\ell, j+\varepsilon\ell)$.

(3) $\{(i+d, j+\varepsilon d) \mid 0 \leq d < \ell\} \subseteq M$.

Note that an enclosed diagonal may consist of a single element, as long as the corresponding box in the mesh pattern contains no element of $P(\pi)$. To illustrate, the following three mesh patterns all have a unique enclosed diagonal:



However, none of the following five mesh patterns have any enclosed diagonals:



The following theorem gives the characterization of when a pattern has superfluous mesh. As a result, we will not focus on any patterns with superfluous mesh, but we will still use the theorem briefly.

Theorem 2.1 [Tenner 2013, Theorem 3.5′]. *A mesh pattern has superfluous mesh if and only if it has no enclosed diagonals.*

Mesh patterns also generalize 1-*barred patterns*, in which a classical pattern is allowed (but not required) to have a bar above one letter. This is a special case of *barred patterns*, in which each letter is allowed to have a bar above it. The bars above letters indicate that certain additional rules are required in order to define containment of the pattern. We will not give the precise definition of containment and avoidance of barred patterns in general, but will observe that if there are two or more bars in the pattern, there is not necessarily a simple translation of the barred pattern into a mesh pattern. In some instances, a barred pattern may be described as a *decorated mesh pattern* [Úlfarsson 2011/12], but this is not always possible. To avoid this difficulty in the statement and proof of Proposition 3.7, we will simply describe here what it means for a permutation to avoid two specific barred patterns.

We say that $\sigma = a_1 \cdots a_n$ *avoids* $\bar{1}\bar{2}43$ if, whenever $a_i a_j$ is an occurrence of 21, then there are some integers k, l such that $k < l < i$ and $a_k a_l a_i a_j$ is an occurrence of 1243. We also say that σ *avoids* $\bar{1}32\bar{4}$ if, whenever $a_i a_j$ is an occurrence of 21, then there are some integers k, l such that $k < i < j < l$ and $a_k a_i a_j a_l$ is an occurrence of 1324. As an example, $\sigma = 124635$ avoids $\bar{1}\bar{2}43$ since all occurrences of 21, which are 43, 63, and 65, extend to an occurrence of 1243 by placing 12 before them. However, σ contains $\bar{1}32\bar{4}$ since 63, which is an occurrence of 21, does not play the role of 32 in any occurrence of 1324 in σ.

3. Main results

We now have all of the tools we need to begin proving results. We begin with a simple application of several known theorems.

Proposition 3.1. *If* $(132, M_1)$ *and* $(312, M_2)$ *are mesh patterns, neither of which contain an enclosed diagonal, then*

$$(132, M_1) \overset{\text{des}}{\equiv} (312, M_2).$$

Proof. By Theorem 2.1, $\mathrm{Av}_n((312, M_2)) = \mathrm{Av}_n(312)$, so $(312, M_2) \overset{\text{des}}{\equiv} 312$. It then follows directly from [Reifegerste 2003, Remark 2.5(b)] that the number of elements in $\mathrm{Av}_n(312)$ with exactly k descents is

$$N_{n,k} := \frac{1}{n}\binom{n}{k}\binom{n}{k+1}.$$

Since the sequence $\{N_{n,k}\}_{k=0}^{n-1}$ is symmetric for fixed n, and since

$$\mathrm{des}(\sigma) = n - 1 - \mathrm{des}(\sigma^c),$$

we have

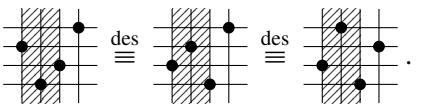
$$(312, M_2) \overset{\text{des}}{\equiv} 312 \overset{\text{des}}{\equiv} 132.$$

Again by Theorem 2.1, we have $\mathrm{Av}_n(132) = \mathrm{Av}_n((132, M_1))$, so these two patterns are des-Wilf equivalent as well. Connecting the equivalences, the claim follows. □

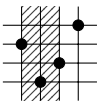
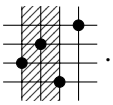
Characterizing the des-Wilf classes for mesh patterns (π, M) where $\pi \in \mathfrak{S}_4$ is difficult, and we will not attempt to fully characterize the des-Wilf equivalence classes of such patterns. In what follows, we merely wish to present a step toward understanding these in more depth, but first we need two more definitions.

If $A \subseteq [n]$, $f \in D_4$, and $\sigma \in \mathfrak{S}_A$, then we let $f^A(\sigma)$ denote the unique element of $\mathfrak{S}_A$ whose standardization is $f(\mathrm{std}(\sigma))$. We say that f^A is a dihedral action *relative to* A. As a simple example, if $7461 \in \mathfrak{S}_{\{1,4,6,7\}}$, then $\mathrm{std}(7461) = 4231$ and $R_{90}^{\{1,4,6,7\}}(\sigma) = 1647$.

Theorem 3.2. *We have*



Proof. First consider

$(\pi_1, M_1) =$  and $(\pi_2, M_2) =$  .

To prove their des-Wilf equivalence, we will form a des-preserving bijection

$$\alpha : \mathfrak{S}_n \setminus \mathrm{Av}_n((\pi_1, M_1)) \to \mathfrak{S}_n \setminus \mathrm{Av}_n((\pi_2, M_2)),$$

that is, a des-preserving bijection between permutations in $\mathfrak{S}_n$ containing (π_1, M_1) and those containing (π_2, M_2).

Suppose $\sigma = a_1 \cdots a_n \in \mathfrak{S}_n$ contains (π_1, M_1). If σ contains (π_2, M_2), then set $\alpha(\sigma) = \sigma$. Otherwise, let j be the smallest index in which an occurrence of (π_1, M_1) begins, and consider $a_i a_{i+1} \cdots a_p$, where

$$p = \min\{m \mid m > j+2,\ a_m > a_j\},$$
$$i = \min\{m \mid m \le j,\ a_m, a_{m+1}, \ldots, a_j < a_p\}.$$

Let $A = \{a_i, a_{i+1}, \ldots, a_p\}$, and set

$$R_{180}^A(a_i \cdots a_p) = b_i \cdots b_p,$$

and further set

$$\alpha(\sigma) = a_1 \cdots a_{i-1} b_i \cdots b_p a_{p+1} \cdots a_n.$$

Since R_{180}^A is a des-preserving map, we have that for any $k \in \{1, \ldots, p-1-i\}$, $i+k \in \mathrm{Des}(\sigma)$ if and only if $p-k \in \mathrm{Des}(\alpha(\sigma))$. Additionally, for any $k \in \{1, \ldots, i-1, p, p+1, \ldots, n-1\}$, $k \in \mathrm{Des}(\sigma)$ if and only if $k \in \mathrm{Des}(\alpha(\sigma))$. Thus, α is des-preserving.

To show that α is invertible, we will construct a map

$$\beta : \mathfrak{S}_n \setminus \mathrm{Av}_n((\pi_2, M_2)) \to \mathfrak{S}_n \setminus \mathrm{Av}_n((\pi_1, M_1))$$

and show that $\beta \circ \alpha$ is the identity map on $\mathfrak{S}_n \setminus \mathrm{Av}_n((\pi_1, M_1))$. If $\sigma' = a'_1 \cdots a'_n$ contains (π_2, M_2), then we create $\beta(\sigma)$ by first testing a construction similar to the one from the previous paragraph. Namely, let j' be the smallest index in which an occurrence of (π_2, M_2) begins, and consider $a'_i a'_{i+1} \cdots a'_p$, where

$$p' = \min\{m \mid m > j'+2,\ a'_m > a'_{j'+1}\},$$
$$i' = \min\{m \mid m \le j',\ a'_m, a'_{m+1}, \ldots, a_{j'} < a'_p\}.$$

This time, let $A' = \{a'_i, a'_{i+1}, \ldots, a'_p\}$, and set

$$R_{180}^{A'}(a'_i \cdots a'_p) = b'_i \cdots b'_p.$$

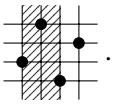
If $a'_1 \cdots a'_{i-1} b'_i \cdots b'_{p'} a'_{p'+1} \cdots a'_n$ contains both (π_2, M_2) and (π_1, M_1), then set $\beta(\sigma') = \sigma'$. Otherwise, set

$$\beta(\sigma') = a'_1 \cdots a'_{i-1} b'_i \cdots b'_{p'} a'_{p'+1} \cdots a'_n.$$

The fact that $\beta \circ \alpha$ is the identity map on $\mathfrak{S}_n \setminus \mathrm{Av}_n((\pi_1, M_1))$ follows from construction.

Now consider (π_2, M_2) and

$$(\pi_3, M_3) = \quad .$$



Suppose $\sigma = a_1 a_2 \cdots a_n$ and $a_j a_{j+1} a_{j+2} a_p$ is the first copy of (π_3, M_3), as identified in the second paragraph in this proof. If a_p is the only a_l for which $l > j+2$ and $a_l > a_j$, then set $\alpha(\sigma)$ to be σ with a_{j+1} and a_p transposed. Otherwise, choose

$$r = \min\{l \mid a_j < a_l < a_{j+1},\ l > j+2\}.$$

Let $S = \{a_r, a_{r+1}, \ldots, a_q\}$ where q is the maximum index for which $\{a_r, a_{r+1}, \ldots, a_q\}$ is increasing and $a_j < a_k < a_{j+1}$ for all $k \in S$. Set $\alpha(\sigma)$ to be σ with a_{j+1} and $\max S$ transposed. By choosing the maximum of S we are guaranteeing that α is des-preserving. By construction, $\alpha(\sigma)$ contains an occurrence of (π_2, M_2). Using an argument similar to the first part of this proof, α is invertible and is therefore a bijection. □

Recall that the *Stirling numbers of the second kind*, denoted by $S(n, k)$, record the number of ways to partition $[n]$ into k nonempty blocks. Here, we will begin to find useful the notation

$$\mathrm{Av}_n^{\mathrm{des},k}(\Pi) = \{\sigma \in \mathrm{Av}_n(\Pi) \mid \mathrm{des}(\sigma) = k\}.$$

Proposition 3.3. *Let*

$$(\pi, M) = \quad .$$

For all n, we have

$$F_n^{\mathrm{des}}((\pi, M); q) = \sum_{k=0}^{n-1} S(n, k+1) q^k.$$

Proof. Let $\Sigma_{n,k}$ denote the collection of set partitions of $[n]$ into exactly k nonempty blocks. We will create a bijection

$$f : \mathrm{Av}_n^{\mathrm{des},k}((\pi, M)) \to \Sigma_{n,k+1},$$

from which the conclusion follows.

First, let $\sigma = a_1 \cdots a_n \in \mathrm{Av}_n^{\mathrm{des},k}((\pi, M))$. It follows from [Burstein and Lankham 2005/07, Theorem 4.1] that any such permutation is the concatenation of substrings

$$\begin{aligned} a_1 &< \cdots < a_{i_0}, \\ a_{i_0+1} &< \cdots < a_{i_1}, \\ &\vdots \\ a_{i_k+1} &< \cdots < a_n, \end{aligned}$$

where $a_1 < a_{i_j+1} > a_{i_{j+1}+1}$ for all j. In particular, the values $a_{i_0}, \ldots, a_{i_k}$ determine the entire permutation.

Associate to σ the set partition

$$f(\sigma) = \{\{a_1, \ldots, a_{i_0}\}, \{a_{i_0+1}, \ldots, a_{i_1}\}, \ldots, \{a_{i_k+1}, \ldots, a_n\}\}.$$

Note that if $\sigma = 12\cdots n$, then $k = 0$, so this partition consists of only one block. Thus, if σ has k descents, then the partition obtained has $k+1$ blocks. Because each choice of the a_{i_j} determines σ, we know that $f(\sigma) \neq f(\sigma')$ whenever $\sigma' \in \mathrm{Av}_n^{\mathrm{des},k}((\pi, M))$ and $\sigma \neq \sigma'$. That is, f is injective.

Now we will show that f is surjective. Consider a set partition $B = \{B_1, \ldots, B_{k+1}\}$ of $[n]$ into $k+1$ blocks. We are free to write the B_i such that

$$B_i = \{b_{i,1} < \cdots < b_{i,i_l}\} \quad \text{and} \quad \min B_i < \min B_{i+1}$$

for all i. Construct the permutation

$$b_{k+1,1}b_{k+1,2}\cdots b_{k+1,i_{k+1}}b_{k,1}b_{k,2}\cdots b_{k,i_k}\cdots b_{1,1}b_{1,2}\cdots b_{1,i_1}.$$

We claim that this permutation is an element of $\mathrm{Av}_n^{\mathrm{des},k}((\pi, M))$.

Any occurrence of

say, $b_\alpha b_\beta b_\gamma$, implies that $b_\alpha \in B_i$, $b_\beta \in B_j$, and $b_\gamma \in B_k$ for some $i \leq j < k$. Since the sequence of minima of the blocks is decreasing, we know that $\min B_k < b_\alpha < b_\gamma$. Thus, the string

$$b_\alpha b_\beta (\min B_k) b_\gamma$$

is an occurrence of

Since the elements of the blocks strictly increase, the minima decrease, and since there are $k+1$ blocks, there are k descents in the permutation. Thus f is surjective, completing the proof. □

Example 3.4. Consider the permutation

$$3427156 \in \mathrm{Av}_6^{\mathrm{des},2}\left(\right.$$  $$\left.\right).$$

Our construction in the previous proof associates to this permutation the partition

$$\{\{3, 4\}, \{2, 7\}, \{1, 5, 6\}\}.$$

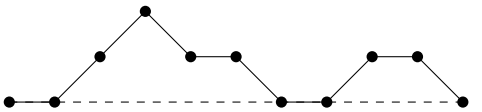
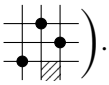


Figure 3. A Motzkin path of length 10 with 3 up-steps.

In the other direction, given the set partition

$$\{\{5\}, \{3, 1, 4\}, \{7, 2, 6\}\} = \{\{5\}, \{2, 6, 7\}, \{1, 3, 4\}\},$$

we obtain the permutation 5267134, which the reader may verify is indeed an element of

$$\mathrm{Av}_7^{\mathrm{des},2}\left(\right.$$  $$\left.\right).$$

A *Motzkin path of length n* is a lattice path from $(0, 0)$ to $(n, 0)$ using only *up-steps* $(1, 1)$, *down-steps* $(1, -1)$, and *horizontal steps* $(1, 0)$ such that the path does not go below the x-axis. An example is shown in Figure 3. We let $\mathcal{M}_{n,k}$ denote the set of Motzkin paths of length n with exactly k up-steps.

The next result we present was first proven in [Chen et al. 2002/03] by writing Motzkin paths according to a "strip decomposition" and by writing permutations according to canonical reduced decompositions. Here, we present a new, simpler proof. To do so, we only need a few more definitions.

If i is a descent of $\sigma = a_1 \cdots a_n$, then we call a_i a *descent top* and a_{i+1} a *descent bottom*. Let $\mathrm{Destop}(\sigma)$ denote the set of descent tops of σ and let $\mathrm{Desbot}(\sigma)$ denote the set of descent bottoms of σ. A *valley* in σ is an element i for which $a_{i-1} > a_i < a_{i+1}$.

Theorem 3.5 [Chen et al. 2002/03, Theorem 3.1]. *Let*

$$\Pi = \{(\pi_4, M_4), (\pi_5, M_5)\},$$

where

$$(\pi_4, M_4) =$$  $$, \quad (\pi_5, M_5) =$$  $$.$$

For all n,

$$F_n^{\mathrm{des}}(\Pi; q) = \sum_{k=0}^{n} |\mathcal{M}_{n,k}| q^k.$$

Proof. We will form a bijection

$$\mu : \mathrm{Av}_n^{\mathrm{des},k}(\Pi) \to \mathcal{M}_{n,k}.$$

For $\sigma = a_1 \cdots a_n \in \mathrm{Av}_n^{\mathrm{des},k}(\Pi)$, let $\mu(\sigma)$ be the lattice path obtained by making step a_i a down-step if a_i is a descent bottom, an up-step if a_i is a descent top, and a horizontal step if a_i is neither.

First, we need to check that μ is well-defined. Note that no letter of σ can be both a descent top and a descent bottom, since this would imply σ contains an instance of π_4, which is forbidden. So, since the sets of descent tops and of descent bottoms are disjoint, and these appear in pairs, we can be certain that the path constructed by μ has length n and ends at $(n, 0)$. Moreover, since a descent top always appears before a descent bottom, at no step of the path can there have been more down-steps than up-steps. This establishes that $\mu(\sigma)$ is a Motzkin path of length n. Finally, since there are k descents, there are k descent tops, and $\mu(\sigma)$ will have k up-steps. Hence, $\mu(\sigma) \in \mathcal{M}_{n,k}$.

Next we will show that μ is injective. To do so, we will determine exactly the structure of the elements in $\mathrm{Av}_n(\Pi)$. Notice that the descent bottoms of σ must appear in increasing order in σ, since, otherwise, there would be an occurrence of π_4. For the same reason, the descent tops must appear in increasing order in σ.

Let $\sigma = a_1 \cdots a_n \in \mathrm{Av}_n^{\mathrm{des},k}(\Pi)$ and suppose that i is neither a descent top nor a descent bottom. Suppose for now that j is the first descent greater than i. If $a_{j+1} < a_i < a_j$, then $a_i a_j a_{j+1}$ is an occurrence of 231. Since σ avoids (π_5, M_5), there must be some l for which σ has the subsequence $a_i a_l a_j a_{j+1}$ and $a_l < a_{j+1}$. This implies that some integer $i+1, i+2, \ldots, l-1$ is a descent, which contradicts the fact that j is the first descent greater than i. So, it must be true that $a_i < a_{j+1} < a_j$. Since j is the first descent greater than i, it follows that $a_i a_{i+1} \cdots a_{j-1} a_{j+1}$ is an increasing sequence. It follows that the subsequence of σ consisting of all letters that are not descent tops is an increasing sequence.

Now we will show that μ is injective. If $\mu(\sigma_1) = \mu(\sigma_2)$ for $\sigma_1, \sigma_2 \in \mathrm{Av}_n(\Pi)$, then $\mathrm{Destop}(\sigma_1) = \mathrm{Destop}(\sigma_2)$ and $\mathrm{Desbot}(\sigma_1) = \mathrm{Desbot}(\sigma_2)$, since these are identified by the up-steps and down-steps in the Motzkin path. Our description of elements of $\mathrm{Av}_n(\Pi)$ shows that once the descent-top sets and descent-bottom sets have been identified, there is a unique σ in the avoidance class with those sets. Therefore, $\sigma_1 = \sigma_2$, and μ is injective.

Finally, we will show that μ is surjective. Let $A \in \mathcal{M}_{n,k}$, and label its steps $1, \ldots, n$ from left to right. We will construct its preimage in stages. First write down $1, \ldots, n$, but exclude the labels on the down-steps. Then insert the label on the i-th down-step immediately before the label of the i-th up-step. Call the resulting permutation σ_A. Using the description of elements of $\mathrm{Av}_n(\Pi)$ from earlier in this proof, we see that $\sigma_A \in \mathrm{Av}_n(\Pi)$. Additionally, it is clear that $\mu(\sigma_A) = A$ by our construction of σ_A and the definition of μ. Therefore, μ is surjective, completing the proof. □

Example 3.6. Let A be the Motzkin path in Figure 3. Steps 2, 3, and 8 are up-steps, and therefore will be descents bottoms. Steps 4, 6, and 10 are down-steps, so these will be descent tops. The remaining numbers will be neither descent tops nor bottoms.

When the descent tops are removed from $\mu^{-1}(A)$, the result will be an increasing string of numbers: 1235789. The descent tops are then placed immediately preceding the descent bottoms, to obtain 1426357(10)89.

For the final result of the section, we make two notes. First, recall that the *Eulerian polynomial* $A_n(q)$ is the polynomial

$$\sum_{\sigma \in \mathfrak{S}_n} q^{\mathrm{des}(\sigma)} = A_n(q).$$

It should be noted that some authors, e.g., in [Stanley 1997], define the Eulerian polynomials using $q^{\mathrm{des}(\sigma)+1}$ rather than the definition given here. So, one should take care when encountering Eulerian polynomials in the literature. Second, recall from the end of Section 2 what it means for a permutation to contain and avoid the barred patterns $\bar{1}\bar{2}43$ or $\bar{1}32\bar{4}$.

Proposition 3.7. *For all* n,

$$F_n^{\mathrm{des}}(\bar{1}\bar{2}43; q) = F_n^{\mathrm{des}}(\bar{1}32\bar{4}; q) = \begin{cases} 1 & \text{if } n = 0, 1, \\ A_{n-2}(q) & \text{if } n \geq 2. \end{cases}$$

Proof. We will first show that $F_n(\bar{1}\bar{2}43; q)$ satisfies the right-hand side. The conclusion is clearly true for $n < 2$, so we will restrict our attention to when $n \geq 2$. Choose $\sigma = a_1 \cdots a_n \in \mathrm{Av}_n(\bar{1}\bar{2}43)$. Note first that $a_1 < a_2$ since, if $a_1 > a_2$, then a_1a_2 would be an occurrence of $u(\bar{1}\bar{2}43) = 21$ but this cannot extend to an occurrence of 1243.

Now, suppose $a_2 > 2$. Setting $a_m = \min\{a_i \mid 3 \leq i \leq n\}$ we have $a_2 > a_m$, so a_2a_m is an occurrence of $u(\bar{1}\bar{2}43)$ in σ. However, there is only letter to the left of a_2, so this pattern does not extend to an instance of 1243. Thus, $a_2 = 2$. Together with the previous paragraph, we know $a_1 = 1$ as well. In particular, $a_1 < a_2 < a_i$ for all $i \geq 3$.

Now, take any occurrence a_ia_j of 21 in which $2 < i < j$. Clearly, $a_1a_2a_ia_j$ is an extension to 1243. This holds for any possible permutation of $3, \ldots, n$ as the final $n-2$ letters. Since 1 and 2 are never descents of these permutations, we have

$$F_n^{\mathrm{des}}(\bar{1}\bar{2}43; q) = A_{n-2}(q),$$

as claimed.

Now we will show that the same formula holds for $\bar{1}32\bar{4}$. This time, assume $\sigma \in \mathrm{Av}_n(\bar{1}32\bar{4})$. If $a_i = 1$ for some $i > 1$, then a_1a_i would be an occurrence of 21. However, this can never extend to 1324 since there is no letter to the left of a_1. Thus, $a_1 = 1$. An analogous argument shows $a_n = n$.

This allows $a_2 \cdots a_{n-1}$ to be any arrangement of $2, 3, \ldots, n-1$, since, whenever a_ia_j is an occurrence of 21 for $2 \leq i, j \leq n-1$, this extends to $1a_ia_jn$. So, we have

the bijection

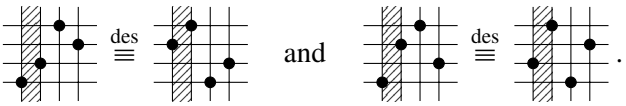
$$a_1 a_2 \cdots a_n \mapsto (a_2 - 1)(a_3 - 1) \cdots (a_{n-1} - 1)$$

with elements of $\mathfrak{S}_{n-2}$. Since 1 and n are never descents in $\mathrm{Av}_n(\bar{1}32\bar{4})$, this is a des-preserving bijection. Therefore, $F_n^{\mathrm{des}}(\bar{1}\bar{2}43; q) = A_{n-2}(q)$. □

4. Conjectures and further directions

In this section, we provide a few conjectures, supporting data, and additional direction in which this work could proceed. In all cases, no closed forms for the functions $F_n^{\mathrm{des}}(\Pi; q)$ are known. We refer the reader to Table 1 for all known polynomials $F_n^{\mathrm{des}}(\Pi; q)$ for $4 \le n \le 8$, since, for these choices of Π, $F_n^{\mathrm{des}}(\Pi; q) = F_n^{\mathrm{des}}(\varnothing; q)$ for $n \le 3$.

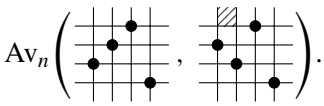
Conjecture 4.1. The following des-Wilf equivalences hold:



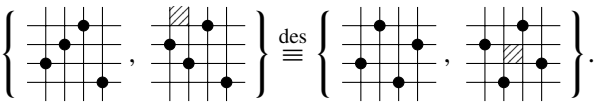
To state our next conjecture, we must discuss a particular sorting of permutations. Let $\sigma = a_1 \cdots a_n \in \mathfrak{S}_n$ and suppose $a_i = n$. Let Γ be the operator defined recursively as

$$\Gamma(\sigma) = \Gamma(a_1 \cdots a_{i-1})\Gamma(a_{i+1} \cdots a_n)n.$$

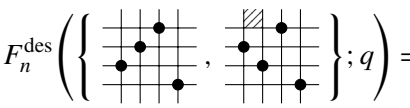
We say that σ is *West-t-stack-sortable* if $\Gamma^t(\sigma)$ is the identity permutation. Note that the 2-West-stack-sortable permutations [West 1990] are exactly those in



Conjecture 4.2. The following des-Wilf equivalence holds:



If this conjecture is true, then from [Bóna 2002] it follows that



$$= \sum_{k=0}^{n-1} \frac{(n+k)!\,(2n-k-1)!}{(k+1)!\,(n-k)!\,(2k+1)!\,(2n-2k-1)!} q^k.$$

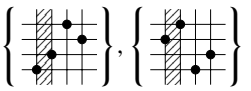
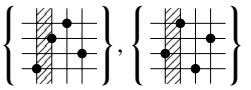
Π	n	$F_n^{\mathrm{des}}(\Pi; q)$
	4	$1+10q+11q^2+q^3$
	5	$1+20q+57q^2+26q^3+q^4$
	6	$1+35q+204q^2+252q^3+57q^4+q^5$
	7	$1+56q+581q^2+1500q^3+969q^4+120q^5+q^6$
	8	$1+84q+1414q^2+6588q^3+9117q^4+3426q^5+247q^6+q^7$
	4	$1+10q+11q^2+q^3$
	5	$1+20q+56q^2+26q^3+q^4$
	6	$1+35q+196q^2+241q^3+57q^4+q^5$
	7	$1+56q+546q^2+1361q^3+897q^4+120q^5+q^6$
	8	$1+84q+1302q^2+5675q^3+7739q^4+3060q^5+247q^6+q^7$

Table 1. The polynomials $F_n^{\mathrm{des}}(\Pi; q)$ for certain sets of patterns Π.

Instead of generalizing the patterns being avoided, one may generalize permutations themselves. One way to do this is to consider the *colored permutations*

$$G_{r,n} := \{(\varepsilon, \sigma) \mid \varepsilon \in \mathbb{Z}_r,\ \sigma \in \mathfrak{S}_n\}.$$

In this case, we say that $(\varepsilon, \sigma) \in G_{r,n}$ *contains* $(\zeta, \pi) \in G_{s,m}$ if there are elements $1 \le i_1 < i_2 < \cdots < i_s \le n$ such that $\mathrm{std}(\sigma_{i_1} \cdots \sigma_{i_s}) = \pi$ and $\varepsilon_{i_j} = \zeta_j$ for all j. If no such choice of i_j exist, then we say (ε, σ) *avoids* (ζ, π). For a set of colored permutations Π, let

$$\mathrm{Av}_{r,n}(\Pi) = \{(\varepsilon, \sigma) \in G_{r,n} \mid (\varepsilon, \sigma) \text{ avoids all } (\zeta, \pi) \in \Pi\}.$$

Question 4.3. What can be said about the polynomials

$$F_{r,n}^{\mathrm{st}}(\Pi; q) = \sum_{(\varepsilon,\sigma) \in \mathrm{Av}_{r,n}(\Pi)} q^{\mathrm{st}(\varepsilon,\sigma)}?$$

We close by noting that $G_{r,n}$ is the set of elements in the wreath product $\mathbb{Z}_r \wr \mathfrak{S}_n$, a fact which may be useful when addressing the above questions.

Acknowledgements

The authors would like to thank the anonymous referee for thoughtful comments which significantly improved this work.

References

[Baxter 2014] A. M. Baxter, "Refining enumeration schemes to count according to permutation statistics", *Electron. J. Combin.* **21**:2 (2014), art. id. 2.50. MR Zbl

[Bóna 2002] M. Bóna, "Symmetry and unimodality in t-stack sortable permutations", *J. Combin. Theory Ser. A* **98**:1 (2002), 201–209. MR Zbl

[Brändén and Claesson 2011] P. Brändén and A. Claesson, "Mesh patterns and the expansion of permutation statistics as sums of permutation patterns", *Electron. J. Combin.* **18**:2 (2011), art. id. 5. MR Zbl

[Burstein and Lankham 2005/07] A. Burstein and I. Lankham, "Combinatorics of patience sorting piles", *Sém. Lothar. Combin.* **54A** (2005/07), art. id. B54Ab. MR Zbl

[Cameron and Killpatrick 2015] N. T. Cameron and K. Killpatrick, "Inversion polynomials for permutations avoiding consecutive patterns", *Adv. in Appl. Math.* **67** (2015), 20–35. MR Zbl

[Chen et al. 2002/03] W. Y. C. Chen, Y.-P. Deng, and L. L. M. Yang, "Motzkin paths and reduced decompositions for permutations with forbidden patterns", *Electron. J. Combin.* **9**:2 (2002/03), art. id. 15. MR Zbl

[Dokos et al. 2012] T. Dokos, T. Dwyer, B. P. Johnson, B. E. Sagan, and K. Selsor, "Permutation patterns and statistics", *Discrete Math.* **312**:18 (2012), 2760–2775. MR Zbl

[Knuth 1969] D. E. Knuth, *The art of computer programming, I: Fundamental algorithms*, Addison-Wesley, Reading, 1969. MR Zbl

[MacMahon 1960] P. A. MacMahon, *Combinatory analysis*, Chelsea, New York, 1960. MR Zbl

[Reifegerste 2003] A. Reifegerste, "On the diagram of 132-avoiding permutations", *European J. Combin.* **24**:6 (2003), 759–776. MR Zbl

[Stanley 1997] R. P. Stanley, *Enumerative combinatorics, I*, Cambridge Studies in Advanced Mathematics **49**, Cambridge University Press, 1997. MR Zbl

[Stanley 2015] R. P. Stanley, *Catalan numbers*, Cambridge University Press, 2015. MR Zbl

[Tenner 2013] B. E. Tenner, "Mesh patterns with superfluous mesh", *Adv. in Appl. Math.* **51**:5 (2013), 606–618. MR Zbl

[Úlfarsson 2011/12] H. Úlfarsson, "Describing West-3-stack-sortable permutations with permutation patterns", *Sém. Lothar. Combin.* **67** (2011/12), art. id. B67d. MR Zbl

[West 1990] J. West, *Permutations with forbidden subsequences and stack-sortable permutations*, Ph.D. thesis, Massachusetts Institute of Technology, 1990, https://search.proquest.com/docview/303904866. MR

Received: 2017-06-23 Revised: 2018-03-25 Accepted: 2018-10-28

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

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

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

qinhan_zhou@qq.com *School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, Shaanxi, China*

INVOLVE 12:4 (2019)
dx.doi.org/10.2140/involve.2019.12.565

The classification of involutions and symmetric spaces of modular groups

Marc Besson and Jennifer Schaefer

(Communicated by Kenneth S. Berenhaut)

The involutions and the symmetric spaces associated to the family of modular groups of order 2^m are explored. We begin by analyzing the structure of the automorphism group and by establishing which automorphisms are involutions. We conclude by calculating the fixed-point group and symmetric spaces determined by each involution.

1. Introduction

A first course in group theory usually provides a short introduction to the idea of the automorphism group of a group. Students often begin by calculating the automorphism group for a few familiar groups of small order, such as the symmetric group S_3 or the dihedral group D_4. Computing the automorphism group of one of these groups is an especially fruitful exercise as it requires a student to understand properties of the group itself and results in students making conjectures about the structure of automorphism groups of similar groups. Though this activity is worthwhile on its own, knowing the structure of the automorphism group of a group has also proven essential in a variety of areas, including the theory of symmetric spaces.

First introduced by Élie Cartan [1926; 1927], real symmetric spaces were a special class of homogeneous Riemannian manifolds. Berger [1957] later generalized these spaces and gave classifications of the irreducible semisimple symmetric spaces. Since then the theory of symmetric spaces has expanded into a field that plays a fundamental role in numerous areas of active research, including Lie theory, number theory, differential geometry, harmonic analysis, and physics; see [Harish-Chandra 1984a; 1984b; 1984c; 1984d; Ōshima and Matsuki 1984; Brylinski and Delorme 1992; Carmona and Delorme 1994; van den Ban and Schlichtkrull 1997a; 1997b; Delorme 1998] for mathematics examples and [Olshanetsky and Perelomov 1983; Zirnbauer 1996] for physics examples. The theory of symmetric

MSC2010: 20D15, 53C35.
Keywords: modular 2-group, symmetric spaces, automorphisms, involutions.

spaces also has many generalizations. Symmetric varieties, symmetric k-varieties, Vinberg's theta-groups, spherical varieties, Gelfand pairs, Bruhat–Tits buildings, Kac–Moody symmetric spaces, and generalized symmetric spaces are among these generalizations which have found importance in various areas of mathematics and physics such as number theory, algebraic geometry, and representation theory.

The majority of these generalizations can be studied in the context of generalized symmetry spaces. Generalized symmetric spaces are defined as the homogeneous spaces G/H with G an arbitrary group and $H = G^\theta = \{g \in G \mid \theta(g) = g\}$ the fixed-point group of an order-n automorphism θ. Of special interest are automorphisms of order 2, also called *involutions*. If G is an algebraic group defined over a field k and θ an involution defined over k, then these spaces are also called symmetric k-varieties, first introduced in [Helminck 1994].

For involutions there is a natural embedding of the homogeneous spaces G/H into the group G as follows. Let $\tau : G \to G$ be a morphism of G given by $\tau(g) = g\,\theta(g)^{-1}$ for $g \in G$, where θ is an involution of G. The map τ induces an isomorphism of the coset space G/H onto $\tau(G) = \{g\,\theta(g)^{-1} \mid g \in G\}$. We will take the image $Q = \{g\,\theta(g)^{-1} \mid g \in G\}$ as our definition of the *generalized symmetric space determined by* (G, θ). In addition, we define the *extended symmetric space determined by* (G, θ) as $R = \{g \in G \mid \theta(g) = g^{-1}\}$. Extended symmetric spaces play an important role in generalizing the Cartan decomposition for real reductive groups to reductive algebraic groups defined over an arbitrary field. While for real groups it suffices to use Q for the Cartan decomposition, in the general case one needs the extended symmetric space R. Symmetric spaces and symmetric k-varieties are well known for their role in many areas of mathematics, but they are probably best known for their fundamental role in representation theory. The generalized symmetric spaces as defined above are of importance in a number of areas as well, including group theory, number theory, and representation theory.

Recently, involutions and symmetric spaces have been determined for dihedral groups [Cunningham et al. 2014], dicyclic groups [Bishop et al. 2013], and semidihedral groups [Schaefer and Schlechtweg 2017]. In this paper, we investigate the involutions and symmetric spaces associated to the modular groups of order 2^m. Since all non-Abelian 2-groups of order 2^m which contain a cyclic subgroup of order 2^{m-1} and where $m \geq 4$ are isomorphic to a dihedral group, a generalized quaternion group (contained in the more general class of dicyclic groups), a semidihedral group, or a modular group by [Gorenstein 1968], this work completes the study of involutions and symmetric spaces for groups of this structure. We begin in Section 2 by analyzing the family of modular groups, $\mathrm{M}_m(2)$, of order 2^m for $m \geqslant 4$. In Section 3, we classify the automorphisms of $\mathrm{M}_m(2)$ and establish which automorphisms are involutions. We also consider which involutions arise from inner automorphisms. In Section 4, we describe the fixed-point group H, the generalized

symmetric space Q, and the extended symmetric space R determined by each involution of $\mathrm{M}_m(2)$. Finally in the Appendix, we provide H, Q, and R for each involution of $\mathrm{M}_4(2)$.

2. Preliminaries

Throughout this paper, we consider the modular 2-group $\mathrm{M}_m(2)$, which can be described using the following presentation from [Gorenstein 1968]:

$$\mathrm{M}_m(2) = \langle x, y \mid x^{2^{m-1}} = y^2 = 1,\ yx = x^{2^{m-2}+1} y \rangle,$$

where $m \geq 4$ is an integer. Defined in terms of generators and relations, this presentation is convenient for determining the automorphism group of $\mathrm{M}_m(2)$ and the fixed-point group and symmetric spaces associated with each involution.

We begin by providing some basic structural properties of $\mathrm{M}_m(2)$ that are prerequisites for the rest of the paper. The group presentation given above clearly shows that $\mathrm{M}_m(2)$ is a non-Abelian group. The next result we state provides a commutation relation which we will use to simplify the structure of the group's elements.

Lemma 1. *For any integer $k \geq 1$, we have $yx^k = x^{(2^{m-2}+1)k} y$.*

Using the outcome of Lemma 1 repeatedly, together with the relations $x^{2^{m-1}} = y^2 = 1$ and the uniqueness of a quotient and a remainder in the quotient-remainder theorem, we have the following results.

Proposition 2. *Every element of $\mathrm{M}_m(2)$ has a unique presentation as $x^i y^j$, where i and j are integers with $0 \leq i < 2^{m-1}$ and $j \in \{0, 1\}$.*

We call the presentation given in Proposition 2 the *normal form* of an element of $\mathrm{M}_m(2)$ and by writing all elements of the group in their normal form, we have the subsequent corollary.

Corollary 3. *The non-Abelian group $\mathrm{M}_m(2)$ has order 2^m and consists of the elements $1,\ x,\ x^2, \ldots, x^{2^{m-1}-1},\ y,\ xy,\ x^2y, \ldots, x^{2^{m-1}-1}y$.*

In order to determine the automorphism group and the symmetric spaces, it will be necessary to know the order and inverse of each group element. The next three results establish this information.

Lemma 4. *For any integer $k \geq 1$,*

$$(x^i y^j)^k = \begin{cases} x^{ik+ij(k-1)2^{m-3}} y^j & \text{when } k \text{ is odd,} \\ x^{ik+ijk2^{m-3}} & \text{when } k \text{ is even.} \end{cases}$$

Proof. Suppose $k \geq 1$ is an integer and $x^i y^j \in \mathrm{M}_m(2)$ for $0 \leq i < 2^{m-1}$ and $j \in \{0, 1\}$. Then $(x^i y^j)(x^i y^j) = x^{2i+ij2^{m-2}}$ by Lemma 1. When k is odd,

$(x^i y^j)^k$ has $\frac{1}{2}(k-1)$ pairs of the form $(x^i y^j)(x^i y^j)$. Thus

$$\begin{aligned}(x^i y^j)^k &= x^{(2i+ij2^{m-2})\frac{1}{2}(k-1)} x^i y^j \\ &= x^{(k-1)(i+ij2^{m-3})+i} y^j = x^{ik+ij(k-1)2^{m-3}} y^j.\end{aligned}$$

When k is even, $(x^i y^j)^k$ has $\frac{1}{2}k$ pairs of the form $(x^i y^j)(x^i y^j)$. In this case

$$(x^i y^j)^k = x^{(2i+ij2^{m-2})\frac{1}{2}k} = x^{ik+ijk2^{m-3}}$$

as desired. □

Proposition 5. *For any integer i with $0 \le i < 2^{m-1}$,*

$$|x^i| = \frac{2^{m-1}}{\gcd(i, 2^{m-1})} \quad \textit{and} \quad |x^i y| = \frac{2^{m-1}}{\gcd(2^{m-2}, i + i2^{m-3})}.$$

Proof. By basic properties of cyclic groups and the fact that $|x| = 2^{m-1}$,

$$|x^i| = \frac{2^{m-1}}{\gcd(i, 2^{m-1})}.$$

Consider $x^i y$. Then $(x^i y)^2 = x^{2i+i2^{m-2}}$ by Lemma 4, and

$$|x^{2i+i2^{m-2}}| = \frac{2^{m-1}}{\gcd(2^{m-1}, 2i + i2^{m-2})}$$

by above. By Lagrange's theorem, $|(x^i y)^2| \le |x^i y|$. Furthermore, $|x^i y| \le 2|(x^i y)^2|$ by properties of order. Hence we have $|(x^i y)^2| \le |x^i y| \le 2|(x^i y)^2|$.

Since $|\mathrm{M}_m(2)| = 2^m$, we know that $|x^i y|$ is a power of 2 by Lagrange's theorem. So either $|x^i y| = |(x^i y)^2|$ or $|x^i y| = 2|(x^i y)^2|$. We can easily rule out the first case, because $\langle (x^i y)^2 \rangle$ is a proper subgroup of $\langle x^i y \rangle$, seeing as it does not contain $x^i y$ for instance. Thus

$$|x^i y| = 2|(x^i y)^2| = 2\frac{2^{m-1}}{\gcd(2^{m-1}, 2i + i2^{m-2})} = \frac{2^{m-1}}{\gcd(2^{m-2}, i + i2^{m-3})}. \quad \square$$

Proposition 6. *For any integer i with $0 \le i < 2^{m-1}$,*

$$(x^i)^{-1} = x^{2^{m-1}-i} \quad \textit{and} \quad (x^i y)^{-1} = x^{(2^{m-1}-i)(2^{m-2}+1)} y.$$

Proof. The result follows immediately from Lemma 1 and the relations $x^{2^{m-1}} = y^2 = 1$. □

The final result of this section describes which elements compose the center of $\mathrm{M}_m(2)$. Knowing the center allows us to simplify calculations in several instances.

Proposition 7. *The center of $\mathrm{M}_m(2)$ consists of all elements of the form x^i where $0 \le i < 2^{m-1}$ is even. Thus $Z(\mathrm{M}_m(2))$ is a cyclic subgroup of order 2^{m-2}.*

Proof. We break this proof into three cases.

Case 1: Consider $x^{2k} \in M_m(2)$, where $0 \leq k < 2^{m-2}$. Then

$$xx^{2k} = x^{1+2k} = x^{2k+1} = x^{2k}x,$$

and by Lemma 1,

$$yx^{2k} = x^{2k(2^{m-2}+1)}y = x^{k2^{m-1}}x^{2k}y = x^{2k}y.$$

Thus x^{2k} commutes with both generators and $\langle x^2 \rangle \leq Z(M_m(2))$.

Case 2: Consider $x^{2k+1} \in M_m(2)$, where $0 \leq k < 2^{m-2}$. Using the commutation relation of Lemma 1,

$$yx^{2k+1} = x^{(2k+1)(2^{m-2}+1)}y = x^{2k+1}x^{2^{m-2}}y \neq x^{2k+1}y,$$

as $x^{2^{m-2}}$ is not equal to the identity. Thus x^{2k+1} is not central.

Case 3: Consider $x^i y \in M_m(2)$, where $0 \leq i < 2^{m-1}$. Then $xx^i y = x^{i+1}y$. However,

$$x^i yx = x^i x^{2^{m-2}+1}y = x^{2^{m-2}}x^{i+1}y.$$

These two expressions cannot be equal because $x^{2^{m-2}}$ is not equal to the identity. Thus elements of the form $x^i y$ are *not* central.

Therefore, $Z(M_m(2)) = \langle x^2 \rangle$. □

Example. The center of $M_4(2)$ is $Z(M_4(2)) = \{1, x^2, x^4, x^6\}$.

3. Automorphisms and involutions of $\mathrm{M}_m(2)$

In this section, we determine the automorphism group of $\mathrm{M}_m(2)$, denoted by $\mathrm{Aut}(\mathrm{M}_m(2))$. We begin by analyzing the structure of each automorphism and then move to proving some properties of the automorphism group as a whole. We conclude this section by establishing which elements of $\mathrm{Aut}(\mathrm{M}_m(2))$ are involutions and what properties two automorphism must satisfy to be equivalent.

Theorem 8. *A homomorphism $\phi : \mathrm{M}_m(2) \to \mathrm{M}_m(2)$ is an automorphism if and only if $\phi(x) = x^a y^b$ and $\phi(y) = x^{c2^{m-2}}y$ where a is odd and $b, c \in \{0, 1\}$.*

Proof. Let $\phi \in \mathrm{Aut}(\mathrm{M}_m(2))$. Then by properties of automorphisms, ϕ must map x to an element of order 2^{m-1} and y to an element of order 2. Thus by Proposition 5, $\phi(x) = x^a$ or $x^a y$, where a is odd and $\phi(y) = y$, $x^{2^{m-2}}$, or $x^{2^{m-2}}y$. However, ϕ would not be injective if y mapped to $x^{2^{m-2}}$. Therefore, if ϕ is an automorphism, $\phi(x) = x^a y^b$ and $\phi(y) = x^{c2^{m-2}}y$, where a is odd and $b, c \in \{0, 1\}$. The converse of this statement can be proven using cases. □

Corollary 9. *The automorphism group* $\mathrm{Aut}(\mathrm{M}_m(2))$ *has order* 2^m.

Proof. Since there are $2^{m-2} \cdot 2$ elements $x^a y^b$, where a is odd and $b \in \{0, 1\}$, and two elements $x^{c2^{m-2}} y$, where $c \in \{0, 1\}$,

$$|\mathrm{Aut}(\mathrm{M}_m(2))| = 2^{m-2} \cdot 2 \cdot 2 = 2^m. \qquad \square$$

Remark. It is interesting that $|\mathrm{Aut}(\mathrm{M}_m(2))| = |\mathrm{M}_m(2)|$. In the cases of dihedral groups [Cunningham et al. 2014], generalized quaternion groups [Bishop et al. 2013], and semidihedral groups [Schaefer and Schlechtweg 2017], the order of the automorphism group is much larger than the order of the group.

Based on the results of Theorem 8, we can represent each automorphism uniquely as $\phi_{a,b,c}$, where $\phi_{a,b,c}(x) = x^a y^b$ and $\phi_{a,b,c}(y) = x^{c2^{m-2}} y$, where a is odd and $b, c \in \{0, 1\}$. Using this notation, we see that $\phi_{1,0,0}$ denotes the identity automorphism. In the following theorem, we determine where $\phi_{a,b,c}$ maps an arbitrary element $x^i y^j \in M_m(2)$.

Theorem 10. *Let $x^i y^j \in M_m(2)$ for $0 \le i < 2^{m-1}$ and $j \in \{0, 1\}$ and $\phi_{a,b,c} \in \mathrm{Aut}(M_m(2))$, where a is odd and $b, c \in \{0, 1\}$. Then*

$$\phi_{a,b,c}(x^i y^j) = \begin{cases} x^{ai + abi2^{m-3} + cj2^{m-2}} y^j & \text{when } i \text{ is even,} \\ x^{ai + ab(i-1)2^{m-3} + cj2^{m-2}} y^{b+j} & \text{when } i \text{ is odd.} \end{cases}$$

Proof. Let $x^i y^j \in M_m(2)$ for $0 \le i < 2^{m-1}$ and $j \in \{0, 1\}$ and $\phi_{a,b,c} \in \mathrm{Aut}(M_m(2))$, where a is odd and $b, c \in \{0, 1\}$. By Theorem 8, we have

$$\phi_{a,b,c}(x^i y^j) = (x^a y^b)^i (x^{c2^{m-2}} y)^j.$$

In Proposition 7, we proved $x^{c2^{m-2}} \in Z(\mathrm{M}_m(2))$. Thus $(x^{c2^{m-2}} y)^j = x^{cj2^{m-2}} y^j$. To understand how the term $(x^a y^b)^i$ interacts with $x^{cj2^{m-2}} y^j$, we split into two cases: i even and i odd.

Case 1: Let i be even. Then by Lemma 4

$$\begin{aligned} \phi_{a,b,c}(x^i y^j) &= (x^a y^b)^i x^{cj2^{m-2}} y^j \\ &= x^{ai + abi2^{m-3}} x^{cj2^{m-2}} y^j \\ &= x^{ai + abi2^{m-3} + cj2^{m-2}} y^j. \end{aligned}$$

Case 2: Let i be odd. Then by Lemma 4

$$\begin{aligned} \phi_{a,b,c}(x^i y^j) &= (x^a y^b)^i x^{cj2^{m-2}} y^j \\ &= x^{ai + ab(i-1)2^{m-3}} y^b x^{cj2^{m-2}} y^j \\ &= x^{ai + ab(i-1)2^{m-3} + cj2^{m-2}} y^{b+j}. \end{aligned}$$

$\square$

Conjugation by a fixed element of a group G is one of the most important examples of an automorphism of a group. Thus it is interesting to determine which elements of $\mathrm{Aut}(\mathrm{M}_m(2))$ are inner automorphisms. Given an arbitrary group G and

an element $g \in G$, we let $\varphi_g \in \mathrm{Aut}(G)$ denote conjugation by g and $\mathrm{Inn}(G)$ denote the collection of inner automorphisms of G.

Theorem 11. *The inner automorphisms of* $\mathrm{M}_m(2)$ *are* $\phi_{1,0,c}$ *and* $\phi_{(2^{m-2}+1),0,c}$, *where* $c \in \{0, 1\}$.

Proof. Consider φ_g for some $g \in \mathrm{M}_m(2)$. Suppose $g = x^i$. Then

$$\varphi_{x^i}(x) = x^i x x^{2^{m-1}-i} = x^{2^{m-1}+1} = x,$$
$$\varphi_{x^i}(y) = x^i y x^{2^{m-1}-i} = x^i x^{(2^{m-2}+1)(2^{m-1}-i)} y = x^{-i2^{m-2}} y.$$

When $-i$ is even, $x^{-i2^{m-2}} y = y$ and when $-i$ is odd, $x^{-i2^{m-2}} y = x^{2^{m-2}} y$. Next, consider $g = x^i y$. Then

$$\varphi_{x^i y}(x) = (x^i y) x (y x^{2^{m-1}-i}) = x^i (x^{2^{m-2}+1} y)(y x^{2^{m-1}-i}) = x^{2^{m-2}+1},$$
$$\varphi_{x^i y}(y) = (x^i y) y (y x^{2^{m-1}-i}) = x^i (x^{(2^{m-1}-i)(2^{m-2}+1)} y) = x^{-i2^{m-2}} y.$$

Again, when $-i$ is even, $x^{-i2^{m-2}} y = y$ and when $-i$ is odd, $x^{-i2^{m-2}} y = x^{2^{m-2}} y$.

Conversely, consider $\phi_{1,0,c} \in \mathrm{Aut}(\mathrm{M}_m(2))$. Note that conjugation by x^{-c} gives

$$x^{-c} x x^c = x,$$
$$x^{-c} y x^c = x^{c(2^{m-2})} y.$$

Thus, $\phi_{1,0,c} \in \mathrm{Inn}(\mathrm{M}_m(2))$. Similarly, consider $\phi_{2^{m-2}+1,0,c} \in \mathrm{Aut}(\mathrm{M}_m(2))$. Then conjugation by $x^{-c} y$ gives

$$(x^{-c} y) x (y x^c) = x^{2^{m-2}+1},$$
$$(x^{-c} y) y (y x^c) = x^{c(2^{m-2})} y.$$

Thus, $\phi_{2^{m-2}+1,0,c} \in \mathrm{Inn}(\mathrm{M}_m(2))$. Therefore, $\phi_{a,b,c}$ is an inner automorphism of $\mathrm{M}_m(2)$ if and only if a is 1 or $2^{m-2} + 1$, $b = 0$, and $c \in \{0, 1\}$. □

It follows from this result that four of the 2^m automorphisms in $\mathrm{Aut}(\mathrm{M}_m(2))$ are inner automorphisms, which we knew would be the case as $\mathrm{Inn}(\mathrm{M}_m(2)) \cong \mathrm{M}_m(2)/Z(\mathrm{M}_m(2))$ and $|Z(\mathrm{M}_m(2))| = 2^{m-2}$ [Gorenstein 1968]. In Section 4, we will find it useful to understand the structure of the involutions arising from inner automorphisms because it will allow us to simplify the presentation of the fixed-point groups, the generalized symmetric spaces, and the extended symmetric spaces in these cases.

Before we can characterize the involutions, we require the following lemmas.

Lemma 12. *For any* $\phi_{a,b,c}, \phi_{d,e,f} \in \mathrm{Aut}(\mathrm{M}_m(2))$, *where* a *and* d *are odd and* $b, c, e, f \in \{0, 1\}$,

$$\phi_{a,b,c} \circ \phi_{d,e,f} = \phi_{ad+ab(d-1)2^{m-3}+ce2^{m-2},\, b+e,\, c+f}.$$

Proof. Let $\phi_{a,b,c}$ and $\phi_{d,e,f} \in \mathrm{Aut}(M_m(2))$. To determine $\phi_{a,b,c} \circ \phi_{d,e,f}$, we examine $\phi_{a,b,c} \circ \phi_{d,e,f}(x)$ and $\phi_{a,b,c} \circ \phi_{d,e,f}(y)$.

By Theorem 10 and d odd,

$$\phi_{a,b,c} \circ \phi_{d,e,f}(x) = \phi_{a,b,c}(x^d y^e) = x^{ad+ab(d-1)2^{m-3}+ce2^{m-2}} y^{b+e}.$$

Next, by Theorem 10 and $f2^{m-2}$ even,

$$\begin{aligned}\phi_{a,b,c} \circ \phi_{d,e,f}(y) &= \phi_{a,b,c}(x^{f2^{m-2}} y) \\ &= x^{af2^{m-2}+abf2^{m-2}2^{m-3}+c2^{m-2}} y = x^{(af+c)2^{m-2}} y.\end{aligned}$$

Because a is odd, $a = 2k+1$ for $k \in \mathbb{Z}$ and we have

$$x^{(af+c)2^{m-2}} y = x^{((2k+1)f+c)2^{m-2}} y = x^{(f+c)2^{m-2}} y.$$

Thus $\phi_{a,b,c} \circ \phi_{d,e,f}(y) = x^{(c+f)2^{m-2}} y$.

Given the images of x and y under $\phi_{a,b,c} \circ \phi_{d,e,f}$, we can define the general form of automorphism composition:

$$\phi_{a,b,c} \circ \phi_{d,e,f} = \phi_{ad+ab(d-1)2^{m-3}+ce2^{m-2},\, b+e,\, c+f}. \qquad \square$$

This result now allows to us to answer our question regarding automorphisms of order 2. We see in the following theorem that this reduces to evaluating an equation modulo 2^{m-1}.

Lemma 13. *Let $\phi_{a,b,c} \in \mathrm{Aut}(\mathrm{M}_m(2))$, where a is odd and $b, c \in \{0,1\}$. Then $(\phi_{a,b,c})^2 = \phi_{1,0,0}$ if and only if*

$$a^2 + ab(a-1)2^{m-3} + bc2^{m-2} \equiv 1 \bmod 2^{m-1}. \tag{1}$$

Proof. Consider $\phi_{a,b,c} \in \mathrm{Aut}(\mathrm{M}_m(2))$. By Lemma 12, we find that

$$\phi_{a,b,c} \circ \phi_{a,b,c} = \phi_{a^2+ab(a-1)2^{m-3}+bc2^{m-2},\, 2b,\, 2c}.$$

Since $b, c \in \{0,1\}$, we have $2b \equiv 2c \equiv 0 \bmod 2$ always. Thus we only need to solve (1) to determine when $\phi_{a,b,c} \circ \phi_{a,b,c} = \phi_{1,0,0}$. $\square$

Theorem 14. *For $m = 4$, $\mathrm{Aut}(\mathrm{M}_4(2))$ contains 11 involutions and for integers $m \geq 5$, $\mathrm{Aut}(\mathrm{M}_m(2))$ contains 15 involutions.*

Proof. Let $\phi_{a,b,c} \in \mathrm{Aut}(\mathrm{M}_m(2))$, where a is odd and $b, c \in \{0,1\}$, such that $(\phi_{a,b,c})^2 = \phi_{1,0,0}$. Then by Lemma 13, (1) holds.

Case 1: Suppose $b = 0$ and $c = 0$. Then (1) reduces to $a^2 \equiv 1 \bmod 2^{m-1}$. There are four elements a in $\mathbb{Z}_{2^{m-1}}$ with $a^2 \equiv 1 \bmod 2^{m-1}$ by [Burton 2010], namely 1, -1, $1 + 2^{m-2}$, and $-1 + 2^{m-2}$. Thus we have four elements of the form $\phi_{a,0,0} \in \mathrm{Aut}(\mathrm{M}_m(2))$ with $(\phi_{a,0,0})^2 = \phi_{1,0,0}$. Because $\phi_{1,0,0}$ has order 1, it follows that there are three involutions of the form $\phi_{a,0,0}$, where $a \in \{-1, 1+2^{m-2}, -1+2^{m-2}\}$.

Case 2: Suppose $b = 0$ and $c = 1$. Then (1) again reduces to $a^2 \equiv 1 \bmod 2^{m-1}$ with solutions 1, -1, $1 + 2^{m-2}$, and $-1 + 2^{m-2}$. Thus in this case we have four involutions of the form $\phi_{a,0,1}$, where $a \in \{1, -1, 1 + 2^{m-2}, -1 + 2^{m-2}\}$.

Case 3: Suppose $b = 1$ and $c = 0$. Then (1) reduces to $a^2 + a(a-1)2^{m-3} \equiv 1 \bmod 2^{m-1}$, which is equivalent to $a^2(1 + 2^{m-3}) - a2^{m-3} - 1 \equiv 0 \bmod 2^{m-1}$. Consider $m = 4$. Then our equation becomes $3a^2 - 2a - 1 \equiv 0 \bmod 8$. It can be shown that 1 and 5 are the only solutions. Thus the only involutions of the form $\phi_{a,1,0}$ when $m = 4$ are $\phi_{1,1,0}$ and $\phi_{5,1,0}$.

Now suppose $m \geq 5$. Because $1 + 2^{m-3}$ is odd, our equation is equivalent to $(1 + 2^{m-3})[a^2(1 + 2^{m-3}) - a2^{m-3} - 1] \equiv 0 \bmod 2^{m-1}$. By using the identity

$$(1 + 2^{m-3})[a^2(1 + 2^{m-3}) - a2^{m-3} - 1] = (a(1 + 2^{m-3}) - 2^{m-4})^2 - (2^{m-4} + 1)^2,$$

our original quadratic equivalence may be expressed as

$$(a(1 + 2^{m-3}) - 2^{m-4})^2 \equiv (2^{m-4} + 1)^2 \bmod 2^{m-1}.$$

Because $(2^{m-4} + 1)^2$ is odd when $m \geq 5$, this congruence has four solutions by [Burton 2010]. It can be shown that 1, $1 + 2^{m-2}$, $-1 - 2^{m-3}$, and $-1 - 2^{m-2} - 2^{m-3}$ are the solutions for a. Thus we have four involutions of the form $\phi_{a,1,0}$, where $a \in \{1, 1 + 2^{m-2}, -1 - 2^{m-3}, -1 - 2^{m-2} - 2^{m-3}\}$.

Case 4: Suppose $b = 1$ and $c = 1$. Then finally (1) reduces to $a^2 + a(a-1)2^{m-3} + 2^{m-2} \equiv 1 \bmod 2^{m-1}$, which is equivalent to

$$a^2(1 + 2^{m-3}) - a2^{m-3} + 2^{m-2} - 1 \equiv 0 \bmod 2^{m-1}.$$

Consider $m = 4$. Then our equation becomes $3a^2 - 2a + 3 \equiv 0 \bmod 8$. It can be shown that 3 and 7 are the only solutions. Thus the only involutions of the form $\phi_{a,1,1}$ when $m = 4$ are $\phi_{3,1,1}$ and $\phi_{7,1,1}$.

Now suppose $m \geq 5$. Because $1 + 2^{m-3}$ is odd, our equation is equivalent to $(1 + 2^{m-3})[a^2(1 + 2^{m-3}) - a2^{m-3} + 2^{m-2} - 1] \equiv 0 \bmod 2^{m-1}$. Using the identity

$$(1 + 2^{m-3})[a^2(1 + 2^{m-3}) - a2^{m-3} + 2^{m-2} - 1] = (a(1 + 2^{m-3}) - 2^{m-4})^2 - (2^{m-4} - 1)^2,$$

our original quadratic equivalence may be expressed as

$$(a(1 + 2^{m-3}) - 2^{m-4})^2 \equiv (2^{m-4} - 1)^2 \bmod 2^{m-1}.$$

Because $(2^{m-4} - 1)^2$ is odd when $m \geq 5$, this congruence has four solutions by [Burton 2010]. It can be shown that -1, $-1 - 2^{m-2}$, $1 + 2^{m-3}$, and $1 + 2^{m-2} + 2^{m-3}$ are the solutions for a. Thus we have four involutions of the form $\phi_{a,1,1}$, where $a \in \{-1, -1 - 2^{m-2}, 1 + 2^{m-3}, 1 + 2^{m-2} + 2^{m-3}\}$.

Considering all cases, it follows that $\mathrm{Aut}(\mathrm{M}_m(2))$ contains 11 involutions when $m = 4$ and 15 involutions $m \geq 5$. □

Remark. Given that the number of involutions increases as m increases in the cases of dihedral groups [Cunningham et al. 2014], generalized quaternion groups [Bishop et al. 2013], and semihedral groups [Schaefer and Schlechtweg 2017], it is a bit surprising that the number of involutions of $\mathrm{M}_m(2)$ is at most 15 for all m.

Example. Consider $\mathrm{M}_4(2)$. Then by Theorem 14 the 11 involutions in $\mathrm{Aut}(\mathrm{M}_4(2))$ are $\phi_{3,0,0}$, $\phi_{5,0,0}$, $\phi_{7,0,0}$, $\phi_{1,0,1}$, $\phi_{3,0,1}$, $\phi_{5,0,1}$, $\phi_{7,0,1}$, $\phi_{1,1,0}$, $\phi_{5,1,0}$, $\phi_{3,1,1}$, and $\phi_{7,1,1}$.

As stated earlier, it is useful to know which of these involutions arise from inner automorphisms. Using the results of Theorems 11 and 14, it is clear that when $a = 1$ or $2^{m-2} + 1$, $b = 0$, and $c = 0$ or 1, equation (1) is satisfied. Thus, we have the following result that characterizes which inner automorphisms are also involutions.

Theorem 15. *All three nonidentity, inner automorphisms of* $\mathrm{M}_m(2)$ *are involutions.*

Example. Consider $\mathrm{M}_4(2)$. It follows by Theorem 15 that the involutions in $\mathrm{Aut}(\mathrm{M}_4(2))$ that arise from inner automorphisms are $\phi_{1,0,1}$, $\phi_{5,0,0}$, and $\phi_{5,0,1}$.

We complete this section by determining which elements of $\mathrm{Aut}(\mathrm{M}_m(2))$ are equivalent, for equivalent involutions produce the same generalized symmetric spaces.

Definition 16. Let G be a group and ϕ, $\sigma \in \mathrm{Aut}(G)$. Then ϕ and σ are said to be isomorphic, written $\phi \sim \sigma$, if and only if there exists $\rho \in \mathrm{Aut}(G)$ such that $\rho\phi\rho^{-1} = \sigma$, i.e., ϕ and σ are conjugate to each other. Two isomorphic automorphisms are said to be in the same equivalence class.

We begin by finding the inverse of an automorphism.

Lemma 17. *For any* $\phi_{a,b,c}, \phi_{d,e,f} \in \mathrm{M}_m(2)$, *where* a *and* d *are odd and* $b, c, e, f \in \{0, 1\}$, *we have*

$$\phi_{d,e,f} = \phi_{a,b,c}^{-1}$$

if and only if

$$d \equiv (a + ab2^{m-3})^{-1}(1 + ab2^{m-3} - bc2^{m-2}) \bmod 2^{m-1}, \quad e = b \quad \text{and} \quad f = c.$$

Proof. Consider $\phi_{a,b,c}, \phi_{d,e,f} \in \mathrm{M}_m(2)$. It follows by Lemma 12 that

$$\phi_{a,b,c} \circ \phi_{d,e,f} = \phi_{ad+ab(d-1)2^{m-3}+ce2^{m-2},\, b+e,\, c+f} = \phi_{1,0,0}$$

if and only if

$$ad + ab(d-1)2^{m-3} + ce2^{m-2} \equiv 1 \bmod 2^{m-1}, \quad b = e \quad \text{and} \quad c = f.$$

Using the fact that $b = e$, the equation

$$ad + ab(d-1)2^{m-3} + ce2^{m-2} \equiv 1 \bmod 2^{m-1}$$

is equivalent to

$$ad + ab(d-1)2^{m-3} + bc2^{m-2} \equiv 1 \bmod 2^{m-1}.$$

Solving for d, we get

$$d \equiv (a + ab2^{m-3})^{-1}(1 + ab2^{m-3} - bc2^{m-2}) \bmod 2^{m-1}. \qquad \square$$

Lemma 18. *For any $\phi_{a,b,c}, \phi_{d,e,f} \in \mathrm{M}_m(2)$, where a and d are odd and $b, c, e, f \in \{0,1\}$, we have*

$$\phi_{a,b,c} \circ \phi_{d,e,f} \circ \phi_{a,b,c}^{-1} = \phi_{\alpha,e,f},$$

where

$$\alpha \equiv (a+ab2^{m-3})^{-1}[ad+c(e-abd)2^{m-2}+(ab(2d-1)+ade(1-a))2^{m-3}] + b(c+f)2^{m-2}. \quad (2)$$

Proof. Consider $\phi_{a,b,c}, \phi_{d,e,f} \in \mathrm{M}_m(2)$. Then

$$\phi_{a,b,c} \circ \phi_{d,e,f} \circ \phi_{a,b,c}^{-1} = \phi_{ad+ab(d-1)2^{m-3}+ce2^{m-2},\, b+e,\, c+f} \circ \phi_{(a+ab2^{m-3})^{-1}(1+ab2^{m-3}-bc2^{m-2}),\, b,\, c}$$

by Lemmas 12 and 17. Utilizing Lemma 12 again, this composition becomes $\phi_{\beta\gamma+\beta(\gamma-1)(b+e)2^{m-3}+b(c+f)2^{m-2},\, 2b+e,\, 2c+f}$, where

$$\beta = ad + ab(d-1)2^{m-3} + ce2^{m-2},$$
$$\gamma = (a + ab2^{m-3})^{-1}(1 + ab2^{m-3} - bc2^{m-2}),$$

which is equivalent to $\phi_{\alpha,e,f}$, where α satisfies (2), by basic algebra and reduction modulo 2^{m-1} and $2b+e \equiv e \bmod 2$ and $2c+f \equiv f \bmod 2$ by reduction modulo 2. $\square$

Proposition 19. *Two elements $\phi_{d,e,f}, \phi_{p,q,r} \in \mathrm{Aut}(\mathrm{M}_m(2))$ are equivalent if there exists an $\phi_{a,b,c} \in \mathrm{Aut}(\mathrm{M}_m(2))$ such that*

$$p \equiv (a+ab2^{m-3})^{-1}[ad+c(e-abd)2^{m-2}+(ab(2d-1)+ade(1-a))2^{m-3}] + b(c+f)2^{m-2} \bmod 2^{m-1}, \quad (3)$$

$q = e$, *and* $r = f$.

Proof. Let $\phi_{d,e,f}, \phi_{p,q,r} \in \mathrm{Aut}(\mathrm{M}_m(2))$, where d and p are odd and $e, f, p, q \in \{0,1\}$. These elements are conjugate if there exists an $\phi_{a,b,c} \in \mathrm{Aut}(\mathrm{M}_m(2))$, where a is odd and $b, c \in \{0,1\}$, such that

$$\phi_{a,b,c} \circ \phi_{d,e,f} \circ \phi_{a,b,c}^{-1} = \phi_{p,q,r}.$$

Using the results of the previous theorem, this is true if and only if p satisfies (3), $q = e$, and $r = f$. □

Example. Consider $M_4(2)$ and the 11 involutions in $\mathrm{Aut}(M_4(2))$, namely $\phi_{3,0,0}$, $\phi_{5,0,0}$, $\phi_{7,0,0}$, $\phi_{1,0,1}$, $\phi_{3,0,1}$, $\phi_{5,0,1}$, $\phi_{7,0,1}$, $\phi_{1,1,0}$, $\phi_{5,1,0}$, $\phi_{3,1,1}$, and $\phi_{7,1,1}$. Take $\phi_{3,0,0}$. Then by Proposition 19 the only involutions $\phi_{3,0,0}$ could be equivalent to are $\phi_{3,0,0}$, $\phi_{5,0,0}$, and $\phi_{7,0,0}$. Using $d = 3$, $e = 0$, and $f = 0$, the equivalence in Proposition 19 reduces to $p \equiv (1 + 2b)^{-1}[3 + 4bc + 2b] + 4bc \mod 8$. Since $b, c \in \{0, 1\}$, the only possible values for p are 3 and 7. Thus $\phi_{3,0,0}$ is equivalent to itself and $\phi_{7,0,0}$ but not $\phi_{5,0,0}$. We can use similar calculations to show the remaining equivalence classes of involutions in $\mathrm{Aut}(M_4(2))$ are $\{\phi_{1,0,1}, \phi_{5,0,1}\}$, $\{\phi_{3,0,1}\}$, $\{\phi_{7,0,1}\}$, $\{\phi_{1,1,0}, \phi_{5,1,0}\}$, and $\{\phi_{3,1,1}, \phi_{7,1,1}\}$.

4. Fixed-point groups and symmetric spaces of $M_m(2)$

Recall from the Introduction that we are interested in determining the fixed-point group H, the generalized symmetric space Q, and the extended symmetric space R for each involution of $M_m(2)$ found in Theorem 14. Please note that for the remainder of this paper the notation "≡" will represent equivalence modulo 2^{m-1}.

Let $\phi_{a,b,c} \in \mathrm{Aut}(M_m(2))$ be an involution. Then we know by Theorem 8 that $b = 0$ or $b = 1$. We begin by considering the fixed-point group for an involution of the form $\phi_{a,0,c}$.

Theorem 20. *For an involution $\phi_{a,0,c} \in \mathrm{Aut}(M_m(2))$, where a is odd and $c \in \{0, 1\}$, the fixed-point group is*

$$H_{\phi_{a,0,c}} = \{x^i y^j \mid i(a-1) + jc2^{m-2} \equiv 0\},$$

where $0 \le i < 2^{m-1}$ and $j \in \{0, 1\}$.

Proof. Let $\phi_{a,0,c} \in \mathrm{Aut}(M_m(2))$ be an involution. By definition, an element $x^i y^j \in M_m(2)$ is in the fixed-point group of $\phi_{a,0,c}$ if $\phi_{a,0,c}(x^i y^j) = x^i y^j$. By Theorem 10, this implies

$$\phi_{a,0,c}(x^i y^j) = x^{ai + cj2^{m-2}} y^j = x^i y^j.$$

For $x^i y^j$ to satisfy this equation, $ai + jc2^{m-2} \equiv i$ or $i(a-1) + jc2^{m-2} \equiv 0$. □

We now consider involutions of the form $\phi_{a,1,c}$.

Theorem 21. *For an involution $\phi_{a,1,c} \in \mathrm{Aut}(M_m(2))$, where a is odd and $c \in \{0, 1\}$, the fixed-point group is*

$$H_{\phi_{a,1,c}} = \{x^i y^j \mid i(a - 1 + a2^{m-3}) + jc2^{m-2} \equiv 0 \text{ for } i \text{ even}\},$$

where $0 \le i < 2^{m-1}$ and $j \in \{0, 1\}$.

Proof. Let $\phi_{a,1,c} \in \mathrm{Aut}(M_m(2))$ be an involution and let $x^i y^j \in M_m(2)$. We break this proof into two cases: i even and i odd.

Case 1: Suppose i is even. Then Theorem 10 implies

$$\phi_{a,1,c}(x^i y^j) = x^{ai+ai2^{m-3}+cj2^{m-2}} y^j = x^i y^j.$$

Thus, $x^i y^j$ is fixed when $ai + ai2^{m-3} + cj2^{m-2} \equiv i$ or $i(a - 1 + a2^{m-3}) + jc2^{m-2} \equiv 0$.

Case 2: Suppose i is odd. Then again Theorem 10 implies

$$\phi_{a,1,c}(x^i y^j) = x^{ai+a(i-1)2^{m-3}+cj2^{m-2}} y^{j+1} = x^i y^j.$$

Because $j + 1 \neq j$, elements of the form $x^i y^j$ with i odd are *never* in the fixed-point group of $\phi_{a,1,c}$. □

Example. Consider $\mathrm{M}_4(2)$ and four of its involutions: $\phi_{3,0,0}$, $\phi_{5,0,1}$, $\phi_{1,1,0}$, and $\phi_{7,1,1}$. Using the results of Theorems 20 and 21, we have

$$\begin{aligned}
H_{\phi_{3,0,0}} &= \{1, x^4, x^4 y, y\},\\
H_{\phi_{5,0,1}} &= \{1, x^2, x^4, x^6, xy, x^3 y, x^5 y, x^7 y\},\\
H_{\phi_{1,1,0}} &= \{1, x^4, x^4 y, y\},\\
H_{\phi_{7,1,1}} &= \{1, x^2, x^4, x^6\}.
\end{aligned}$$

Theorem 22. *For an involution $\phi_{a,0,c} \in \mathrm{Aut}(M_m(2))$, where a is odd and $c \in \{0, 1\}$, the generalized symmetric space is*

$$Q_{\phi_{a,0,c}} = \{x^{i(1-a)-jc2^{m-2}} \mid 0 \leq i < 2^{m-1} \text{ and } j \in \{0, 1\}\}.$$

Proof. Let $\phi_{a,0,c} \in \mathrm{Aut}(M_m(2))$ be an involution and let $x^i y^j \in M_m(2)$. Using Theorem 10 and Proposition 6, we have

$$\begin{aligned}
x^i y^j (\phi_{a,0,c}(x^i y^j))^{-1} &= x^i y^j (x^{ai+cj2^{m-2}} y^j)^{-1}\\
&= x^i y^j (y^j x^{-(ai+cj2^{m-2})})\\
&= x^{i(1-a)-jc2^{m-2}}.
\end{aligned}$$

□

Recall by Proposition 7 that elements of the form x^i where i is even are in the center $Z(M_m(2))$. Since for any involution $\phi_{a,b,c}$ the value of a is odd, we have the following corollary:

Corollary 23. *For an involution $\phi_{a,0,c} \in \mathrm{Aut}(M_m(2))$, where a is odd and $c \in \{0, 1\}$, the generalized symmetric space satisfies $Q_{\phi_{a,0,c}} \subseteq Z(M_m(2))$.*

Now we will examine the generalized symmetric spaces for involutions of the form $\phi_{a,1,c}$.

Theorem 24. *For an involution $\phi_{a,1,c} \in \mathrm{Aut}(M_m(2))$, where a is odd and $c \in \{0,1\}$, the generalized symmetric space is*

$$Q_{\phi_{a,1,c}} = \{x^{a2^{m-3}+i(1-a-a2^{m-3}-a2^{m-2})-jc2^{m-2}} y \mid i \text{ is odd}\} \cup \{x^{i(1-a-a2^{m-3})-jc2^{m-2}} \mid i \text{ is even}\},$$

where $0 \le i < 2^{m-1}$ and $j \in \{0,1\}$.

Proof. Let $\phi_{a,1,c} \in \mathrm{Aut}(M_m(2))$ and $x^i y^j \in M_m(2)$.

Case 1: Suppose i is even and $j = 0$. By Theorem 10 and Proposition 6,

$$\begin{aligned} x^i(\phi_{a,1,c}(x^i))^{-1} &= x^i(x^{ai+ai2^{m-3}})^{-1} \\ &= x^{i(1-a-a2^{m-3})}. \end{aligned}$$

Case 2: Suppose i is odd and $j = 0$. By Theorem 10, Proposition 6, and Lemma 1,

$$\begin{aligned} x^i(\phi_{a,1,c}(x^i))^{-1} &= x^i(x^{ai+a(i-1)2^{m-3}} y)^{-1} \\ &= x^i x^{(-ai-a(i-1)2^{m-3})(2^{m-2}+1)} y \\ &= x^{i-ai2^{m-2}-ai-a(i-1)2^{m-3}} y \\ &= x^{a2^{m-3}+i(1-a-a2^{m-3}-a2^{m-2})} y. \end{aligned}$$

Case 3: Suppose i is even and $j = 1$. By Theorem 10 and Proposition 6,

$$\begin{aligned} x^i y(\phi_{a,1,c}(x^i y))^{-1} &= x^i y(x^{ai+ai2^{m-3}+c2^{m-2}} y)^{-1} \\ &= x^i(y^2)x^{-ai-ai2^{m-3}-c2^{m-2}} \\ &= x^{i-ai-ai2^{m-3}-c2^{m-2}} \\ &= x^{i(1-a-a2^{m-3})-c2^{m-2}}. \end{aligned}$$

Case 4: Suppose i is odd and $j = 1$. By Theorem 10, Proposition 6, and Lemma 1,

$$\begin{aligned} x^i y(\phi_{a,1,c}(x^i y))^{-1} &= x^i y(x^{ai+a(i-1)2^{m-3}+c2^{m-2}})^{-1} \\ &= x^i(x^{(-ai-a(i-1)2^{m-3}-c2^{m-2})(2^{m-2}+1)}) y \\ &= x^{i-ai2^{m-2}-ai-a(i-1)2^{m-3}-c2^{m-2}} y \\ &= x^{a2^{m-3}+i(1-a-a2^{m-3}-a2^{m-2})-c2^{m-2}} y. \end{aligned}$$

□

We now determine the extended symmetric spaces for each involution. We begin with involutions of the form $\phi_{a,0,c} \in \mathrm{Aut}(\mathrm{M}_m(2))$.

Theorem 25. *For an involution $\phi_{a,0,c} \in \mathrm{Aut}(\mathrm{M}_m(2))$, where a is odd and $c \in \{0,1\}$, the extended symmetric space is*

$$R_{\phi_{a,0,c}} = \{x^i y^j \mid i(a + (2^{m-2}+1)^j) + jc2^{m-2} \equiv 0\},$$

where $0 \le i < 2^{m-1}$ and $j \in \{0,1\}$.

Proof. Let $\phi_{a,0,c} \in \mathrm{Aut}(M_m(2))$ and $x^i y^j \in M_m(2)$. To solve the equation $\phi_{a,0,c}(x^i y^j) = (x^i y^j)^{-1}$, we solve the equivalent equation $\phi_{a,0,c}(x^i y^j)x^i y^j = 1$.

By Theorem 10 and Lemma 1, we have

$$\begin{aligned}\phi_{a,0,c}(x^i y^j)x^i y^j &= x^{ai+cj2^{m-2}} y^j x^i y^j \\ &= x^{ai+cj2^{m-2}} x^{i(2^{m-2}+1)^j} y^{2j} \\ &= x^{ai+cj2^{m-2}+i(2^{m-2}+1)^j} = 1\end{aligned}$$

when $i(a + (2^{m-2}+1)^j) + jc2^{m-2} \equiv 0$. □

Next we turn our attention to the extended symmetric spaces of involutions of the form $\phi_{a,1,c}$. As in the fixed-point group case, we find that the extended symmetric spaces of these involutions do not contain elements of the form $x^i y^j$ for i odd.

Theorem 26. *For an involution $\phi_{a,1,c} \in \mathrm{Aut}(M_m(2))$, where a is odd and $c \in \{0, 1\}$, the extended symmetric space is*

$$R_{\phi_{a,1,c}} = \{x^i y^j \mid i(a + a2^{m-3} + (2^{m-2}+1)^j) + jc2^{m-2} \equiv 0 \text{ and } i \text{ is even}\},$$

where $0 \le i < 2^{m-1}$ and $j \in \{0, 1\}$.

Proof. Let $\phi_{a,1,c} \in \mathrm{Aut}(M_m(2))$ and $x^i y^j \in M_m(2)$. We again split into two cases: i even and i odd.

Case 1: Suppose i is even. Using Theorem 10 and Lemma 1, we have

$$\begin{aligned}\phi_{a,1,c}(x^i y^j)x^i y^j &= x^{ai+ai2^{m-3}+cj2^{m-2}} y^j x^i y^j \\ &= x^{ai+ai2^{m-3}+cj2^{m-2}+i(2^{m-2}+1)^j} y^{2j} \\ &= x^{ai+ai2^{m-3}+cj2^{m-2}+i(2^{m-2}+1)^j} = 1\end{aligned}$$

when $i(a + a2^{m-3} + (2^{m-2}+1)^j) + jc2^{m-2} \equiv 0$.

Case 2: Suppose i is odd. Using Theorem 10 and Lemma 1, we have

$$\begin{aligned}\phi_{a,1,c}(x^i y^j)x^i y^j &= x^{ai+a(i-1)2^{m-3}+cj2^{m-2}} y^{j+1} x^i y^j \\ &= x^{ai+a(i-1)2^{m-3}+cj2^{m-2}+i(2^{m-2}-1)^{j+1}} y.\end{aligned}$$

An element of this form can never be equivalent to the identity. Thus, when i is odd, $x^i y^j \notin R_{\phi_{a,1,c}}$. □

Example. Consider $\mathrm{M}_4(2)$ and four of its involutions: $\phi_{3,0,0}$, $\phi_{5,0,1}$, $\phi_{1,1,0}$, and $\phi_{7,1,1}$. Using the results of Theorems 22 and 24, we have

$$\begin{aligned}Q_{\phi_{3,0,0}} &= \{1, x^2, x^4, x^6\}, \\ Q_{\phi_{5,0,1}} &= \{1, x^4\},\end{aligned}$$

$$Q_{\phi_{1,1,0}} = \{1, x^4, x^4y, y\},$$
$$Q_{\phi_{7,1,1}} = \{1, x^4, x^2y, x^6y\}.$$

In addition, we have

$$R_{\phi_{3,0,0}} = \{1, x^2, x^4, x^6, xy, x^2y, x^3y, x^4y, x^5y, x^6y, x^7y\},$$
$$R_{\phi_{5,0,1}} = \{1, x^4, x^2y, x^6y\},$$
$$R_{\phi_{1,1,0}} = \{1, x^2, x^4, x^6, y, x^2y, x^4y, x^6y, \},$$
$$R_{\phi_{7,1,1}} = \{1, x^4, x^2y, x^6y\}$$

by Theorems 25 and 26.

Remark. In general, $Q \subseteq R$ for all arbitrary groups and all of their respective involutions. Thus it is not a surprise that $Q_{\phi_{a,b,c}} \subseteq R_{\phi_{a,b,c}}$ in these instances. However, it is usually the case that $Q \neq R$. Thus the fact that $Q_{\phi_{7,1,1}} = R_{\phi_{7,1,1}}$ for $\mathrm{M}_4(2)$ is notable. The fixed-point group, the generalized symmetric space, and the extended symmetric space for all involutions of $\mathrm{M}_4(2)$ are provided in the Appendix.

The descriptions of H, Q, and R can be simplified when $\phi_{a,b,c}$ is an inner automorphism. Recall from Theorem 15 that an involution arising from an inner automorphism is of the form $\phi_{1,0,1}$ or $\phi_{2^{m-2}+1,0,c}$, where $c \in \{0, 1\}$.

Theorem 27. *Let $\phi_{a,0,c}$ be an involution of $\mathrm{M}_{m-1}(2)$ which arises from an inner automorphism.*

(1) *If $a = 1$ and $c = 1$, then*

$$H_{\phi_{1,0,1}} = \{1, x, x^2, \ldots, x^{2^{m-1}-1}\},$$
$$Q_{\phi_{1,0,1}} = \{1, x^{2^{m-2}}\},$$
$$R_{\phi_{1,0,1}} = \{1, x^{2^{m-2}}, x^{2^{m-3}}y, x^{3\cdot 2^{m-3}}y\}.$$

(2) *If $a = 2^{m-2} + 1$ and $c = 0$, then*

$$H_{\phi_{2^{m-2}+1,0,0}} = \{x^i y^j \mid i \text{ is even and } j \in \{0, 1\}\},$$
$$Q_{\phi_{2^{m-2}+1,0,0}} = \{1, x^{2^{m-2}}\},$$
$$R_{\phi_{2^{m-2}+1,0,0}} = \{1, x^{2^{m-2}}, y, x^{2^{m-2}}y\}.$$

(3) *If $a = 2^{m-2} + 1$ and $c = 1$, then*

$$H_{\phi_{2^{m-2}+1,0,1}} = \{x^i y^j \mid i + j \text{ is even and } j \in \{0, 1\}\},$$
$$Q_{\phi_{2^{m-2}+1,0,1}} = \{1, x^{2^{m-2}}\},$$
$$R_{\phi_{2^{m-2}+1,0,1}} = \{1, x^{2^{m-2}}, x^{2^{m-3}}y, x^{3\cdot 2^{m-3}}y\}.$$

Appendix: Fixed-point groups and symmetric spaces for involutions of $\mathbf{M_4(2)}$

	H	Q	R
$\phi_{3,0,0}$	$\{1, x^4, y, x^4y\}$	$\{1, x^2, x^4, x^6\}$	$\{1, x^2, x^4, x^6, xy, x^2y, x^3y, x^4y, x^5y, x^6y, x^7y\}$
$\phi_{5,0,0}$	$\{1, x^2, x^4, x^6, y, x^2y, x^4y, x^6y\}$	$\{1, x^4\}$	$\{1, x^4, y, x^4y\}$
$\phi_{7,0,0}$	$\{1, x^4, y, x^4y\}$	$\{1, x^2, x^4, x^6\}$	$\{1, x, x^2, x^3, x^4, x^5, x^6, x^7, y, x^2y, x^4y, x^6y\}$
$\phi_{1,0,1}$	$\{1, x, x^2, x^3, x^4, x^5, x^6, x^7\}$	$\{1, x^4\}$	$\{1, x^4, x^2y, x^6y\}$
$\phi_{3,0,1}$	$\{1, x^4, x^2y, x^6y\}$	$\{1, x^2, x^4, x^6\}$	$\{1, x^2, x^4, x^6\}$
$\phi_{5,0,1}$	$\{1, x^2, x^4, x^6, xy, x^3y, x^5y, x^7y\}$	$\{1, x^4\}$	$\{1, x^4, x^2y, x^6y\}$
$\phi_{7,0,1}$	$\{1, x^4, x^2y, x^6y\}$	$\{1, x^2, x^4, x^6\}$	$\{1, x, x^2, x^3, x^4, x^5, x^6, x^7, xy, x^3y, x^5y, x^7y\}$
$\phi_{1,1,0}$	$\{1, x^4, y, x^4y\}$	$\{1, x^4, y, x^4y\}$	$\{1, x^2, x^4, x^6, y, x^2y, x^4y, x^6y\}$
$\phi_{5,1,0}$	$\{1, x^4, x^4y, y\}$	$\{1, x^4, y, x^4y\}$	$\{1, x^2, x^4, x^6, y, x^2y, x^4y, x^6y\}$
$\phi_{3,1,1}$	$\{1, x^2, x^4, x^6\}$	$\{1, x^4, x^2y, x^6y\}$	$\{1, x^4, x^2y, x^6y\}$
$\phi_{7,1,1}$	$\{1, x^2, x^4, x^6\}$	$\{1, x^4, x^2y, x^6y\}$	$\{1, x^4, x^2y, x^6y\}$

Acknowledgements

This paper is based on the undergraduate honors thesis of Besson under the supervision of Schaefer. Besson would like to thank Professor Schaefer for her consistent advice, support, mentoring and mathematical guidance. He'd also like to thank Professors Hermann and Tesman for inspiring his love of algebra. Schaefer would like to thank the Research Experiences for Undergraduate Faculty (REUF) program, a joint program of the American Institute of Mathematics and the Institute for Computational and Experimental Research in Mathematics, and Aloysius G. Helminck, in particular, for introducing her to the deep and rich theory of generalized symmetric spaces.

References

[van den Ban and Schlichtkrull 1997a] E. P. van den Ban and H. Schlichtkrull, "The most continuous part of the Plancherel decomposition for a reductive symmetric space", *Ann. of Math.* (2) **145**:2 (1997), 267–364. MR Zbl

[van den Ban and Schlichtkrull 1997b] E. van den Ban and H. Schlichtkrull, "Fourier transforms on a semisimple symmetric space", *Invent. Math.* **130**:3 (1997), 517–574. MR Zbl

[Berger 1957] M. Berger, "Les espaces symétriques noncompacts", *Ann. Sci. École Norm. Sup.* (3) **74** (1957), 85–177. MR Zbl

[Bishop et al. 2013] A. Bishop, C. Cyr, J. Hutchens, C. May, N. Schwartz, and B. Turner, "On involutions and generalized symmetric spaces of dicyclic groups", preprint, 2013. arXiv

[Brylinski and Delorme 1992] J.-L. Brylinski and P. Delorme, "Vecteurs distributions H-invariants pour les séries principales généralisées d'espaces symétriques réductifs et prolongement méromorphe d'intégrales d'Eisenstein", *Invent. Math.* **109**:3 (1992), 619–664. MR Zbl

[Burton 2010] D. M. Burton, *Elementary number theory*, 7th ed., McGraw-Hill, Boston, 2010.

[Carmona and Delorme 1994] J. Carmona and P. Delorme, "Base méromorphe de vecteurs distributions H-invariants pour les séries principales généralisées d'espaces symétriques réductifs: equation fonctionnelle", *J. Funct. Anal.* **122**:1 (1994), 152–221. MR Zbl

[Cartan 1926] E. Cartan, "Sur une classe remarquable d'espaces de Riemann", *Bull. Soc. Math. France* **54** (1926), 214–264. MR Zbl

[Cartan 1927] E. Cartan, "Sur une classe remarquable d'espaces de Riemann, II", *Bull. Soc. Math. France* **55** (1927), 114–134. MR Zbl

[Cunningham et al. 2014] K. K. A. Cunningham, T. Edgar, A. G. Helminck, B. F. Jones, H. Oh, R. Schwell, and J. F. Vasquez, "On the structure of involutions and symmetric spaces of dihedral groups", *Note Mat.* **34**:2 (2014), 23–40. MR Zbl

[Delorme 1998] P. Delorme, "Formule de Plancherel pour les espaces symétriques réductifs", *Ann. of Math.* (2) **147**:2 (1998), 417–452. MR Zbl

[Gorenstein 1968] D. Gorenstein, *Finite groups*, Harper & Row, New York, 1968. MR Zbl

[Harish-Chandra 1984a] Harish-Chandra, *Collected papers, I: 1944–1954*, edited by V. S. Varadarajan, Springer, 1984. MR Zbl

[Harish-Chandra 1984b] Harish-Chandra, *Collected papers, II: 1955–1958*, edited by V. S. Varadarajan, Springer, 1984. MR Zbl

[Harish-Chandra 1984c] Harish-Chandra, *Collected papers, III: 1959–1968*, edited by V. S. Varadarajan, Springer, 1984. MR Zbl

[Harish-Chandra 1984d] Harish-Chandra, *Collected papers, IV: 1970–1983*, edited by V. S. Varadarajan, Springer, 1984. MR Zbl

[Helminck 1994] A. G. Helminck, "Symmetric k-varieties", pp. 233–279 in *Algebraic groups and their generalizations: classical methods* (University Park, PA, 1991), edited by W. J. Haboush and B. J. Parshall, Proc. Sympos. Pure Math. **56**, Amer. Math. Soc., Providence, RI, 1994. MR Zbl

[Olshanetsky and Perelomov 1983] M. A. Olshanetsky and A. M. Perelomov, "Quantum integrable systems related to Lie algebras", *Phys. Rep.* **94**:6 (1983), 313–404. MR

[Ōshima and Matsuki 1984] T. Ōshima and T. Matsuki, "A description of discrete series for semisimple symmetric spaces", pp. 331–390 in *Group representations and systems of differential equations* (Tokyo, 1982), edited by K. Okamoto, Adv. Stud. Pure Math. **4**, North-Holland, Amsterdam, 1984. MR Zbl

[Schaefer and Schlechtweg 2017] J. Schaefer and K. Schlechtweg, "On the structure of symmetric spaces of semidihedral groups", *Involve* **10**:4 (2017), 665–676. MR Zbl

[Zirnbauer 1996] M. R. Zirnbauer, "Riemannian symmetric superspaces and their origin in random-matrix theory", *J. Math. Phys.* **37**:10 (1996), 4986–5018. MR Zbl

Received: 2017-07-06 Revised: 2018-08-21 Accepted: 2018-10-30

marmarc@live.unc.edu *Mathematics Department, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States*

schaefje@dickinson.edu *Department of Mathematics and Computer Science, Dickinson College, Carlisle, PA, United States*

INVOLVE 12:4 (2019)
dx.doi.org/10.2140/involve.2019.12.585

When is $a^n + 1$ the sum of two squares?

Greg Dresden, Kylie Hess, Saimon Islam, Jeremy Rouse,
Aaron Schmitt, Emily Stamm, Terrin Warren and Pan Yue

(Communicated by Kenneth S. Berenhaut)

Using Fermat's two squares theorem and properties of cyclotomic polynomials, we prove assertions about when numbers of the form $a^n + 1$ can be expressed as the sum of two integer squares. We prove that $a^n + 1$ is the sum of two squares for all $n \in \mathbb{N}$ if and only if a is a square. We also prove that if $a \equiv 0, 1, 2 \pmod 4$, n is odd, and $a^n + 1$ is the sum of two squares, then $a^\delta + 1$ is the sum of two squares for all $\delta \mid n,\ \delta > 1$. Using Aurifeuillian factorization, we show that if a is a prime and $a \equiv 1 \pmod 4$, then there are either zero or infinitely many odd n such that $a^n + 1$ is the sum of two squares. When $a \equiv 3 \pmod 4$, we define m to be the least positive integer such that $(a + 1)/m$ is the sum of two squares, and prove that if $a^n + 1$ is the sum of two squares for n odd, then $m \mid n$, and both $a^m + 1$ and n/m are sums of two squares.

1. Introduction

Many facets of number theory revolve around investigating terms of a sequence that are *interesting*. For example, if $a_n = 2^n - 1$ is prime (called a Mersenne prime), then n itself must be prime [Hardy and Wright 1979, Theorem 18, p. 15]. In this case, the property that is interesting is primality. Ramanujan was interested in the terms of the sequence $b_n = 2^n - 7$ that are squares. He conjectured that the only such terms are those with $n = 3, 4, 5, 7$ and 15, and it was later proved by Nagell [1948]; a modern reference is [Stewart and Tall 2002, p. 96]. Finally, if the Fibonacci sequence is defined by $F_0 = 0,\ F_1 = 1$ and $F_n = F_{n-1} + F_{n-2}$ for $n \geq 2$, then F_n is prime only if n is prime or $n = 4$ [Hardy and Wright 1979, Theorem 179, p. 148], and the only powers in the Fibonacci sequence are 0, 1, 8 and 144, which was proven by Bugeaud, Mignotte, and Siksek [Bugeaud et al. 2006] using similar tools to the proof of Fermat's last theorem.

In this paper, we will consider a number to be *interesting* if it can be expressed as the sum of two squares. The earliest work on this topic relates to Pythagorean

MSC2010: primary 11E25; secondary 11C08, 11R18.
Keywords: cyclotomic polynomials, Fermat's two squares theorem.
Hess, Rouse, Stamm and Warren were supported by the NSF grant DMS-1461189.

triples, which are integer solutions to $a^2 + b^2 = c^2$. Euclid supplied an infinite family of solutions: $a = m^2 - n^2$, $b = 2mn$ and $c = m^2 + n^2$.

Fermat's two squares theorem classifies which numbers can be written as the sum of two squares. Fermat claimed to have proven this theorem in his 1640 letter to Mersenne, but never shared the proof. The first published proof is attributed to Euler and was completed in 1749; see [Cox 1989, p. 11].

Theorem (Fermat's two squares theorem). *A positive integer N can be written as the sum of two squares if and only if in the prime factorization of N,*

$$N = \prod_{i=1}^{k} p_i^{e_i},$$

we have $p_i \equiv 3 \pmod 4$ only if e_i is even.

In light of Fermat's theorem, integers that can be expressed as the sum of two squares become increasingly rare. In particular, if $S(x)$ denotes the number of integers $n \leq x$ that are expressible as a sum of two squares, then Landau [1908] proved

$$\lim_{x\to\infty} \frac{S(x)}{x/\sqrt{\ln(x)}} = K \approx 0.764.$$

This can be stated more colloquially as "the probability that a number n is the sum of two squares is $K/\sqrt{\ln(n)}$."

A lot of progress has recently been made in understanding the gaps between prime numbers. In particular, [Zhang 2014; Maynard 2015] prove there are bounded gaps between primes infinitely often. The analogous questions for sums of two squares are much easier: problem A2 from the 2000 Putnam competition asked participants to show that there are infinitely many n such that n, $n+1$ and $n+2$ are all sums of two squares.

The culmination of several papers on large gaps between primes is [Ford et al. 2018], where it is proven that there are infinitely many n such that

$$p_{n+1} - p_n \gg \frac{\log p_n \, \log\log p_n \, \log\log\log\log p_n}{\log\log\log p_n},$$

where p_n is the n-th prime. This is still quite a ways from the conjectured statement that $p_{n+1} - p_n \gg \log^2 p_n$ holds infinitely often. For sums of two squares, the analogue of this conjecture is that if q_n is the n-th positive integer that is a sum of two squares, one should have $q_{n+1} - q_n \gg \log q_n$ infinitely often. This result was proved in [Richards 1982], and some recent work [Kalmynin 2017] has been done on estimating the moments

$$\sum_{q_{n+1}\leq x} (q_{n+1} - q_n)^\gamma,$$

extending [Hooley 1971].

We are interested in which terms in sequences of the form a^n+1 can be written as a sum of two squares. Curtis [2014] showed that 2^n+1 is the sum of two squares if and only if n is even or $n=3$. Additionally, if n is odd and 3^n+1 is the sum of two squares, then n must be the sum of two squares, and 3^p+1 is the sum of two squares for all prime numbers $p \mid n$.

The focus of the present paper is to say as much as possible about when a^n+1 is the sum of two squares for a general positive integer a. This paper is the result of two undergraduate research teams working simultaneously and independently over two months in the summer of 2016. The first team, from Wake Forest University, consisted of students Hess, Stamm, and Warren, and was led by Jeremy Rouse; the second team, from Washington and Lee University, consisted of students Islam, Schmitt, and Yue, and was led by Greg Dresden. Remarkably, the two teams ended up covering many of the same topics. Some of the results are unique to the Wake Forest team, while other results were proved by both teams using different methods.

In the case that $n=2k$ is even, $a^n+1=(a^k)^2+1^2$ is trivially the sum of two squares. For this reason, we focus on cases when n is odd. Our first result is the following.

Theorem 1.1. *If $a \in \mathbb{Z}$, then a^n+1 is the sum of two squares for every $n \in \mathbb{N}$ if and only if a is a square or $a=-1$.*

This result parallels Artin's conjecture that an integer a is a primitive root modulo every prime if and only if a is not a square and $a \neq -1$.

Example. (1) If $a=9$, then $9^n+1=(3^n)^2+1^2$.

(2) If $a=7$, then there is some odd n such that 7^n+1 is not the sum of two squares. For example, 7^3+1 is not the sum of two squares.

For the remainder of the paper, we assume that a is a positive integer. Our next result gives specific criteria that handle the case when a is even.

Theorem 1.2. *Suppose a is even, n is odd, and a^n+1 is the sum of two squares*:

- *If $a+1$ is the sum of two squares, then $a^\delta+1$ is the sum of two squares for all $\delta \mid n$.*
- *If $a+1$ is not the sum of two squares, then there is a unique prime number $p \equiv 3 \pmod 4$ such that $p^r \| a+1$ for some odd r, and $n=p$.*

Example. (1) If $a \equiv 2 \pmod 4$, then $a+1$ is not the sum of two squares and so there is at most one odd exponent n such that a^n+1 is the sum of two squares. For example, with $a=6$, since $a+1=7$ is divisible by the unique prime $p=7 \equiv 3 \pmod 4$, $n=7$ is the only possible odd n for which a^n+1 is the sum of two squares. Indeed, $6^7+1=476^2+231^2$.

(2) For $a \equiv 0 \pmod 4$, there are more options. If we let $a = 20$, then since $a + 1 = 3 \cdot 7$ has two prime factors congruent to 3 mod 4 that divide it to an odd power, we conclude that $20^n + 1$ is not the sum of two squares for any odd n. On the other hand, for $a = 24$, since $24^{77} + 1$ is the sum of two squares, we must also have that $24^{11} + 1$, $24^7 + 1$, and $24^1 + 1$ are each the sum of two squares. In general, the most efficient way to test if a positive integer n is a sum of two squares is to compute its prime factorization and use Fermat's two squares theorem.

We consider a special case when a is a multiple of 4.

Theorem 1.3. *Let $a = 4x$, where $x \equiv 3 \pmod 4$ and x is square-free. If n is odd, then $a^{nx} + 1$ is not the sum of two squares.*

Example. (1) Let $a = 12 = 4 \cdot 3$. Then $12^{3n} + 1$ is not the sum of two squares for any odd n. Note that Theorem 1.2 implies that since $12^3 + 1$ is not the sum of two squares, then $12^{3n} + 1$ is not the sum of two squares for any odd n. However, Theorem 1.3 guarantees, without any computation necessary, that $12^3 + 1$ is not the sum of two squares.

(2) Let $a = 28 = 4 \cdot 7$. Then $28^{7n} + 1$ is not the sum of two squares for any odd n.

The factorization tables for $12^n + 1$ [Brillhart et al. 2002; Wagstaff] imply that there are sixteen exponents $1 \le n < 293$ for which $12^n + 1$ is the sum of two squares, which are all prime except for $n = 1$. The two smallest composite exponents n for which $12^n + 1$ could possibly be the sum of two squares are $n = 473 = 11 \cdot 43$ and $n = 545 = 5 \cdot 109$; so far, of those two, we have confirmed only that $12^{545} + 1$ is the sum of two squares.

We now consider the case when a is odd. We split this into three subcases: $a \equiv 1 \pmod 8$, $a \equiv 5 \pmod 8$, and $a \equiv 3 \pmod 4$.

Theorem 1.4. *Let $a \equiv 1 \pmod 8$. If $a^n + 1$ is the sum of two squares for n odd, then $a^\delta + 1$ is the sum of two squares for all $\delta \mid n$.*

Example. (1) Let $a = 33$. Since $33^{119} + 1$ is the sum of two squares, $33^1 + 1$, $33^7 + 1$, and $33^{17} + 1$ must also be sum of two squares. Since $33^3 + 1$ is not the sum of two squares, we know $33^{3n} + 1$ is not the sum of two squares for any odd n.

(2) Let $a = 41$. Since $42 = 2 \cdot 3 \cdot 7$ is not the sum of two squares, $41^1 + 1$ is not the sum of two squares, and hence $41^n + 1$ is not the sum of two squares for any odd n.

Note that (as seen in the example with $a = 41$) the above theorem implies that if $a \equiv 1 \pmod 8$ and $a + 1$ is not the sum of two squares, then $a^n + 1$ is not the sum of two squares for any odd n. The next theorem addresses the case that $a \equiv 5 \pmod 8$.

Theorem 1.5. *Let $a \equiv 5 \pmod 8$. Then, $a^n + 1$ is never the sum of two squares for n odd.*

Example. Since $13 \equiv 5 \pmod 8$, we know $13^n + 1$ is not the sum of two squares for any odd n.

It follows that if $a \equiv 0, 1, 2 \pmod 4$, n is odd, and $a^n + 1$ is the sum of two squares, then $a^\delta + 1$ is the sum of two squares for all $\delta \mid n$. (The case when a is even follows from Theorem 1.2, $a \equiv 1 \pmod 8$ from Theorem 1.4, and $a \equiv 5 \pmod 8$ from Theorem 1.5.)

Finally, we consider $a \equiv 3 \pmod 4$, as covered in three separate results. These first two place considerable restrictions on the values of n for which $a^n + 1$ can be a sum of two squares.

Lemma 1.6. *Let $a \equiv 3 \pmod 4$, and let m be the smallest integer such that $(a+1)/m$ is the sum of two squares. If $a^n + 1$ is the sum of two squares, then $n \equiv m \pmod 4$.*

Theorem 1.7. *Let $a \equiv 3 \pmod 4$, and let m be the smallest integer such that $(a+1)/m$ is the sum of two squares. If $a^n + 1$ is a sum of two squares for some odd n, then*

- *n/m is a sum of two squares, and*
- *$a^m + 1$ is the sum of two squares, and*
- *if $\delta \mid (n/m)$ and δ is the sum of two squares, then $a^{m\delta} + 1$ is the sum of two squares.*
- *Moreover, if $a^{np^2} + 1$ is the sum of two squares for some $p \equiv 3 \pmod 4$, then $p \mid (a^n + 1)$.*

Theorem 1.7 showcases the advantages of having two teams working independently. When we first shared our results in late July, the Wake Forest group had only the first two parts of the above theorem, and the Washington and Lee group had a weaker version of the third part that was restricted to $m = 1$ and to δ being a prime equivalent to 1 (mod 4). Two weeks later, both teams had improved their results, with Wake Forest coming up with both the fourth part and the stronger version of the third part, as seen here. The proof that resulted from this collaboration is a nice combination of ideas from both teams.

Example. (1) Let $a = 11$. Then $m = 3$, and since $11^3 + 1$ is the sum of two squares, if $11^n + 1$ is the sum of two squares, then $3^j \| n$, j odd.

(2) Let $a = 43$. Then $m = 11$, and since $43^{11} + 1$ is not the sum of two squares, we conclude that $43^n + 1$ is not the sum of two squares for any odd n.

(3) If $a = 4713575$, then $m = 21$. It turns out that $a^{21} + 1$ is the sum of two squares, and so if $a^n + 1$ is the sum of two squares, then $21 \mid n$. Sure enough, $a^{105} + 1$ is the sum of two squares (and has 701 decimal digits).

We pause for a moment to remind the reader that Theorem 1.1 states that if a is not a square, then there exists some odd n such that $a^n + 1$ is not the sum of two squares. We can now extend this theorem and demonstrate that in fact there will be infinitely many such exponents:

- If a is even with $a + 1$ not the sum of two squares, or if $a \equiv 5 \pmod 8$, then Theorems 1.2 and 1.5 tell us that $a^n + 1$ fails to be the sum of two squares for infinitely many odd n (in fact, for all but at most one odd exponent n).
- If a is even with $a + 1$ the sum of two squares, or if $a \equiv 1 \pmod 8$, then we can use Theorems 1.2 and 1.4 to state that if $a^\delta + 1$ is not the sum of two squares for some odd exponent δ, then $a^{\delta N} + 1$ fails to be the sum of two squares for all odd integers N.
- Finally, if $a \equiv 3 \pmod 4$, we call upon Lemma 1.6 to state that $a^n + 1$ can only be a sum of two squares for $n \equiv m \pmod 4$.

This next result allows one to state that for certain special values of a, there is an infinite collection of odd values of n for which $a^n + 1$ is the sum of two squares.

Theorem 1.8. *Suppose n is odd, $p \equiv 1 \pmod 4$ is a prime number and $a = px^2$. Then $a^n + 1$ is the sum of two squares if and only if $a^{np} + 1$ is the sum of two squares.*

The above theorem implies that for those specific values of a, there are either no odd n, or an infinite number of odd n, for which $a^n + 1$ is the sum of two squares. In particular, if $a + 1$ is the sum of two squares, then $a^{p^n} + 1$ is the sum of two squares for all $n \geq 0$. If $a + 1$ is not the sum of two squares, one of Theorems 1.2, 1.4, or 1.5 implies that $a^n + 1$ is not the sum of two squares for any odd n.

Example. (1) Let $a = 17$, where $p = 17$ and $x = 1$. Since 18 is the sum of two squares, $17^{17^n} + 1$ is the sum of two squares for any n.

(2) Let $a = 117$, where $p = 13$ and $x = 3$. Since $a + 1 = 2 \cdot 59$ is not the sum of two squares, $117^{13^n} + 1$ is not the sum of two squares for any n.

Remark. In light of the above theorem, it is natural to ask if there are infinitely many $a \equiv 1 \pmod 8$ such that $a^n + 1$ is the sum of two squares for infinitely many odd n. This is indeed the case. In particular, the main theorem of [Iwaniec 1972] implies that if x is a real number ≥ 17, then the number of primes $p \leq x$ with $p \equiv 1 \pmod 8$ for which $p + 1$ is the sum of two squares is at least $cx/\log(x)^{3/2}$ for some positive constant c.

We can use the ideas from Theorem 1.8 to construct an infinite family of numbers a such that $a^p + 1$ is the sum of two squares. This is our next result.

Theorem 1.9. *If $p \equiv 1 \pmod 4$ is prime, there is a degree-4 polynomial $f(X)$ with integer coefficients such that $f(X)^p + 1 = g(X)^2 + h(X)^2$ for some $g(X)$ and $h(X)$*

with integer coefficients. Moreover, there is no positive integer n such that $f(n)$ is a square.

Example. If $p = 13$, then $f(X) = 13(13X^2 + 3X)^2$. Then $f(n)^{13} + 1$ is the sum of two squares for every $n \in \mathbb{N}$.

We end with a conjecture about the number of odd n for which $a^n + 1$ is the sum of two squares.

Conjecture 1.10. *Suppose a is a positive integer and $a \neq c^k$ for any positive integer c and $k > 1$. Let m be the smallest positive integer such that $(a+1)/m$ is the sum of two squares*:

- *If $m = 1$, then there are infinitely many odd n such that $a^n + 1$ is the sum of two squares.*
- *If $a \equiv 3 \pmod 4$, $a^m + 1$ is the sum of two squares, and m is prime, then there are infinitely many odd n such that $a^n + 1$ is the sum of two squares. (In fact, there should be infinitely many $p \equiv 1 \pmod 4$ such that $a^{mp} + 1$ is the sum of two squares.)*
- *If $a \equiv 3 \pmod 4$ and m is composite, then there are only finitely many odd n such that $a^n + 1$ is the sum of two squares.*

The main theoretical tools we use in this paper are the theory of cyclotomic polynomials, and in particular, a classification of which primes divide $\Phi_n(a)$ (see Theorem 2.1). Theorems 1.3 and 1.8 also use the identity $\Phi_n(x) = F(x)^2 - kx^q G(x)^2$ that arises in Aurifeuillian factorization.

The rest of the paper will proceed as follows. In Section 2, we review previous results which we will use. In Section 3, we prove a few facts that will be used in the remainder of the proofs. In Section 4, we prove Theorem 1.1. In Section 5, we prove Theorems 1.2 and 1.3. In Section 6, we prove Theorems 1.4, 1.5, and 1.7, along with Lemma 1.6, and we include a heuristic supporting Conjecture 1.10. In Section 7, we prove Theorems 1.8 and 1.9. We conclude with a chart listing all $a \leq 50$ and the first few odd integers n such that $a^n + 1$ is the sum of two squares, as well as a reference to one our theorems.

2. Background

If n is a positive integer and p is a prime number, we write $p^r \| n$ if $p^r \mid n$ but $p^{r+1} \nmid n$. If n is a positive integer and we write that n is not a sum of two squares because of the prime p, we mean that $p \equiv 3 \pmod 4$ and there is an odd r such that $p^r \| n$. If a and m are integers with $\gcd(a, m) = 1$, we define $\mathrm{ord}_m(a)$ to be the smallest positive integer k such that $a^k \equiv 1 \pmod m$. It is well known that $a^r \equiv 1 \pmod m$ if and only if $\mathrm{ord}_m(a) \mid r$. Fermat's little theorem states that if $\gcd(a, p) = 1$, then $a^{p-1} \equiv 1 \pmod p$; it follows that $\mathrm{ord}_p(a) \mid (p-1)$.

We will make use of the identity (originally due to Diophantus)

$$(a^2+b^2)(c^2+d^2) = (ac+bd)^2 + (ad-bc)^2.$$

This applies if $a, b, c, d \in \mathbb{Z}$, and also if a, b, c and d are polynomials.

Let $\Phi_n(x)$ denote the n-th cyclotomic polynomial; recall that $\Phi_n(x)$ is the unique irreducible factor of $x^n - 1$ with integer coefficients that does not divide $x^k - 1$ for any proper divisor k of n. We have $\prod_{d \mid n} \Phi_d(x) = x^n - 1$ and from this it follows that when n is odd,

$$x^n + 1 = \frac{x^{2n}-1}{x^n-1} = \prod_{d \mid 2n,\, d \nmid n} \Phi_d(x) = \prod_{\delta \mid n} \Phi_{2\delta}(x).$$

We will make use of the facts that for n odd, $\Phi_{2n}(x) = \Phi_n(-x)$ and that if $n = p^k$ is prime, then

$$\Phi_{p^k}(1) = \lim_{x \to 1} \frac{x^{p^k}-1}{x^{p^{k-1}}-1} = p.$$

The following theorem classifies prime divisors of $\Phi_n(a)$.

Theorem 2.1. *Assume that $a \geq 2$ and $n \geq 2$:*

- *If p is a prime and $p \nmid n$, then $p \mid \Phi_n(a)$ if and only if $n = \operatorname{ord}_p(a)$.*
- *If p is a prime and $p \mid n$, then $p \mid \Phi_n(a)$ if and only if $n = \operatorname{ord}_p(a) \cdot p^k$. In this case, if $n \geq 3$, then $p^2 \nmid \Phi_n(a)$.*

The authors have not been able to trace the origin of the result above, but it is certainly quite old, and may be contained in the work of A. S. Bang [1886a; 1886b]. This theorem arises in connection with Zsigmondy's work showing that for any a, $n \geq 2$ there is a prime p for which $\operatorname{ord}_p(a) = n$ unless $n = 2$ and $a + 1$ is a power of 2. One can find a proof of the result above in Trygve Nagell's textbook [1964] (see Theorems 94 and 95).

We will also make use of certain identities for cyclotomic polynomials that arise in Aurifeuillian factorization. If k is a square-free positive integer, let $d(k)$ be the discriminant of $\mathbb{Q}(\sqrt{k})$, that is,

$$d(k) = \begin{cases} k & \text{if } k \equiv 1 \pmod 4, \\ 4k & \text{if } k \equiv 2, 3 \pmod 4. \end{cases}$$

Suppose that $n \equiv 2 \pmod 4$, and $d(k) \nmid n$ but $d(k) \mid 2n$. Write the prime factorization of n as $n = 2\prod_{i=1}^{k} p_i^{e_i}$ and define $q = \prod_{i=1}^{k} p_i^{e_i - 1}$. Then Theorem 2.1 of [Stevenhagen 1987] states that

$$\Phi_n(x) = F(x)^2 - kx^q G(x)^2$$

for some polynomials $F(x), G(x) \in \mathbb{Z}[x]$. If $x = -kv^2$ for some integer v, we get

$$\Phi_n(-kv^2) = F(-kv^2)^2 + (k^{(q+1)/2} v^q G(-kv^2))^2$$

is the sum of two squares. In the case that $x = kv^2$ for some integer v, we get the factorization

$$\begin{aligned}\Phi_n(kv^2) &= F(kv^2)^2 - k(kv^2)^q G(kv^2)^2 \\ &= (F(kv^2) + k^{(q+1)/2} v^q G(kv^2))(F(kv^2) - k^{(q+1)/2} v^q G(kv^2)).\end{aligned}$$

Theorem 2.7 of [Stevenhagen 1987] states that these two factors are relatively prime.

We will also require some basic facts about quadratic residues. If p is an odd prime, we define $\left(\frac{a}{p}\right)$ to be 1 if $\gcd(a, p) = 1$ and there is some $x \in \mathbb{Z}$ such that $x^2 \equiv a \pmod{p}$. We define $\left(\frac{a}{p}\right)$ to be -1 if $\gcd(a, p) = 1$ and there is no such x, and we set $\left(\frac{a}{p}\right) = 0$ if $p \mid a$. Euler's criterion gives the congruence $\left(\frac{a}{p}\right) \equiv a^{(p-1)/2} \pmod{p}$. The definition of the quadratic residue symbol can be extended. If n is an odd integer with prime factorization $n = \prod_{i=1}^{k} p_i^{e_i}$, define the Jacobi symbol by

$$\left(\frac{a}{n}\right) = \prod_{i=1}^{k} \left(\frac{a}{p}\right)^{e_i}.$$

The quadratic reciprocity law for Jacobi symbols states that if a and b are both positive and odd, then

$$\left(\frac{a}{b}\right) = (-1)^{\frac{1}{2}(a-1)\cdot\frac{1}{2}(b-1)} \left(\frac{b}{a}\right).$$

3. General results

The following general lemmas pertain primarily to how the divisors of n affect the divisors of $a^n + 1$, and are used in rest of the sections of the paper. Results of this type are well known and date back to [Lucas 1878; Carmichael 1913/14]. A more modern source is [Stewart 1977]. We provide our own simple and short proofs of these facts to keep the paper self-contained.

Lemma 3.1. *Let $b, n \in \mathbb{Z}$, and n be odd and suppose $b \mid (x+1)$. Then*

$$b \mid (x^{n-1} - x^{n-2} + x^{n-3} - \cdots + 1)$$

if and only if $b \mid n$.

Proof. Let $b \mid (x+1)$. Then $x + 1 \equiv 0 \pmod{b}$, so $x \equiv -1 \pmod{b}$. Thus,

$$\begin{aligned}x^{n-1} - x^{n-2} + x^{n-3} - \cdots - x + 1 &\\ &\equiv (-1)^{n-1} - (-1)^{n-2} + (-1)^{n-3} - \cdots - (-1) + 1 \pmod{b} \\ &\equiv 1 + 1 + 1 + \cdots + 1 + 1 \pmod{b} \\ &\equiv n \pmod{b}.\end{aligned}$$

Therefore $b \mid (x^{n-1} - x^{n-2} + x^{n-3} - \cdots - x + 1)$ if and only if $n \equiv 0 \pmod{b}$, or equivalently, $b \mid n$. □

We obtain the following corollary as a result of the above lemma.

Corollary 3.2. *Suppose that n is odd, $\delta \mid n$ and $x^\delta + 1$ is not the sum of two squares because of some prime p. If $p \nmid n$, then $x^n + 1$ is not the sum of two squares.*

Proof. Consider

$$x^n + 1 = (x^\delta + 1)(x^{n-\delta} - x^{n-2\delta} + x^{n-3\delta} - \cdots - x^\delta + 1).$$

Since $x^\delta + 1$ is not the sum of two squares because of p, we have $p \equiv 3 \pmod 4$, r odd and $p^r \| (x^\delta + 1)$. Then $p \nmid n$ implies $p \nmid (x^{n-\delta} - x^{n-2\delta} + x^{n-3\delta} - \cdots - x^\delta + 1)$ by Lemma 3.1, and thus $p^r \| (x^n + 1)$ implying that $x^n + 1$ is not the sum of two squares. □

Lemma 3.3. *Let p be a prime such that $p^e \| (a^m + 1)$ for some $e \in \mathbb{N}$, and let $n = mcp^k$ be odd with $\gcd(c, p) = 1$ and $k \geq 0$. Then $p^{e+k} \| (a^n + 1)$.*

Proof. Using notation from the statement of the theorem, we can write

$$a^n + 1 = (a^m + 1) \cdot \frac{a^n + 1}{a^m + 1}.$$

Then, recalling how $a^m + 1$ factors into cyclotomics, we let d be the smallest divisor of m such that $p \mid \Phi_{2d}(a)$. Thanks to Theorem 2.1, we know that $p \| \Phi_{2dp}(a)$, $p \| \Phi_{2dp^2}(a)$, and so on, yet p does not divide into any other cyclotomic expressions not of that form. Now, choose i as large as possible such that $2dp^i \mid m$. Then, by our definition of n, we know that everything in the set $\{dp^{i+1}, dp^{i+2}, \ldots, dp^{i+k}\}$ divides into n yet none of them divide into m, and we also know from Theorem 2.1 (as mentioned above) that each of the k expressions $\Phi_{2dp^{i+1}}(a), \Phi_{2dp^{i+2}}(a), \ldots, \Phi_{2dp^{i+k}}(a)$ contains exactly one copy of the prime p and that no other cyclotomic divisors of $(a^n + 1)/(a^m + 1)$ contain this prime p. Hence, since $p^e \| (a^m + 1)$, we know $p^{e+k} \| (a^n + 1)$. □

4. Proof of Theorem 1.1

We begin with a lemma constructing an odd n such that $a^n + 1$ is not the sum of two squares.

Lemma 4.1. *Suppose there exists a prime $p \equiv 3 \pmod 4$ such that $\left(\frac{a}{p}\right) = -1$. Then either $a^{(p-1)/2} + 1$ or $a^{p(p-1)/2} + 1$ is not a sum of two squares.*

Proof. If $a^{(p-1)/2} + 1$ is not a sum of two squares, then we are done. Suppose $a^{(p-1)/2} + 1$ is a sum of two squares. By Euler's criterion, we have $a^{(p-1)/2} \equiv -1 \pmod p$, and it follows therefore that for some $k \in \mathbb{N}$, $p^{2k} \| a^{(p-1)/2} + 1$. By Lemma 3.3, letting $m = (p-1)/2$ and $n = p(p-1)/2$, we know that $p^{2k+1} \| (a^{p(p-1)/2} + 1)$. Thus, by Fermat's two squares theorem, $a^{p(p-1)/2} + 1$ is not the sum of two squares. □

As an example, we examine 148^n+1. We can conclude from the prime factorization of 148^n+1 that 148^n+1 is a sum of two squares for all odd $n<9$. Note that $9=\frac{19-1}{2}$ and that 19 is the smallest prime $p\equiv 3 \pmod 4$ for which $\left(\frac{148}{p}\right)=-1$. Calculation and Fermat's two square theorem reveal that $148^{(19-1)/2}+1=148^9+1$ is not a sum of two squares.

Proof of Theorem 1.1. Write $a=2^k a'$, where a' is odd. If $q\equiv 3 \pmod 4$ is prime, then

$$\left(\frac{a}{q}\right)=\left(\frac{2^k}{q}\right)\left(\frac{a'}{q}\right)=\left(\frac{2^k}{q}\right)\cdot(-1)^{(a'-1)/2}\left(\frac{q}{a'}\right)=\left(\frac{2^k}{q}\right)\left(\frac{-q}{a'}\right).$$

If a' is not a square and $a'\neq -1$, then there is a prime $r\mid a'$ that occurs to an odd power. The system of congruences

$$\begin{aligned} q &\equiv 7 \pmod 8,\\ -q &\equiv \text{quadratic nonresidue} \pmod r,\\ -q &\equiv 1 \pmod s \quad \text{for all prime } s\mid a',\ s\neq r, \end{aligned}$$

has a solution $q\equiv x \pmod{8a'}$ with $\gcd(x,8a')=1$. Therefore there is a prime q satisfying these congruences, and we have $\left(\frac{a}{q}\right)=-1$.

In the case that a' is a square but a is not, k is odd. In this case we choose $q\equiv 3 \pmod 8$ and $-q\equiv 1 \pmod s$ for all prime $s\mid a'$. This likewise yields a prime q such that $\left(\frac{a}{q}\right)=-1$.

By Lemma 4.1, either $a^{(p-1)/2}+1$ or $a^{p(p-1)/2}+1$ is not a sum of two squares and so there is at least one value of n for which a^n+1 is not a sum of two squares. □

5. Even

Now we consider the case when a is even. We prove Theorems 1.2 and 1.3.

Proof of Theorem 1.2. Suppose that a^n+1 is the sum of two squares. If $a^\delta+1$ is also the sum of two squares for every divisor δ of n, then we are done. If not, then let δ be the largest divisor of n such that $a^\delta+1$ is not the sum of two squares. Thus, $\delta<n$ and so there is a prime p that divides n/δ. By assumption, we have that $a^{\delta p}+1$ is the sum of two squares and

$$a^{\delta p}+1=(a^\delta+1)(a^{\delta(p-1)}-a^{\delta(p-2)}+\cdots+1).$$

Lemma 3.1 implies that $\gcd(a^\delta+1,(a^{\delta p}+1)/(a^\delta+1))$ divides p. Since $a^\delta+1$ is not the sum of two squares, the gcd cannot be 1 and so it must be p. Moreover,

$$\frac{a^{\delta p}+1}{p^2}=\frac{a^\delta+1}{p}\cdot\frac{a^{\delta p}+1}{p(a^\delta+1)}$$

is a sum of two squares and the product of two relatively prime integers. Thus, $(a^\delta + 1)/p$ is the sum of two squares. It follows that $p \equiv 3 \pmod 4$ and since $a^\delta + 1$ is odd, we get

$$a^\delta + 1 = p \times \text{sum of two squares} \equiv 3 \pmod 4.$$

However, since a is even, we must have that $\delta = 1$ and the previous equation implies that p is the unique prime congruent to 3 mod 4 that divides $a + 1$ to an odd power. □

Now we prove Theorem 1.3 involving a special case when $a \equiv 0 \pmod 4$.

Proof of Theorem 1.3. First, we show that $a^x + 1$ is not the sum of two squares. We have

$$a^x + 1 = \prod_{d \mid 2x,\, d \nmid x} \Phi_d(a).$$

We apply Theorem 2.1 of [Stevenhagen 1987] to $\Phi_{2x}(y) \in \mathbb{Z}[y]$. We set $n = 2x$, $k = x$, $d(k) = 4x$. Then $d(k) \nmid n$ but $d(k) \mid 2n$. We have

$$\Phi_{2x}(y) = F(y)^2 - xyG(y)^2.$$

Assume without loss of generality that the leading coefficient of $F(y)$ is positive. Note that since $\Phi_{2x}(y)$ has even degree, the degree of $F(y)$ is larger than that of $G(y)$.

Replacing y with xy^2 we get

$$\begin{aligned}\Phi_{2x}(xy^2) &= F(xy^2)^2 - x(xy^2)G(xy^2) \\ &= (F(xy^2) + xyG(xy^2))(F(xy^2) - xyG(xy^2)).\end{aligned}$$

Let $f(y)$ and $g(y)$ be the first and second factors above, respectively. We have $\Phi_{2x}(a) = \Phi_{2x}(4x) = f(2)g(2)$. From Theorem 2.7 of [Stevenhagen 1987] we know $\gcd(f(2), g(2)) = 1$. We claim $f(2) \equiv g(2) \equiv 3 \pmod 4$. This will follow if we show that the constant coefficients of $f(y)$ and $g(y)$ are both 1, and the linear coefficients of $f(y)$ and $g(y)$ are both odd.

We have $f(y) = a_0 + a_1 y + a_2 y^2 + \cdots$ and $g(y) = a_0 - a_1 y + a_2 y^2 + \cdots$. Since the constant coefficient of $\Phi_{2x}(y)$ is 1, we have $a_0^2 = 1$ and so $a_0 = \pm 1$. If $a_0 = -1$, then since the leading coefficient of $F(y)$ is positive, $f(y)$ and $g(y)$ have positive leading coefficients. However, then $\lim_{y\to\infty} f(y) = \lim_{y\to\infty} g(y) = \infty$ but $f(0) = g(0) = -1$. This implies that $f(y)$ and $g(y)$ both have a positive real root, but $f(y)g(y) = \Phi_{2x}(xy^2)$ has no real roots. This is a contradiction and so $a_0 = 1$.

It is well known that if $n > 1$, the coefficient of y in $\Phi_n(y)$ is $-\mu(n)$; see for example, the last equation on page 107 of [Lehmer 1966]. Multiplying $f(y)$ and $g(y)$, we get

$$\Phi_{2x}(xy^2) = 1 - \mu(2x)xy^2 + \cdots = a_0^2 + (2a_0a_2 - a_1^2)y^2 + \cdots.$$

We have that $\mu(2x) = \pm 1$ is odd and $-\mu(2x) = 2a_0a_2 - a_1^2$. Thus, $a_1^2 \equiv \mu(2x) \pmod 2$ and so a_1 is odd. Thus, $f(2) \equiv a_0 + 2a_1 \equiv 1 + 2 \equiv 3 \pmod 4$ and likewise $g(2) \equiv a_0 - 2a_1 \equiv 1 - 2 \equiv 3 \pmod 4$.

Thus, there is a prime $p \equiv 3 \pmod 4$ and an odd j such that $p^j \| f(2)$ and a prime $q \equiv 3 \pmod 4$ and an odd k such that $q^k \| g(2)$. Since $\gcd(f(2), g(2)) = 1$, we have $p \neq q$.

We claim that at most one of p or q divides x. Suppose to the contrary that $p \mid x$ and $q \mid x$. Since $p \mid \Phi_{2x}(a)$, Theorem 2.1 implies $2x = p \cdot \operatorname{ord}_p(a)$ and since $q \mid \Phi_{2x}(a)$, we get $2x = q \cdot \operatorname{ord}_q(a)$. This implies that $\operatorname{ord}_p(a) = 2x/p$ is a multiple of q and $\operatorname{ord}_q(a) = 2x/q$ is a multiple of p. This is a contradiction, because either $p < q$ (in which case $q \leq \operatorname{ord}_p(a) \leq p-1$) or $q < p$ (in which case $p \leq \operatorname{ord}_q(a) \leq q-1$).

Thus, at most one of p or q divides x. Assume without loss of generality that $p \nmid x$. Then we have $p^j \| \Phi_{2x}(a)$ and Theorem 2.1 gives $\operatorname{ord}_p(a) = 2x$. This implies $p \nmid \Phi_{2\delta}(a)$ for $\delta \mid x$ with $\delta \neq x$. As a consequence, $p^j \| (a^x + 1)$ and so $a^x + 1$ is not the sum of two squares.

Now, let $A = a^x$. Then $A+1$ is not the sum of two squares, and $A+1 \equiv 1 \pmod 4$. Thus, there are at least two primes $\equiv 3 \pmod 4$ that divide $A+1$ to an odd power, and Theorem 1.2 implies that $A^n + 1$ is never the sum of two squares for n odd. □

6. Odd

This section contains proofs of Theorems 1.4, 1.5, and 1.7, along with Lemma 1.6, which pertain to when $a^n + 1$ can be written as a sum of two squares when a is an odd integer. In this section, we define m to be the least positive integer such that $(a+1)/m$ is the sum of two squares.

We begin with $a \equiv 1 \pmod 4$. We prove Theorem 1.4 which handles the case $a \equiv 1 \pmod 8$, and Theorem 1.5 which handles $a \equiv 5 \pmod 8$.

Proof of Theorem 1.4. Let $a \equiv 1 \pmod 8$. Then $a^n + 1 \equiv 2 \pmod 8$ for all n, so $(a^n+1)/2 \equiv 1 \pmod 4$. Suppose $a^n + 1$ is the sum of two squares, and assume by contradiction that δ is the largest divisor of n such that $a^\delta + 1$ is not the sum of two squares. Since $(a^\delta + 1)/2 \equiv 1 \pmod 4$, there exist distinct primes $q_1 \equiv q_2 \equiv 3 \pmod 4$ such that $q_1^{j_1} \| (a^\delta + 1)$ and $q_2^{j_2} \| (a^\delta + 1)$, j_1, j_2 odd.

We know from Lemma 3.3 that since $a^n + 1$ is the sum of two squares, $q_1^{l_1} \| n$ and $q_2^{l_2} \| n$ for some odd l_1 and l_2. Without loss of generality, suppose $q_1 > q_2$, and consider

$$a^{\delta q_1} + 1 = (a^\delta + 1) \prod_{\delta_x \mid \delta q_1,\, \delta_x \nmid \delta} \Phi_{2\delta_x}(a).$$

Since $q_1 > q_2$, we know $q_1 \nmid \operatorname{ord}_{q_2}(a)$, and Theorem 2.1 implies $q_2 \nmid (a^{\delta q_1}+1)/(a^\delta+1)$. Then $q_2^{j_2} \| (a^{\delta q_1}+1)$, so $a^{\delta q_1}+1$ is not the sum of two squares. This is a contradiction because $\delta q_1 > \delta$ and $\delta q_1 \mid n$. Thus $a^\delta + 1$ is the sum of two squares for all $\delta \mid n$. □

Proof of Theorem 1.5. Suppose $a \equiv 5 \pmod 8$ and n is odd. Then

$$\begin{aligned} a^n + 1 &= a^{2k+1} + 1 \\ &\equiv 5^{2k} \cdot 5 + 1 \pmod 8 \\ &\equiv 6 \pmod 8. \end{aligned}$$

This implies that $(a^n + 1)/2 \equiv 3 \pmod 4$, so by Fermat's two squares theorem we know that $a^n + 1$ is never the sum of two squares when n is odd. □

Next, the following lemmas will be useful in forming contradictions in the proof of Theorem 1.7 because of the restrictions they place on n in order for $a^n + 1$ to be the sum of two squares, where $a \equiv 3 \pmod 4$ and n odd.

We begin with two lemmas that cover the modulus of permissible exponents n when $a \equiv 3 \pmod 4$.

Lemma 6.1. *For $a = 4 \cdot 2^i \cdot (4j + 1) - 1$ with $i, j \geq 0$, then $a^n + 1$ can only be written as the sum of two squares (for n odd) if $n \equiv 1$ mod 4.*

Note that this covers values of a such as $a = 3$, 7, 15, 19, 31, and 35. This explains why $35^9 + 1$ is a sum of two squares but $35^3 + 1$ is not.

Proof. Let us argue by contradiction. Suppose $n \equiv 3 \bmod 4$. Write $n = 4k + 3$, and note that $a \equiv 4 \cdot 2^i - 1 \bmod 16 \cdot 2^i$. Then, making liberal use of the binomial theorem on $a^3 \equiv (4 \cdot 2^i - 1)^3$ and $a^4 \equiv (4 \cdot 2^i - 1)^4$, we have:

$$\begin{aligned} a^n + 1 &= a^{4k+3} + 1 \\ &= (a^3) \cdot (a^4)^k + 1 \\ &\equiv (\cdots + 3 \cdot (4 \cdot 2^i) - 1) \cdot (\cdots - 4 \cdot (4 \cdot 2^i) + 1)^k + 1 \bmod 16 \cdot 2^i \\ &\equiv (3 \cdot 4 \cdot 2^i - 1) \cdot (1)^k + 1 \bmod 16 \cdot 2^i \\ &\equiv 12 \cdot 2^i \bmod 16 \cdot 2^i. \end{aligned}$$

This implies that $(a^n + 1)/(4 \cdot 2^i)$ is equivalent to 3 mod 4. Then there must be at least one prime equivalent to 3 mod 4 that appears in the factorization of $(a^n + 1)/(4 \cdot 2^i)$ an odd number of times. This implies the same for $a^n + 1$ and thus by Fermat, $a^n + 1$ is not the sum of two squares. This is a contradiction to our assumption and thus n cannot be equivalent to 3 mod 4. □

Lemma 6.2. *For $a = 4 \cdot 2^i \cdot (4j + 3) - 1$ with $i, j \geq 0$, then $a^n + 1$ can only be written as the sum of two squares (for n odd) if $n \equiv 3$ mod 4.*

Note that this covers values of a such as $a = 11$, 23, 27, 43, and so on, including 191 which gives us two values $n = 3$ and $n = 15$ such that $191^n + 1$ is the sum of two squares. Both 3 and 15, of course, are equivalent to 3 mod 4.

Proof. Keeping in mind that $a \equiv -1 \bmod 4$, we have

$$\begin{aligned} a^n + 1 &= (a+1) \cdot (a^{n-1} - a^{n-2} + \cdots + 1) \\ &= 4 \cdot 2^i \cdot (4j+3) \cdot (a^{n-1} - a^{n-2} + \cdots + 1). \end{aligned}$$

Since $a \equiv -1 \bmod 4$, the last expression, $(a^{n-1} - a^{n-2} + \cdots + 1)$, is equivalent to $n \bmod 4$. The only hope, then, for $a^n + 1$ to be the sum of two squares is for n to be 3 mod 4, as then $(a^n+1)/(4 \cdot 2^i)$ will be the product of two expressions both equivalent to 3 mod 4, resulting in $(a^n+1)/(4 \cdot 2^i)$ being equivalent to 1 mod 4. □

The last two lemmas allow us to now prove one of our earlier lemmas:

Proof of Lemma 1.6. For $a \equiv 3 \pmod 4$, we can write $a = 4K - 1$, where K can be split into an even part (which we write as 2^i) and an odd part (which we write as either $4j+1$ or $4j+3$). In the first case, $a+1$ equals $4 \cdot 2^i \cdot (4j+1)$ and since m is the smallest integer such that $(a+1)/m$ is the sum of two squares, m must be equivalent to $1 \pmod 4$, and by Lemma 6.1 we have $n \equiv 1 \pmod 4$ in this case, and so $n \equiv m \pmod 4$. A similar argument applies to the second case. □

This lemma places further restrictions on n. Recall that m is the smallest positive integer such that $(a+1)/m$ is the sum of two squares.

Lemma 6.3. *Let $a \equiv 3 \pmod 4$. If $a^n + 1$ is the sum of two squares, then for all primes $p \equiv 3 \pmod 4$ such that $p^e \| (a+1)$, e odd, we have $p^k \| n$, k odd. In particular, if $a^n + 1$ is the sum of two squares, then $m \mid n$.*

Proof. Let $a^n + 1$ be the sum of two squares and suppose $p^e \| (a+1)$, e odd, and $p \equiv 3 \pmod 4$. Select k such that $p^k \| n$. Then, Lemma 3.3 implies $p^{e+k} \| (a^n+1)$. Since $a^n + 1$ is the sum of two squares, we know $e+k$ is even, which makes k odd. It follows that since $m = \prod p$ for p such primes of this type, if $a^n + 1$ is the sum of two squares, then $m \mid n$. □

We will now prove Theorem 1.7, which applies to all $a \equiv 3 \pmod 4$.

Proof of Theorem 1.7. First we will prove that n/m is the sum of two squares. Suppose that $a^n + 1$ is the sum of two squares and recall that by Lemma 6.3, $m \mid n$. Assume by contradiction that n/m is not the sum of two squares. Then let q be the greatest prime such that $q \equiv 3 \pmod 4$ and $q^j \parallel n/m$, j odd. If $q \mid m$, then Lemma 3.3 implies that an even power of q divides $a^m + 1$, and so if an odd power of q divides $a^n + 1$, then $q^r \parallel n$, r odd. But m is square-free, so $q \parallel m$. Then $q^{r-1} \parallel n/m$, $r-1$ even, which is a contradiction. Therefore we can assume $q \nmid m$, so $q^j \parallel n$.

We know that $\Phi_{2q^j}(a)$ divides $a^n + 1$. We have $\Phi_{2q^j}(a) \equiv \Phi_{2q^j}(-1) \equiv \Phi_{q^j}(1) \equiv q \equiv 3 \pmod 4$. This implies that there exists a prime $p \equiv 3 \pmod 4$ such that $p^k \parallel \Phi_{2q^j}(a)$, k odd. We can consider two cases: $p \neq q$ and $p = q$.

Suppose $p \neq q$. Then $p \nmid q^j$, so $\operatorname{ord}_p(a) = 2q^j$, which implies $p > q$. Since $a^n + 1$ is the sum of two squares, Lemma 3.3 implies $p^l \parallel n$, l odd. Since $\operatorname{ord}_p(a) > 2$,

$p \nmid a+1$, so $p \nmid m$. Then p is a prime congruent to 3 (mod 4) that divides n/m to an odd power, and $p > q$, which is a contradiction because we assumed q is the largest such prime.

Now suppose $p = q$. Since $p \mid \Phi_{2p^j}(a)$, it follows that $a^{p^j} + 1 \equiv 0 \pmod{p}$. Repeatedly applying Fermat's little theorem, $a^p \equiv a \pmod{p}$, we find that $p \mid (a+1)$. Since $p \nmid m$, $p^k \| (a+1)$, k even. Then Lemma 3.3 implies that $p^{k+j} \| a^n + 1$, where $k + j$ is odd, which is a contradiction. Thus if $a^n + 1$ is the sum of two squares, then n/m is also the sum of two squares.

Next we'll prove that $a^m + 1$ is the sum of two squares. Suppose $a^n + 1$ is the sum of two squares, where $n = ms$, and assume by contradiction that $a^m + 1$ is not the sum of two squares. Then there exists some prime $q \equiv 3 \pmod{4}$ such that $q^j \| (a^m + 1)$, j odd. Since $s = n/m$ is the sum of two squares, we know $q^k \| s$, k even. Then $n = mq^k s'$, where $\gcd(s', q) = 1$, so $q^{k+j} \| (a^n + 1)$, $k + j$ odd (Lemma 3.3). This is a contradiction because we assumed $a^n + 1$ is the sum of two squares. Therefore if $a^n + 1$ is the sum of two squares for some odd n, then $a^m + 1$ is also the sum of two squares.

Let $\delta \mid (n/m)$, where δ is the sum of two squares, and suppose $a^n + 1$ is the sum of two squares. We will show that $a^{m\delta} + 1$ is the sum of two squares. Assume by contradiction that there exists a prime $q \equiv 3 \pmod{4}$ such that $q^j \| a^{m\delta} + 1$, j odd.

Since δ is the sum of two squares, we know $q^k \| \delta$, k even, $k \geq 0$. Because q must divide $a^n + 1$ to an even power, Lemma 3.3 implies $q^l \| n/(m\delta)$, l odd, so $q^{l+k} \| n/m$, $l + k$ odd, which is a contradiction because n/m is the sum of two squares. Thus if $a^n + 1$ is the sum of two squares, $a^{m\delta} + 1$ is the sum of two squares for all $\delta \mid n/m$ such that δ is the sum of two squares.

Finally, we will show that if $a^{np^2} + 1$ is the sum of two squares for some $p \equiv 3 \pmod{4}$, then $p \mid (a^n + 1)$. By Lemma 1.6 we know $a^{np} + 1$ is not the sum of two squares, so there exists some $q \equiv 3 \pmod{4}$ with $q^j \| (a^{np} + 1)$, j odd. If $q \neq p$, then by Lemma 3.3 we have $q^j \| (a^{np^2} + 1)$, j odd, which contradicts $a^{np^2} + 1$ being the sum of two squares. Hence $q = p$, and since $p \mid (a^{np} + 1)$ and $a^{np} \equiv a^n \pmod{p}$, we have $p \mid (a^n + 1)$, as desired. □

We conclude this section with a heuristic giving evidence for Conjecture 1.10. Suppose first that $a \equiv 0$ or 1 mod 4. In this case, if $a^n + 1$ is the sum of two squares for any n, then $a + 1$ is the sum of two squares. Let A_p be the event that $\Phi_{2p}(a)$ is the sum of two squares. It seems plausible that the probability that this occurs is approximately

$$\frac{K}{\sqrt{\ln(\Phi_{2p}(a))}} \approx \frac{K}{\sqrt{p}}.$$

Since $\sum_{p \equiv 1 \pmod{4}} 1/\sqrt{p}$ diverges, we should expect an infinite number of the events A_p to occur, and this would yield infinitely many primes p for which $a^p + 1$ is the sum of two squares.

If $a \equiv 2 \pmod 4$, then Theorem 1.2 implies there is at most one n such that a^n+1 is the sum of two squares.

In the case that $a \equiv 3 \pmod 4$, let m denote the smallest positive integer such that $(a+1)/m$ is the sum of two squares. First, consider primes $p \equiv 1 \pmod 4$ such that $a^{mp}+1$ is the sum of two squares. We have

$$\frac{a^{mp}+1}{a^m+1} = \prod_{d \mid 2mp,\, d \nmid 2m} \Phi_d(a).$$

Theorem 2.1 implies that if we write $\Phi_d(a) = \gcd(\Phi_d(a), m)c_d$, then the c_d are pairwise coprime and this implies that c_d is the sum of two squares for all d. It seems plausible that the c_d being the sum of two squares are independent, and so the probability that $a^{mp}+1$ is the sum of two squares is approximately

$$\prod_d \frac{1}{\sqrt{\ln(c_d)}} \approx p^{-\tau(m)/2},$$

where $\tau(m)$ is the number of divisors of m. The sum $\sum_{p \equiv 1 \pmod 4 \text{ prime}} p^{-\tau(m)/2}$ diverges if $m = 1$ or m is prime, and converges if m is composite. In particular, in the case that m is composite, there are only finitely many primes p such that $a^{mp}+1$ is the sum of two squares.

Then, Theorem 1.7 implies that there are only finitely many primes that can divide some number n such that a^n+1 is the sum of two squares. If there are infinitely many n such that a^n+1 is the sum of two squares, it follows then that there is a prime p such that $a^{p^r}+1$ is the sum of two squares for infinitely many r. We have $a^{p^r}+1 = \prod_{i=0}^{r} \Phi_{2p^i}(a)$. If we write

$$r_i = \frac{\Phi_{2p^i}(a)}{\gcd(\Phi_{2p^i}(a),\, p)},$$

then Theorem 2.1 implies $\gcd(r_i, r_j) = 1$. It follows from this that r_i is the sum of two squares for all $i \geq 1$. Assuming that these events are independent, the probability this occurs is $\sum_i K/\sqrt{\ln(r_i)}$. But this sum converges. Therefore the "probability is zero" that there are infinitely many n such that a^n+1 is the sum of two squares in the case when $a \equiv 3 \pmod 4$ and m is composite.

As an example, we consider $a = 4713575$, with a composite m value of $m = 21$. We conjecture that there are finitely many n such that a^n+1 is the sum of two squares. So far, we know only of $n = 21$ and $n = 105$.

7. $p \equiv 1 \pmod 4$

The previous theorems put constraints on when a^n+1 can be the sum of two squares for different categories of a. The following proof of Theorem 1.8 uses Aurifeuillian

factorization to show that when $a = pv^2$, where $p \equiv 1 \pmod 4$ is a prime and $p \nmid v$, there are either zero or infinitely many odd integers n such that $a^n + 1$ is the sum of two squares.

Proof of Theorem 1.8. Let $a = pv^2$, where $p \equiv 1 \pmod 4$ is prime. Suppose $a^n + 1$ is the sum of two squares and consider

$$a^{np} + 1 = \prod_{\delta \mid n} \Phi_{2\delta}(a) \prod_{\delta \mid np,\, \delta \nmid n} \Phi_{2\delta}(a).$$

We know $\prod_{\delta \mid n} \Phi_{2\delta}(a) = a^n + 1$ is the sum of two squares. Consider the Aurifeuillian factorization of $\Phi_{2\delta}(a)$, where $\delta \mid np,\ \delta \nmid n,\ x = -kv^2,\ k = -p \equiv 3 \pmod 4$, and q is odd:

$$\begin{aligned}
\Phi_{2\delta}(x) &= (F(x))^2 - kx^q (G(x))^2, \\
\Phi_{2\delta}(-kv^2) &= (F(-kv^2))^2 - k(-kv^2)^q (G(-kv^2))^2 \\
&= (F(-kv^2))^2 + k^{q+1} v^{2q} (G(-kv^2))^2 \\
&= (F(-kv^2))^2 + (k^{(q+1)/2} v^q G(-kv^2))^2 \\
&= \Phi_{2\delta}(a).
\end{aligned}$$

Therefore $\Phi_{2\delta}(a)$ is the sum of two squares for any $\delta \mid np$ with $\delta \nmid n$. Thus $a^{np} + 1$ is the sum of two squares. Conversely, suppose that $a^{np} + 1$ is the sum of two squares. Then we can see again that $\Phi_{2\delta p}(a)$ is the sum of two squares for any factor δ. This implies that $\prod_{\delta \mid n} \Phi_{2\delta}(a) = a^n + 1$ is the sum of two squares. □

Now, we will construct an infinite family of numbers $a = f(X)$ such that $a^p + 1$ is the sum of two squares.

Proof of Theorem 1.9. If $p \equiv 1 \pmod 4$, then there exists an even integer u and an odd integer v such that $p = u^2 + v^2$. Then consider the polynomials

$$\begin{aligned}
A(X) &= \tfrac{1}{2} upX^2 + vX, \\
B(X) &= \tfrac{1}{2} u^2 pX^2 - 1, \\
C(X) &= \tfrac{1}{2} uvpX^2 + pX.
\end{aligned}$$

Let $f(X) = pA(X)^2$; then we have

$$\begin{aligned}
f(X)^p + 1 &= (f(X) + 1)\Phi_{2p}(f(X)) \\
&= (pA(X)^2 + 1)\Phi_{2p}(pA(X)^2).
\end{aligned}$$

It is straightforward to check that $f(X) + 1$ can be written as the sum of two squares: $pA(X)^2 + 1 = B(X)^2 + C(X)^2$. Then consider the Aurifeuillian

factorization of $\Phi_{2p}(x)$, where we let $k=-p$ and $x=pA(X)^2$. Then we get

$$\begin{aligned}\Phi_{2p}(x) &= F(x)^2 - kxG(x)^2,\\ \Phi_{2p}(pA(X)^2) &= (F(pA(X)^2))^2 - p(-pA(X)^2)(G(pA(X)^2))^2\\ &= (F(pA(X)^2))^2 + (p^2A(X)^2)(G(pA(X)^2))^2\\ &= (F(pA(X)^2))^2 + \big(pA(X)(G(pA(X)^2))\big)^2.\end{aligned}$$

Therefore, $\Phi_{2p}(f(X))$ can be written as the sum of two squares as well. This implies that $f(X)^p+1$ is the product of two terms, each of which can be written as the sum of two squares. □

8. Chart

Here we illustrate the first few odd integers n such that a^n+1 is the sum of two squares for all integers $a \in [1, 50]$.

a	n	Theorem	a	n	Theorem
1	all	1.1	26	–	1.2
2	3	1.2	27	–	1.7
3	1, 5, 13, 65, …	1.7	28	1, 3, 11, 19, …	1.2
4	all	1.1	29	–	1.5
5	–	1.5	30	31	1.2
6	7	1.2	31	1, 5, 25, 41, …	1.7
7	1, 13, 17, 29, …	1.7	32	–	1.2
8	1	1.2	33	1, 5, 7, 17, …	1.4
9	all	1.1	34	–	1.2
10	–	1.2	35	1, 9, 13, 29, …	1.7
11	3, 159, …	1.7	36	all	1.1
12	1, 5, 11, 23, …	1.2	37	–	1.5
13	–	1.5	38	–	1.2
14	3	1.2	39	1, 13, 37, 61, …	1.7
15	1, 29, 89, 97, …	1.7	40	1, 5, 13, 53, …	1.2
16	all	1.1	41	–	1.4
17	1, 7, 17, 23, …	1.8	42	–	1.2
18	19	1.2	43	–	1.7
19	1, 17, 29, 37, …	1.7	44	1, 5, 7, 17, …	1.2
20	–	1.2	45	–	1.5
21	–	1.5	46	–	1.2
22	–	1.2	47	–	1.7
23	3, 123, …	1.7	48	1, 3, 5, 17, …	1.2
24	1, 7, 11, 19, …	1.2	49	all	1.1
25	all	1.1	50	–	1.2

Acknowledgements

The authors would like to thank the anonymous referees for helpful comments that improved the exposition.

References

[Bang 1886a] A. S. Bang, "Taltheoretiske undersøgelser", *Tidsskrift Math.* (5) **4** (1886), 70–80.

[Bang 1886b] A. S. Bang, "Taltheoretiske undersøgelser (fortsat, se. s. 80)", *Tidsskrift Math.* (5) **4** (1886), 130–137.

[Brillhart et al. 2002] J. Brillhart, D. H. Lehmer, J. L. Selfridge, B. Tuckerman, and S. S. Wagstaff, Jr., *Factorizations of $b^n \pm 1$: $b = 2, 3, 5, 6, 7, 10, 11, 12$ up to high powers*, 3rd ed., Contemporary Mathematics **22**, American Mathematical Society, Providence, RI, 2002.

[Bugeaud et al. 2006] Y. Bugeaud, M. Mignotte, and S. Siksek, "Classical and modular approaches to exponential Diophantine equations, I: Fibonacci and Lucas perfect powers", *Ann. of Math.* (2) **163**:3 (2006), 969–1018. MR Zbl

[Carmichael 1913/14] R. D. Carmichael, "On the numerical factors of the arithmetic forms $\alpha^n \pm \beta^n$", *Ann. of Math.* (2) **15**:1-4 (1913/14), 49–70. MR Zbl

[Cox 1989] D. A. Cox, *Primes of the form $x^2 + ny^2$: Fermat, class field theory and complex multiplication*, John Wiley & Sons, New York, 1989. MR Zbl

[Curtis 2014] K. Curtis, "Sums of two squares: an analysis of numbers of the form $2^n + 1$ and $3^n + 1$", preprint, 2014.

[Ford et al. 2018] K. Ford, B. Green, S. Konyagin, J. Maynard, and T. Tao, "Long gaps between primes", *J. Amer. Math. Soc.* **31**:1 (2018), 65–105. MR Zbl

[Hardy and Wright 1979] G. H. Hardy and E. M. Wright, *An introduction to the theory of numbers*, 5th ed., Oxford University Press, New York, 1979. MR Zbl

[Hooley 1971] C. Hooley, "On the intervals between numbers that are sums of two squares", *Acta Math.* **127** (1971), 279–297. MR Zbl

[Iwaniec 1972] H. Iwaniec, "Primes of the type $\phi(x, y) + A$ where ϕ is a quadratic form", *Acta Arith.* **21** (1972), 203–234. MR Zbl

[Kalmynin 2017] A. Kalmynin, "Intervals between numbers that are sums of two squares", preprint, 2017. arXiv

[Landau 1908] E. Landau, "Über die Einteilung der positiven ganzen Zahlen in vier Klassen nach der Mindeszahl der zu ihrer additiven Zusammensetzung erforderlichen quadrate", *Arch. Math. Phys.* **13** (1908), 305–312.

[Lehmer 1966] D. H. Lehmer, "Some properties of the cyclotomic polynomial", *J. Math. Anal. Appl.* **15** (1966), 105–117. MR Zbl

[Lucas 1878] E. Lucas, "Theorie des fonctions numeriques simplement periodiques", *Amer. J. Math.* **1**:2 (1878), 184–196. MR Zbl

[Maynard 2015] J. Maynard, "Small gaps between primes", *Ann. of Math.* (2) **181**:1 (2015), 383–413. MR Zbl

[Nagell 1948] T. Nagell, "Løsning till oppgave nr 2", *Norsk Mat. Tidsskr.* **30** (1948), 62–64.

[Nagell 1964] T. Nagell, *Introduction to number theory*, 2nd ed., Chelsea, New York, 1964. MR Zbl

[Richards 1982] I. Richards, "On the gaps between numbers which are sums of two squares", *Adv. in Math.* **46**:1 (1982), 1–2. MR Zbl

[Stevenhagen 1987] P. Stevenhagen, "On Aurifeuillian factorizations", *Nederl. Akad. Wetensch. Indag. Math.* **49**:4 (1987), 451–468. MR Zbl

[Stewart 1977] C. L. Stewart, "On divisors of Fermat, Fibonacci, Lucas, and Lehmer numbers", *Proc. London Math. Soc.* (3) **35**:3 (1977), 425–447. MR Zbl

[Stewart and Tall 2002] I. Stewart and D. Tall, *Algebraic number theory and Fermat's last theorem*, 3rd ed., A K Peters, Natick, MA, 2002. MR Zbl

[Wagstaff] S. S. Wagstaff, Jr., "The Cunningham Project", website, http://homes.cerias.purdue.edu/~ssw/cun/index.html.

[Zhang 2014] Y. Zhang, "Bounded gaps between primes", *Ann. of Math.* (2) **179**:3 (2014), 1121–1174. MR Zbl

Received: 2017-10-11 Revised: 2018-06-20 Accepted: 2018-06-24

dresdeng@wlu.edu *Department of Mathematics, Washington and Lee University, Lexingston, VA, United States*

kylie.hess@emory.edu *Department of Mathematics and Computer Science, Emory University, Atlanta, GA, United States*

islams19@wlu.edu *Department of Mathematics, Washington and Lee University, Lexington, VA, United States*

rouseja@wfu.edu *Department of Mathematics and Statistics, Wake Forest University, Winston-Salem, NC, United States*

aaronschmitt96@gmail.com *Department of Mathematics, Washington and Lee University, Lexington, VA, United States*

emily.stamm.12@gmail.com *Department of Mathematics and Statistics, Vassar College, Poughkeepsie, NY, United States*

warrentm@uga.edu *Department of Mathematics, University of Georgia, Athens, GA, United States*

pany19@wlu.edu *Department of Mathematics, Washington and Lee University, Lexington, VA, United States*

INVOLVE 12:4 (2019)
dx.doi.org/10.2140/involve.2019.12.607

Irreducible character restrictions to maximal subgroups of low-rank classical groups of types B and C

Kempton Albee, Mike Barnes, Aaron Parker,
Eric Roon and A. A. Schaeffer Fry

(Communicated by Scott T. Chapman)

Representations are special functions on groups that give us a way to study abstract groups using matrices, which are often easier to understand. In particular, we are often interested in irreducible representations, which can be thought of as the building blocks of all representations. Much of the information about these representations can then be understood by instead looking at the trace of the matrices, which we call the character of the representation. This paper will address restricting characters to subgroups by shrinking the domain of the original representation to just the subgroup. In particular, we will discuss the problem of determining when such restricted characters remain irreducible for certain low-rank classical groups.

1. Introduction

Given a finite group G, a (*complex*) *representation* of G is a homomorphism $\Psi : G \to \mathrm{GL}_n(\mathbb{C})$. By summing the diagonal entries of the images $\Psi(g)$ for $g \in G$ (that is, taking the trace of the matrices), we obtain the corresponding *character*, $\chi = \mathrm{Tr} \circ \Psi$ of G. The *degree* of the representation Ψ or character χ is $n = \chi(1)$. It is well known that any character of G can be written as a sum of so-called *irreducible* characters of G. In this sense, irreducible characters are of particular importance in representation theory, and we write $\mathrm{Irr}(G)$ to denote the set of irreducible characters of G.

Given a subgroup H of G, we may view Ψ as a representation of H as well, simply by restricting the domain. As such, we will write $\chi|_H$ to denote the corresponding character of H, called the restricted character or character restriction. In this paper, we are interested in the general problem of classifying triples (G, H, χ), where G is a finite group, H is a maximal subgroup, and χ is an irreducible character of G

MSC2010: 20C15, 20C33.
Keywords: irreducible characters, classical groups.

whose restriction to H remains irreducible. There is a large body of work on this topic, see [Brundan and Kleshchev 2003; Kleshchev and Sheth 2002; Kleshchev and Tiep 2004; 2010; Liebeck 1985; Seitz 1987; Seitz and Testerman 1990; Nguyen and Tiep 2008; Himstedt et al. 2009; Nguyen 2008; Seitz 1990; Schaeffer Fry 2013], but several interesting cases remain unsolved.

We remark that although the general problem of classifying irreducible restrictions is of interest for representations over general fields, we work in this paper only with complex representations, and therefore the term "character" will refer specifically to complex characters here.

In this paper, we are concerned with the case that G is a classical group and H is a maximal subgroup of G. (For a brief introduction to classical groups, see Section 2A below.) In [Schaeffer Fry 2013], the faculty author classified all triples as above in the case $G = \mathrm{Sp}_4(q)$ or $\mathrm{Sp}_6(q)$, where q is a power of 2. There, and in many of the other articles on the topic, there are relatively few maximal subgroups that need to be considered using advanced techniques. In [Schaeffer Fry 2013], the process of reducing to these more difficult cases is referred to as the "initial reduction". Since $\mathrm{Sp}_6(2^a) \cong \Omega_7(2^a)$, the natural next step is to address the cases $G = \mathrm{Sp}_6(q)$ or $\Omega_7(q)$ with q odd.

Hence, here we work with symplectic groups $\mathrm{Sp}_{2n}(q)$ and orthogonal groups $\Omega_{2n+1}(q)$ with $1 \leq n \leq 3$ and q a power of an odd prime, which corresponds to the groups of Lie type B and C. Specifically, the goal of this paper is to provide the "initial reduction" for these groups, which leaves a short list of more difficult subgroups to be addressed. Our main results, providing this "initial reduction", are found in Theorems 4.1, 5.1, 6.1, and 7.1. Further, we provide complete classifications of irreducible restrictions for small values of q, which is found in Section 8.

The organization of the paper is as follows. In Section 2, we introduce some background material regarding finite classical groups and representations. In Section 3, we discuss the code used in the computer algebra system GAP for the cases that q is small. The remainder of the paper is dedicated to the main results.

1A. ***Notation.*** Here we introduce some basic notation for products and extensions of groups, which will be found throughout the paper. If H is a subgroup of G, we denote by $[G : H]$ the index of H in G. Given two groups X and Y, we denote the direct product of X and Y by $X \times Y$. The notation $X \circ Y$ will denote any central product of X and Y as defined in [Gorenstein 1968, Theorem 5.3]. Such a group is defined with respect to a subgroup Z of $Z(X)$ that may be identified under an isomorphism with a subgroup of $Z(Y)$. Then X and Y generate the group $X \circ Y$ and centralize each other, and $Z = X \cap Y \subseteq Z(X \circ Y)$.

If X acts on Y, we denote the semidirect product of Y with X by $Y : X$, defined as in [Gorenstein 1968, Theorem 5.1]. Here Y and X may be viewed as subgroups

of $Y : X$ satisfying that Y is normal in $Y : X$ and $Y \cap X = \{1\}$. More generally, if Y is a normal subgroup of G with quotient $G/Y \cong X$, we write $G = Y.X$ or $Y^{\cdot}X$, where we use the latter if we specifically know that Y has no complement in G. If r and m are positive integers, we may simply write r^m for the direct product, $(C_r)^m$, of m copies of a cyclic group of order r.

If $q = p^a$ is a power of a prime, an elementary abelian group C_p^a of order q will be denoted by E_q. We will use S_n and A_n to denote the symmetric and alternating groups, respectively, on n letters. The *wreath product* of a group X and S_n will be denoted by $X \wr S_n$. This can be thought of as a semidirect product $X^n : S_n$, where X^n denotes the direct product of n copies of X. Further, D_n will denote the dihedral group of order $2n$. Given two integers r and m, we will write (r, m) for the gcd of the two integers.

2. Background material

2A. ***The finite classical groups.*** In this section, we introduce the main groups of study in this paper. Readers familiar with the construction of the finite classical groups may feel free to disregard this section. We will view the classical groups here as groups of matrices, although they may also be viewed as groups of Lie type or as certain groups of linear transformations. For a more in-depth discussion of these groups, we refer the reader to [Grove 2002; Kleidman and Liebeck 1990, Section 2].

Let q be a power of a prime p, and let $\mathbb{F}_q$ denote a finite field of size q. In general, the finite classical groups can be viewed as subgroups or subquotients of the *general linear group* $\mathrm{GL}_n(q)$, which is composed of all invertible $n \times n$ matrices with entries in $\mathbb{F}_q$. The *special linear group* is the normal subgroup, $\mathrm{SL}_n(q)$, of matrices with determinant 1. We obtain the *projective special linear group* as the quotient $\mathrm{PSL}_n(q) = \mathrm{SL}_n(q)/Z(\mathrm{SL}_n(q))$. The sizes of these groups are

$$|\mathrm{GL}_n(q)| = q^{\frac{1}{2}n(n-1)} \prod_{k=1}^{n} (q^k - 1),$$

$$|\mathrm{SL}_n(q)| = \frac{|\mathrm{GL}_n(q)|}{q-1}, \quad |\mathrm{PSL}_n(q)| = \frac{|\mathrm{SL}_n(q)|}{(n, q-1)}.$$

The *general unitary group* is a subgroup of $\mathrm{GL}_n(q^2)$, and can be defined as

$$\mathrm{GU}_n(q) := \{A \in \mathrm{GL}_n(q^2) : \bar{A}^T A = I_n\},$$

where $\bar{A}^T$ is the matrix obtained from A by raising each entry to the q-th power and taking the transpose. The *special unitary group*, $\mathrm{SU}_n(q)$, is the subgroup of $\mathrm{GU}_n(q)$ of matrices with determinant 1, and the *projective special unitary group* is

the quotient $\mathrm{PSU}_n(q) = \mathrm{SU}_n(q)/Z(\mathrm{SU}_n(q))$. The corresponding sizes are

$$|\mathrm{GU}_n(q)| = q^{\frac{1}{2}n(n-1)} \prod_{k=1}^{n} (q^k - (-1)^k),$$

$$|\mathrm{SU}_n(q)| = \frac{|\mathrm{GU}_n(q)|}{q+1}, \quad |\mathrm{PSU}_n(q)| = \frac{|\mathrm{SU}_n(q)|}{(n, q+1)}.$$

The *symplectic group* can be viewed as the subgroup

$$\mathrm{Sp}_{2n}(q) = \{g \in \mathrm{GL}_{2n}(q) : g^T J g = J\},$$

where J is the matrix

$$J = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix},$$

I_n is the $n \times n$ identity matrix, and g^T is the transpose of g. Note here that the dimension must be even. The *projective symplectic group* is then $\mathrm{PSp}_{2n}(q) = \mathrm{Sp}_{2n}(q)/Z(\mathrm{Sp}_{2n}(q))$. We have $|\mathrm{Sp}_{2n}(q)| = q^{n^2} \prod_{k=1}^{n} (q^{2k} - 1)$ and $|\mathrm{PSp}_{2n}(q)| = \frac{1}{2}|\mathrm{Sp}_{2n}(q)|$ when q is odd. For most values of n, q, the groups $\mathrm{PSL}_n(q)$, $\mathrm{PSU}_n(q)$, and $\mathrm{PSp}_{2n}(q)$ are simple.

The last type of finite classical group comes from the *orthogonal groups*. We will be primarily interested in odd-dimensional orthogonal groups. In this case, we can define $O_{2n+1}(q) := \{g \in \mathrm{GL}_{2n+1}(q) : g^T M g = M\}$, where M is the matrix

$$M = \begin{bmatrix} 0 & I_n & 0 \\ I_n & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The size of this group is $|O_{2n+1}(q)| = 2q^{n^2} \prod_{k=1}^{n} (q^{2k} - 1)$.

Taking the subgroup of elements with determinant 1, we get the *special orthogonal group*, denoted by $\mathrm{SO}_{2n+1}(q)$. Now, for $n \geq 1$ and q odd, $\mathrm{SO}_{2n+1}(q)$ contains a unique subgroup of index 2, which we denote by $\Omega_{2n+1}(q)$. The size of $\Omega_{2n+1}(q)$ is $\frac{1}{2} q^{n^2} \prod_{k=1}^{n} (q^{2k} - 1)$.

We remark that in even dimension, there are similar constructions, but this leads to two isomorphism classes: $\Omega^-_{2n}(q)$ and $\Omega^+_{2n}(q)$. We do not discuss these groups further, as for our purposes the isomorphisms mentioned below will suffice.

When the rank of the matrices is small, there are "accidental" isomorphisms between classical groups. The next theorem, found as part of [Kleidman and Liebeck 1990, Proposition 2.9.1] lists several of these isomorphisms relevant to the current work.

Theorem 2.1. *The following isomorphisms hold*:

- $\mathrm{SL}_2(q) \cong \mathrm{Sp}_2(q) \cong \mathrm{SU}_2(q)$.
- $\mathrm{PSL}_2(q) \cong \Omega_3(q)$ *for q odd.*

- $\Omega_4^-(q) \cong \mathrm{PSL}_2(q^2)$.
- $\Omega_4^+(q) \cong \mathrm{SL}_2(q) \circ \mathrm{SL}_2(q) \cong 2.(\mathrm{PSL}_2(q) \times \mathrm{PSL}_2(q))$.
- $\mathrm{PSp}_4(q) \cong \Omega_5(q)$ *for* q *odd.*

2A1. *Maximal subgroups of finite classical groups.* For the purposes of this paper, we are interested in restricting irreducible characters of finite classical groups to maximal subgroups. Understanding and classifying the maximal subgroups of the finite classical groups has been a topic of particular importance in group theory and representation theory. We encourage the interested reader to explore the texts discussed here.

Aschbacher [1984] showed that a maximal subgroup of a finite classical group lies either in the class $\mathcal{C} = \mathcal{C}_1 \cup \cdots \cup \mathcal{C}_8$ composed of eight naturally defined subclasses of subgroups or a collection $\mathcal{S}$ of almost quasisimple groups satisfying certain properties. Kleidman and Liebeck [1990] classify which groups in $\mathcal{C}$ are indeed maximal. For low-rank classical groups, all maximal subgroups have been classified by Bray, Holt, and Roney-Dougal [Bray et al. 2013].

2B. ***Preliminary observations on characters.*** Throughout, we denote by

$$b(G) := \max\{\chi(1) : \chi \in \mathrm{Irr}(G)\}$$

the largest irreducible character degree of G. It is well known that an upper bound for $b(G)$ is given by $\sqrt{|G|}$, which follows from the fact that $|G|$ can be expressed as the sum of the squares of the irreducible character degrees.

Now, note that if $\chi \in \mathrm{Irr}(G)$ restricts irreducibly to a subgroup H, then $\chi(1) = \chi|_H(1)$ must be at most $b(H)$. As we will use this fact throughout the paper, we record it here:

Lemma 2.2. *Let $H \leq G$ and $\chi \in \mathrm{Irr}(G)$. If $\chi(1) > b(H)$, we must have $\chi|_H$ is reducible.*

This yields the following corollary, which will be essential throughout the following sections.

Corollary 2.3. *Let $H \leq G$ and $\chi \in \mathrm{Irr}(G)$. Then if $\chi(1) \geq \sqrt{|H|}$, we must have $\chi|_H$ is reducible.*

One of our primary tools will be to use Lemma 2.2 or Corollary 2.3. It will be useful to have more efficient bounds for $b(H)$, however. The following well-known results from character theory will be crucial in this regard.

Theorem 2.4 (Itô's theorem, [Isaacs 1976, Theorem 6.15]). *If $H \lhd G$ is a normal abelian subgroup with $[G : H] = n$, then $\chi(1)$ divides n for all $\chi \in \mathrm{Irr}(G)$.*

Theorem 2.5 (Clifford's theorem, [James and Liebeck 2001, Theorem 20.8]). *If $H \lhd G$ are groups with H normal, and $\chi \in \mathrm{Irr}(G)$, then the restriction $\chi|_H$ satisfies*:

(1) $\chi|_H = e\left(\sum_{i=1}^{m} \psi_i\right)$ *for some irreducible characters* $\psi_i \in \mathrm{Irr}(H)$, $1 \leq i \leq m$, *and some positive integers* e *and* m.

(2) *All constituents* ψ_i *of* $\chi|_H$ *have the same degree.*

Theorem 2.6 [Isaacs 1976, Corollary 11.29]. *Let* $N \lhd G$ *and* $\chi \in \mathrm{Irr}(G)$. *Let* $\theta \in \mathrm{Irr}(N)$ *be a constituent of* $\chi|_N$. *Then,* $\chi(1)/\theta(1)$ *divides the index* $[G : N]$.

We remark that Theorem 2.4 can be viewed as a corollary to Theorem 2.6 and the fact that irreducible characters of an abelian group are always linear. (That is, every irreducible character of an abelian group has degree 1.)

It will also be beneficial to understand the characters of certain products of groups.

Theorem 2.7 [James and Liebeck 2001, Theorem 19.18]. *Let* G_1 *and* G_2 *be groups with corresponding irreducible characters* $\chi \in \mathrm{Irr}(G_1)$ *and* $\psi \in \mathrm{Irr}(G_2)$. *Then the function* $\chi \times \psi : G_1 \times G_2 \to \mathbb{C}$, *defined by* $(\chi \times \psi)(g, h) = \chi(g)\psi(h)$ *for* $g \in G_1$, $h \in G_2$, *is an irreducible character of the direct product* $G_1 \times G_2$. *Moreover, every irreducible character of* $G_1 \times G_2$ *is of this form.*

We remark that given two groups G_1, G_2, a central product $G_1 \circ G_2$ is, in a sense, lateral to other types of products we have come to understand, since the groups G_1 and G_2 do *not* have a trivial intersection. However, central products do have the property that all the elements in common commute with all other elements of the larger group. For a discussion of the representation theory of these objects, we refer the reader to [Gorenstein 1968, Chapter 3.7]. As discussed there, the irreducible characters of a central product $G_1 \circ G_2$ can be viewed as irreducible characters of a factor group when a suitable normal subgroup is chosen from the kernel of the representation. For our purposes, this means that irreducible characters of a central product $G_1 \circ G_2$ can again be taken to be products of characters of G_1 and G_2.

We end this section with the following lemma recording a divisibility property for the first several cyclotomic polynomials.

Lemma 2.8. *Let* $q \geq 3$ *be an odd number*:

(a) *If* $\ell \geq 5$ *is prime, then* ℓ *divides at most one of* q, $q-1$, $q+1$, q^2+1, q^2+q+1, *and* q^2-q+1.

(b) 3 *divides at most one of* q, $q-1$, *and* $q+1$, *and does not divide* q^2+1.

(c) *If* 3 *divides* $q - \epsilon$ *with* $\epsilon \in \{\pm 1\}$, *then*
- 3 *divides* $q^2 + \epsilon q + 1$;
- 9 *does not divide both* $q - \epsilon$ *and* $q^2 + \epsilon q + 1$; *and*
- 3 *does not divide* $q^2 - \epsilon q + 1$.

Proof. Let ℓ be a prime. Suppose first that ℓ divides q. We see easily that this implies that the other listed values are congruent to $\pm 1 \pmod{\ell}$, and hence cannot

be divisible by ℓ. Similarly, if ℓ divides $q \pm 1$, then the remaining values are congruent to 1, ± 2, or 3 (mod ℓ), and hence cannot be divisible by ℓ unless $\ell = 2$ or 3. When $\ell = 3$, we see from this that the only possibilities are that 3 divides $q-1$ and q^2+q+1 or that 3 divides $q+1$ and q^2-q+1. If ℓ divides q^2+1, then it cannot divide q, $q-1$, or $q+1$ from above, and the remaining two values are congruent to $\pm q$ (mod ℓ). Since ℓ does not divide q, the latter are also not divisible by ℓ. If ℓ divides one of the last two values listed and is larger than 3, then it cannot divide any of the first four by the previous arguments. Further, the remaining value is congruent to $\pm 2q$ (mod ℓ), and hence again cannot be divisible by ℓ. Finally, if 9 divides $q \pm 1$, then $q^2 \mp q + 1 \equiv 3 \pmod 9$ and hence is not divisible by 9. □

3. Using GAP

GAP is a computer algebra system ("Groups, algorithms, and programming") that is extremely useful for computing with finite groups and their characters. For the purposes of this paper, we especially make use of the character table library package [Breuer 2013], which builds on the results in the ATLAS [Conway et al. 1985] and contains several character tables, lists of maximal subgroups, and other useful information about certain small groups. Our goal in this section is to describe some of the functions and commands that will be useful for our results.

We can obtain the character table and corresponding irreducible character values for groups stored in the character table library by using the commands `CharacterTable` and `Irr`, respectively. For many groups stored in the library, the list of maximal subgroups is also available, which can be obtained using the `Maxes` command. Given the character table for a maximal subgroup H of G stored in GAP, the library also has the fusion of classes stored (that is, the way conjugacy classes of H embed into those of G), which is necessary for comparing the characters of H to those of G for the purposes of understanding the restrictions. This is obtained using the command `GetFusionMap`.

Below is the code used to generate our results in Section 8, given the character tables stored in the library, `ctg` and `cth` for G and H, respectively. This gives the indices of the nonlinear irreducible characters of G that restrict irreducibly to H:

```
irrg:=Irr(ctg);
irrh:=Irr(cth);
fus:=GetFusionMap(cth,ctg);
PositionsProperty(irrg, x -> x[1] > 1 and x{fus} in irrh);
```

4. Restrictions from $G = \Omega_3(q) \cong \mathrm{PSL}_2(q)$

In this section, we let G be the finite group $\Omega_3(q) \cong \mathrm{PSL}_2(q)$, where $q \geq 5$ is a power of an odd prime p. The character table for G is well-known, and the set of

nontrivial character degrees is $\left\{\frac{1}{2}(q+\epsilon), q-1, q, q+1\right\}$, where $\epsilon \in \{\pm 1\}$ is such that $q \equiv \epsilon \pmod 4$.

From [Bray et al. 2013, Table 8.7], we see that a maximal subgroup H of G is isomorphic to one of the following:

(1) A_5 for $q = p \equiv \pm 1 \pmod{10}$ or $q = p^2$, with $p \equiv \pm 3 \pmod{10}$.

(2) S_4 for $q = p \equiv \pm 1 \pmod 8$.

(3) A_4 for $q = p \equiv \pm 3, 5, \pm 11, \pm 13, \pm 19 \pmod{40}$.

(4) $D_{q\pm 1}$ for $q > 5$.

(5) $\Omega_3(q_0)$ for $q = q_0^r$, with r an odd prime.

(6) $\mathrm{SO}_3(q_0)$ for $q = q_0^2$.

(7) $E_q : \left(\frac{1}{2}(q-1)\right)$.

The goal of this section is to prove the following theorem:

Theorem 4.1. *Let $q \geq 13$ and $G \cong \Omega_3(q)$. Let H be a maximal subgroup and $\chi \in \mathrm{Irr}(G)$ such that $\chi(1) \neq 1$ and $\chi|_H$ is irreducible. Then $q \equiv 3 \pmod 4$, $H \cong E_q : \left(\frac{1}{2}(q-1)\right)$, and $\chi(1) = \frac{1}{2}(q-1)$.*

We prove Theorem 4.1 in Lemmas 4.2–4.6 below by addressing the cases (1)–(7) individually. We address the case $5 \leq q \leq 11$ in Section 8. Throughout the remainder of the section, let $\mathcal{L}$ denote the real-valued function

$$\mathcal{L}(x) = \tfrac{1}{2}(x-1).$$

Note that $\chi(1) \geq \mathcal{L}(q)$ for any $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$ and that $\mathcal{L}$ is increasing for all real x.

Lemma 4.2. *Let $H \cong A_5$, if $q = p \equiv \pm 1 \pmod{10}$, or $q = p^2$, with $p \equiv \pm 3 \pmod{10}$ and let $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$. Then $\chi|_H$ is reducible, except possibly when $q = 11$ or $q = 9$.*

Proof. First, note that $|H| = 60$ and that $q > 19$ unless $q = 9$ or 11. Since $\mathcal{L}(19) = 9 > \sqrt{60} = \sqrt{|H|}$, we have that $\chi|_H$ is reducible by Corollary 2.3, except possibly for the stated exceptions of q. □

Lemma 4.3. *Let $H \cong A_4$ with $q = p \equiv \pm 3, 5, \pm 11, \pm 13, \pm 19 \pmod{40}$ or $H \cong S_4$ with $q = p \equiv \pm 1 \pmod 8$, and let $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$. Then $\chi|_H$ is reducible, except possibly when $q \leq 7$.*

Proof. Note that $|H| \leq 24$ and that $q \geq 11$ unless q is 3, 5, or 7. Hence, since $\mathcal{L}(11) = 5 > \sqrt{24} \geq \sqrt{|H|}$, arguing as in the proof of Lemma 4.2, we see that $\chi|_H$ is reducible for all $\chi \in \mathrm{Irr}(G)$, except possibly in the case of the stated exceptions. □

Lemma 4.4. *Let $H \cong D_{q\pm 1}$ with $q \geq 11$, and let $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$. Then $\chi|_H$ is reducible.*

Proof. Note that $|H| = 2(q \pm 1)$. We claim that $\mathcal{L}(q) \geq \sqrt{2(q+1)} > \sqrt{2(q-1)}$ for all $q \geq 11$, which will prove the statement by Corollary 2.3. Since the second inequality clearly holds, we will work with the first. This is equivalent to solving the inequality $\frac{1}{4}(x-1)^2 \geq 2(x+1)$, and hence to solving $x^2 - 10x - 7 \geq 0$ or $(x-5)^2 - 32 \geq 0$. Since $x^2 - 10x - 7$ is increasing for $x > 5$, we see that $\mathcal{L}(x) \geq \sqrt{2(x+1)}$ whenever $x \geq 5 + 4\sqrt{2}$, which is satisfied by $x \geq 11$. □

Lemma 4.5. *Let $H \cong \Omega_3(q_0)$, where $q = q_0^r$ and r is an odd prime, or let $H \cong \mathrm{SO}_3(q_0)$, where $q = q_0^2$. Then $\chi|_H$ is reducible for every $\chi \in \mathrm{Irr}(G)$ such that $\chi(1) \neq 1$.*

Proof. First consider the case that $q_0 = 3$. If $q \geq 27$, then $\mathcal{L}(q) \geq 13 > \sqrt{24} \geq \sqrt{|H|}$. If $q = 9$, then $\mathcal{L}(9) = 4$, but we have $H \cong \mathrm{SO}_3(3)$, which is isomorphic to the symmetric group S_4. The character degrees of S_4 are $\{1, 2, 3\}$, so the claim holds in this case.

Hence we may assume that $q_0 \geq 5$. We have $|H| \leq q_0(q_0^2 - 1)$ and $\frac{1}{2}(q-1) \geq \frac{1}{2}(q_0^2 - 1)$, so by Corollary 2.3, it suffices to show that

$$\tfrac{1}{2}(q_0^2 - 1) \geq \sqrt{q_0(q_0^2 - 1)}.$$

We will do this by showing that the quotient $(x^2-1)/(2\sqrt{x(x^2-1)})$ is at least 1 for $x \geq 5$. We have

$$\left(\frac{(x^2-1)}{2\sqrt{x(x^2-1)}}\right)^2 = \frac{1}{4}\left(x - \frac{1}{x}\right) \geq \frac{1}{4}(x-1) \geq 1$$

for all such x, completing the proof. □

Lemma 4.6. *Let $H \cong E_q : \left(\frac{1}{2}(q-1)\right)$. Then $\chi|_H$ is reducible for every $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$, unless $q \equiv 3 \pmod 4$ and $\chi(1) = \frac{1}{2}(q-1)$.*

Proof. Since E_q is normal and abelian in H and $[H : E_q] = \frac{1}{2}(q-1)$, Theorem 2.4 implies $b(H) \leq \frac{1}{2}(q-1)$. Hence by Lemma 2.2, all irreducible nonlinear characters χ of G restrict reducibly to H, except possibly if $\chi(1) = \frac{1}{2}(q-1)$. □

5. Restrictions from $G = \Omega_5(q) \cong \mathrm{PSp}_4(q)$

Throughout this section, let q be a power of an odd prime p and let G be the group $\mathrm{PSp}_4(q)$, which is isomorphic to $\Omega_5(q)$. In this section, we prove the following:

Theorem 5.1. *Let $G = \Omega_5(q)$ with $q \geq 7$ odd. Let H be a maximal subgroup of G and $\chi \in \mathrm{Irr}(G)$ such that $\chi(1) \neq 1$ and $\chi|_H$ is irreducible. Then one of the following holds*:

- *H is isomorphic to $\Omega_4^{\pm}(q).2$ or a maximal parabolic subgroup of G.*
- *H is isomorphic to $\mathrm{SO}_5(q^{1/2})$.*

maximal $H \cong$	condition on q	treated in Lemma
A_6	$q \neq 7$ and $q = p \equiv \pm 5 \pmod{12}$	5.2
A_7	$q = 7$	5.2
S_6	$q = p \equiv \pm 1 \pmod{12}$	5.2
$\mathrm{PSL}_2(q)$	$q \geq 7$ and $p \geq 5$	5.3
$2^4 : A_5$	$p = q \equiv \pm 3 \pmod{8}$	5.4
$2^4 : S_5$	$p = q \equiv \pm 1 \pmod{8}$	5.4
$\mathrm{PSp}_4(q_0)$,	$q = q_0^r$ and r is an odd prime	5.5
$\left(\frac{1}{2}(q \pm 1) \times \mathrm{PSL}_2(q)\right).2^2$	$q \geq 5$	5.6

Table 1

We note that by [Bray et al. 2013], the groups excluded by the first item of Theorem 5.1 are the groups in Aschbacher class $\mathcal{C}_1$ but not isomorphic to $\left(\frac{1}{2}(q \pm 1) \times \mathrm{PSL}_2(q)\right).2^2$. We also remark that the exceptions listed in Theorem 5.1 are beyond the scope of this work, and that the cases $q = 3, 5$ are addressed in Section 8. By [Bray et al. 2013], to prove Theorem 5.1, we must show that $\chi|_H$ is reducible for $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$ and H isomorphic to one of the groups shown in Table 1.

By the main theorem of [Landazuri and Seitz 1974], we have a lower bound on the degree of the nonlinear irreducible characters for G given by the function

$$\mathcal{L}(q) := \tfrac{1}{2}(q^2 - 1).$$

Note that the continuous function $\mathcal{L} : \mathbb{R} \to \mathbb{R}$ given by $\mathcal{L}(x) = \frac{1}{2}(x^2 - 1)$ is everywhere differentiable and $\mathcal{L}'(x) = x$, which we know to be greater than zero on the interval $(0, \infty)$. Hence we see that $\mathcal{L}$ is increasing on $(0, \infty)$.

As in the previous section, our main strategy is to determine an upper bound for $b(H)$ and to show that $\mathcal{L}(q)$ is larger than this bound, implying that the nontrivial characters of $G \cong \mathrm{PSp}_4(q)$ restrict reducibly to H.

Lemma 5.2. *Let $H \cong A_6$ with $q = p \equiv \pm 5 \pmod{12}$ and $q \neq 7$, $H \cong A_7$ with $q = 7$, or $H \cong S_6$ with $q = p \equiv \pm 1 \pmod{12}$. Then $\chi|_H$ is reducible for every $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$.*

Proof. First, if $q \neq 5, 7$, note that since $\mathcal{L}$ is increasing, $\mathcal{L}(q) \geq \mathcal{L}(11) = 60 > \sqrt{720} \geq \sqrt{|H|}$. Hence in these cases, the statement follows from Corollary 2.3.

When $q = 5$, we have $H \cong A_6$ and $\mathcal{L}(q) = 12$. However, the largest irreducible character degree of A_6, as seen in the ATLAS and the GAP character table library [Conway et al. 1985; Breuer 2013], is 10. When $q = 7$, we can see using the character tables for A_7 and G in GAP [Breuer 2013] that none of the nonlinear

irreducible character degrees match. Hence in any case, $\chi|_H$ is reducible for any $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$. □

Lemma 5.3. *Let q be a power of a prime p with $p \geq 5$ and $q \geq 7$, and let $H \cong \mathrm{PSL}_2(q)$. Then $\chi|_H$ is reducible for every $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$.*

Proof. Set $u(x) = \sqrt{\frac{1}{2}x(x^2-1)}$, so $u(q) = \sqrt{|H|}$. We claim that $\mathcal{L}(x) \geq u(x)$ for the relevant values of x, implying the statement by Corollary 2.3. Note that

$$\left(\frac{\mathcal{L}(x)}{u(x)}\right)^2 = \left(\frac{(x^2-1)}{2\sqrt{(1/2)x(x^2-1)}}\right)^2 = \frac{1}{2}\left(x - \frac{1}{x}\right) \geq \frac{1}{2}(x-1) \geq 1$$

for all $x \geq 3$, which proves the claim. □

Lemma 5.4. *Suppose $q \geq 7$. Let $q = p \equiv \pm 1 \pmod 8$ and $H \cong 2^4 : S_5$, or let $p = q \equiv \pm 3 \pmod 8$ and $H \cong 2^4 : A_5$. Then $\chi|_H$ is reducible for every $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$.*

Proof. Since $\mathcal{L}$ is increasing, we see $\mathcal{L}(q) \geq \mathcal{L}(11) = 60$ unless $q = 7$. Since $\sqrt{|H|} \leq \sqrt{2^4 \cdot 120} < 60$, we see that the nontrivial characters of $\mathrm{PSp}_4(q)$ restrict reducibly in this case by Corollary 2.3. From the character tables available in GAP [Breuer 2013] for $2^4 : S_5$ and $\mathrm{PSp}_4(7)$, we see further that when $q = 7$, there are no nontrivial irreducible character degrees for G that are also degrees for H. □

Lemma 5.5. *Let $H \cong \mathrm{PSp}_4(q_0)$, where $q = q_0^r$ and r is an odd prime. Then $\chi|_H$ is reducible for every $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$.*

Proof. In this case, $|H| = \frac{1}{2}q_0^4(q_0^2-1)(q_0^4-1)$. We define real-valued functions l and u by $l(x) = \frac{1}{2}(x^{2r}-1)$ and $u(x) = \left(\frac{1}{2}x^4(x^2-1)(x^4-1)\right)^{1/2}$. We will show that $l(x) > u(x)$ whenever $r \geq 3$ and $x \geq 3$, which will establish the statement by Corollary 2.3.

Indeed, notice that for $x > 1$,

$$u(x)^2 = \tfrac{1}{2}x^4(x^2-1)(x^4-1) < x^4(x^2-1)(x^4-1) = x^{10} - (x^8 + x^6 - x^4) < x^{10},$$

where the last inequality follows from the fact that $x^8 + x^6 - x^4 > 0$ for $x > 1$. Further, for $r \geq 3$, we have $l(x) \geq \frac{1}{2}(x^6-1)$, which is larger than x^5 for $x \geq 3$. This shows that $l(x)^2 > u(x)^2$ for $x \geq 3$ and $r \geq 3$, which completes the claim. □

In the final case, $q \geq 5$ and $H \cong \left(\frac{1}{2}(q \pm 1) \times \mathrm{PSL}_2(q)\right).2^2$. Then H is an extension of $K := \frac{1}{2}(q \pm 1) \times \mathrm{PSL}_2(q)$ by the direct product $C_2 \times C_2$. That is, $H/K \cong C_2 \times C_2$.

Lemma 5.6. *Let $q \geq 7$ and $H \cong \left(\frac{1}{2}(q \pm 1) \times \mathrm{PSL}_2(q)\right).2^2$. Then $\chi|_H$ is reducible for every $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$.*

Proof. First, recall that for abelian groups, the degree of every irreducible character is exactly 1. Let C be the cyclic group of size $\frac{1}{2}(q \pm 1)$. Thus for all $\lambda \in \mathrm{Irr}(C)$, we have $\lambda(1) = 1$. Since K is the direct product of C with $\mathrm{PSL}_2(q)$, we know that all elements of $\mathrm{Irr}(K)$ are of the form $\lambda \times \varphi$ by Theorem 2.7, where $\lambda \in \mathrm{Irr}(C)$ and $\varphi \in \mathrm{Irr}(\mathrm{PSL}_2(q))$. In particular, the degree of each of these characters is simply given by $\varphi(1)$.

Further, since $[H : K] = 4$ and $K \lhd H$, we see using Theorem 2.6 that if $\chi \in \mathrm{Irr}(H)$ and $\chi|_H$ contains $\theta \in \mathrm{Irr}(K)$ as a constituent, then $\chi(1)/\theta(1)$ divides 4. However, since C is abelian and we know the maximal degree given by the generic character table of $\mathrm{PSL}_2(q)$ (see, for example, [Geck et al. 1996]) is $q + 1$, we have an upper bound for $b(H)$ given by $b(H) \leq 4(q + 1)$.

Letting $u(x) = 4(x + 1)$, notice that

$$\frac{\mathcal{L}(x)}{u(x)} = \frac{(x^2 - 1)}{8(x + 1)} = \frac{1}{8}(x - 1) > 1$$

whenever $x > 9$, and hence $\mathcal{L}(x) > u(x)$, proving the statement for $q > 9$ by Lemma 2.2. Further, using the GAP character table library [Breuer 2013], we see that $\mathrm{PSp}_4(9)$ has smallest nontrivial degree $41 > u(9)$, finishing the case $q = 9$.

Finally, consider the case $q = 7$. Note that the character degrees of K must be in the set $\{d, 2d, 4d\}$, where d ranges over the irreducible character degrees of $\mathrm{PSL}_2(7)$. Utilizing GAP, we see that none of these numbers occur as character degrees of G larger than 1, completing the proof. □

6. Restrictions from $G = \mathrm{Sp}_6(q)$

In this section, let G be the symplectic group $\mathrm{Sp}_6(q)$, where q is a power of an odd prime p. We prove the following:

Theorem 6.1. *Let $G = \mathrm{Sp}_6(q)$, where $q \geq 5$ is a power of an odd prime, and let $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$. Suppose $H \leq G$ is a maximal subgroup such that the restriction $\chi|_H$ is irreducible. Then one of the following holds*:

- *H is isomorphic to $\mathrm{Sp}_2(q) \times \mathrm{Sp}_4(q)$ or a maximal parabolic subgroup.*
- *$q = 5$, $H \cong 2^{\cdot}J_2$, and $\chi(1) = 63$.*
- *$q = 5$, $H \cong \mathrm{GL}_3(5).2$, and $\chi(1) = 62$.*
- *$H \cong \mathrm{Sp}_2(q^3) : 3$ and $\chi(1) = \frac{1}{2}(q^3 \pm 1)$.*
- *$H \cong \mathrm{Sp}_6(q_0).2$, where $q = q_0^2$, and $q_0 = 5$ or $\chi(1) = \frac{1}{2}(q^3 \pm 1)$.*

As in the case of Theorem 5.1, the groups excluded by the first item of Theorem 6.1 are those found in Aschbacher class $\mathcal{C}_1$. We remark that the case $q = 3$ will be considered in Section 8 and that addressing the exceptions listed in Theorem 6.1, aside from $2^{\cdot}J_2$ addressed in Lemmas 6.3 and 6.4 below, will require methods

maximal $H \cong$	condition on q	restriction behavior	treated in Lemma
$2^{\cdot}A_5$	$q = p \equiv \pm 3, \pm 11, \pm 13, \pm 19 \pmod{40}$	always red.	6.2
$2^{\cdot}S_5^-$	$q = p \equiv \pm 1 \pmod 8$	always red.	6.2
$2^{\cdot}\mathrm{PSL}_2(7)^{\cdot}2^+$	$q = p \equiv \pm 1 \pmod{16}$	always red.	6.2
$2^{\cdot}\mathrm{PSL}_2(7)$	$q = p \equiv \pm 7 \pmod{16}$, $q \neq 7$	always red.	6.2
$2^{\cdot}\mathrm{PSL}_2(7)$	$q = p^2$, $p \equiv \pm 3, \pm 5 \pmod{16}$	always red.	6.2
$2^{\cdot}\mathrm{PSL}_2(13)$	$q = p \equiv \pm 1, \pm 3, \pm 4 \pmod{13}$	always red.	6.2
$2^{\cdot}\mathrm{PSL}_2(13)$	$q = p^2$, $p \equiv \pm 2, \pm 5, \pm 6 \pmod{13}$	always red.	6.2
$2^{\cdot}A_7$	$q = 9$	always red.	6.2
$2\times\mathrm{PSU}_3(3)$	$q = p \equiv \pm 7, \pm 17, \pm 19, \pm 29 \pmod{60}$	always red.	6.2
$(2\times\mathrm{PSU}_3(3)).2$	$q = p \equiv \pm 1 \pmod{12}$	always red.	6.2
$2^{\cdot}J_2$	$q = p \equiv \pm 1 \pmod 5$	always red.	6.3
$2^{\cdot}J_2$	$q = 5$	red. unless $\chi(1) = 63$	6.3, 6.4
$2^{\cdot}J_2$	$q = p^2$, $p \equiv \pm 2 \pmod 5$	always red.	6.3
$2^{\cdot}\mathrm{PSL}_2(q)$	$p \geq 7$	always red.	6.5

Table 2

beyond the scope of this article. The remainder of this section is devoted to proving Theorem 6.1.

By the proof of [Tiep and Zalesskii 1996, Theorem 5.2], a lower bound for the nontrivial character degrees of G is

$$\mathcal{L}(q) := \tfrac{1}{2}(q^3 - 1).$$

As in the previous sections, our new lower bound $\mathcal{L}$ is an increasing function for $x > 1$.

We will first investigate the character restrictions to the subgroups listed in Table 2, which according to [Bray et al. 2013] are the maximal subgroups in the Aschbacher class $\mathcal{S}$. Recall here that we are assuming $q \geq 5$.

We may treat the first several cases using the strategies from the previous sections.

Lemma 6.2. *Let $q \geq 5$ and let H be one of the maximal subgroups listed above, aside from $2^{\cdot}J_2$ or $2^{\cdot}\mathrm{PSL}_2(q)$. Then $\chi|_H$ is reducible for every $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$.*

Proof. In each case, $\mathcal{L}(q) > \sqrt{|H|}$, which can be seen using the same arguments as in the previous sections. Hence the statement follows from Corollary 2.3. □

Now, we consider the second Janko group, J_2, which is one of the sporadic finite simple groups. Also called the Janko–Hall group, J_2 was one of the first simple

sporadic groups discovered and its order is $|J_2| = 604800$. The group $H \cong 2^{\cdot}J_2$ is the so-called Schur cover or universal covering group of J_2.

Lemma 6.3. *Let $H \cong 2^{\cdot}J_2$, where $p = q \equiv \pm 1 \pmod 5$, $q = 5$, or $q = p^2$ and $p \equiv \pm 2 \pmod 5$. Let $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$. Then $\chi|_H$ is reducible, except in the case $q = 5$ and $\chi(1) = 63$.*

Proof. Note that $b(H) < \sqrt{2 \cdot 604800} \approx 1099.8$. Since $\mathcal{L}$ is increasing and $\mathcal{L}(19) > \sqrt{2 \cdot 604800}$, the statement follows by Corollary 2.3 as long as $q \neq 5, 9$, or 11. Further, the character table of H is available in the ATLAS or the GAP character table library [Conway et al. 1985; Breuer 2013], from which we see that $b(H) = 448$. Since $\mathcal{L}(11) = 665$, we are finished in this case.

Now, let $q = 9$. Using the character degrees for G available from [Lübeck 2007], we see that the only degrees below the maximal degree of $2^{\cdot}J_2$ are 364 and 365, but neither of these appear in the list of degrees from $2^{\cdot}J_2$, so they must restrict reducibly.

When $q = 5$, again using [Lübeck 2007], we must consider characters of G of degrees 62 and 63. However, only 63 occurs as a character degree for $2^{\cdot}J_2$, completing the proof. □

Lemma 6.4. *The irreducible characters of $G = \mathrm{Sp}_6(5)$ with degree 63 restrict irreducibly to $H \cong 2^{\cdot}J_2$.*

Proof. From [Lübeck 2007], we see there are two characters of G of degree 63. Further, from the character table of $\mathrm{PSp}_6(5)$ available in GAP, we see that $G/Z(G)$ also has two irreducible characters of degree 63. That is, the two irreducible characters of G of degree 63 are trivial on the center. We also see, using the character tables available in GAP, that the character of H of degree 63 is trivial on the center. Hence the characters of degree 63 of G and H can be considered as characters of $\mathrm{PSp}_6(5)$ and J_2, respectively.

Implementing the algorithm described in Section 3, we see that the characters χ_2 and χ_3 of degree 63 of $\mathrm{PSp}_6(5)$ restrict irreducibly to the character χ_7 of J_2. Hence the inflations to G will restrict irreducibly to H as well. □

Lemma 6.5. *Let $p \geq 7$ and let $H \cong 2^{\cdot}\mathrm{PSL}_2(q)$. Then $\chi|_H$ is reducible for every irreducible character χ of $G = \mathrm{Sp}_6(q)$ with $\chi(1) \neq 1$.*

Proof. From the character table for $H \cong \mathrm{SL}_2(q)$, we know $b(H) = q + 1$. Since $x + 1 < \frac{1}{2}(x^3 - 1) = \mathcal{L}(x)$ whenever $x > 2$, the statement follows from Lemma 2.2. □

We now consider the maximal subgroups of $G = \mathrm{Sp}_6(q)$ from Aschbacher class $\mathcal{C}$ given in Table 3. Recall here that $q \geq 5$.

We remark that these are all of the maximal subgroups in $\mathcal{C}$, with the exception of those in $\mathcal{C}_1$ and $\mathcal{C}_8$. Addressing these omitted groups and those for which we only attain partial results will require methods beyond the scope of this article.

Aschbacher class	maximal $H \cong$	condition on q	restriction behavior	treated in Lemma
$\mathcal{C}_2$	$\mathrm{Sp}_2(q) \wr S_3$		always red.	6.6
$\mathcal{C}_2$	$\mathrm{GL}_3(q).2$		red. unless $q = 5$, $\chi(1) = 62$	6.7
$\mathcal{C}_3$	$\mathrm{Sp}_2(q^3) : 3$		partial results	6.8
$\mathcal{C}_3$	$\mathrm{GU}_3(q).2$		always red.	6.9
$\mathcal{C}_4$	$\mathrm{Sp}_2(q) \circ \mathrm{GO}_3(q)$		always red.	6.10
$\mathcal{C}_5$	$\mathrm{Sp}_6(q_0)$	$q = q_0^r$, r odd prime	always red.	6.11
$\mathcal{C}_5$	$\mathrm{Sp}_6(q_0).2$	$q = q_0^2$	partial results	6.12

Table 3

For the remainder of the section recall that $q \geq 5$ and define d_i to be the i-th irreducible character degree of $\mathrm{Sp}_6(q)$ as obtained from the list generated by [Lübeck 2007]. In particular, we have

$$d_2 := \tfrac{1}{2}(q^3 - 1), \quad d_4 := \tfrac{1}{2}q(q-1)(q^3-1),$$
$$d_3 := \tfrac{1}{2}(q^3 + 1), \quad d_5 := \tfrac{1}{2}(q-1)(q^2+q+1)(q^2-q+1).$$

Certainly $d_2 < d_3$ and $d_4 < d_5$, since $q^3 - 1 < q^3 + 1$ and $q^2 - q < q^2 - q + 1$. Further, using the upper bound $\sum_{i=0}^{n-1} |a_i|$ for the positive roots of a polynomial $x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$, with real coefficients, we see that for $q > 5$, $2(d_4 - d_3) = q^5 - q^4 - q^3 - q^2 + q - 1$ must be positive. Since $d_4 > d_3$ when $q = 5$ by checking directly, we therefore see that $d_5 > d_4 > d_3 > d_2$ for all $q \geq 5$. Further, similar arguments using the polynomials in q in the list generated by [Lübeck 2007] yield that $d_i \geq d_5$ for each $i \geq 5$.

Lemma 6.6. *If $q \geq 5$ and $H \cong \mathrm{Sp}_2(q) \wr S_3$, then $\chi|_H$ is reducible for each $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$.*

Proof. Recall that H can be viewed as the semidirect product $\mathrm{Sp}_2(q)^3 : S_3$. Theorem 2.7, combined with the fact that q is odd and $\mathrm{Sp}_2(q) \cong \mathrm{SL}_2(q)$, yields that the irreducible characters of $\mathrm{Sp}_2(q)^3$ have degree at most $(q+1)^3$. Then, Theorem 2.6 implies $b(H) \leq 6(q+1)^3$.

Now, note that $b(H) < d_4$ for $q > 5$ and that $b(H) < d_5$ for $q \geq 5$. Further, when $q = 5$, we have $d_4 = 1240$, which has 31 as a prime factor. Since 31 does not divide $|H|$ in this case, we see d_4 cannot be a character degree for H. Hence it suffices to show neither d_2 nor d_3 can be a character degree for H when $q \geq 5$.

Assume by way of contradiction that d_2, respectively d_3, is the degree of some irreducible character of H. Note that this means $d := d_2$, respectively d_3, must divide $|H| = 6q^3(q-1)^3(q+1)^3$. Letting $\epsilon = 1$ in the case $d = d_2$ and $\epsilon = -1$ in the case $d = d_3$, recall that $d = \frac{1}{2}(q-\epsilon)(q^2+\epsilon q+1)$. In particular, we see that any prime dividing $q^2+\epsilon q+1$, must also divide $|H|$. Applying Lemma 2.8, it follows that d must be a product of powers of 2 and 3. Since $q \geq 5$ is odd, this means that the odd number $q^2+\epsilon q+1$ is of the form 3^r with $r > 3$. However, by Lemma 2.8(c), this means that 3^5 divides d but that the largest power of 3 dividing $|H|$ is 3^4, a contradiction. □

Lemma 6.7. *If $q \geq 5$ and $H \cong \mathrm{GL}_3(q).2$, then $\chi|_H$ is reducible for each $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$, except possibly if $q = 5$ and $\chi(1) = 62$.*

Proof. From the generic character table available for $\mathrm{GL}_3(q)$ in CHEVIE [Geck et al. 1996], we see that the irreducible character degrees for $\mathrm{GL}_3(q)$ are

$$\{1,\ q(q+1),\ q^3,\ q^2+q+1,\ q(q^2+q+1),\ (q\pm1)(q^2+q+1),\ (q-1)^2(q+1)\} \quad (1)$$

and that the largest of these is $(q+1)(q^2+q+1)$. Then $b(H) \leq 2(q+1)(q^2+q+1)$ by Theorem 2.6.

Recall that $d_4 = \frac{1}{2}q(q^2+q+1)(q-1)^2$ and that for $i \geq 4$, we have $d_i \geq d_4$. Notice also that d_4 is an increasing function and $d_4 > b(H)$ whenever $q \geq 5$. This shows that every irreducible character of degree larger than d_3 must restrict reducibly to H, by Lemma 2.2.

Now, applying Theorem 2.6, we see that every member of $\mathrm{Irr}(H)$ has degree of the form m or $2m$ for some m in the set (1). Arguing as in Lemma 6.6, using Lemma 2.8, we see that for each member $m \neq 1$ in this list, there is some odd divisor of $d_3 = \frac{1}{2}(q+1)(q^2-q+1)$ that does not divide m, and hence no character of degree d_3 restricts irreducibly to H. The same argument yields the same statement for $d_2 = \frac{1}{2}(q-1)(q^2+q+1)$, except possibly if m is one of the numbers in the list with divisor q^2+q+1. But since $q \geq 5$, we further have $\frac{1}{2}(q-1)$ cannot be in the set $\{1,\ 2,\ q,\ 2q,\ q\pm1,\ 2(q\pm1)\}$, and hence d_2 also cannot coincide with any character degree of H, unless $q = 5$ and $d_2 = 62 = 2(q^2+q+1)$. □

We remark that the character degree 62 does not appear for the simple group $\mathrm{PSp}_6(5)$, so the unsolved exception in Lemma 6.7 is irrelevant for the problem of determining irreducible restrictions from the simple group $G/Z(G)$.

Lemma 6.8. *If $q \geq 5$ is odd and $H \cong \mathrm{Sp}_2(q^3):3$, then $\chi|_H$ is reducible for each $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$, with the possible exception of those with degree equal to d_2 or d_3.*

Proof. Recall from above that $\mathrm{Sp}_2(q) \cong \mathrm{SL}_2(q)$ and whenever q is odd, the maximum degree is $q+1$. A quick application of Theorem 2.6 gives us that the characters

of H have degree at most $3(q^3+1)$. Since $q>3$, it is easy to see that the inequality

$$3(q^3+1) < \tfrac{1}{2}q(q^2+q+1)(q-1)^2$$

is true. Since the degree d_4 is an increasing function and for $i \geq 4$ we have $d_i \geq d_4$, we get that the characters of $\mathrm{Sp}_6(q)$ with degrees greater than or equal to d_4 will restrict reducibly to H. □

Lemma 6.9. *Let $q \geq 5$ and $H \cong \mathrm{GU}_3(q).2$. Then $\chi|_H$ is reducible for each $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$.*

Proof. The set of irreducible character degrees for $\mathrm{GU}_3(q)$ is

$$\{1,\ q(q-1),\ q^3,\ q^2-q+1,\ q(q^2-q+1),\ (q\pm1)(q^2-q+1),\ (q+1)^2(q-1)\}, \quad (2)$$

which can be seen from the generic character table for $\mathrm{GU}_3(q)$ available in CHEVIE [Geck et al. 1996]. Then the maximum character degree of $\mathrm{GU}_3(q)$ is $(q+1)^2(q-1)$, so Theorem 2.6 implies that $b(H)$ is at most $2(q+1)^2(q-1)$.

Since the inequality

$$2(q+1)^2(q-1) < d_4 = \tfrac{1}{2}q(q^2+q+1)(q-1)^2$$

is true for $q \geq 3$, we see by Lemma 2.2 that the statement is true, except possibly for characters of G with degree d_2 or d_3.

Arguing exactly as in Lemma 6.7 with the roles of d_2 and d_3 reversed, we see that no character of degree d_2 or d_3 may restrict irreducibly to H. Note that in this case, we do not need to make an exception like that in Lemma 6.7, since when $q \geq 5$, $\frac{1}{2}(q+1)$ cannot be in the set $\{1,\ 2,\ q,\ 2q,\ q\pm1,\ 2(q\pm1)\}$. □

Lemma 6.10. *If $q \geq 5$ and $H \cong \mathrm{Sp}_2(q) \circ \mathrm{GO}_3(q)$, then $\chi|_H$ is reducible for each $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$.*

Proof. Let $q \geq 5$ be odd and note that $\mathrm{Sp}_2(q) \cong \mathrm{SL}_2(q)$ and $\Omega_3(q) \cong \mathrm{PSL}_2(q)$ and that the largest irreducible character degree of either of these groups is at most $q+1$. Further, note that $\mathrm{GO}_3(q)$ contains a normal subgroup of index 4 isomorphic to the latter group. Using this information and Theorem 2.6, we see that $b(\mathrm{GO}_3(q)) \leq 4(q+1)$.

Now, recalling that the irreducible characters of H are products of the irreducible characters of the groups $\mathrm{GO}_3(q)$ and $\mathrm{Sp}_2(q)$ since it is a central product, we obtain an upper bound on the character degrees of $H \cong \mathrm{Sp}_2(q) \circ \mathrm{GO}_3(q)$ given by $b(H) \leq 4(q+1)^2$.

Solving the inequality computationally, we get that $4(q+1)^2 < \mathcal{L}(q)$ whenever $q \geq 11$, proving the statement for $q \geq 11$ by Lemma 2.2. Further, note that if $q \geq 3$, the inequality

$$4(q+1)^2 < \tfrac{1}{2}q(q^2+q+1)(q-1)^2$$

is satisfied, so any irreducible character of G of degree at least d_4 must restrict reducibly to H.

Let d be d_2 or d_3. Note that the character degrees for $\mathrm{Sp}_2(q)$ and $\mathrm{GO}_3(q)$ are 1, q, $q-1$, and $q+1$, up to multiplying or dividing by powers of 2. Since H is a central product, its irreducible character degrees are composed of products of two of these values, up to multiplying or dividing by powers of 2. Using Lemma 2.8 and arguing exactly as in Lemma 6.6, we again see that either some prime $\ell \geq 5$ or some power of 3 divides d but not any of the irreducible character degrees of H. Hence we see that d cannot be the degree of any irreducible character of H, and each $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$ must therefore restrict reducibly to H. □

We next turn our attention to the subgroups of the form $\mathrm{Sp}_6(q_0).(2, r)$, where $q = q_0^r$ and r is prime. Recall that $|\mathrm{Sp}_6(q_0)| = q_0^9 \prod_{i=1}^3 (q_0^{2i} - 1)$. We begin with the case that r is odd.

Lemma 6.11. *Let q_0 be a prime power such that $q = q_0^r$, where r is an odd prime. Let $H \cong \mathrm{Sp}_6(q_0)$. Then $\chi|_H$ is reducible for each $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$.*

Proof. From the list available at [Lübeck 2007], we see that the largest irreducible character degree for $\mathrm{Sp}_6(q_0)$ is at most $(q_0^2+1)(q_0^2+q_0+1)(q_0^2-q_0+1)(q_0+1)^3$, which is smaller than $\mathcal{L}(q) = \frac{1}{2}(q_0^{3r} - 1)$ when $r \geq 5$ and $q_0 \geq 3$. Then by Corollary 2.3, we are done in the case $r > 3$.

Now let $r = 3$ and notice that

$$d_4 = \tfrac{1}{2}(q^5 - q^4 - q^2 + q) = \tfrac{1}{2}(q_0^{15} - q_0^{12} - q_0^6 + q_0^3).$$

Then, we see computationally that $b(H) < d_4$ for all $q_0 \geq 3$. Hence it suffices to show that the character degrees d_2 and d_3 for G do not appear as irreducible character degrees for H.

Notice that

$$d_2 = \tfrac{1}{2}(q_0^9 - 1) = \tfrac{1}{2}(q_0 - 1)(q_0^2 + q_0 + 1)(q_0^6 + q_0^3 + 1),$$
$$d_3 = \tfrac{1}{2}(q_0^9 + 1) = \tfrac{1}{2}(q_0 + 1)(q_0^2 - q_0 + 1)(q_0^6 - q_0^3 + 1).$$

Further, observing Lübeck's list of character degrees, see [Lübeck 2007], we see that, up to dividing by 2, every degree for H is a product of the cyclotomic polynomials dividing $q_0(q_0^6 - 1)$, which are those listed in Lemma 2.8 in terms of q_0. Using Lemma 2.8 applied to q_0^3 and the arguments used before, we see that $(q_0^6 + q_0^3 + 1)$ and $(q_0^6 - q_0^3 + 1)$ have odd divisors that do not divide $q_0(q_0^6 - 1) = q_0(q_0^3 - 1)(q_0^3 + 1)$, and therefore the same is true for d_2 and d_3. Hence these do not appear as character degrees for $\mathrm{Sp}_6(q_0)$, completing the proof. □

Finally, we address the more delicate case that $H \cong \mathrm{Sp}_6(q_0).2$, where $q = q_0^2$. In this case, we only achieve partial results.

Lemma 6.12. *Let $q = q_0^2$ be odd such that $q_0 \geq 7$ and let $H \cong \mathrm{Sp}_6(q_0).2$. Then $\chi|_H$ is reducible for each $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$, except possibly those with degree d_2 or d_3.*

Proof. First, consider a character $\chi \in \mathrm{Irr}(\mathrm{Sp}_6(q))$ with degree greater than or equal to

$$d_4 = \tfrac{1}{2}q(q^2+q+1)(q-1)^2 = \tfrac{1}{2}q_0^2(q_0^4+q_0^2+1)(q_0^2-1)^2.$$

Again, from the list available at [Lübeck 2007], we see that the largest irreducible character degree for $\mathrm{Sp}_6(q_0)$ is at most $(q_0^2+1)(q_0^2+q_0+1)(q_0^2-q_0+1)(q_0+1)^3$. Hence by Theorem 2.6, we see that

$$b(H) \leq 2(q_0^2+1)(q_0^2+q_0+1)(q_0^2-q_0+1)(q_0+1)^3.$$

When $q_0 \geq 7$, we therefore have $d_4 > b(H)$, which completes the proof by Lemma 2.2. □

7. Restrictions from $G = \Omega_7(q)$

In this section, let G be the group $\Omega_7(q)$, where q is the power of an odd prime p. We prove the following:

Theorem 7.1. *Let $G = \Omega_7(q)$, where $q \geq 5$ is a power of an odd prime, and let $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$. Suppose $H \leq G$ is a maximal subgroup such that the restriction $\chi|_H$ is irreducible. Then one of the following holds*:

- *H is isomorphic to $\Omega_6^{\pm}(q).2$, $(\Omega_2^{\pm}(q) \times \Omega_5(q)).2^2$, or a maximal parabolic subgroup*;
- *H is isomorphic to $\mathrm{SO}_7(q^{1/2})$; or*
- *H is isomorphic to the exceptional group of Lie type $G_2(q)$.*

The groups listed in the first item of Theorem 7.1 are the maximal subgroups in Aschbacher class $\mathcal{C}_1$ other than $(\Omega_3(q) \times \Omega_4^{\pm}(q)).2^2$, by [Bray et al. 2013]. As before, addressing the exceptions listed in Theorem 7.1 is beyond the scope of this article. We further remark that the case $H \cong G_2(q) \leq \Omega_7(q)$ is pointed out in [Seitz 1990] as one of very few embeddings of groups of Lie type into finite classical groups defined in the same characteristic that produce examples of irreducible restrictions, and is the topic of a forthcoming paper by the faculty author. The reader may also note that the exceptions listed are similar to those that must be carefully treated in [Schaeffer Fry 2013] in the case $p = 2$. The remainder of this section is devoted to proving Theorem 7.1.

Note that the smallest nontrivial irreducible character degree of G is $\mathcal{L}(q) = q^4+q^2+1$, see [Tiep and Zalesskii 1996, Theorem 1.1], and that the real-valued function $\mathcal{L}(x)$ is an increasing function for positive x. Our methods in this section will largely be similar to those in previous sections.

Aschbacher class	maximal $H \cong$	condition on q	treated in Lemma
$\mathcal{S}$	$\Omega_7(2)$	$q = p$	7.2
$\mathcal{C}_2$	$2^6 : A_7$	$p = q \equiv \pm 3 \pmod 8$	7.3
$\mathcal{C}_2$	$2^6 : S_7$	$p = q \equiv \pm 1 \pmod 8$	7.3
$\mathcal{C}_5$	$\Omega_7(q_0)$	$q = q_0^r$, r odd prime	7.4
$\mathcal{C}_1$	$(\Omega_3(q) \times \Omega_4^{\pm}(q)).2^2$		7.5

Table 4

From [Bray et al. 2013], we see that Table 4 lists all maximal subgroups when $q \geq 5$, aside from those excepted in Theorem 7.1. Note that we will treat the case $q = 3$ in Section 8 below.

Lemma 7.2. *Let $H \cong \Omega_7(2)$ and let $q \geq 5$ be an odd prime. Then $\chi|_H$ is reducible for each $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$.*

Proof. Using the character table for $H \cong \Omega_7(2) \cong \mathrm{Sp}_6(2)$ available in GAP [Breuer 2013], we see that $b(H) = 512$. The statement follows since $\mathcal{L}(5) = 651 > b(H)$ and $\mathcal{L}$ is increasing. □

Lemma 7.3. *Let $q \geq 5$ be an odd prime and let H be a maximal subgroup of G isomorphic to $2^6 : A_7$, where $q = p \equiv \pm 3 \pmod 8$ or $2^6 : S_7$, where $q = p \equiv \pm 1 \pmod 8$. Then $\chi|_H$ is reducible for each $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$.*

Proof. Since C_2^6 is abelian and normal in H, we may use Theorem 2.4 to see that $b(H) \leq [H : C_2^6] \leq |S_7| = 5040$. Note that $\mathcal{L}(9) = 6643 > b(H)$, so that the statement follows for $q \geq 9$ since $\mathcal{L}$ is increasing. When $q = 5$, we may obtain the character table for G using GAP [Breuer 2013], from which we see that the character degrees of G that are less than $[H : C_2^6] = |A_7| = 2520$ do not divide 2520. Similarly, using GAP and [Lübeck 2007], we see that the only character degree of G when $q = 7$ that is less than 5040 is 2451, which does not divide 5040. Hence applying Theorem 2.4 again yields that these cannot be character degrees of H. □

Lemma 7.4. *Let $q = q_0^r$ for a power of an odd prime q_0 and an odd prime r and let $H \cong \Omega_7(q_0)$. Then $\chi|_H$ is reducible for each $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$.*

Proof. We have

$$|H| = |\Omega_7(q_0)| = q_0^9(q_0^2 - 1)(q_0^4 - 1)(q_0^6 - 1),$$
$$\mathcal{L}(q) = q^4 + q^2 + 1 = q_0^{4r} + q_0^{2r} + 1 \geq q_0^{12} + q_0^6 + 1.$$

Consider the real-valued functions

$$g(x) = x^{12} + x^6 + 1,$$
$$h(x) = \sqrt{x^9(x^2 - 1)(x^4 - 1)(x^6 - 1)}.$$

Since both functions are positive for $x > 1$ and

$$h(x)^2 = x^{21} - (x^{19} + x^{17} - x^{13} - x^{11} - x^{13} + x^9) < x^{24} < g(x)^2,$$

we see that $g(x) > h(x)$ for all $x \geq 1$, and hence $\mathcal{L}(q) > b(H)$. □

Lemma 7.5. *Let $q \geq 5$ and let $H \cong (\Omega_3(q) \times \Omega_4^{\pm}(q)).2^2$. Then $\chi|_H$ is reducible for each $\chi \in \mathrm{Irr}(G)$ with $\chi(1) \neq 1$.*

Proof. Let $N \lhd H$ be the normal subgroup $N \cong \Omega_3(q) \times \Omega_4^-(q)$ and let $\chi \in \mathrm{Irr}(H)$. From Clifford's theorem, Theorem 2.5, $\chi|_N = e(\psi_1 + \cdots + \psi_m)$ for some positive integer e and some $\psi_1, \ldots, \psi_m \in \mathrm{Irr}(N)$ such that $\psi_1(1) = \cdots = \psi_m(1)$. Then

$$\chi(1) = \chi|_N(1) = e(\psi_1(1) + \cdots + \psi_m(1)) \leq e \cdot m \cdot b(N) \leq 4b(N),$$

where the last inequality follows from Theorem 2.6.

Now, by Theorem 2.7, each character ψ_i of N is of the form $\phi_i \times \varphi_i$, where ϕ_i is an irreducible character of $\Omega_3(q) \cong \mathrm{PSL}_2(q)$ and φ_i is an irreducible character of $\Omega_4^{\pm}(q)$, which is isomorphic to $\mathrm{PSL}_2(q^2)$ in the case "$-$" and to $\mathrm{SL}_2(q) \circ \mathrm{SL}_2(q) \cong 2.(\mathrm{PSL}_2(q) \times \mathrm{PSL}_2(q))$ in the case "$+$".

Then in the case "$-$", we have

$$b(N) = b(\mathrm{PSL}_2(q))b(\mathrm{PSL}_2(q^2)) = (q+1)(q^2+1) = q^3 + q^2 + q + 1,$$

and so $\chi(1) \leq 4(q^3 + q^2 + q + 1) < \mathcal{L}(q)$, for $q \geq 5$, where the last inequality follows by analyzing the corresponding real-valued functions. This completes the proof in the case "$-$".

In the case "$+$", we have $b(N) = (q+1)^3$, using the discussion after Theorem 2.7, so that $\chi(1) \leq 4(q+1)^3 < \mathcal{L}(q)$ whenever $q \geq 7$. Now, when $q = 5$, the only nontrivial character degree of G that is at most $4(5+1)^3 = 864$ is 651, using the character table available in GAP [Breuer 2013]. However, note that 651 is not divisible by 2, and that $651 > (5+1)^3 = 216$. This yields that 651 is not a character degree of N or of H, again using Theorem 2.6, and completes the proof. □

8. Results for small values of q

Here we address the case that q is "small". That is, we consider the exceptional values of q from Theorems 4.1, 5.1, 6.1, and 7.1. We do this using GAP [Breuer 2013] and our algorithm discussed in Section 3. For $G \cong \Omega_3(q) \cong \mathrm{PSL}_2(q)$, we obtain the results for $5 \leq q \leq 11$, which are summarized in Table 5. For $G \cong \Omega_5(q) \cong \mathrm{PSp}_4(q)$, we summarize the results for $q = 3, 5$ in Table 6. The results for $G = \mathrm{PSp}_6(3)$ and $G = \Omega_7(3)$ are summarized in Tables 7 and 8, respectively.

In the tables, Maxes[i] means the i-th maximal subgroup of G found using the `Maxes` command in GAP and the labeling for the characters is as in the GAP character table for G.

value of q	maximal subgroup with irreducible restrictions	irreducible restrictions	degree
$q = 5$	Maxes[1] $\cong A_4$	χ_2, χ_3	3
$q = 7$	Maxes[1] $\cong S_4$	χ_2, χ_3	3
$q = 7$	Maxes[2] $\cong S_4$	χ_2, χ_3	3
$q = 7$	Maxes[3] $\cong 7:3$	χ_2, χ_3	3
$q = 9$	Maxes[1] $\cong A_5$	χ_3	5
$q = 9$	Maxes[2]	χ_2	5
$q = 11$	Maxes[1] $\cong A_5$	χ_2, χ_3	5
$q = 11$	Maxes[2] $\cong A_5$	χ_2, χ_3	5
$q = 11$	Maxes[3] $\cong 11:5$	χ_2, χ_3	5

Table 5. Irreducible restrictions for $\mathrm{PSL}_2(q)$ for small q.

value of q	maximal subgroup with irreducible restrictions	irreducible restrictions	degree
$q = 3$	Maxes[1]	χ_2, χ_3 χ_5, χ_6	5 10
$q = 3$	Maxes[2] $\cong A_6.2_1$	χ_2, χ_3 χ_5, χ_6	5 10
$q = 3$	Maxes[4] $\cong 3^3 : S_4'$	χ_4	6
$q = 5$	Maxes[2] $\cong 5^3 : (2 \times A_5).2$	χ_4	40
$q = 5$	Maxes[3] $\cong \mathrm{PSL}_2(25).2_2$	χ_2, χ_3	13

Table 6. Irreducible restrictions for $\mathrm{PSp}_4(q)$ for small q.

maximal subgroup with irreducible restrictions	irreducible restrictions	degree
$3^6 : \mathrm{PSL}_3(3)$	χ_2, χ_3 χ_4	13 78
$\mathrm{PSL}_2(27).3$	χ_2, χ_3 χ_4	13 78
$\mathrm{PSL}_3(3).2$	χ_2, χ_3	13
Maxes[9] $\cong \mathrm{PSL}_2(13)$	χ_2, χ_3	13
Maxes[10] $\cong \mathrm{PSL}_2(13)$	χ_2, χ_3	13

Table 7. Irreducible restrictions from $\mathrm{PSp}_6(3)$ to maximal subgroups.

maximal subgroup with irreducible restrictions	irreducible restrictions	degree
Maxes[3] $\cong \mathrm{PSL}_4(3).2$	χ_8, χ_9	260
Maxes[4] $\cong G_2(3)$	χ_2 χ_3 χ_5 χ_6 χ_{10} χ_{11}	78 91 168 182 273 546
Maxes[5]	χ_2 χ_3 χ_5 χ_6 χ_{10} χ_{11}	78 91 168 182 273 546
Maxes[6] $\cong (C_3^3.C_3^3) : \mathrm{PSL}_3(3)$	χ_2	78
Maxes[7] $\cong \mathrm{Sp}_6(2)$	χ_4	105
Maxes[8]	χ_4	105
Maxes[10] $\cong A_9.2$	χ_4	105
Maxes[11]	χ_4	105

Table 8. Irreducible restrictions for $\Omega_7(3)$.

Acknowledgments

The authors would like to thank the two referees for their careful reading of the manuscript and several suggestions and comments that greatly helped the clarity and accuracy of the paper, including comments that significantly simplified the GAP code.

The authors were supported by the MSU Denver Faculty Scholars Program and a Research at Undergraduate Institutions grant from the National Science Foundation (Award #DMS-1801156).

Albee and Roon were also supported by an MAA Undergraduate Travel Grant. They would also like to thank fellow MSU Denver undergraduate Luke Smith for productive discussions and lending his computer knowledge and hardware.

References

[Aschbacher 1984] M. Aschbacher, "On the maximal subgroups of the finite classical groups", *Invent. Math.* **76**:3 (1984), 469–514. MR Zbl

[Bray et al. 2013] J. N. Bray, D. F. Holt, and C. M. Roney-Dougal, *The maximal subgroups of the low-dimensional finite classical groups*, Lond. Math. Soc. Lect. Note Series **407**, Cambridge Univ. Press, 2013. MR Zbl

[Breuer 2013] T. Breuer, "CTblLib", 2013, available at https://tinyurl.com/ctbllib. GAP package, version 1.2.2.

[Brundan and Kleshchev 2003] J. Brundan and A. Kleshchev, "Representation theory of symmetric groups and their double covers", pp. 31–53 in *Groups, combinatorics & geometry* (Durham, NC, 2001), edited by A. A. Ivanov et al., World Sci. Publ., River Edge, NJ, 2003. MR Zbl

[Conway et al. 1985] J. H. Conway, R. T. Curtis, S. P. Norton, R. A. Parker, and R. A. Wilson, *Atlas of finite groups*, Oxford Univ. Press, 1985. MR Zbl

[Geck et al. 1996] M. Geck, G. Hiss, F. Lübeck, G. Malle, and G. Pfeiffer, "CHEVIE: a system for computing and processing generic character tables", *Appl. Algebra Engrg. Comm. Comput.* **7**:3 (1996), 175–210. MR Zbl

[Gorenstein 1968] D. Gorenstein, *Finite groups*, Harper & Row, New York, 1968. MR Zbl

[Grove 2002] L. C. Grove, *Classical groups and geometric algebra*, Graduate Studies in Math. **39**, Amer. Math. Soc., Providence, RI, 2002. MR Zbl

[Himstedt et al. 2009] F. Himstedt, H. N. Nguyen, and P. H. Tiep, "On the restriction of cross characteristic representations of ${}^2F_4(q)$ to proper subgroups", *Arch. Math.* (*Basel*) **93**:5 (2009), 415–423. MR Zbl

[Isaacs 1976] I. M. Isaacs, *Character theory of finite groups*, Pure Appl. Math. **69**, Academic Press, New York, 1976. MR Zbl

[James and Liebeck 2001] G. James and M. Liebeck, *Representations and characters of groups*, 2nd ed., Cambridge Univ. Press, 2001. MR Zbl

[Kleidman and Liebeck 1990] P. Kleidman and M. Liebeck, *The subgroup structure of the finite classical groups*, Lond. Math. Soc. Lect. Note Series **129**, Cambridge Univ. Press, 1990. MR Zbl

[Kleshchev and Sheth 2002] A. S. Kleshchev and J. Sheth, "Representations of the alternating group which are irreducible over subgroups", *Proc. London Math. Soc.* (3) **84**:1 (2002), 194–212. MR Zbl

[Kleshchev and Tiep 2004] A. S. Kleshchev and P. H. Tiep, "On restrictions of modular spin representations of symmetric and alternating groups", *Trans. Amer. Math. Soc.* **356**:5 (2004), 1971–1999. MR Zbl

[Kleshchev and Tiep 2010] A. S. Kleshchev and P. H. Tiep, "Representations of the general linear groups which are irreducible over subgroups", *Amer. J. Math.* **132**:2 (2010), 425–473. MR Zbl

[Landazuri and Seitz 1974] V. Landazuri and G. M. Seitz, "On the minimal degrees of projective representations of the finite Chevalley groups", *J. Algebra* **32** (1974), 418–443. MR Zbl

[Liebeck 1985] M. W. Liebeck, "On the orders of maximal subgroups of the finite classical groups", *Proc. London Math. Soc.* (3) **50**:3 (1985), 426–446. MR Zbl

[Lübeck 2007] F. Lübeck, "Character degrees and their multiplicities for some groups of Lie type of rank < 9", online notes, 2007, available at https://tinyurl.com/lublierank.

[Nguyen 2008] H. N. Nguyen, "Irreducible restrictions of Brauer characters of the Chevalley group $G_2(q)$ to its proper subgroups", *J. Algebra* **320**:4 (2008), 1364–1390. MR Zbl

[Nguyen and Tiep 2008] H. N. Nguyen and P. H. Tiep, "Cross characteristic representations of ${}^3D_4(q)$ are reducible over proper subgroups", *J. Group Theory* **11**:5 (2008), 657–668. MR Zbl

[Schaeffer Fry 2013] A. A. Schaeffer Fry, "Cross-characteristic representations of $\mathrm{Sp}_6(2^a)$ and their restrictions to proper subgroups", *J. Pure Appl. Algebra* **217**:8 (2013), 1563–1582. MR Zbl

[Seitz 1987] G. M. Seitz, *The maximal subgroups of classical algebraic groups*, Mem. Amer. Math. Soc. **365**, Amer. Math. Soc., Providence, RI, 1987. MR Zbl

[Seitz 1990] G. M. Seitz, "Cross-characteristic embeddings of finite groups of Lie type", *Proc. London Math. Soc.* (3) **60**:1 (1990), 166–200. MR Zbl

[Seitz and Testerman 1990] G. M. Seitz and D. M. Testerman, "Extending morphisms from finite to algebraic groups", *J. Algebra* **131**:2 (1990), 559–574. MR Zbl

[Tiep and Zalesskii 1996] P. H. Tiep and A. E. Zalesskii, "Minimal characters of the finite classical groups", *Comm. Algebra* **24**:6 (1996), 2093–2167. MR Zbl

Received: 2018-02-12 Revised: 2018-10-11 Accepted: 2018-10-12

kalbee@msudenver.edu *Metropolitan State University of Denver, Denver, CO, United States*

jmichaelbarnes@gmail.com *Metropolitan State University of Denver, Denver, CO, United States*

aaronparker2319@gmail.com *Metropolitan State University of Denver, Denver, CO, United States*

ebroon@math.arizona.edu *University of Arizona, Tucson, AZ, United States*

aschaef6@msudenver.edu *Department of Mathematical and Computer Sciences, Metropolitan State University of Denver, Denver, CO, United States*

INVOLVE 12:4 (2019)
dx.doi.org/10.2140/involve.2019.12.633

Prime labelings of infinite graphs

Matthew Kenigsberg and Oscar Levin

(Communicated by Kenneth S. Berenhaut)

A finite graph on n vertices has a prime labeling provided there is a way to label the vertices with the integers 1 through n such that every pair of adjacent vertices has relatively prime labels. We extend the definition of prime labeling to infinite graphs and give a simple necessary and sufficient condition for an infinite graph to have a prime labeling. We then measure the complexity of prime labelings of infinite graphs using techniques from computability theory to verify that our condition is as simple as possible.

1. Introduction

A *graph labeling* is essentially an assignment of integers to the vertices (or sometimes edges or both) of a graph subject to certain conditions. In the last 50 or so years, a multitude of graph labelings have been described and studied. The dynamic survey [Gallian 1998] describes over 50 types of graph labelings with results drawn from over 2000 papers. All but a handful of these consider only *finite* graphs. Here we consider one type of graph labeling and see how we can extend the definition to infinite graphs, with the hope that understanding this limit case might shed some light on open problems for finite graphs.

For a finite graph $G(V, E)$, a *prime labeling* is a bijection $f : V \to \{1, 2, \ldots, |V|\}$ such that for all $\{u, v\} \in E$, $f(u)$ and $f(v)$ are relatively prime ($\gcd(f(u), f(v)) = 1$). If a graph admits a prime labeling, we call the graph *prime*. This notion of graph labeling originates with Entringer, and was first described in a paper by Tout, Dabboucy, and Howalla [Tout et al. 1982]. Most of the results on prime labelings have been to show that large classes of graphs are in fact prime, but little is known in general. For example, Pikhurko [2007] proved that all trees with up to 50 vertices are prime. Recently Haxell, Pikhurko, and Taraz [Haxell et al. 2011] proved that all large trees are prime. However, the Entringer–Tout conjecture, that all trees are prime, remains open.

MSC2010: 05C78, 05C63, 05C85, 03D80.
Keywords: graph labelings, infinite graphs, prime labelings, computability theory.

A similar story emerges for another class of graphs: ladders ($P_n \square P_2$ for some n). T. Varkey conjectured in an unpublished work that all ladders are prime. Work on this question was done in [Berliner et al. 2016; Sundaram et al. 2006; 2007], and a recent preprint [Ghorbani and Kamali 2016] claims to prove the conjecture.

In this present work, we ask which *infinite* graphs admit prime labelings. As far as we know, this is the first attempt at such an investigation, although we note that other types of labelings have successfully been extended to infinite graphs, such as in [Combe and Nelson 2006] for magic labelings or [Chan et al. 2009] for graceful labelings. The latter is particularly interesting in that it classifies precisely which infinite trees have graceful labelings, despite the long open conjecture that all (finite) trees are graceful. In Section 4, we will similarly prove that all infinite trees and all infinite ladders are prime.

We will start in Section 2 with some preliminary definitions and notation. Then in Section 3 we give an algorithm which produces a prime labeling of many infinite graphs that have prime labelings. This will lead us to a classification theorem for which infinite graphs are prime, which we state and prove in Section 4. We consider issues of complexity in Section 5. Finally, we conclude with some open questions in Section 6.

2. Preliminaries

Before we can study prime labelings of infinite graphs, we must decide what exactly we mean by this. First, by an infinite graph $G = (V, E)$ we will always mean a countably infinite graph (while there are uncountable graphs, it does not make sense to label these with integers). We could safely take $V = \mathbb{N} = \{0, 1, 2, \ldots\}$, but we will usually use $v_0, v_1, v_2, \ldots$ for the names of the vertices to avoid confusion with their labels. The edge set E will simply be a set of two-element subsets of V. Note this allows for finite or countably infinite numbers of edges, and does not prohibit vertices having infinite degree.

We will freely generalize standard notation for graphs to the infinite case: $K_{2,\infty}$, for example, will be the complete bipartite graph which has two vertices in one part and infinitely many in the other. The only time standard notation becomes ambiguous is with infinite paths: since P_n is a path with n edges, it makes sense to consider P_∞ as a path with infinitely many edges. However, there are two options here. The path could extend infinitely in both directions (a *two-way infinite path*) or just one (a *one-way infinite path*). We will use P_∞ to represent the one-way infinite path and not adopt a notation for the former.

It is then reasonable to extend the definition of prime labeling to infinite graphs as follows:

Definition. Given an infinite graph $G = (V, E)$, a *prime labeling* is a bijection $f : V \to \{1, 2, \ldots\}$ such that $\gcd(f(u), f(v)) = 1$ for all $\{u, v\} \in E$.

In what follows, it will sometimes be useful to exclude 1 from the codomain. Following Vaidya and Prajapati [2011], who introduced and studied k-prime labelings for finite graphs, we define k-prime labelings of infinite graphs as follows:

Definition. Given an infinite graph $G = (V, E)$, a *k-prime labeling* is a bijection $f : V \to \{k, k+1, k+2, \ldots\}$ such that $\gcd(f(u), f(v)) = 1$ for all $\{u, v\} \in E$.

Note that a 1-prime labeling is the same as a prime labeling. Thus trivially, every prime graph is k-prime for some k, and every graph that is k-prime for all k will be prime. We will see shortly that there are infinite graphs that are prime but not 2-prime. However, it turns out that every infinite 2-prime graph is k-prime for *all* k. This can be seen by considering an algorithm for producing a k-prime labeling, as we now proceed to do.

3. An algorithm for prime labelings

We begin by describing a procedure which we think is a reasonable way to produce a k-prime labeling of an infinite graph. As usual, we take the vertex set to be $V = \{v_0, v_1, \ldots\}$.

We will proceed in stages, so that the every vertex is assigned some label at a finite stage, and in the limit, the labeling of the graph is k-prime. At the start of stage s, we will assume that we have labeled finite subsets $V_s \subseteq V$ without mistakes (i.e., the greatest common divisor of labels on any two adjacent vertices in V_s is 1), and proceed to find and label two vertices appropriately.

Algorithm 3.1. Proceed in stages.

Stage $s = 0$: label v_0 with k and set $V_1 = \{v_0\}$.

Stage $s > 0$: Given labeled $V_s \subset V$:

(1) Find the least natural number i such that v_i is not adjacent to any vertex in V_s, and label it with the least integer greater than k not yet used as a label.

(2) Find the least integer j such that v_j is unlabeled, and label it with a prime not yet used as a label, larger than any label of vertices adjacent to v_j.

(3) Let $V_{s+1} = V_s \cup \{v_i, v_j\}$ and proceed to the next stage.

By design, this algorithm will always label adjacent vertices with numbers that are relatively prime. Since there are infinitely many prime numbers, it is always possible to complete step (2) of each stage. Thus, in order to show that this algorithm produces a k-prime labeling for a graph, it is only necessary to show that it is always possible to find a vertex v_i such that v_i is not adjacent to any vertex in V_s.

To illustrate the algorithm, we give some examples of infinite graphs that have prime labelings, as well as some that do not.

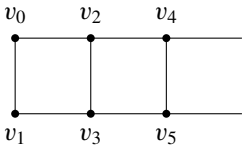


Figure 1. A (one-way) infinite ladder.

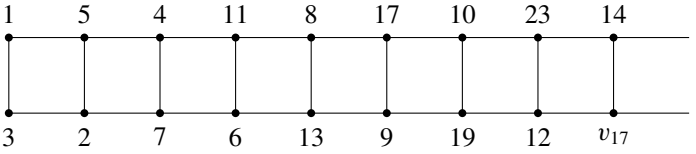


Figure 2. The result of the first eight stages of the algorithm.

Example 3.2. The graph $P_\infty \square P_2$ with vertices arranged as in Figure 1 receives a prime labeling from Algorithm 3.1.

The result of the first eight stages of the algorithm is shown in Figure 2. Since the graph extends infinitely, it will always be possible to find a vertex not adjacent to any of the already labeled vertices. This means the algorithm will produce a prime labeling.

Example 3.3. An infinite complete binary tree with vertices arranged as in Figure 3 receives a prime labeling from Algorithm 3.1.

Once again, it will always be possible to find a vertex not connected to the labeled part of the graph, so the algorithm produces a prime labeling. The result of the first four stages of the algorithm is shown in Figure 4.

Example 3.4. Algorithm 3.1 does not produce a prime labeling for an infinite star (the graph $K_{1,\infty}$).

In order to produce a prime labeling, the algorithm must label the center of the star. After labeling the center of the star, step (1) of the next stage will attempt to find the least natural number i such that v_i is not adjacent to any vertex in the set

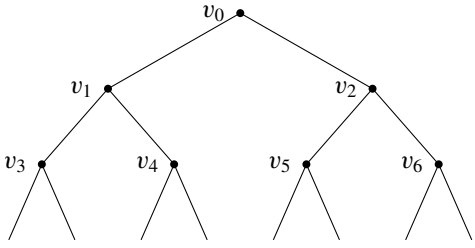


Figure 3. The top of a complete infinite binary tree.

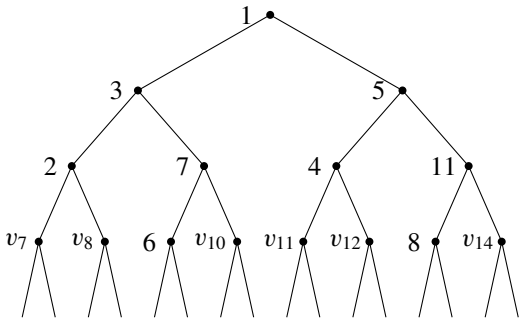


Figure 4. The labeling after four stages.

of already labeled vertices, which includes the center of the star. Since the center of the star is adjacent to all other vertices, this is impossible, and the algorithm will not produce a prime labeling.

Note that if the infinite vertex was removed from the graph, the algorithm could easily produce a 2-prime labeling for the resulting graph. If the center of the star was then labeled with 1, the union of the two labelings would be a prime labeling for $K_{1,\infty}$.

Example 3.5. Algorithm 3.1 does not produce a prime labeling for the infinite bipartite graph $K_{\infty,\infty}$.

To see this, consider any graph $K_{\infty,\infty}$. Let a be the least natural number such that the vertex v_a is adjacent to v_0.

After a finite number of stages, v_a will be labeled. At the next stage, step (1) will look for the least natural number i such that v_i is not adjacent to any element of the set of labeled vertices $V_s \supset \{v_0, v_a\}$. Since every vertex is adjacent to either v_0 or v_a, this is not possible, and as such the algorithm will not be able to label the rest of the graph.

Unlike with the infinite star, there is no way to adjust the algorithm to produce a prime labeling of $K_{\infty,\infty}$.

Proposition 3.6. *$K_{\infty,\infty}$ has no prime labeling.*

Proof. Let $a \neq 1$ and $b \neq 1$ be any two labels of a pair of vertices in separate partite sets, and consider $n = ab$. Whatever vertex gets labeled with n (or indeed, any multiple of n) cannot be adjacent to either of the vertices labeled a or b. However, every vertex is adjacent to one of these vertices, a contradiction. Thus the graph has no prime labeling. □

4. Classification of infinite graphs

We have seen that not all graphs have prime labelings. The issue illustrated in Proposition 3.6 demonstrates a particular obstruction, which we summarize in the

following lemma. Let $N(S)$ denote the set of vertices adjacent to one or more vertices in S (the *open neighborhood* of S) and $N[S] = N(S) \cup S$ (the *closed neighborhood* of S).

Lemma 4.1. *If an infinite graph $G = (V, E)$ has a finite set $S \subset V$, for which $N[S]$ contains all but finitely many vertices of G, then G does not have a k-prime labeling.*

Proof. Suppose G has a k-prime labeling, and consider such a finite set $S \subset V$. Let n be the product of the labels on the vertices of S. As such the infinitely many multiples of n must be assigned to vertices not in $N[S]$. Thus $N[S]$ cannot be cofinite, contrary to hypothesis. □

Note that if S is finite and $N[S]$ is cofinite, then there is a finite set S' for which $N[S'] = V$ (add to S all finitely many elements not in $N[S]$). Such a set S' is called a *dominating set*. Thus another way to describe the obstruction to a graph having a k-prime labeling is to say the graph has a finite dominating set. We will see that graphs that avoid this obstruction will always have a k-prime labeling at least for each $k \geq 2$. Thus we make the following definition.

Definition. An infinite graph $G = (V, E)$ is called *finitely dominated* provided there is some finite dominating set S, that is, a finite S such that $N[S] = V$.

Theorem 4.2. *An infinite graph G has a k-prime labeling for $k \geq 2$ if and only if G is not finitely dominated.*

Proof. The forward direction is Lemma 4.1.

Conversely, if G is not finitely dominated, then for any finite set S of vertices there is a vertex not adjacent to any element in S. This means that Algorithm 3.1 will produce a k-prime labeling: at each stage, V_s is finite, so it is always possible to find the least natural number i such that v_i is not adjacent to any vertex in the set V_s of already labeled vertices. □

We saw in Example 3.4 that the infinite star does not get a k-prime labeling from Algorithm 3.1, and by this theorem, we see that in fact it cannot have a k-prime labeling for any $k \geq 2$ (the center vertex is dominating). However, the infinite star *is* prime, since we can eliminate the "problem" by labeling the center vertex 1. This works in general and provides our main classification theorem.

We write $G - v$ for the graph resulting from removing the vertex v (and all incident edges).

Theorem 4.3. *An infinite graph G has a prime labeling if and only if there is a vertex v such that $G - v$ is not finitely dominated.*

Proof. Suppose first that G has a prime labeling f for which $f(v) = 1$. Then $G^- = G - v$ is 2-prime, witnessed by $f|_{G^-}$. By Theorem 4.2, G^- is not finitely dominated, as required.

Conversely, if $G - v$ is not finitely dominated, then $G - v$ has a 2-prime labeling by Theorem 4.2. The vertex that was removed can be labeled with 1, giving a prime labeling of G. □

Note, another way to state this result is that a graph will have a prime labeling if and only if it is possible to remove one vertex such that the remaining graph has a 2-prime labeling.

We can now state the relationship between k-prime graphs for different values of k.

Corollary 4.4. *If a graph has a k-prime labeling for any $k \geq 2$, it has a k-prime labeling for all k.*

Proof. According to Theorem 4.2, the condition for a graph to have a k-prime labeling is exactly the same for any $k \geq 2$. So if a graph satisfies that condition for any $k \geq 2$, it satisfies it for all $k \geq 2$. Further, if a graph is 2-prime, then it is not finitely dominated. But then $G - v_0$ will also not be finitely dominated, so by Theorem 4.3, G will have a prime labeling. □

As a result of our classification theorem, some natural classes of graphs will clearly have prime labelings.

Corollary 4.5. *All infinite trees are prime.*

We say a graph is *locally finite* if every vertex has finite degree.

Corollary 4.6. *All infinite locally finite graphs are prime. In particular, the infinite ladder is prime.*

The reason locally finite graphs allow our algorithm to work is that the neighborhood of any finite set must be finite. But even if this doesn't happen, we could always have enough vertices not adjacent to the finite set for other reasons. For example, the graph could have infinitely many connected components or one of the connected components could have infinite diameter.

Corollary 4.7. *All infinite graphs with infinitely many connected components or containing a connected component with infinite diameter have prime labelings.*

5. Computable graphs

We turn now to the question of complexity of prime labelings for infinite graphs. In the finite case, we would consider computational complexity: you might ask whether deciding if a finite graph has a prime labeling is NP-complete. For infinite graphs, we use ideas from *computability theory*.

To do this, we must restrict our attention to *computable* graphs. Essentially, we identify graphs with their edge set, taking the vertex set to be $\mathbb{N}$, and require the edge set to be a computable set. This means that there is an algorithm that,

given any two vertices (natural numbers) as input, returns whether the two vertices are adjacent. A more precise definition is beyond the scope of this paper, but the interested reader can see [Soare 1987] for background on computability theory in general or [Gasarch 1998] for a survey of the use of computability theory in combinatorics.

The first natural question to consider in this context is whether all computable graphs that have prime labelings have *computable* prime labelings (note that since we insist $V = \mathbb{N}$, a computable graph must necessarily be infinite). In other words, if the graph is nicely presented, will it always be possible to nicely describe a prime labeling? Somewhat surprisingly, the answer here is yes. (This is surprising given that many graph-theoretic properties do not behave so nicely: there are computable graphs with 3-colorings with no computable 3-coloring [Bean 1976a] and computable graphs with Euler paths with no computable Euler path [Bean 1976b], for example.)

Proposition 5.1. *If G is a computable graph which admits a prime labeling, then G has a computable prime labeling.*

Proof. Let G be a computable graph with a prime labeling. By Theorem 4.3, we know that there is a vertex v such that $G - v$ is not finitely dominated. Label v with 1, then proceed with Algorithm 3.1. At step (1) of stage s, we are looking for a vertex not in $N[V_s]$. This can be found in finite time by asking whether v_i is adjacent to v_j for each $v_j \in V_s$, and if ever the answer is yes, we move on to the next potential v_i, which we know we must eventually find since V_s is not dominating. □

The procedure outlined above relies on a certain amount of *nonuniformity*: we must know where to place the label 1. This does not prevent the prime labeling from being computable, since we are only asking for the existence of an algorithm for the prime labeling, not for a procedure to *find* that algorithm. But could we? Is it possible, given the algorithm for a particular graph, to produce the algorithm that gives the prime labeling? Here, we find the answer is negative.

Theorem 5.2. *There is no computable function which, given any computable graph admitting a prime labeling, produces the prime labeling for that graph.*

Before we give the proof, we need a little more background from computability theory. They key fact we will use is that there is an effective list $\varphi_0, \varphi_1, \varphi_2, \ldots$ of all *partial* computable functions (again, see [Soare 1987] for details). The intuition here is that we can consider every possible algorithm, perhaps written in Java, arranged alphabetically and by length (all algorithms have finite length). Of course, for any given algorithm, we have no reason to think that this algorithm will halt on all inputs, and this is why we are only considering *partial* computable functions (if it does halt on all inputs, we call it *total*). However, since the list contains every

algorithm, partial or total, we know that if there were a computable function which gave the computable prime labeling of every computable graph (admitting a prime labeling), it must be somewhere on the list. Our goal then is to ensure every partial computable function on the list is wrong at least once.

Proof. We will build a sequence $G_0, G_1, \ldots$ of computable graphs, each admitting a prime labeling. While doing so, we will ensure that, for each $e \in \mathbb{N}$, the partial computable function φ_e is not a prime labeling of the graph G_e.

The construction will "dove-tail" the construction of the infinitely many graphs, so that by the end of stage s, we will have described the first s vertices of the first s graphs. The construction of each graph in the sequence will be independent of the others, so we need only describe how we build an arbitrary graph G_e.

In the limit, the graph G_e will be the union of two stars with centers v_0 and v_1, at least one of which is infinite. Notice that such a graph will have a prime labeling, as removing the center of an infinite star produces an infinite set of isolated vertices (we are appealing to Theorem 4.3 here). At each stage, we check whether φ_e has returned the label 1 for either v_0 or v_1. If this has not yet occurred, we add a new vertex adjacent to either v_0 or v_1, whichever we did not add to in the previous stage. If φ_e returns 1 for the label of v_i with $i \in \{0, 1\}$, then we only ever add new vertices adjacent to v_{1-i}.

Note that it is possible that φ_e will never return 1 for v_0 or v_1 (perhaps φ_e is not total, or it labels a different vertex with 1). In this case, G_e will consist of two infinite stars, but there is no way for φ_e to be a prime labeling (the product of the labels of the two centers has nowhere to go, as in Proposition 3.6). On the other hand, if φ_e does label one of the vertices v_0 or v_1 with a 1, then we never add any more neighbors to that vertex, and only the other vertex will be an infinite star. In this case, φ_e also cannot be a prime labeling. Whatever the label of the center of the infinite star is, there are only finitely many vertices (on the other star) that the infinitely many multiples of this label can be assigned to. This completes the proof. □

The proof above relies on the inability of computable functions to predict whether a vertex of a graph will have infinite degree, and as such, the computable function does not know which vertex to label with 1. However, this is the only barrier to uniformity. If we consider instead 2-prime labelings, then we get uniformity.

The other computability question we should consider is the *decision problem*: given a computable graph, how hard is it to decide whether the graph has a prime labeling? The usual way to analyze this in computability theory is to determine where the decision problem lies inside (or above) the arithmetical hierarchy. One way to think of this task is that we are assessing the complexity of the condition which is equivalent to a graph having a prime labeling. We have a condition given

in Theorem 4.3. Is this the simplest necessary and sufficient condition to a graph having a prime labeling?

Notice that by Theorem 4.2, a graph has a k-prime labeling for $k \geq 2$ if and only if for all finite sets of vertices, there is at least one vertex not in the neighborhood of the set. Analyzing the quantifiers, we can state this condition as

$$\forall n \, \exists k \, (k > n \wedge k \notin N(\{0, 1, \ldots, n\})).$$

Since saying that a vertex is not in the neighborhood of a finite set of vertices is computable, we see that a graph having a 2-prime labeling is Π^0_2. Similarly, to say a graph has a prime labeling, we need it to be the case that there is a vertex, the removal of which, leaves a 2-prime graph. Thus a graph having a prime labeling is Σ^0_3.

Can we do better? For 2-prime labelings, the answer is no.

Theorem 5.3. *The decision problem for a graph having a k-prime labeling for $k \geq 2$ is Π^0_2-complete.*

Proof. Fix $k \geq 2$. We argued above that having a k-prime labeling is Π^0_2, so we need only show completeness. We will do this by giving a 1-reduction to the known Π^0_2-complete index set $\mathrm{INF} = \{e : |W_e| = \infty\}$, where W_e is the domain of φ_e. That is, we build a sequence of computable graphs $\{G_i\}$ such that G_e has a k-prime labeling if and only if $e \in \mathrm{INF}$.

We build the graphs simultaneously, as in the proof of Theorem 5.2, but this time each graph will either be the disjoint union of an infinite star with a finite path, or the disjoint union of an infinite star with a (one-way) infinite path. In the former case, the graph will not be k-prime, and in the latter it will be k-prime, by Theorem 4.2.

The procedure for building the graph G_e is as follows. Initialize G_e with a center vertex for its star and an initial vertex for its path. At stage s of the construction we assume that we have built a finite star and a finite path. Run $\varphi_e(x)$ on all $x < s$ for which $\varphi_e(x)$ has not already halted at some earlier stage. We continue to run these computations until either $\varphi_e(x)$ halts for some input x, or until each computation has run for s steps, whichever comes first. If we see some $\varphi_e(x)$ halt, this will be the first time we realize that $x \in W_e$, so we have further evidence that $|W_e|$ might be infinite. Thus we add a vertex to the end of the finite path. On the other hand, if no (new) x appears in W_e (i.e., $\varphi_e(x)$ does not halt for any new x by stage s) we work off the assumption that $|W_e|$ is finite and add a vertex to the finite star in G_e.

To verify that this procedure gives us what we want, suppose first that $|W_e| = \infty$. Then there will be infinitely many stages at which we add a vertex to the end of the path, since at each stage we "discover" at most one new x in W_e. Thus in the limit, the path will be infinite (the star will likely be infinite as well, but regardless, G_e will have a k-prime labeling). Conversely, suppose $|W_e|$ is finite. Then there

is a last stage at which any x appears in W_e, and so after that stage, we never add vertices to the path, making the path finite. □

What about prime labelings? By the quantifier analysis above, we know that the decision problem cannot be harder than Σ_3^0. Further, a simple modification of the proof for Theorem 5.3 shows that the decision problem is at least Π_2^0-hard. We would expect the decision problem to in fact be Σ_3^0-complete, but a proof that it is Σ_3^0-hard goes beyond the scope of this paper. We leave this as an open question.

Question 1. Is the decision problem for a graph having a prime labeling Σ_3^0-complete?

6. Conclusion and open questions

We have considered a natural extension of the definition of prime labelings to infinite graphs. For 2-prime labelings, we have a simple necessary and sufficient condition and a condition only slightly less simple for prime labelings. By using tools from computability theory, we see that producing a 2-prime labeling of a 2-prime graph is as straightforward as possible, and only slightly less so for producing prime labelings of prime graphs. We also have that our criterion for 2-prime labelings is as simple as possible, and conjecture that the same is true for prime labelings.

These results mirror those for graceful labelings of infinite graphs, in that working with labelings of infinite graphs seems quite a bit easier than their finite counterparts. This suggests that the difficulty with working with finite graphs is very much tied to finiteness itself. The feeling of "running out of room" is exactly why labeling results are difficult.

We wonder however, whether a more restrictive definition of labelings for infinite graphs might serve as a better infinite analogue to the finite case. Note that for vertex coloring, it turns out that an infinite graph is k-colorable if and only if every finite subgraph is 4-colorable. Such a result for prime (and other) labelings would be very nice, but with our definition, is clearly false.

We do not know what the "right" definition would be, but we conclude by considering one possible variant of prime labeling that might be a step in the right direction and encourage others to pursue this further.

Definition. Let G be a graph, v_c be a vertex of that graph (c for *center*), and G_r be the subgraph of G that includes all vertices within distance r of v_c. Then G has a *limitwise prime labeling* if it is possible to choose v_c and label the graph such that for infinitely many r, G_r has been given a prime labeling.

We call a graph *limitwise prime* if it has a limitwise prime labeling.

To get a feel for this, consider the complete infinite binary tree.

Example 6.1. A complete infinite binary tree has a limitwise prime labeling.

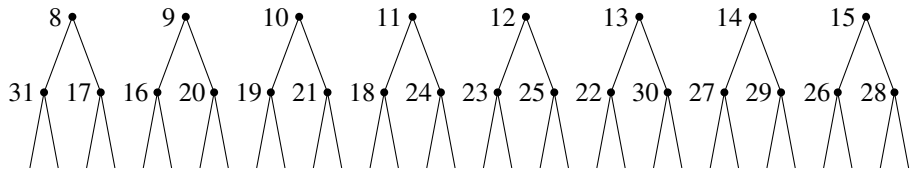


Figure 5. A limitwise prime labeling of rows 3 and 4 of the complete binary tree.

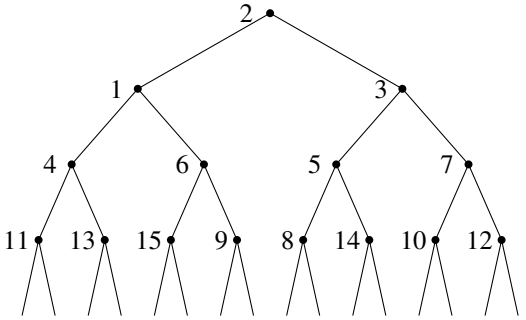


Figure 6. The start of a limitwise prime labeled tree.

Proof. For all $r \geq 3$, each row of the graph can have children labeled with the integers from 2^{r+1} to $2^{r+2} - 1$ as follows:

The lowest even number e has children $2e + 1$ and $4e - 1$. All other evens e have children $2e - 1$ and $2e + 1$. The lowest odd number o has children $2o - 2$ and $2o + 2$. The second-greatest odd number o has children $2o - 4$ and $2o + 4$. The greatest odd number o has children $2o - 4$ and $2o - 2$. All others odd numbers o have children $2o - 4$ and $2o + 2$.

The process is shown here for $r = 3$ in Figure 5.

It is straightforward but tedious to show that this will produce a limitwise prime labeling for the tree after the first four rows are labeled with the numbers 1 to 15 in any manner that is prime. One possibility is shown in Figure 6 □

It certainly appears that giving a limitwise prime labeling is more difficult that giving a prime labeling. Indeed, there are prime graphs that are not limitwise prime.

Example 6.2. Let G be the square of the two-way infinite path, as in Figure 7. Then G has a prime labeling, but not a limitwise prime labeling

Proof. Since G is locally finite, it has a prime labeling.

To show that G has no limitwise prime labeling, choose any vertex for v_c and let G_r be the subgraph that includes all vertices within distance r of v_c. G_r contains $4r + 1$ vertices. This means that if G_r has a prime labeling, then $2r$ even labels must be used.

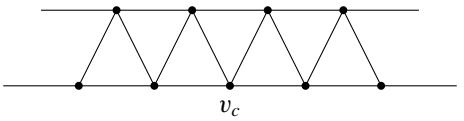


Figure 7. A prime graph that is not limitwise prime.

Without loss of generality, let v_c be on the bottom of the graph as shown in Figure 7, and let b and t be the number of vertices with even labels on the bottom and top of the graph respectively. Since there are $2r+1$ vertices on the bottom and adjacent vertices cannot have even labels, $b \leq r+1$. Similarly, $t \leq r$. Since $2r$ total even labels must be used, $b+t=2r$, so we have only two cases to consider: either $b=t=r$ or $b=r+1$ and $t=r-1$. We will argue that as soon as $r \geq 2$, both of these cases are impossible.

If $t=r$, then it must be that exactly every other vertex on top is even. Since each of these are adjacent to two different vertices on bottom, there is only one vertex on the bottom that can be even, so $b=1 \neq r$. On the other hand, if $b=r+1$, then every other vertex on bottom is even, leaving no vertices on top for even vertices, so $t=0 \neq r$.

So for $r>1$, G_r does not have a prime labeling, which means G does not have a limitwise prime labeling, even though it does have a prime labeling. □

There are plenty of questions to consider about limitwise prime labelings including whether this is even a useful variant of prime labeling of infinite graphs. Here are a few to get the ambitious reader started.

Question 2. Are all infinite trees limitwise prime?

Question 3. What are reasonable necessary and/or sufficient conditions for a graph to be limitwise prime?

Note that if every finite subgraph of an infinite graph is prime, then the graph is limitwise prime. However, the converse is likely false. This could be investigated further.

There are also questions of complexity:

Question 4. Does every computable graph with a limitwise prime labeling have a computable limitwise prime labeling?

Question 5. How hard is it to decide whether a computable graph is limitwise prime?

References

[Bean 1976a] D. R. Bean, "Effective coloration", *J. Symbolic Logic* **41**:2 (1976), 469–480. MR Zbl

[Bean 1976b] D. R. Bean, "Recursive Euler and Hamilton paths", *Proc. Amer. Math. Soc.* **55**:2 (1976), 385–394. MR Zbl

[Berliner et al. 2016] A. H. Berliner, N. Dean, J. Hook, A. Marr, A. Mbirika, and C. D. McBee, "Coprime and prime labelings of graphs", *J. Integer Seq.* **19**:5 (2016), art. id. 16.5.8. MR Zbl

[Chan et al. 2009] T. L. Chan, W. S. Cheung, and T. W. Ng, "Graceful tree conjecture for infinite trees", *Electron. J. Combin.* **16**:1 (2009), art. id. 65. MR Zbl

[Combe and Nelson 2006] D. Combe and A. M. Nelson, "Magic labellings of infinite graphs over infinite groups", *Australas. J. Combin.* **35** (2006), 193–210. MR Zbl

[Gallian 1998] J. A. Gallian, "A dynamic survey of graph labeling", *Electron. J. Combin.* **5** (1998), art. id. 6. MR Zbl

[Gasarch 1998] W. Gasarch, "A survey of recursive combinatorics", pp. 1041–1176 in *Handbook of recursive mathematics, II*, edited by Y. L. Ershov et al., Stud. Logic Found. Math. **139**, North-Holland, Amsterdam, 1998. MR Zbl

[Ghorbani and Kamali 2016] E. Ghorbani and S. Kamali, "Prime labeling of ladders", preprint, 2016. arXiv

[Haxell et al. 2011] P. Haxell, O. Pikhurko, and A. Taraz, "Primality of trees", *J. Comb.* **2**:4 (2011), 481–500. MR Zbl

[Pikhurko 2007] O. Pikhurko, "Trees are almost prime", *Discrete Math.* **307**:11-12 (2007), 1455–1462. MR Zbl

[Soare 1987] R. I. Soare, *Recursively enumerable sets and degrees*, Springer, 1987. MR Zbl

[Sundaram et al. 2006] M. Sundaram, R. Ponraj, and S. Somasundaram, "On a prime labeling conjecture", *Ars Combin.* **79** (2006), 205–209. MR Zbl

[Sundaram et al. 2007] M. Sundaram, R. Ponraj, and S. Somasundaram, "A note on prime labeling of ladders", *Acta Cienc. Indica Math.* **33**:2 (2007), 471–477. MR Zbl

[Tout et al. 1982] R. Tout, A. N. Dabboucy, and K. Howalla, "Prime labeling of graphs", *Nat. Acad. Sci. Lett. India* **5**:11 (1982), 365–368. Zbl

[Vaidya and Prajapati 2011] S. Vaidya and U. Prajapati, "Some results on prime and k-prime labeling", *J. Math. Res.* **3**:1 (2011), 66. Zbl

Received: 2018-02-22 Revised: 2018-07-09 Accepted: 2018-11-08

matthew.kenigsberg@vanderbilt.edu *Vanderbilt University, Nashville, TN, United States*

oscar.levin@unco.edu *School of Mathematical Sciences, University of Northern Colorado, Greeley, CO, United States*

INVOLVE 12:4 (2019)
dx.doi.org/10.2140/involve.2019.12.647

Positional strategies in games of best choice

Aaron Fowlkes and Brant Jones

(Communicated by Kenneth S. Berenhaut)

We study a variation of the game of best choice (also known as the secretary problem or game of googol) under an additional assumption that the ranks of interview candidates are restricted using permutation pattern-avoidance. We describe the optimal positional strategies and develop formulas for the probability of winning.

1. Introduction

The game of best choice, also known as the "secretary problem," appeared in Martin Gardner's 1960 Scientific American column (reprinted in [Gardner 1995]), although it has a history which predates this; see, e.g., [Kadison 1994]. Gilbert and Mosteller [1966] gave a nice survey of the problem and solved some variations. The basic idea is to try to hire the best candidate out of N applicants for a job, each candidate having a specific ranking 1 (worst) through N (best). When interviewing the candidates, the decision must be made to hire them or not, on the spot, and candidates cannot be recalled later. The order of the interviews is (uniformly) random and so the interviewer does not know when the top candidate will come in.

As an example, suppose the interviews have rank order 574239618. The interviewer will be able to rank each initial segment of candidates relative to each other, but will not know their rank overall out of N. So the interviewer will see

$$1,\ 12,\ 231,\ 3421,\ 45312,\ 453126,\ \ldots$$

and must decide when to stop and hire. We count the game as a win if the best candidate out of N is hired and as a loss otherwise, with all losses having equal value. The optimal strategy, for N sufficiently large, turns out to be to reject the first N/e of the candidates (about 37%) and then hire the next candidate who is better than all earlier candidates.

Now, suppose that a consulting firm (with some oracular powers) agrees to filter candidates for the interviewer. They offer two strategies. In the first strategy, they

MSC2010: 05A05, 91A60.
Keywords: secretary problem, random permutation, permutation pattern.

will guarantee that each time a candidate B ranks lower than some candidate A already interviewed ("disappointing"), no future candidates will rank lower than B. In the second strategy, they guarantee that each time a candidate B ranks higher than some candidate A already interviewed ("raising the bar"), no future candidates will rank lower than A. All other aspects of the game remain the same.

Is there any difference between these? Are they better or worse than the classical case?

2. Refinement

Interview rank orders are *permutations* of some fixed size N which we write using the notation $p_1 p_2 \cdots p_N$, where the p_i are the values $1, 2, \ldots, N$ arranged in some order. In this work, we restrict the interview rank orders using pattern-avoidance.

Definition 2.1. We say that the permutation $p = p_1 p_2 \cdots p_N$ contains the pattern $q = q_1 q_2 q_3$ if there exist $i < j < k$ such that p_i, p_j, p_k are in the same relative order as q_1, q_2, q_3.

So, the "disappointment-free" consulting strategy is equivalent to requiring the interview rank orders to be 321-avoiding. Similarly, the "bar-raising" situation is the same as 231-avoiding. See the textbook [Bóna 2012] for a gentle introduction to pattern-avoidance. Putting aside the story about the consultants, we believe that pattern-avoidance is a natural mechanism for modeling the effect of domain learning by the player during the game. More precisely, as the interviewer ranks the current candidates at each step, they acquire information that allows them to hone the pool to include more relevant candidates at future time steps. We represent this honing process using pattern-avoidance.

The *left-to-right maxima* in a permutation p consist of elements p_j that are larger in value than every element p_i to the left (i.e., for $i < j$). In the game of best choice, it is never optimal to select a candidate that is not a left-to-right maximum. A *positional* strategy for the game of best choice is one in which the interviewer transitions from rejection to hiring based on the position of the interview. More precisely, the interviewer may play the *k-positional* strategy on a permutation p by rejecting candidates $p_1, p_2, \ldots, p_k$ and then accepting the next left-to-right maximum thereafter. If k is set too high, it is likely the player will miss the best candidate. If k is set too low, they will probably not have set their standards high enough to capture the best candidate. We say that a particular interview rank order is *k-winnable* if transitioning from rejection to hiring after the k-th interview captures the best candidate. For example, 574239618 is k-winnable for $k = 2, 3, 4$, and 5. It is straightforward to verify that a permutation p is k-winnable precisely when k lies between the last two left-to-right maxima in p.

In this paper, we restrict to using these positional strategies applied to a permutation chosen uniformly at random among those avoiding 321 (or, alternatively, 231) in order to facilitate comparison with the classical case. For each model, we seek to determine the optimal transition position k and probability of winning for finite N and asymptotically as $N \to \infty$.

We now mention some ties to recent work. Several authors have investigated the distribution of various permutation statistics for a random model in which a pattern-avoiding permutation is chosen uniformly at random. For example, [Miner and Pak 2014] finds the positions of smallest and largest elements as well as the number of fixed points in a random permutation avoiding a single pattern of size 3; [Madras and Pehlivan 2016] finds the probability that one or two specified points occur in a random permutation avoiding 312; and the work of several authors [Deutsch et al. 2002/03; Firro et al. 2007] determines the lengths of the longest monotone and alternating subsequences in a random permutation avoiding a single pattern of size 3. We also consider uniformly random 321-avoiding and 231-avoiding permutations in our work, but the statistics we are concerned with arise from the game of best choice. In some sense, our results refine the question of where a uniformly random pattern-avoiding permutation achieves its maximum because in our problem we want to transition so as to capture the maximum value. We also consider asymptotics for both of our models, thus obtaining a "limit-strategy," just as in the classical game.

In addition, Wilf [1995] has collected some results on distributions of left-to-right maxima and Prodinger [2002] has studied these under a geometric random model. Although we phrase our results in terms of the game of best choice, they may also be viewed as an extension of the literature on distributions of left-to-right maxima to subsets of pattern-avoiding permutations.

3. Raising the bar

An *extension* of a permutation $p = p_1 p_2 \cdots p_{N-1}$ is the result of inserting value N into one of the N positions before, between, or after entries in p.

Lemma 3.1. *Let p be a 231-permutation of size $N-1$ and $0 \le k \le N-1$. Then there exists a unique extension of p that is k-winnable for N.*

Proof. Fix N and k. Let $p_1 p_2 \cdots p_k \,|\, p_{k+1} \cdots p_{N-1}$ be a 231-avoiding permutation of size $N-1$, with $p_m = \max\{p_1, p_2, \ldots, p_k\}$.

Define p_w to be the leftmost value greater than p_m among $\{p_{k+1}, p_{k+2}, \ldots, p_{N-1}\}$, and let q be the result of inserting N into the position directly prior to p_w (or into the last position if p_w does not exist). So we have

$$q = p_1 p_2 \cdots p_m \cdots p_k \,|\, p_{k+1} \cdots p_{w-1} N p_w \cdots p_{N-1}.$$

We claim that q is the unique 231-avoiding k-winnable extension of p. To see this, observe that:

- By construction, all elements of $\{p_{k+1}, \ldots, p_{w-1}\}$ are less than p_m, so q is k-winnable.
- We began with a 231-avoiding permutation p. If q contains 231, the value N must play the role of "3". Therefore, it suffices to show that all of the values lying to the left of N are less than all values lying to the right of N. By construction, $p_m = \max\{p_1, p_2, \ldots, p_{w-1}\}$ and $p_m < p_w$. If there exists some element $y < p_m$ among the entries $p_{w+1}, p_{w+2}, \ldots, p_{N-1}$ then (p_m, p_w, y) forms a 231-instance, contradicting that p is 231-avoiding. Hence, no such y exists and q is 231-avoiding.
- If the extension q were not unique, we would have two positions L_1 and L_2, say, where N could be inserted to the right of p_k to produce distinct k-winnable permutations of size N. In particular, there must exist at least one element p_v between L_1 and L_2. But the previous paragraph shows that we would require $p_m < p_v$ for the extension q using L_1 to be 231-avoiding, so the extension using L_2 is not k-winnable, a contradiction. Hence, the extension is unique. □

It is well known that the Catalan numbers

$$C_N = \frac{1}{N+1}\binom{2N}{N}$$

count the number of 231-avoiding permutations of size N; see, e.g., [Bóna 2012]. Hence, we obtain the following result.

Corollary 3.2. *There are exactly C_{N-1} permutations of size N that are* 231*-avoiding and k-winnable.*

Proof. For fixed k, the set of 231-avoiding permutations of size $N-1$ are in bijection with the set of 231-avoiding k-winnable permutations of size N by Lemma 3.1. □

Notice the curious consequence that *it does not matter which positional strategy we use*: for fixed N, the probability of selecting the best candidate is the same for all k. From the explicit formula, it is straightforward to work out the asymptotic probability of success

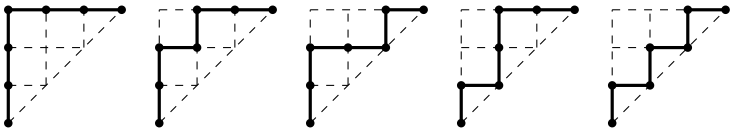
$$\lim_{N\to\infty} \frac{C_{N-1}}{C_N} = \frac{1}{4}.$$

4. Avoiding disappointment

Next, we consider positional strategies for the 321-avoiding interview rank orders. Recall that a permutation is k-winnable if and only if k lies between its last two left-to-right maxima. Hence, we study the distribution of left-to-right maxima in 321-avoiding permutations. For this, we make use of *Dyck paths*. These may be

viewed as paths in the Cartesian plane from $(0, 0)$ to (N, N), consisting of $(0, 1)$ steps (i.e., north) and $(1, 0)$ steps (i.e., east), staying above the line $y = x$. The *northeast corners* in a Dyck path consist of a north step immediately followed by an east step. We label each northeast corner by the column and height at the end of its east step.

Example 4.1. The Dyck paths for $N = 3$ are shown below:



Their sets of northeast corners are

$$\{(1,3)\},\ \ \{(1,2),(2,3)\},\ \ \{(1,2),(3,3)\},\ \ \{(1,1),(2,3)\},\ \ \{(1,1),(2,2),(3,3)\}$$

respectively.

Lemma 4.2. *The possible sets $\{p_{i_1}, p_{i_2}, \ldots, p_{i_m}\}$ of values and positions of left-to-right maxima arising from the various permutations of N are in bijection with the sets of northeast corners*

$$\{(i_j, p_{i_j}) : j = 1, \ldots, m\}$$

of Dyck paths of size N.

Proof. The defining property for a Dyck path is that at each step along the path, the number of east steps taken so far is less than or equal to the number of north steps taken so far. Equivalently, we may consider paths whose northeast corners satisfy the following two conditions:

- There is always a northeast corner in the first column.
- Whenever we add a northeast corner corresponding to p_{i_j}, we take at most $p_{i_j} - i_j$ east steps until we reach the next column with a northeast corner.

But this is precisely equivalent to the conditions that define sets of left-to-right maxima in a permutation:

- The first position is always a left-to-right maximum.
- Whenever we add a left-to-right maximum corresponding to p_{i_j}, we have (by definition) at most $p_{i_j} - i_j$ complementary values that are smaller than p_{i_j} and have not yet been used. Hence, there are at most $p_{i_j} - i_j$ entries until we reach the next left-to-right maximum.

Given a Dyck path representing a set of left-to-right maxima, we can produce a canonical permutation p that realizes this set of left-to-right maxima as follows: Place each p_{i_j} into position i_j and then fill the complementary positions with

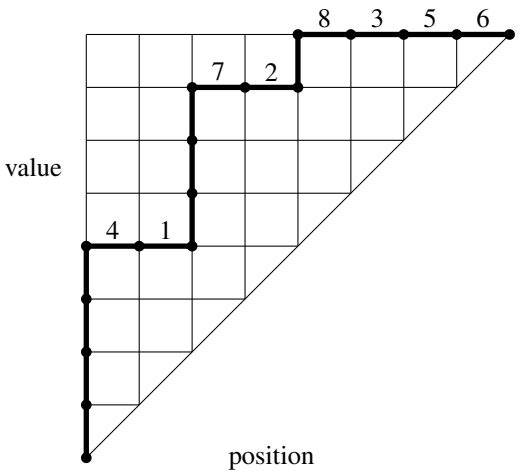


Figure 1. Completing the set of left-to-right maxima $\{p_1=4,\ p_3=7,\ p_5=8\}$.

the complementary values $\{1, 2, \ldots, N\} \setminus \{p_{i_1}, \ldots, p_{i_m}\}$ arranged increasingly. In terms of the Dyck path, we can label each northeast corner by the value of its corresponding left-to-right maximum, and then label the remaining horizontal edges with the complementary values, arranged increasingly as we read north and east along the path. Thus, the label for column i of the Dyck path gives the value for the i-th position of the permutation. □

As an example in $N = 8$, if $p_1 = 4$, $p_3 = 7$, and $p_5 = 8$ are the p_{i_j}, we obtain $p = 41728356$; this is illustrated in Figure 1.

Recall that the Catalan numbers C_N count 321-avoiding permutations of size N, and also count the number of Dyck paths of size N; see, e.g., [Bóna 2012]. Hence, we obtain the following result.

Corollary 4.3. *A* 321*-avoiding permutation p of size N is uniquely determined by the values and positions of its left-to-right maxima.*

Proof. The construction in the previous proof produces C_N distinct permutations of size N that have the structure of two increasing sequences shuffled together (namely, the sequence of left-to-right maxima, and the sequence of complementary values). Hence, the permutations constructed from Dyck paths in the previous result are all 321-avoiding. Since there are Catalan-many of each, there must be exactly one 321-avoiding permutation for each Dyck path. □

Definition 4.4. For $1 \le i \le N-1$ define $T_i(N)$ to be the total number of partial Dyck paths from $(0, 0)$ to $(N-1-i, N-1)$, and define $S_i(N)$ to be the number of Dyck paths from $(0, 0)$ to (N, N) where column $N-i$ lies weakly right of the next-to-last northeast corner and strictly left of the last northeast corner in the path.

$N\backslash k$	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1
2											**1**
3										**3**	2
4									6	**8**	5
5								10	20	**23**	14
6							15	40	65	**70**	42
7						21	70	145	214	**222**	132
8					28	112	280	514	717	**726**	429
9				36	168	490	1064	1817	**2442**	2431	1430
10			45	240	798	1988	3962	6446	**8437**	8294	4862
11		55	330	1230	3444	7784	14636	22997	**29510**	28730	16796
12	66	440	1815	5628	14154	29924	53937	82550	**104312**	100776	58786

Table 1. Number of k-winnable 321-avoiding permutations of N.

By Corollary 4.3, $S_i(N)$ is the number of $(N-i)$-winnable permutations of N. For example, the path in Figure 1 would be counted in $S_i(N)$ for $N-i \in \{3,4\}$ because the last two northeast corners occur in columns 3 and 5, respectively. Some initial values are given in Table 1. If we divide by the N-th Catalan number we obtain the probability of success for the corresponding $(N-i)$-positional strategy. These are illustrated in Table 2. It turns out that the $T_i(N)$ are Catalan triangle entries at $(N-1, i)$, namely

$$T_i(N) = \frac{i+1}{N}\binom{2(N-1)-i}{N-1},$$

but we do not use this in our development.

Now, define an operation Δ that acts on a function of N by replacing N with $N-1$. That is, $\Delta f(N) = f(N-1)$. We prefer to use this operator, with the argument N suppressed, as a notational convenience for our formulas and figures (although all of our results can be obtained without it). We next prove recurrences for the S_i and T_i that will facilitate their computation.

Theorem 4.5. *We have*

$$T_i = T_{i-1} - \Delta T_{i-2},$$

with $T_1 = C_{N-1}$ *and* $T_2 = C_{N-1} - C_{N-2}$, *and*

$$S_i = i\ T_i + \Delta S_{i-1},$$

with $S_1 = C_{N-1}$.

Proof. See Figure 2 for a schematic illustrating these recurrences.

The recurrence for T follows because each path counted by $T_{i-1}(N)$ must end with a vertical step or a horizontal step; these are counted by $\Delta T_{i-2}(N) = T_{i-2}(N-1)$ and $T_i(N)$, respectively.

$N\backslash k$	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1
2											**50.0**
3										**60.0**	40.0
4									42.8	**57.1**	35.7
5								23.8	47.6	**54.7**	33.3
6							11.3	30.3	49.2	**53.0**	31.8
7						4.89	16.3	33.7	49.8	**51.7**	30.7
8					1.95	7.83	19.5	35.9	50.1	**50.7**	30.0
9				0.74	3.45	10.0	21.8	37.3	**50.2**	50.0	29.4
10			0.26	1.42	4.75	11.8	23.5	38.3	**50.2**	49.3	28.9
11		0.09	0.56	2.09	5.85	13.2	24.8	39.1	**50.1**	48.8	28.5
12	0.03	0.21	0.87	2.7	6.8	14.3	25.9	39.6	**50.1**	48.4	28.2
13	0.07	0.34	1.18	3.26	7.61	15.3	26.7	40.1	**50.0**	48.0	28.0
14	0.13	0.49	1.47	3.76	8.31	16.1	27.4	40.4	**50.0**	47.7	27.7
15	0.19	0.63	1.75	4.21	8.92	16.8	28.0	40.7	**49.9**	47.5	27.5
16	0.26	0.78	2.01	4.61	9.46	17.3	28.5	41.0	**49.9**	47.2	27.4
17	0.33	0.92	2.26	4.98	9.93	17.8	28.9	41.2	**49.8**	47.0	27.2
18	0.4	1.05	2.48	5.31	10.3	18.3	29.2	41.3	**49.7**	46.8	27.1
$\vdots$											
10^5	2.73	4.49	7.22	11.3	17.1	24.9	34.2	43.3	**48.4**	43.7	25.0

Table 2. Percentage of k-winnable 321-avoiding permutations of N.

The recurrence for S follows because each path counted by $S_i(N)$ passes through column $N-i$ at level $N-1$ or passes through column $N-i$ below level $N-1$. The first set of paths is counted by $iT_i(N)$ because any path ending at $(N-1-i, N-1)$ can be extended in i ways depending on which of the columns $N-i$, $N-i+2, \ldots,$ $N-1$ is used for the last vertical step. The second set of paths is counted by $\Delta S_{i-1}(N) = S_{i-1}(N-1)$ because we can bijectively extend any path passing the required column and ending at $(N-1, N-1)$ to end at (N, N) instead by inserting one more pair of vertical/horizontal steps at the last northeast corner. □

Using this theorem, we may write each S_i and T_i as a linear combination of Catalan numbers. On the one hand, applying Δ to S_i, say, simply restricts the Dyck paths we are counting to end at $(N-1, N-1)$ instead of (N, N). Algebraically, applying Δ replaces each Catalan number in the linear combination with the previous Catalan number.

Example 4.6. Applying the recurrences from Theorem 4.5, we have

$$T_3 = (C_{n-1} - C_{n-2}) - \Delta(C_{n-1}) = C_{n-1} - 2C_{n-2},$$

$$T_4 = (C_{n-1} - 2C_{n-2}) - \Delta(C_{n-1} - C_{n-2}) = C_{n-1} - 3C_{n-2} + C_{n-3},$$

$$T_5 = (C_{n-1} - 3C_{n-2} + C_{n-3}) - \Delta(C_{n-1} - 2C_{n-2}) = C_{n-1} - 4C_{n-2} + 3C_{n-3}$$

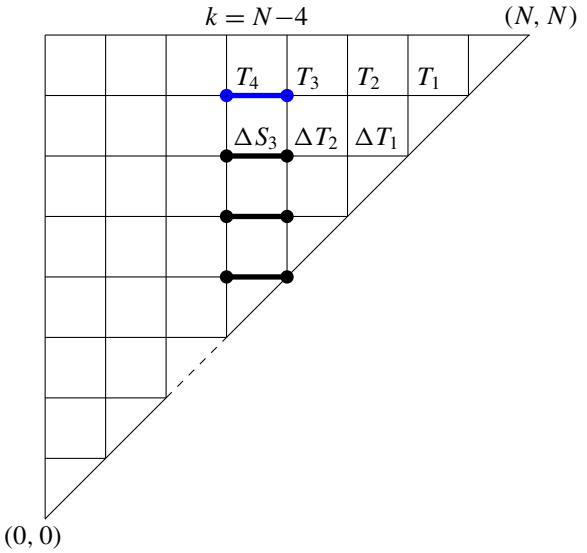


Figure 2. Schematic for path recurrences.

and

$$S_2 = 2(C_{n-1}-C_{n-2})+\Delta(C_{n-1}) = 2C_{n-1}-C_{n-2},$$

$$S_3 = 3(C_{n-1}-2C_{n-2})+\Delta(2C_{n-1}-C_{n-2}) = 3C_{n-1}-4C_{n-2}-C_{n-3},$$

$$S_4 = 4(C_{n-1}-3C_{n-2}+C_{n-3})+\Delta(3C_{n-1}-4C_{n-2}-C_{n-3}) = 4C_{n-1}-9C_{n-2}-C_{n-4},$$

$$\begin{aligned} S_5 &= 5(C_{n-1}-4C_{n-2}+3C_{n-3})+\Delta(4C_{n-1}-9C_{n-2}-C_{n-4}) \\ &= 5C_{n-1}-16C_{n-2}+6C_{n-3}-C_{n-5}. \end{aligned}$$

Lemma 4.7. *Let $i \le N-5$ and X_i be a linear combination of the Catalan numbers $C_{N-1}, C_{N-2}, \ldots, C_{N-i}$. Then,*

$$\frac{1}{4}\frac{X_i}{C_N} < \frac{\Delta X_i}{C_N} \le \frac{1}{3}\frac{X_i}{C_N}.$$

Proof. Observe that

$$\frac{1}{4} < \frac{C_{N-1}}{C_N} \le \frac{1}{3}$$

for all $N \ge 5$. Since

$$\frac{\Delta C_{N-i}}{C_N} = \frac{C_{N-i-1}}{C_N} = \frac{C_{N-i-1}}{C_{N-i}}\frac{C_{N-i}}{C_N},$$

we have

$$\frac{1}{4}\frac{C_{N-i}}{C_N} < \frac{\Delta C_{N-i}}{C_N} \le \frac{1}{3}\frac{C_{N-i}}{C_N}$$

for all $N-i \ge 5$, and the result follows by linearity. □

Lemma 4.8. *For all $i \leq N-5$, we have*

$$\frac{T_i}{C_N} \leq \frac{1}{3}\left(\frac{3}{4}\right)^{i-1}.$$

Proof. It is straightforward to verify that the result holds for $i = 1$ and $i = 2$. Suppose the result holds for $i-1$. Then,

$$\frac{T_i}{C_N} = \frac{T_{i-1}}{C_N} - \frac{\Delta T_{i-2}}{C_N} < \frac{T_{i-1}}{C_N} - \frac{1}{4}\frac{T_{i-2}}{C_N}$$

by Lemma 4.7. From their definition in terms of lattice paths, it is also clear that the T_i are decreasing in i (for each fixed N). Hence,

$$\frac{T_{i-1}}{C_N} - \frac{1}{4}\frac{T_{i-2}}{C_N} \leq \frac{T_{i-1}}{C_N} - \frac{1}{4}\frac{T_{i-1}}{C_N} = \frac{3}{4}\frac{T_{i-1}}{C_N} \leq \frac{1}{3}\left(\frac{3}{4}\right)^{i-1}$$

by induction. □

Theorem 4.9. *We have*

$$\frac{S_3}{C_N} > \frac{S_i}{C_N}$$

for all $N \geq 9$ and all $i > 3$.

Proof. We have

$$\frac{S_i}{C_N} = \frac{iT_i + \Delta S_{i-1}}{C_N} \leq \frac{i}{3}\left(\frac{3}{4}\right)^{i-1} + \frac{1}{3}\frac{S_{i-1}}{C_N}.$$

An exercise using calculus proves $\frac{i}{3}\left(\frac{3}{4}\right)^{i-1}$ is decreasing once $i > -1/\ln\left(\frac{3}{4}\right)$ (which is between 3 and 4) and $\frac{i}{3}\left(\frac{3}{4}\right)^{i-1}$ is less than $\frac{1}{4}$ for all $i \geq 11$. Consequently, once $S_i/C_N < \frac{3}{8}$, it remains so as i increases for all $i \geq 11$.

In fact, using the linear combinations of Catalan numbers obtained from Theorem 4.5 as in Example 4.6, we can verify that $S_i/C_N < \frac{3}{8}$ for all $5 \leq i \leq 11$ as illustrated in Table 2. More precisely, when we express S_i/C_N as a linear combination of ratios of Catalan numbers, the limiting value as $N \to \infty$ can be obtained by plugging in powers of $\frac{1}{4}$ for each ratio of Catalan numbers; as these limits are each smaller than $\frac{3}{8}$, we reduce to a finite computation. In detail, we use the bounds

$$0.25^j < \frac{C_{N-j}}{C_N} < 0.254^j$$

for $N > 95 + j$ to verify that $S_i/C_N < \frac{3}{8}$ for each of the linear combinations $i = 5, 6, \ldots, 11$ (and check remaining finite cases for N manually).

Thus, the optimal value of S_i/C_N must occur in $i \leq 4$ for all N. Using the formulas from Example 4.6 again, we then find that S_1/C_N is optimal for $N = 2$, that S_2/C_N is optimal for $3 \leq N \leq 8$, and that S_3/C_N is optimal for all $N \geq 9$. □

Corollary 4.10. *The optimal k-positional strategy for the game of best choice restricted to the* 321*-avoiding interview rank orders is*

$$k = \begin{cases} N-1 & \text{if } N = 2, \\ N-2 & \text{if } 3 \le N \le 8, \\ N-3 & \text{otherwise.} \end{cases}$$

The asymptotic probability of success is

$$\lim_{N\to\infty} \frac{3C_{N-1} - 4C_{N-2} - C_{N-3}}{C_N} = \frac{31}{64} = 0.484375.$$

Using André's reflection method or a straightforward induction argument, one can show that the number of partial Dyck paths (i.e., lying above the line $y = x$) from $(0, 0)$ to (a, b) (where $a < b$) is given by the formula

$$C_{(a,b)} = \binom{a+b}{a} \frac{b-a+1}{b+1}.$$

Using this, we can also give a direct count of the Dyck paths for which column k lies between the last two northeast corners of the path.

Theorem 4.11. *The probability that a* 321*-avoiding permutation of length N is k-winnable is*

$$\frac{1}{C_N} \sum_{i=1}^{N-k} \binom{(k-1)+(N-i)}{k-1} \frac{(N-k-i+2)}{(N-i)+1} (N-k-i+1).$$

Proof. Set $a = k-1$, and let b range over $k, k+1, k+2, \ldots, N-1$. Once the path passes through (a, b), there are $b-k+1$ ways to complete it so that it is k-winnable. □

5. Conclusions

It seems fair to say that these results are somewhat surprising and further investigation is warranted. The "bar-raising" model has a robust strategy but only allows a 25% success rate. The optimal strategy in the "disappointment-free" model reviews and rejects most of the applicants yet has a success rate that is close to 50%. Remarkably, these are not mutually exclusive and the $k = N-3$ positional strategy is asymptotically optimal in both models simultaneously.

Acknowledgements

This project was supported by the James Madison University Program of Grants for Faculty Assistance. We are grateful to an anonymous referee for several insightful suggestions on an earlier draft of this work.

References

[Bóna 2012] M. Bóna, *Combinatorics of permutations*, 2nd ed., CRC Press, Boca Raton, FL, 2012. MR Zbl

[Deutsch et al. 2002/03] E. Deutsch, A. J. Hildebrand, and H. S. Wilf, "Longest increasing subsequences in pattern-restricted permutations", *Electron. J. Combin.* **9**:2 (2002/03), art. id. 12. MR Zbl

[Firro et al. 2007] G. Firro, T. Mansour, and M. C. Wilson, "Longest alternating subsequences in pattern-restricted permutations", *Electron. J. Combin.* **14**:1 (2007), art. id. 34. MR Zbl

[Gardner 1995] M. Gardner, *New mathematical diversions*, revised ed., Math. Assoc. Amer., Washington, DC, 1995. MR Zbl

[Gilbert and Mosteller 1966] J. P. Gilbert and F. Mosteller, "Recognizing the maximum of a sequence", *J. Amer. Statist. Assoc.* **61** (1966), 35–73. MR

[Kadison 1994] R. V. Kadison, "Strategies in the secretary problem", *Exposition. Math.* **12**:2 (1994), 125–144. MR Zbl

[Madras and Pehlivan 2016] N. Madras and L. Pehlivan, "Structure of random 312-avoiding permutations", *Random Structures Algorithms* **49**:3 (2016), 599–631. MR Zbl

[Miner and Pak 2014] S. Miner and I. Pak, "The shape of random pattern-avoiding permutations", *Adv. in Appl. Math.* **55** (2014), 86–130. MR Zbl

[Prodinger 2002] H. Prodinger, "Combinatorics of geometrically distributed random variables: value and position of large left-to-right maxima", *Discrete Math.* **254**:1-3 (2002), 459–471. MR Zbl

[Wilf 1995] H. S. Wilf, "On the outstanding elements of permutations", preprint, 1995, available at https://tinyurl.com/wilfoutst.

Received: 2018-02-25 Revised: 2018-10-23 Accepted: 2018-12-04

afowlkes@math.sc.edu *Department of Mathematics, University of South Carolina, Columbia, SC, United States*

jones3bc@jmu.edu *Department of Mathematics and Statistics, James Madison University, Harrisonburg, VA, United States*

msp

INVOLVE 12:4 (2019)
dx.doi.org/10.2140/involve.2019.12.659

Graphs with at most two trees in a forest-building process

Steve Butler, Misa Hamanaka and Marie Hardt

(Communicated by Glenn Hurlbert)

Given a graph, we can form a spanning forest by first sorting the edges in a random order, and then only keeping edges incident to a vertex which is not incident to any previous edge. The resulting forest is dependent on the ordering of the edges, and so we can ask, for example, how likely is it for the process to produce a graph with k trees.

We look at all graphs which can produce at most two trees in this process and determine the probabilities of having either one or two trees. From this we construct infinite families of graphs which are nonisomorphic but produce the same probabilities.

1. Introduction

We consider the following *forest-building process*:

(1) Take all of the edges of the graph, remove them and sort them in a random order.

(2) Go through the edges in this order and only put those edges back in which connect to some vertex not previously seen by any edge.

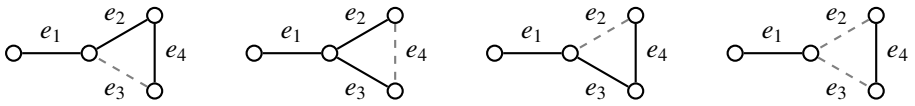
From this, we must end up with a forest (a graph without cycles) since we can never add an edge that closes a cycle. As an example, in Figure 1 we list all 24 different ways to order the edges and group them based on the resulting forest formed.

We will consider this problem: how many different edge orderings produce a given number, say k, of trees in the resulting graph? Equivalently, what is the probability that if we take a random ordering of the edges, we produce a forest with k trees? We will let $P(G, k)$ denote this probability. As an example when G is the paw graph, we see that $P(G, 1) = \frac{5}{6}$ and $P(G, 2) = \frac{1}{6}$ (see Figure 1).

This process was implicitly used in [Butler et al. 2015] for the complete graph, and explicitly introduced in [Berikkyzy et al. 2018], where some basic properties

MSC2010: 05C05.

Keywords: forests, edge ordering, components, probability.



(e_1, e_2, e_4, e_3)	(e_1, e_2, e_3, e_4)	(e_1, e_3, e_4, e_2)	(e_1, e_4, e_2, e_3)
(e_2, e_1, e_4, e_3)	(e_1, e_3, e_2, e_4)	(e_3, e_1, e_4, e_2)	(e_1, e_4, e_3, e_2)
(e_2, e_4, e_1, e_3)	(e_2, e_1, e_3, e_4)	(e_3, e_4, e_1, e_2)	(e_4, e_1, e_2, e_3)
(e_2, e_4, e_3, e_1)	(e_2, e_3, e_1, e_4)	(e_3, e_4, e_2, e_1)	(e_4, e_1, e_3, e_2)
(e_4, e_2, e_1, e_3)	(e_2, e_3, e_4, e_1)	(e_4, e_3, e_1, e_2)	
(e_4, e_2, e_3, e_1)	(e_3, e_1, e_2, e_4)	(e_4, e_3, e_2, e_1)	
	(e_3, e_2, e_1, e_4)		
	(e_3, e_2, e_4, e_1)		

Figure 1. The results from different edge orderings of the paw graph.

were established and the probabilities for complete bipartite graphs were determined. We summarize these results here.

Theorem 1 [Butler et al. 2015]. *We have*

$$P(K_n, k) = \frac{\binom{n-1}{n-2k,k,k-1}2^{n-2k}}{\binom{2n-2}{n}}.$$

Theorem 2 [Berikkyzy et al. 2018]. *We have*

$$P(K_{s,t}, k) = \frac{(s+t)\binom{s}{k}\binom{t}{k}}{st\binom{s+t}{s}}.$$

For small graphs (at most five vertices), the probabilities are given in [Berikkyzy et al. 2018]. There are a few instances where two graphs would have the same probabilities for all k, and most of those are edge-transitive graphs. More generally, the following was observed.

Lemma 3 [Berikkyzy et al. 2018]. *If G is an edge-transitive graph with minimum degree of at least 2 and e is any edge, then we have $P(G, k) = P(G - e, k)$ for all k.*

In essence, this follows by noting that the last edge in an ordering is never kept, *and* by symmetry every edge is the last edge in an ordering equally often.

The goal of this note is to compute the probabilities for more families of graphs, namely graphs which can produce at most two trees in the forest-building process. Using this, we will produce infinitely many examples of nonisomorphic graphs G and H where the probabilities agree and neither G or H are edge-transitive.

2. Graphs with at most two trees

We are interested in exploring the graphs which can produce at most two trees in the forest-building process. Equivalently, this means that there are at most two disjoint edges in the graph (disjoint in the sense that they share no vertex).

Proposition 4. *The only nonempty graphs without isolated vertices, which contain no pair of disjoint edges, are star graphs* ($K_{1,n}$) *and the triangle graph* (K_3).

Proof. If the graph is not connected, then taking one edge from two different components gives two disjoint edges. So we may assume the graph is connected.

If the graph has two disjoint edges, then we can connect these together by a path, creating a path with at least four vertices. Conversely, if the graph has a path with at least four vertices, then it must contain two disjoint edges. So we can conclude the longest path is a path with at most three vertices. If the longest path has two vertices, then the graph is a K_2.

If the longest path has three vertices and the ends of the path are not leaves, then it must be that the ends connect and form a triangle. Since the paw graph has two disjoint edges, this can only happen if the graph is a K_3.

Finally, if the graph is not a triangle and doesn't have a path of length 4 (and hence has no cycles), then it must be a star. □

Proposition 5. *If a graph without isolated vertices has a vertex v of degree at least* 5 *and contains no set of three disjoint edges, then deleting v and all incident edges, and removing any isolated vertices results in either an empty graph, a star, or a K_3.*

Proof. Let v be a vertex of degree at least 5 in the graph. Suppose that the graph resulting from deleting the vertex v and all incident edges, and removing any isolated vertices, is not the empty graph, a star, or a K_3. Proposition 4 states that the only nonempty graphs without isolated vertices, that contain no set of two disjoint edges, are the star graph and K_3. Thus the resulting graph will contain at least two disjoint edges. Call these edges e_1 and e_2; note neither of these edges are incident to v.

At most four edges incident to v are also incident to the edges e_1 and e_2. Since v has degree at least 5, this leaves at least one edge, e_3, that is connected to v and not incident to e_1 or e_2. Thus the original graph contains a set of three disjoint edges. □

Finally, we observe that if all the degrees are bounded by at most 4 and the graph is connected, then as n gets large, the diameter must also grow — which forces three disjoint edges. In particular there are finitely many graphs with maximum degree at most 4 which produce at most two trees.

Putting this all together, we see that for n sufficiently large ($n \geq 6$; verified computationally) the only connected graphs which produce at most two trees, and are not stars, are the following five families, shown in Figure 2:

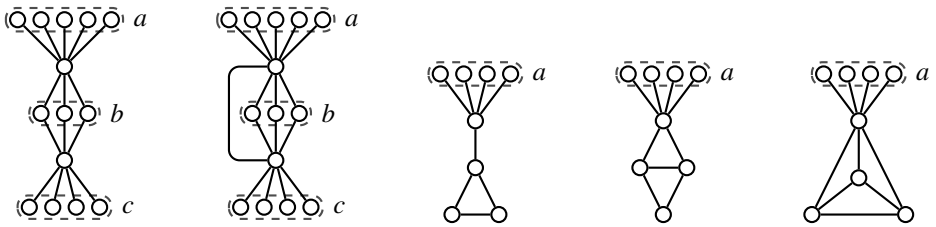


Figure 2. The five families of graphs.

- GS$_{a,b,c}$: the stars $K_{1,a+b}$ and $K_{1,b+c}$ which have b leaves glued together (glued stars).
- GS$^{+}_{a,b,c}$: the stars $K_{1,a+b}$ and $K_{1,b+c}$ which have b leaves glued together and the centers joined by an edge (glued stars with an edge).
- Paw$_a$: the paw graph with a leaves appended to the vertex of degree 1.
- Di$_a$: the diamond graph (a four-cycle with an extra edge) with a leaves appended to one of the vertices of degree 2.
- $(K_4)_a$: the complete graph on four vertices with a leaves appended to one of the vertices.

Note that when the degree is at least 5, the first two of these correspond to Proposition 5 where the remaining graph is a star and the last three of these correspond to Proposition 5 where the remaining graph is a K_3.

3. Computing probabilities for the families

We now turn our attention to computing the probabilities that a graph ends in one or two trees in the forest-building process. We can find these probabilities by noting that if there are m edges in the graph, then the probability that we end with two trees is

$$P(G, 2) = \frac{|\{\text{rearrangements with two trees}\}|}{m!}.$$

We will focus on counting the rearrangements which produce two trees. Particularly, we want to count rearrangements where an edge occurs exactly once, not at the start, and involves two vertices which have not been previously seen.

Since we will be counting rearrangements, we will find it useful to know how to manipulate binomial coefficients. Recall that

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

is the number of ways to choose k elements (in our case this will usually be locations) out of an n element set. There are many binomial coefficient identities (see the book of Graham, Knuth, and Patashnik [Graham et al. 1994, Chapter 5] for a good introduction); we will need to make repeated use of the following well-known result.

Proposition 6. *We have*

$$\sum_j \binom{\ell - j}{m}\binom{q+j}{n} = \binom{\ell+q+1}{m+n+1}, \tag{1}$$

where the sum ranges over all values where the summands are nonzero.

Theorem 7. *We have the following probabilities*:

$$P(\mathrm{GS}_{a,b,c}, 1) = \frac{b}{(b+c+1)(b+c)} + \frac{b}{(a+b+1)(a+b)},$$

$$P(\mathrm{GS}^{+}_{a,b,c}, 1) = \frac{2b+c+2}{(b+c+1)(b+c+2)} + \frac{2b+a+2}{(b+a+1)(b+a+2)} - \frac{1}{a+2b+c+1}.$$

Proof. Since $P(\mathrm{GS}_{a,b,c}, 1) + P(\mathrm{GS}_{a,b,c}, 2) = 1$, we can focus on computing the probability of resulting in two trees. We now claim

$$\begin{aligned} &P(\mathrm{GS}_{a,b,c}, 2) \\ &= \sum_{i=0}^{b}\sum_{j=0}^{a} \frac{\binom{a}{j}\binom{b}{i}(i+j)!\,(b+c-i)(a+2b+c-i-j-1)!}{(a+2b+c)!} - \frac{b+c}{a+2b+c} \\ &\quad + \sum_{i=0}^{b}\sum_{j=0}^{c} \frac{\binom{c}{j}\binom{b}{i}(i+j)!\,(a+b-i)(a+2b+c-i-j-1)!}{(a+2b+c)!} - \frac{a+b}{a+2b+c}. \end{aligned} \tag{2}$$

This comes from the two cases, namely where our first edge initially comes from the "top half" (i.e., edges coming from the star $K_{1,a+b}$), and where our first edge initially comes from the "bottom half" (i.e., edges coming from the star $K_{1,b+c}$). We focus on the top-half case, as the bottom half follows by an identical argument by interchanging the roles of a and c.

Determining if we have two trees comes down to what happens when we pick our first edge from the star $K_{1,b+c}$. We look at all ways that this occurs by first picking edges from $K_{1,a+b}$ and then considering what happens when we pick our edge from $K_{1,b+c}$. In particular, we will pick j edges from the a leaf vertices and i edges from the b gluing vertices. We now run over all possibilities for i and j.

For each choice of edges we now consider all possible orderings as follows:

- $\binom{a}{j}$ corresponds to which of the j edges among the a were chosen.
- $\binom{b}{i}$ corresponds to which of the i edges among the b were chosen.

- $(i+j)!$ indicates how many ways to order these $i+j$ edges (note that these $i+j$ edges are all of the initial edges).
- $(b+c-i)$ indicates how many edges disjoint from the ones above are available to choose, if we want to create two trees.
- $(a+2b+c-i-j-1)!$ is the number of ways to rearrange the remaining edges.

This gives all orderings of edges possible; to get the probability we now divide by the total number of orderings, which is $(a+2b+c)!$.

Note that in the summation we need to correct for $i=0$, $j=0$, which does not fall into the case where the first edge is from the top. So we subtract this term, which gives the $-(b+c)/(a+2b+c)$ at the end.

To now simplify these sums we can repeatedly apply (1). So we have

$$\begin{aligned}
&\sum_{i=0}^{b}\sum_{j=0}^{a}\frac{\binom{a}{j}\binom{b}{i}(i+j)!\,(b+c-i)(a+2b+c-i-j-1)!}{(a+2b+c)!}\\
&\quad=\sum_{i=0}^{b}\sum_{j=0}^{a}\frac{\frac{a!}{j!\,(a-j)!}\frac{b!}{i!\,(b-i)!}(i+j)!\,(b+c-i)(a+2b+c-i-j-1)!}{(a+2b+c)!}\\
&\quad=\sum_{i=0}^{b}\frac{a!\,b!\,(b+c-i)}{(a+2b+c)!\,(b-i)!}\sum_{j=0}^{a}\frac{(i+j)!}{i!\,j!}\frac{(a+2b+c-i-j-1)!}{(a-j)!}\\
&\quad=\sum_{i=0}^{b}\frac{a!\,b!\,(b+c-i)(2b+c-i-1)!}{(a+2b+c)!\,(b-i)!}\sum_{j=0}^{a}\frac{(i+j)!}{i!\,j!}\frac{(a+2b+c-i-j-1)!}{(a-j)!\,(2b+c-i-1)!}\\
&\quad=\sum_{i=0}^{b}\frac{a!\,b!\,(b+c-i)(2b+c-i-1)!}{(a+2b+c)!\,(b-i)!}\sum_{j=0}^{a}\binom{i+j}{i}\binom{a+2b+c-i-1-j}{2b+c-i-1}\\
&\quad=\sum_{i=0}^{b}\frac{a!\,b!\,(b+c-i)(2b+c-i-1)!}{(a+2b+c)!\,(b-i)!}\binom{a+2b+c}{2b+c}\\
&\quad=\sum_{i=0}^{b}\frac{a!\,b!\,(b+c-i)(2b+c-i-1)!}{(a+2b+c)!\,(b-i)!}\frac{(a+2b+c)!}{(2b+c)!\,a!}\\
&\quad=\sum_{i=0}^{b}\frac{b!\,(b+c-i)(2b+c-i-1)!}{(b-i)!\,(2b+c)!}\\
&\quad=\frac{b!\,(b+c-1)!}{(2b+c)!}\sum_{i=0}^{b}\frac{(2b+c-i-1)!}{(b-i)!\,(b+c-1)!}(b+c-i)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{b!\,(b+c-1)!}{(2b+c)!}\left((b+c)\sum_{i=0}^{b}\binom{2b+c-1-i}{b+c-1}\binom{i}{0}-\sum_{i=0}^{b}\binom{2b+c-1-i}{b+c-1}\binom{i}{1}\right)\\
&= \frac{b!\,(b+c-1)!}{(2b+c)!}\left((b+c)\binom{2b+c}{b+c}-\binom{2b+c}{b+c+1}\right)\\
&= \frac{b!\,(b+c-1)!}{(2b+c)!}\left((b+c)\frac{(2b+c)!}{(b+c)!\,b!}-\frac{(2b+c)!}{(b+c+1)!\,(b-1)!}\right)\\
&= 1-\frac{b}{(b+c+1)(b+c)}.
\end{aligned}
$$

By a similar process, the other double sum becomes

$$
\sum_{i=0}^{b}\sum_{j=0}^{c}\frac{\binom{c}{j}\binom{b}{i}(i+j)!\,(a+b-i)(a+2b+c-i-j-1)!}{(a+2b+c)!}=1-\frac{b}{(b+a+1)(b+a)}.
$$

Now replacing the double sums by these simplified expressions we have

$$
\begin{aligned}
&P(\mathrm{GS}_{a,b,c},2)\\
&\quad=\left(1-\frac{b}{(b+c+1)(b+c)}\right)-\frac{b+c}{a+2b+c}+\left(1-\frac{b}{(b+a+1)(b+a)}\right)-\frac{a+b}{a+2b+c}\\
&\quad=1-\frac{b}{(b+c+1)(b+c)}-\frac{b}{(b+a+1)(b+a)}.
\end{aligned}
$$

Finally we note

$$
P(\mathrm{GS}_{a,b,c},1)=1-P(\mathrm{GS}_{a,b,c},2)=\frac{b}{(b+c+1)(b+c)}+\frac{b}{(b+a+1)(b+a)},
$$

establishing the result for $\mathrm{GS}_{a,b,c}$.

The result for $P(\mathrm{GS}^{+}_{a,b,c},2)$ follows by a similar argument, the only difference being the additional edge which *cannot* be used in order to result in two trees. So (2) would now become

$$
\begin{aligned}
P(\mathrm{GS}^{+}_{a,b,c},2)=&\sum_{i=0}^{b}\sum_{j=0}^{a}\frac{\binom{a}{j}\binom{b}{i}(i+j)!\,(b+c-i)(a+2b+c-i-j)!}{(a+2b+c+1)!}-\frac{b+c}{a+2b+c+1}\\
&+\sum_{i=0}^{b}\sum_{j=0}^{c}\frac{\binom{c}{j}\binom{b}{i}(i+j)!\,(a+b-i)(a+2b+c-i-j)!}{(a+2b+c+1)!}-\frac{a+b}{a+2b+c+1}.
\end{aligned}
$$

The rest of the argument works in the same way as before. □

Our approach was to focus on the probability of producing two trees. It is possible to establish the result by focusing on one tree, and the following proof was communicated to us by a referee of the paper.

Alternative proof of Theorem 7. To compute the probability of producing a single tree in the graph $\mathrm{GS}_{a,b,c}$ we focus on finding the probability of producing a "bridge", a path of length 2 connecting the vertices v_a and v_c (the vertices with a and b leaves respectively). Label the b vertices which connect to both v_a, and v_c as $v_1, \ldots, v_b$.

The probability that we first pick an edge incident to a *and* form a bridge going through v_i is

$$\frac{1}{(b+c+1)(b+c)}.$$

To see this, once we have picked the first edge there are $b+c+1$ important edges remaining, namely the $b+c$ edges incident to v_c and the edge $\{v_a, v_i\}$. In order to form the bridge among these $b+c+1$ edges we must first choose $\{v_a, v_i\}$ (probability $1/(b+c+1)$) and must second choose $\{v_i, v_c\}$ (probability $1/(b+c)$), establishing the above probability. This is independent of v_i and so going over all b of the v_i we have that the probability that we first pick an edge incident to a and form a bridge is

$$\frac{b}{(b+c+1)(b+c)}.$$

A symmetrical argument gives that the probability that we first pick an edge incident to c and form a bridge is

$$\frac{b}{(b+a+1)(b+a)}.$$

Finally these events are disjoint and cover all ways to form a bridge. So we can conclude that the probability of forming a bridge, and hence the probability of having one component, is

$$\frac{b}{(b+a+1)(b+a)} + \frac{b}{(b+c+1)(b+c)}.$$

A similar approach works for $\mathrm{GS}^{+}_{a,b,c}$ once we also account for the edge $\{v_a, v_c\}$ being a bridge. We leave the details of this to the interested reader. □

Theorem 8. *We have the following probabilities*:

$$P(\mathrm{Paw}_a, 1) = \frac{1}{6} - \frac{1}{a+3} + \frac{1}{a+1},$$
$$P(\mathrm{Di}_a, 1) = \frac{3}{10} - \frac{2}{a+4} + \frac{2}{a+2},$$
$$P((K_4)_a, 1) = \frac{2}{5} - \frac{3}{a+5} + \frac{3}{a+3}.$$

Proof. We will again compute the probability that there are two trees in the process. However, in these cases there are many more possibilities to consider. To simplify the situation, we make the following observation: every edge which is a leaf in the original graph will always be kept in the forest-building process. This indicates if there

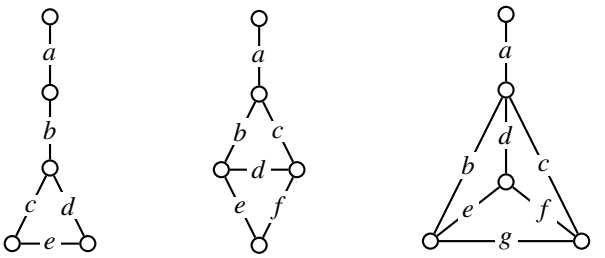


Figure 3. The remaining three graphs with the leaves collapsed to a single edge a.

are multiple leaves off of a single vertex v, then we only need to know when the *first* leaf was chosen. This is because, after first leaf, v will have been seen by some edge.

So we now represent the remaining graphs from Figure 2, as shown in Figure 3, where a corresponds to all of the a leaves condensed down, and the remaining edges are labeled as indicated with each label other than a corresponding to a single edge.

For each graph, we now look at all possible ways to start selecting edges and end with a pair of disjoint edges. We also find the probability of starting our selection in a particular way. Recall that an edge marked a corresponds to a different edges, and so until we select that edge, we assume all a of them haven't been seen and are available for picking; after selection, by the observation we can assume they have all been seen. (In other words, it is only the relative ordering of the different types of edges that matter.)

For the paw graph, we have the possibilities shown in Table 1 (the first column indicates every possible sequence of choices of edges until two trees are formed, while the second column indicates the probability of any one of those sequence of

start of edge orderings resulting in two trees	probabilities of an ordering
ac, ad, ae	$\frac{a}{a+4} \cdot \frac{1}{4}$
be, eb	$\frac{1}{a+4} \cdot \frac{1}{a+3}$
ca, da, ea	$\frac{1}{a+4} \cdot \frac{a}{a+3}$
abe	$\frac{a}{a+4} \cdot \frac{1}{4} \cdot \frac{1}{3}$
bae	$\frac{1}{a+4} \cdot \frac{a}{a+3} \cdot \frac{1}{3}$
$cda, cea, dca, dea, eda, eca$	$\frac{1}{a+4} \cdot \frac{1}{a+3} \cdot \frac{a}{a+2}$
$ceda, cdea, dcea, deca, ecda, edca$	$\frac{1}{a+4} \cdot \frac{1}{a+3} \cdot \frac{1}{a+2} \cdot \frac{a}{a+1}$

Table 1. Probabilities associated with Paw$_a$.

start of edge orderings resulting in two trees	probabilities of an ordering
ad, ae, af	$\frac{a}{a+5} \cdot \frac{1}{5}$
bf, ce, ec, fb	$\frac{1}{a+5} \cdot \frac{1}{a+4}$
da, ea, fa	$\frac{1}{a+5} \cdot \frac{a}{a+4}$
abf, ace	$\frac{a}{a+5} \cdot \frac{1}{5} \cdot \frac{1}{4}$
baf, cae	$\frac{1}{a+5} \cdot \frac{a}{a+4} \cdot \frac{1}{4}$
$dea, dfa, eda, efa, fda, fea$	$\frac{1}{a+5} \cdot \frac{1}{a+4} \cdot \frac{a}{a+3}$
$defa, dfea, edfa, efda, fdea, feda$	$\frac{1}{a+5} \cdot \frac{1}{a+4} \cdot \frac{1}{a+3} \cdot \frac{a}{a+2}$

Table 2. Probabilities associated with Di_a.

start of edge orderings resulting in two trees	probabilities of an ordering
ae, af, ag	$\frac{a}{a+6} \cdot \frac{1}{6}$
bf, ce, dg, ec, fb, gd	$\frac{1}{a+6} \cdot \frac{1}{a+5}$
ea, fa, ga	$\frac{1}{a+6} \cdot \frac{a}{a+5}$
abf, adg, ace	$\frac{a}{a+6} \cdot \frac{1}{6} \cdot \frac{1}{5}$
baf, dag, cae	$\frac{1}{a+6} \cdot \frac{a}{a+5} \cdot \frac{1}{5}$
$efa, ega, fea, fga, gea, gfa$	$\frac{1}{a+6} \cdot \frac{1}{a+5} \cdot \frac{a}{a+4}$
$efga, egfa, fega, fgea, gefa, gfea$	$\frac{1}{a+6} \cdot \frac{1}{a+5} \cdot \frac{1}{a+4} \cdot \frac{a}{a+3}$

Table 3. Probabilities associated with $(K_4)_a$.

choices being made). If we now sum all of these probabilities together, we get

$$P(\mathrm{Paw}_a, 2) = \frac{5}{6} + \frac{1}{a+3} - \frac{1}{a+1},$$

establishing the result (recall that $P(\mathrm{Paw}_a, 1) + P(\mathrm{Paw}_a, 2) = 1$).

The results for the remaining two graphs are established in the same way and the corresponding probabilities are given in Tables 2 and 3. □

4. Examples of graphs with the same probabilities

Using the formulas from the theorems in the preceding section, we can now compute the probabilities for a large number of these graphs efficiently. In particular, we examined all graphs up through 500 vertices in these families, and discovered several examples of families of nonisomorphic graphs which produce the same probabilities.

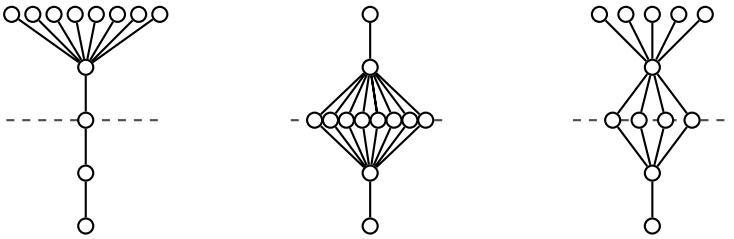


Figure 4. Examples of a set of three graphs with the same probabilities from Proposition 9. In this case $s = t = 1$.

Proposition 9. *Given $s, t \geq 1$ with s dividing $2t(t+1)$, let $r = 2t(t+1)/s$. Then we have for all k*

$$P(\mathrm{GS}_{r+3t+1,s,t}, k) = P(\mathrm{GS}_{t,r+s+2t+1,t}, k) = P(\mathrm{GS}_{3t+s+1,r,t}, k).$$

This immediately follows by applying the formulas for the probabilities from Theorem 7.

Proposition 10. *Given $t \geq 1$, we have for all k*

$$P(\mathrm{GS}_{5t+3,t,2t}, k) = P(\mathrm{GS}_{5t+1,t+1,2t+1}, k).$$

This also immediately follows by applying the formulas for probabilities from Theorem 7. We note that there were many other examples of pairs of glued star graphs which are not explained by Propositions 9 and 10. A complete characterization of all such pairs of glued stars remains elusive.

Looking beyond glued stars, we found very few pairs of graphs with the same probabilities and the results do not seem to fit any patterns. As an example, all pairs of graphs from the $\mathrm{GS}^+_{a,b,c}$ family up through 500 vertices with the same probabilities are listed below (it is possible that this is a complete list for this family; showing this would relate to solving a system of diophantine equations).

$$P(\mathrm{GS}^+_{17,3,9}, k) = P(\mathrm{GS}^+_{10,9,10}, k), \qquad P(\mathrm{GS}^+_{28,5,9}, k) = P(\mathrm{GS}^+_{26,8,8}, k),$$
$$P(\mathrm{GS}^+_{103,15,48}, k) = P(\mathrm{GS}^+_{63,71,32}, k), \quad P(\mathrm{GS}^+_{95,23,53}, k) = P(\mathrm{GS}^+_{53,66,52}, k).$$

5. Conclusion

We found the probabilities for all connected graphs which can form at most two trees in this forest-building process. A natural next step is to consider graphs with at most three trees. As an example, two pairs of graphs with at most three trees and matching probabilities are given in Figure 5. This is suggestive that these are the start of an infinite family of such graphs, but we have not yet established this. One difficulty is that unlike the situation for two trees where only one probability

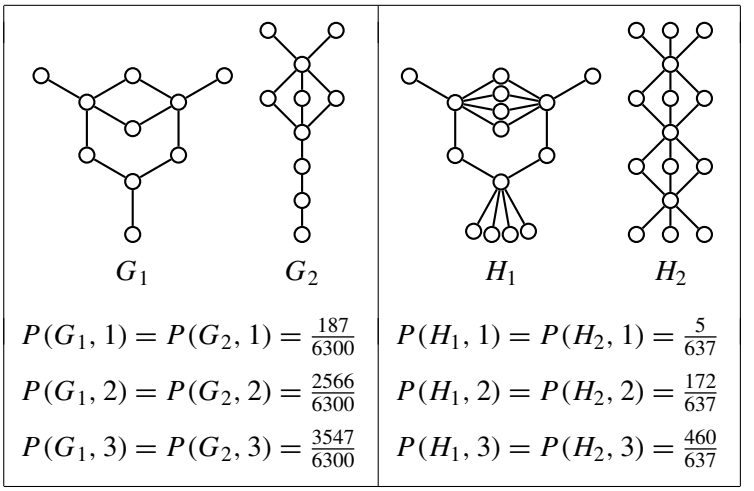


Figure 5. Two pairs of graphs with at most three trees and producing the same probabilities.

needed to be computed (since the probabilities sum to 1), this requires that two probabilities be computed.

Acknowledgements

The authors are grateful to the referees for their thoughtful reading of the paper and many suggestions for improvement, including the alternative proof of Theorem 7.

References

[Berikkyzy et al. 2018] Z. Berikkyzy, S. Butler, J. Cummings, K. Heysse, P. Horn, R. Luo, and B. Moran, "A forest building process on simple graphs", *Discrete Math.* **341**:2 (2018), 497–507. MR Zbl

[Butler et al. 2015] S. Butler, F. Chung, J. Cummings, and R. Graham, "Edge flipping in the complete graph", *Adv. in Appl. Math.* **69** (2015), 46–64. MR Zbl

[Graham et al. 1994] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete mathematics: a foundation for computer science*, 2nd ed., Addison-Wesley, Reading, MA, 1994. MR Zbl

Received: 2018-03-30 Revised: 2018-09-10 Accepted: 2018-10-28

butler@iastate.edu *Department of Mathematics, Iowa State University, Ames, IA, United States*

hamanaka@iastate.edu *Department of Mathematics, Iowa State University, Ames, IA, United States*

hardtme@iastate.edu *Department of Mathematics, Iowa State University, Ames, IA, United States*

INVOLVE 12:4 (2019)
dx.doi.org/10.2140/involve.2019.12.671

Log-concavity of Hölder means and an application to geometric inequalities

Aurel I. Stan and Sergio D. Zapeta-Tzul

(Communicated by Sever S. Dragomir)

The log-concavity of the Hölder mean of two numbers, as a function of its index, is presented first. The notion of α-cevian of a triangle is introduced next, for any real number α. We use this property of the Hölder mean to find the smallest index $p(\alpha)$ such that the length of an α-cevian of a triangle is less than or equal to the $p(\alpha)$-Hölder mean of the lengths of the two sides of the triangle that are adjacent to that cevian.

1. Introduction

All parts of mathematics are interconnected, including two important branches, geometry and analysis. Continuity, which is a fundamental notion in real analysis, is used in Euclidean geometry as one axiom in Hilbert axiomatization, and in proving Thales' theorem for irrational ratios. On the other hand, geometry helps real analysis by providing pictures that help us understand certain theorems. For example, Euler's theorem, which says that in any parallelogram the sum of the squares of the lengths of its sides is equal to the sum of the squares of its diagonals, provides a visual representation for the parallelogram identity that characterizes the norms of inner product spaces.

There is an abundant literature of geometric inequalities concerning important line segments in a triangle; see [Bottema et al. 1969; Mitrinović et al. 1989], for example. Some of these inequalities improve previously existing inequalities.

In this paper we present an application of the log-concavity of the Hölder mean with positive index, of two numbers, to find sharp inequalities relating lengths of cevians and sides of a triangle. Using these inequalities we find the best possible index for the Hölder mean, in a certain sense.

The paper is divided as follows:

In Section 2, we prove that the Hölder mean of two positive numbers, viewed as a function of its index, is logarithmically concave on $[0, \infty)$. In Section 3,

MSC2010: 26A06, 26D99.
Keywords: Hölder mean, log-concavity, Jensen inequality, triangle, cevian.

we define the notion of an α-cevian in a triangle, and find the smallest index $p(\alpha)$ such that the length of every α-cevian is less than or equal to the $p(\alpha)$-Hölder mean of the lengths of the two sides of the triangle that are adjacent to that cevian.

2. Log-concavity of Hölder means

Let a and b be two positive numbers. For any $p \in [-\infty, \infty]$, we define the p-Hölder mean of a and b, as

$$H_p(a,b) := \begin{cases} \left(\frac{1}{2}a^p + \frac{1}{2}b^p\right)^{1/p} & \text{if } p \in \mathbb{R} \setminus \{0\}, \\ \lim_{p\to 0} H_p(a,b) = \sqrt{ab} & \text{if } p = 0, \\ \lim_{p\to -\infty} H_p(a,b) = \min\{a,b\} & \text{if } p = -\infty, \\ \lim_{p\to \infty} H_p(a,b) = \max\{a,b\} & \text{if } p = \infty. \end{cases} \tag{2-1}$$

It follows from Jensen's inequality that for all $-\infty \le p < q \le \infty$, we have

$$H_p(a,b) \le H_q(a,b), \tag{2-2}$$

and this inequality is strict if $a \ne b$; see [Bullen 1998; Bullen et al. 1988; Pólya and Szegő 1972].

We prove now that the Hölder mean of two positive numbers, viewed as a function of its index, is logarithmically concave on $[0, \infty)$.

Lemma 2.1. *For all positive numbers a and b, the function $f : [0,\infty) \to \mathbb{R}$, defined by*

$$f(x) := \ln(H_x(a,b)), \tag{2-3}$$

is concave downward.

Proof. If $a = b$, then the lemma is obvious since f is a constant function, and its value is $f(x) = \ln(a)$ for all x in $[0, \infty)$.

Let us assume now that $0 < a < b$. Then, defining $c := \frac{b}{a} > 1$ for all $x \ge 0$, we have

$$f(x) = \ln(H_x(a,b)) = \ln\left(a H_x\left(1, \frac{b}{a}\right)\right) = \ln(H_x(1,c)) + \ln(a).$$

Thus the graph of f is just a vertical translation by $\ln(a)$ of the graph of $g : [0,\infty) \to \mathbb{R}$, defined by

$$g(x) = \ln(H_x(1,c)). \tag{2-4}$$

Therefore, it suffices to show that g is concave downward on $[0, \infty)$.

We know that g is continuous on $[0, \infty)$, and so to achieve our goal we need to prove that the second derivative of g is negative on $(0, \infty)$.

Indeed, if $'$ denotes the derivative with respect to x, then we have

$$g'(x) = \frac{d}{dx}\left[\frac{1}{x}\ln(1+c^x) - \frac{1}{x}\ln(2)\right]$$
$$= -\frac{1}{x^2}\ln(1+c^x) + \frac{1}{x}\frac{c^x\ln(c)}{1+c^x} + \frac{\ln(2)}{x^2}. \tag{2-5}$$

Differentiating one more time, we obtain

$$g''(x) = \frac{2}{x^3}\ln(1+c^x) - \frac{2}{x^2}\frac{c^x\ln(c)}{1+c^x} + \frac{1}{x}\frac{c^x\ln^2(c)}{(1+c^x)^2} - \frac{2\ln(2)}{x^3}. \tag{2-6}$$

We make now the change of variable

$$y := c^x \in (1, \infty), \tag{2-7}$$

which means

$$x = \frac{\ln(y)}{\ln(c)}. \tag{2-8}$$

Substituting back in the formula of $g''(x)$, we obtain

$$g''(x) = \frac{2\ln^3(c)}{\ln^3(y)}\ln(1+y) - \frac{2\ln^2(c)}{\ln^2(y)}\frac{y\ln(c)}{1+y} + \frac{\ln(c)}{\ln(y)}\frac{y\ln^2(c)}{(1+y)^2} - \frac{2\ln(2)\ln^3(c)}{\ln^3(y)}. \tag{2-9}$$

Thus, to show that, for all $x > 0$, we have $g''(x) < 0$, by multiplying both sides by the positive number $(1+y)^2\ln^3(y)/\ln^3(c)$, we have to prove that for all $y > 1$

$$h(y) := 2(1+y)^2\ln(1+y) - 2y(1+y)\ln(y) + y\ln^2(y) - 2(1+y)^2\ln(2) \tag{2-10}$$

is negative.

The function h is defined even for $y = 1$, and we have $h(1) = 0$.

We will study the sign of the first, second, and third derivatives of h on $[1, \infty)$.

Using the product rule of differentiation, the derivative of h with respect to y is

$$h'(y) = 4(1+y)\ln(1+y) + 2(1+y)^2\frac{1}{1+y} - 2(1+y)\ln(y) - 2y\ln(y)$$
$$- 2y(1+y)\frac{1}{y} + \ln^2(y) + 2y\ln(y)\frac{1}{y} - 4(1+y)\ln(2)$$
$$= 4(1+y)\ln(1+y) - 4y\ln(y) + \ln^2(y) - 4(1+y)\ln(2). \tag{2-11}$$

Let us observe that $h'(1) = 0$.

Differentiating again, we obtain

$$h''(y) = 4\ln(1+y) + 4(1+y)\frac{1}{1+y} - 4\ln(y) - 4y\frac{1}{y} + 2\frac{1}{y}\ln(y) - 4\ln(2)$$
$$= 4\ln(1+y) - 4\ln(y) + \frac{2\ln(y)}{y} - 4\ln(2). \tag{2-12}$$

We observe that $h''(1) = 0$.

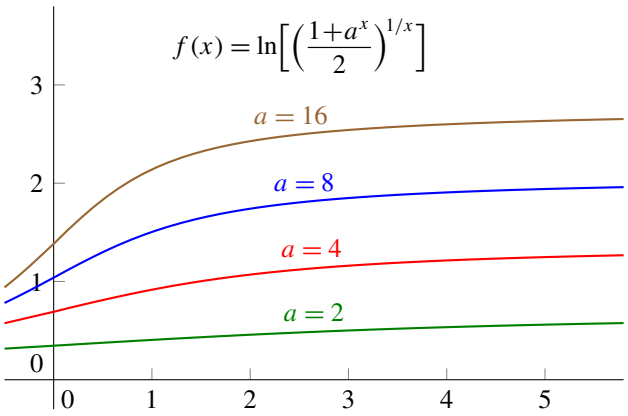


Figure 1. Graph of $y = \ln[((1+a^x)/2)^{1/x}]$ for various values of a.

Finally, differentiating one more time, we obtain

$$h'''(y) = 2\left[\frac{2}{1+y} - \frac{2}{y} + \frac{1}{y^2} - \frac{\ln(y)}{y^2}\right] = 2\left[\frac{1-y}{y^2(y+1)} - \frac{\ln(y)}{y^2}\right] < 0 \qquad (2\text{-}13)$$

for all $y > 1$, since $1 - y < 0$ and $-\ln(y) < 0$.

Thus, we conclude that h'' is strictly decreasing on $[1, \infty)$. This implies that for all $y > 1$, we have $h''(y) < h''(1) = 0$. Hence, h' is strictly decreasing on $[1, \infty)$. This implies that for all $y > 1$, we have $h'(y) < h'(1) = 0$. Therefore, h is strictly decreasing on $[1, \infty)$. Finally, from this assertion we conclude that $h(y) < h(1) = 0$ for all $y > 1$. The last statement is equivalent to the fact that $g''(x) < 0$ for all $x > 0$, and this proves that f is strictly concave on $[0, \infty)$. Therefore, the Hölder mean function of two positive, distinct numbers is strictly logarithmically concave downward on $[0, \infty)$. □

A graphical illustration of the logarithmic concavity of the Hölder means of two positive numbers 1 and a, for various values of a, is presented in Figure 1.

We make now the following simple observation.

Observation 2.2. The Hölder mean of two positive numbers is logarithmically symmetric about the geometric mean of the two numbers. That means, if a and b are positive numbers, then for all $x \in [-\infty, \infty]$, we have

$$H_x(a, b)H_{-x}(a, b) = H_0^2(a, b). \qquad (2\text{-}14)$$

Proof. Indeed, if $x = \infty$, then

$$\begin{aligned} H_\infty(a, b)H_{-\infty}(a, b) &= \max\{a, b\} \min\{a, b\} \\ &= ab = H_0^2(a, b). \end{aligned}$$

On the other hand, for all $x \in \mathbb{R} \setminus \{0\}$, we have

$$H_x(a,b)H_{-x}(a,b) = \left(\frac{a^x + b^x}{2}\right)^{1/x} \left(\frac{a^{-x} + b^{-x}}{2}\right)^{-1/x}$$
$$= \frac{(a^x + b^x)^{1/x}}{2^{1/x}} \frac{(2a^x b^x)^{1/x}}{(a^x + b^x)^{1/x}} = ab = H_0^2(a,b). \qquad \square$$

Corollary 2.3. *Since for any two positive numbers a and b, the function $x \mapsto \ln(H_x(a,b))$ is concave downward on $[0,\infty)$, and its graph is symmetric about the point $(0, \ln(\sqrt{ab}))$, this function is concave upward on $(-\infty, 0]$.*

3. Sharp inequalities concerning α-cevians in a triangle

In this section we use the logarithmic concavity property of the Hölder mean, of two positive numbers, as a function of the index, to prove a sharp inequality for the length of an α-cevian in a triangle.

We give first some definitions.

Definition 3.1. Given a triangle ABC in the plane, for any point M on the side BC, we call AM a *cevian*.

If $M \in BC$, meaning M is between B and C, then we say that AM is an *interior* cevian.

We say that sides AB and AC of the triangle ABC are *adjacent* to the cevian AM.

Definition 3.2. Given a triangle ABC in the plane and α a real number, if $M_\alpha \in BC$, then we say that AM_α is an *α-interior cevian* if

$$\frac{\overline{BM}_\alpha}{\overline{CM}_\alpha} = \left(\frac{\overline{AB}}{\overline{AC}}\right)^\alpha. \tag{3-1}$$

Here $\overline{PQ}$ denotes the length of the segment PQ for any two points P and Q in the plane. See Figure 2.

Observation 3.3. For any real number α, the three α-interior cevians AM_α, BN_α, and CP_α of a triangle ABC are concurrent.

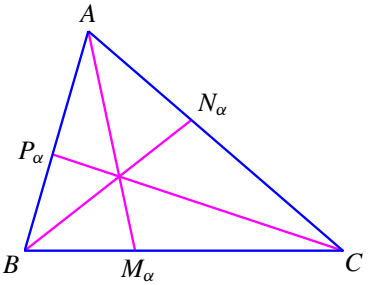


Figure 2. A triangle and its three α-cevians.

Proof. Indeed, we have (see Figure 2)

$$\frac{\overline{BM}_\alpha}{\overline{CM}_\alpha}\cdot\frac{\overline{CN}_\alpha}{\overline{AN}_\alpha}\cdot\frac{\overline{AP}_\alpha}{\overline{BP}_\alpha}=\frac{\overline{AB}^\alpha}{\overline{AC}^\alpha}\cdot\frac{\overline{BC}^\alpha}{\overline{BA}^\alpha}\cdot\frac{\overline{CA}^\alpha}{\overline{CB}^\alpha}=1.$$

It follows now from Ceva's theorem that AM_α, BN_α, and CP_α are concurrent. □

Observation 3.4. We make the following observations:

• For $\alpha=0$, AM_0, BN_0, and CP_0 are the medians of the triangle ABC and they are concurrent in the *centroid* of the triangle ABC. The centroid of a triangle is denoted by $X(2)$ in [Kimberling 1994].

• For $\alpha=1$, AM_1, BN_1, and CP_1 are the inner bisectors of the angles of the triangle ABC and they are concurrent in the *incenter* of the triangle ABC. The incenter of a triangle is denoted by $X(1)$ in [Kimberling 1994].

• For $\alpha=2$, AM_2, BN_2, and CP_2 are the symmedians (symmetric to the medians about the corresponding bisectors) of the triangle ABC and they are concurrent in the *Lemoine point*, also called the *Grebe point*, of the triangle ABC. The Lemoine (Grebe) point of a triangle is denoted by $X(6)$ in [Kimberling 1994].

Let us observe that if AM is an interior cevian of a triangle ABC, then at least one of the angles $\sphericalangle AMB$ and $\sphericalangle AMC$ is obtuse or right. If the angle $\sphericalangle AMB$ is obtuse or right, then in the triangle AMB, the side AB opposite to this angle, with say $\overline{AB}=c$, is the largest side of the triangle. Thus, we have $\overline{AM}<c$.

Similarly, if the angle $\sphericalangle AMC$ is obtuse or right, then $\overline{AM}<b$.

Therefore, in both cases we conclude that

$$\overline{AM}<\max\{b,c\}=H_\infty(b,c).$$

Starting from this simple inequality, we can ask the question:

Question 3.5. *Given a real number α, what is the smallest number $p=p(\alpha)\in[-\infty,\infty]$ such that for all triangles ABC, if AM_α is an α-interior cevian, we have*

$$\overline{AM}_\alpha\le H_p(\overline{AB},\overline{AC})? \tag{3-2}$$

We have the following proposition:

Proposition 3.6. *Let b and c be two fixed positive numbers. We denote by $\mathcal{T}_{b,c}$ the set of all triangles ABC in the plane such that $\overline{AB}=c$ and $\overline{AC}=b$. Then, we have*

$$\sup_{ABC\in\mathcal{T}_{b,c}}\{\overline{AM}_\alpha\mid AM_\alpha \textit{is an } \alpha\textit{-interior cevian in } ABC\}=bc\,\frac{b^{\alpha-1}+c^{\alpha-1}}{b^\alpha+c^\alpha}. \tag{3-3}$$

Proof. We give a vectorial proof.

In triangle ABM_α we have

$$\overrightarrow{AM_\alpha}=\overrightarrow{AB}+\overrightarrow{BM_\alpha}. \tag{3-4}$$

In triangle ACM_α we have

$$\overrightarrow{AM_\alpha} = \overrightarrow{AC} + \overrightarrow{CM_\alpha}. \tag{3-5}$$

Let us first multiply both sides of (3-4) by b^α, and both sides of (3-5) by c^α, and then add the two resulting equations. We obtain

$$(b^\alpha + c^\alpha)\overrightarrow{AM_\alpha} = b^\alpha \overrightarrow{AB} + c^\alpha \overrightarrow{AC} + (b^\alpha \overrightarrow{BM_\alpha} + c^\alpha \overrightarrow{CM_\alpha}). \tag{3-6}$$

Since AM_α is an α-interior cevian, we have

$$\frac{\overline{BM}_\alpha}{\overline{CM}_\alpha} = \frac{c^\alpha}{b^\alpha}.$$

This is equivalent to

$$b^\alpha \overrightarrow{BM_\alpha} + c^\alpha \overrightarrow{CM_\alpha} = 0. \tag{3-7}$$

It follows now from (3-6) that

$$\overrightarrow{AM_\alpha} = \frac{1}{b^\alpha + c^\alpha}(b^\alpha \overrightarrow{AB} + c^\alpha \overrightarrow{AC}). \tag{3-8}$$

Applying the triangle inequality in (3-8), we conclude that

$$\begin{aligned} \overline{AM}_\alpha &\le \frac{1}{b^\alpha + c^\alpha}(b^\alpha \overline{AB} + c^\alpha \overline{AC}) \\ &= \frac{1}{b^\alpha + c^\alpha}(b^\alpha c + c^\alpha b) = bc\,\frac{b^{\alpha-1} + c^{\alpha-1}}{b^\alpha + c^\alpha}. \end{aligned} \tag{3-9}$$

Since this happens for all triangles ABC such that $\overline{AB} = c$ and $\overline{AC} = b$, we conclude that

$$\mathcal{S} \le bc\,\frac{b^{\alpha-1} + c^{\alpha-1}}{b^\alpha + c^\alpha}, \tag{3-10}$$

where

$$\mathcal{S} = \sup_{ABC \in \mathcal{T}_{b,c}} \{\overline{AM}_\alpha \mid AM_\alpha \text{ is an } \alpha\text{-interior cevian in } ABC\}.$$

On the other hand, we have

$$\begin{aligned} \mathcal{S} &\ge \lim_{m(\sphericalangle BAC) \to 0^+} \overline{AM}_\alpha \\ &= \lim_{m(\sphericalangle BAC) \to 0^+} \left[\frac{1}{b^\alpha + c^\alpha} \left| b^\alpha \overrightarrow{AB} + c^\alpha \overrightarrow{AC} \right| \right] \\ &= \left[\frac{1}{b^\alpha + c^\alpha}(b^\alpha \overline{AB} + c^\alpha \overline{AC}) \right] = bc\,\frac{b^{\alpha-1} + c^{\alpha-1}}{b^\alpha + c^\alpha}, \end{aligned} \tag{3-11}$$

where $|\vec{v}|$ denotes the length of the vector $\vec{v}$ for any vector $\vec{v}$ in $\mathbb{R}^2$.

The result of our proposition follows now from inequalities (3-10) and (3-11). □

We can write

$$bc\frac{b^{\alpha-1}+c^{\alpha-1}}{b^{\alpha}+c^{\alpha}}=bc\frac{(b^{\alpha-1}+c^{\alpha-1})/2}{(b^{\alpha}+c^{\alpha})/2}=H_0^2(b,c)\frac{H_{\alpha-1}^{\alpha-1}(b,c)}{H_{\alpha}^{\alpha}(b,c)}. \tag{3-12}$$

Thus, we obtain

$$\mathcal{S}=H_0^2(b,c)\frac{H_{\alpha-1}^{\alpha-1}(b,c)}{H_{\alpha}^{\alpha}(b,c)}. \tag{3-13}$$

Now, Question 3.5 becomes:

Question 3.7. *Given a real number α, what is the smallest number $p=p(\alpha)\in[-\infty,\infty]$ such that for all b and c positive, we have*

$$H_0^2(b,c)\frac{H_{\alpha-1}^{\alpha-1}(b,c)}{H_{\alpha}^{\alpha}(b,c)}\le H_p(b,c)? \tag{3-14}$$

Before answering this question, we present the following necessary condition for an inequality between two functions, whose graphs touch at one point, to hold.

Lemma 3.8. *Let $I\subseteq R$ be an interval, and let*

$$\overset{\circ}{I}:=\{x\in I\mid \text{there exists } r>0 \text{ such that } (x-r,x+r)\subset I\}$$

be the set of the interior points of I. Suppose f and g are two real-valued functions such that:

(1) *$f(x)\le g(x)$ for all $x\in I$.*

(2) *f and g are continuous on I.*

(3) *f and g are twice-differentiable on $\overset{\circ}{I}$.*

(4) *There exists $x_0\in\overset{\circ}{I}$ such that $f(x_0)=g(x_0)$.*

(5) *f'' is continuous at x_0.*

Then, we must have $f'(x_0)=g'(x_0)$ and $f''(x_0)\le g''(x_0)$.

Proof. Let $h(x):=g(x)-f(x)$. Then, for all $x\in I$, we have

$$h(x)\ge 0=h(x_0).$$

Thus, h has an absolute minimum value at x_0, and since x_0 is a point in the interior of I, Fermat's theorem implies $h'(x_0)=0$. This is equivalent to $f'(x_0)=g'(x_0)$.

Since $x_0\in\overset{\circ}{I}$, there exists $r>0$ such that $(x_0-r,x_0+r)\subset I$. Because the function f is dominated by function g, for all $0<h<r$, we have

$$\begin{aligned} f(x_0+h)&\le g(x_0+h),\\ f(x_0-h)&\le g(x_0-h),\\ -2f(x_0)&=-2g(x_0). \end{aligned}$$

Adding these three relations and dividing both sides by the positive number h^2, we obtain

$$\frac{f(x_0+h)+f(x_0-h)-2f(x_0)}{h^2} \le \frac{g(x_0+h)+g(x_0-h)-2g(x_0)}{h^2}.$$

Passing to the limit as $h \to 0^+$, we obtain

$$\lim_{h\to 0^+} \frac{f(x_0+h)+f(x_0-h)-2f(x_0)}{h^2} \le \lim_{h\to 0^+} \frac{g(x_0+h)+g(x_0-h)-2g(x_0)}{h^2}. \quad (3\text{-}15)$$

Applying L'Hôpital's rule in the $\frac{0}{0}$ case, twice, or using Taylor's formula with Lagrange's remainder, it is not hard to see that due to the continuity of f'' at x_0, the last inequality becomes

$$f''(x_0) \le g''(x_0). \qquad \square$$

To answer Question 3.7, we will analyze four cases.

Case 1. If $\alpha \ge 1$, then the answer of Question 3.7 is given by the following proposition.

Proposition 3.9. *If $\alpha \ge 1$, then the smallest number $p = p(\alpha) \in [-\infty, \infty]$ such that for all positive numbers b and c, we have*

$$H_0^2(b,c)\frac{H_{\alpha-1}^{\alpha-1}(b,c)}{H_\alpha^\alpha(b,c)} \le H_p(b,c) \quad (3\text{-}16)$$

is

$$p(\alpha) = 1-2\alpha. \quad (3\text{-}17)$$

Proof. Step 1: We prove first the inequality $p(\alpha) \le 1-2\alpha$.

Indeed, using Observation 2.2, we have

$$\begin{aligned} H_0^2(b,c)\frac{H_{\alpha-1}^{\alpha-1}(b,c)}{H_\alpha^\alpha(b,c)} &= (H_{1-2\alpha}(b,c)H_{2\alpha-1}(b,c))\frac{H_{\alpha-1}^{\alpha-1}(b,c)}{H_\alpha^\alpha(b,c)} \\ &= H_{1-2\alpha}(b,c)\left[\frac{H_{\alpha-1}^{(\alpha-1)/\alpha}(b,c)H_{2\alpha-1}^{1/\alpha}(b,c)}{H_\alpha(b,c)}\right]^\alpha \\ &\le H_{1-2\alpha}(b,c)\cdot 1^\alpha = H_{1-2\alpha}(b,c), \end{aligned} \quad (3\text{-}18)$$

since $0 \le \alpha-1 < \alpha \le 2\alpha-1$ (due to the fact that $\alpha \ge 1$),

$$\frac{\alpha-1}{\alpha}(\alpha-1)+\frac{1}{\alpha}(2\alpha-1) = \alpha, \quad (3\text{-}19)$$

and so, because $x \mapsto H_x(b,c)$ is logarithmically concave on $[0,\infty)$, we have

$$H_{\alpha-1}^{(\alpha-1)/\alpha}(b,c)H_{2\alpha-1}^{1/\alpha}(b,c) \le H_\alpha(b,c). \quad (3\text{-}20)$$

Step 2: We prove now that if p is a positive number such that for all positive numbers b and c, we have

$$H_0^2(b,c)\,\frac{H_{\alpha-1}^{\alpha-1}(b,c)}{H_\alpha^\alpha(b,c)} \le H_p(b,c),$$

then $p \ge 1-2\alpha$.

Choosing $b=1$ and $c=x$, where x is an arbitrary positive number, the above inequality becomes

$$x\,\frac{1+x^{\alpha-1}}{1+x^\alpha} \le \left(\frac{1+x^p}{2}\right)^{1/p}. \tag{3-21}$$

We can see now that the hypotheses of Lemma 3.8 are satisfied for the functions

$$f(x) := \frac{x+x^\alpha}{1+x^\alpha} = 1+\frac{x-1}{1+x^\alpha} = 1+\frac{1}{2}(x-1)+(x-1)\left(\frac{1}{1+x^\alpha}-\frac{1}{2}\right) \tag{3-22}$$

and

$$g(x) := \left(\frac{1+x^p}{2}\right)^{1/p}, \tag{3-23}$$

and the point

$$x_0 := 1. \tag{3-24}$$

Thus, we obtain

$$f''(1) \le g''(1). \tag{3-25}$$

Using Leibniz's rule of differentiation and keeping only the nonzero terms, we obtain

$$\begin{aligned} f''(1) &= \frac{d^2}{dx^2}\left[1+\frac{1}{2}(x-1)+(x-1)\left(\frac{1}{1+x^\alpha}-\frac{1}{2}\right)\right]\Big|_{x=1} \\ &= \frac{d^2}{dx^2}\left[(x-1)\left(\frac{1}{1+x^\alpha}-\frac{1}{2}\right)\right]\Big|_{x=1} \\ &= \binom{2}{1}\frac{d}{dx}(x-1)\Big|_{x=1}\frac{d}{dx}\left(\frac{1}{1+x^\alpha}-\frac{1}{2}\right)\Big|_{x=1} \\ &= 2\left(\frac{-\alpha x^{\alpha-1}}{(1+x^\alpha)^2}\right)\Big|_{x=1} = -\frac{\alpha}{2}. \end{aligned} \tag{3-26}$$

On the other hand, we have

$$\begin{aligned} g'(x) &= \frac{1}{2^{1/p}}\frac{d}{dx}[(1+x^p)^{1/p}] \\ &= \frac{1}{2^{1/p}}\frac{1}{p}(1+x^p)^{(1/p)-1}\,px^{p-1} \\ &= \frac{1}{2^{1/p}}\left(\frac{1+x^p}{x^p}\right)^{(1-p)/p} = \frac{1}{2^{1/p}}(x^{-p}+1)^{(1-p)/p}. \end{aligned}$$

Thus, we obtain

$$g''(x) = \frac{1}{2^{1/p}} \frac{1-p}{p} (x^{-p}+1)^{(1-2p)/p} (-p) x^{-p-1}$$
$$= \frac{p-1}{2^{1/p}} (x^{-p}+1)^{(1-2p)/p} x^{-p-1}.$$

Hence, we have

$$g''(1) = \frac{p-1}{4}. \tag{3-27}$$

Therefore, inequality (3-25) becomes

$$-\frac{\alpha}{2} \leq \frac{p-1}{4}. \tag{3-28}$$

This inequality is equivalent to

$$p \geq 1 - 2\alpha, \tag{3-29}$$

and so, our proof is complete. □

Case 2. If $\frac{1}{2} < \alpha < 1$, then the answer to Question 3.7 is given by the following proposition.

Proposition 3.10. *If* $\frac{1}{2} < \alpha < 1$, *then the smallest number* $p = p(\alpha) \in [-\infty, \infty]$ *such that for all positive numbers b and c, we have*

$$H_0^2(b,c) \frac{H_{\alpha-1}^{\alpha-1}(b,c)}{H_\alpha^\alpha(b,c)} \leq H_p(b,c)$$

is

$$p(\alpha) = 0. \tag{3-30}$$

Proof. Step 1: We prove first the inequality $p(\alpha) \leq 0$. That means, we show that for all positive numbers b and c we have

$$H_0^2(b,c) \frac{H_{\alpha-1}^{\alpha-1}(b,c)}{H_\alpha^\alpha(b,c)} \leq H_0(b,c).$$

Indeed, using the symmetry of the function $x \mapsto \ln(H_x(b,c))$ with respect to the origin

$$H_x(b,c) H_{-x}(b,c) = H_0^2(b,c), \tag{3-31}$$

for $x = \alpha - 1$, we obtain

$$H_{\alpha-1}(b,c) = \frac{H_0^2(b,c)}{H_{1-\alpha}(b,c)}. \tag{3-32}$$

Thus, we have

$$H_0^2(b,c)\frac{H_{\alpha-1}^{\alpha-1}(b,c)}{H_\alpha^\alpha(b,c)} = H_0^2(b,c)\left[\frac{H_0^2(b,c)}{H_{1-\alpha}(b,c)}\right]^{\alpha-1}\frac{1}{H_\alpha^\alpha(b,c)}$$
$$= \frac{H_0^{2\alpha}(b,c)H_{1-\alpha}^{1-\alpha}(b,c)}{H_\alpha^\alpha(b,c)}$$
$$= H_0(b,c)\left[\frac{H_0(b,c)}{H_\alpha(b,c)}\right]^{2\alpha-1}\left[\frac{H_{1-\alpha}(b,c)}{H_\alpha(b,c)}\right]^{1-\alpha}$$
$$\le H_0(b,c)\cdot 1^{2\alpha-1}\cdot 1^{1-\alpha} = H_0(b,c), \tag{3-33}$$

since $0<\alpha$, $1-\alpha<\alpha$, the function $x\mapsto H_x(b,c)$ is increasing, $2\alpha-1>0$, and $1-\alpha>0$.

Step 2: We show now that if $p<0$, then the inequality

$$H_0^2(b,c)\frac{H_{\alpha-1}^{\alpha-1}(b,c)}{H_\alpha^\alpha(b,c)} \le H_p(b,c)$$

cannot hold for all positive numbers b and c.

Indeed, if we assume by contradiction that it holds for all positive numbers b and c, then choosing $b=1$ and $c=x$, where x is an arbitrary positive number, we obtain

$$x\frac{1+x^{\alpha-1}}{1+x^\alpha} \le \left(\frac{1+x^p}{2}\right)^{1/p}. \tag{3-34}$$

Passing to the limit as $x\to\infty$, we get

$$\lim_{x\to\infty}\frac{x+x^\alpha}{1+x^\alpha} \le \lim_{x\to\infty}\left(\frac{1+x^p}{2}\right)^{1/p}. \tag{3-35}$$

Since $\alpha<1$ and $p<0$, the last inequality becomes

$$\infty \le \left(\tfrac{1}{2}\right)^{1/p},$$

which is a contradiction.

Thus the smallest number p for which inequality (3-14) holds is $p(\alpha)=0$. □

Case 3. If $0\le\alpha\le\frac{1}{2}$, then the answer to Question 3.7 is given by the following proposition.

Proposition 3.11. *If $0\le\alpha\le 1$, then the smallest number $p=p(\alpha)\in[-\infty,\infty]$ such that for all positive numbers b and c we have*

$$H_0^2(b,c)\frac{H_{\alpha-1}^{\alpha-1}(b,c)}{H_\alpha^\alpha(b,c)} \le H_p(b,c)$$

is

$$p(\alpha) = 1-2\alpha. \tag{3-36}$$

Proof. Step 1: We show first that $p(\alpha) \leq 1-2\alpha$. Using the logarithmic symmetry of the function $x \mapsto H_x(b,c)$, we have

$$H_0^2(b,c)\,\frac{H_{\alpha-1}^{\alpha-1}(b,c)}{H_\alpha^\alpha(b,c)} = H_0^2(b,c)\left[\frac{H_0^2(b,c)}{H_{1-\alpha}(b,c)}\right]^{\alpha-1}\frac{1}{H_\alpha^\alpha(b,c)}$$
$$= \frac{H_0^{2\alpha}(b,c)H_{1-\alpha}^{1-\alpha}(b,c)}{H_\alpha^\alpha(b,c)}. \tag{3-37}$$

Since $0 \leq \alpha \leq \frac{1}{2}$, we have $0 \leq \alpha \leq 1-\alpha$, and α can be written as a convex combination of 0 and $1-\alpha$ in the following way:

$$\alpha = \left(1-\frac{\alpha}{1-\alpha}\right)\cdot 0 + \frac{\alpha}{1-\alpha}\cdot(1-\alpha). \tag{3-38}$$

Since $x \mapsto H_x(b,c)$ is logarithmically concave on $[0,\infty)$, applying Jensen's inequality, we obtain

$$H_\alpha \geq H_0^{1-\alpha/(1-\alpha)}H_{1-\alpha}^{\alpha/(1-\alpha)}. \tag{3-39}$$

Thus, using (3-37) and (3-39), we have

$$H_0^2(b,c)\,\frac{H_{\alpha-1}^{\alpha-1}(b,c)}{H_\alpha^\alpha(b,c)} = \frac{H_0^{2\alpha}(b,c)H_{1-\alpha}^{1-\alpha}(b,c)}{H_\alpha^\alpha(b,c)}$$
$$\leq \frac{H_0^{2\alpha}(b,c)H_{1-\alpha}^{1-\alpha}(b,c)}{[H_0^{1-\alpha/(1-\alpha)}(b,c)H_{1-\alpha}^{\alpha/(1-\alpha)}(b,c)]^\alpha}$$
$$= H_0^{\alpha/(1-\alpha)}(b,c)H_{1-\alpha}^{(1-2\alpha)/(1-\alpha)}(b,c). \tag{3-40}$$

Let us observe that $\alpha/(1-\alpha) \in [0,1]$, $(1-2\alpha)/(1-\alpha) \in [0,1]$, and

$$\frac{\alpha}{1-\alpha} + \frac{1-2\alpha}{1-\alpha} = 1. \tag{3-41}$$

Applying, Jensen's inequality again, we obtain

$$H_0^2(b,c)\,\frac{H_{\alpha-1}^{\alpha-1}(b,c)}{H_\alpha^\alpha(b,c)} \leq H_0^{\alpha/(1-\alpha)}(b,c)H_{1-\alpha}^{(1-2\alpha)/(1-\alpha)}(b,c)$$
$$\leq H_{[\alpha/(1-\alpha)]\cdot 0+[(1-2\alpha)/(1-\alpha)]\cdot(1-\alpha)}(b,c)$$
$$= H_{1-2\alpha}(b,c). \tag{3-42}$$

Step 2: We can prove now in exactly the same way as in the proof of Proposition 3.9 that if p is real number such that inequality (3-14) holds for all positive numbers b and c, then

$$p \geq 1-2\alpha. \qquad \square$$

Case 4. If $\alpha < 0$, then the answer to Question 3.7 is given by the following proposition.

Proposition 3.12. *If $\alpha < 0$, then the smallest (only) number $p = p(\alpha) \in [-\infty, \infty]$ such that for all positive numbers b and c, we have*

$$H_0^2(b,c)\,\frac{H_{\alpha-1}^{\alpha-1}(b,c)}{H_\alpha^\alpha(b,c)} \le H_p(b,c)$$

is

$$p(\alpha) = \infty. \tag{3-43}$$

Proof. Indeed, we saw geometrically at the beginning of the paper that for all triangles ABC, and all interior cevians AM, we have

$$\overline{AM} \le \max\{\overline{AB}, \overline{AC}\} = H_\infty(b,c),$$

where $b := \overline{AC}$ and $c := \overline{AB}$.

To show that $p(\alpha) = \infty$, we must prove that for all $p < \infty$, inequality (3-14) cannot hold for all positive numbers b and c.

Supposing that for some $p < \infty$ (we may assume $p > 0$) inequality (3-14) holds for all positive numbers b and c, we can choose $b = 1$ and $c = x$, where x is an arbitrary positive number. That means, for all $x > 0$, we have

$$\frac{x + x^\alpha}{1 + x^\alpha} \le \left(\frac{1 + x^p}{2}\right)^{1/p}.$$

Passing to the limit in this inequality as $x \to 0^+$, we obtain

$$\lim_{x\to 0^+} \frac{x + x^\alpha}{1 + x^\alpha} \le \lim_{x\to 0^+} \left(\frac{1 + x^p}{2}\right)^{1/p}.$$

Since $\alpha < 0$, the last inequality is equivalent to

$$1 \le \left(\tfrac{1}{2}\right)^{1/p}.$$

This inequality is impossible, since $0 < \frac{1}{2} < 1$ and $\frac{1}{p} > 0$. □

Therefore, the function $\alpha \mapsto p(\alpha)$ that gives the smallest p such that in any triangle ABC the α-interior cevian starting from A, AM_α, has a length less than or equal to the p-Hölder mean of $\overline{AB}$ and $\overline{AC}$ is $P : \mathbb{R} \to [-\infty, \infty]$, defined by

$$P(\alpha) = \begin{cases} \infty & \text{if } \alpha < 0, \\ 1 - 2\alpha & \text{if } 0 \le \alpha \le \frac{1}{2}, \\ 0 & \text{if } \frac{1}{2} < \alpha < 1, \\ 1 - 2\alpha & \text{if } \alpha > 1. \end{cases} \tag{3-44}$$

See the graph of P in Figure 3.

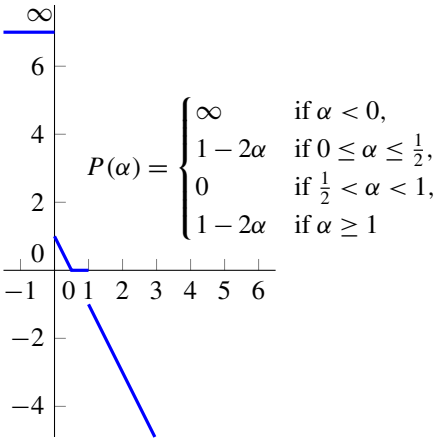


Figure 3. The graph of function $y = P(\alpha)$.

We observe that the function P is nonincreasing and lower semicontinuous.

The branching point $\alpha = 0$ of the piecewise-defined function P corresponds to the median AM_0 of the triangle ABC.

The branching point $\alpha = 1$ corresponds to the bisector AM_1 of the angle $\sphericalangle BAC$.

The branching point $\alpha = \frac{1}{2}$ corresponds to a cevian $AM_{1/2}$ that is concurrent with the corresponding cevians $BN_{1/2}$ and $CP_{1/2}$ in the point $X(366)$ from [Kimberling 1994]. The point $X(366)$ is the isogonal conjugate of $X(365)$, the square root point, which is the intersection point of the three $\frac{3}{2}$-interior cevians of the triangle ABC.

We summarize below our results, in the case of some classic cevians:

Proposition 3.13. *Let ABC be a triangle with sides, starting from A, of lengths $\overline{AC} = b$ and $\overline{AB} = c$. Let M be a point on the side BC of this triangle. Then*:

(1) *If AM is the median corresponding to the vertex A, then its length satisfies*

$$\overline{AM} < \frac{b+c}{2}. \tag{3-45}$$

Moreover, for every $p < 1$, there exists a triangle ABC (depending on p) such that

$$\overline{AM} > \left(\frac{b^p + c^p}{2}\right)^{1/p}. \tag{3-46}$$

(2) *If AM is the interior bisector of the angle $\sphericalangle(BAC)$, then its length satisfies*

$$\overline{AM} < \frac{2}{\frac{1}{a} + \frac{1}{b}}. \tag{3-47}$$

Moreover, for every $p < -1$, there exists a triangle ABC such that

$$\overline{AM} > \left(\frac{b^p + c^p}{2}\right)^{1/p}. \tag{3-48}$$

(3) *If AM is the symmedian corresponding to the vertex A, then its length satisfies*

$$\overline{AM} < \left(\frac{b^{-3}+c^{-3}}{2}\right)^{-1/3}. \tag{3-49}$$

Moreover, for every $p < -3$, there exists a triangle ABC such that

$$\overline{AM} > \left(\frac{b^{p}+c^{p}}{2}\right)^{1/p}. \tag{3-50}$$

Acknowledgements

This research was carried out during the Sampling Advanced Mathematics for Minority Students (SAMMS) program, organized by The Ohio State University (OSU), Department of Mathematics, in Columbus, Ohio, July 10–August 4, 2017. The authors would like to express their gratitude to the OSU Department of Mathematics for supporting this research.

The authors would also like to thank Professor Clark Kimberling, from the University of Evansville, Indiana, for providing them the necessary information about the point $X(366)$.

The authors are extremely grateful to Professor Edward Overman, from the OSU Department of Mathematics for greatly helping them with the figures for this paper.

References

[Bottema et al. 1969] O. Bottema, R. Ž. Djordjević, R. R. Janić, D. S. Mitrinović, and P. M. Vasić, *Geometric inequalities*, Wolters-Noordhoff, Groningen, Netherlands, 1969. MR Zbl

[Bullen 1998] P. S. Bullen, *A dictionary of inequalities*, Pitman Monographs Surv. Pure Appl. Math. **97**, Longman, Harlow, UK, 1998. MR Zbl

[Bullen et al. 1988] P. S. Bullen, D. S. Mitrinović, and P. M. Vasić, *Means and their inequalities*, Math. Appl. (East Eur. Series) **31**, Reidel, Dordrecht, 1988. MR Zbl

[Kimberling 1994] C. Kimberling, "Encyclopedia of triangle centers", website, 1994, available at https://tinyurl.com/encytria.

[Mitrinović et al. 1989] D. S. Mitrinović, J. E. Pečarić, and V. Volenec, *Recent advances in geometric inequalities*, Math. Appl. (East Eur. Series) **28**, Kluwer, Dordrecht, 1989. MR Zbl

[Pólya and Szegő 1972] G. Pólya and G. Szegő, *Problems and theorems in analysis, I: Series, integral calculus, theory of functions*, Grundlehren der Math. Wissenschaften **193**, Springer, 1972. MR Zbl

Received: 2018-05-23 Revised: 2018-11-09 Accepted: 2018-11-15

stan.7@osu.edu *Department of Mathematics, Ohio State University at Marion, Marion, OH, United States*

zap14480@uvg.edu.gt *Department of Mathematics, Universidad del Valle de Guatemala, Guatemala, Guatemala*

INVOLVE 12:4 (2019)
dx.doi.org/10.2140/involve.2019.12.687

Applying prospect theory to multiattribute problems with independence assumptions

Jack Stanley and Frank P. A. Coolen

(Communicated by Sat N. Gupta)

We discuss a descriptive theory of decision making which has received much attention in recent decades: prospect theory. We specifically focus on applying the theory to problems with two attributes, assisted by different independence assumptions. We discuss a process for solving decision problems using the theory before applying it to a real life example of purchasing breakdown cover.

1. Introduction

In this paper, we apply prospect theory (PT) to multiattribute problems, specifically those with two attributes. We will consider levels of independence between the attributes in the problem and will split the corresponding value function into different parts. When discussing independence between attributes in multiattribute expected utility theory (EUT), Keeney and Raiffa [1976] use the term "utility independence". However, in this paper, we use the term "independence" to represent utility or value independence, dependent on whether we are in the EUT or PT case. This can be seen later in Definition 1.

Within PT, the reference point is chosen to be the point from which you consider gains and losses. As such, the different parts of the value function are all based on whether outcomes are better or worse than the reference point with respect to each attribute. Following this, we will design a process which can be used to effectively and efficiently solve a multiattribute decision problem. We show how this can be applied to a real-life problem of purchasing breakdown cover.

We will begin by covering some background information in Section 2, introducing notation and explaining how EUT deals with multiattribute problems. We also introduce PT and how it is applied to single attribute problems. In Section 3, we derive formulas for applying PT to multiattribute problems under different levels of independence. Following this, in Section 4, we introduce a standard process which

MSC2010: 91B06, 91B16.
Keywords: decision theory, independence, multiple attributes, prospect theory, utility theory.

can be used to solve multiattribute decision problems. We illustrate the approach with an example of purchasing breakdown cover in Section 5.

2. Background

We begin by introducing some definitions and notation. The following notation is based on [Starmer 2000, p. 334]. We define a prospect to consist of a set of outcomes (e.g., $x_1, x_2, \ldots, x_n$) with probabilities corresponding to them (e.g., $p_1, p_2, \ldots, p_n$). At this point, it should be made clear that prospects will be present for both EUT and PT. A prospect simply represents what it is that we are considering and should not be thought of as being linked exclusively to PT.

Notationally, we will consider prospects using capital letters (e.g., A, B, C) and will consider probabilities using p with subscripts (e.g., p_1, p_2, p_3). Therefore, an example of a prospect is A = $(x_1, p_1; x_2, p_2; \ldots; x_n, p_n)$. Here, $(p_1, p_2, \ldots, p_n)$ is a probability distribution (hence $p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$) and $(x_1, x_2, \ldots, x_n)$ are the associated outcomes. So, with prospect A as described above, we would expect outcome x_i with probability p_i for $i = 1, \ldots, n$. Interestingly, we can consider some of the outcomes within a prospect to be prospects themselves. So, for example, we could have prospects B, C and D with D = (B, 0.25; C, 0.75). This would yield prospect B with probability 0.25 and prospect C with probability 0.75.

When comparing prospects A and B:

- A $\prec$ B denotes B is preferred to A.
- A $\preccurlyeq$ B denotes B is at least as preferable as A.
- A $\sim$ B denotes indifference between A and B.

We now discuss prospect theory (PT) in the single-attribute case, which we later extend to be applicable to multiattribute decision problems. PT was introduced by Daniel Kahneman and Amos Tversky [1979]. It was designed to be a direct alternative to EUT, dealing with some of the issues where EUT fails to reflect the preferences of the majority of individuals. In reflecting these attitudes, it is clear that PT was designed as a descriptive theory.

PT is made up of two stages: an editing phase and an evaluation phase. The editing phase of PT is designed to simplify the decision-making process. Using several different rules and operations, the individual can change the prospects which they have to choose between and eliminate any that should never be picked. After the editing phase has been completed, the decision maker will then complete an evaluation phase in which they will decide what the best decision will be for them, based on their opinions and beliefs. Within the evaluation phase, there are two main functions which are considered and combined to help evaluate each prospect numerically: a decision-weighting function, represented by $\pi(p)$, and a value function, represented by $v(x)$.

The reference point in PT is of great importance as it allows us to choose a neutral point and then consider better points as gains and worse points as losses. This is because it was found in [Kahneman and Tversky 1979, pp. 268–269] that individuals are risk-averse when it comes to gains, whereas they are risk-seeking when dealing with losses. To reflect this, the value function is generally designed to be concave above the reference point and convex below the reference point.

Now that we have introduced PT, we will discuss the independence assumptions which we will later use in deriving the formulas for multiattribute PT. Note here that we are going to denote the values the attributes can take by Y and Z. In this paper, we will be focussing on the instances where we have certain levels of independence between the attributes Y and Z. This independence is used in multiattribute EUT for utility functions and later we will use it for multiattribute PT when we work with value functions. We define it as follows:

Definition 1 (adapted from [Keeney and Raiffa 1976, pp. 226, 229] to be applicable to PT).

- Y is independent of Z when conditional preferences for outcomes on Y given z do not depend on the particular level of z.
- Y and Z are mutually independent attributes if Y is independent of Z and Z is independent of Y.

We will now briefly discuss multiattribute EUT and explain how with certain levels of independence, it is possible to construct a utility function which helps to choose the best possible alternative for a decision problem. This focusses on [Keeney and Raiffa 1976].

To begin with, we discuss how Keeney and Raiffa show that independence between attributes can be represented effectively with an equation. This can be found in [loc. cit., p. 144] where it is suggested that two strategically equivalent utility functions can be linked by

$$u_1(x) = h + ku_2(x) \quad \text{for all } x. \tag{1}$$

Now, let us consider the instance where Y is independent of Z. By the definition of independence, this means that $u(y, z_0)$ and $u(y, z_1)$ are strategically equivalent (i.e., $u(y, z_0) \sim u(y, z_1)$) for any $z_0, z_1 \in \mathrm{Z}$. Due to this, we can represent the utility function of any $y \in \mathrm{Y}$ and any $z \in \mathrm{Z}$ as a transformation of another utility function of $y \in \mathrm{Y}$ and a different $z' \in \mathrm{Z}$. Namely

$$u(y, z) = g(z) + h(z)u(y, z') \quad \text{for any } y \in \mathrm{Y} \text{ and } z \in \mathrm{Z}, \tag{2}$$

with the functions g and h only functions of z (i.e., constant in y). Following this, we can say that an attribute Y is independent of Z if and only if (2) holds. This will later be applied to PT and form the basis of the derivation of the formulas for multiattribute PT.

If we have two attributes which are mutually independent, we can use Theorem 5.2 from [loc. cit., p. 234] to evaluate the utility of any alternative. Similarly, if the attributes only have Z independent of Y, we use Theorem 5.6 from [loc. cit., p. 244]. This briefly describes how EUT can be applied to multiattribute problems to assist in choosing the best alternative in a decision problem. We will now show how PT can be applied in a similar way.

3. Applying PT to multiattribute problems

As PT is the natural alternative to EUT and EUT has been applied to multiattribute problems, it seems logical to apply PT to multiattribute problems. One method which uses PT is the TODIM method [Gomes and Lima 1991]. This uses pairwise comparisons over each alternative based on each attribute, choosing one attribute to be the reference attribute. Other than that, a lot of the research in multiattribute problems using prospect theory involves combining the theory itself with the theory of fuzzy sets. The theory of fuzzy sets was introduced in [Zadeh 1979] and applied to multiattribute problems by [Bellman and Zadeh 1970]. Combining prospect theory with the theory of fuzzy sets will not be covered in this paper, research into this area can be found in [Krohling and de Souza 2012; Liu et al. 2011; Wang and Sun 2008].

There is limited research looking into applying PT in the same way as EUT is applied to multiattribute decision problems in [Keeney and Raiffa 1976]. The closest thing currently available is [Hu and Zhou 2009], which is described in [Liu et al. 2011] as being a "multiple criteria decision-making method for the risk decision-making problem based on prospect theory". However, it still does not produce a piecewise form of the value function and instead focusses on using the weighting function from cumulative PT, introduced in [Tversky and Kahneman 1992]. Further research which uses PT methods with multiattribute problems can be found in [Egozcue et al. 2014].

In introducing the theory, we begin by considering what the reference point will be for each attribute Y and Z. Denote these as (y_0, z_0) and let $v(y_0, z_0) = 0$. In this paper, we keep the reference point constant and as such, we do not include it in the notation. It should be emphasised here how important the choice of reference point is. Different choices of reference point have a significant impact on the outcome of a decision problem. This can be seen later in Section 5.

We begin by defining strategic equivalence of two value functions (similarly to [Keeney and Raiffa 1976, p. 144]) as follows:

Definition 2. Two value functions, v and v^*, are *strategically equivalent*, denoted as $v \sim v^*$, if and only if they imply the same preference ranking for any two prospects or outcomes.

Suppose we have two value functions, v and v^*, which are strategically equivalent. We assume that there are constants h_1, k_1, h_2, k_2 with $k_1 > 0$ and $k_2 > 0$ such that

$$v(x) = \begin{cases} h_1 + k_1 v^*(x) & \text{if } x \succcurlyeq x_0, \\ h_2 + k_2 v^*(x) & \text{if } x \preccurlyeq x_0. \end{cases} \tag{3}$$

This is a natural assumption for the value functions v and v^*, because multiplication by a positive constant will not affect the preference ordering over outcomes and neither will a transformation by an additive constant.

Similarly to the multiattribute EUT case, we can represent attribute Y being independent of Z with the equation

$$v(y, z) = \begin{cases} c_1^+(z) + c_2^+(z) v(y, z_0) & \text{if } y \succcurlyeq y_0, \\ c_1^-(z) + c_2^-(z) v(y, z_0) & \text{if } y \preccurlyeq y_0, \end{cases} \tag{4}$$

and attribute Z being independent of Y with

$$v(y, z) = \begin{cases} d_1^+(y) + d_2^+(y) v(y_0, z) & \text{if } z \succcurlyeq z_0, \\ d_1^-(y) + d_2^-(y) v(y_0, z) & \text{if } z \preccurlyeq z_0. \end{cases} \tag{5}$$

Notice here the similarities with (3). For example, $c_1^+(z)$ and $d_1^+(y)$ are similar to h_1 in (3). Note that $c_1^+(z)$ is a function of z meaning it is constant in y, as is required for the value function which is only a function of y. The same logic applies for the rest of the c and d values in (4) and (5) respectively.

3.1. ***Working with mutually independent attributes.*** To begin with, we are going to consider the case where we have attributes Y and Z which are mutually independent. Assuming this, we are going to loosely follow the proof for the multiattribute EUT formula in [Keeney and Raiffa 1976, p. 234–235] but apply it to PT.

Y and Z being mutually independent means we can represent the independence using (4) and (5). Substituting y_0 into (4) gives that $c_1^+(z) = c_1^-(z) = v(y_0, z)$. This is to retain a level of continuity and to ensure that the limit as y tends to y_0 from above or below is the same. Then, consider a value y_1 such that $y_1 \succ y_0$ and substitute this value into (4) to get

$$v(y_1, z) = v(y_0, z) + c_2^+(z) v(y_1, z_0) \quad \Longrightarrow \quad c_2^+(z) = \frac{v(y_1, z) - v(y_0, z)}{v(y_1, z_0)}. \tag{6}$$

Similarly, considering a value y_{-1} with $y_{-1} \prec y_0$ and substituting this into (4) gives

$$v(y_{-1}, z) = v(y_0, z) + c_2^-(z) v(y_{-1}, z_0) \quad \Longrightarrow \quad c_2^-(z) = \frac{v(y_{-1}, z) - v(y_0, z)}{v(y_{-1}, z_0)}. \tag{7}$$

Notice that both constants $c_2^+(z)$ and $c_2^-(z)$ are positive as is required in (3). The function $c_2^-(z)$ is positive as both the numerator and denominator are negative as $v(y_0, z_0) = 0$, so $v(y_{-1}, z_0) < 0$.

Now that we have evaluated the values of the constants $c_1^+(z)$, $c_1^-(z)$, $c_2^+(z)$ and $c_2^-(z)$, we can substitute these back into (4) to get

$$v(y,z)=\begin{cases} v(y_0,z)+\left(\dfrac{v(y_1,z)-v(y_0,z)}{v(y_1,z_0)}\right)v(y,z_0) & \text{if } y\succcurlyeq y_0,\\ v(y_0,z)+\left(\dfrac{v(y_{-1},z)-v(y_0,z)}{v(y_{-1},z_0)}\right)v(y,z_0) & \text{if } y\preccurlyeq y_0. \end{cases} \tag{8}$$

Similar logic can be used to rewrite (5) as

$$v(y,z)=\begin{cases} v(y,z_0)+\left(\dfrac{v(y,z_1)-v(y,z_0)}{v(y_0,z_1)}\right)v(y_0,z) & \text{if } z\succcurlyeq z_0,\\ v(y,z_0)+\left(\dfrac{v(y,z_{-1})-v(y,z_0)}{v(y_0,z_{-1})}\right)v(y_0,z) & \text{if } z\preccurlyeq z_0. \end{cases} \tag{9}$$

Evaluating (9) at y_1 for any point $z\in \mathrm{Z}$ gives

$$v(y_1,z)=\begin{cases} v(y_1,z_0)+\left(\dfrac{v(y_1,z_1)-v(y_1,z_0)}{v(y_0,z_1)}\right)v(y_0,z) & \text{if } z\succcurlyeq z_0,\\ v(y_1,z_0)+\left(\dfrac{v(y_1,z_{-1})-v(y_1,z_0)}{v(y_0,z_{-1})}\right)v(y_0,z) & \text{if } z\preccurlyeq z_0. \end{cases} \tag{10}$$

Similarly, evaluating (9) at y_{-1} for any point $z\in \mathrm{Z}$ gives

$$v(y_1,z)=\begin{cases} v(y_{-1},z_0)+\left(\dfrac{v(y_{-1},z_1)-v(y_{-1},z_0)}{v(y_0,z_1)}\right)v(y_0,z) & \text{if } z\succcurlyeq z_0,\\ v(y_{-1},z_0)+\left(\dfrac{v(y_{-1},z_{-1})-v(y_{-1},z_0)}{v(y_0,z_{-1})}\right)v(y_0,z) & \text{if } z\preccurlyeq z_0. \end{cases} \tag{11}$$

Substituting (10) and (11) into (8) and simplifying leads to the following theorem for calculating the value function for a multiattribute prospect theory problem:

Theorem 3. *For attributes* Y *and* Z *which are mutually independent, the value function required for multiattribute prospect theory for any point* (y,z) *with* $y\in \mathrm{Y}$ *and* $z\in \mathrm{Z}$ *can be evaluated as*

$$\begin{aligned} v(y,z)=\ & v(y_0,z)+v(y,z_0)\\ & +v(y_0,z)v(y,z_0)\\ & \times\begin{cases} \left(\dfrac{v(y_1,z_1)-v(y_1,z_0)-v(y_0,z_1)}{v(y_0,z_1)v(y_1,z_0)}\right) & \text{if } y\succcurlyeq y_0,\ z\succcurlyeq z_0,\\ \left(\dfrac{v(y_1,z_{-1})-v(y_1,z_0)-v(y_0,z_{-1})}{v(y_0,z_{-1})v(y_1,z_0)}\right) & \text{if } y\succcurlyeq y_0,\ z\preccurlyeq z_0,\\ \left(\dfrac{v(y_{-1},z_1)-v(y_{-1},z_0)-v(y_0,z_1)}{v(y_0,z_1)v(y_{-1},z_0)}\right) & \text{if } y\preccurlyeq y_0,\ z\succcurlyeq z_0,\\ \left(\dfrac{v(y_{-1},z_{-1})-v(y_{-1},z_0)-v(y_0,z_{-1})}{v(y_0,z_{-1})v(y_{-1},z_0)}\right) & \text{if } y\preccurlyeq y_0,\ z\preccurlyeq z_0. \end{cases} \end{aligned} \tag{12}$$

So, provided we have mutual independence, this theorem allows us to assign a value to any alternative (y, z) for any $y \in \mathrm{Y}$ and $z \in \mathrm{Z}$. The values in the piecewise function which are the coefficients of the $v(y_0, z)v(y, z_0)$ term are similar to the constant k_{YZ} in the EUT case [Keeney and Raiffa 1976, p. 234, Theorem 5.2]. In fact, if you were to choose the reference point as the worst possible values in Y and Z, you would find that you are only in the top part of the piecewise function of Theorem 3. This would give no significant differences between the utility function in EUT and the value function in PT as all outcomes would be considered as gains. As such, everyone would display risk-aversion to all options, as is the case in EUT. This shows how the significant difference in the theories results from the use and choice of a reference point.

In deriving this formula, we assigned the reference point to be (y_0, z_0). However, suppose we decide to change the reference point to another point (y, z) for any $y \in \mathrm{Y}$ and $z \in \mathrm{Z}$. This would then potentially need different points y_1, y_{-1}, z_1, z_{-1} to be chosen, which changes the constant term. So, an individual who has the same preference ordering for each attribute could change their decision using this formula based on the reference point that they choose. This shows how important the reference point is.

3.2. ***One independent attribute.*** Let us consider the case where we only have one attribute being independent of the other. Without loss of generality, we assume that attribute Z is independent of Y, meaning we can write the value function of y and z (with reference point (y_0, z_0)) in the form of (5). From here, we will use similar steps as in the proof of Theorem 5.6 in [Keeney and Raiffa 1976, pp. 244–245].

We begin with (5) and choose z_1 and z_{-1} with $z_1 \succ z_0 \succ z_{-1}$. They are chosen to satisfy the following equations which fix the origin and unit of measure of $v(y_0, z)$:

$$v(y_0, z_0) = 0, \tag{13}$$

$$v(y_0, z_1) = 1, \tag{14}$$

$$v(y_0, z_{-1}) = -1. \tag{15}$$

We can now evaluate (5) at the point $z = z_0$ for any value of $y \in \mathrm{Y}$ which leads to (using (13))

$$d_1^+(y) = d_1^-(y) = v(y, z_0). \tag{16}$$

This can now be combined with (5) to give

$$v(y, z) = \begin{cases} v(y, z_0) + d_2^+(y)v(y_0, z) & \text{if } z \succcurlyeq z_0, \\ v(y, z_0) + d_2^-(y)v(y_0, z) & \text{if } z \preccurlyeq z_0. \end{cases} \tag{17}$$

Now evaluate (17) at the points $z = z_1$ and $z = z_{-1}$. Using (14) and (15), we get

$$d_2^+(y) = v(y, z_1) - v(y, z_0), \tag{18}$$

$$d_2^-(y) = v(y, z_0) - v(y, z_{-1}). \tag{19}$$

Notice here that we have $d_2^+(y) > 0$ and $d_2^-(y) > 0$, as is required in the assumption in (3). Substituting these into (17) and rearranging leads to the following theorem.

Theorem 4. *For attributes* Y *and* Z *with* Z *independent of* Y *but without* Y *being independent of* Z*, the value function required for multiattribute prospect theory for any point* (y, z) *with* $y \in \mathrm{Y}$ *and* $z \in \mathrm{Z}$ *can be evaluated as*

$$v(y,z) = \begin{cases} v(y,z_0)[1 - v(y_0,z)] + v(y,z_1)v(y_0,z) & \text{if } z \succcurlyeq z_0, \\ v(y,z_0)[1 + v(y_0,z)] - v(y,z_{-1})v(y_0,z) & \text{if } z \preccurlyeq z_0. \end{cases} \tag{20}$$

If an individual is going to use Theorem 4 to help solve a multiattribute problem, they are required to specify a few values initially to fix the unit of measure for each of the value functions. This is something which is also done in the EUT case with one independent attribute [Keeney and Raiffa 1976, p. 244, Theorem 5.6]. We have already fixed the unit of measure for $v(y_0, z)$ for any $z \in \mathrm{Z}$. However, we still need to specify a unit of measure for the other value functions in formula (20), namely $v(y, z_0)$, $v(y, z_1)$ and $v(y, z_{-1})$.

At this point, we have already specified one value for each of these functions (each of them evaluated at y_0). Another two points are required for each value function to specify the unit of measure. This is because $v(y, z)$ is, by definition, different for gains and losses. Therefore, we must fix the unit of measure for $y \succ y_0$ and also for $y \prec y_0$. We can do this in a similar way to the EUT case.

For fixing the unit of measure of $v(y, z_0)$, we find points $y_2 \succ y_0 \succ y_{-2} \in \mathrm{Y}$ and $z_2 \succ z_0 \succ z_{-2} \in \mathrm{Z}$ such that the individual is indifferent between (y_0, z_2) and (y_2, z_0) and between (y_0, z_{-2}) and (y_{-2}, z_0). The indifference means that $v(y_0, z_2) = v(y_2, z_0)$ and $v(y_0, z_{-2}) = v(y_{-2}, z_0)$. Then, whether dealing with gains or losses in the attribute Y, we will have fixed the unit of measure for $v(y, z_0)$. The same can then be done for $v(y, z_1)$ and $v(y, z_{-1})$ to fix their unit of measure. Assuming that such points exist is a trivial assumption to make as if this is not the case, you are in a much simpler situation and do not require the theories introduced in this paper.

Clearly, if attribute Y is independent of Z instead of Z being independent of Y, (20) would have y and z swapped, with points y_1, y_{-1} chosen instead of z_1, z_{-1}. Therefore, Theorem 4 allows us to evaluate a value for all alternatives y, z with a weaker requirement of only one attribute being independent of the other. However, in exchange for this, there are more values that an individual will have to specify.

3.3. ***Calculating the coefficients for multiattribute PT.*** In this section, we are going to focus on methods that can be used to evaluate the coefficients of the value functions in (12). Within this, let us suppose that the individual has already chosen the forms that their value functions will take. This means they will have already specified $v(y, z_0)$ for all $y \in \mathrm{Y}$ and $v(y_0, z)$ for all $z \in \mathrm{Z}$. This means we

have four points which remain to be specified: $v(y_1, z_1)$, $v(y_1, z_{-1})$, $v(y_{-1}, z_1)$ and $v(y_{-1}, z_{-1})$ for appropriately chosen y_1, y_{-1}, z_1, z_{-1}. But how do we make these choices of points and resulting values?

Choosing what points to use as y_1, y_{-1}, z_1 and z_{-1} is a free choice for the decision maker, provided that they satisfy $y_1 \succ y_0 \succ y_{-1}$ and $z_1 \succ z_0 \succ z_{-1}$. This is because the important consideration here is the values associated with them, not the actual points. The values then form the coefficient of the value functions in the piecewise part of (12).

Let us initially focus on working out $v(y_1, z_1)$. A simple way to do this is to attempt to find equivalences. For example, the decision maker should consider what value of $y \in \mathrm{Y}$ means that they are indifferent between (y_1, z_1) and (y, z_0). As we have already specified the value function $v(y, z_0)$, we now have a value for $v(y_1, z_1)$. The same can be done for fixing Y at y_0 and considering what value of $z \in \mathrm{Z}$ leads to indifference between (y_0, z) and (y_1, z_1). A sensible check which the decision maker can complete is to do both and ensure that the values are the same (or at least very similar). This simple method allows the decision maker to accurately fix the coefficients of their value function in (12) and ensures that the values chosen are in line with their beliefs.

If the value functions $v(y, z_0)$ and $v(y_0, z)$ have not already been specified, this leads to a slightly more complicated situation. In this case, equivalence relations would not be very useful and as such, it is likely that the decision maker will simply have to choose these values. However, logic checks can be used to ensure that the values chosen are appropriate. For example, $v(y_1, z_1)$ should be greater than $v(y_1, z_0)$ provided that $z_1 \succ z_0$. So, checking that more preferred values are given a higher value is a simple logic check. Furthermore, the decision maker could use the equivalence relations stated earlier once the value functions have been fully specified. This will then ensure that the values that they have chosen are appropriate for their beliefs.

4. Process of solving a multiattribute decision problem

We now outline a "standard" process which can be used to help an individual solve a decision problem with two attributes. The aim of this is to make it quicker and easier for the individual to solve their decision problem and to ensure the best outcome based on their beliefs. We suggest that the process of solving a multiattribute decision problem can be broken down into a few stages. They are

- formulating the problem,
- independence and choosing a theory,
- applying the theory of choice.

We will now discuss the first two stages in the following two subsections. We will not discuss applying the theory of choice as this is a very simple process once the

previous two stages have been completed. However, when applying the theory of choice, logic checks should be carried out to ensure that the individual is assigning utilities/values in line with their underlying preferences and beliefs.

4.1. ***Formulating the problem.*** Formulating a multiattribute decision problem involves several stages which all need to be completed fully and carefully. In formulating the problem in this way, it will save time later in evaluating alternatives and will ensure the individual fully understands the decision problem they are faced with. At this point, we are going to assume that we know the individual who is faced with the decision problem.

First, we need to establish what the decision problem actually is. What are the alternatives that the individual is aiming to choose between? For example, the alternatives could be different treatments which a doctor is choosing between to give to a patient.

Following this, we need to be clear as to what is required as the outcome from the decision problem. This could involve choosing a single best alternative, choosing an acceptable region of alternatives or simply giving a ranking for all of the possible alternatives. This helps the individual understand the problem they are faced with and ensures they are getting the output that they desire.

It should also be specified to the decision maker whether there are any probabilities involved in the decision problem. With probabilities involved, this would then be an extra consideration for the decision maker to have when deciding between EUT and PT. This is because they will have to consider the probability distribution itself and consider whether the weighting function in PT better reflects how they would view the probabilities.

We now need to understand what the attributes are that we are using to compare the alternatives. In this paper, we have focussed on the case where there are two attributes. It is important at this point to understand what these attributes actually mean and the possible values they could take in the decision problem. Knowing the domain of values for each attribute allows us to accurately scale the utility/value function based around the best and worst possible outcomes, according to the individual. Without this, mistakes could be made in that the individual may expect a much higher value for an attribute than is actually possible. Therefore, it is vital that the individual understands the possible values each attribute can take.

The individual should also be clear as to whether the attributes are continuous or discrete. For example, a monetary attribute is continuous to two decimal places. So, if the decision problem has two of the possible outcomes as £10.00 and £15.00, we understand the values that would go between them. As such, it is conceivable to have £12.50 as the reference point if we were using PT. However, if an attribute is considered to be discrete, this would not be the case. For example, suppose the attribute is

different types of fruit and that two of the possible outcomes are apple and orange. Clearly, it is not feasible to choose a reference point between these two outcomes.

It is also useful to understand a rough preference ordering for the attributes from the decision maker. For some attributes, this is simpler than others. For example, a monetary attribute is one where it is easy to suggest that an individual will simply aim to minimise expenditure or maximise income. However, other attributes such as colours of paint are much more subjective.

It could be suggested that a rough preference ordering is not needed at this stage. However, we believe there to be a couple of reasons why it would be useful to consider this now. First, it gives us a basis on which to ensure the individual is acting logically. Secondly, it is useful at this stage to allow for cancellation. Once the preference ordering has been considered, it may be that some alternatives are better than others for both attributes. If this is the case (and the decision problem is simply choosing a single best decision), then the alternative which is worse for both attributes could be ignored. So, if we preferred blue paint to red paint and blue was cheaper than red, then the red paint would be dominated and removed from the decision problem.

We have now completed the process of formulating the problem. This was done with the aim of simplifying the process later and assisting the decision maker in understanding the problem with which they are faced.

4.2. ***Independence and choosing a theory.*** Now that we have formulated the decision problem, we have to consider whether either of the attributes are independent of each other. This can be tested in the following way. Consider two particular values of Y, say $y_1, y_2 \in \mathrm{Y}$, that you can specify a preference and strength of preference between. Now consider any value $z_1 \in \mathrm{Z}$. What is your preference and level of preference between (y_1, z_1) and (y_2, z_1)? Now suppose that we choose a different $z_2 \in \mathrm{Z}$ with $z_2 \neq z_1$. What is your preference and level of preference between (y_1, z_2) and (y_2, z_2)? If the preference ordering and strength of preference remains the same between the two options, irrespective of the value of Z, then we can say that Y is independent of Z. As you would expect, to show that Z is independent of Y uses a similar logic, but with the y and z values switched in the above.

However, can you be 100% sure of independence without testing the indifference for all of the possible values of Y or Z? There is no guarantee that the indifference will necessarily hold for all Y and Z. Unfortunately, attempting to test this for all possible values of Y and Z would take a significant amount of time and potentially be impossible. As such, it is easier to either test for a couple of potential Y, Z combinations or to simply make an assumption.

It is worth considering at this point how individuals actually face multiattribute decision problems — do most consider attributes as independent? Clearly the test above can be used to see whether independence is a reasonable assumption. The

theory	level of independence	number of choices
EUT	mutual independence	$2n+2$
	single independence	$3n+1$
PT	mutual independence	$2n+8$
	single independence	$3n+7$

Table 1. Comparison of the effect of independence on the number of choices for both methods.

advantage of it is that it creates a simpler model from which to evaluate the possible alternatives. Without this assumption, fitting a model to evaluate how good (or bad) each alternative is becomes difficult. One possible alternative is to assume independence and perform logic checks on the outcome to ensure rational decisions have been made.

Suppose that we are in a situation where we assume some level of independence between the attributes. A key consideration when choosing what level of independence to assume is the number of choices that will be required. Clearly, this will be different for EUT and PT and for different levels of independence. Let us consider the general case of having attributes Y and Z, with n different alternatives to choose between. We begin by considering the case of mutual independence in PT and following Theorem 3 from Section 3.1.

To begin with, we make a choice of what the individual's reference points $y_0 \in \mathrm{Y}$ and $z_0 \in \mathrm{Z}$ are. Once the individual has made these two choices, we have to choose $y_1, y_{-1} \in \mathrm{Y},\ z_1, z_{-1} \in \mathrm{Z}$ that satisfy $y_1 \succ y_0 \succ y_{-1}$ and $z_1 \succ z_0 \succ z_{-1}$. Following this, the individual will make eight choices to evaluate $v(y_i, z_j)$ for $i, j = -1, 0, 1$ using the methods in Section 3.3 (note, we already have $v(y_0, z_0) = 0$). This fixes the constant values in the piecewise function of Theorem 3.

The individual will now need to specify conditional value functions $v(y_0, z)$ for all $z \in \mathrm{Z}$ and, similarly, specify $v(y, z_0)$ for all $y \in \mathrm{Y}$. This could be considered to be four choices of value functions for each domain of Y and Z. However, they would then need to be evaluated for all appropriate $y \in \mathrm{Y}$, $z \in \mathrm{Z}$ that we have not already specified. As such, we suggest that evaluating $v(y, z_0)$ for all $y \in \mathrm{Y}$ will be $n-3$ choices and similarly for $v(y_0, z)$ for all $z \in \mathrm{Z}$.

The individual is now ready to evaluate $v(y, z)$ for any $y \in \mathrm{Y}$ and any $z \in \mathrm{Z}$ and choose the pairing (y, z) with the highest value. To get to this point, the individual has made $2+4+8+2(n-3) = 2n+8$ choices. Applying similar logic to the other three cases gives us the number of choices for each case; see Table 1.

Notice here that when n is not significantly large, there is not a huge difference between the different levels of independence. However, if n gets large, this is where having mutual independence would save significantly more time. We can

also notice here that the number of choices is unlikely to have an impact when choosing between EUT and PT. These extra choices for PT can be considered to be a trade-off for the potentially more realistic modelling that PT provides.

When choosing between EUT and PT, it is worth considering whether there are probabilities involved in the decision problem. With probabilities involved, the decision can become more complicated. For example, if the probability distribution has some events occurring with certainty or with very low probabilities, the weighting function in PT treats these differently [Kahneman and Tversky 1979, pp. 280–284], making the decision more complicated. However, if the probabilities are away from 0 or 1, there is no significant difference between the probabilities and the weighting function and as such, the choice remains the same as without probabilities.

For now, we are going to consider that we do not have probabilities involved and compare the theories at a base level. The main difference comes in that PT has a reference point from which to consider gains and losses, whereas EUT does not. As such, if the alternatives contain gains and losses, PT is likely to be more useful as it was designed to reflect how individuals deal with gains and losses better than EUT. On the other hand, EUT is designed as a normative theory and as such, you would expect that the decision made will be logical and rational. If we use PT, the decision can be affected by how most people act.

Once the individual has decided the level of independence and which of the theories they prefer, we are ready to apply the theory of choice and get the outcome of the decision problem.

5. Solving multiattribute problems: breakdown cover

5.1. ***Introducing the example: purchasing breakdown cover.*** Let us consider a real-life application where an individual is aiming to purchase breakdown cover for their car. At this point, it should be emphasised that the policies that we will be comparing are for vehicle cover for a general vehicle, with the prices a basic quote. If an individual were to actually purchase breakdown cover, they would be required to publish details of the vehicle that they have. Furthermore, we are only going to consider vehicle breakdown cover as quoted by the AA[1] and the RAC[2] on the 26th February 2018.

When deciding upon which cover to purchase, there are two things the individual must consider: the cost and the extent of the cover. This is therefore an example of a problem with two attributes to consider, as we require. To simplify notation, we are going to assign the cost per month attribute to be represented by Y and the level of cover attribute to be represented by Z.

[1] https://www.theaa.com/breakdown-cover/

[2] https://www.rac.co.uk/breakdown-cover/

policy number	level of cover	cost per month	notation
P0	no cover	£0	(£0, 0)
P1	RA	£5.50	(−£5.50, 1)
P2	RA and KR	£7.50	(−£7.50, 2)
P3	RA and AH	£8.00	(−£8.00, 3)
P4	RA, AH and NR	£10.00	(−£10.00, 4)
P5	RA, NR and OT	£12.00	(−£12.00, 5)
P6	RA, AH and KR	£12.00	(−£12.00, 6)
P7	RA, NR and KR	£12.00	(−£12.00, 7)
P8	RA, AH, NR and OT	£12.50	(−£12.50, 8)
P9	RA, AH, NR and KR	£13.50	(−£13.50, 9)
P10	RA, AH, NR, OT and KR	£14.50	(−£14.50, 10)

Table 2. Summary of coverage costs for roadside assistance (RA), key replace (KR), at home (AH), national recovery (NR), and onward travel (OT).

Clearly, the attribute for cost is one which is going to be fairly simple to evaluate as it is a quantitative one and utilities/values can be based off the cost. However, the extent of the cover needs further specification. Using the AA and the RAC as the potential providers, the possible policies we are going to compare and evaluate can be seen in Table 2. To simplify the notation, we have used a numbering system to represent each level of cover. For policy PI where $I \in \{0, 1, 2, \ldots, 10\}$, we represent the corresponding level of cover for that policy with the number I. Note that the cost per month is a negative value as it is an amount of money that the individual will spend on the cover. This is all summarised in Table 2.

Full details of what is included in each level of cover can be found in footnotes 1 and 2. We are going to assume that individuals prefer to minimise expenditure (and hence maximise the attribute of money). Furthermore, we will assume individuals prefer having a higher level of cover. However, differences in preferences between things such as key replace and at home cover will need to be specified by the individual in the stating of their utilities/values.

A significant difference between attributes Y and Z comes in that Y is a continuous attribute (to 2 decimal places), whereas Z is a discrete attribute. This has a significant impact on the choice of the reference point for PT, as was discussed in Section 4.1.

5.2. ***Formulating the problem.*** We are now going to apply the process discussed in Section 4 to the example of purchasing breakdown cover for a vehicle. In the introduction to the example, we have completed many of the tasks required in the formulating of the problem. We (the authors) will be the ones who will be faced with the decision problem and hence the outcome will be based on our preferences

and beliefs. We have specified that we are aiming to choose a single best policy from the options P0–P10, have identified that the attributes are cost per month (Y) and level of cover (Z), and identified the relevant values they can take.

The only part of formulating the problem which we have not completed in the introduction is specifying a rough preference ordering between the values that the attributes can take. For attribute Y which represents cost per month, it is trivial to suggest that the decision maker will aim to reduce the expenditure and hence maximise the money they have. However, for attribute Z which represents level of cover, it is slightly more difficult. While we can easily say that the decision maker would prefer more cover to less cover, it is hard to distinguish between some of the options which have similar levels of cover but with slightly different things involved.

Let us suppose we have our rough preference ordering as follows:

$$0 \prec 1 \prec 2 \prec 3 \prec 6 \prec 7 \prec 4 \prec 5 \prec 9 \prec 8 \prec 10.$$

This is built on the rough idea that we prefer onward travel (OT) to national recovery (NR) to at home (AH) cover to key replace (KR). Now that we have specified this rough preference ordering, we can simplify the problem significantly. For example, we can remove P9 as it is dominated by P8 due to having a lower cost per month and a more preferable level of cover. Similarly, P6 and P7 are dominated by P4. As such, we can remove policies P6, P7 and P9 from the problem and we are left with eight different policies to choose between. We have now fully formulated the problem and can move on to the next stage.

5.3. ***Independence and choosing a theory.*** Now that we have formulated the problem, we need to consider whether we have any independence between the attributes Y and Z. At this point, we can notice we have eight alternatives and, as such, the difference in the number of choices between the different methods is not significant. Therefore, we should base the choice of independence on the tests discussed at the start of Section 4.2. For this particular example, we are going to cover both levels of independence assumption and compare the outcomes.

We are now going to move on to applying the theory. Usually at this stage, we would make a decision on whether we want to use EUT or PT. However, we are going to apply both theories for each case of independence and compare the different outcomes.

5.4. ***Applying the theories.*** In this section we will be applying the multiattribute PT formulas to the example. Within this, we will analyse the impact that each choice we make will have and will compare the outcomes to those when using multiattribute EUT. Applying multiattribute EUT to the example is presented in Appendix A.

5.4.1. *Mutually independent attributes.* We will begin by considering the attributes as mutually independent and look to apply Theorem 3 and PT to the example.

The instance of mutually independent attributes and applying EUT can be seen in Appendix A.1. An interesting consideration is how the choice of reference point impacts the decision that is made. This is something we will explore within this section. With the eight policies we are considering, we know that we will have 24 $(= 2n + 8)$ choices that we need to make.

We consider three different choices for the reference points. They are as follows:

Choice 1: $y_0 = -£10.00, \quad z_0 = 4.$

Choice 2: $y_0 = -£12.00, \quad z_0 = 3.$

Choice 3: $y_0 = -£9.00, \quad z_0 = 5.$

At this point, we can decide on what points we want to choose as y_1, y_{-1}, z_1, z_{-1} with $y_1 \succ y_0 \succ y_{-1}$ and $z_1 \succ z_0 \succ z_{-1}$. We will choose these points now and use the same points for each of the choices of reference points. We will choose $y_1 = -£7.50$, $y_{-1} = -£12.50$, $z_1 = 8$ and $z_{-1} = 2$. As in the EUT case, these choices are arbitrary.

However, unlike the EUT case, we choose what values we give to them rather than in the EUT case where it had utility 1. The values that are chosen have an impact on the coefficient of the $v(y_0, z)v(y, z_0)$ term (which could be considered as an interaction term) and so have a similar impact to the impact that the constants chosen in the EUT case have. These will therefore need to be specified for each choice of reference point. The values chosen can be seen in Appendix B.1.

Now that we have decided upon these values, we are ready to state the conditional value functions for each attribute Y and Z. In other words, we are now going to decide the values $v(y, z_0)$ for all $y \in \mathrm{Y}$ and $v(y_0, z)$ for all $z \in \mathrm{Z}$. The choices made were based on the previously stated preference ordering and can be found in Appendix B.1.

After these values have been stated, we can now evaluate the values of each of the policies using (12) from Theorem 3. This then gives us the following:

$v(\text{policy}) = v(y, z)$	choice 1	choice 2	choice 3
P0 = (£0, 0)	0.645	−3.501	−0.663
P1 = (−£5.50, 1)	0.401	−0.973	−0.370
P2 = (−£7.50, 2)	0.200	.750	−0.350
P3 = (−£8.00, 3)	0.252	**1.350**	−0.276
P4 = (−£10.00, 4)	0.000	1.080	−0.632
P5 = (−£12.00, 5)	−0.038	0.800	−0.750
P8 = (−£12.50, 8)	0.400	0.700	−0.150
P10 = (−£14.50, 10)	**0.972**	−0.238	**1.210**

We have highlighted the best policy under each choice in bold. We can see that with choices 1 and 3 for the reference point, we choose policy P10 to be the best option. However, when we have choice 2 for the reference point, we choose policy P3. In fact, for choice 2, policy P10 was one of the worst options. This is clearly significantly different and shows how changing the reference point can impact what decision is made regarding an "acceptable" level of cover for the appropriate cost.

5.4.2. *One independent attribute.* We are now going to focus on the instance where we have independence in one direction. In this paper, we will focus on the case where Z is independent of Y. This means we will be applying Theorem 4 from Section 3.2 to the breakdown cover example. As such, it will require 31 ($= 3n + 7$) choices to be made by the decision maker, more than has been required in any other circumstance. We begin by deciding on the reference point (y_0, z_0). When making this decision in the mutual independence case, we considered several different choices of reference point and considered the impact they would have on the final decision. Therefore, we are once again going to consider the same three pairs of reference points and compare the outcomes.

Following the choice of reference point, we must choose $z_1, z_{-1} \in \mathrm{Z}$ such that $z_1 \succ z_0 \succ z_{-1}$. These choices need to be carefully made as we need z_1 such that $v(y_0, z_1) = 1$ and z_{-1} such that $v(y_0, z_{-1}) = -1$. Let our choices be the following:

Choice 1: $y_0 = -£10.00, \quad z_0 = 4, \quad z_1 = 8, \quad z_{-1} = 2.$

Choice 2: $y_0 = -£12.00, \quad z_0 = 3, \quad z_1 = 5, \quad z_{-1} = 1.$

Choice 3: $y_0 = -£9.00, \quad z_0 = 5, \quad z_1 = 8, \quad z_{-1} = 3.$

Once these choices have been made, we need to fix the units of measure of $v(y, z_0)$, $v(y, z_1)$ and $v(y, z_{-1})$. As with the other theories, this is done by finding equivalences. So, for example, to fix the unit of measure for $v(y, z_0)$, we need to choose $y_2, y_{-2} \in \mathrm{Y}$ and $z_2, z_{-2} \in \mathrm{Z}$ such that $(y_0, z_2) \sim (y_2, z_0)$ and $(y_0, z_{-2}) \sim (y_{-2}, z_0)$. Similarly, we need $y_3, y_{-3}, y_4, y_{-4} \in \mathrm{Y}$ and $z_3, z_{-3}, z_4, z_{-4} \in \mathrm{Z}$ such that we have $(y_3, z_1) \sim (y_0, z_3)$, $(y_{-3}, z_1) \sim (y_0, z_{-3})$, $(y_4, z_{-1}) \sim (y_0, z_4)$ and $(y_{-4}, z_{-1}) \sim (y_0, z_{-4})$. The choices we will make for each choice of reference point are:

Choice 1:
$$z_2 = z_3 = z_4 = 5, \quad z_{-2} = z_{-3} = z_{-4} = 3,$$
$$y_2 = -£8.00, \quad y_3 = -£12.00, \quad y_4 = -£6.00,$$
$$y_{-2} = -£12.50, \quad y_{-3} = -£14.50, \quad y_{-4} = -£9.00.$$

Choice 2:
$$z_2 = z_3 = z_4 = 4, \quad z_{-2} = z_{-3} = z_{-4} = 2,$$
$$y_2 = -£10.00, \quad y_3 = -£13.50, \quad y_4 = -£8.00,$$
$$y_{-2} = -£13.50, \quad y_{-3} = -£16.50, \quad y_{-4} = -£10.00.$$

Choice 3: $z_2 = z_3 = z_4 = 10, \quad z_{-2} = z_{-3} = z_{-4} = 4,$
$y_2 = -£7.50, \quad y_3 = -£8.00, \quad y_4 = -£2.50$
$y_{-2} = -£11.00, \quad y_{-3} = -£13.00, \quad y_{-4} = -£8.00.$

Notice here that similar values of attribute Z are chosen to formulate equivalences with. This is because Z is a discrete attribute with only eight conceivable outcomes so equivalences would be difficult to work with if changing Z at all times. Furthermore, as Y is continuous to two decimal places, we can fix Z at certain levels and change the attribute Y to allow for the required equivalence relations.

Now that these choices have been made, we can state the conditional value functions $v(y_0, z)$ for all $z \in \mathrm{Z}$ and $v(y, z_{-1})$, $v(y, z_0)$, $v(y, z_1)$ for all $y \in \mathrm{Y}$. These can be found in Appendix B.2.

Now that all these have been specified, we are in a position to be able to evaluate the value of each of the policies. The results are as follows:

$v(\text{policy}) = v(y, z)$	choice 1	choice 2	choice 3
P0 = (£0, 0)	**0.735**	1.100	**0.715**
P1 = (−£5.50, 1)	0.413	0.900	0.100
P2 = (−£7.50, 2)	−0.100	0.685	−0.725
P3 = (−£8.00, 3)	−0.020	0.750	−0.450
P4 = (−£10.00, 4)	0.000	0.780	−0.695
P5 = (−£12.00, 5)	−0.140	1.000	−0.600
P8 = (−£12.50, 8)	0.200	**1.110**	−0.100
P10 = (−£14.50, 10)	−0.375	0.475	−0.380

As we can see, in this situation, the best policy is P0 for choices 1 and 3 and P8 for choice 2. It is interesting to see here that the best policy for choices 1 and 3 is to have no breakdown cover at all. Furthermore, for choice 2, policy P0 is very close to being chosen as the best option. This clearly suggests that for this context, there is an underlying attitude that losing money is worse than having a lower level of cover.

While choosing cheaper policies has an obvious financial benefit in the short term, it could lead to financial problems in the long term if the vehicle encountered problems. This is something we could potentially account for if we included probabilities and the corresponding financial outlay if problems occurred. However, it would reduce it to a single attribute problem in which the costs and corresponding probabilities are difficult to construct.

6. Concluding remarks

In this paper, we have applied PT to multiattribute problems with different independence assumptions. We have also introduced a process which can be followed to

assist in effectively and efficiently solving a multiattribute decision problem. Using these two developments allows an individual to apply PT to a decision problem with two attributes. This is especially useful for problems where there are gains and losses to consider with respect to certain attributes.

An example of where this can be useful is seen in Section 5 where we apply the results of this paper to a real-life example of purchasing breakdown cover. If you already had a certain level of breakdown cover, it is easy to see how you could consider the different costs and levels of cover as gains and losses. This shows how multiattribute PT can be useful in solving real-life problems. However, there is a lot more further research which can be done into applying PT to multiattribute problems.

First, extending PT so that it can be applied to problems with more than two attributes is interesting. This may require different methods to be used, although independence assumptions could still be useful. Secondly, research into the reference point in PT will be interesting. Can the reference point be adapted so that you can choose a set of values as a reference point? This would allow the individual to select an acceptable region as their reference point and then consider gains and losses from that set. Finally, this paper focusses on the instance where we have at least a certain level of independence assumption between the attributes. Could PT be adapted so that we do not require any independence between the attributes?

Appendix A: Breakdown cover EUT calculations

A.1. ***Mutually independent attributes — multiattribute EUT.*** Using multiattribute EUT for this context of mutually independent attributes, we will be applying Theorem 5.2 from [Keeney and Raiffa 1976, p. 234]. For this instance, we choose $y_0 = -£14.50,\ y_1 = -£12.00,\ y_* = 0,\ z_0 = 0,\ z_1 = 4$ and $z_* = 10$.

We are now required to choose the values for the constants k_Y and k_Z which subsequently decide k_{YZ} and k. These choices have a significant impact on the overall utilities we assign to each of the policies. As such, we are going to consider three possible pairs of values and see how changing these values affects the overall utilities and decisions. The pairs of values we are going to consider are as follows:

Choice 1: $k_Y = 0.2, \quad k_Z = 0.35.$

Choice 2: $k_Y = k_Z = 0.5.$

Choice 3: $k_Y = 0.6, \quad k_Z = 0.8.$

So, with choice 1, we have that $k_{YZ} = 0.45$ and $k = \frac{45}{7} > 0$, which implies that the attributes Y and Z are complimentary. Choice 2 would give the formulas from additive independence and would also imply there is no interaction between the

attributes. Choice 3 would make $k_{YZ} = -0.4$ and $k = -\frac{5}{6} < 0$, which implies the attributes Y and Z are substitutes.

After specifying the values for the constants, we have now fully defined what we need for the theorem and hence can begin to evaluate the utilities for all of the policies. To do this, we are going to consider the conditional utility functions for each of the attributes Y and Z and evaluate a utility for each alternative. We choose the utility values as follows:

$$\begin{aligned} u_Y(-£14.50) &= 0, & u_Y(-£8.00) &= 1.75, \\ u_Y(-£12.50) &= 0.8, & u_Y(-£7.50) &= 1.9, \\ u_Y(-£12.00) &= 1, & u_Y(-£5.50) &= 2.2, \\ u_Y(-£10.00) &= 1.35, & u_Y(£0) &= 2.9, \end{aligned}$$

$$\begin{aligned} u_Z(0) &= 0, & u_Z(4) &= 1, \\ u_Z(1) &= 0.25, & u_Z(5) &= 1.25, \\ u_Z(2) &= 0.4, & u_Z(8) &= 1.6, \\ u_Z(3) &= 0.65, & u_Z(10) &= 2.2. \end{aligned}$$

The result of choosing these utility values and applying Theorem 5.2 is the following:

$u(\text{policy}) = u(y, z)$	choice 1	choice 2	choice 3
P0 = (£0, 0)	0.580	**1.450**	1.740
P1 = (−£5.50, 1)	0.775	1.225	1.300
P2 = (−£7.50, 2)	0.862	1.150	1.156
P3 = (−£8.00, 3)	1.089	1.200	1.115
P4 = (−£10.00, 4)	1.228	1.175	1.070
P5 = (−£12.00, 5)	1.200	1.125	1.100
P8 = (−£12.50, 8)	**1.296**	1.200	1.248
P10 = (−£14.50, 10)	0.770	1.100	**1.760**

We have highlighted the best option for each choice in bold. We can see that for choice 1, the best option is policy P8; for choice 2, the best option is policy P0 and for choice 3, the best option is P10. It is not a surprise to see one of the more extreme options (i.e., P0 or P10) being the chosen option for choice 3 as this was the instance where the attributes are substitutes.

A.2. ***One independent attribute — multiattribute EUT.*** We are now going to focus on the instance where we only have one attribute which is independent of the other. For this paper, we are going to focus on the instance where we have Z independent of Y. This involves applying Theorem 5.6 from [Keeney and Raiffa 1976, p. 244].

We begin by choosing $y_0 \in \mathrm{Y}$ and $z_0, z_1 \in \mathrm{Z}$ such that $u(y_0, z_0) = 0$ and $u(y_0, z_1) = 1$. In this instance, we are going to choose $y_0 = -£14.50,\ z_0 = 0$ (as in the mutual independence case) and $z_1 = 4$. Following this, we are required to fix the unit of measure of $u(y, z_0)$ and $u(y, z_1)$. This comes from choosing $y_2 \in \mathrm{Y}$, $z_2 \in \mathrm{Z}$ such that $(y_0, z_2) \sim (y_2, z_0)$ and, similarly, $y_3 \in \mathrm{Y}$, $z_3 \in \mathrm{Z}$ such that $(y_0, z_3) \sim (y_3, z_1)$. Let us choose $y_2 = -£5.50,\ z_2 = 3,\ y_3 = -£12.00$ and $z_3 = 8$.

Now that these decisions have been made, we are required to evaluate the conditional utility functions $u(y_0, z)$ for all $z \in \mathrm{Z}$ and $u(y, z_0), u(y, z_1)$ for all $y \in \mathrm{Y}$. To begin with, we assign the utilities $u(y_0, z)$ for all $z \in \mathrm{Z}$ and using these, then assign the remaining $u(y, z_0)$ and $u(y, z_1)$ for all $y \in \mathrm{Y}$. The choices we make are as follows:

$$
\begin{aligned}
u(y_0, 0) &= 0, & u(y_0, 4) &= 1,\\
u(y_0, 1) &= 0.2, & u(y_0, 5) &= 1.25,\\
u(y_0, 2) &= 0.45, & u(y_0, 8) &= 1.6,\\
u(y_0, 3) &= 0.8, & u(y_0, 10) &= 2.1,
\end{aligned}
$$

$$
\begin{aligned}
u(-£14.50, z_0) &= 0, & u(-£8.00, z_0) &= 0.45,\\
u(-£12.50, z_0) &= 0.1, & u(-£7.50, z_0) &= 0.55,\\
u(-£12.00, z_0) &= 0.15, & u(-£5.50, z_0) &= 0.8,\\
u(-£10.00, z_0) &= 0.25, & u(£0.00, z_0) &= 1.25,
\end{aligned}
$$

$$
\begin{aligned}
u(-£14.50, z_1) &= 1, & u(-£8.00, z_1) &= 2.6,\\
u(-£12.50, z_1) &= 1.6, & u(-£7.50, z_1) &= 2.7,\\
u(-£12.00, z_1) &= 1.8, & u(-£5.50, z_1) &= 3.1,\\
u(-£10.00, z_1) &= 2.25, & u(£0.00, z_1) &= 3.7.
\end{aligned}
$$

Now that these choices have been made, we can calculate the utilities of each of the policies and choose the one with the highest utility value. The corresponding utilities are

$$
\begin{aligned}
u(\mathrm{P0}) &= 1.25, & u(\mathrm{P4}) &= 2.25,\\
u(\mathrm{P1}) &= 1.26, & u(\mathrm{P5}) &= 2.2125,\\
u(\mathrm{P2}) &= 1.5175, & \boldsymbol{u}(\mathbf{P8}) &= \mathbf{2.5},\\
u(\mathrm{P3}) &= 2.17, & u(\mathrm{P10}) &= 2.1.
\end{aligned}
$$

We can see from this that when we have Z independent of Y, we choose policy P8 to be the best policy. P8 was the same policy that we chose for choice 1 of the mutually independent EUT case but different from all other choices we have made so far. This shows how different choices can come from the same decision maker dependent on how they formulate the problem and choose independence.

Appendix B: Breakdown cover PT choices

B.1. *Mutually independent attributes.*

Choice 1: $v(y_1, z_1) = 1.7$, $v(y_1, z_0) = 0.8$, $v(y_1, z_{-1}) = 0.2$,
$v(y_0, z_1) = 0.85$, $v(y_0, z_0) = 0$, $v(y_0, z_{-1}) = -0.7$,
$v(y_{-1}, z_1) = 0.4$, $v(y_{-1}, z_0) = -0.5$, $v(y_{-1}, z_{-1}) = -1.3$.

Choice 2: $v(y_1, z_1) = 2.5$, $v(y_1, z_0) = 1.6$, $v(y_1, z_{-1}) = 0.75$,
$v(y_0, z_1) = 1.4$, $v(y_0, z_0) = 0$, $v(y_0, z_{-1}) = -0.3$,
$v(y_{-1}, z_1) = 0.7$, $v(y_{-1}, z_0) = -0.4$, $v(y_{-1}, z_{-1}) = -0.7$.

Choice 3: $v(y_1, z_1) = 1.1$, $v(y_1, z_0) = 0.65$, $v(y_1, z_{-1}) = -0.35$,
$v(y_0, z_1) = 0.4$, $v(y_0, z_0) = 0$, $v(y_0, z_{-1}) = -0.95$,
$v(y_{-1}, z_1) = -0.15$, $v(y_{-1}, z_0) = -0.9$, $v(y_{-1}, z_{-1}) = -1.7$.

Our choices for $v(y, z_0)$ for all possible $y \in \mathrm{Y}$ are as follows:

Choice 1: $v(£0, z_0) = 1.9$, $v(-£10.00, z_0) = 0$,
$v(-£5.50, z_0) = 1.35$, $v(-£12.00, z_0) = -0.35$,
$v(-£7.50, z_0) = 0.8$, $v(-£12.50, z_0) = -0.5$,
$v(-£8.00, z_0) = 0.65$, $v(-£14.50, z_0) = -1.05$.

Choice 2: $v(£0, z_0) = 3.1$, $v(-£10.00, z_0) = 0.7$,
$v(-£5.50, z_0) = 2.35$, $v(-£12.00, z_0) = 0$,
$v(-£7.50, z_0) = 1.6$, $v(-£12.50, z_0) = -0.4$,
$v(-£8.00, z_0) = 1.35$, $v(-£14.50, z_0) = -1.1$.

Choice 3: $v(£0, z_0) = 1.8$, $v(-£10.00, z_0) = -0.25$,
$v(-£5.50, z_0) = 1.1$, $v(-£12.00, z_0) = -0.75$,
$v(-£7.50, z_0) = 0.65$, $v(-£12.50, z_0) = -0.9$,
$v(-£8.00, z_0) = 0.45$, $v(-£14.50, z_0) = -1.6$.

Similarly, we can state our values $v(y_0, z)$ for all $z \in \mathrm{Z}$ as:

Choice 1: $v(y_0, 0) = -1.9$, $v(y_0, 3) = -0.45$, $v(y_0, 8) = 0.85$,
$v(y_0, 1) = -1.25$, $v(y_0, 4) = 0$, $v(y_0, 10) = 1.8$.
$v(y_0, 2) = -0.7$, $v(y_0, 5) = 0.3$,

Choice 2: $v(y_0, 0) = -1.45$, $v(y_0, 3) = 0$, $v(y_0, 8) = 1.4$,
$v(y_0, 1) = -0.9$, $v(y_0, 4) = 0.45$, $v(y_0, 10) = 2.1$.
$v(y_0, 2) = -0.3$, $v(y_0, 5) = 0.8$,

Choice 3: $v(y_0, 0) = -2.15$, $v(y_0, 3) = -0.7$, $v(y_0, 8) = 0.4$,
$v(y_0, 1) = -1.35$, $v(y_0, 4) = -0.4$, $v(y_0, 10) = 1.1$.
$v(y_0, 2) = -0.95$, $v(y_0, 5) = 0$,

B.2. ***Single independence.*** We evaluate $v(y_0, z)$ for all $z \in \mathrm{Z}$ as:

Choice 1: $v(y_0, 0) = -1.7$, $v(y_0, 3) = -0.6$, $v(y_0, 8) = 1$,
$v(y_0, 1) = -1.25$, $v(y_0, 4) = 0$, $v(y_0, 10) = 1.45$.
$v(y_0, 2) = -1$, $v(y_0, 5) = 0.4$,

Choice 2: $v(y_0, 0) = -1.5$, $v(y_0, 3) = 0$, $v(y_0, 8) = 1.4$,
$v(y_0, 1) = -1$, $v(y_0, 4) = 0.4$, $v(y_0, 10) = 1.75$.
$v(y_0, 2) = -0.55$, $v(y_0, 5) = 1$,

Choice 3: $v(y_0, 0) = -2.1$, $v(y_0, 3) = -1$, $v(y_0, 8) = 1$,
$v(y_0, 1) = -1.7$, $v(y_0, 4) = -0.45$, $v(y_0, 10) = 1.45$.
$v(y_0, 2) = -1.45$, $v(y_0, 5) = 0$,

Following this, we evaluate $v(y, z_0)$ for all $y \in \mathrm{Y}$ as:

Choice 1: $v(-£14.50, z_0) = -1.1$, $v(-£8.00, z_0) = 0.4$,
$v(-£12.50, z_0) = -0.6$, $v(-£7.50, z_0) = 0.45$,
$v(-£12.00, z_0) = -0.5$, $v(-£5.50, z_0) = 0.85$,
$v(-£10.00, z_0) = 0$, $v(£0.00, z_0) = 1.5$.

Choice 2: $v(-£14.50, z_0) = -0.75$, $v(-£8.00, z_0) = 0.75$,
$v(-£12.50, z_0) = -0.15$, $v(-£7.50, z_0) = 0.85$,
$v(-£12.00, z_0) = 0$, $v(-£5.50, z_0) = 1.2$,
$v(-£10.00, z_0) = 0.4$, $v(£0.00, z_0) = 1.85$.

Choice 3: $v(-£14.50, z_0) = -1.25$, $v(-£8.00, z_0) = 1.05$,
$v(-£12.50, z_0) = -0.8$, $v(-£7.50, z_0) = 1.45$,
$v(-£12.00, z_0) = -0.6$, $v(-£5.50, z_0) = 1.8$,
$v(-£10.00, z_0) = -0.2$, $v(£0.00, z_0) = 2.5$.

Similarly, we evaluate $v(y, z_1)$ for all $y \in \mathrm{Y}$ as:

Choice 1: $v(-£14.50, z_1) = -0.6$, $v(-£8.00, z_1) = 1.4$,
$v(-£12.50, z_1) = 0.2$, $v(-£7.50, z_1) = 1.55$,
$v(-£12.00, z_1) = 0.4$, $v(-£5.50, z_1) = 1.8$,
$v(-£10.00, z_1) = 1$, $v(£0.00, z_1) = 2.4$.

Choice 2: $v(-£14.50, z_1) = -0.05$, $v(-£8.00, z_1) = 1.8$,
$v(-£12.50, z_1) = 0.75$, $v(-£7.50) = 1.9$,
$v(-£12.00, z_1) = 1$, $v(-£5.50, z_1) = 2.35$,
$v(-£10.00, z_1) = 1.35$, $v(£0.00, z_1) = 2.9$.

Choice 3: $v(-£14.50, z_1) = -0.65$, $v(-£8.00, z_1) = 1.45$,
$v(-£12.50, z_1) = -0.1$, $v(-£7.50, z_1) = 1.6$,
$v(-£12.00, z_1) = 0.3$, $v(-£5.50, z_1) = 2.05$,
$v(-£10.00, z_1) = 0.85$, $v(£0.00, z_1) = 2.75$.

Finally, we evaluate $v(y, z_{-1})$ for all $y \in \mathrm{Y}$ as:

Choice 1: $v(-£14.50, z_{-1}) = -2.05$, $v(-£8.00, z_{-1}) = -0.3$,
$v(-£12.50, z_{-1}) = -1.65$, $v(-£7.50, z_{-1}) = -0.1$,
$v(-£12.00, z_{-1}) = -1.5$, $v(-£5.50, z_{-1}) = 0.5$,
$v(-£10.00, z_{-1}) = -1$, $v(£0, z_{-1}) = 1.05$.

Choice 2: $v(-£14.50, z_{-1}) = -1.6$, $v(-£8.00, z_{-1}) = 0.4$,
$v(-£12.50, z_{-1}) = -1.15$, $v(-£7.50, z_{-1}) = 0.55$,
$v(-£12.00, z_{-1}) = -1$, $v(-£5.50, z_{-1}) = 0.9$,
$v(-£10.00, z_{-1}) = -0.55$, $v(£0.00, z_{-1}) = 1.35$.

Choice 3: $v(-£14.50, z_{-1}) = -2.55$, $v(-£8.00, z_{-1}) = -0.45$,
$v(-£12.50, z_{-1}) = -2$, $v(-£7.50, z_{-1}) = -0.05$,
$v(-£12.00, z_{-1}) = -1.85$, $v(-£5.50, z_{-1}) = 0.8$,
$v(-£10.00, z_{-1}) = -1.3$, $v(£0, z_{-1}) = 1.65$.

Acknowledgements

The authors thank two reviewers for their enthusiastic comments on this work and for useful suggestions that have led to improved presentation.

References

[Bellman and Zadeh 1970] R. E. Bellman and L. A. Zadeh, "Decision-making in a fuzzy environment", *Management Sci.* **17**:4 (1970), B141–B164. MR Zbl

[Egozcue et al. 2014] M. Egozcue, S. Massoni, W.-K. Wong, and R. Zitikis, "Integration-segregation decisions under general value functions: 'Create your own bundle: choose 1, 2 or all 3"', *IMA J. Manag. Math.* **25**:1 (2014), 57–72. MR Zbl

[Gomes and Lima 1991] L. F. A. M. Gomes and M. M. P. P. Lima, "Todim: basics and application to multicriteria ranking of projects with environmental impacts", *Found. Comput. Decision Sci.* **16**:3-4 (1991), 113–127. Zbl

[Hu and Zhou 2009] H. Jun-hua and Z. Yi-wen, "Prospect theory based multi-criteria decision making method", pp. 2930–2934 in *Chinese Control and Decision Conference* (Guilin, China, 2009), IEEE, Piscataway, NJ, 2009. In Chinese.

[Kahneman and Tversky 1979] D. Kahneman and A. Tversky, "Prospect theory: an analysis of decision under risk", *Econometrica* **47**:2 (1979), 263–292. Zbl

[Keeney and Raiffa 1976] R. L. Keeney and H. Raiffa, *Decisions with multiple objectives: preferences and value tradeoffs*, Wiley, New York, 1976.

[Krohling and de Souza 2012] R. A. Krohling and T. T. M. de Souza, "Combining prospect theory and fuzzy numbers to multi-criteria decision making", *Expert Syst. Appl.* **39**:13 (2012), 11487–11493.

[Liu et al. 2011] P. Liu, F. Jin, X. Zhang, Y. Su, and M. Wang, "Research on the multi-attribute decision-making under risk with interval probability based on prospect theory and the uncertain linguistic variables", *Knowledge-Based Syst.* **24**:4 (2011), 554–561.

[Starmer 2000] C. Starmer, "Developments in non-expected utility theory: the hunt for a descriptive theory of choice under risk", *J. Econ. Literature* **38**:2 (2000), 332–382.

[Tversky and Kahneman 1992] A. Tversky and D. Kahneman, "Advances in prospect theory: cumulative representation of uncertainty", *J. Risk Uncertainty* **5**:4 (1992), 297–323. Zbl

[Wang and Sun 2008] J.-Q. Wang and T. Sun, "Fuzzy multiple criteria decision making method based on prospect theory", pp. 288–291 in *Int. Conf. Information Management, Innovation Management and Industrial Engineering, I* (Taipei, 2008), IEEE, Piscataway, NJ, 2008.

[Zadeh 1979] L. A. Zadeh, "Fuzzy sets", pp. 569–606 in *Operations research support methodology*, edited by A. G. Holzman, Dekker, New York, 1979.

Received: 2018-07-20 Revised: 2018-10-04 Accepted: 2018-11-15

jackstanley12@hotmail.co.uk *Department of Mathematical Sciences, Durham University, Durham, United Kingdom*

frank.coolen@durham.ac.uk *Department of Mathematical Sciences, Durham University, Durham, United Kingdom*

INVOLVE 12:4 (2019)
dx.doi.org/10.2140/involve.2019.12.713

On weight-one solvable configurations of the Lights Out puzzle

Yuki Hayata and Masakazu Yamagishi

(Communicated by Kenneth S. Berenhaut)

We show that the center-one configuration is always solvable in the Lights Out puzzle on a square grid with odd vertices.

1. Introduction

Let $\Gamma = (V, E)$ be a finite undirected simple graph, $n = \#V$ the number of vertices, and $\mathscr{F}$ the set of functions on V with values in $\mathbb{F}_2$, the field with two elements. We define the Laplacian $\Delta : \mathscr{F} \to \mathscr{F}$ by

$$(\Delta f)(v) := f(v) + \sum_{(v,w)\in E} f(w)$$

for $f \in \mathscr{F}$, $v \in V$. Let $\boldsymbol{e}_v$ denote the characteristic function of $v \in V$. Then $\{\boldsymbol{e}_v : v \in V\}$ is a basis of $\mathscr{F}$ as a vector space over $\mathbb{F}_2$, and by means of this basis we identify $\mathscr{F}$ with $\mathbb{F}_2^n$. Under this identification, Δ is a linear map represented by $I_n + \mathrm{adj}(\Gamma)$, where I_n denotes the identity matrix of degree n and $\mathrm{adj}(\Gamma)$ the adjacency matrix of Γ. Let the image and the kernel of Δ be denoted by $\mathscr{C}$ and $\mathscr{H}$, respectively. $\mathscr{C}$ is the set of solvable configurations of the Lights Out puzzle on Γ; see [Fleischer and Yu 2013; Goldwasser and Klostermeyer 1997; Goshima and Yamagishi 2010]. It is known that the all-one configuration is always solvable:

Theorem 1.1 [Sutner 1989]. *For any* Γ, *it holds that* $(1\ 1\ \cdots\ 1) \in \mathscr{C}$.

Since $\mathscr{C}$ is a linear subspace of $\mathbb{F}_2^n$, we may regard it as a binary linear code; see [Goldwasser and Klostermeyer 1997] for this point of view. The weight enumerator of $\mathscr{C}$ is defined by

$$W_{\mathscr{C}}(x, y) = \sum_{i=0}^{n} A_i x^{n-i} y^i,$$

MSC2010: primary 05C57; secondary 05C38, 91A46, 94B60.
Keywords: Lights Out, path graph, Cartesian product, linear code.

where A_i is the number of vectors in $\mathscr{C}$ which have Hamming weight i. By Sutner's theorem, we have $A_{n-i} = A_i$. If Δ is bijective, then $\mathscr{C} = \mathbb{F}_2^n$ and we have

$$A_i = \binom{n}{i}, \quad W_{\mathscr{C}}(x, y) = (x + y)^n.$$

In this paper, we are interested in A_1 of the classical $n \times n$ Lights Out puzzle. Our main result is Theorem 3.1, which states that the center-one configuration is always solvable when n is odd. Our proof is a neat application of Sutner's theorem and is not constructive. Theorem 3.1 implies in particular that the minimal distance of $\mathscr{C}$ is 1 when n is odd. For even n, it turns out that the minimal distance is at most 2.

We then look at the case $A_1 \leq 1$ more closely, and make some conjectures based on numerical computations. We also make an attempt to "explain" the value of A_1.

2. Path and cycle graphs

Before proceeding to the main result, we consider the case of path and cycle graphs as first examples.

Let $\Gamma = \boldsymbol{P}_n$ be the path graph with n vertices. We have

$$\operatorname{adj}(\Gamma) = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \ddots & \vdots \\ 0 & 1 & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix}$$

under an obvious ordering of vertices. It is well known, see [Yamagishi 2015, Lemma 3.1], that the characteristic polynomial of $\operatorname{adj}(\Gamma)$ is $S_n(x)$, the n-th Chebyshev polynomial of the second kind, defined by

$$S_0(x) = 1, \quad S_1(x) = x, \quad S_n(x) = xS_{n-1}(x) - S_{n-2}(x) \quad (n \geq 2).$$

So we see that Δ is bijective if and only if $S_n(-1) \not\equiv 0 \pmod 2$ if and only if $n \not\equiv 2 \pmod 3$.

In the case $n \equiv 2 \pmod 3$, it is easy to see that $\mathscr{H}$ is one-dimensional, spanned by the vector

$$(1\ 1\ 0\ 1\ 1\ 0\ \cdots\ 0\ 1\ 1), \tag{2-1}$$

so that

$$W_{\mathscr{H}}(x, y) = x^n + x^{(n-2)/3} y^{(2n+2)/3}.$$

Since $\mathscr{C} = \mathscr{H}^{\perp}$, we have

$$W_{\mathscr{C}}(x, y) = \tfrac{1}{2}((x + y)^n + (x + y)^{(n-2)/3}(x - y)^{(2n+2)/3}) \tag{2-2}$$

by the MacWilliams identity [MacWilliams and Sloane 1977, p. 127]. In particular, expanding (2-2), we find that

$$A_1 = \tfrac{1}{3}(n-2), \quad A_2 = \tfrac{1}{18}(5n^2 - 5n + 8).$$

Note that A_1 and A_2 can be seen more quickly as follows. In the general setting, we have $\mathscr{C} = \mathscr{H}^{\perp}$ since $\mathrm{adj}(\Gamma)$ is a symmetric matrix. Suppose $\dim \mathscr{C} = k < n$, so that $\dim \mathscr{H} = n - k > 0$. Any basis of $\mathscr{H}$ gives a parity check matrix H (of size $(n-k) \times n$) of $\mathscr{C}$, and A_i is the number of unordered i-tuples of columns of H whose sum is the zero vector. In the case $\Gamma = \boldsymbol{P}_n$, $n \equiv 2 \pmod 3$, the vector (2-1) itself is a parity check matrix, and one easily sees that

$$A_1 = \tfrac{1}{3}(n-2), \quad A_2 = \binom{\frac{1}{3}(n-2)}{2} + \binom{\frac{1}{3}(2n+2)}{2}.$$

Next let $\Gamma = \boldsymbol{C}_n$ be the cycle graph with n vertices ($n \geq 3$). It is also well known, see [Yamagishi 2015, Lemma 3.1], that Δ is bijective if and only if $C_n(-1) \not\equiv 0 \pmod 2$ if and only if $n \not\equiv 0 \pmod 3$, where $C_n(x)$ is the n-th Chebyshev polynomial of the first kind, defined by

$$C_0(x) = 2, \quad C_1(x) = x, \quad C_n(x) = xC_{n-1}(x) - C_{n-2}(x) \quad (n \geq 2).$$

In the case $n \equiv 0 \pmod 3$, it is easy to see that $\mathscr{H}$ is two-dimensional, spanned by the row vectors of

$$\begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 0 & \cdots & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & \cdots & 1 & 0 & 1 \end{pmatrix}, \tag{2-3}$$

so that

$$W_{\mathscr{H}}(x, y) = x^n + 3x^{n/3}y^{2n/3},$$
$$W_{\mathscr{C}}(x, y) = \tfrac{1}{4}((x+y)^n + 3(x+y)^{n/3}(x-y)^{2n/3}).$$

In particular, we obtain

$$A_1 = 0, \quad A_2 = \tfrac{1}{6}(n^2 - 3n).$$

As explained above, A_1 and A_2 can be seen directly from (2-3). This is clear for A_1. Since i-th and j-th columns add to zero if and only if $i \equiv j \pmod 3$, we see that $A_2 = \frac{1}{2}n\big(\frac{1}{3}n - 1\big)$. We also have an alternative proof for $A_1 = 0$ as follows. Suppose there is a vector in $\mathscr{C}$ with Hamming weight 1. Then any vector with Hamming weight 1 belongs to $\mathscr{C}$ since Δ commutes with "shifts". This implies $\mathscr{C} = \mathbb{F}_2^n$, which contradicts $n \equiv 0 \pmod 3$.

3. The main theorem

In the following, we let Γ be the Cartesian product $\boldsymbol{P}_n \times \boldsymbol{P}_n$, forgetting the previous meaning of n as the number of vertices. The corresponding objects V, $\mathscr{F}$, Δ, $\mathscr{C}$, $\mathscr{H}$,

and A_i will be denoted by V_n, $\mathscr{F}_n$, Δ_n, $\mathscr{C}_n$, $\mathscr{H}_n$, and $A_i(n)$, respectively. We use double indices for the vertices in a natural way:

$$V_n = \{v_{i,j} : 1 \le i, j \le n\},$$
$$v_{i,j} \text{ and } v_{k,l} \text{ are adjacent} \iff |i-k| + |j-l| = 1.$$

Let $\boldsymbol{e}_{i,j}$ denote the characteristic function of $v_{i,j}$.

The main result of this paper is the following, which states that the center-one configuration is always solvable in the Lights Out puzzle on $\boldsymbol{P}_n \times \boldsymbol{P}_n$ when n is odd.

Theorem 3.1. *If $n = 2m+1$ ($m \ge 0$), then $\boldsymbol{e}_{m+1,m+1} \in \mathscr{C}_n$.*

Proof. The case $m = 0$ is trivial since Δ_1 is the identity map, so we suppose $m \ge 1$. We identify a function $f \in \mathscr{F}_n$ with the matrix $(a_{i,j})$ such that

$$f = \sum_{1 \le i,j \le n} a_{i,j} \boldsymbol{e}_{i,j} \quad (a_{i,j} \in \mathbb{F}_2).$$

Let $\mathbf{1}_{a,b}$ denote the $a \times b$ matrix whose entries are all 1, and $\mathbf{0}$ the zero matrix whose size will be clear from the context. Sutner's theorem states that $\mathbf{1}_{n,n} \in \mathscr{C}_n$. Applying Sutner's theorem to $\boldsymbol{P}_m \times \boldsymbol{P}_m$, we see that

$$f_1 := \begin{pmatrix} \mathbf{1}_{m,m} & \boldsymbol{x} & \mathbf{0} \\ \boldsymbol{y} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \in \mathscr{C}_n$$

for a suitable column vector $\boldsymbol{x}$ and a row vector $\boldsymbol{y}$. Since $\mathscr{C}_n$ is invariant under horizontal reflection, say α, and vertical reflection, say β, we find that

$$f_2 := \begin{pmatrix} \mathbf{1}_{m,m} & \mathbf{0} & \mathbf{1}_{m,m} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{1}_{m,m} & \mathbf{0} & \mathbf{1}_{m,m} \end{pmatrix} = f_1 + \alpha(f_1) + \beta(f_1) + \alpha\beta(f_1) \in \mathscr{C}_n.$$

Similarly, we have

$$f_3 := (\mathbf{1}_{n,m} \;\; \boldsymbol{z} \;\; \mathbf{0}) \in \mathscr{C}_n$$

for a suitable column vector $\boldsymbol{z}$, so that

$$f_4 := (\mathbf{1}_{n,m} \;\; \mathbf{0} \;\; \mathbf{1}_{n,m}) = f_3 + \alpha(f_3) \in \mathscr{C}_n,$$

and likewise,

$$f_5 := \begin{pmatrix} \mathbf{1}_{m,n} \\ \mathbf{0} \\ \mathbf{1}_{m,n} \end{pmatrix} \in \mathscr{C}_n.$$

Therefore we have

$$\boldsymbol{e}_{m+1,m+1} = f_2 + f_4 + f_5 + \mathbf{1}_{n,n} \in \mathscr{C}_n$$

as desired. □

Remark 3.2. Our proof is not constructive; in the context of Lights Out puzzle, we only know that $\boldsymbol{e}_{m+1,m+1}$ is solvable, but do not know any solution (an inverse image of $\boldsymbol{e}_{m+1,m+1}$ under Δ_n). It would be interesting to find out a unified description of a solution of $\boldsymbol{e}_{m+1,m+1}$.

Remark 3.3. The center-one configuration is the only universal solvable configuration of weight 1, since $A_1(n) = 1$ for some (infinitely many, under Conjecture 4.4 below) odd integers n.

Since $A_1(n)$ is the number of $\boldsymbol{e}_{i,j}$'s contained in $\mathscr{C}_n$, taking symmetry (i.e., invariance of $\mathscr{C}_n$ under the horizontal and vertical reflections) into account, we have:

Corollary 3.4. $A_1(n) \equiv 1 \pmod 4$ *if* n *is odd.* $A_1(n) \equiv 0 \pmod 4$ *if* n *is even.*

Let d_n denote the minimal distance of the linear code $\mathscr{C}_n$. By Theorem 3.1, we have $d_n = 1$ for odd n. We see that $d_n \leq 2$ in general by the following:

Lemma 3.5. *For* $n \geq 4$, *we have* $\boldsymbol{e}_{1,4} + \boldsymbol{e}_{3,2} \in \mathscr{C}_n$.

Proof. We have $\boldsymbol{e}_{1,4} + \boldsymbol{e}_{3,2} = \Delta_n(\boldsymbol{e}_{1,1} + \boldsymbol{e}_{1,2} + \boldsymbol{e}_{1,3} + \boldsymbol{e}_{2,2}) \in \mathscr{C}_n$. □

Note that $d_2 = 1$ since Δ_2 is bijective. Thus the determination of d_n is equivalent to answering the following:

Problem 3.6. Characterize (necessarily even) n such that $A_1(n) = 0$.

4. The case $A_1(n) \leq 1$

With the same notation as in the previous section, we consider the case $A_1(n) \leq 1$.

A first look at Table 1 leads to the following two conjectures.

Conjecture 4.1. *If* $A_1(n) = 0$, *then* $n + 1 = 2^l \pm 1$ *for some* $l \geq 2$.

Conjecture 4.2. *Let* $n \geq 2$. *We have* $A_1(n) \leq 1$ *if and only if* $A_1(2n+1) \leq 1$.

The "if" part of Conjecture 4.2 follows from:

Proposition 4.3. *We have* $A_i(n) \leq A_i(2n+1)$ *for* $n \geq 1$ *and* $0 \leq i \leq n$.

Proof. We define a map $\iota_n : \mathscr{F}_n \to \mathscr{F}_{2n+1}$ by

$$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{pmatrix} \mapsto \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & a_{1,1} & 0 & a_{1,2} & \cdots & a_{1,n} & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & a_{2,1} & 0 & a_{2,2} & \cdots & a_{2,n} & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & a_{n,1} & 0 & a_{n,2} & \cdots & a_{n,n} & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix},$$

which is an analog of $\iota^{\pm}_{m,n}$ used in [Goshima and Yamagishi 2010] for $\boldsymbol{C}_m \times \boldsymbol{C}_n$. One can then verify the identity $\iota_n \Delta_n = \Delta^2_{2n+1} \iota_n$, so it follows that $\iota_n(\mathscr{C}_n) \subset \mathscr{C}_{2n+1}$. Since ι_n preserves the Hamming weight, we have $A_i(n) \leq A_i(2n+1)$ for $0 \leq i \leq n$. □

n	$A_1(n)$	$\dim \mathscr{H}_n$	n	$A_1(n)$	$\dim \mathscr{H}_n$	n	$A_1(n)$	$\dim \mathscr{H}_n$	n	$A_1(n)$	$\dim \mathscr{H}_n$
1	1	0	41	701	2	81	6561	0	121	14641	0
2	4	0	42	1764	0	82	6724	0	122	14884	0
3	9	0	43	1849	0	83	1401	6	123	1	80
4	0	4	44	640	4	84	128	12	124	5376	4
5	5	2	45	2025	0	85	7225	0	125	1	50
6	36	0	46	2116	0	86	7396	0	126	0	56
7	49	0	47	9	30	87	7569	0	127	16129	0
8	64	0	48	2304	0	88	7744	0	128	0	56
9	1	8	49	401	8	89	829	10	129	1	56
10	100	0	50	196	8	90	8100	0	130	16900	0
11	9	6	51	2601	0	91	8281	0	131	1	86
12	144	0	52	2704	0	92	364	20	132	17424	0
13	169	0	53	1189	2	93	8649	0	133	17689	0
14	52	4	54	980	4	94	3060	4	134	6292	4
15	225	0	55	3025	0	95	9	62	135	1	64
16	0	8	56	3136	0	96	9216	0	136	18496	0
17	109	2	57	3249	0	97	9409	0	137	8189	2
18	324	0	58	3364	0	98	388	20	138	19044	0
19	1	16	59	53	22	99	801	16	139	1681	16
20	400	0	60	3600	0	100	10000	0	140	19600	0
21	441	0	61	1	40	101	197	18	141	19881	0
22	484	0	62	0	24	102	10404	0	142	20164	0
23	9	14	63	3969	0	103	10609	0	143	649	30
24	176	4	64	0	28	104	3760	4	144	7280	4
25	625	0	65	1	42	105	11025	0	145	21025	0
26	676	0	66	4356	0	106	11236	0	146	21316	0
27	729	0	67	1	32	107	2377	6	147	21609	0
28	784	0	68	4624	0	108	11664	0	148	21904	0
29	53	10	69	841	8	109	2201	8	149	2501	10
30	0	20	70	4900	0	110	12100	0	150	22500	0
31	961	0	71	361	14	111	12321	0	151	22801	0
32	0	20	72	5184	0	112	12544	0	152	2368	8
33	1	16	73	5329	0	113	5549	2	153	23409	0
34	372	4	74	1876	4	114	4532	4	154	240	24
35	217	6	75	5625	0	115	13225	0	155	5097	6
36	1296	0	76	5776	0	116	13456	0	156	24336	0
37	1369	0	77	2549	2	117	13689	0	157	24649	0
38	1444	0	78	6084	0	118	1380	8	158	24964	0
39	1	32	79	1	64	119	53	46	159	1	128
40	1600	0	80	6400	0	120	14400	0	160	25600	0

Table 1

Applying Conjecture 4.2 repeatedly and using Corollary 3.4, we easily arrive at the following:

Conjecture 4.4. *Let $n \geq 3$ be odd and let d be the maximal odd divisor of $n+1$. Then we have $A_1(n) = 1$ if and only if $d > 1$ and $A_1(d-1) = 0$.*

Proposition 4.5. *Conjectures 4.2 and 4.4 are equivalent.*

Proof. It suffices to show the implication Conjecture 4.4 $\Rightarrow$ Conjecture 4.2. Let $n \geq 2$ and let d be the maximal odd divisor of $n+1$ (and hence of $2n+2$). By Corollary 3.4, $A_1(2n+1) \leq 1$ is equivalent to $A_1(2n+1) = 1$, which, in turn, is equivalent to $d > 1$ and $A_1(d-1) = 0$ by Conjecture 4.4. If n is odd, then the same reasoning shows $A_1(n) \leq 1 \Longleftrightarrow d > 1$ and $A_1(d-1) = 0$, so we are done. If n is even, then $d = n+1 > 1$ and we have $A_1(n) \leq 1 \Longleftrightarrow A_1(d-1) = 0$ by Corollary 3.4. □

Next we make an attempt to "explain" the value of $A_1(n)$. If the Laplacian Δ_n is bijective, then we have $\mathscr{C}_n = \mathbb{F}_2^{n^2}$ and hence $A_1(n) = n^2$. We comment here on the bijectivity of Δ_n. Sutner [2000] proved

$$\dim \mathscr{H}_n = \deg \gcd(S_n(x), S_n(x+1)),$$

where S_n is the n-th Chebyshev polynomial of the second kind, regarded as a polynomial over $\mathbb{F}_2$. Some sufficient conditions for the bijectivity of Δ_n follow from this identity and well-known properties of Chebyshev polynomials. For example, $n = 2^l - 1$ ($l \geq 1$) is sufficient [Yamagishi 2015, Corollary 4.3]. Note that this confirms Conjecture 4.4 for $n = 2^l - 1$, as $A_1(n) = n^2$ and $d = 1$. There seems to be no simple characterization of n for which Δ_n is bijective.

Now we consider the case where Δ_n is not bijective, i.e., $\dim \mathscr{H}_n > 0$. As in Conjecture 4.4, the divisors d of $n+1$ with $A_1(d-1) = 0$ play an important role in the following two conjectures.

Conjecture 4.6. *Let n be even. Then Δ_n is not bijective if and only if there exists a (necessarily odd) divisor $d > 1$ of $n+1$ such that $A_1(d-1) = 0$.*

Conjecture 4.7. *Suppose n is even and Δ_n is not bijective. Assume Conjecture 4.6, and let d_k ($1 \leq k \leq t$) be the divisors of $n+1$ such that $d_k > 1$ and $A_1(d_k - 1) = 0$. Then for $1 \leq i, j \leq n$, we have $\boldsymbol{e}_{i,j} \in \mathscr{C}_n$ if and only if*

$$i \equiv 0 \pmod{d_k} \quad \text{or} \quad j \equiv 0 \pmod{d_k} \tag{4-1}$$

for $k = 1, 2, \ldots, t$.

Example 4.8. If $A_1(n) = 0$, then we can take $d_1 = n+1$ and Conjecture 4.7 is trivially true. But this gives no explanation of why $A_1(n) = 0$. We exclude this case in the following examples.

Example 4.9. Suppose $t = 1$ and put $b = (n+1)/d_1$. The number of pairs (i, j) for which (4-1) with $k = 1$ fails is $(n-b+1)^2$, so we have $A_1(n) = n^2 - (n-b+1)^2$.

This applies for $n = 14, 24, 34, 44, 54, 74, 94, 104, 114, 124, 134, 144$ ($d_1 = 5$), $n = 50, 118, 152$ ($d_1 = 17$), $n = 92$ ($d_1 = 31$), and $n = 98$ ($d_1 = 33$).

Example 4.10. For $n = 84$, we have $t = 2$, $d_1 = 5$, $d_2 = 17$, and (4-1) for $k = 1, 2$ reads as $ij \equiv 0 \pmod{85}$. Thus we have $A_1(84) = 2(5-1)(17-1) = 128$. The same reasoning applies for $n = 154$: $t = 2$, $d_1 = 5$, $d_2 = 31$ and $A_1(154) = 2(5-1)(31-1) = 240$.

Finally, we note that an answer to Problem 3.6 would give, under Conjecture 4.4, a characterization of (necessarily odd) n with $A_1(n) = 1$, and, under Conjecture 4.6, a characterization of even n with nonbijective Δ_n.

We also point out that, in Table 1, there are four exceptions $n = 2, 6, 8, 14$ for the converse statement of Conjecture 4.1. Problem 3.6 would be settled if they are the only exceptions.

Acknowledgments

Yamagishi was supported by JSPS KAKENHI Grant Number JP17K05168. The authors are grateful to Professor Norihiro Nakashima for informing them of Lemma 3.5.

References

[Fleischer and Yu 2013] R. Fleischer and J. Yu, "A survey of the game 'Lights Out!"', pp. 176–198 in *Space-efficient data structures, streams, and algorithms* (Waterloo, ON, 2013), edited by A. Brodnik et al., Lecture Notes in Comput. Sci. **8066**, Springer, 2013. MR Zbl

[Goldwasser and Klostermeyer 1997] J. Goldwasser and W. Klostermeyer, "Maximization versions of 'lights out' games in grids and graphs", *Congr. Numer.* **126** (1997), 99–111. MR Zbl

[Goshima and Yamagishi 2010] M. Goshima and M. Yamagishi, "On the dimension of the space of harmonic functions on a discrete torus", *Experiment. Math.* **19**:4 (2010), 421–429. MR Zbl

[MacWilliams and Sloane 1977] F. J. MacWilliams and N. J. A. Sloane, *The theory of error-correcting codes*, North-Holland Math. Library **16**, North-Holland, Amsterdam, 1977. MR Zbl

[Sutner 1989] K. Sutner, "Linear cellular automata and the Garden-of-Eden", *Math. Intelligencer* **11**:2 (1989), 49–53. MR Zbl

[Sutner 2000] K. Sutner, "σ-automata and Chebyshev-polynomials", *Theoret. Comput. Sci.* **230**:1-2 (2000), 49–73. MR Zbl

[Yamagishi 2015] M. Yamagishi, "Periodic harmonic functions on lattices and Chebyshev polynomials", *Linear Algebra Appl.* **476** (2015), 1–15. MR Zbl

Received: 2018-09-22 Accepted: 2018-10-25

29414088@stn.nitech.ac.jp *Field of Mathematics and Mathematical Science, Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology, Nagoya, Japan*

yamagishi.masakazu@nitech.ac.jp *Department of Mathematics, Nagoya Institute of Technology, Nagoya, Japan*

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

2019 vol. 12 no. 4

1944-4176(2019)12:4;1-1