# involve

## a journal of mathematics





msp

2020 vol. 13, no. 2

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

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

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

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

INVOLVE 13:2 (2020)
dx.doi.org/10.2140/involve.2020.13.181

# Arithmetic functions of higher-order primes

Kyle Czarnecki and Andrew Giddings

(Communicated by Filip Saidak)

The sieve of Eratosthenes (SoE) is a well-known method of extracting the set of prime numbers $\mathbb{P}$ from the set positive integers $\mathbb{N}$. Applying the SoE again to the index of the prime numbers will result in the set of prime-indexed primes $\mathbb{P}_2 = \{3, 5, 11, 17, 31, \ldots\}$. More generally, the application of the SoE $k$-times will yield the set $\mathbb{P}_k$ of $k$-th order primes. In this paper, we give an upper bound for the $n$-th $k$-order prime as well as some results relating to number-theoretic functions over $\mathbb{P}_k$.

## 1. Introduction

This paper lies within the intersection of two topics in analytic number theory: abstract analytic number theory and the prime-indexed primes. In this section we give brief summaries of these topics followed by our results, whose proofs are presented in the subsequent sections.

**1.1.** ***Abstract analytic number theory.*** We begin with a set of *generalized primes* $\mathcal{P} := \{p_i \in \mathbb{R} \mid 1 < p_1 \le p_2 \le \cdots < \infty\}$ out of which the set of *generalized integers* $\mathcal{N} := \{n = p_1^{\alpha_1} \cdots p_k^{\alpha_k} \mid p_i \in \mathcal{P},\ \alpha_i \in \mathbb{N}\}$ is constructed; that is, $\mathcal{N}$ consists of all possible finite words over $\mathcal{P}$. One of the main goals of abstract analytic number theory is to describe the asymptotic behavior of, and the relationship between, the two counting functions

$$\pi_{\mathcal{P}}(x) := \sum_{\substack{p<x \\ p\in\mathcal{P}}} 1 \quad \text{and} \quad N_{\mathcal{P}}(x) := \sum_{\substack{n<x \\ n\in\mathcal{N}}} 1. \tag{1}$$

For example, what growth conditions must we impose on $N_{\mathcal{P}}(x)$ so that $\pi_{\mathcal{P}}(x)$ satisfies the classical prime number theorem (PNT) $\pi_{\mathcal{P}}(x) \sim x/\log x$? One might naturally suppose that $N_{\mathcal{P}}(x) \sim x$ may be enough to guarantee the PNT, but Beurling [1937] showed this is not the case. In particular, he proved that if $N_{\mathcal{P}}(x) = Ax + O(x/\log^{\gamma} x)$, $A > 0$, and $\gamma > \frac{3}{2}$, then the PNT follows; however, he also showed

*MSC2010:* 11A41, 11N37, 11N80.
*Keywords:* prime-indexed primes, abstract analytic number theory, Beurling zeta function.

the PNT can fail if $\gamma = \frac{3}{2}$. For more on the PNT and Beurling's generalization of it, we refer the reader to Sections 1.1, 6.2, and 8.4 of [Montgomery and Vaughan 2007].

Central to the proof of Beurling's generalized PNT, and arguably the most fundamental object in abstract analytic number theory, is the *Beurling zeta function* $\zeta_{\mathcal{P}}(s)$, which is defined as

$$\zeta_{\mathcal{P}}(s) := \prod_{p\in\mathcal{P}} \left(1 - \frac{1}{p^s}\right)^{-1} = \sum_{n\in\mathcal{N}} \frac{1}{n^s} \tag{2}$$

for $s \in \mathbb{C}$ wherever it converges.[1] Classically, many important arithmetic functions arise as the coefficients of Dirichlet series involving the Riemann zeta function $\zeta(s) := \zeta_{\mathbb{P}}(s)$; see Chapter 1 of [Titchmarsh 1986]. The von Mangoldt function $\Lambda(n)$, for instance, satisfies

$$\frac{\zeta'(s)}{\zeta(s)} = -\sum_{n=1}^{\infty} \frac{\Lambda(n)}{n^s}, \quad \text{where } \Lambda(n) = \begin{cases} \log p & \text{if } n = p^k \text{ for } k \in \mathbb{N}, \\ 0 & \text{otherwise.} \end{cases}$$

Replacing the Riemann zeta function with the Beurling zeta function, we arrive at the generalized von Mangoldt function $\Lambda_{\mathcal{P}}(n)$ defined over $\mathcal{N}$; that is,

$$\frac{\zeta'_{\mathcal{P}}(s)}{\zeta_{\mathcal{P}}(s)} = -\sum_{n\in\mathcal{N}} \frac{\Lambda_{\mathcal{P}}(n)}{n^s}, \quad \text{where } \Lambda_{\mathcal{P}}(n) = \begin{cases} \log p & \text{if } n = p^k \text{ for } p \in \mathcal{P},\ k \in \mathbb{N}, \\ 0 & \text{otherwise.} \end{cases}$$

In this paper we will restrict our attention to the generalized prime counting function $\pi_{\mathcal{P}}(n)$, the $n$-th prime function $\pi_{\mathcal{P}}^{-1}(n)$, the von Mangoldt function $\Lambda_{\mathcal{P}}(n)$, and the Chebyshev theta function

$$\theta_{\mathcal{P}}(n) := \sum_{\substack{p<x \\ p\in\mathcal{P}}} \log p.$$

For further information on the Beurling zeta function and the various arithmetic functions associated to it, see the excellent summary given in Chapter 1 of [Diamond and Zhang 2016], as well as the more detailed Chapter 2 of [Knopfmacher 1990].

**1.2.** ***The k-th order primes.*** The sieve of Eratosthenes (SoE) is a well-known method of extracting the set of prime numbers $\mathbb{P}$ from the set of positive integers $\mathbb{N}$. Applying the SoE again to the index of the prime numbers will result in the set of prime-indexed primes $\mathbb{P}_2 = \{3, 5, 11, 17, 31, 41, 59, 67, 83, \ldots\}$. More generally, the application of the SoE $k$-times will yield the set $\mathbb{P}_k$ of $k$-th order primes. Regarding $\mathbb{P}_k$ as a set of generalized primes, one also has the corresponding set $\mathbb{N}_k$

[1] It is known (see Proposition 2.1 in Chapter 4 of [Knopfmacher 1990]) that the Euler product and the Dirichlet series converge in the same half-plane and are equal.

of generalized integers (as constructed above in Section 1.1). It follows by definition that the corresponding prime counting functions are given by

$$\pi_k(x) := \pi_{\mathbb{P}_k}(x) = \underbrace{(\pi \circ \cdots \circ \pi)}_{k-\text{times}}(x),$$

and the PNT implies $\pi_k(x) \sim x/\log^k x$ (see Lemma 8 below). Furthermore, if we set $\pi^{-1}(n) := p_n$, where $p_n$ is the $n$-th prime, then the $n$-th $k$-order prime is given by

$$\pi_{-k}(n) := \pi_{\mathbb{P}_k}^{-1}(n) = \underbrace{(\pi^{-1} \circ \cdots \circ \pi^{-1})}_{k-\text{times}}(n).$$

The $k$-th order primes $\mathbb{P}_k$ and corresponding functions $\pi_{\pm k}(n)$ have been studied recently in [Broughan and Barnett 2009; Bayless et al. 2013], with emphasis on the $k = 2$, prime-indexed prime case. In addition to generalizing fundamental questions like the Goldbach and twin prime conjectures to $\mathbb{P}_2$, bounds have been derived for $\pi_{\pm 2}(n)$, as well as $\pi_k(n)$ for $k \geq 2$. For instance, Broughan and Barnett give the bounds

$$\pi_{-2}(n) < n(\log n + \log\log n)(\log n + 2\log\log n) - n\log n + O(n\log\log n),$$

$$\pi_{-2}(n) > n(\log n + \log\log n)(\log n + 2\log\log n) - 3n\log n + O(n\log\log n).$$

The goal of this paper is to continue this line of inquiry against the backdrop of abstract analytic number theory.

**1.3.** ***Results.*** Our main result is a bound for $\pi_{-k}(n)$ for all $k \geq 1$. Consequently, the following theorem, together with the PNT, bound $\pi_k(n)$ for all $k \in \mathbb{Z}$.

**Theorem 1.** *For any integer $k \geq 1$ the function $\pi_{-k}(n)$ satisfies*

$$\pi_{-k}(n) = n \prod_{j=1}^{k} f(n, j) + O_k(n\log^{k-2} n \log\log n), \tag{3}$$

*where $f(n, j) := \log n + j\log\log n - 1$.*

The proof of Theorem 1 is given below in Section 2. Ideally, one would like to determine the precise relation between $k$ and the error constants in the theorem.

The next two theorems are generalizations of Theorems 9.1 and 9.2 in [De Koninck and Ivić 1980] involving sums of arithmetic functions. The first involves the $k$-th order prime counting function $\pi_k(n)$.

**Theorem 2.** *For any integer $k \geq 1$ the function $\pi_k(x)$ satisfies*

$$\sum_{n \leq x}{}' \frac{1}{\pi_k(n)} = \frac{\log^{k+1} x}{k+1} + O_k(\log^k x \log\log x), \tag{4}$$

*where the primed summation means any terms where the denominator is zero are ignored.*

The proof of Theorem 2 is given below in Section 3 and essentially relies on the PNT. From this theorem and the fact that $\pi_k(n) \ll n/\log^k n$ we have the following aesthetically pleasing corollary.

**Corollary 3.** *For any integer $k \geq 1$,*

$$\sum_{n\leq x}{}' \frac{1}{\pi_k(n)} \ll_k \frac{x}{\pi_{k+1}(x)}.$$

Note that this corollary can be interpreted as saying the average value of $1/\pi_k(n)$ on $[1, x]$ is bounded by a constant multiple of $1/\pi_{k+1}(x)$. Lastly, we give a similar theorem involving the von Mangoldt function $\Lambda_k(n) := \Lambda_{\mathbb{P}_k}(n)$.

**Theorem 4.** *For any integer $k \geq 1$ and a fixed integer $N > k$ the function $\Lambda_k(n)$ satisfies*

$$\sum_{\substack{n\leq x\\ n\in\mathbb{N}_k}}{}' \frac{1}{\Lambda_k(n)} = \sum_{j=k+1}^{N} \frac{c_j x}{\log^j x} + O_k\left(\frac{x\log\log x}{\log^{k+2} x}\right), \tag{5}$$

*where $c_{k+1} = 1$ and the remaining $c_j$ are computable constants.*

The proof relies on the following bound for the generalized Chebyshev theta function $\theta_k(x) := \theta_{\mathbb{P}_k}(x)$.

**Lemma 5.** *For any integer $k \geq 1$ and a fixed positive integer $N$ the function $\theta_k(x)$ satisfies*

$$\theta_k(x) = \frac{x}{\log^{k-1} x} - \sum_{j=k}^{N} \frac{(j-1)!}{(k-1)!}\frac{x}{\log^j x} + O_k\left(\frac{x\log\log x}{\log^k x}\right), \tag{6}$$

*where the summation is understood to be zero if $N < k$.*

Both proofs of Theorem 4 and Lemma 5 are given below in Section 4.

## 2. Proof of Theorem 1

The basis of the theorem is the following bound for the $n$-th prime.

**Lemma 6.** *The function $\pi_{-1}(n)$ satisfies*

$$\frac{\pi_{-1}(n)}{n} = \log n + \log\log n - 1 + O\left(\frac{\log\log n}{\log n}\right).$$

*Proof.* This follows from a bound of Pervouchine which can be found in [Cesàro 1894]. □

Next, we give a lemma that will be used to bound compositions of $\pi_{-1}(n)$.

**Lemma 7.** *Let $f(n, j)$ be defined as*

$$f(n, j) := \log n + j \log\log n - 1.$$

*Then for any fixed integers $j, l \geq 1$ the following three equations hold*:

$$\log(nf(n, j)) = \log n + \log\log n + O_j\left(\frac{\log\log n}{\log n}\right), \tag{7}$$

$$\log\log(nf(n, j)) = \log\log n + O_j\left(\frac{\log\log n}{\log n}\right), \tag{8}$$

$$f(nf(n, j), l) = f(n, l+1) + O_{j,l}\left(\frac{\log\log n}{\log n}\right). \tag{9}$$

*Proof.* The proof is a straightforward computation, but we include it here for completeness. First, using the definition of $f(n, j)$ and properties of logarithms, we have

$$\begin{aligned}\log(nf(n, j)) &= \log n + \log(\log n + j\log\log n - 1)\\ &= \log n + \log\left(\log n\left(1 + j\frac{\log\log n}{\log n} - \frac{1}{\log n}\right)\right)\\ &= \log n + \log\log n + \log\left(1 + j\frac{\log\log n}{\log n} - \frac{1}{\log n}\right).\end{aligned}$$

Now (7) follows since

$$\log\left(1 + j\frac{\log\log n}{\log n} - \frac{1}{\log n}\right) \ll_j \frac{\log\log n}{\log n}$$

by expanding the logarithm. Similarly, from (7) we obtain

$$\begin{aligned}\log\log(nf(n, j)) &= \log\left(\log n + \log\log n + O_j\left(\frac{\log\log n}{\log n}\right)\right)\\ &= \log\left(\log n\left(1 + \frac{\log\log n}{\log n} + O_j\left(\frac{\log\log n}{\log^2 n}\right)\right)\right)\\ &= \log\log n + \log\left(1 + \frac{\log\log n}{\log n} + O_j\left(\frac{\log\log n}{\log^2 n}\right)\right),\end{aligned}$$

and (8) follows from expanding the last term. Finally, (9) follows immediately from the definition of $f$ and (7) and (8):

$$\begin{aligned}f(nf(n, j), l) &= \log(nf(n, j)) + l\log\log(nf(n, j)) - 1\\ &= \log n + \log\log n + O_j\left(\frac{\log\log n}{\log n}\right) + l\left(\log\log n + O_j\left(\frac{\log\log n}{\log n}\right)\right) - 1\\ &= \log n + (l+1)\log\log n - 1 + O_{j,l}\left(\frac{\log\log n}{\log n}\right).\end{aligned}$$

□

**Remark.** The three equations in Lemma 7 remain valid if we replace $f(n, j)$ by $f(n, j) + O(\log\log n/\log n)$ as the computations remain unchanged.

*Proof of Theorem 1.* We proceed via induction on $k$. Since the $k = 1$ case reduces to Lemma 6, suppose (3) holds for some $k = m > 1$; that is,

$$\pi_{-m}(n) = n\prod_{j=1}^{m} f(n, j) + O_m(n\log^{m-2} n\log\log n).$$

Recall that $\pi_{-(m+1)} = \pi_{-m}\circ\pi_{-1}$, so we have

$$\pi_{-(m+1)}(n) = \pi_{-1}(n)\prod_{j=1}^{m} f(\pi_{-1}(n), j) + O_m(\pi_{-1}(n)\log^{m-2}(\pi_{-1}(n))\log\log(\pi_{-1}(n))).$$

Note that by Lemmas 6 and 7

$$f(\pi_{-1}(n), l) = f\left(nf(n, 1)+O\left(\frac{n\log\log n}{\log n}\right), l\right) = f(n, l+1)+O_l\left(\frac{\log\log n}{\log n}\right)$$

for any $l \geq 1$. Applying this and Lemma 6 to the expression for $\pi_{-(m+1)}(n)$ above yields

$$\pi_{-(m+1)}(n) = n\prod_{j=1}^{m+1}\left(f(n, j) + O_j\left(\frac{\log\log n}{\log n}\right)\right) + O_m\big(nf(n, 1)\log^{m-2}(nf(n, 1))\log\log(nf(n, 1))\big).$$

The error terms arising from the product are at most

$$n\log^m n\, O_m\left(\frac{\log\log n}{\log n}\right) = O_m(n\log^{m-1} n\log\log n),$$

and by Lemma 7 we have

$$nf(n, 1)\log^{m-2}(nf(n, 1))\log\log(nf(n, 1)) \ll n\log^{m-1} n\log\log n.$$

Thus,

$$\pi_{-(m+1)}(n) = n\prod_{j=1}^{m+1} f(n, j) + O_m(n\log^{m-1} n\log\log n),$$

which completes the proof. □

## 3. Proof of Theorem 2

The proof depends on the following two lemmas.

**Lemma 8.** *For any integer $k \geq 1$ the function $\pi_k(n)$ satisfies*

$$\pi_k(n) = \frac{n}{\log^k n} + O_k\left(\frac{n \log\log n}{\log^{k+1} n}\right). \tag{10}$$

*Proof.* This follows from the PNT; see Theorem 7.1 in [Bayless et al. 2013]. □

**Lemma 9.** *For any integer $k \geq 1$ we have*

$$\sum_{n \leq x} \frac{\log^k n}{n} = \frac{\log^{k+1} x}{k+1} + O_k(1). \tag{11}$$

*Proof.* Recall Abel's summation formula[2]

$$\sum_{y < n \leq x} a_n \phi(n) = A(x)\phi(x) - A(y)\phi(y) - \int_y^x A(t)\phi'(t)\,dt, \tag{12}$$

where $\phi \in \mathcal{C}^1$ and $A(x) = \sum_{n \leq x} a_n$ for some sequence $(a_n)$. Taking $a_n \equiv 1$ we have $A(x) = \lfloor x \rfloor = x - \{x\}$, where $\{x\}$ denotes the fractional part of $x$, and so

$$\sum_{y < n \leq x} \phi(n) = (x - \{x\})\phi(x) - (y - \{y\})\phi(y) - \int_y^x (t - \{t\})\phi'(t)\,dt.$$

Observing that $\int_y^x t\phi'(t)\,dt = x\phi(x) - y\phi(y) - \int_y^x \phi(t)\,dt$, we arrive at the following form of Abel's summation formula:

$$\sum_{y < n \leq x} \phi(n) = \int_y^x \phi(t)\,dt - \{x\}\phi(x) + \{y\}\phi(y) + \int_y^x \{t\}\phi'(t)\,dt.$$

Setting $y = 1$ and $\phi(x) = \log^k x / x$ yields

$$\sum_{n \leq x} \frac{\log^k n}{n} = \frac{\log^{k+1} x}{k+1} - \{x\}\frac{\log^k x}{x} + \int_1^x \{t\}\phi'(t)\,dt,$$

or, in other words,

$$\sum_{n \leq x} \frac{\log^k n}{n} - \frac{\log^{k+1} x}{k+1} \ll \phi(x) + \int_1^x |\phi'(t)|\,dt. \tag{13}$$

Now $\phi'(x) > 0$ on the interval $(1, e^k)$, and $\phi'(x) < 0$ on $(e^k, \infty)$. Thus, by the fundamental theorem of calculus, the right-hand side of (13) becomes

$$\phi(x) + \phi(e^k) - \phi(1) - \phi(x) + \phi(e^k) = 2\phi(e^k) = 2\left(\frac{k}{e}\right)^k,$$

which proves the lemma. □

[2] Abel's summation formula is integration by parts of a Riemann–Stieltjes integral; see Theorem 4.2 in [Apostol 1976] and Appendix A in [Montgomery and Vaughan 2007]. Several of the computations here, especially those in Section 4 below, become more straightforward with this machinery.

**Remark.** Lemma 9 also follows from the fact that

$$\lim_{x\to\infty}\left(\sum_{n\le x}\frac{\log^k n}{n}-\frac{\log^{k+1}x}{k+1}\right)=\gamma_k,$$

where $\gamma_k$ are constants, called Stieltjes constants, that occur in the Laurent series expansion of the Riemann zeta function about $s=1$; in particular,

$$\zeta(s)=\frac{1}{s-1}+\sum_{k=0}^{\infty}\frac{(-1)^k}{k!}\gamma_k(s-1)^k;$$

see Example 1.3.12 on p. 29 of [Montgomery and Vaughan 2007]. Consequently, the proof of Lemma 9 also gives a simple upper bound for $|\gamma_k|$.

*Proof of Theorem 2.* The $k=1$ case is Theorem 9.1 in [De Koninck and Ivić 1980], so assume $k\ge 2$. From Lemmas 8 and 9 we have

$$\sum_{n\le x}{}'\left(\frac{1}{\pi_k(n)}-\frac{1}{n/\log^k n}\right)=O_k\left(\sum_{n\le x}{}'\frac{\log\log n\log^{k-1}n}{n}\right).$$

The theorem now follows from the fact that

$$\begin{aligned}\sum_{n\le x}{}'\frac{\log\log n\log^{k-1}n}{n}&\ll\int_1^x\frac{\log\log t\log^{k-1}t}{t}\,dt\\&=\log\log t\,\frac{\log^k t}{k}\bigg|_1^x-\frac{1}{k}\int_1^x\frac{\log^{k-1}t}{t}\,dt\\&=\frac{1}{k}\log\log x\log^k x-\frac{\log^k x}{k^2}.\end{aligned}$$

□

## 4. Proofs of Theorem 4 and Lemma 5

We first prove the bound for the Chebyshev theta function over $\mathbb{P}_k$.

*Proof of Lemma 5.* Using Riemann–Stieltjes integration, or Abel's summation formula in (12) with $\phi(x)=\log x$ and $A(x)=\pi_k(x)$, we have

$$\theta_k(x)=\int_2^x\log x\,d\pi_k(x)=\pi_k(x)\log x-\int_2^x\frac{\pi_k(t)}{t}\,dt$$

(note this reduces to Theorem 4.3 in [Apostol 1976] when $k=1$). By Lemma 8 this becomes

$$\theta_k(x)=\frac{x}{\log^{k-1}x}+O_k\left(\frac{x\log\log x}{\log^k x}\right)-\int_2^x\frac{dt}{\log^k t}-O_k\left(\int_2^x\frac{\log\log t}{\log^{k+1}t}\,dt\right).\quad(14)$$

Next, integration by parts yields the formula

$$\int \frac{dt}{\log^k t} = \frac{t}{\log^k t} + k \int \frac{dt}{\log^{k+1} t},$$

which through repeated use gives

$$\int_2^x \frac{dt}{\log^k t} = \sum_{j=k}^{N} \frac{(j-1)!}{(k-1)!} \frac{x}{\log^j x} + O_k\left(\frac{x}{\log^{N+1} x}\right). \tag{15}$$

Similarly, using integration by parts on the integral in the error term, we have

$$\int \frac{\log\log t}{\log^{k+1} t}\, dt = \frac{t \log\log t}{\log^{k+1} t} - \int \frac{dt}{\log^{k+2} t} + (k+1) \int \frac{\log\log t}{\log^{k+2} t}\, dt,$$

which implies

$$\int_2^x \frac{\log\log t}{\log^{k+1} t}\, dt \ll \frac{x \log\log x}{\log^{k+1} x} \ll \frac{x \log\log x}{\log^k x}. \tag{16}$$

The lemma now follows from using (16) and (15) in (14). □

*Proof of Theorem 4.* The $k = 1$ case is Theorem 9.2 in [De Koninck and Ivić 1980], so assume $k \geq 2$. The idea of the proof is to use the definition of $\Lambda_k$ to write

$$\sum_{\substack{n \leq x \\ n \in \mathbb{N}_k}}{}' \frac{1}{\Lambda_k(n)} = \sum_{\substack{p^\alpha \leq x \\ \alpha \geq 1 \\ p \in \mathbb{P}_k}} \frac{1}{\log p} = \sum_{\substack{p \leq x \\ p \in \mathbb{P}_k}} \frac{1}{\log p} + \sum_{\substack{p^2 \leq x \\ p \in \mathbb{P}_k}} \frac{1}{\log p} + \cdots$$

and then bound the summations on the right-hand side.[3] To this end, we again use Riemann–Stieltjes integration, or Abel's summation formula in (12) with $\phi(x) = 1/\log^2 x$ and $A(x) = \theta_k(x)$, to obtain

$$\sum_{\substack{p \leq x \\ p \in \mathbb{P}_k}} \frac{1}{\log p} = \int_2^x \frac{1}{\log^2 t}\, d\theta_k(t) = \frac{\theta_k(x)}{\log^2 x} + 2 \int_2^x \frac{\theta_k(t)}{t \log^3 t}\, dt,$$

which by Lemma 5 becomes

$$\begin{aligned} \sum_{\substack{p \leq x \\ p \in \mathbb{P}_k}} \frac{1}{\log p} = {} & \frac{x}{\log^{k+1} x} - \sum_{j=k}^{N-2} \frac{(j-1)!}{(k-1)!} \frac{x}{\log^{j+2} x} + O_k\left(\frac{x \log\log x}{\log^{k+2} x}\right) \\ & + 2 \int_2^x \frac{dt}{\log^{k+2} t} - 2 \sum_{j=k}^{N-3} \frac{(j-1)!}{(k-1)!} \int_2^x \frac{dt}{\log^{j+3} t} + O_k\left(\int_2^x \frac{\log\log t}{\log^{k+3} t}\, dt\right). \end{aligned} \tag{17}$$

[3] Note that there are only finitely many summations for a given $x$ because $p^\alpha \leq x$ implies $\alpha \leq \log x / \log p \leq \log x / \log 2$.

Note that the upper bounds on the summations have been chosen so that they terminate on the order of $x/\log^N x$. We can combine error terms by taking $k \to k+2$ in (16); that is,

$$\int_2^x \frac{\log\log t}{\log^{k+3} t}\,dt \ll \frac{x\log\log x}{\log^{k+3} x} \ll \frac{x\log\log x}{\log^{k+2} x}. \tag{18}$$

Now by (15) we have

$$\begin{aligned}\int_2^x \frac{dt}{\log^{k+2} t} &= \sum_{j=k+2}^{N} \frac{(j-1)!}{(k+1)!}\frac{x}{\log^j x} + O_k\left(\frac{x}{\log^{N+1} x}\right)\\ &= \sum_{j=k}^{N-2} \frac{(j+1)!}{(k+1)!}\frac{x}{\log^{j+2} x} + O_k\left(\frac{x}{\log^{N+1} x}\right),\end{aligned} \tag{19}$$

as well as

$$\int_2^x \frac{dt}{\log^{j+3} t} = \frac{x}{\log^{j+3} x} + O_j\left(\frac{x}{\log^{j+4} x}\right). \tag{20}$$

Putting (18)–(20) into (17) yields

$$\begin{aligned}\sum_{\substack{p\le x\\ p\in\mathbb{P}_k}} \frac{1}{\log p} = \frac{x}{\log^{k+1} x} &+ \sum_{j=k}^{N-2}\left(2\frac{(j+1)!}{(k+1)!} - \frac{(j-1)!}{(k-1)!}\right)\frac{x}{\log^{j+2} x}\\ &- 2\sum_{j=k}^{N-3}\frac{(j-1)!}{(k-1)!}\frac{x}{\log^{j+3} x} + O_k\left(\frac{x\log\log x}{\log^{k+2} x}\right),\end{aligned}$$

which can be rewritten as

$$\sum_{\substack{p\le x\\ p\in\mathbb{P}_k}} \frac{1}{\log p} = \sum_{j=k+1}^{N} \frac{c_j x}{\log^j x} + O_k\left(\frac{x\log\log x}{\log^{k+2} x}\right).$$

The theorem now follows from the fact that

$$\sum_{\substack{n\le x\\ n\in\mathbb{N}_k}}{}' \frac{1}{\Lambda_k(n)} = \sum_{\substack{p^\alpha\le x\\ \alpha\ge 1\\ p\in\mathbb{P}_k}} \frac{1}{\log p} = \sum_{\substack{p\le x\\ p\in\mathbb{P}_k}} \frac{1}{\log p} + O\left(\sum_{\substack{p^2\le x\\ p\in\mathbb{P}_k}} \frac{1}{\log p}\right),$$

and by observing that

$$\sum_{\substack{p\le\sqrt{x}\\ p\in\mathbb{P}_k}} \frac{1}{\log p} \ll \frac{\sqrt{x}}{\log^k\sqrt{x}} \ll \frac{x\log\log x}{\log^{k+2} x}$$

by Lemma 8. □

## References

[Apostol 1976] T. M. Apostol, *Introduction to analytic number theory*, Springer, 1976. MR Zbl

[Bayless et al. 2013] J. Bayless, D. Klyve, and T. Oliveira e Silva, "New bounds and computations on prime-indexed primes", *Integers* **13** (2013), art. id. A43. MR Zbl

[Beurling 1937] A. Beurling, "Analyse de la loi asymptotique de la distribution des nombres premiers généralisés, I", *Acta Math.* **68**:1 (1937), 255–291. MR Zbl

[Broughan and Barnett 2009] K. A. Broughan and A. R. Barnett, "On the subsequence of primes having prime subscripts", *J. Integer Seq.* **12**:2 (2009), art. id. 09.2.3. MR Zbl

[Cesàro 1894] E. Cesàro, "Sur une formule empirique de M. Pervouchine", *C. R. Acad. Sci. Paris* **119** (1894), 848–849. JFM

[De Koninck and Ivić 1980] J.-M. De Koninck and A. Ivić, *Topics in arithmetical functions*: *asymptotic formulae for sums of reciprocals of arithmetical functions and related results*, Notas de Matemática **72**, North-Holland, Amsterdam, 1980. MR Zbl

[Diamond and Zhang 2016] H. G. Diamond and W.-B. Zhang, *Beurling generalized numbers*, Mathematical Surveys and Monographs **213**, American Mathematical Society, Providence, RI, 2016. MR Zbl

[Knopfmacher 1990] J. Knopfmacher, *Abstract analytic number theory*, 2nd ed., Dover Books on Advanced Mathematics, Dover, New York, 1990. 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

[Titchmarsh 1986] E. C. Titchmarsh, *The theory of the Riemann zeta-function*, 2nd ed., Clarendon, Oxford, 1986. MR Zbl

Received: 2018-08-29 Revised: 2019-07-25 Accepted: 2019-12-28

czarneckik@uwplatt.edu *University of Wisconsin, Platteville, WI, United States*

giddingsa@uwplatt.edu *University of Wisconsin, Platteville, WI, United States*

INVOLVE 13:2 (2020)
dx.doi.org/10.2140/involve.2020.13.193

# Spherical half-designs of high order

Daniel Hughes and Shayne Waldron

(Communicated by David Royal Larson)

We give some explicit examples of putatively optimal spherical half-designs, i.e., ones for which there is numerical evidence that they are of minimal size. These include a 16-point weighted spherical half-design of order 8 for $\mathbb{R}^3$ based on the pentakis dodecahedron. This gives rise to a 32-point weighted spherical 9-design for the sphere.

## 1. Introduction

Let $\mathbb{S}$ be the unit sphere in $\mathbb{R}^d$ and $\sigma$ be the surface area measure on $\mathbb{S}$, normalised to have $\sigma(\mathbb{S}) = 1$, i.e., to be a probability measure. A "spherical design" is a sequence of points $v_1, \dots, v_n$ in $\mathbb{S}$ for which the integration (cubature) rule

$$\int_{\mathbb{S}} p(x)\, d\sigma(x) = \frac{1}{n} \sum_{j=1}^{n} p(v_j) \tag{1-1}$$

holds for all $p$ in some finite-dimensional space of polynomials $P$. The existence of such a spherical design for $n$ sufficiently large was proved in [Seymour and Zaslavsky 1984]. When $P$ is the space of polynomials of degree $\le t$, one has a *spherical $t$-design*. Suppose that $P = \Pi_k^\circ(\mathbb{R}^d)$, the space of homogeneous polynomials of degree $k$. If $k$ is even, then (1-1) integrates all the homogeneous polynomials $q$ of even degrees $2m \le k = 2t$, since taking

$$p(x) = (x_1^2 + \cdots + x_d^2)^{t-m} q(x)$$

in (1-1) gives

$$\int_{\mathbb{S}} q(x)\, d\sigma(x) = \int_{\mathbb{S}} p(x)\, d\sigma(x) = \frac{1}{n} \sum_{j=1}^{n} p(v_j) = \frac{1}{n} \sum_{j=1}^{n} q(v_j).$$

For this reason, the spherical designs which integrate the homogeneous polynomials of degree $2t$ (and hence of degrees $0, 2, \dots, 2t$) are called *spherical half-designs*

*MSC2010:* primary 05B30, 42C15, 65D30; secondary 94A12.
*Keywords:* spherical $t$-designs, spherical half-designs, tight spherical designs, finite tight frames, integration rules, cubature rules, cubature rules for the sphere, pentakis dodecahedron.

of order $2t$. The terms spherical $2t$-design [Seidel 2001] and spherical $(t, t)$-design [Waldron 2017] are also used. The related spherical designs of harmonic index $2t$ [Bannai et al. 2015] integrate the subspace of harmonic homogeneous polynomials of degree $2t$.

We also observe that if $q$ is homogeneous of odd degree, then $q(-x) = -q(x)$ and its integral over $\mathbb{S}$ is zero. Thus if a spherical design is *centrally symmetric*, i.e., of the form $\{\pm v_j\}$, then it integrates all homogeneous polynomials of odd degree. Thus:

**Proposition 1.1.** *The following are equivalent*:

(i) $(\pm v_j)$ *is a centrally symmetric spherical* $(2t+1)$*-design of* $2n$ *vectors for* $\mathbb{R}^d$.

(ii) $(v_j)$ *is a spherical half-design of order* $2t$ *of* $n$ *vectors for* $\mathbb{R}^d$.

The analogue of Proposition 1.1 for complex spherical designs is discussed in [Roy and Suda 2014, Lemma 3.4; Mohammadpour and Waldron 2019].

There are various equivalent conditions to being a spherical design [Delsarte et al. 1977; Bannai and Bannai 2009]. These include being an integration rule for a subspace of harmonic polynomials, and a variational characterisation. The spherical half-designs $(v_j)$ of order $2t$ are characterised in [Waldron 2017] as the vectors in $\mathbb{R}^d$ which give equality in the inequality

$$\sum_{j=1}^{n}\sum_{k=1}^{n} |\langle v_j, v_k\rangle|^{2t} \geq \frac{1\cdot 3\cdot 5\cdots(2t-1)}{d(d+2)\cdots(d+2(t-1))}\left(\sum_{\ell=1}^{n}\|v_\ell\|^{2t}\right)^2. \tag{1-2}$$

This implies that a spherical half-design $(v_j)$ is *projectively unitarily invariant*; i.e., $(c_j U v_j)$ is also a half-design when $c_j \in \{-1, 1\}$ and $U$ is unitary. A spherical $t$-design has this property if and only if it is centrally symmetric. In view of this (and their definitions), a spherical $t$-design can be thought of as a set of points that are evenly spaced on the sphere, and a spherical half-design as a set of lines (antipodal points) which are evenly spaced on the sphere. Using results from Brouwer degree theory, [Bondarenko et al. 2013] showed that the minimum number of points in a spherical $t$-design (and hence in a spherical half-design of order $2t$) grows like $t^{d-1}$ (with $d$ fixed).

When the vectors $v_1, \ldots, v_n$ in $\mathbb{R}^d$ giving equality in (1-2) are not all of unit norm and not all zero, then one has the weighted integration rule

$$\int_{\mathbb{S}} p\, d\sigma = \frac{1}{\sum_k \|v_k\|^{2t}}\sum_{j=1}^{n} p(v_j) = \sum_{\substack{j=1\\ v_j\neq 0}}^{n} w_j\, p\left(\frac{v_j}{\|v_j\|}\right) \quad \text{for all } p\in\Pi^{\circ}_{2t}(\mathbb{R}^d),$$

where

$$w_j := \frac{\|v_j\|^{2t}}{\sum_k \|v_k\|^{2t}},$$

and we call $(v_j)$ a *weighted* spherical half-design of order $2t$ with weights $(w_j)$ [Kotelina and Pevnyi 2011].

Since spherical half-designs $(v_j)$ satisfy (1-2) and are determined by the equation

$$\sum_{j=1}^{n}\sum_{k=1}^{n}|\langle v_j, v_k\rangle|^{2t} = \frac{1\cdot 3\cdot 5\cdots(2t-1)}{d(d+2)\cdots(d+2(t-1))}\left(\sum_{\ell=1}^{n}\|v_\ell\|^{2t}\right)^2, \tag{1-3}$$

it is possible to find them numerically, for $n$ sufficiently large [Bramwell 2011]. In this way, [Hughes and Waldron 2018] found *putatively optimal* spherical half-designs of order $2t$ for a given $t$ and $d$, i.e., those with the smallest number of vectors. This was done by using an iterative algorithm that attempts to minimise the difference between the left-hand and right-hand sides of (1-2) by making appropriate perturbations, starting from an initial guess. A similar search for putatively optimal spherical $t$-designs for the sphere ($d = 3$) was done in [Hardin and Sloane 1996]. Analogous searches for complex (projective) spherical designs include [Renes et al. 2004] (complex equiangular lines) and [Roy and Scott 2007] (complex weighted $t$-designs, with predetermined weights).

In this paper, we give some explicit spherical half-designs motivated by the putatively optimal ones found in [Hughes and Waldron 2018]. Since these exhibit a high degree of symmetry, we believe them to be optimal, i.e., to have the minimum number of vectors possible. Before doing this, we give a couple of examples.

**Example 1.2** ($t = 1$). The spherical half-designs of order 2 are precisely the *tight frames* for $\mathbb{R}^d$ [Waldron 2003]. A sequence of vectors $(v_j)$ is a *tight frame* for $\mathbb{R}^d$, see [Waldron 2018], if it satisfies the generalised Parseval identity

$$x = \frac{1}{A}\sum_{j=1}^{n}\langle x, v_j\rangle v_j \quad \text{for all } x\in\mathbb{R}^d, \text{ where } dA = \sum_j \|v_j\|^2.$$

**Example 1.3** (tight spherical $t$-designs). The term "tight" is also used for a spherical $t$-design which gives equality in the estimate

$$N(d,t) \geq \begin{cases} \binom{d-1+k}{d-1} + \binom{d-2+k}{d-1}, & t = 2k, \\ 2\binom{d-1+k}{d-1}, & t = 2k+1, \end{cases} \tag{1-4}$$

of [Delsarte et al. 1977] for the minimal number $N(d,t)$ of vectors in a spherical $t$-design for $\mathbb{R}^d$. A tight spherical $(2t+1)$-design is centrally symmetric. Therefore, if $(\pm v_j)$ is a tight spherical $(2t+1)$-design for $\mathbb{R}^d$, then $(v_j)$ is an optimal spherical half-design of order $2t$.

From the known tight spherical $(2t+1)$-designs for $\mathbb{R}^d$, $d \geq 3$ [Bannai and Bannai 2009; Nebe and Venkov 2012], we have the following optimal spherical half-designs: an orthonormal basis (Example 1.2) which comes from the cross polytope $(\pm e_j)$, 6 vectors in $\mathbb{R}^3$ (order 4) obtained from 12 vertices of the icosahedron, 28 vectors in $\mathbb{R}^7$ (order 4), 276 vectors in $\mathbb{R}^{23}$ (order 4), 120 vectors in $\mathbb{R}^8$ (order 6), 2300 vectors in $\mathbb{R}^{23}$ (order 6), 98280 vectors in $\mathbb{R}^{24}$ (order 10).

## 2. Optimal spherical half-designs for $\mathbb{R}^2$

The putatively optimal spherical half-designs for $\mathbb{R}^2$ are given by uniformly spaced lines.

**Proposition 2.1.** *The $n = t+1$ uniformly spaced lines in $\mathbb{R}^2$ given by the vectors*

$$(v_j) = \left\{ \left( \cos\frac{\pi}{n} j, \sin\frac{\pi}{n} j \right) : j = 0, \dots, n-1 \right\}$$

*are a spherical half-design of order $2t$.*

*Proof.* We will use the cubature rule that for all bivariate polynomials $p \in \Pi_{n-1}(\mathbb{R}^2)$

$$\int_{\mathbb{S}(\mathbb{R}^2)} p(x,y)\, d\sigma(x,y) = \frac{1}{2\pi} \int_0^{2\pi} p(\cos\theta, \sin\theta)\, d\theta = \frac{1}{n} \sum_{j=1}^{n} p\left( \cos\frac{2\pi}{n} j, \sin\frac{2\pi}{n} j \right).$$

We now verify that $(w_j)$ is a spherical $(t,t)$-design for $\mathbb{R}^2$, i.e., (1-3) holds. Using the trigonometric identity $\cos^2\theta = (\cos 2\theta + 1)/2$, and the cubature rule, we have

$$\begin{aligned}
\sum_j \sum_k |\langle v_j, v_k \rangle|^{2t} &= \sum_j \sum_k \left( \cos j\frac{\pi}{n} \cos k\frac{\pi}{n} + \sin j\frac{\pi}{n} \sin k\frac{\pi}{n} \right)^{2t} \\
&= \sum_j \sum_k \left( \cos(j-k)\frac{\pi}{n} \right)^{2t} = \sum_j \sum_k \left( \frac{\cos\frac{2\pi}{n}(j-k)+1}{2} \right)^t \\
&= n^2 \frac{1}{n} \sum_j \left( \frac{\cos\frac{2\pi}{n} j + 1}{2} \right)^t = n^2 \frac{1}{2\pi} \int_0^{2\pi} \left( \frac{\cos\theta + 1}{2} \right)^t d\theta.
\end{aligned}$$

The integral above simplifies to

$$\begin{aligned}
\frac{1}{2\pi} \int_0^{2\pi} \left( \frac{\cos\theta+1}{2} \right)^t d\theta &= \frac{1}{\pi} \int_0^{2\pi} \left( \cos\frac{\theta}{2} \right)^{2t} \frac{d\theta}{2} = \frac{1}{\pi} \int_0^{\pi} (\cos x)^{2t}\, dx \\
&= \frac{1}{2\pi} \int_0^{2\pi} (\cos x)^{2t}\, dx = \frac{1}{2} \cdot \frac{3}{4} \cdots \frac{2t-3}{2t-2} \cdot \frac{2t-1}{2t},
\end{aligned}$$

which gives the result. □

The corresponding spherical $(2t+1)$-design of $2(t+1)$ vectors for the circle is a *tight* spherical design (see Example 1.3), and so this configuration is an optimal spherical half-design, which is unique in the class of *rigid* spherical designs; see [Bannai and Bannai 2009].

## 3. Optimal spherical half-designs for $\mathbb{R}^3$ and $\mathbb{R}^5$

The putatively optimal spherical half-designs we present here are *weighted*. There is a corresponding notion for $t$-designs: a sequence of points $v_1, \dots, v_n$ in $\mathbb{S} \subset \mathbb{R}^d$

and *weights* $w_1, \dots, w_n \ge 0$, $w_1 + \cdots + w_n = 1$, is said to be a *weighted spherical $t$-design* if

$$\int_{\mathbb{S}} p(x)\, d\sigma(x) = \sum_{j=1}^{n} w_j\, p(v_j) \tag{3-1}$$

holds for all polynomials of degree $\le t$. The correspondence of Proposition 1.1 extends. The following is proved in [Kotelina and Pevnyi 2011] using a different definition of spherical half-designs.

**Theorem 3.1.** *Let $(v_j)$ be a sequence of $n$ vectors in $\mathbb{R}^d$ and*

$$w_j = w_j^{(t)} := \frac{\|v_j\|^{2t}}{\sum_k \|v_k\|^{2t}}$$

*be its weights as a spherical half-design of order $2t$. Then following are equivalent*:

(i) *$(\pm v_j/\|v_j\|)$, $(w_j/2)$ is a weighted spherical $(2t+1)$-design of $2n$ vectors for $\mathbb{R}^d$.*

(ii) *$(v_j)$ is a weighted spherical half-design of order $2t$ of $n$ vectors for $\mathbb{R}^d$.*

*Proof.* First suppose that $(\pm v_j/\|v_j\|)$, $(w_j/2)$ is a weighted spherical $(2t+1)$-design of $2n$ vectors for $\mathbb{R}^d$ (the weight for $\pm v_j/\|v_j\|$ is $w_j/2$). Then

$$\begin{aligned}\int_{\mathbb{S}} p(x)\, d\sigma(x) &= \sum_{j=1}^{n} \frac{w_j^{(t)}}{2}\left\{p\left(\frac{v_j}{\|v_j\|}\right) + p\left(-\frac{v_j}{\|v_j\|}\right)\right\} \\ &= \sum_{j=1}^{n} w_j^{(t)}\, p\left(\frac{v_j}{\|v_j\|}\right) \quad \text{for all } p \in \Pi_{2t}^{\circ}(\mathbb{R}^d),\end{aligned}$$

so that $(v_j)$ is a weighted spherical half-design of order $2t$.

Now suppose that $(v_j)$ is a weighted spherical half-design of order $2t$. Then the integration rule with points $(\pm v_j/\|v_j\|)$ and weights $(w_j/2)$ integrates $\Pi_{2t}^{\circ}(\mathbb{R}^d)$ (by the above calculation). It also integrates the homogeneous polynomials of odd order (since $p(x) + p(-x) = 0$ when $p$ is odd) and the constants (since the weights add to 1). It therefore only remains to show that this rule integrates $\Pi_{2r}^{\circ}(\mathbb{R}^d)$, $1 \le r < t$. A direct calculation, see [Waldron 2017], of the condition (1-3) shows that $(\|v_j\|^{t/r-1} v_j)$ is a spherical half-design of order $2r$, $1 \le r \le t$. Thus, we have the integration rule

$$\int_{\mathbb{S}} p(x)\, d\sigma(x) = \sum_{j=1}^{n} w_j^{(r)}\, p\left(\frac{v_j}{\|v_j\|}\right) \quad \text{for all } p \in \Pi_{2r}^{\circ}(\mathbb{R}^d),$$

where

$$w_j^{(r)} = \frac{\|\|v_j\|^{t/r-1} v_j\|^{2r}}{\sum_k \|\|v_k\|^{t/r-1} v_k\|^{2r}} = \frac{\|v_j\|^{2t}}{\sum_k \|v_k\|^{2t}} = w_j^{(t)}, \quad 1 \le r \le t.$$

Therefore, for $p \in \Pi^{\circ}_{2r}(\mathbb{R}^d)$, we have

$$\sum_{j=1}^{n} \frac{w_j^{(t)}}{2}\left\{p\left(\frac{v_j}{\|v_j\|}\right) + p\left(-\frac{v_j}{\|v_j\|}\right)\right\} = \sum_{j=1}^{n} w_j^{(r)}\, p\left(\frac{v_j}{\|v_j\|}\right) = \int_{\mathbb{S}} p(x)\, d\sigma(x),$$

as desired. □

This gives a 1-1 correspondence, where the weighted spherical half-designs are given up to multiplication of its vectors by $\pm 1$ and fixed nonzero scalar. In particular, if $(\pm u_j)$, $(w_j)$ is a weighted spherical $(2t+1)$-design, then $v_j := w_j^{1/(2t)} u_j$ gives the corresponding spherical half-design of order $2t$. We note that the minimal number of vectors in a *weighted* spherical half-design of order $2t$ is an increasing function of $t$ (as it is for weighted spherical $t$-designs). We also observe that a (weighted) spherical $2t$-design is a (weighted) spherical half-design of order $2t$.

When describing our weighted spherical designs, we will use the normalised weights

$$\hat{w}_j := n w_j = \frac{n\|v_j\|^{2t}}{\sum_k \|v_k\|^{2t}}, \quad 1 \le j \le n,$$

which are all 1 for an unweighted spherical design.

We now summarise our new constructions, with details and explanation to follow.

**Theorem 3.2.** *There exist*

(i) *a weighted spherical half-design of* 16 *vectors for* $\mathbb{R}^3$ *of order* 8 (*Example 3.3*),

(ii) *a weighted spherical half-design of* 16 *vectors for* $\mathbb{R}^5$ *of order* 4 (*Example 3.5*),

*and correspondingly* (*by Theorem 3.1*)

(iii) *a weighted spherical* 9*-design of* 32 *vectors for* $\mathbb{R}^3$,

(iv) *a weighted spherical* 5*-design of* 32 *vectors for* $\mathbb{R}^5$.

*The normalised weights* $\hat{w}_j$ *for all of these designs are* $\frac{20}{21} \approx 0.9523$, $\frac{36}{35} \approx 1.0286$.

In particular, we observe that tight $t$-designs for $\mathbb{R}^n$ can exist only for $t \le 5$, $t = 7$, or $t = 11$, and for $n = (2m+1)^2 - 2$ [Bannai and Damerell 1980; Bannai et al. 2004], so that there is no tight spherical 5-design of 30 points for $\mathbb{R}^5$, and no tight spherical 9-design of 30 points for $\mathbb{R}^3$. This suggests the weighted spherical 5-design and 9-design of 32 points are indeed optimal.

Let $U$ be unitary. Since the unitary image $(Uv_j)$ of a spherical half-design $(v_j)$ for $\mathbb{R}^d$ is also a spherical half-design, one cannot recognise an exact form for spherical half-design from the individual coordinates of the vectors of a numerically generated one. Instead, one must consider the Gramian matrix $(\langle v_j, v_k\rangle)$ of the design, which determines it up to the above unitary equivalence.

In the following example, we first found an exact form for the Gramian, and then recognised the vectors to be the vertices of a pentakis dodecahedron (a Catalan

solid). A *pentakis dodecahedron* (or *kisdodecahedron*) is a dodecahedron with a pentagonal pyramid covering each face, i.e., the Kleetope of the dodecahedron.

**Example 3.3.** There is a weighted spherical half-design $(v_j)$ of 16 vectors for $\mathbb{R}^3$ of order 8, which is given by the lines through the antipodal vertices of the pentakis dodecahedron (take one of the two vertices) as follows (the six vertices/lines of the icosahedron are the first columns):

$$[v_j] := \frac{1}{\sqrt{3}} \begin{pmatrix} 0 & 1 & \tau & 0 & -1 & \tau & 1 & 1 & 1 & 1 & 0 & 0 & \frac{1}{\tau} & \frac{1}{\tau} & \tau & -\tau \\ \tau & 0 & 1 & \tau & 0 & -1 & 1 & 1 & -1 & -1 & \frac{1}{\tau} & \frac{1}{\tau} & \tau & -\tau & 0 & 0 \\ 1 & \tau & 0 & -1 & \tau & 0 & 1 & -1 & 1 & -1 & \tau & -\tau & 0 & 0 & \frac{1}{\tau} & \frac{1}{\tau} \end{pmatrix} \begin{pmatrix} \alpha\Lambda_1 & \\ & \Lambda_2 \end{pmatrix},$$

$$\tau := \frac{1+\sqrt{5}}{2} \text{ (the golden ratio)}, \quad \alpha := \sqrt{\frac{3}{1+\tau^2}}, \quad \Lambda_1 := \left(\tfrac{20}{21}\right)^{\frac{1}{8}} I_6, \quad \Lambda_2 := \left(\tfrac{36}{35}\right)^{\frac{1}{8}} I_{10}.$$

It is easy to verify that (1-3) holds. These vectors have lengths $\|v_j\| = \left(\frac{20}{21}\right)^{1/8}, \left(\frac{36}{35}\right)^{1/8}$ (respectively), and the corresponding normalised weights are

$$\frac{16\left(\frac{20}{21}\right)}{6\left(\frac{20}{21}\right)+10\left(\frac{36}{35}\right)} = \tfrac{20}{21} \approx 0.9523, \qquad \frac{16\left(\frac{36}{35}\right)}{6\left(\frac{20}{21}\right)+10\left(\frac{36}{35}\right)} = \tfrac{36}{35} \approx 1.0286. \tag{3-2}$$

By Theorem 3.1, this gives a weighted spherical 9-design of 32 points for $\mathbb{R}^3$. By way of comparison, [Hardin and Sloane 1996] gives numerical evidence for a spherical 8-design of $n = 36, 40, 42, \geq 44$ points, and of a spherical 9-design of $n = 48, 50, 52, \geq 54$ points for $\mathbb{R}^3$; [Womersley 2018] suggests $n = 50$.

Equiangular lines have long been studied in relation to spherical designs. The unit vectors $(v_j)$ in $\mathbb{R}^d$ (or the lines that they give) are said to be *equiangular* if they have equal cross-correlation, i.e.,

$$|\langle v_j, v_k \rangle| = \alpha, \quad j \neq k, \text{ for some } \textit{angle } \alpha > 0.$$

**Example 3.4** (maximal lines). The number $n$ of equiangular lines in $\mathbb{R}^d$ satisfies the absolute (or Gerzon) bound $n \leq d(d+1)/2$. When this bound is attained, the set of lines has angle $1/\sqrt{d+2}$, and hence is a spherical half-design of order 4, by checking (1-3), i.e.,

$$n \cdot 1 + (n^2 - n)\left(\frac{1}{\sqrt{d+2}}\right)^4 = \frac{3}{4}\frac{d(d+1)^2}{d+2} = \frac{1 \cdot 3}{d(d+2)} n^2.$$

Such lines can exist only when $d = 2, 3$ or $d+2$ is the square of an odd integer. On the other hand, the spherical 5-design of the $2n = d(d+1)$ vectors these lines give is tight, since (1-4) holds as

$$N(d, 5) = 2\binom{d-1+2}{d-1} = d(d+1) = 2n.$$

Therefore a set of $n = d(d+1)/2$ equiangular lines in $\mathbb{R}^d$ gives an optimal spherical half-design of order 4. These are known to exist for $d = 2, 3, 7, 23$ (there are just a few cases of tight spherical $t$-designs known; see Example 1.3).

In our final example, various subsets of equiangular lines were recognised from the Gramian, which ultimately led to the presentation we now give.

**Example 3.5.** There is a weighted spherical half-design $(v_j)$ of 16 vectors for $\mathbb{R}^5$ of order 4. This consists of 6 equiangular lines in $\mathbb{R}^5$ at an angle of $\frac{1}{5}$ (the vertices of a simplex) given by vectors of length $\left(\frac{20}{21}\right)^{1/4}$, and 10 equiangular lines in $\mathbb{R}^5$ at an angle of $\frac{1}{3}$ given by vectors of length $\left(\frac{36}{35}\right)^{1/4}$, where the angle between lines from different families is $1/\sqrt{5}$. A direct calculation shows that (1-3) holds for $t = 2$; i.e.,

$$\left(\tfrac{20}{21}\right)^{\frac{1}{2}4}\left(30\left(\tfrac{1}{5}\right)^4+6\right)+\left(\tfrac{36}{35}\right)^{\frac{1}{2}4}\left(90\left(\tfrac{1}{3}\right)^4+10\right)+\left(\tfrac{20}{21}\tfrac{36}{35}\right)^{\frac{1}{4}4}\left(120\left(\tfrac{1}{\sqrt{5}}\right)^4\right)=\tfrac{3}{35}\left(6\tfrac{20}{21}+10\tfrac{36}{35}\right)^2.$$

The Gramian can be presented as the (rank-5) block matrix (the six lines first)

$$\begin{pmatrix}\Lambda_1 & \\ & \Lambda_2\end{pmatrix}\begin{pmatrix}\frac{1}{2}BB^T & B \\ B^T & \frac{5}{6}B^TB\end{pmatrix}\begin{pmatrix}\Lambda_1 & \\ & \Lambda_2\end{pmatrix}, \qquad \Lambda_1 := \left(\tfrac{20}{21}\right)^{\frac{1}{4}} I_6, \quad \Lambda_2 := \left(\tfrac{36}{35}\right)^{\frac{1}{4}} I_{10},$$

where

$$B = \frac{1}{\sqrt{5}}\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & -1 & 1 \\ -1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \end{pmatrix}. \tag{3-3}$$

The weights for this design are $\frac{20}{21} \approx 0.9523$, $\frac{36}{35} \approx 1.0286$, the same as in Example 3.3. This design gives a 32-point weighted spherical 5-design for $\mathbb{R}^5$. By way of comparison, a tight spherical 5-design for $\mathbb{R}^5$ (which does not exist) would have $N(5, 5) = 30$ points.

The $6 \times 10$ matrix $B$ of (3-3) is very interesting, since its columns and its rows give equiangular lines in $\mathbb{R}^5$, i.e.,

$$A = \tfrac{1}{2}BB^T = \begin{pmatrix} 1 & -\frac{1}{5} & -\frac{1}{5} & -\frac{1}{5} & -\frac{1}{5} & -\frac{1}{5} \\ -\frac{1}{5} & 1 & -\frac{1}{5} & -\frac{1}{5} & -\frac{1}{5} & -\frac{1}{5} \\ -\frac{1}{5} & -\frac{1}{5} & 1 & -\frac{1}{5} & -\frac{1}{5} & -\frac{1}{5} \\ -\frac{1}{5} & -\frac{1}{5} & -\frac{1}{5} & 1 & -\frac{1}{5} & -\frac{1}{5} \\ -\frac{1}{5} & -\frac{1}{5} & -\frac{1}{5} & -\frac{1}{5} & 1 & -\frac{1}{5} \\ -\frac{1}{5} & -\frac{1}{5} & -\frac{1}{5} & -\frac{1}{5} & -\frac{1}{5} & 1 \end{pmatrix}, \tag{3-4}$$

$$C = \tfrac{5}{6}B^TB = \begin{pmatrix} 1 & -\frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} & \frac{1}{3} \\ -\frac{1}{3} & 1 & \frac{1}{3} & \frac{1}{3} & -\frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & -\frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & 1 & -\frac{1}{3} & -\frac{1}{3} & \frac{1}{3} & \frac{1}{3} & -\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & -\frac{1}{3} & 1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} \\ \frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} & \frac{1}{3} & 1 & -\frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & -\frac{1}{3} & 1 & -\frac{1}{3} & \frac{1}{3} & -\frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & -\frac{1}{3} & 1 & -\frac{1}{3} & \frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{1}{3} & -\frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & -\frac{1}{3} & 1 & \frac{1}{3} & \frac{1}{3} \\ -\frac{1}{3} & \frac{1}{3} & \frac{1}{3} & -\frac{1}{3} & \frac{1}{3} & -\frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 1 & \frac{1}{3} \\ \frac{1}{3} & -\frac{1}{3} & \frac{1}{3} & -\frac{1}{3} & \frac{1}{3} & \frac{1}{3} & -\frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 1 \end{pmatrix}. \tag{3-5}$$

The presentation of these lines also seems to be new, as they do not appear in Janet Tremain's list [2008] of concrete constructions of equiangular lines.

The weighted 16-point designs for $\mathbb{R}^3$ and $\mathbb{R}^5$ share the following properties:

- They have the same weights $\frac{20}{21} \approx 0.9523$, $\frac{36}{35} \approx 1.0286$ (which are rationals).
- For the 32-point weighted spherical $t$-designs that they correspond to, the number of points in an unweighted tight spherical design is $30 = N(3, 9) = N(5, 5)$.
- Both are the orbit of two vectors, under the projective symmetry group.

This seems to be a curious coincidence, since the designs are of different orders and are in different dimensions. Moreover, the projective symmetry groups, see [Chien and Waldron 2018], of the designs are different: for the design for $\mathbb{R}^3$ it is $A_5$ (the symmetries of the dodecahedron factored by $\langle -I \rangle$), and for the design for $\mathbb{R}^5$ it is $S_6$.

Motivated by the fact that the projective symmetry group of the design for $\mathbb{R}^5$ of Example 3.5 is (isomorphic to) $S_6$, we can give the following neat presentation of it:

$$V = [v_1, \dots, v_{16}] = [\alpha BB^T, \beta B] \in \mathbb{R}^{6\times 16}, \quad \tfrac{12}{5}\alpha^2 = \tfrac{1}{2}\sqrt{\tfrac{20}{21}}, \quad \beta^2 = \tfrac{5}{6}\sqrt{\tfrac{36}{35}},$$

where $B$ is given by (3-3), since the Gramian of this $V$ is

$$V^TV = \begin{pmatrix} \alpha BB^T \\ \beta B^T \end{pmatrix} \begin{pmatrix} \alpha BB^T & \beta B \end{pmatrix} = \begin{pmatrix} \alpha^2 (BB^T)^2 & \alpha\beta BB^TB \\ \alpha\beta B^TBB^T & \beta^2 B^TB \end{pmatrix}$$
$$= \begin{pmatrix} \frac{12}{5}\alpha^2 BB^T & \frac{12}{5}\alpha\beta B \\ \frac{12}{5}\alpha\beta B^T & \beta^2 B^TB \end{pmatrix}.$$

The vectors $(v_j)$ are in the 5-dimensional subspace $\{x \in \mathbb{R}^6 : x_1 + \cdots + x_6 = 0\}$ of $\mathbb{R}^6$, and $S_6$ acts on them by permutation of the coordinates (the first six vectors are the

orthogonal projections of the standard basis vectors onto the subspace). This action on the subspace is irreducible; indeed it is the complex reflection group $G(1,1,6)$ in the first infinite family of the Shephard–Todd classification of complex reflection groups [Lehrer and Taylor 2009].

## 4. Conclusion

We gave explicit examples of putatively optimal weighted spherical half-designs, which are the orbit of *two* vectors (Examples 3.3 and 3.5) of close to equal norm. This suggests that the *weighted* spherical designs with a high degree of symmetry and a small number of vectors are natural in some cases. The study of such spherical $t$-designs is still in its infancy; see [Sloan and Womersley 2004; Bondarenko and Gorbachev 2012; Womersley 2018; Zhou and Chen 2018].

We also clarified the very close relationship between (centrally symmetric) weighted spherical $(2t+1)$-designs and weighted spherical half-designs of order $2t$ (Theorem 3.1). In this regard, it would be interesting to know if there are any optimal (weighted) spherical $(2t+1)$-designs which are not centrally symmetric for $t$ large.

## References

[Bannai and Bannai 2009] E. Bannai and E. Bannai, "A survey on spherical designs and algebraic combinatorics on spheres", *European J. Combin.* **30**:6 (2009), 1392–1425. MR Zbl

[Bannai and Damerell 1980] E. Bannai and R. M. Damerell, "Tight spherical designs, II", *J. London Math. Soc.* (2) **21**:1 (1980), 13–30. MR Zbl

[Bannai et al. 2004] E. Bannai, A. Munemasa, and B. Venkov, "The nonexistence of certain tight spherical designs", *Algebra i Analiz* **16**:4 (2004), 1–23. In Russian; translated in *St. Petersburg Math. J.* **16**:4 (2005), 609–625. MR Zbl

[Bannai et al. 2015] E. Bannai, T. Okuda, and M. Tagami, "Spherical designs of harmonic index $t$", *J. Approx. Theory* **195** (2015), 1–18. MR Zbl

[Bondarenko and Gorbachev 2012] A. V. Bondarenko and D. V. Gorbachev, "Minimal weighted 4-designs on the sphere $S^2$", *Mat. Zametki* **91**:5 (2012), 787–790. In Russian; translated in *Math. Notes* **91**:5-6 (2012), 738–741.

[Bondarenko et al. 2013] A. Bondarenko, D. Radchenko, and M. Viazovska, "Optimal asymptotic bounds for spherical designs", *Ann. of Math.* (2) **178**:2 (2013), 443–452. MR Zbl

[Bramwell 2011] J. Bramwell, *On the existence of spherical $(t,t)$-designs*, honours project, University of Auckland, 2011, available at https://www.math.auckland.ac.nz/~waldron/Students/Jennifer/JenniferBramwelldissertation.pdf.

[Chien and Waldron 2018] T.-Y. Chien and S. Waldron, "The projective symmetry group of a finite frame", *New Zealand J. Math.* **48** (2018), 55–81. MR Zbl

[Delsarte et al. 1977] P. Delsarte, J. M. Goethals, and J. J. Seidel, "Spherical codes and designs", *Geometriae Dedicata* **6**:3 (1977), 363–388. MR Zbl

[Hardin and Sloane 1996] R. H. Hardin and N. J. A. Sloane, "McLaren's improved snub cube and other new spherical designs in three dimensions", *Discrete Comput. Geom.* **15**:4 (1996), 429–441. MR Zbl

[Hughes and Waldron 2018] D. Hughes and S. Waldron, "Spherical $(t, t)$-designs with a small number of vectors", preprint, 2018, available at https://www.math.auckland.ac.nz/~waldron/Preprints/Numerical-t-designs/numerical-t-designs.html.

[Kotelina and Pevnyi 2011] N. O. Kotelina and A. B. Pevnyi, "The Venkov inequality with weights and weighted spherical half-designs", *J. Math. Sci. (N.Y.)* **173**:6 (2011), 674–682. MR Zbl

[Lehrer and Taylor 2009] G. I. Lehrer and D. E. Taylor, *Unitary reflection groups*, Australian Mathematical Society Lecture Series **20**, Cambridge University Press, 2009. MR Zbl

[Mohammadpour and Waldron 2019] M. Mohammadpour and S. Waldron, "Complex spherical designs from group orbits", preprint, 2019. arXiv

[Nebe and Venkov 2012] G. Nebe and B. Venkov, "On tight spherical designs", *Algebra i Analiz* **24**:3 (2012), 163–171. In Russian; translated in *St. Petersberg Math. J.* **24**:3 (2013), 485–491. MR Zbl

[Renes et al. 2004] J. M. Renes, R. Blume-Kohout, A. J. Scott, and C. M. Caves, "Symmetric informationally complete quantum measurements", *J. Math. Phys.* **45**:6 (2004), 2171–2180. MR Zbl

[Roy and Scott 2007] A. Roy and A. J. Scott, "Weighted complex projective 2-designs from bases: optimal state determination by orthogonal measurements", *J. Math. Phys.* **48**:7 (2007), art. id. 072110. MR Zbl

[Roy and Suda 2014] A. Roy and S. Suda, "Complex spherical designs and codes", *J. Combin. Des.* **22**:3 (2014), 105–148. MR Zbl

[Seidel 2001] J. J. Seidel, "Definitions for spherical designs", *J. Statist. Plann. Inference* **95**:1-2 (2001), 307–313. MR Zbl

[Seymour and Zaslavsky 1984] P. D. Seymour and T. Zaslavsky, "Averaging sets: a generalization of mean values and spherical designs", *Adv. in Math.* **52**:3 (1984), 213–240. MR Zbl

[Sloan and Womersley 2004] I. H. Sloan and R. S. Womersley, "Extremal systems of points and numerical integration on the sphere", *Adv. Comput. Math.* **21**:1-2 (2004), 107–125. MR Zbl

[Tremain 2008] J. C. Tremain, "Concrete constructions of real equiangular line sets", preprint, 2008. arXiv

[Waldron 2003] S. Waldron, "Generalized Welch bound equality sequences are tight frames", *IEEE Trans. Inform. Theory* **49**:9 (2003), 2307–2309. MR Zbl

[Waldron 2017] S. Waldron, "A sharpening of the Welch bounds and the existence of real and complex spherical $t$-designs", *IEEE Trans. Inform. Theory* **63**:11 (2017), 6849–6857. MR Zbl

[Waldron 2018] S. F. D. Waldron, *An introduction to finite tight frames*, Springer, 2018. MR Zbl

[Womersley 2018] R. S. Womersley, "Efficient spherical designs with good geometric properties", pp. 1243–1285 in *Contemporary computational mathematics: a celebration of the 80th birthday of Ian Sloan*, edited by J. Dick et al., Springer, 2018. MR Zbl

[Zhou and Chen 2018] Y. Zhou and X. Chen, "Spherical $t_\epsilon$-designs for approximations on the sphere", *Math. Comp.* **87**:314 (2018), 2831–2855. MR Zbl

Received: 2018-09-05 Revised: 2019-06-03 Accepted: 2019-11-04

dhug729@aucklanduni.ac.nz *Department of Mathematics, University of Auckland, Auckland, New Zealand*

waldron@math.auckland.ac.nz *Department of Mathematics, University of Auckland, Auckland, New Zealand*

INVOLVE 13:2 (2020)
dx.doi.org/10.2140/involve.2020.13.205

# A series of series topologies on $\mathbb{N}$

Jason DeVito and Zachary Parker

(Communicated by Józef H. Przytycki)

Each series $\sum_{n=1}^{\infty} a_n$ of real strictly positive terms gives rise to a topology on $\mathbb{N} = \{1, 2, 3, \dots\}$ by declaring a proper subset $A \subseteq \mathbb{N}$ to be closed if $\sum_{n\in A} a_n < \infty$. We explore the relationship between analytic properties of the series and topological properties on $\mathbb{N}$. In particular, we show that, up to homeomorphism, $|\mathbb{R}|$-many topologies are generated. We also find an uncountable family of examples $\{\mathbb{N}_\alpha\}_{\alpha\in[0,1]}$ with the property that for any $\alpha < \beta$, there is a continuous bijection $\mathbb{N}_\beta \to \mathbb{N}_\alpha$, but the only continuous functions $\mathbb{N}_\alpha \to \mathbb{N}_\beta$ are constant.

## 1. Introduction

Consider a series $\sum_{n=1}^{\infty} a_n$ whose terms are strictly positive real numbers. For any $A \subseteq \mathbb{N} = \{1, 2, 3, \dots\}$, we will use the notation $\sum_A a_n$ as a shorthand for $\sum_{n\in A} a_n$.

If $A_1, A_2 \subseteq \mathbb{N}$ are chosen so that $\sum_{A_1} a_n < \infty$ and $\sum_{A_2} a_n < \infty$, then

$$\sum_{A_1\cup A_2} a_n \le \sum_{A_1} a_n + \sum_{A_2} a_n < \infty.$$

Further, if $\sum_A a_n < \infty$ and $B \subseteq A$, then $\sum_B a_n \le \sum_A a_n < \infty$. Thus, the collection $\left\{A : \sum_A a_n < \infty\right\} \cup \{\mathbb{N}\}$ forms a topology of closed sets on $\mathbb{N}$. We use the notation $\mathbb{N}_{a_n}$ to refer to $\mathbb{N}$ equipped with this topology.

If one does not include $\mathbb{N}$ in the topology, then one obtains an ideal of sets $\mathcal{I} \subseteq \mathcal{P}(\mathbb{N})$, called a summable ideal. Summable ideals have been studied considerably by set theorists (see, for example [Flašková 2008; Hrušák 2011; Katětov 1968; Kwela and Tryba 2017; Mrożek 2016]), but do not seem to have been considered from the topological perspective before.

One focus of our paper is to relate analytic properties of the series with some of the usual basic topological properties. For example, we can characterize when $\mathbb{N}_{a_n}$ is connected.

**Theorem 1.1.** *The series $\sum_{\mathbb{N}} a_n$ diverges if and only if $\mathbb{N}_{a_n}$ is connected.*

*MSC2010:* 54A10, 54G99.
*Keywords:* series, countable topologies.

Likewise, we have a characterization of compactness.

**Theorem 1.2.** *The series $\sum_{\mathbb{N}} a_n$ has $\inf\{a_n\} > 0$ if and only if $\mathbb{N}_{a_n}$ is compact.*

Another focus of this paper is on continuity of maps to and from an $\mathbb{N}_{a_n}$. We first establish that continuous maps between an $\mathbb{N}_{a_n}$ and a "nice" topological space are constant, unless $\mathbb{N}_{a_n}$ is discrete. More specifically, we prove the following theorem.

**Theorem 1.3.** *If $X$ is compact, connected, and Hausdorff, or $X$ is path-connected then any continuous map $f : X \to \mathbb{N}_{a_n}$ is constant. If $Y$ is metrizable and $\sum_{\mathbb{N}} a_n$ diverges, then any continuous function $f : \mathbb{N}_{a_n} \to Y$ is constant.*

We also investigate continuous maps $\mathbb{N}_{a_n} \to \mathbb{N}_{b_n}$. Our main theorem finds sufficient conditions for $\mathbb{N}_{a_n}$ and $\mathbb{N}_{b_n}$ to have distinct homeomorphism types.

**Theorem 1.4.** *Suppose $\sum_{\mathbb{N}} a_n$ and $\sum_{\mathbb{N}} b_n$ are both series consisting of positive terms and assume $\lim_{n\to\infty} a_n = \lim_{n\to\infty} a_n/b_n = 0$ and that $\sum_{\mathbb{N}} b_n$ diverges. Then $\mathbb{N}_{a_n}$ and $\mathbb{N}_{b_n}$ are not homeomorphic.*

By considering $a_n = 1/n^p$ and $b_n = 1/n^q$ for $0 \le p < q \le 1$, we show, as a simple corollary, that this theorem gives $|\mathbb{R}|$-many homeomorphism types among the $\mathbb{N}_{a_n}$.

In fact, Theorem 1.4 can be strengthened in the case where $a_n = 1/n^p$ and $b_n = 1/n^q$ for $p \in [0, 1]$, $p < q$.

**Theorem 1.5.** *Suppose $0 \le p \le 1$ and $p < q$. Then the only continuous functions $\mathbb{N}_{1/n^p} \to \mathbb{N}_{1/n^q}$ are constant.*

Often when initially learning topology, one questions whether a continuous bijection is necessarily a homeomorphism. Of course, there are simple counterexamples to this conjecture, e.g., $e^{i\theta} : [0, 2\pi) \to S^1$, but Theorem 1.5 provides an interesting uncountable family of counterexamples: the identity function $\mathbb{N}_{1/n^q} \to \mathbb{N}_{1/n^p}$ is a continuous bijection, but the only continuous functions $\mathbb{N}_{1/n^p} \to \mathbb{N}_{1/n^q}$ are constants.

We also find an interesting uncountable family of spaces with diversity one. Recall that a topological space is said to have *diversity one* or to have *homeomorphic open sets* if any two nonempty open sets are homeomorphic. Such topological spaces have been previously studied [Rajagopalan and Franklin 1990], under the added restriction that the underlying topology is Hausdorff.

**Theorem 1.6.** *Suppose $\inf\{a_n\} = 0$ and $\sum_{\mathbb{N}} a_n$ diverges. Then every nonempty open subset of $\mathbb{N}_{a_n}$ is homeomorphic to $\mathbb{N}_{a_n}$.*

Theorems 1.4, 1.5, and 1.6 can be deduced from more general results in [Kwela and Tryba 2017; Mrożek 2016]. We choose to include proofs of these theorems for two reasons. First, our proofs use simpler tools. In fact, they employ nothing beyond the curriculum of a standard American calculus II course. Second, [Kwela

and Tryba 2017; Mrożek 2016] are intended for set theorists. While the intersection of topologists and set theorists is large, we hope that by expressing the results in topological language, they become more widely accessible.

The outline of this paper is as follows. In Section 2, we establish the connection between analytic properties of the series and basic topological properties (compactness, connectedness, etc.), proving Theorems 1.1 and 1.2. In Section 3, we establish properties about continuous maps between "nice spaces" and $\mathbb{N}_{a_n}$, proving Theorem 1.3. Finally, in Section 4, we study continuous functions $\mathbb{N}_{a_n} \to \mathbb{N}_{b_n}$, proving Theorems 1.4, 1.5, and 1.6.

## 2. Basic topological properties

As mentioned in the introduction, given any series of positive terms $\sum_{\mathbb{N}} a_n$, one can construct a topology on $\mathbb{N}$ by declaring a subset $A \subseteq \mathbb{N}$ to be closed if and only if $\sum_A a_n < \infty$ or $A = \mathbb{N}$. The restriction that all $a_n$ be positive serves two purposes. First, it means that the convergence of a series of the form $\sum_A a_n$ for $A \subseteq \mathbb{N}$ is independent of the order we take the sum. Second, as the following example shows, allowing a mixture of infinitely many positive and negative terms can lead to cases where the above prescription does not actually define a topology.

**Proposition 2.1.** *The set* $\left\{A \subseteq \mathbb{N} : \sum_A (-1)^{n+1}/n < \infty\right\} \cup \{\mathbb{N}\}$ *is not a topology of closed sets on* $\mathbb{N}$.

*Proof.* For each $n \in \mathbb{N}$, let $A_n$ denote the set $\mathbb{N}\setminus\{2, 4, 6, \ldots, 2n\}$. Then $\sum_{A_n} a_n < \infty$, because $A_n$ differs from $\mathbb{N}$ only on a finite set and $\sum_{\mathbb{N}} a_n = \ln(2) < \infty$.

On the other hand $\bigcap_{n\in\mathbb{N}} A_n$ is precisely the odd numbers $O$, and

$$\sum_O a_n = \sum_{\mathbb{N}} \frac{1}{2n+1} = \infty. \qquad \square$$

Of course, the above proof can be adapted to any conditionally convergent series. Henceforth, all the terms in any sum will be positive real numbers.

We begin with some basic topological properties of $\mathbb{N}_{a_n}$.

**Proposition 2.2.** *For any series of positive terms* $\sum_{\mathbb{N}} a_n$, *the topological space* $\mathbb{N}_{a_n}$ *has the following properties*:

(1) *Points are closed.*

(2) *If* $B \subseteq A \subsetneq \mathbb{N}$ *with* $A$ *closed, then* $B$ *is closed.*

(3) *For any* $A \subseteq \mathbb{N}$, *either* $\bar{A} = A$ *or* $\bar{A} = \mathbb{N}$.

*Proof.* First, (1) follows because every finite sum automatically converges.

For (2), let $B \subseteq A \subsetneq \mathbb{N}$ with $A$ closed. Then $\sum_B a_n \leq \sum_A a_n < \infty$, so $B$ is closed as well.

For (3), if $\bar{A} \neq \mathbb{N}$, then $\sum_A a_n \leq \sum_{\bar{A}} a_n < \infty$, so $A$ is closed, that is, $A = \bar{A}$. $\square$

In fact, the second and third properties in Proposition 2.2 are equivalent in any topological space, as shown in the next proposition.

**Proposition 2.3.** *Suppose $X$ is a topological space. Then $X$ has the property that every subset is closed or dense if and only if $X$ has the property that every subset of a proper closed set is closed.*

*Proof.* Assume initially that the closure of every set is itself or $X$. Let $A \subsetneq X$ be a closed set and let $B \subseteq A$ be arbitrary. Then, $\overline{B} \subseteq A$, so $\overline{B} \neq X$. By assumption, $\overline{B} = B$ so $B$ is closed.

Conversely, assume every subset of a proper closed set is closed and let $A \subseteq X$ be arbitrary. If $\overline{A} \neq X$, then $\overline{A}$ is a proper closed set. Since $A \subseteq \overline{A}$, the set $A$ must be closed. □

By taking complements, it follows immediately that supersets of nonempty open sets are open, and, equivalently, that the interior of any subset is either itself or empty.

We now show that many familiar topological properties are equivalent to natural analytic properties. We begin with a characterization of the discrete topology, which in particular, proves Theorem 1.1.

**Theorem 2.4.** *The following are equivalent*:

(1) $\sum_{\mathbb{N}} a_n$ *converges.*

(2) $\mathbb{N}_{a_n}$ *is discrete.*

(3) $\mathbb{N}_{a_n}$ *is disconnected.*

(4) $\mathbb{N}_{a_n}$ *is Hausdorff*

*Proof.* To show (1) $\Longrightarrow$ (2), suppose $\sum_{\mathbb{N}} a_n$ converges and let $A \subseteq \mathbb{N}$ be arbitrary. Then $\sum_A a_n \leq \sum_{\mathbb{N}} a_n < \infty$, so every set is closed. This shows $\mathbb{N}_{a_n}$ is discrete.

The implication (2) $\Longrightarrow$ (4) is obvious, as is (2) $\Longrightarrow$ (3) for any set with more than one point.

To see that (3) $\Longrightarrow$ (1) and (4) $\Longrightarrow$ (1), we note that both conditions (3) and (4) imply there are two nonempty proper closed $A_1, A_2$ for which $A_1 \cup A_2 = \mathbb{N}$. For (3), one can obtain $A_1$ and $A_2$ as complements of any two disconnecting open sets. For (4), one can obtain $A_1$ and $A_2$ as complements of any two disjoint proper open sets. Because both $A_i$ are proper, $\sum_{A_i} a_n < \infty$. Thus,

$$\sum_{\mathbb{N}} a_n \leq \sum_{A_1} a_n + \sum_{A_2} a_n < \infty,$$

so the series converges. □

We now give a similar characterization of when the topology on $\mathbb{N}$ is the cofinite topology, which will encompass Theorem 1.2.

**Theorem 2.5.** *The following are equivalent*:

(1) $\inf\{a_n\} > 0$.

(2) $\mathbb{N}_{a_n}$ *is the cofinite topology.*

(3) $\mathbb{N}_{a_n}$ *is compact.*

Before proving this, we need a lemma which will be used again in Section 4.

**Lemma 2.6.** *Suppose* $\inf\{a_n\} = 0$. *Then there is an infinite closed set.*

*Proof.* Because $\inf\{a_n\} = 0$, there is an $n_1 \in \mathbb{N}$ with $a_{n_1} \leq 1/1^2$. Continuing inductively, there is an $n_k \in \mathbb{N}$ for which both $n_k > n_{k-1}$ and $a_{n_k} \leq 1/k^2$. Setting $A = \{n_1, \ldots, n_k, \ldots\}$, we see that $\sum_A a_n \leq \sum_{\mathbb{N}} 1/k^2 < \infty$. Thus, $A \subseteq \mathbb{N}_{a_n}$ is closed. □

We now prove Theorem 2.5.

*Proof.* For (1) $\Longrightarrow$ (2), let $L = \inf\{a_n\}$ and assume $L > 0$. If $A \subseteq \mathbb{N}$ is infinite, then $\sum_A a_n \geq \sum_A L = \infty$, so $A$ is not closed, unless $A = \mathbb{N}$. Since we have previously shown that finite sets are closed, this is precisely the cofinite topology.

For (2) $\Longrightarrow$ (3), we note that any nonempty open set can only miss finitely many points. It follows easily that the cofinite topology on any set is compact.

Finally, we show (3) $\Longrightarrow$ (1) via the contrapositive. So, assume $\inf\{a_n\}$ is not greater than 0. Since the terms $a_n$ are all positive, this implies $\inf\{a_n\} = 0$. By Lemma 2.6, there is an infinite closed subset $A = \{n_1, n_2, \ldots, n_k, \ldots\}$.

Now, for each $i \in \mathbb{N}$, we let $U_i = A^c \cup \{n_1, \ldots, n_i\}$. We claim $\{U_i\}$ forms an open cover of $\mathbb{N}_{a_n}$ with no finite subcover. To see that each $U_i$ is open, simply note that $U_i^c = A \setminus \{n_1, \ldots, n_i\}$ is a subset of the closed set $A$, and so is closed. Further, the $U_i$ cover $\mathbb{N}$ because for each $n_i \in A$, we have $n_i \in U_i$ and for $n \notin A$, we have $n \in U_1$.

On the other hand, if $\{U_{i_1}, \ldots, U_{i_k}\}$ is a finite collection of the $U_i$ and $l > \max\{i_1, \ldots, i_k\}$, then $n_l \notin U_{i_1} \cup \cdots \cup U_{i_k}$. So, there is no finite subcover. □

## 3. Continuous functions to and from "nice" spaces

In this section, we study continuous functions between "nice" spaces and an $\mathbb{N}_{a_n}$, proving Theorem 1.3. In fact, Propositions 3.1 and 3.2 together are equivalent to Theorem 1.3.

We begin this section with a characterization of continuous functions $f : X \to \mathbb{N}_{a_n}$ where $X$ is a continuum or $X$ is path-connected. We recall that a topological space is called a *continuum* if it is compact, connected, and Hausdorff.

**Proposition 3.1.** *Suppose* $\sum_{\mathbb{N}} a_n$ *is a series of positive terms. If* $X$ *is a continuum or path-connected, then any continuous function* $f : X \to \mathbb{N}_{a_n}$ *is constant.*

In fact, the result holds more generally (with the same proof) if $\mathbb{N}_{a_n}$ is replaced with any countable topological space for which points are closed.

*Proof.* Assume initially that $X$ is a continuum. As discussed previously, in any series topology, finite sets are closed. Now, since $f$ is continuous, for each $n \in \mathbb{N}$, $f^{-1}(n)$ is a closed subset of $X$. Thus, we may write $X$ as a countable disjoint union of closed sets: $X = \coprod_{n\in\mathbb{N}} f^{-1}(n)$. According to [Sierpinski 1918], this is only possible if at most one of the closed sets is nonempty. That is, $f$ is constant.

Now, assume $X$ is path connected and $f : X \to \mathbb{N}_{a_n}$ is continuous. Let $p, q \in X$ be arbitrary and let $\gamma : [0, 1] \to X$ be a curve with $\gamma(0) = p$ and $\gamma(1) = q$. Then, since $f \circ \gamma$ is continuous and $[0, 1]$ is a continuum, $f \circ \gamma$ must be constant. Thus, $f(p) = f(q)$ as claimed. It follows that $f$ is constant. □

In particular, each $\mathbb{N}_{a_n}$ is totally path disconnected. This contrasts with Theorem 1.1, which asserts that $\mathbb{N}_{a_n}$ is connected if $\sum_{\mathbb{N}} a_n$ diverges.

We now change focus to the case where the domain is an $\mathbb{N}_{a_n}$. If $\sum_{\mathbb{N}} a_n$ converges, then the induced topology is discrete (Theorem 2.4), so any function $f : \mathbb{N} \to Y$ is continuous for any topological space $Y$. In particular, there are many nonconstant continuous functions $f : \mathbb{N}_{a_n} \to Y$. Nonetheless, we now show this behavior is limited to convergent series.

**Proposition 3.2.** *Suppose $\mathbb{N}_{a_n}$ is associated to a divergent series $\sum_{\mathbb{N}} a_n$. If $X$ is metrizable, then any continuous function $f : \mathbb{N} \to X$ is constant.*

*Proof.* By Theorem 1.1, the topology on $\mathbb{N}$ is connected, so $f(\mathbb{N})$ is a connected subset of $X$. We recall that a connected set in a metrizable space is either a single point, or has cardinality at least that of $\mathbb{R}$. Indeed, suppose $C$ is connected with $1 < |C| < |\mathbb{R}|$, and let $c_1, c_2 \in C$ be distinct. Let $d : X \times X \to \mathbb{R}$ be any compatible metric. Then, because the interval $[0, d(c_1, c_2))$ has the cardinality of $\mathbb{R}$, there must be an $r \in (0, d(c_1, c_2))$ for which $d(c_1, c) \neq r$ for any $c \in C$. Then $U = \{c \in C : d(c, c_1) < r\}$ and $V = \{c \in C : d(p, c_1) > r\}$ disconnect $C$.

So, as $f(\mathbb{N})$ is connected and countable, it must consist of a single point. □

## 4. Continuous functions $\mathbb{N}_{a_n} \to \mathbb{N}_{b_n}$

We begin with the following observation about continuous maps $\mathbb{N}_{a_n} \to \mathbb{N}_{b_n}$.

**Proposition 4.1.** *Suppose $f : \mathbb{N}_{a_n} \to \mathbb{N}_{b_n}$ is continuous and $\sum_{\mathbb{N}} a_n$ diverges. If $f$ is nonconstant, then the image of $f$ is dense. In particular, $f(\mathbb{N})$ must be an infinite subset of $\mathbb{N}$.*

*Proof.* This is obvious if $f$ is surjective, so we assume $f(\mathbb{N}) \subsetneq \mathbb{N}_{b_n}$. By Theorem 2.4, since $\sum_{\mathbb{N}} a_n$ diverges, $\mathbb{N}_{a_n}$ is connected. Now, if $f(\mathbb{N}) \subseteq \mathbb{N}_{b_n}$ is closed, then in the subspace topology, the connected set $f(\mathbb{N})$ is discrete, and so must consist of a single point. In particular, if $f$ is nonconstant, $f(\mathbb{N})$ cannot be closed. But, we showed in Proposition 2.2 that in any series topology, a subset is either closed or dense. □

We now work towards providing partial results characterizing the homeomorphism type of $\mathbb{N}_{a_n}$. Our first proposition provides some simple sufficient conditions which guarantee $\mathbb{N}_{a_n}$ is homeomorphic to $\mathbb{N}_{b_n}$.

**Proposition 4.2.** *$\mathbb{N}_{a_n}$ and $\mathbb{N}_{b_n}$ are homeomorphic if any of the following conditions is satisfied*:

(1) *There is a bijection $\sigma : \mathbb{N} \to \mathbb{N}$ for which $a_n = b_{\sigma(n)}$ for all $n$.*

(2) *The limit $\lim_{n\to\infty} a_n/b_n$ exists and is positive.*

(3) *The space $\mathbb{N}_{b_n}$ is not cofinite, and $a_{2n-1} = b_n$, while $\sum_{\mathbb{N}} a_{2n}$ converges.*

The last condition essentially says that a convergent series and a series $\sum_{\mathbb{N}} b_n$ with $\inf\{b_n\} = 0$ can be spliced together to give a space homeomorphic to $\mathbb{N}_{b_n}$. Using the first condition, it can be inserted anywhere and in any order.

*Proof.* For the first statement, $\sigma : \mathbb{N}_{a_n} \to \mathbb{N}_{b_n}$ is a homeomorphism. In more detail, suppose $X \subseteq \mathbb{N}_{b_n}$ is proper.

Since

$$\sum_{\sigma^{-1}(X)} a_n = \sum_{\sigma^{-1}(X)} b_{\sigma(n)} = \sum_{X} b_n,$$

we see that $X \subseteq \mathbb{N}_{b_n}$ is closed if and only if $\sigma^{-1}(X) \subseteq \mathbb{N}_{a_n}$ is closed.

For the second statement, suppose $L = \lim_{n\to\infty} a_n/b_n$ with $0 < L < \infty$. Then there is a natural number $N$ with the property that for any $n > N$,

$$\left| \frac{a_n}{b_n} - L \right| < \frac{L}{2}.$$

In particular,

$$\frac{L}{2} b_n < a_n < \frac{3L}{2} b_n$$

for all $n > N$.

We claim the identity function $\mathbb{N}_{a_n} \to \mathbb{N}_{b_n}$ is a homeomorphism. Suppose $A \subseteq \mathbb{N}_{a_n}$ is proper and closed. Decompose $A$ as $A = A_0 \cup A_1$, with $A_0 = A \cap \{0, 1, \ldots, N\}$ and $A_1 = A \setminus A_0$. Then

$$\sum_{n\in A} b_n = \sum_{A_0} b_n + \sum_{A_1} b_n < \sum_{A_0} b_n + \frac{2}{L} \sum_{A_1} a_n.$$

The first sum contains only finitely many terms, and so converges, and the second is bounded by

$$\frac{2}{L} \sum_{A} a_n < \infty,$$

so $A$ is also closed in $\mathbb{N}_{b_n}$. This shows the identity map is a closed map. Interchanging $a_n$ and $b_n$ and using the fact that $a_n < (3L/2)b_n$ shows the identity is continuous, so it is a homeomorphism.

For the last statement, since $\mathbb{N}_{b_n}$ is not cofinite, $\inf\{b_n\} = 0$ (Theorem 2.5), so, by Lemma 2.6 there is an infinite closed subset of $\mathbb{N}_{b_n}$. By shrinking this subset, we may assume the complement is infinite. Then, by (1) of this proposition, we may assume this subset is $E$, the even numbers. Let $\psi : E \to D$, where $D = \{n \in \mathbb{N} : n \not\equiv 1 \pmod 4\}$, denote any bijection. Now, our homeomorphism $\sigma : \mathbb{N}_{b_n} \to \mathbb{N}_{a_n}$ is the map

$$\sigma(n) = \begin{cases} 2n-1 & n \text{ odd}, \\ \psi(n) & n \text{ even}. \end{cases}$$

To see that $\sigma$ is surjective, note that by the definition of $\psi$, we need only show that $2n-1$, $n$ odd, represents every number which is congruent to $1 \pmod 4$. But $2n-1 = 4k+1$ is solved by the odd number $n = 2k+1$. The fact that $\sigma$ is injective follows because both $n \mapsto 2n-1$ and $\psi$ are injective, together with the fact that $2n-1$ is always congruent to $1 \pmod 4$ if $n$ is odd.

We now show that $\sigma$ and $\sigma^{-1}$ are continuous, beginning with $\sigma$. So, suppose $A \subseteq \mathbb{N}_{a_n}$ is proper and closed. Decompose $A$ as $A = A_0 \cup A_1$, where $A_0 = A \cap D$ and $A_1 = A \setminus A_0$. Then, $\sigma^{-1}(A_0) = \psi^{-1}(A_0) \subseteq E$. By assumption, $\sum_E b_n < \infty$, so $\sum_{\sigma^{-1}(A_0)} b_n \leq \sum_E b_n < \infty$.

Also, we compute that

$$\sum_{\sigma^{-1}(A_1)} b_n = \sum_{\sigma^{-1}(A_1)} a_{2n-1} = \sum_{A_1} a_n < \infty.$$

Thus,

$$\sum_{\sigma^{-1}(A)} b_n = \sum_{\sigma^{-1}(A_0)} b_n + \sum_{\sigma^{-1}(A_1)} b_n < \infty.$$

This shows that $\sigma^{-1}(A) \subseteq \mathbb{N}_{b_n}$ is closed.

Finally, assume $A \subseteq \mathbb{N}_{b_n}$ is closed and proper. We must show $\sigma(A)$ is closed. Decompose $A$ as $A = A_0 \cup A_1$, where $A_0 = A \cap E$ and $A_1 = A \setminus A_0$. Then

$$\sum_{\sigma(A_0)} a_n = \sum_{\psi(A_0)} a_n \leq \sum_D a_n = \sum_E b_n + \sum_E a_n < \infty.$$

Also,

$$\sum_{\sigma(A_1)} a_n = \sum_{A_1} a_{\sigma(n)} = \sum_{A_1} a_{2n-1} = \sum_{A_1} b_n \leq \sum_A b_n < \infty.$$

Thus,

$$\sum_{\sigma(A)} a_n = \sum_{\sigma(A_0)} a_n + \sum_{\sigma(A_1)} a_n < \infty. \qquad \square$$

Theorem 1.6 is a corollary of Proposition 4.2(3).

*Proof of Theorem 1.6.* Suppose $U \subseteq \mathbb{N}_{a_n}$ is a nonempty open set. Then $U^c$ is a proper closed set. If $U$ is finite, then $\sum_{\mathbb{N}} a_n = \sum_U a_n + \sum_{U^c} a_n < \infty$, a contradiction, so $U$ must be infinite. List the elements of $U$ as $U = \{u_1, u_2, \dots\}$, with $u_1 < u_2 < \cdots$. Consider the series $\sum_{\mathbb{N}} b_n$, where $b_n = a_{u_n}$. Then $\sum_{\mathbb{N}} a_n$ is obtained from $\sum_{\mathbb{N}} b_n$ by adjoining the convergent series $\sum_{U^c} a_n$. Proposition 4.2(c) now implies that $U$ an $\mathbb{N}_{a_n}$ are homeomorphic. □

***A proof of Theorem 1.4.*** We now establish Theorem 1.4, using a proof due to Alexey Lebedev.[1] For the remainder of this section, we will assume $\lim_{n\to\infty} a_n = \lim_{n\to\infty} a_n/b_n = 0$ and that $\sum_{\mathbb{N}} b_n$ diverges. Notice in this case that $\{a_n\}$ has a largest element, and a second largest element, etc. Thus, by Proposition 4.2(1), we assume without loss of generality that the sequence $\{a_n\}$ is nonincreasing.

We prove Theorem 1.4 via a sequence of lemmas. We let $S_n = \sum_{i=1}^n a_i$ and $T_n = \sum_{j=1}^n b_j$ be the partial sums.

**Lemma 4.3.** *We have* $\lim_{n\to\infty} S_n/T_n = 0$.

*Proof.* Because $\lim_{n\to\infty} a_n/b_n = 0$, for any $\epsilon > 0$, there is an $N$ with the property that $n > N$ implies $a_n/b_n < \epsilon$. Then, for $n > N$, we have

$$\frac{S_n}{T_n} = \frac{S_N + \sum_{N+1\le i\le n} a_i}{T_n} < \frac{S_N + \epsilon \sum_{N+1\le i\le n} b_i}{T_n} = \frac{S_N + \epsilon(T_n - T_N)}{T_n}.$$

Thus, $S_n/T_n < S_N - \epsilon T_N/T_n + \epsilon$. Now, $S_N$ and $T_N$ are constants, and $T_n \to \infty$ as $n \to \infty$, so clearly, we can make $S_n/T_n$ as small as we like by taking $n$ large enough. □

Now, suppose $f : \mathbb{N} \to \mathbb{N}$ is any bijection. For any $\epsilon > 0$, we define

$$M_\epsilon = \{i \in \mathbb{N} : a_{f(i)} < \epsilon b_i\}.$$

**Lemma 4.4.** *For any* $\epsilon > 0$, *the sum* $\sum_{M_\epsilon} b_i$ *diverges.*

*Proof.* Since $f$ is a bijection and $a_n$ is nonincreasing, we have

$$\sum_{i\le n} a_{f(i)} \le \sum_{i\le n} a_i = S_n.$$

Thus, we compute

$$\frac{S_n}{T_n} \ge \frac{\sum_{i\le n} a_{f(i)}}{T_n} \ge \frac{\sum_{i\le n, i\notin M_\epsilon} a_{f(i)}}{T_n} \ge \frac{\epsilon \sum_{i\le n, i\notin M_\epsilon} b_i}{T_n} = \frac{\epsilon\left(T_n - \sum_{i\le n, i\in M_\epsilon} b_i\right)}{T_n}.$$

Now, if $\sum_{M_\epsilon} b_i$ converges, then the fact that $T_n \to \infty$ as $n \to \infty$ implies that $S_n/T_n \ge \epsilon/2$ for $n$ large enough. This contradicts Lemma 4.3. □

We can now finish the proof of Theorem 1.4.

[1] https://math.stackexchange.com/q/2429818.

*Proof.* We now construct a set $A$ for which $\sum_{f(A)} a_n$ converges, but $\sum_A b_n$ diverges. Assuming we can do this, this shows that $f$ cannot be a homeomorphism.

We construct $A$ as a disjoint union $A = \bigcup_m A_m$, where each $A_m$ is finite. The sets $A_m$ are defined inductively, beginning with $A_0 = \varnothing$.

Now, assuming $A_0, A_1, \ldots, A_{m-1}$ have been defined and are finite, we now define $A_m$. To do so, consider the set

$$X_m := M_{2^{-m}} \setminus \left( \{i \in \mathbb{N} : a_{f(i)} \geq 2^{-m}\} \cup \bigcup_{i<m} A_i \right).$$

Notice $\{i \in \mathbb{N} : a_{f(i)} \geq 2^{-m}\}$ is a finite set since $\lim_{n\to\infty} a_n = 0$. Likewise, inductively, $\bigcup_{i<m} A_i$ is a finite set. Since $X_m$ and $M_{2^{-m}}$ differ in only a finite set, Lemma 4.4 implies that $\sum_{X_m} b_n$ diverges. In particular, there is a finite subset $Y_m \subseteq X_m$ for which $\sum_{Y_m} b_n > 1$. Among all the subsets of $Z_m \subseteq Y_m$ for which $\sum_{Z_m} b_n > 1$, we let $A_m$ denote one with minimal cardinality. Then $\sum_{A_m} b_n > 1$ but, for any $x \in A_m$, $\sum_{A_m \setminus \{x\}} b_n \leq 1$.

Now, set $A = \bigcup A_m$. We claim that $\sum_A b_n$ diverges. Indeed, we have

$$\sum_A b_n = \sum_{m=1}^{\infty} \sum_{A_m} b_n \geq \sum_{m=1}^{\infty} 1.$$

We next claim that $\sum_{f(A)} a_n$ converges. To see this, first, let $x_m$ be any element in $A_m$. Since $x_m \in A_m \subseteq X_m$, we have $a_{f(x_m)} < 2^{-m}$. Further, since $A_m \subseteq M_{2^{-m}}$ and $\sum_{A_m \setminus \{x_m\}} b_n \leq 1$,

$$\sum_{A_m \setminus \{x_m\}} a_{f(n)} \leq 2^{-m} \sum_{A_m \setminus \{x_m\}} b_n \leq 2^{-m}.$$

Thus, we see that

$$\sum_{f(A_m)} a_n = a_{f(x_m)} + \sum_{f(A_m \setminus \{x_m\})} a_n < 2^{-m} + \sum_{A_m \setminus \{x_m\}} a_{f(n)} \leq 2^{-m} + 2^{-m} = 2^{-m+1}.$$

So,

$$\sum_{f(A)} a_n = \sum_{m=1}^{\infty} \sum_{f(A_m)} a_n < \sum_{m=1}^{\infty} 2^{-m+1} < \infty. \qquad \square$$

As a simple corollary, we see that, for $0 < p < q \leq 1$, the spaces $\mathbb{N}_{1/n^p}$ and $\mathbb{N}_{1/n^q}$ are not homeomorphic. Thus, forming a series topology generates at least $|\mathbb{R}|$-many topologies which are distinct up to homeomorphism. On the other hand, there are only $|\mathbb{R}|^{|\mathbb{N}|} = |2^{|\mathbb{N}|}|^{|\mathbb{N}|} = |2^{|\mathbb{N}||\mathbb{N}|}| = |\mathbb{R}|$ series. So there are precisely $|\mathbb{R}|$-many topologies on $\mathbb{N}$ derived from series.

**4.1.** ***A proof of Theorem 1.5.*** We conclude with a proof of Theorem 1.5. We will henceforth assume $p \in [0, 1]$ and $p < q$. We must prove that any continuous map $f : \mathbb{N}_{1/n^p} \to \mathbb{N}_{1/n^q}$ is constant.

We handle the case $p = 0$ separately from the case $p > 0$. For, if $p = 0$, then $\mathbb{N}_{1/n^0}$ is the $\mathbb{N}$ with the cofinite topology, whereas $\mathbb{N}_{1/n^q}$ is not cofinite. Since any nonconstant continuous function $f : \mathbb{N}_{1/n^0} \to \mathbb{N}_{1/n^q}$ has infinite image (Proposition 4.1), Lemma 2.6 gives an infinite closed $A \subseteq f(\mathbb{N}_{1/n^0})$. Then $f^{-1}(A)$ is infinite, and so not closed in $\mathbb{N}_{1/n^0}$.

We prove the case $p > 0$ following an approach of Georgiy Shevchenko.[2] The following lemma is due to Georgiy Shevchenko.

**Lemma 4.5.** *Suppose $a_n$ and $b_n$ are positive, $\sum_{\mathbb{N}} b_n = \infty$, and $\lim_{n\to\infty} b_n = \lim_{n\to\infty} a_n/b_n = 0$. Then there is a subset $A \subseteq \mathbb{N}$ with the property that $\sum_A b_n$ diverges, while $\sum_A a_n$ converges.*

*Proof.* For each $k \in \mathbb{N}$, we let $N_k$ be a natural number with the property that $a_n/b_n \leq 1/k$ for any $n \geq N_k$. Because $\lim_{n\to\infty} b_n = 0$, but $\sum_{\mathbb{N}} b_n = \infty$, we can find $n_1', n_1 \geq N_1$ with $n_1' > n_1$ for which

$$\sum_{i=n_1}^{n_1'} b_i \in (1, 2).$$

Inductively, we can find $n_k', n_k \geq N_k$ with $n_k' > n_k > n_{k-1}'$ for which

$$\sum_{i=n_k}^{n_k'} b_i \in \left(\frac{1}{k}, \frac{2}{k}\right).$$

Finally, we set $A = \bigcup_k \{n_k, n_k + 1, n_k + 2, \ldots, n_k'\}$. Because

$$\sum_{\{n_k, n_k+1, \ldots, n_k'\}} b_n > \frac{1}{k},$$

$\sum_A b_n > \sum_{\mathbb{N}} 1/n$, so diverges. On the other hand,

$$\sum_{\{n_k, n_k+1, \ldots, n_k'\}} a_n \leq \frac{1}{k} \sum_{\{n_k, n_k+1, \ldots, n_k'\}} b_n \in \left(\frac{1}{k^2}, \frac{2}{k^2}\right),$$

so $\sum_A a_n \leq 2 \sum_{\mathbb{N}} 1/n^2 < \infty$. □

Now, suppose $0 < p \leq 1$ and $p < q$ and $f : \mathbb{N}_{1/n^p} \to \mathbb{N}_{1/n^q}$ is any nonconstant continuous function. For each $i \in \mathbb{N}$, we let $u_i = \sum_{f^{-1}(i)} 1/n^p$, with the convention that this sum is zero if $f^{-1}(i)$ is empty. Note $\sum_{f^{-1}(i)} 1/n^p$ is finite, as $f^{-1}(i)$ is a closed subset of $\mathbb{N}_{1/n^p}$.

[2] https://math.stackexchange.com/q/2434530.

**Proposition 4.6.** *If $f$ is continuous, then* $\lim_{n\to\infty} u_n = 0$.

*Proof.* Assume not. Then there is some $\epsilon > 0$ with the property that $u_n \geq \epsilon$ for $n$ in some infinite set $A \subseteq \mathbb{N}$. From Lemma 2.6, we can find an infinite subset $B \subseteq A$ for which $\sum_B 1/n^q < \infty$. Then $B$ is closed in $\mathbb{N}_{1/n^q}$, but

$$\sum_{f^{-1}(B)} \frac{1}{n^p} = \sum_B u_n \geq \sum_{\mathbb{N}} \epsilon,$$

which diverges. This contradicts the fact that $f$ is continuous. □

Now, we claim that $\sum_{\mathbb{N}} u_n^{1/p}$ diverges. Indeed, since $1/p \geq 1$, the function $x \mapsto x^{1/p}$ is convex, and thus, superadditive. In particular, for each term

$$u_i^{1/p} = \left( \sum_{f^{-1}(i)} \frac{1}{n^p} \right)^{1/p} \geq \sum_{f^{-1}(i)} \frac{1}{n}.$$

Then

$$\sum_{\mathbb{N}} u_i^{1/p} \geq \sum_{\mathbb{N}} \sum_{n \in f^{-1}(i)} \frac{1}{n} = \sum_{\mathbb{N}} \frac{1}{n},$$

so $\sum_{\mathbb{N}} u_n^{1/p}$ diverges.

Now, let $r \in (1, q/p)$. Then clearly

$$\sum_{\{n : u_n^{1/p} \leq 1/n^r\}} u_n^{1/p} \leq \sum_{\mathbb{N}} \frac{1}{n^r} < \infty,$$

so it follows that

$$\sum_{\{n : u_n^{1/p} > 1/n^r\}} u_n^{1/p}$$

diverges. We let

$$C = \left\{ n \in \mathbb{N} : u_n^{1/p} > \frac{1}{n^r} \right\}$$

and write $C = \{c_1, c_2, \dots\}$, where $c_1 < c_2$, etc. Thus $\sum_{\mathbb{N}} u_{c_n}^{1/p}$ diverges. Since $1/p \geq 1$ and $u_n \to 0$ as $n \to \infty$, it follows that for large $n$, we have $u_{c_n} \geq u_{c_n}^{1/p}$. In particular, we have the following proposition.

**Proposition 4.7.** *The series $\sum_{\mathbb{N}} u_{c_n}$ diverges.*

Finally, we have the following proposition.

**Proposition 4.8.** *We have*

$$\lim_{n\to\infty} \frac{1/(c_n)^q}{u_{c_n}} = 0.$$

*Proof.* Since $c_n \in C$, each $u_{c_n} > 0$, so the terms in the limit are defined. Now, note that if $u_n^{1/p} > 1/n^r$, then $u_n > 1/n^{rp}$. For large $n$, we see

$$0 \leq \frac{1/(c_n)^q}{u_{c_n}} \leq \frac{1/(c_n)^q}{1/(c_n)^{rp}} = c_n^{rp-q}.$$

Since $r < q/p$, we have $rp - q < 0$. Further, $c_n \to \infty$ as $n \to \infty$. Thus $\lim_{n\to\infty} c_n^{rp-q} = 0$. The result now follows from the squeeze theorem. □

If $f$ is continuous, then Propositions 4.6, 4.7, and 4.8 exactly show that the sequences $a_n = 1/(c_n)^q$ and $b_n = u_{c_n}$ satisfy the conditions of Lemma 4.5. In particular, there is a set $A \subseteq \mathbb{N}$ for which $\sum_A u_{c_n}$ diverges, but $\sum_A 1/(c_n)^q$ converges. This, in turn, means there is a subset $B \subseteq \mathbb{N}$ for which $\sum_B u_n$ diverges, but $\sum_B 1/n^q$ converges. However, we have

$$\sum_{f^{-1}(B)} \frac{1}{n^p} = \sum_{i\in B} \sum_{n\in f^{-1}(i)} \frac{1}{n^p} = \sum_B u_n = \infty.$$

Thus, $B$ is closed in $\mathbb{N}_{1/n^q}$, but $f^{-1}(B)$ is not closed in $\mathbb{N}_{1/n^p}$. This contradicts the fact that $f$ is continuous, and concludes the proof of Theorem 1.5.

## Acknowledgements

We would like to acknowledge support from the Bill and Roberta Blankenship Undergraduate Research Endowment. We would also like to thank Alexey Lebedev and Georgiy Shevchenko for several proofs and ideas appearing in Section 4. Finally, we would like to thank an anonymous referee who informed us that the main theorems in Section 4 are special cases of theorems found in [Kwela and Tryba 2017; Mrożek 2016].

## References

[Flašková 2008] J. Flašková, "Description of some ultrafilters via i-ultrafilters", pp. 1–7 in *Combinatorial and descriptive set theory*, edited by J. Brendle, Sūrikaisekikenkyūsho Kōkyūroku **1619**, RIMS, Kyoto, 2008. Zbl

[Hrušák 2011] M. Hrušák, "Combinatorics of filters and ideals", pp. 29–69 in *Set theory and its applications*, edited by L. Babinkostova et al., Contemp. Math. **533**, Amer. Math. Soc., Providence, RI, 2011. MR Zbl

[Katětov 1968] M. Katětov, "Products of filters", *Comment. Math. Univ. Carolinae* **9** (1968), 173–189. MR Zbl

[Kwela and Tryba 2017] A. Kwela and J. Tryba, "Homogeneous ideals on countable sets", *Acta Math. Hungar.* **151**:1 (2017), 139–161. MR Zbl

[Mrożek 2016] N. Mrożek, "Some applications of the Katětov order on Borel ideals", *Bull. Pol. Acad. Sci. Math.* **64**:1 (2016), 21–28. MR Zbl

[Rajagopalan and Franklin 1990] M. Rajagopalan and S. P. Franklin, "Spaces of diversity one", *J. Ramanujan Math. Soc.* **5**:1 (1990), 7–31. MR Zbl

[Sierpinski 1918] W. Sierpinski, "Un théorème sur les continus", *Tôhoku Math. J.* **13** (1918), 300–305. Zbl

Received: 2019-02-27 Revised: 2019-07-07 Accepted: 2019-12-03

jdevito@utm.edu *University of Tennessee at Martin, Martin, TN, United States*

zacppark@ut.utm.edu *University of Tennessee at Martin, Martin, TN, United States*

msp

INVOLVE 13:2 (2020)
dx.doi.org/10.2140/involve.2020.13.219

# Discrete Morse functions, vector fields, and homological sequences on trees

Ian Rand and Nicholas A. Scoville

(Communicated by Colin Adams)

We construct a discrete Morse function which induces both a specified gradient vector field and homological sequence on a given tree. After reviewing the basics of discrete Morse theory, we provide an algorithm to construct a discrete Morse function on a tree inducing a desired gradient vector field and homological sequence. We prove that our algorithm is correct, and conclude with an example to illustrate its use.

## 1. Introduction

Let $f$ and $g$ be two discrete Morse functions defined on a 1-dimensional simplicial complex, i.e., a graph. Inspired by Nicolaescu [2008], R. Ayala et al. [2009a] studied the homological sequence of a discrete Morse function by introducing the notion of $f$ and $g$ being homologically equivalent, and they counted the number of excellent discrete Morse functions on all graphs [Ayala et al. 2011]. This result was used to compute the number of excellent discrete Morse functions on all collapsible 2-dimensional complexes in [Agiorgousis et al. 2019]. Furthermore, R. Ayala et al. [2008] studied gradient vector fields on trees by considering discrete Morse functions up to Forman equivalence. In this paper, we desire to combine these two ideas by investigating the combination of the two notions of equivalence. To this end, let $\mathcal{V}$ be a fixed gradient vector field on a graph with $m > 1$ critical values. Which homological sequences can arise from a discrete Morse function that induces $\mathcal{V}$ as its gradient vector field? Conversely, fix a homological sequence $B$ with $m > 1$ critical values. Which discrete vector fields can arise from a discrete Morse function with $B$ as its homological sequence? In the case that $G$ is a tree, we show in Proposition 3.5 that any feasible gradient vector field and homological sequence with the same number of critical values can be realized by a discrete Morse function. This is via a simple algorithmic construction, the details of which are given in Algorithm 2. In addition, this algorithm provides a new proof of Theorem 6.1 in

*MSC2010:* primary 05E45; secondary 57M15, 05C05, 68R10.
*Keywords:* discrete Morse theory, homological sequence, gradient vector field, trees, Dyck path.

[Ayala et al. 2009a] in the case where $G$ is a tree. The layout of the paper is as follows. Section 2 is devoted to introducing the notation and terminology that will be used, as well as background results. Section 3 is the main section where we give an algorithm that yields a desired homological sequence and gradient vector field on a tree. Finally, we give an example illustrating the utility of our algorithm.

## 2. Background

We now establish notation and terminology that will be used in the body of this paper. Our main reference for graph theory is [Chartrand 1985], while we refer the reader to [Forman 2002; Scoville 2019] for the basics of discrete Morse theory.

### *Graphs.*

**Definition 2.1.** Let $V \neq \varnothing$ be a finite set called the *vertex set*. A *graph* $G = (V, E)$ is a collection of distinct subsets of $V$ of size 2, denoted by $E(G)$, called the *edge set*. Elements of $V = V(G)$ are *vertices*, while elements of $E$ are called *edges*. If $e = \{a, b\}$ is any edge, $a$ and $b$ are *endpoints* of $e$. We also say that $a$ and $b$ are *incident* with edge $e$. We write $ab = \{a, b\}$ for an edge when there is no possibility of confusion.

**Definition 2.2.** Let $u, v$ be two distinct vertices of $G$. A $(u, v)$-*path* is a sequence

$$u = u_0,\ e_0,\ u_1,\ e_1,\ \ldots,\ e_n,\ u_{n+1} = v$$

of distinct vertices and edges such that $u_i, u_{i+1}$ are the endpoints of $e_i$. Suppose $G$ is a graph such that for every pair of distinct vertices $u, v$ there exists a unique $(u, v)$-path. Then $G$ is a *tree*. A disconnected graph in which each component is a tree is called a *forest*.

There are several other characterizations of trees. They will be utilized below without further reference.

**Theorem 2.3** (characterization of trees). *Let $G$ be a connected graph with $v$ vertices and $e$ edges. The following are equivalent*:

(a) *$G$ is a tree.*

(b) $v = e + 1$.

(c) *$G$ contains no cycles.*

(d) $b_1(G) = 0$, *where $b_1$ is the first Betti number of $G$* (see [Ferrario and Piccinini 2011, Chapter II.4]).

(e) *The removal of any edge from $G$ results in a disconnected graph.*

Proofs of the equivalence of the statements may be found in any graph theory textbook; e.g., [Chartrand et al. 2016, Chapter 2.2].

***Discrete Morse theory.*** We now introduce discrete Morse theory on graphs.

**Definition 2.4.** Let $G$ be a graph. A *discrete Morse function* $G$ is a function $f : G \to \mathbb{R}$ such that for every $v \in G$ we have

$$|\{e \in E : f(e) \geq f(v), v \text{ an endpoint of } e\}| \leq 1$$

and for every $e \in G$

$$|\{v \in V : f(v) \geq f(e), v \text{ an endpoint of } e\}| \leq 1.$$

If vertex $v$ satisfies

$$|\{e \in E : f(e) \geq f(v), v \text{ an endpoint of } e\}| = 0$$

then $v$ is a *critical vertex* and the value $f(v)$ is a *critical value*. If an edge $e$ satisfies

$$|\{v \in V : f(v) \geq f(e), v \text{ an endpoint of } e\}| = 0$$

then $e$ is a *critical edge* and $f(e)$ is a *critical value*. A vertex or edge that is not critical is called a *regular vertex* or *regular edge*, respectively. If all the critical values of $f$ are distinct, $f$ is said to be *excellent*.

We will assume that all discrete Morse functions are excellent. The critical values of a discrete Morse function tell us how to "build" the graph $G$ in stages. This is formally accomplished through the level subcomplexes.

**Definition 2.5.** Let $f : G \to \mathbb{R}$ be a discrete Morse function, $c \in \mathbb{R}$. The smallest subgraph of $G$ containing all $v, e$ such that $f(e), f(v) \leq c$ is called a *level subcomplex*, denoted by $G(c)$.

Although it can be defined more generally, for the purposes of this paper we restrict the following definition to the case when $G = T$ is a tree. Recall that $b_0(T)$ is the 0*-th Betti number of* $T$; i.e., $b_0(T)$ is the number of connected components of $T$.

**Definition 2.6.** Let $f : T \to \mathbb{R}$ be an excellent discrete Morse function with critical values $c_0 < c_1 < c_2 < \cdots < c_{m-1}$. The *homological sequence* of $f$, denoted by $B_f$, is given by $b_0(T(c_0)), b_0(T(c_1)), \ldots, b_0(T(c_{m-1}))$. We write $B(i) = B_f(i) := b_0(T(c_i))$ when the discrete Morse function and tree are clear from the context. Two discrete Morse functions $f, g : T \to \mathbb{R}$ are said to be *homologically equivalent* if $f$ and $g$ induce the same homological sequence.

In other words, the homological sequence records the number of connected components at each level subcomplex, associating a sequence to a discrete Morse function. For an excellent discrete Morse function, exactly one single component will be added or removed at each level subcomplex. Formally:

**Proposition 2.7** [Ayala et al. 2009a]. *If $f$ is an excellent discrete Morse function with $m$ critical values on a connected tree $T$, then the homological sequence of $f$ satisfies $|B(i+1) - B(i)| = 1$.*

In addition to homological equivalence, there is the original notion of equivalence of two discrete Morse functions due to Robin Forman [1998].

**Definition 2.8.** Let $f, g : G \to \mathbb{R}$ be discrete Morse functions. Then $f$ and $g$ are *Forman equivalent* if for every pair $(v, e)$ consisting of a vertex $v$ and an incident edge $e$, we have $f(v) < f(e)$ if and only if $g(v) < g(e)$.

It turns out that this definition has a nice geometric characterization by passing to the gradient vector field.

**Definition 2.9.** Let $f : G \to \mathbb{R}$ be a discrete Morse function. The *induced gradient vector field* is defined by

$$\mathcal{V}_f := \{(v, e) : f(v) \geq f(e), v \text{ is incident with } e\}.$$

**Remark 2.10.** In light of the above, we may view a discrete Morse function as a pairing of a vertex with an incident edge (a regular pair) with each vertex and edge appearing in at most one pair. It then follows that any vertex or edge not in a regular pair is critical.

**Example 2.11.** Let $f$ be the discrete Morse function defined on the graph to the left. The induced gradient vector field is shown on the right:


As mentioned above, the following theorem, due to Ayala et al., shows that Forman equivalence can be characterized in terms of the gradient vector field.

**Theorem 2.12** [Ayala et al. 2009b, Theorem 3.1]. *Two discrete Morse functions $f$ and $g$ defined on $G$ are equivalent if and only if $\mathcal{V}_f = \mathcal{V}_g$.*

An immediate corollary of this result is that a gradient vector field is completely determined by the critical simplices. We will use this fact below.

**Corollary 2.13.** *Two discrete Morse functions $f$ and $g$ are Forman equivalent if and only if they have the same critical simplices.*

## 3. Main result

In this section, we combine the notions of Forman equivalence and homological equivalence to produce a discrete Morse function on a tree with a desired gradient vector field and homological equivalence.

**Definition 3.1.** Let $T$ be a tree, $v \in V(T)$, and $f : T \to \mathbb{R}$ a discrete Morse function. We say that $f$ *roots* $T$ *in* $v$ or $T$ *is rooted in* $v$ if $v$ is the unique critical simplex of $f$. Such a vertex is called the *root* of $f$.

**Lemma 3.2.** *Let $f, g : T \to \mathbb{R}$ both root $T$ in $v$. Then $f$ and $g$ are Forman equivalent.*

*Proof.* By Corollary 2.13, a gradient vector field is uniquely determined by its critical simplices. Since by definition both $f$ and $g$ have the exact same set of critical simplices (namely, a single vertex), the result follows. □

We first give an algorithm to explicitly construct a discrete Morse function on a tree $T$ rooted in a particular vertex $v$. If $G$ is a graph and $u$, $v$ are vertices of $G$, the *distance* between $u$ and $v$ is the minimum number of edges traversed over all paths from $u$ to $v$.

**Algorithm 1** (rooted-tree algorithm).
Input: A tree $T$ with a given root vertex $v_0$ and a nonnegative integer $k$.
Output: A discrete Morse function $f : T \to \mathbb{R}$ rooted in $v_0$ with $f(v_0) = k$.

(1) Label $f(v_0) = k$.

(2) For any vertex $u \neq v_0$, label $f(u) = d(v_0, u) + k$.

(3) For any edge $e = uw$, label $f(e) = \max\{f(u), f(w)\}$.

Note that hidden in (2) of Algorithm 1 is the breadth-first search algorithm, which computes the distances from a fixed vertex to all other vertices. The time complexity for this is $O(|E|)$ [Agnarsson and Greenlaw 2007, p. 400]. Since the rest of the rooted-tree algorithm is constant time, the total cost of Algorithm 1 is the same.

**Proposition 3.3.** *Algorithm 1 is correct; that is, it produces a discrete Morse function rooted in $v_0$.*

*Proof.* Clearly $f(v_0) = k$ is critical since, for any edge $uv_0$, we have $f(uv_0) = \max\{f(u), f(v_0)\} = k + 1 > k = f(v_0)$. Let $u \neq v_0$ be any vertex in $T$. We claim that there exists a unique edge $e = v_0w$ such that $f(e) \leq f(v_0)$. If so, then $f$ satisfies the definition of a discrete Morse function on each vertex and each vertex (other than $v_0$) is regular. By induction on $n$, the distance from a vertex to $v_0$, we see that there exists an edge $e$ incident with $u$ such that $f(e) = f(u)$. If $e'$ is another edge such that $f(e') = f(u)$, then we obtain two paths from $u$ to $v_0$, namely, one through $e$ and another through $e'$, contradicting the fact that paths in a tree are unique.

Now let $e = uw$ be any edge. Since the shortest path from $v_0$ to any other vertex $u$ is unique, either $f(u) > f(w)$ or $f(u) < f(w)$. Hence each edge satisfies the definition of a discrete Morse function and is regular. We conclude that $f$ is a discrete Morse function rooted in $v_0$. □

**Lemma 3.4.** *Let $T$ be a tree and $f : T \to \mathbb{R}$ a discrete Morse function on $T$ with $m$ critical vertices. Let $C_e$ be the set of critical edges of $f$. Then $T - C_e$ is a forest with exactly $T_1, T_2, \ldots, T_m$ trees, each of which contains exactly one critical vertex.*

*Proof.* Clearly $T - C_e$ is a forest. It remains to show that each tree in the forest has exactly one critical vertex. If so, then the number of trees is in bijective correspondence with the number of critical vertices and hence there would be $m$ trees. Now the discrete Morse function on $T$ restricts to a discrete Morse function on any $T_i$. By Remark 2.10, regular simplices come in vertex/edge pairs. Since all critical edges have been removed, $T_i$ is partitioned into vertex/edge pairs along with critical vertices. But since $|V(T_i)| = |E(T_i)| + 1$, such a partition can only happen if there is exactly one critical vertex. □

It then follows from Lemma 3.2 that the gradient vector field on each tree in the forest is uniquely determined.

Before giving the main algorithm, observe that a homological sequence from an excellent discrete Morse function on a tree $T$ must start with 1 and end with 1. Furthermore, by Proposition 2.7, each value increases or decreases by exactly 1, never going below 1. Such a sequence is known as a *Dyck path* [Brualdi 2010]. This was first observed in [Ayala et al. 2009a].

We are now ready to give an algorithm which constructs an excellent discrete Morse function on a tree $T$ whose induced homological sequence $B$ is a given Dyck path beginning and ending at 1 and which induces a given gradient vector field $\mathcal{V}$. Subdivide $T$ if more critical values are needed and by abuse of notation, call the resulting graph $T$. We call a Dyck path and a gradient vector field $\mathcal{V}$ *consistent* if the number of points in the Dyck path is the same as the number of critical simplices of $\mathcal{V}$.

**Algorithm 2** (homological and Forman algorithm).
Input: A tree $T$ with a given Dyck path beginning and ending at 1 and gradient vector field $\mathcal{V}$ consistent with the Dyck path.
Output: An excellent discrete Morse function $f : T \to \mathbb{R}$ whose homological sequence $B$ is the given Dyck path and gradient vector field is $\mathcal{V}$.

(a) Remove all critical edges from $T$. Initialize $n = -1$.

(b) Choose an unlabeled tree $T'$ and apply Algorithm 1 with root vertex the unique critical vertex of $T'$ with $k = n + 1 = 0$.

(c) If $b_0$ increases, choose an unlabeled tree $T'$ such that there exists a critical edge between $T'$ and a labeled tree. Apply Algorithm 1 with root vertex the unique critical vertex of $T'$ with $k = n + 1$ and go to step (e). Otherwise, go to step (d).

(d) If $b_0$ decreases, choose a critical edge connecting two labeled trees, and label this edge $n+1$. Go to step (e).

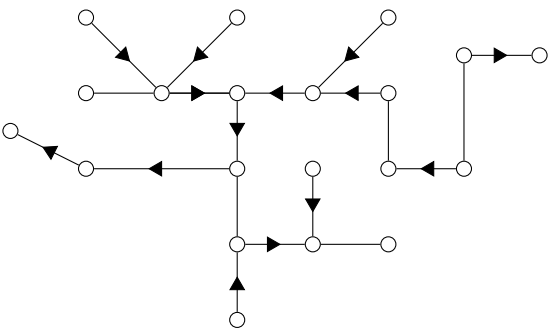
(e) If all vertices and edges are labeled, terminate. Otherwise, go to step (c).

**Proposition 3.5.** *Algorithm 2 is correct.*

*Proof.* By Algorithm 1, $T - C_e$ is labeled with a discrete Morse function with precisely the desired critical vertices. By step (d), each of the edges in $C_e$ is labeled strictly greater than its two incident vertices, and hence the edges in $C_e$ are the critical edges of $f$. Since $f$ is a discrete Morse function with precisely the desired critical simplices, Theorem 2.12 implies that $\mathcal{V}_f = \mathcal{V}$.

To see that $B = B_f$, we note that when $b_0$ increases, a new tree is added in step (c) which increases $b_0$, and when $b_0$ decreases, two trees become a single tree in step (d). Thus the result. □

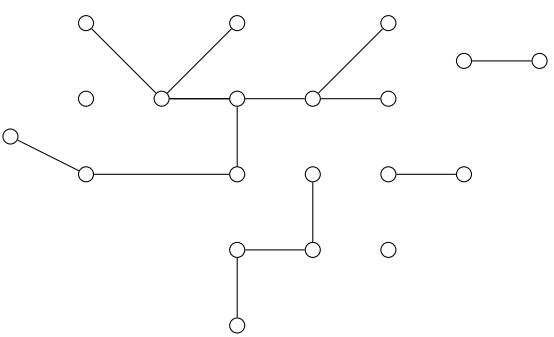
**Example 3.6.** We give an example to illustrate the construction of a discrete Morse function with a desired homological sequence and gradient vector field. Let $T$ be the tree with desired gradient vector field given below:



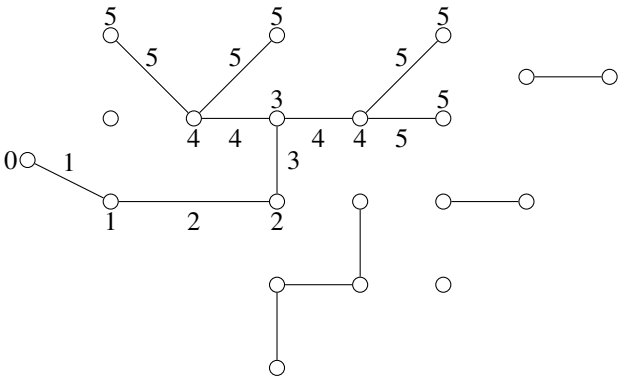
We use Algorithm 2 to construct a discrete Morse function on $T$ inducing the above gradient vector field with homological sequence (Dyck path)

$$B_0 : 1\ \ 2\ \ 3\ \ 2\ \ 1\ \ 2\ \ 3\ \ 4\ \ 3\ \ 2\ \ 1.$$

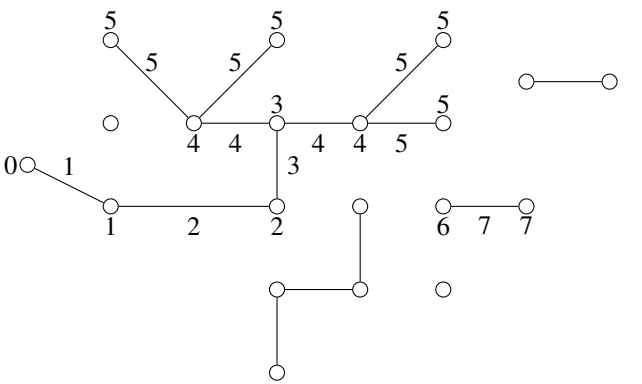
Step (a) of 2 is to remove all critical edges and initialize $n = -1$:



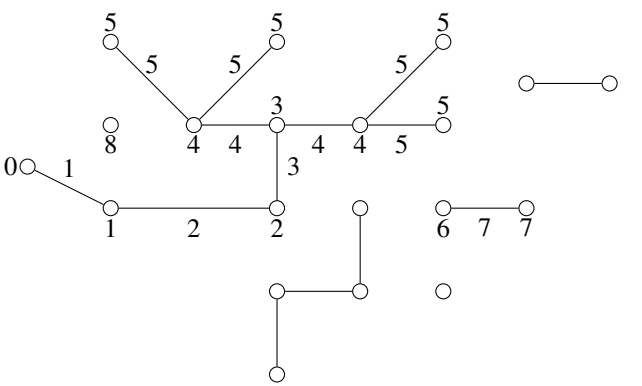
For step (b) of Algorithm 2, we pick a tree and apply Algorithm 1 with $k = n + 1 = 0$. This yields the following labeling:



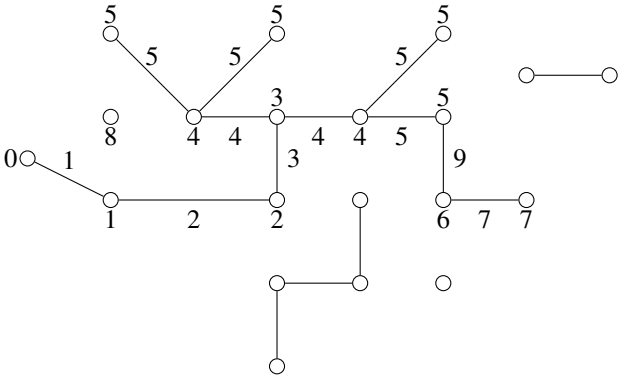

Because $B_0$ increases in the next stage of the specified homological sequence, we apply step (c) of Algorithm 2 with $k = n + 1 = 6$:



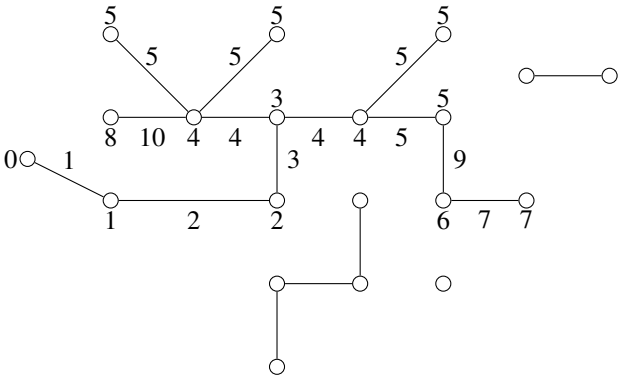

The homological sequence increases again, so we repeat with $k = n + 1 = 8$ on a tree with a critical edge connecting to a labeled tree:



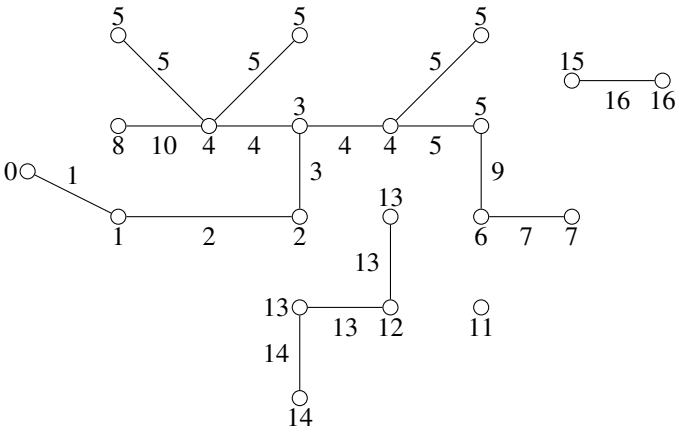
The desired homological sequence decreases at this point, so by step (d), we choose a critical edge connecting two labeled trees and label this $n+1=9$:



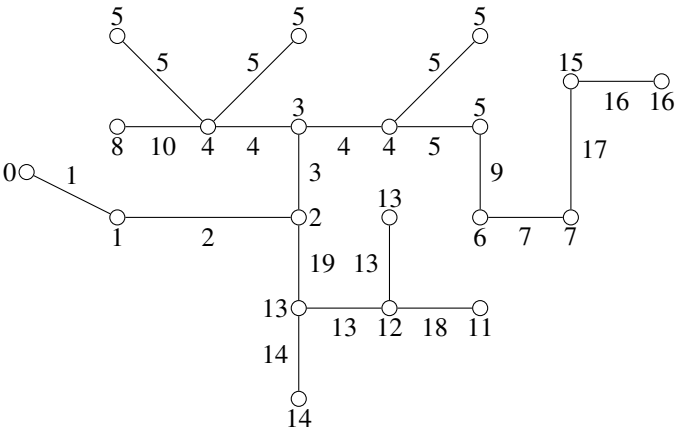

The sequence decreases again, so we repeat:




The homological sequence increase three times in a row, which we show all at once:



Finally, $B_0$ decreases three times in a row, which we show all at once:




It is now easy to check that the discrete Morse function has both the desired induced gradient vector field and homological sequence.

## Acknowledgements

This work was partially supported by a grant from the Ursinus College Summer Fellows program.

## References

[Agiorgousis et al. 2019] M. Agiorgousis, B. Green, A. Onderdonk, K. Rich, and N. A. Scoville, "Homologically equivalent discrete Morse functions", *Topology Proc.* **54** (2019), 283–294. MR Zbl

[Agnarsson and Greenlaw 2007] G. Agnarsson and R. Greenlaw, *Graph theory: modeling, applications, and algorithms*, Pearson Prentice Hall, Upper Saddle River, NJ, 2007. MR Zbl

[Ayala et al. 2008] R. Ayala, L. M. Fernández, A. Quintero, and J. A. Vilches, "A note on the pure Morse complex of a graph", *Topology Appl.* **155**:17-18 (2008), 2084–2089. MR Zbl

[Ayala et al. 2009a] R. Ayala, L. M. Fernández, D. Fernández-Ternero, and J. A. Vilches, "Discrete Morse theory on graphs", *Topology Appl.* **156**:18 (2009), 3091–3100. MR Zbl

[Ayala et al. 2009b] R. Ayala, L. M. Fernández, and J. A. Vilches, "Characterizing equivalent discrete Morse functions", *Bull. Braz. Math. Soc. (N.S.)* **40**:2 (2009), 225–235. MR Zbl

[Ayala et al. 2011] R. Ayala, D. Fernández-Ternero, and J. A. Vilches, "The number of excellent discrete Morse functions on graphs", *Discrete Appl. Math.* **159**:16 (2011), 1676–1688. MR Zbl

[Brualdi 2010] R. A. Brualdi, *Introductory combinatorics*, 5th ed., Pearson Prentice Hall, Upper Saddle River, NJ, 2010. MR Zbl

[Chartrand 1985] G. Chartrand, *Introductory graph theory*, Dover, New York, 1985. MR

[Chartrand et al. 2016] G. Chartrand, L. Lesniak, and P. Zhang, *Graphs & digraphs*, 6th ed., CRC Press, Boca Raton, FL, 2016. MR

[Ferrario and Piccinini 2011] D. L. Ferrario and R. A. Piccinini, *Simplicial structures in topology*, Springer, 2011. MR Zbl

[Forman 1998] R. Forman, "Morse theory for cell complexes", *Adv. Math.* **134**:1 (1998), 90–145. MR Zbl

[Forman 2002] R. Forman, "A user's guide to discrete Morse theory", *Sém. Lothar. Combin.* **48** (2002), art. id. B48c. MR Zbl

[Nicolaescu 2008] L. I. Nicolaescu, "Counting Morse functions on the 2-sphere", *Compos. Math.* **144**:5 (2008), 1081–1106. MR Zbl

[Scoville 2019] N. A. Scoville, *Discrete Morse theory*, Student Mathematical Library **90**, American Mathematical Society, Providence, RI, 2019. MR Zbl

Received: 2019-03-15 Accepted: 2020-03-05

i.bradford.r@gmail.com *University of Deleware, Newark, DE, United States*

nscoville@ursinus.edu *Mathematics and Computer Science Department, Ursinus College, Collegeville, PA, United States*

INVOLVE 13:2 (2020)
dx.doi.org/10.2140/involve.2020.13.231

# An explicit third-order one-step method for autonomous scalar initial value problems of first order based on quadratic Taylor approximation

Thomas Krainer and Chenzhang Zhou

(Communicated by Kenneth S. Berenhaut)

We present an explicit one-step numerical method of third order that is error-free on autonomous scalar Riccati equations such as the logistic equation. The method replaces the differential equation by its quadratic Taylor polynomial in each step and utilizes the exact solution of that equation for the calculation of the next approximation.

## 1. Introduction

One of the basic ordinary differential equations in quantitative population dynamics is the logistic differential equation

$$\begin{cases} \dot{y} = ry\left(1 - \dfrac{y}{K}\right) - qy, \\ y|_{t=0} = y_0, \end{cases}$$

where $y = y(t)$ is the size of the population at time $t \geq 0$, $r > 0$ is the maximal growth rate for the population, and $K > 0$ is the carrying capacity of the habitat for the population under study. We modified the equation in this example by a harvesting term with harvesting rate $q > 0$ as it appears, for example, in fishery models. We refer to [Thieme 2003] as a general reference for ordinary differential equation models in ecology. While the logistic model, as well as its variations and perturbations, are classical cornerstones of ecological quantitative modeling, it is remarkable that the standard numerical methods for approximating solutions to ordinary differential equations do not solve the logistic equation error-free. Motivated by this observation, we are presenting here an explicit third-order one-step numerical method that is applicable to scalar autonomous initial value problems of the form

$$\begin{cases} \dot{y} = f(y), \\ y|_{t=0} = y_0, \end{cases} \tag{1.1}$$

*MSC2010:* 65L05.
*Keywords:* numerical methods for ODEs, initial value problems.

with sufficiently smooth real-valued $f$ that approximates the solution $y = y(t)$ on the compact interval $[0, T]$ by a sequence of values $y_0, y_1, y_2, \ldots$ based on equidistant time-stepping with step size $h > 0$, and whose distinguishing feature is that the method is error-free if $f$ is a polynomial up to degree 2 such as in the logistic equation; i.e., our method solves autonomous Riccati equations exactly. The idea for this method is simple:

(1) Replace $f(y)$ in (1.1) by its quadratic Taylor polynomial $T_{y_0}(y)$ centered at $y_0$.

(2) Solve $\dot{u} = T_{y_0}(u)$ exactly with initial condition $u(0) = y_0$.

(3) Set $y_1 = u(h)$ and repeat with $y_1$ in place of $y_0$, etc.

There are several issues that arise upon implementation of this basic idea. Most importantly, it must be noted that solutions to Riccati equations can blow up in finite time, so integrity checks on the step size $h > 0$ are needed to preclude a potential blow-up of the approximate solution $u$ on the interval $(0, h]$ as otherwise the calculated term $y_1$ is invalid (and likewise in subsequent steps). To make this more transparent, consider the example

$$\begin{cases} \dot{y} = (y - \lambda)(1 - y)e^{-y^4}, \\ y|_{t=0} = 0, \end{cases}$$

where $\lambda \gg 1$. The maximal solution to this differential equation exists on $(-\infty, \infty)$ because the function

$$f(y) = (y - \lambda)(1 - y)e^{-y^4}, \quad -\infty < y < \infty,$$

is bounded. However, the maximal solution to $\dot{u} = T_0(u) = (u-\lambda)(1-u)$ with initial value $u(0) = 0$ blows up at $t = \ln(\lambda)/(\lambda - 1)$. In view of $\lim_{\lambda\to\infty} \ln(\lambda)/(\lambda - 1) = 0$ we see that blow-up does occur on $(0, h)$ if $\lambda \gg 1$ is large enough.

We provide two options to deal with this problem in the implementation, a priori or at run time. The a priori option calculates a threshold for how small the step size $h > 0$ ought to be chosen at the outset to avoid invalid approximating terms throughout and is based on the differential equation (1.1) and the viewing window $[0, T] \times [y_{\min}, y_{\max}]$ where its solution is supposed to be approximated as inputs, while the run time option checks validity of each approximation value at the time when it is calculated. The issue of blow-up is germane to the method we discuss in this note; it does not occur in standard Runge–Kutta methods or exponential integrators.

A second issue that we needed to address in the implementation concerns evaluation of the formula for the approximate solution $u$ itself. If the roots of the Taylor polynomial $T_{y_0}$ are distinct but close, the exact formula for $u$ would require evaluation of quotients nearly of the form $0/0$ (but with defined limiting value corresponding to the double-root case). We deal with this by introducing a tolerance

parameter $0 < \mathrm{tol}_0 \ll 1$ and replace evaluation of the exact formula for $u$ by appropriate expansions once the critical expressions fall under tolerable thresholds. Problems of similar kind are well known to arise elsewhere in numerical ODEs, for example in exponential integrators where evaluation of $\phi_1(z) = (e^z - 1)/z$ for $z$ near zero occurs; see [Hochbruck and Ostermann 2010; Kassam and Trefethen 2005].

The general idea of utilizing zeroth- and first-order Taylor approximations in the differential equation is well-established both in theoretical and computational ODEs. In computational ODEs, adaptive first-order Taylor approximation (linearization) is the basis for exponential integrators, classically rooted in the Rosenbrock–Euler method (observe that adaptive zeroth-order Taylor approximation in the differential equation yields the Euler method). Exponential integrators [Hochbruck and Ostermann 2010] have been widely used for stiff problems over the past 30 years; they are generally more effective for these problems than standard Runge–Kutta methods because the linearization of the differential equation is solved exactly. Since the theoretical underpinning for exponential integrators is linear theory, they have been developed into a versatile family of methods applicable to single equations and systems alike. The method we present in this note based on adaptive quadratic Taylor approximation of the differential equation is qualitatively more accurate than methods rooted in linearization, but does not exhibit the same degree of versatility and universality, and the applicability is strictly limited to autonomous scalar equations. The reason is that ordinary differential equations with quadratic nonlinearities generally do not allow for closed solution formulas, the autonomous case of a single unknown function being an exception.

We note that our work relates to nonstandard finite difference models and their applications to numerical ODEs as pioneered by Mickens [1994; 2000]; see also [Patidar 2005]. In particular, exact nonstandard finite difference models for the logistic equation and many other ODEs where explicit solution formulas are available are well known [Vigo-Aguiar and Ramos 2011].

The paper is structured as follows: Section 2 covers the theoretical part. We prove, more generally than what has been stated above, that when the function $f$ in (1.1) is adaptively replaced by its $r$-th order Taylor polynomial and the exact solution to the modified ODE is used to calculate the next approximating value, we obtain a well-defined convergent explicit numerical method of order $r + 1$. More precisely, when the exact solution is supposed to be approximated in the window $[0, T] \times [y_{\min}, y_{\max}]$, we show that there is a threshold $h_0 > 0$ such that the method is defined everywhere in that window for step sizes $0 < h < h_0$ and allows calculation of the next approximating value to the solution. This qualitatively addresses the aforementioned blow-up issue (that is not present for $r = 0$ and $r = 1$ of course). The proofs utilize some results about ODEs depending on parameters

and an abstract theorem about the convergence of one-step methods, stated in the needed forms in Appendices A and B.

Section 3 contains the core of this paper. We discuss the formulas of the method based on quadratic Taylor approximation and their adjustments based on the aforementioned tolerance considerations, the quantitative a priori as well as run time aspects of step size control to address the blow-up issue, and discuss in detail the numerical algorithms. The Matlab code of the programs can be found in the online supplement.

Section 4 contains the results of numerical tests of the method, using Matlab, with benchmarks against some Runge–Kutta methods of orders 3 and 4, respectively. We have tested the quadratic Taylor method on some standard equations from population dynamics, in line with our original motivation, as well as other equations. Our results on the tested equations confirm that the method based on quadratic Taylor expansion can fare better on the global error by several orders of magnitude when compared to the tested Runge–Kutta methods.

As was mentioned before, we only consider equidistant time-stepping in this paper. We also do not utilize any extrapolation techniques to further improve our method. There are certainly several avenues of investigation, in parallel to established ones for standard numerical methods, that could be pursued to augment the method presented in this paper and improve it further. However, the fact that general Riccati equations do not allow for closed solution formulas is going to remain a limiting factor.

## 2. Convergence of explicit methods based on exactly solving Taylor approximations of the differential equation

Let $r \in \mathbb{N}_0$, and let $f : D \to \mathbb{R}$ be $(r+1)$-times continuously differentiable on the open set $D \subset \mathbb{R}$, and suppose $y : [0, T] \to D$ solves the initial value problem

$$\begin{cases} \dot{y}(t) = f(y(t)) & \text{on } 0 \le t \le T, \\ y|_{t=0} = y_0 \in D. \end{cases} \tag{2.1}$$

As mentioned in the Introduction, an idea for an explicit method is to locally replace $f$ by its $r$-th order Taylor polynomial and take the exact solution of the resulting differential equation with the Taylor polynomial instead of $f$ as numerical approximation for $y : [0, T] \to D$ over small time steps. To pursue this idea, define $F : \mathbb{R} \times D \to \mathbb{R}$ via

$$F(w, y) = \sum_{j=0}^{r} \frac{f^{(j)}(y)}{j!} w^j, \tag{2.2}$$

and let $w(h, y)$ for $(h, y) \in U_{\max}$ be the maximally extended solution of

$$\begin{cases} \dfrac{\partial w}{\partial h}(h, y) = F(w(h, y), y), \\ w(0, y) = 0. \end{cases}$$

This differential equation for $w$ depends on $y$ as a parameter, and we have summarized some results about differential equations depending on parameters that we will use below in Appendix B. Since $F$ and all its partial $w$-derivatives are continuously differentiable with respect to $(w, y)$ in $\mathbb{R} \times D$, we obtain that $\partial_h^k w$ is continuously differentiable with respect to $(h, y) \in U_{\max}$ for all $k \in \mathbb{N}_0$. Define $\Phi : U_{\max} \to \mathbb{R}$ via

$$\Phi(h, y) = y + w(h, y). \tag{2.3}$$

Observe that $\Phi$ solves

$$\begin{cases} \dfrac{\partial \Phi}{\partial h}(h, y) = \displaystyle\sum_{j=0}^{r} \frac{f^{(j)}(y)}{j!} (\Phi(h, y) - y)^j, \\ \Phi(0, y) = y. \end{cases}$$

**Proposition 2.4.** *$\Phi$ and all its partial $h$-derivatives are continuously differentiable with respect to $(h, y) \in U_{\max}$. For every compact subset $K \Subset D$ there exists $h_0 > 0$ such that $\Phi : [0, h_0] \times K \to \mathbb{R}$ is defined, and $\partial\Phi/\partial h : [0, h_0] \times K \to \mathbb{R}$ satisfies a Lipschitz condition with respect to $y$ in $[0, h_0] \times K$.*

*Proof.* By Theorem B.2 and Remark B.4, $w$ and all its partial $h$-derivatives exist and are continuously differentiable on $U_{\max}$, and for every $K \Subset D$ there exists $h_0 > 0$ such that $w : [0, h_0] \times K \to \mathbb{R}$ is defined. All this is therefore also true for $\Phi$. Since $\partial^2\Phi/(\partial y\, \partial h) : U_{\max} \to \mathbb{R}$ exists and is continuous, $\partial\Phi/\partial h : [0, h_0] \times K \to \mathbb{R}$ satisfies a Lipschitz condition with respect to $y$ as claimed. □

**Proposition 2.5** (local truncation error). *For any compact neighborhood $K \Subset D$ with $y([0, T]) \subset \mathring{K}$ there exist $h_0 > 0$ such that $\Phi : [0, h_0] \times K \to \mathbb{R}$ is defined, and a constant $C \geq 0$ independent of $0 \leq h \leq h_0$ and $0 \leq t \leq T$ such that*

$$|y(t + h) - \Phi(h, y(t))| \leq Ch^{r+2}$$

*whenever $0 \leq t + h \leq T$.*

*If $f$ is a polynomial of degree $\leq r$ we have $y(t+h) = \Phi(h, y(t))$; i.e., the method is locally exact.*

*Proof.* Recall that if $u$ and $v$ are $n$-times differentiable, $n \in \mathbb{N}$, Faà di Bruno's formula asserts that

$$\frac{d^n}{dh^n}(u \circ v)(h) = \sum_{k=1}^{n} u^{(k)}(v(h)) B_{n,k}(v'(h), v''(h), \ldots, v^{(\mu_k^n)}(h))$$

with the partial Bell polynomials

$$B_{n,k}(x_1,\dots,x_{\mu_k^n}) = \sum_{\substack{\alpha\in\mathbb{N}_0^{\mu_k^n},\,|\alpha|=k\\ 1\alpha_1+2\alpha_2+\cdots+\mu_k^n\alpha_{\mu_k^n}=n}} \frac{n!}{\alpha!}\cdot\Big(\frac{x_1}{1!}\Big)^{\alpha_1}\Big(\frac{x_2}{2!}\Big)^{\alpha_2}\cdots\Big(\frac{x_{\mu_k^n}}{\mu_k^n!}\Big)^{\alpha_{\mu_k^n}},$$

where $\mu_k^n = n-k+1$. We now proceed to use this formula in order to show inductively that

$$\frac{d^n}{dh^n}y(t+h)\Big|_{h=0} = \frac{\partial^n}{\partial h^n}\Phi(h,y(t))\Big|_{h=0} \tag{2.6}$$

for $n=0,\dots,r+1$ (note that $y\in C^{r+2}([0,T])$ since $f\in C^{r+1}(D)$ by assumption). For $n=0$ this follows immediately from the definition in (2.3), keeping in mind that $w(0,y)=0$. For $n=1$ we have

$$\begin{aligned}\frac{d}{dh}y(t+h)\Big|_{h=0} &= f(y(t+h))|_{h=0} = f(y(t)),\\ \frac{\partial}{\partial h}\Phi(h,y(t))\Big|_{h=0} &= F(w(h,y(t)),y(t))|_{h=0} = F(0,y(t)) = f(y(t)).\end{aligned}$$

So suppose we know (2.6) for all $n\le n_0$ for some $1\le n_0\le r$. Now

$$\begin{aligned}\frac{d^{n_0+1}}{dh^{n_0+1}}y(t+h) &= \frac{d^{n_0}}{dh^{n_0}}\Big(\frac{d}{dh}y(t+h)\Big) = \frac{d^{n_0}}{dh^{n_0}}(f\circ y)(t+h)\\ &= \sum_{k=1}^{n_0} f^{(k)}(y(t+h))B_{n_0,k}(y'(t+h),y''(t+h),\dots,y^{(\mu_k^{n_0})}(t+h)).\end{aligned}$$

Evaluation at $h=0$ gives

$$\frac{d^{n_0+1}}{dh^{n_0+1}}y(t+h)\Big|_{h=0} = \sum_{k=1}^{n_0} f^{(k)}(y(t))B_{n_0,k}(y'(t),y''(t),\dots,y^{(\mu_k^{n_0})}(t)). \tag{2.7}$$

Using the differential equation for $w(h,y)$ and (2.3) we get

$$\frac{\partial^{n_0+1}}{\partial h^{n_0+1}}\Phi(h,y(t))\Big|_{h=0} = \frac{\partial^{n_0}}{\partial h^{n_0}}F(w(h,y(t)),y(t))\Big|_{h=0},$$

which by Faà di Bruno's formula equals

$$\sum_{k=1}^{n_0}(\partial_w^k F)(w(0,y(t)),y(t))\,B_{n_0,k}\big((\partial_h w)(0,y(t)),\dots,(\partial_h^{\mu_k^{n_0}}w)(0,y(t))\big). \tag{2.8}$$

By induction, $(\partial_h^j w)(0,y(t)) = y^{(j)}(t)$ for $j=1,\dots,\mu_k^{n_0}$, and thus the arguments in the partial Bell polynomials $B_{n_0,k}$ in (2.7) and (2.8) agree. Moreover,

$$(\partial_w^k F)(w(0,y(t)),y(t)) = (\partial_w^k F)(0,y(t)) = f^{(k)}(y(t))$$

for $k=0,\dots,r$ in view of (2.2) and Taylor's formula. This shows that (2.6) holds for $n=n_0+1$ and finishes the induction.

In view of (2.6), Taylor's formula now implies

$$|y(t+h)-\Phi(h,y(t))| = \left| \int_0^h \frac{y^{(r+2)}(t+s)-(\partial_h^{r+2}\Phi)(s,y(t))}{(r+1)!}(h-s)^{r+1}\,ds \right|$$
$$\le \underbrace{\left[ \frac{1}{(r+2)!}\Big( \max_{0\le s\le T} |y^{(r+2)}(s)| + \max_{\substack{0\le s\le h_0 \\ y\in K}} |(\partial_h^{r+2}\Phi)(s,y)| \Big) \right]}_{=:C} \cdot h^{r+2}$$

for all $0 \le t \le T$ and $0 \le h \le h_0$ such that $t+h \le T$.

If $f$ is a polynomial of degree $\le r$ we have

$$f(z) = \sum_{j=0}^{r} \frac{f^{(j)}(y)}{j!}(z-y)^j$$

for all $z, y \in \mathbb{R}$, and consequently both $h \mapsto y(t+h)$ and $h \mapsto \Phi(h, y(t))$ solve

$$\begin{cases} \dot{u}(h) = f(u(h)), & h \ge 0, \\ u|_{h=0} = y(t). \end{cases}$$

By uniqueness we must therefore have $y(t+h) = \Phi(h, y(t))$. □

**Theorem 2.9.** *The method $\Phi$ defined in (2.3) is a convergent method of order $r+1$ for the approximation of the solution $y : [0, T] \to \mathbb{R}$ of (2.1). The method is exact for differential equations (2.1) when $f$ is a polynomial of degree at most $r$.*

*Proof.* This follows with Propositions 2.4 and 2.5 from Theorem A.1. □

**Example 2.10.** If we specialize to $r = 0$ and $r = 1$ in (2.2) we find familiar methods.

- If $r = 0$ we have $F(w, h) = f(y)$ in (2.2), and so $\Phi(h, y)$ solves

$$\begin{cases} \dfrac{\partial \Phi}{\partial h}(h, y) = f(y), \\ \Phi(0, y) = y. \end{cases}$$

Thus $\Phi(h, y) = y + hf(y)$ is the Euler method.

- If $r = 1$ we have $F(w, y) = f(y) + f'(y)w$ in (2.2), and so $\Phi(h, y)$ solves

$$\begin{cases} \dfrac{\partial \Phi}{\partial h}(h, y) = f(y) + f'(y)(\Phi(h, y) - y), \\ \Phi(0, y) = y. \end{cases}$$

Thus $\Phi(h, y) = y + h\phi_1(f'(y)h)f(y)$ with

$$\phi_1(z) = \begin{cases} (e^z - 1)/z & \text{for } z \ne 0, \\ 1 & \text{for } z = 0 \end{cases}$$

is the Rosenbrock–Euler method [Hochbruck and Ostermann 2010, Section 2.4].

In this paper, we present and analyze the method based on adaptive Taylor approximation in detail for $r = 2$. While the theoretical result in Theorem 2.9 holds for all $r \in \mathbb{N}_0$, it is not feasible for the implementation of methods for larger $r$ as one generally does not have explicit solution formulas for polynomial ordinary differential equations.

## 3. Third-order scheme based on quadratic Taylor approximation

We begin by defining and analyzing the analytic function

$$\psi(u, v) = \frac{\sinh(v)}{u \sinh(v) + v \cosh(v)} \tag{3.1}$$

depending on two complex variables $(u, v) \in \mathbb{C}^2$. Initially, this function is undefined on

$$S = \{(u, v) \in \mathbb{C}^2 : u \sinh(v) + v \cosh(v) = 0\}.$$

It is easy to see that $S$ consists of the complex line $v = 0$ and the complex surface

$$S_\psi : u = -v \coth(v).$$

The singularities of $\psi$ where $v = 0$ are removable, except for the singularity at the single branch point $(-1, 0)$. To see this note that for $v$ near 0 we can write

$$\psi(u, v) = \frac{1}{u + v \coth(v)}.$$

The function $v \mapsto v \coth(v)$ has a removable singularity at $v = 0$. The first few terms of the Taylor series are

$$v \coth(v) = 1 + \frac{v^2}{3} - \frac{v^4}{45} + O(v^6),$$

which shows that $\psi$ has a removable singularity at all points $(u, 0)$ if $u \neq -1$. More precisely, we get

$$\psi(u, v) = \frac{1}{1+u} - \frac{v^2}{3(1+u)^2} + \frac{(u+6)v^4}{45(1+u)^3} + O(v^6) \tag{3.2}$$

locally uniformly in $u$, and, in particular, we see that the definition

$$\psi(u, 0) = \frac{1}{1+u}, \quad u \neq -1, \tag{3.3}$$

extends $\psi$ analytically to $v = 0$ except the branch point. From (3.1) and (3.3) we obtain that $\psi$ remains singular on $S_\psi : u = -v \coth(v)$, but now with the understanding that the singularity of $v \coth(v)$ when $v = 0$ has been removed. We are going to need the function $\psi$ only in the cases that both $u$ and $v$ are real, or that

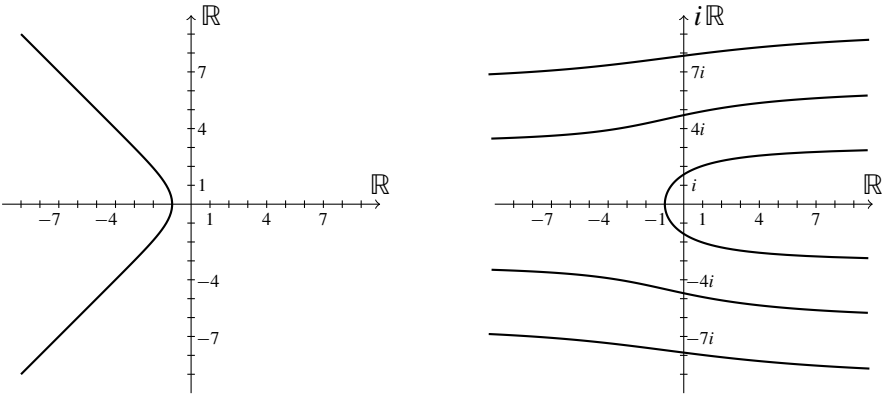



**Figure 1.** Singularities of $\psi$: $S_\psi \cap \mathbb{R}^2$ (left) and $S_\psi \cap (\mathbb{R} \times i\mathbb{R})$ (right).

$u$ is real and $v$ is imaginary. Figure 1 shows parts of the intersection of the singular set $S_\psi$ with $\mathbb{R}^2$ and with $\mathbb{R} \times i\mathbb{R}$, respectively.

The relevance of the function $\psi$ for us is clarified by the following lemma. The proof is straightforward and will be omitted.

**Lemma 3.4.** *Consider the initial value problem for the Riccati ordinary differential equation*

$$\begin{cases} \dot{w} = aw^2 + bw + c, \\ w|_{t=0} = 0, \end{cases} \tag{3.5}$$

*with constant coefficients $a, b, c \in \mathbb{R}$. Let $\Delta = b^2 - 4ac$, and define $\alpha = -\frac{1}{2}b \in \mathbb{R}$, $\beta = \frac{1}{2}\sqrt{\Delta} \in \mathbb{C}$. Note that $\beta \in \mathbb{R}$ if $\Delta \geq 0$, and in the case $\Delta < 0$ we choose $\sqrt{\Delta}$ to be the root with positive imaginary part,*[1] *so $\beta \in i\mathbb{R}_+$.*

*The maximal solution to (3.5) is given by*

$$w(t) \equiv w(t; a, b, c) = ct\psi(t\alpha, t\beta), \quad t_{\min} < t < t_{\max},$$

*where*

$$t_{\min} = \sup\{t < 0 : (t\alpha, t\beta) \in S_\psi\} \in \mathbb{R}_- \cup \{-\infty\},$$

$$t_{\max} = \inf\{t > 0 : (t\alpha, t\beta) \in S_\psi\} \in \mathbb{R}_+ \cup \{\infty\}.$$

It is important to note that the solution $w(t)$ to (3.5) can blow up in finite time depending on the values of the constants $a, b, c$. This has a serious impact on the method presented here in that additional integrity checks on the step size must be performed (a priori or at run time) that do not appear in Runge–Kutta methods or exponential integrators. The way the solution $w(t)$ is represented in Lemma 3.4 utilizing the function $\psi$ facilitates a simple visualization of the existence interval. As

[1] We could choose either, really, since $\psi(u, v)$ is even in $v$.

described in the lemma, the coefficients $a, b, c \in \mathbb{R}$ determine a point $(\alpha, \beta) \in \mathbb{R}^2$, or $(\alpha, \beta) \in \mathbb{R} \times i\mathbb{R}$, respectively, and the solution involves evaluation of the function $\psi$ restricted to the line passing through the origin in $\mathbb{R}^2$ (or $\mathbb{R} \times i\mathbb{R}$) and that point, parametrized by $t$. Then $t_{\max}$ is precisely the first positive $t$-value when that line crosses the singular set $S_\psi$ of the function $\psi$ (and $t_{\max} = \infty$ if there is no such crossing point), and similarly for $t_{\min}$. It is easy to visualize this behavior in Figure 1. We summarize, focusing on $t_{\max}$:

- $\Delta \geq 0$, so $(\alpha, \beta) \in \mathbb{R}^2$: $t_{\max} < \infty$ precisely when $\alpha < 0$ and $|\beta| < |\alpha|$. These conditions are equivalent to $b > 0$ and $0 < ac \leq \frac{1}{4}b^2$. By the definition of $S_\psi$, $t_{\max}$ is then the (unique) solution to

$$\alpha t_{\max} = -\beta t_{\max} \coth(\beta t_{\max})$$

(recall that $v \coth(v) = 1$ when $v = 0$). Consequently,

$$t_{\max} = \begin{cases} -\dfrac{1}{\alpha} = \dfrac{2}{b} & \text{if } \Delta = 0, \\ \dfrac{1}{\beta} \operatorname{arcoth}\left(-\dfrac{\alpha}{\beta}\right) = \dfrac{1}{\sqrt{\Delta}} \ln\left(\dfrac{b+\sqrt{\Delta}}{b-\sqrt{\Delta}}\right) & \text{if } \Delta > 0. \end{cases}$$

- If $\Delta < 0$ we have $t_{\max} < \infty$, and

$$\alpha t_{\max} = -\beta t_{\max} \coth(\beta t_{\max}) = -(-i\beta) t_{\max} \cot(-i\beta t_{\max}).$$

We get[2]

$$t_{\max} = \frac{1}{(-i\beta)} \operatorname{arccot}\left(-\frac{\alpha}{(-i\beta)}\right) = \frac{2}{\sqrt{-\Delta}} \operatorname{arccot}\left(\frac{b}{\sqrt{-\Delta}}\right).$$

In summary,

$$t_{\max} = \begin{cases} \dfrac{2}{b} & \text{if } \Delta = 0,\ b > 0, \\ \dfrac{1}{\sqrt{\Delta}} \ln\left(\dfrac{b+\sqrt{\Delta}}{b-\sqrt{\Delta}}\right) & \text{if } \Delta > 0,\ \sqrt{\Delta} < b, \\ \dfrac{2}{\sqrt{-\Delta}} \operatorname{arccot}\left(\dfrac{b}{\sqrt{-\Delta}}\right) & \text{if } \Delta < 0, \\ \infty & \text{otherwise.} \end{cases}$$

In either case, since the closest point of $S_\psi \cap \mathbb{R}^2$, or $S_\psi \cap (\mathbb{R} \times i\mathbb{R})$, respectively, to the origin with respect to the Euclidean distance $|(\cdot, \cdot)|$ is the point $(-1, 0)$, we note that the solution is guaranteed to exist while $|(t\alpha, t\beta)| < 1$. This gives a rough estimate,

$$t_{\max} \geq \frac{1}{|(\alpha, \beta)|} = \frac{2}{\sqrt{b^2 + |\Delta|}}, \tag{3.6}$$

[2]In the formula for $t_{\max}$ and elsewhere we use the real arccot : $\mathbb{R} \to (0, \pi)$.

where we understand the right-hand side of (3.6) to be $\infty$ if $b = \Delta = 0$. This estimate can be used to derive an a priori estimate on valid step sizes of the method, as described below.

***Description of the method.*** The goal is to approximate the solution $y = y(t)$ to the initial value problem

$$\begin{cases} \dot{y} = f(y), \\ y|_{t=0} = y_0, \end{cases}$$

for $(t, y) \in [0, T] \times [A, B]$. We assume that $f$ is $C^3$ in an open neighborhood of $[A, B]$, and $y_0 \in [A, B]$. To fix notation, define functions $a, b, c, \Delta, h_{\max} : [A, B] \to \mathbb{R} \cup \{\infty\}$ via

$$a(y) = \frac{f''(y)}{2}, \quad b(y) = f'(y), \quad c(y) = f(y), \quad \Delta = b^2 - 4ac,$$

$$h_{\max}(y) = \begin{cases} \dfrac{2}{b(y)} & \text{if } \Delta(y) = 0,\ b(y) > 0, \\ \dfrac{1}{\sqrt{\Delta(y)}} \ln\left(\dfrac{b(y) + \sqrt{\Delta(y)}}{b(y) - \sqrt{\Delta(y)}}\right) & \text{if } \Delta(y) > 0,\ \sqrt{\Delta(y)} < b(y), \\ \dfrac{2}{\sqrt{-\Delta(y)}} \operatorname{arccot}\left(\dfrac{b(y)}{\sqrt{-\Delta(y)}}\right) & \text{if } \Delta(y) < 0, \\ \infty & \text{otherwise.} \end{cases}$$

Choose a tolerance $0 < \mathrm{tol}_0 \ll 1$. Quantities that in absolute value are less than $\mathrm{tol}_0$ are considered numerically zero. Evaluation of approximate $0/0$ expressions with denominators of magnitude $< \mathrm{tol}_0$, such as occur in the evaluation of $\psi(u, v)$ given by (3.1) for $v$ near zero, should be avoided to improve stability. For this reason, we define

$$\Phi(h, y) = \begin{cases} y + \dfrac{2c(y)\sinh\left[\frac{1}{2}\sqrt{\Delta(y)}h\right]}{\sqrt{\Delta(y)}\cosh\left[\frac{1}{2}\sqrt{\Delta(y)}h\right] - b(y)\sinh\left[\frac{1}{2}\sqrt{\Delta(y)}h\right]} & \text{for } (h, y) \in U_+, \\ y + \dfrac{2c(y)\sin\left[\frac{1}{2}\sqrt{-\Delta(y)}h\right]}{\sqrt{-\Delta(y)}\cos\left[\frac{1}{2}\sqrt{-\Delta(y)}h\right] - b(y)\sin\left[\frac{1}{2}\sqrt{-\Delta(y)}h\right]} & \text{for } (h, y) \in U_-, \\ y + \dfrac{2c(y)h}{2 - b(y)h} - \dfrac{h^3 c(y)\Delta(y)}{3(2 - b(y)h)^2} & \text{for } (h, y) \in U_0, \end{cases} \tag{3.7}$$

where

$$U_+ = \{(h, y) \in [0, \infty) \times [A, B] : \Delta(y) \geq 4\mathrm{tol}_0,\ h < h_{\max}(y),\ 2 - hb(y) \geq \sqrt{\mathrm{tol}_0}\},$$

$$U_- = \{(h, y) \in [0, \infty) \times [A, B] : \Delta(y) \leq -4\mathrm{tol}_0,\ h < h_{\max}(y),\ 2 - hb(y) \geq \sqrt{\mathrm{tol}_0}\},$$

$$U_0 = \{(h, y) \in [0, \infty) \times [A, B] : |\Delta(y)| < 4\mathrm{tol}_0,\ 2 - hb(y) \geq \sqrt{\mathrm{tol}_0}\}.$$

Some comments are in order:

• By Lemma 3.4 and formula (3.1), the first two cases in (3.7) are the exact formulas of the general method (2.3) discussed in Section 2 with $r = 2$. In the second case we merely converted to trigonometric functions in the formulas since the argument of the hyperbolic trigonometric functions would be imaginary.

• Taylor expansion of the hyperbolic trigonometric functions in the denominator in the first case gives

$$\begin{aligned}
&\sqrt{\Delta(y)}\cosh\left[\tfrac{1}{2}\sqrt{\Delta(y)}h\right] - b(y)\sinh\left[\tfrac{1}{2}\sqrt{\Delta(y)}h\right] \\
&\qquad = \sqrt{\Delta(y)}\sum_{k=0}^{\infty}\frac{\left[\tfrac{1}{2}\sqrt{\Delta(y)}h\right]^{2k}}{(2k)!} - b(y)\sum_{k=0}^{\infty}\frac{\left[\tfrac{1}{2}\sqrt{\Delta(y)}h\right]^{2k+1}}{(2k+1)!} \\
&\qquad = \frac{\sqrt{\Delta(y)}}{2}\sum_{k=0}^{\infty}\frac{1}{(2k)!}\left[2 - \frac{hb(y)}{2k+1}\right]\left[\tfrac{1}{2}\sqrt{\Delta(y)}h\right]^{2k} \\
&\qquad \geq \mathrm{tol}_0 \cdot \cosh\left[\tfrac{1}{2}\sqrt{\Delta(y)}h\right] \geq \mathrm{tol}_0
\end{aligned}$$

under the assumption that both $\frac{1}{2}\sqrt{\Delta(y)} \geq \sqrt{\mathrm{tol}_0}$ and $2 - hb(y) \geq \sqrt{\mathrm{tol}_0}$, which explains the definition of $U_+$.

• In the second case, we note that when $b(y) > 0$ we have $h_{\max}(y) < 2/b(y)$ directly from the definition when $\Delta(y) < 0$, and consequently $2 - hb(y) > 0$ is implied by $h < h_{\max}(y)$ (the inequality is trivially fulfilled for $b(y) \leq 0$). The condition $2 - hb(y) \geq \sqrt{\mathrm{tol}_0}$ gives an extra buffer. We also note, arguing analogous to the first case, that

$$\begin{aligned}
&\sqrt{-\Delta(y)}\cos\left[\tfrac{1}{2}\sqrt{-\Delta(y)}h\right] - b(y)\sin\left[\tfrac{1}{2}\sqrt{-\Delta(y)}h\right] \\
&\qquad = \tfrac{1}{2}\sqrt{-\Delta(y)}\sum_{k=0}^{\infty}\frac{(-1)^k}{(2k)!}\left[2 - \frac{hb(y)}{2k+1}\right]\left[\tfrac{1}{2}\sqrt{-\Delta(y)}h\right]^{2k} \\
&\qquad = \tfrac{1}{2}\sqrt{-\Delta(y)}\left[2 - hb(y) + O\left[\left[\tfrac{1}{2}\sqrt{-\Delta(y)}h\right]^2\right]\right],
\end{aligned}$$

and thus the denominator is asymptotically $\geq \mathrm{tol}_0$ under the restrictions placed on $U_-$.

• If the discriminant term $\Delta(y)$ is too small, evaluation of $\psi(u, v)$ as given by (3.1) is unstable, so we opt to use the expansion (3.2) instead for such terms, leading to the definition of $\Phi(h, y)$ in the third case. By Lemma 3.4 and expansion (3.2), we note that the theoretical method as determined by (2.3) and our definition for $\Phi(h, y)$ in the third case of (3.7) coincide to third order in $h$ as $h \to 0$, showing that the local truncation error in our definition is still $O(h^4)$ as required. Moreover, our definition for $\Phi(h, y)$ in the third case matches the general method from (2.3) if $\Delta(y) = 0$.

*The algorithm.* Besides the differential equation $\dot{y} = f(y)$ and the initial value $y_0$, the inputs are $0 < \mathrm{tol}_0 \ll 1$, the window $[0, T] \times [A, B]$ where the solution $y(t)$ is supposed to be approximated, and the chosen step size $h > 0$ for constructing an approximating sequence $y_0, y_1, \ldots$ of values for the solution at equidistant points $t = jh$, $j = 0, 1, \ldots$.

(1) Check whether $y_0 \in [A, B]$. If not, the algorithm terminates with an error message that the initial value lies outside of the chosen tracking window.

Now suppose that an approximating partial sequence $y_0, \ldots, y_n$ for some $n \in \mathbb{N}_0$ has already been successfully constructed.

(2) If $(n + 1)h > T$, the algorithm terminates with success and displays the approximation $y_0, \ldots, y_n$ of the solution.

(3) *Integrity check on the step size*: Check whether $(h, y_n) \in U_+ \cup U_- \cup U_0$. If not, the program terminates with the message that the algorithm stops after $n$ steps, approximating the solution on $[0, nh]$, as the method becomes undefined in the next step due to the chosen step size. The approximation of the solution thus far is displayed, and it is suggested to run the program again with a smaller step size $h > 0$.

(4) Check whether $\Phi(h, y_n) \in [A, B]$. If not, the program is terminated with the message that the algorithm stops after $n$ steps, approximating the solution on $[0, nh]$, as the approximate solution is leaving the designated tracking window in the next step. The approximation of the solution thus far is displayed.

(5) If the program reaches this step, it accepts $y_{n+1} = \Phi(h, y_n)$ as the next value of the approximating sequence, and recursively resumes at step (2) with $n$ incremented by one.

Instead of performing the integrity check on the step size in (3) at run time during every execution of the recursive loop, an a priori estimate can be obtained prior to building the approximating sequence to determine a value $h_0 > 0$ that only depends on $f$, $\mathrm{tol}_0$, and the chosen viewing window such that all step sizes $0 < h < h_0$ work. Following this procedure and skipping the integrity checks at run time increases the speed of the program. The a priori estimate utilizes (3.6), as follows:

(i) Find the maximum value $b_{\max}$ of $b : [A, B] \to \mathbb{R}$.

(ii) Find the maximum value $s_{\max}$ of $s := b^2 + |\Delta| : [A, B] \to \mathbb{R}$.

(iii) Set

$$h_0 = \begin{cases} \min\{2/\sqrt{s_{\max}}, (2 - \sqrt{\mathrm{tol}_0})/b_{\max}, T\} & \text{if } s_{\max} > \mathrm{tol}_0,\ b_{\max} > \mathrm{tol}_0, \\ \min\{2/\sqrt{s_{\max}}, T\} & \text{if } s_{\max} > \mathrm{tol}_0,\ b_{\max} \le \mathrm{tol}_0, \\ T & \text{otherwise.} \end{cases}$$

## 4. Numerical tests of the quadratic Taylor method

In all tests described below we used the tolerance $\mathrm{tol}_0 = 1 \cdot 10^{-14}$ and recorded the global error and the run time of the method on the indicated interval for the problem with various step sizes $h$. Errors in magnitude less than $\mathrm{tol}_0$ have been recorded as zero. All tests were performed using Matlab (Version 2018a, Update 4) on a ThinkPad laptop computer equipped with an Intel i7-7500U CPU @ 2.70GHz and 16.0 GB RAM, running Windows 10. We are benchmarking our third-order method, labeled QT3 below, against the following standard methods from the Runge–Kutta family:

• Kutta third-order method (K3), see [Butcher 2008, Section 233]: the Butcher tableau for this method is

$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{2} & \frac{1}{2} & & \\ 1 & -1 & 2 & \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array}$$

• Bogacki–Shampine third-order method (BS3) [1989]: the Butcher tableau for this method is

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{3}{4} & 0 & \frac{3}{4} & & \\ 1 & \frac{2}{9} & \frac{1}{3} & \frac{4}{9} & \\ \hline & \frac{2}{9} & \frac{1}{3} & \frac{4}{9} & 0 \end{array}$$

Embedded in a 3(2) pair, this method is built into one of the standard algorithms, `ode23`, of the Matlab suite [Shampine and Reichelt 1997]. In [Bogacki and Shampine 1989] the third-order formulas are credited to [Ralston 1965].

• Classical Runge–Kutta fourth-order method (RK4), see [Hairer et al. 1993, Section II.1]: the Butcher tableau for this method is

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

***Logistic equation.*** As expected, the quadratic Taylor method outperforms standard Runge–Kutta methods for quadratic ordinary differential equations with regards to

| $h$ | K3 | BS3 | RK4 | QT3 |
|---|---|---|---|---|
| 0.1 | $9.0574 \cdot 10^{-2}$ | $4.9747 \cdot 10^{-2}$ | $1.3532 \cdot 10^{-2}$ | 0 |
| 0.05 | $1.3495 \cdot 10^{-2}$ | $8.2625 \cdot 10^{-3}$ | $1.0941 \cdot 10^{-3}$ | 0 |
| 0.02 | $9.6842 \cdot 10^{-4}$ | $6.3000 \cdot 10^{-4}$ | $3.3012 \cdot 10^{-5}$ | 0 |
| 0.01 | $1.2579 \cdot 10^{-4}$ | $8.3520 \cdot 10^{-5}$ | $2.1834 \cdot 10^{-6}$ | 0 |

**Table 1.** Global error vs. step size for (4.1).

| $h$ | K3 | BS3 | RK4 | QT3 |
|---|---|---|---|---|
| 0.1 | 0.396497 | 0.507937 | 0.498741 | 0.823234 |
| 0.05 | 0.652171 | 0.860344 | 0.896938 | 1.625599 |
| 0.02 | 1.681197 | 2.138214 | 2.149054 | 3.842002 |
| 0.01 | 3.413213 | 4.193556 | 4.182386 | 7.75049 |

**Table 2.** Run time (in seconds) vs. step size for (4.1).

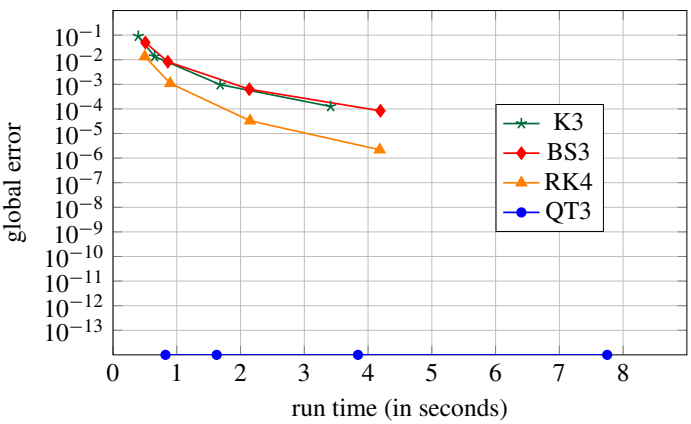



**Figure 2.** Work-precision diagram for (4.1) from Tables 1 and 2.

accuracy. Consider

$$\begin{cases} \dot{y} = y(10 - y), \\ y|_{t=0} = 0.5, \end{cases} \qquad 0 \le t \le 2. \tag{4.1}$$

The exact solution is

$$y(t) = \frac{10e^{10t}}{19 + e^{10t}}, \quad 0 \le t \le 2.$$

The results are displayed in Tables 1 and 2, and in Figures 2 and 3.

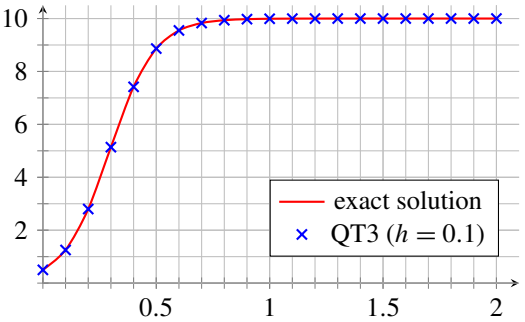



**Figure 3.** Exact solution and QT3 values for (4.1).

| $h$ | K3 | BS3 | RK4 | QT3 |
|---|---|---|---|---|
| 0.1 | $2.8543 \cdot 10^{-6}$ | $2.8543 \cdot 10^{-6}$ | $5.6900 \cdot 10^{-8}$ | $9.6127 \cdot 10^{-13}$ |
| 0.05 | $3.7135 \cdot 10^{-7}$ | $3.7135 \cdot 10^{-7}$ | $3.7073 \cdot 10^{-9}$ | $1.2390 \cdot 10^{-13}$ |
| 0.02 | $2.4343 \cdot 10^{-8}$ | $2.4343 \cdot 10^{-8}$ | $9.7307 \cdot 10^{-11}$ | 0 |
| 0.01 | $3.0673 \cdot 10^{-9}$ | $3.0673 \cdot 10^{-9}$ | $6.1326 \cdot 10^{-12}$ | 0 |

**Table 3.** Global error vs. step size for (4.2).

| $h$ | K3 | BS3 | RK4 | QT3 |
|---|---|---|---|---|
| 0.1 | 0.862458 | 1.125509 | 1.131106 | 2.028254 |
| 0.05 | 1.622196 | 2.122481 | 2.171911 | 3.874084 |
| 0.02 | 4.218344 | 5.616584 | 5.505497 | 10.525293 |
| 0.01 | 8.255611 | 11.230867 | 11.15308 | 20.897061 |

**Table 4.** Run time (in seconds) vs. step size for (4.2).

***Bernoulli equation.*** Consider

$$\begin{cases} \dot{y} = y\big(1 - \big(\frac{1}{20}y\big)^2\big), \\ y|_{t=0} = 1 \cdot 10^{-4}, \end{cases} \qquad 0 \le t \le 5. \tag{4.2}$$

The exact solution is

$$y(t) = \frac{20}{\sqrt{(4 \cdot 10^{10} - 1)e^{-2t} + 1}}, \quad 0 \le t \le 5.$$

The results are displayed in Tables 3 and 4, and in Figures 4 and 5.

Let's also consider the same differential equation

$$\begin{cases} \dot{y} = y\big(1 - \big(\frac{1}{20}y\big)^2\big), \\ y|_{t=0} = 1, \end{cases} \qquad 0 \le t \le 5, \tag{4.3}$$

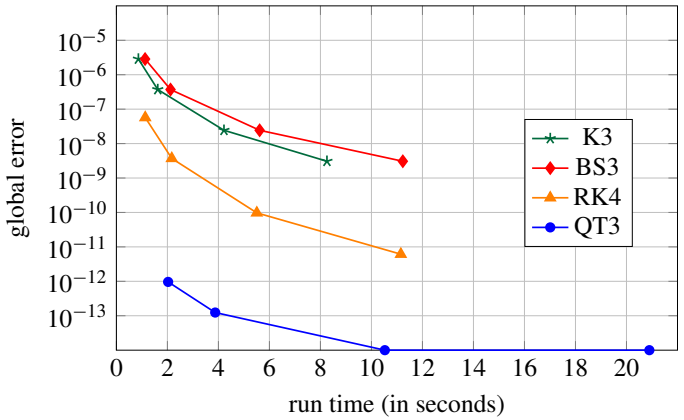



**Figure 4.** Work-precision diagram for (4.2) from Tables 3 and 4.

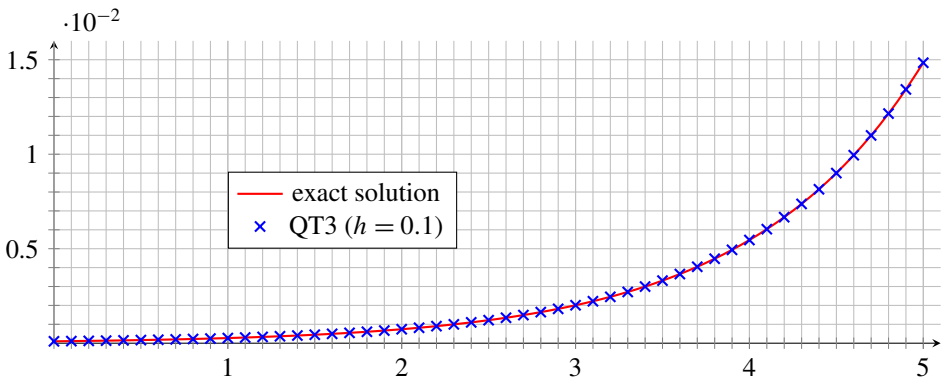



**Figure 5.** Exact solution and QT3 values for (4.2).

but with a different initial value that is farther away from the equilibrium solutions. The exact solution is then

$$y(t) = \frac{20}{\sqrt{399e^{-2t} + 1}}, \quad 0 \le t \le 5.$$

The results are displayed in Tables 5 and 6, and in Figures 6 and 7.

| $h$ | K3 | BS3 | RK4 | QT3 |
|---|---|---|---|---|
| 0.1 | $6.3817 \cdot 10^{-4}$ | $4.5295 \cdot 10^{-4}$ | $1.5055 \cdot 10^{-5}$ | $3.2525 \cdot 10^{-4}$ |
| 0.05 | $8.1554 \cdot 10^{-5}$ | $5.8683 \cdot 10^{-5}$ | $9.2633 \cdot 10^{-7}$ | $4.1018 \cdot 10^{-5}$ |
| 0.02 | $5.2845 \cdot 10^{-6}$ | $3.8374 \cdot 10^{-6}$ | $2.3554 \cdot 10^{-8}$ | $2.6396 \cdot 10^{-6}$ |
| 0.01 | $6.6341 \cdot 10^{-7}$ | $4.8314 \cdot 10^{-7}$ | $1.4695 \cdot 10^{-9}$ | $3.3052 \cdot 10^{-7}$ |

**Table 5.** Global error vs. step size for (4.3).

| $h$ | K3 | BS3 | RK4 | QT3 |
|---|---|---|---|---|
| 0.1 | 0.871428 | 1.135167 | 1.204247 | 1.991003 |
| 0.05 | 1.705217 | 2.190155 | 2.190203 | 4.047121 |
| 0.02 | 4.002053 | 5.512081 | 5.56601 | 9.936015 |
| 0.01 | 8.80262 | 11.240246 | 13.179334 | 18.890316 |

**Table 6.** Run time (in seconds) vs. step size for (4.3).

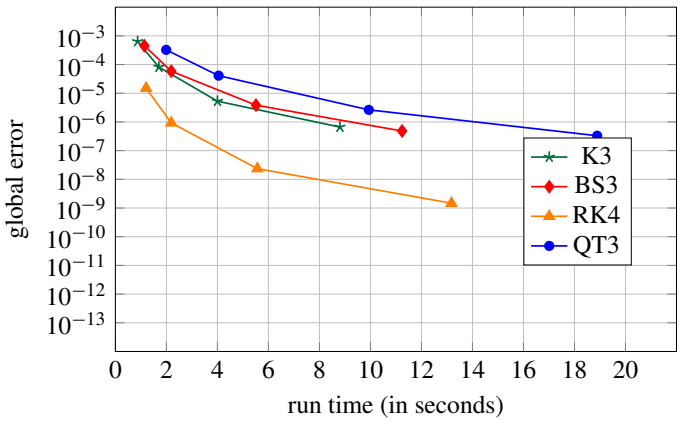



**Figure 6.** Work-precision diagram for (4.3) from Tables 5 and 6.

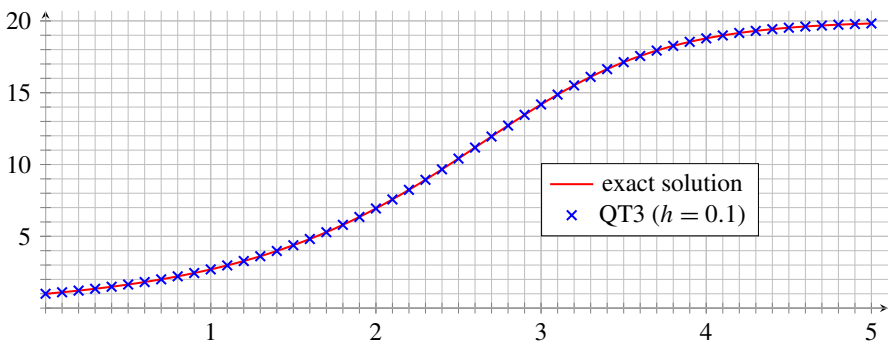



**Figure 7.** Exact solution and QT3 values for (4.3).

***Gompertz equation.*** Consider

$$\begin{cases} \dot{y} = y \ln(30/y), \\ y|_{t=0} = 29, \end{cases} \qquad 0 \le t \le 2. \tag{4.4}$$

The exact solution is

$$y(t) = 30\left(\tfrac{29}{30}\right)^{e^{-t}}, \quad 0 \le t \le 2.$$

The results are displayed in Tables 7 and 8, and in Figures 8 and 9.

| $h$ | K3 | BS3 | RK4 | QT3 |
|---|---|---|---|---|
| 0.1 | $1.5931 \cdot 10^{-5}$ | $1.5604 \cdot 10^{-5}$ | $3.1690 \cdot 10^{-7}$ | $9.7263 \cdot 10^{-9}$ |
| 0.05 | $1.9169 \cdot 10^{-6}$ | $1.8770 \cdot 10^{-6}$ | $1.9019 \cdot 10^{-8}$ | $1.1837 \cdot 10^{-9}$ |
| 0.02 | $1.1990 \cdot 10^{-7}$ | $1.1734 \cdot 10^{-7}$ | $4.7509 \cdot 10^{-10}$ | $7.4419 \cdot 10^{-11}$ |
| 0.01 | $1.4873 \cdot 10^{-8}$ | $1.4554 \cdot 10^{-8}$ | $2.9431 \cdot 10^{-11}$ | $9.2619 \cdot 10^{-12}$ |

**Table 7.** Global error vs. step size for (4.4).

| $h$ | K3 | BS3 | RK4 | QT3 |
|---|---|---|---|---|
| 0.1 | 0.389789 | 0.499958 | 0.483573 | 0.801802 |
| 0.05 | 0.678447 | 0.890931 | 0.892354 | 1.570274 |
| 0.02 | 1.572932 | 2.228926 | 2.24712 | 4.003964 |
| 0.01 | 3.329032 | 4.399356 | 4.406119 | 8.440522 |

**Table 8.** Run time (in seconds) vs. step size for (4.4).

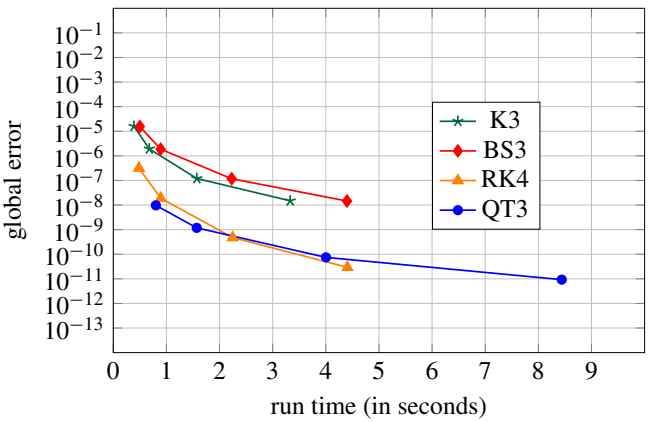



**Figure 8.** Work-precision diagram for (4.4) from Tables 7 and 8.

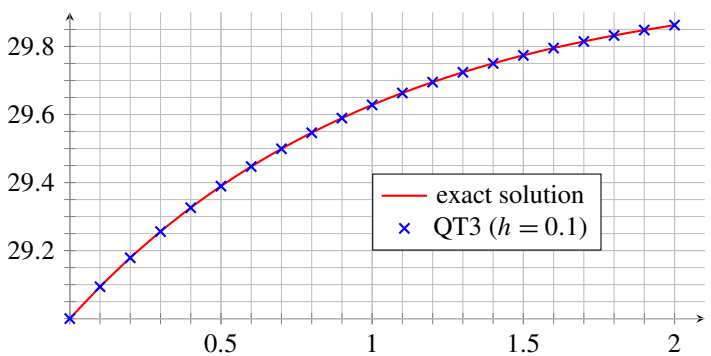



**Figure 9.** Exact solution and QT3 values for (4.4).

| $h$ | K3 | BS3 | RK4 | QT3 |
|---|---|---|---|---|
| 0.1 | $3.0134 \cdot 10^{-7}$ | $2.8743 \cdot 10^{-7}$ | $5.9219 \cdot 10^{-9}$ | $3.8462 \cdot 10^{-10}$ |
| 0.05 | $3.6318 \cdot 10^{-8}$ | $3.4589 \cdot 10^{-8}$ | $3.5555 \cdot 10^{-10}$ | $4.6768 \cdot 10^{-11}$ |
| 0.02 | $2.2745 \cdot 10^{-9}$ | $2.1638 \cdot 10^{-9}$ | $8.8861 \cdot 10^{-12}$ | $2.9453 \cdot 10^{-12}$ |
| 0.01 | $2.8224 \cdot 10^{-10}$ | $2.6843 \cdot 10^{-10}$ | $5.5067 \cdot 10^{-13}$ | $3.6637 \cdot 10^{-13}$ |

**Table 9.** Global error vs. step size for (4.5).

| $h$ | K3 | BS3 | RK4 | QT3 |
|---|---|---|---|---|
| 0.1 | 1.635349 | 2.132353 | 2.109736 | 3.99236 |
| 0.05 | 3.471573 | 4.173892 | 4.026089 | 8.406138 |
| 0.02 | 8.569636 | 10.934895 | 10.836708 | 19.228195 |
| 0.01 | 16.240501 | 20.771019 | 20.858536 | 40.610184 |

**Table 10.** Run time (in seconds) vs. step size for (4.5).

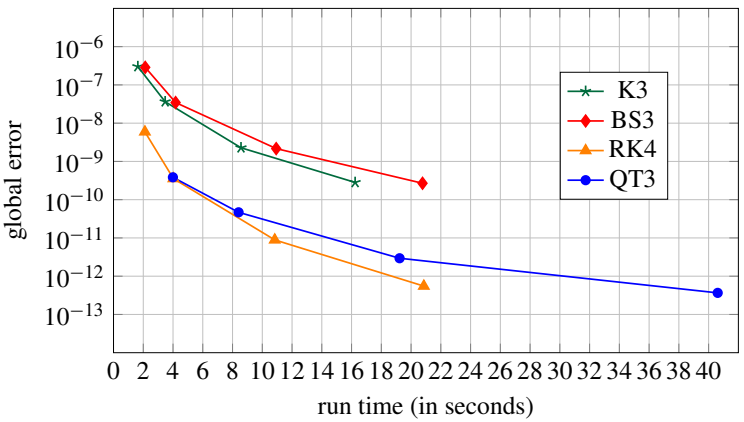



**Figure 10.** Work-precision diagram for (4.5) from Tables 9 and 10.

***Flame propagation.*** The following example is taken from a *Cleve's Corner* blog post on the MathWorks web page; see [Moler 2003]. It is attributed there to L. Shampine. Consider

$$\begin{cases} \dot{y} = y^2 - y^3, \\ y|_{t=0} = 0.98, \end{cases} \qquad 0 \le t \le 10. \tag{4.5}$$

The exact solution is

$$y(t) = \frac{1}{1 + W\left(\frac{1}{49} e^{1/49 - t}\right)}, \quad 0 \le t \le 10,$$

where $W$ is the Lambert W function; see [Corless et al. 1996]. The results are displayed in Tables 9 and 10, and in Figures 10 and 11.

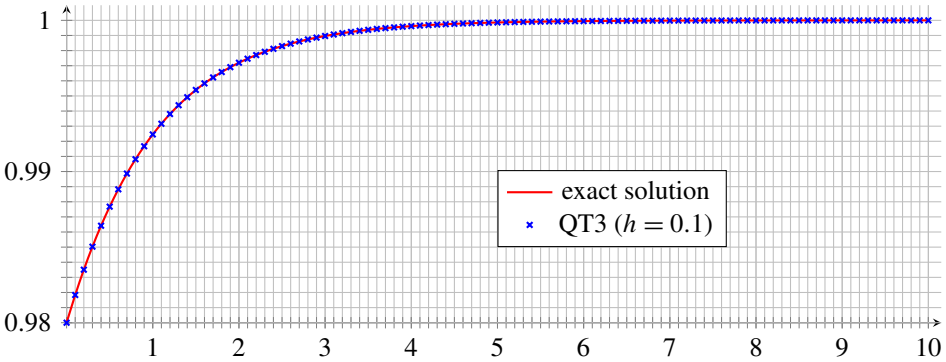



**Figure 11.** Exact solution and QT3 values for (4.5).

| $h$ | K3 | BS3 | RK4 | QT3 |
|---|---|---|---|---|
| 0.1 | $1.0453 \cdot 10^{-6}$ | $1.0450 \cdot 10^{-6}$ | $2.0837 \cdot 10^{-8}$ | $3.4029 \cdot 10^{-10}$ |
| 0.05 | $1.3599 \cdot 10^{-7}$ | $1.3594 \cdot 10^{-7}$ | $1.3576 \cdot 10^{-9}$ | $4.3857 \cdot 10^{-11}$ |
| 0.02 | $8.9142 \cdot 10^{-9}$ | $8.9111 \cdot 10^{-9}$ | $3.5634 \cdot 10^{-11}$ | $2.8583 \cdot 10^{-12}$ |
| 0.01 | $1.1232 \cdot 10^{-9}$ | $1.1228 \cdot 10^{-9}$ | $2.2457 \cdot 10^{-12}$ | $3.5945 \cdot 10^{-13}$ |

**Table 11.** Global error vs. step size for (4.6).

| $h$ | K3 | BS3 | RK4 | QT3 |
|---|---|---|---|---|
| 0.1 | 0.222536 | 0.257829 | 0.267671 | 0.400077 |
| 0.05 | 0.353233 | 0.458556 | 0.455446 | 0.771582 |
| 0.02 | 0.86932 | 1.096014 | 1.093489 | 1.805714 |
| 0.01 | 1.571522 | 2.134188 | 2.135892 | 3.661295 |

**Table 12.** Run time (in seconds) vs. step size for (4.6).

***An equation involving a sine function.*** The following initial value problem is qualitatively similar to the logistic equation as well. Consider

$$\begin{cases} \dot{y} = \sin(y), \\ y|_{t=0} = 0.01, \end{cases} \qquad 0 \le t \le 1. \tag{4.6}$$

The exact solution is

$$y(t) = 2\arctan(\tan(0.005)e^t), \quad 0 \le t \le 1.$$

The results are displayed in Tables 11 and 12, and in Figures 12 and 13.

***Conclusion.*** In the tested cases, the global error of our third-order QT3 method is comparable to and often smaller by several orders of magnitude than the global error of the other tested methods of the same order from the Runge–Kutta family. We even observed it to be smaller or comparable to the global error of the classical

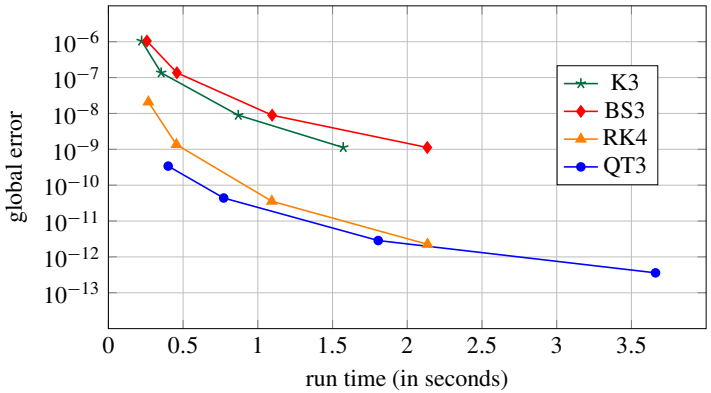



**Figure 12.** Work-precision diagram for (4.6) from Tables 11 and 12.

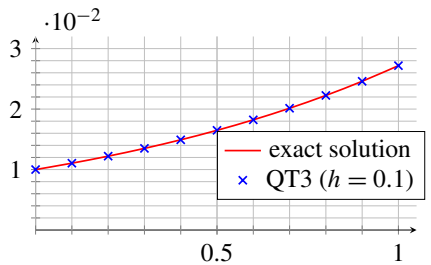



**Figure 13.** Exact solution and QT3 values for (4.6).

Runge–Kutta method of order 4 in most cases. This effect is most pronounced near equilibrium solutions of the tested equations.

Due to the complexity of the formulas the computational cost of QT3 is high. QT3 emphasizes accuracy over computational cost. Nonetheless, as exhibited by the work-precision data for the tested equations, QT3 not only exhibits superior accuracy but also efficiency for approximating solutions to ODEs with initial values near equilibrium solutions as larger step sizes and fewer computations suffice. QT3 thus may be particularly useful in stiff situations (we note that (4.5) is discussed in [Moler 2003] in this context).

In nonstiff situations, such as the initial value problem (4.3) for the Bernoulli equation, the increased computational cost of QT3 over standard algorithms such as BS3 does not appear to be adequately compensated for by gains in accuracy.

## Appendix A: Convergence of one-step methods

Let $D \subset \mathbb{R}$ be open, and suppose $f : D \to \mathbb{R}$ satisfies a local Lipschitz condition in $D$. Let $y : [0, T] \to D$ be the solution to the initial value problem

$$\begin{cases} \dot{y}(t) = f(y(t)) & \text{on } 0 \le t \le T, \\ y|_{t=0} = y_0 \in D. \end{cases}$$

Theorem A.1 below is a general convergence result of abstract numerical one-step methods for the approximation of the solution $y$ on partitions of the interval $[0, T]$; see, for example, [Kress 1998, Section 10.3]. It is the basis for proving Theorem 2.9 in Section 2. We restrict our attention to equidistant partitions of step size $h > 0$.

**Theorem A.1.** *Let $K \Subset D$ be a compact neighborhood with $y([0, T]) \subset \mathring{K}$, and let*

$$\Phi : [0, h_0] \times K \to \mathbb{R}$$

*be continuous, $h_0 > 0$. Assume*:

• ***Consistency***: *$\Phi(0, y) = y$ for all $y \in K$, and $\partial\Phi/\partial h : (0, h_0) \times K \to \mathbb{R}$ exists and extends to a continuous function on $[0, h_0] \times K$ such that $(\partial\Phi/\partial h)(0, y) = f(y)$ for all $y \in K$.*

• ***Lipschitz condition***: *the function $\partial\Phi/\partial h : [0, h_0] \times K \to \mathbb{R}$ satisfies a Lipschitz condition with respect to $y$; i.e., there exists a constant $L > 0$ such that*

$$\left| \frac{\partial\Phi}{\partial h}(h, y) - \frac{\partial\Phi}{\partial h}(h, y') \right| \le L|y - y'|$$

*for all $0 \le h \le h_0$ and $y, y' \in K$.*

• ***Local truncation error***: *there exists $p \ge 1$ and a constant $C \ge 0$ independent of $0 \le h \le h_0$ and $0 \le t \le T$ such that*

$$|y(t + h) - \Phi(h, y(t))| \le Ch^{p+1}$$

*whenever $0 \le t + h \le T$.*

*Then $\Phi$ yields a one-step method of order $p$ for the approximation of $y$ on $[0, T]$; i.e., there exist $N_0 \in \mathbb{N}$ and a constant $M \ge 0$ such that for all $N \ge N_0$, $h = T/N$ the following holds*:

*The sequence of numbers $y_0^{(N)}, \dots, y_N^{(N)}$ defined via*

$$\begin{cases} y_0^{(N)} = y_0, \\ y_{n+1}^{(N)} = \Phi(h, y_n^{(N)}), \quad n = 0, \dots, N - 1, \end{cases}$$

*is well-defined, all $y_n^{(N)} \in \mathring{K}$, and the global error satisfies*

$$\max_{n=0}^{N} |y(nh) - y_n^{(N)}| \le Mh^p. \tag{A.2}$$

*A valid choice for the constant in (A.2) is $M = (C/L)(e^{LT} - 1)$.*

## Appendix B: Differential equations depending on parameters

Let $\Lambda \subset \mathbb{R}^q$ be open, and $V \subset \mathbb{R}$ be an open interval with $0 \in V$. Suppose $F(w; \lambda)$ is continuously differentiable with respect to the variables $(w; \lambda) \in V \times \Lambda$. Consider

the family of ordinary differential equations

$$\begin{cases} \dfrac{\partial w}{\partial t}(t;\lambda) = F(w(t;\lambda);\lambda), \\ w(0;\lambda) = 0, \end{cases} \tag{B.1}$$

for the unknown function $t \mapsto w(t;\lambda)$ depending on the parameter $\lambda \in \Lambda$. The following holds; see [Walter 1998].

**Theorem B.2.** *For each $\lambda \in \Lambda$ there exists a unique maximally extended solution*

$$w(\,\cdot\,;\lambda) : (t_{\min}(\lambda), t_{\max}(\lambda)) \to V$$

*to (B.1), where $-\infty \le t_{\min}(\lambda) < 0 < t_{\max}(\lambda) \le \infty$.*

*The functions $t_{\max}, t_{\min} : \Lambda \to \mathbb{R} \cup \{\pm\infty\}$ are lower and upper semicontinuous, respectively, and the set*

$$U_{\max} = \{(t,\lambda) : \lambda \in \Lambda, t_{\min}(\lambda) < t < t_{\max}(\lambda)\} \subset \mathbb{R} \times \mathbb{R}^q$$

*is open. The solution $w$ to (B.1) defines a map $U_{\max} \to V$, and both $w$ and $\partial w/\partial t$ are continuously differentiable in $U_{\max}$. The partial derivatives of $w$ satisfy*

$$\begin{aligned} \frac{\partial w}{\partial t}(t;\lambda) &= F(w(t;\lambda);\lambda) \quad \textit{(this is just (B.1))}, \\ (\nabla_\lambda w)(t;\lambda) &= \int_0^t e^{\int_s^t F_w(w(u;\lambda);\lambda)\,du} (\nabla_\lambda F)(w(s;\lambda);\lambda)\,ds. \end{aligned} \tag{B.3}$$

*In particular, if $F$ is more than once continuously differentiable, then so is $w$, and formulas for higher partial derivatives of $w$ follow from (B.3) with the chain rule.*

**Remark B.4.** The upper and lower semicontinuity of the endpoint functions of the maximal existence interval follow from the openness of $U_{\max}$. Semicontinuity implies that $t_{\min}$ attains its maximum value $t_{\min}(K) \in \mathbb{R} \cup \{-\infty\}$ and $t_{\max}$ attains its minimum value $t_{\max}(K) \in \mathbb{R} \cup \{\infty\}$ on every compact subset $K \Subset \Lambda$. In particular, $w(t;\lambda)$ is defined (and differentiable) for all $(t;\lambda) \in (t_{\min}(K), t_{\max}(K)) \times K$. Thus, for every compact subset $K \Subset \Lambda$, we are guaranteed that $w(t;\lambda)$ exists on $[0,T] \times K$ for some $T > 0$ (depending on $K$). We make use of this in the theoretical Section 2 of this paper.

## References

[Bogacki and Shampine 1989] P. Bogacki and L. F. Shampine, "A 3(2) pair of Runge–Kutta formulas", *Appl. Math. Lett.* **2**:4 (1989), 321–325. MR Zbl

[Butcher 2008] J. C. Butcher, *Numerical methods for ordinary differential equations*, 2nd ed., John Wiley & Sons, Chichester, 2008. MR Zbl

[Corless et al. 1996] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert $W$ function", *Adv. Comput. Math.* **5**:4 (1996), 329–359. MR Zbl

[Hairer et al. 1993] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving ordinary differential equations, I: Nonstiff problems*, 2nd ed., Springer Series in Computational Mathematics **8**, Springer, 1993. MR Zbl

[Hochbruck and Ostermann 2010] M. Hochbruck and A. Ostermann, "Exponential integrators", *Acta Numer.* **19** (2010), 209–286. MR Zbl

[Kassam and Trefethen 2005] A.-K. Kassam and L. N. Trefethen, "Fourth-order time-stepping for stiff PDEs", *SIAM J. Sci. Comput.* **26**:4 (2005), 1214–1233. MR Zbl

[Kress 1998] R. Kress, *Numerical analysis*, Graduate Texts in Mathematics **181**, Springer, 1998. MR Zbl

[Mickens 1994] R. E. Mickens, *Nonstandard finite difference models of differential equations*, World Scientific, River Edge, NJ, 1994. MR Zbl

[Mickens 2000] R. E. Mickens, "Nonstandard finite difference schemes", pp. 1–54 in *Applications of nonstandard finite difference schemes* (Atlanta, GA, 1999), edited by R. E. Mickens, World Scientific, River Edge, NJ, 2000. MR Zbl

[Moler 2003] C. Moler, "Stiff differential equations", blog post, 2003, https://www.mathworks.com/company/newsletters/articles/stiff-differential-equations.html.

[Patidar 2005] K. C. Patidar, "On the use of nonstandard finite difference methods", *J. Difference Equ. Appl.* **11**:8 (2005), 735–758. MR Zbl

[Ralston 1965] A. Ralston, *A first course in numerical analysis*, McGraw-Hill, New York, 1965. MR Zbl

[Shampine and Reichelt 1997] L. F. Shampine and M. W. Reichelt, "The MATLAB ODE suite", *SIAM J. Sci. Comput.* **18**:1 (1997), 1–22. MR Zbl

[Thieme 2003] H. R. Thieme, *Mathematics in population biology*, Princeton University Press, 2003. MR Zbl

[Vigo-Aguiar and Ramos 2011] J. Vigo-Aguiar and H. Ramos, "A numerical ODE solver that preserves the fixed points and their stability", *J. Comput. Appl. Math.* **235**:7 (2011), 1856–1867. MR Zbl

[Walter 1998] W. Walter, *Ordinary differential equations*, Graduate Texts in Mathematics **182**, Springer, 1998. MR Zbl

Received: 2019-03-19 Accepted: 2020-02-25

krainer@psu.edu *Penn State Altoona, Altoona, PA, United States*

cjz5145@psu.edu *Penn State Altoona, Altoona, PA, United States*

INVOLVE 13:2 (2020)
dx.doi.org/10.2140/involve.2020.13.257

# New generalized secret-sharing schemes with points on a hyperplane using a Wronskian matrix

Weston Loucks and Bahattin Yildiz

(Communicated by Kenneth S. Berenhaut)

A new secret-sharing scheme is constructed using elementary tools from different fields of mathematics. A method is introduced which uses the assignment of points on a hyperplane, serving as terminal points of vectors which meet an outlined criterion for linear independence. Submatrices of a Wronskian matrix are used in the assignment of these points. This method is also generalized to include a weighted scheme and a multilevel hierarchical model.

## 1. Introduction

The first secret-sharing schemes were introduced by Shamir [1979] and Blakley [1979] independently. While Shamir used polynomial interpolation for the basis of his algorithm, Blakley used the intersection of hyperplanes to describe the secret-sharing scheme.

A secret-sharing scheme is a scheme where the key $d$ (called the *secret*) is distributed by a *dealer* to a number of participants in such a way that allows only an authorized subset of participants (called an *access group*) to discover the secret. In a *fully democratic scheme*, $n$ participants each receive a share of the secret and a threshold level of at least $k$ of them must collaborate to recover it.

Since their inception, both Shamir's and Blakley's schemes have inspired other secret-sharing schemes which take different approaches to solving this problem or generalize it into more complex access structures. A *weighted scheme* allows certain individuals to have as much information as multiple participants. The number of shares a participant holds is referred to as their *weight*. A *multilevel hierarchy* requires a specific number of individuals to be present at various levels of a hierarchy, where participants at a higher level may replace those at a lower level but not vice versa. A *dictatorial scheme* is a special case of this model, where one or more participants must be required in all access groups. Different approaches have

*MSC2010:* primary 94A60; secondary 11T71, 34A30.
*Keywords:* secret-sharing schemes, hyperplane, Wronskian, hierarchical secret-sharing schemes, generalized cross product.

been employed to construct schemes that have these properties or other variants. For some of the work done in this context we refer the reader to [Benaloh and Leichter 1990; Beutelspacher and Vedder 1989; Brickell 1989; Charnes et al. 1997; Dawson and Donovan 1994; Ito et al. 1989; Lai and Ding 2004; Simmons 1990; 1992; Tassa 2007].

In this work, we describe a new approach to construct a secret-sharing scheme which uses points on a hyperplane as shares of information. Two main generalizations are explored with this scheme: the weighted scheme and the multilevel hierarchy. The fully democratic scheme is a special case of each of these models. Our proposed scheme uses elementary tools from an undergraduate curriculum and is accessible to readers with a limited technical background.

The rest of the work is organized as follows. In Section 2, we give the necessary preliminaries needed for our construction. In Section 3 we discuss how the dealer constructs a hyperplane equation to contain the secret. Section 4 describes how points may be found which allow for the hyperplane's equation to be reconstructed. Section 5 explains how participants would recover the equation of a hyperplane and discover the secret. Section 6 explains how a weighted scheme is implemented, and Section 8 explores the multilevel hierarchy. An example follows each generalization. Finally, Section 10 presents a proof for why a pair of assigned points will never be collinear with the secret when the threshold level of participants exceeds 2. We finish the paper with concluding remarks and directions for potential future research on the topic.

## 2. Preliminaries

As two distinct points define a unique line in $\mathbb{R}^2$, and three distinct points which are not collinear define a unique plane in $\mathbb{R}^3$, this method uses the principle that the terminal points of $k$ distinct, linearly independent vectors define a unique hyperplane in $\mathbb{R}^k$. Consequently, taking one of these points as a *base point*, the vectors in the direction of the remaining $k-1$ points form $k-1$ linearly independent vectors which can be used to find a vector normal to the hyperplane.

A *$k$-dimensional hyperplane* is defined as the collection of points in $\mathbb{R}^k$ satisfying the equation

$$a_1x_1 + a_2x_2 + \cdots + a_kx_k = b, \tag{2-1}$$

where the constants $a_i \in \mathbb{R}$ are not all equal to zero and $b \in \mathbb{R}$. Each point $(x_1, x_2, \ldots, x_k)$ is in $\mathbb{R}^k$. A hyperplane, in general, is an *affine space* which has one fewer dimension than the ambient vector space in which it exists. An affine space, in this context, is a vector subspace which does not necessarily contain the origin point $\vec{O}$. For example, even though three points are required to define a plane in $\mathbb{R}^3$, only two linearly independent vectors are needed to define a 2-dimensional affine

space of $\mathbb{R}^3$, which is a plane. If one has the knowledge of a point which lies on some plane, and two linearly independent vectors parallel to that plane, then one can define a plane in $\mathbb{R}^3$ and write its equation. The following vector will be normal to the hyperplane:

$$\vec{n} = [a_1 \;\; a_2 \;\; \cdots \;\; a_k]^\top. \tag{2-2}$$

For computational purposes in applications, a hyperplane will be constructed by the dealer such that $a_1, \ldots, a_k \in \mathbb{Z}_+$. Furthermore, the dealer will always set $a_k = 1$. In executing the scheme, participants will calculate a normal vector $\vec{N} = c\vec{n}$ for some $c \in \mathbb{R} \setminus \{0\}$. Since $a_k = 1$, they will divide $\vec{N}$ by its last component $c$ to find $\vec{n}$.

If a point $P = (x_1, x_2, \ldots, x_k) \in \mathbb{R}^k$ satisfies (2-1), it *lies on the hyperplane*. The secret $d \in \mathbb{Z}$ will be associated with the $x_1$-intercept of the hyperplane. If $d = 0$ was never to be used as the secret, the dealer could simply set $b = a_1 d$, where $d$ is the $x_1$-intercept directly. However, when $d = 0$, using this rule causes a problem when the hyperplane equation is being recovered by the participants. This problem is explained in Section 4. To allow 0 as a secret, the dealer can use a bijective mapping from $d \in \mathbb{Z}$ to $b \in \mathbb{Z} \setminus \{0\}$. The following mapping shall be used:

$$b = \begin{cases} a_1 d & \text{if } d > 0, \\ a_1(d-1) & \text{if } d \le 0. \end{cases} \tag{2-3}$$

Therefore, an $x_1$-intercept of $-1$ implies $d = 0$. In general, when the participants find an $x_1$-intercept with a negative value, they must add 1 to it to reveal the secret.

In a fully democratic scheme, each of the $n$ participants holds the same level of information; any participant can be substituted in place of another. In this case, $n$ distinct points that lie on the hyperplane will be assigned to $n$ individuals, whereby any collection of $k$ points defines the same hyperplane in $\mathbb{R}^k$, when expressed in the dealer's form ($a_k = 1$)

$$a_1 x_1 + a_2 x_2 + \cdots + a_{k-1} x_{k-1} + x_k = b. \tag{2-4}$$

## 3. Constructing a suitable hyperplane

A hyperplane can be generated with random parameters while retaining a fixed $x_1$-intercept at $b/a_1$. The dealer creates an equation of the form (2-4) with each integer coefficient $a_i$ chosen randomly from the interval $(0, \max(|d|, n)]$ for $1 \le i < k$. This will generate an *askew hyperplane*, which has an intercept on each axis. A normal vector $\vec{n}$ created by the dealer has the form

$$\vec{n} = [a_1 \;\; a_2 \;\; \cdots \;\; a_{k-1} \;\; 1]^\top. \tag{3-1}$$

The vector $\vec{n}' \in \mathbb{R}^{k-1}$ is defined by the first $k-1$ components of (3-1). Let $x_{j_i}$ denote the $j$-th coordinate of a point assigned to the $i$-th participant. By choosing 1

as the last component of $\vec{n}$, the dealer is then able to calculate $x_{k_i}$ according to the formula

$$x_{k_i} = b - (a_1 x_{1_i} + a_2 x_{2_i} + \cdots + a_{k-1} x_{(k-1)_i}), \tag{3-2}$$

and $x_{k_i}$ retains the randomness from (3-1). The coordinates $x_{1_i}, x_{2_i}, \ldots, x_{(k-1)_i}$ will be chosen from an *assignment matrix* presented in Section 4.

## 4. Assigning suitable points

If a system of random assignment were used for points on the hyperplane, it is possible that some collection of $k$ points cannot be used to reconstruct the hyperplane's equation. For example, three points chosen randomly on a plane in $\mathbb{R}^3$ might be collinear. Therefore, a need arises for a systematic way to create a rectangular matrix of size $(k-1) \times n$, with $k \leq n$, whereby any collection of $k-1$ columns forms a basis for $\mathbb{R}^{k-1}$. This matrix shall be called the *assignment matrix* $A'$. In other words, $A'$ should be constructed so that all submatrices of size $(k-1) \times (k-1)$ will be nonsingular. Applying (3-2) for each participant is comparable to appending an extra row to the bottom of the assignment matrix, creating a matrix of size $k \times n$; then, each column represents an assigned point on the hyperplane. This matrix shall be called the *augmented assignment matrix* $A$. All $k \times k$ submatrices of $A$ must also be nonsingular.

Let $A_{h*}$ represent the $h$-th row of $A$. Using elementary row operations

$$A_{k*} + a_1 A_{1*} + a_2 A_{2*} + a_3 A_{3*} + \cdots + a_{k-1} A_{(k-1)*} = [b \;\; b \;\; \cdots \;\; b]. \tag{4-1}$$

Due to (4-1), $A'$ cannot have a row of 1s, or equivalently, a scalar multiple of such a row because all resulting $A$ would be singular. Equation (4-1) also shows why $b = a_1 d$ cannot be used when $d = 0$.

One way to obtain matrices that satisfy the conditions that we want is by the use of *Wronskian matrices* arising from differential equations. In other words, $A'$ can be taken as a submatrix of a certain Wronskian matrix $W$ when $\det W \neq 0$ can be guaranteed under certain conditions. In general, when $\{y_1, y_2, \ldots, y_n\}$ form a fundamental set of solutions for an ordinary differential equation, the Wronskian matrix is defined as follows [Krusemeyer 1988]:

$$W(y_1, y_2, \ldots, y_n)(x) = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \\ \frac{d}{dx} y_1 & \frac{d}{dx} y_2 & \cdots & \frac{d}{dx} y_n \\ \frac{d^2}{dx^2} y_1 & \frac{d^2}{dx^2} y_2 & \cdots & \frac{d^2}{dx^2} y_n \\ \vdots & \vdots & \ddots & \vdots \\ \frac{d^{n-1}}{dx^{n-1}} y_1 & \frac{d^{n-1}}{dx^{n-1}} y_2 & \cdots & \frac{d^{n-1}}{dx^{n-1}} y_n \end{bmatrix}. \tag{4-2}$$

The first $k-1$ rows of such a matrix will be used for $A'$. The chosen input $x = x_0$ in (4-2) must guarantee a nonzero determinant for $W$ and for all square submatrices coming from its first $k-1$ rows. Consequently, every subset of size $k-1$ from $\{y_1, y_2, \ldots, y_n\}$ must be a set of linearly independent solutions to some $(k-1)$-th order differential equation which can be evaluated at the same input $x_0$.

A special class of differential equations on real-valued functions yields a Wronskian matrix which has a nonzero determinant everywhere the functions are defined. Specifically, this is a property of the fundamental set of solutions to a linear, homogeneous differential equation with constant coefficients, or *LHCC* [Nijenhuis 1980].

An $n$-th order LHCC can be expressed in the form

$$c_n \frac{d^n y}{dx^n} + c_{n-1} \frac{d^{n-1} y}{dx^{n-1}} + \cdots + c_0 y = 0. \tag{4-3}$$

In differential operator notation, $D^n(y) = d^n y/dx^n$, with $D$ being the differential operator on $y$ with respect to $x$. The *characteristic polynomial form* is created by substituting a real-valued variable to the $n$-th power, $r^n$, in place of $D^n(y)$. Consequently, for each distinct real root $r_i$ with $1 \le i \le n$, we know $y_i = e^{r_i x}$ is a fundamental solution for (4-3). Without constraint on the real-valued coefficients $c_1, c_2, \ldots, c_n$ it is possible to construct such a polynomial with exactly $n$ real, distinct roots $r_1, r_2, \ldots, r_n$. Specifically, a differential equation can be written in operator notation as

$$[D - r_1][D - r_2][\cdots][D - r_n](y) = 0. \tag{4-4}$$

It is always possible for an $n$-th order LHCC to have $n$ linearly independent solutions $y_1 = e^{r_1 x}$, $y_2 = e^{r_2 x}$, $\ldots$, $y_n = e^{r_n x}$ because it is always possible to construct (4-4) for any value of $n$. Note that the coefficients $c_i$ would not be chosen by the dealer, but rather the roots $r_i$ would be chosen. The coefficients in (4-3) may be calculated using *elementary symmetric polynomials*, but for the purposes of secret-sharing it is sufficient to note that (4-4) can be created for any collection of $n$ real, distinct roots.

The Wronskian matrix from these functions at $x = 0$ results in the *Vandermonde matrix* [Pólya 1922]:

$$W(e^{r_1 x}, e^{r_2 x}, \ldots, e^{r_n x})(0) = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ r_1 & r_2 & \cdots & r_n \\ r_1^2 & r_2^2 & \cdots & r_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ r_1^{n-1} & r_2^{n-1} & \cdots & r_n^{n-1} \end{bmatrix}. \tag{4-5}$$

The determinant of this matrix is nonzero whenever $r_1, r_2, \ldots, r_n$ are distinct values [Rushanan 1989]. Evaluating $W$ resulting from (4-4) at $x_0 \neq 0$ produces

a matrix which will be equivalent to (4-5) after multiplying a scalar of $e^{-r_i x_0}$ to the $i$-th column, where $1 \le i \le n$ [Pólya 1922]. The Vandermonde matrix is an underlying feature of Shamir's scheme [1979], but appears as a result of polynomial interpolation rather than a Wronskian matrix [Lai and Ding 2004]. The connection between the Vandermonde matrix and the Wronskian matrix for LHCCs is noteworthy as a starting point for the results in this work. However, LHCCs on their own are not practicable for this scheme. Evaluation at $x_0 = 0$ results in a row of 1s, which is problematic when the augmented assignment matrix $A$ is created; $A$ will be singular. Evaluation at $x_0 \in \mathbb{Z} \setminus \{0\}$ yields entries which are not integers when $r_i \in \mathbb{Z}_+$. Integer entries are preferable for computing applications.

These issues can be resolved by using *Cauchy–Euler differential equations* instead of LHCCs. The form of these differential equations is related to that of LHCCs, using a change of variables $x = \ln(t)$ for $t > 0$ [Zill and Cullen 2006]:

$$c_n t^n \frac{d^n y}{dt^n} + c_{n-1} t^{n-1} \frac{d^{n-1} y}{dt^{n-1}} + \cdots + c_0 y = 0. \tag{4-6}$$

For each distinct real root $r$ of a Cauchy–Euler differential equation, $y(t) = t^r$ is a solution. Each term of (4-6) can be computed for $0 < j \le n$ [Zill and Cullen 2006]:

$$c_j t^j \frac{d^j y}{dt^j} = c_j r(r-1)(r-2)\cdots(r-j+1)t^r. \tag{4-7}$$

After applying (4-7) to each term in (4-6), $t^r$ may be factored out. This leaves a polynomial in $r$, with each term having $c_j$ as a leading coefficient. Since the dealer can calculate the values of each $c_j$ from any chosen set of real, distinct roots $r_i$, some polynomial of degree $n$ can always exist with $n$ distinct roots. The dealer can construct a polynomial equation

$$(r - r_1)(r - r_2)\cdots(r - r_n) = 0.$$

Expanding this using elementary symmetric polynomials would reveal exactly what $c_j$ values are necessary to achieve the desired roots. However, since such a polynomial can always exist in theory, the dealer would simply assume that a Cauchy–Euler differential equation exists with the desired roots, without explicitly calculating $c_j$ values.

Let $F$ be defined as the fundamental set of solutions to (4-6). Then, the following property holds regarding its Wronskian determinant [Zill and Cullen 2006]:

$$\det W(F)(t) \neq 0 \quad \Longleftrightarrow \quad t \in (0, +\infty).$$

For this secret-sharing scheme, the dealer will use the Wronskian matrix $W$ of $F$ with distinct $r_i \in \mathbb{Z}_+$. Any resulting $(k-1) \times (k-1)$ submatrix taken from the first $k-1$ rows, evaluated at $t_0 \in (0, +\infty)$, will be nonsingular because the first row will

constitute a fundamental set of solutions for some $(k-1)$-th order Cauchy–Euler differential equation. As a convention for standardizing point assignment, $W$ will be evaluated at $t_0 = 2$, although any $t_0 \in \mathbb{Z}_+ \setminus \{1\}$ could be considered.

When each root $r_i$ is distinct, the first $k-1$ rows of $W$ lead to

$$A'(t) = \begin{bmatrix} t^{r_1} & t^{r_2} & \cdots & t^{r_n} \\ \frac{d}{dt}t^{r_1} & \frac{d}{dt}t^{r_2} & \cdots & \frac{d}{dt}t^{r_n} \\ \frac{d^2}{dt^2}t^{r_1} & \frac{d^2}{dt^2}t^{r_2} & \cdots & \frac{d^2}{dt^2}t^{r_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{d^{k-2}}{dt^{k-2}}t^{r_1} & \frac{d^{k-2}}{dt^{k-2}}t^{r_2} & \cdots & \frac{d^{k-2}}{dt^{k-2}}t^{r_n} \end{bmatrix}. \tag{4-8}$$

Since the dealer chooses each $r_i \in \mathbb{Z}_+$, the *permutation function* $P(n, j) = n!/(n-j)!$ may be utilized in the calculation of (4-8). For all $t^r$, with $r \in \mathbb{Z}_+$ such that $j \le r$,

$$\frac{d^j}{dt^j}t^r = P(n, j)t^{r-j}.$$

When $j > r$,

$$\frac{d^j}{dt^j}t^r = 0.$$

When each $r_i$ is relatively small, many entries of (4-8) will be zero. However, determinants of $(k-1) \times (k-1)$ submatrices taken from the first $k-1$ rows will remain nonzero where $t_0 > 0$. In particular, a submatrix populated by a maximal amount of zero entries would present a worst-case scenario. It would result from the solution set $F_0 = \{1, t, t^2, \ldots, t^{n-1}\}$. One might ask if $\det W(F_0)(t_0) \neq 0$.

From (4-2) the $n$-th row of $W$ is populated by the $(n-1)$-th order derivatives. Specifically, this row would include

$$\frac{d^{n-1}}{dt^{n-1}}t^{n-1} = P(n-1, n-1) = (n-1)!\,.$$

$W(F_0)(t)$ is *upper-triangular*,

$$W(F_0)(t) = \begin{bmatrix} 1 & x & t^2 & \cdots & t^{n-1} \\ 0 & 1 & 2t & \cdots & P(n-1, 1)t^{n-2} \\ 0 & 0 & 2 & \cdots & P(n-1, 2)t^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & P(n-1, n-1) \end{bmatrix}, \tag{4-9}$$

and

$$\det W(F_0)(t) = 0!\,1!\,2! \cdots (n-1)!\,. \tag{4-10}$$

In practice, the points generated by (4-9) would not be used in this secret-sharing scheme; it represents a theoretical scenario. Furthermore, if $1 \in F$ then $d = 1$ could never be used as it would give the secret directly to a participant. A systematic method of assigning each $r_i$ using prime numbers, presented in Section 6, will be used instead. This method will be more advantageous for a weighted scheme and can be applied to the multilevel hierarchy as well.

**4.1.** ***Submatrices of A are nonsingular when constructed from Cauchy–Euler differential equations.*** Using the Wronskian matrix of Cauchy–Euler differential equations with distinct, real roots, all $(k-1) \times (k-1)$ submatrices of $A'(t)$ with $t \in \mathbb{Z}_+ \setminus \{1\}$ will be nonsingular. However, it is not immediately obvious that all $k \times k$ submatrices of $A$ will be nonsingular after the extra row is appended using (3-2). Each entry in this row $x_{k_i}$ will follow the formula

$$x_{k_i} = b - \left( a_1 t^{r_i} + a_2 \frac{d}{dt} t^{r_i} + \cdots + a_{k-1} \frac{d^{k-2}}{dt^{k-2}} t^{r_i} \right). \tag{4-11}$$

Therefore, $A(t)$ can be written as

$$A(t) = \begin{bmatrix} t^{r_1} & t^{r_2} & \cdots & t^{r_n} \\ \frac{d}{dt} t^{r_1} & \frac{d}{dt} t^{r_2} & \cdots & \frac{d}{dt} t^{r_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{d^{k-2}}{dt^{k-2}} t^{r_1} & \frac{d^{k-2}}{dt^{k-2}} t^{r_2} & \cdots & \frac{d^{k-2}}{dt^{k-2}} t^{r_n} \\ x_{k_1} & x_{k_2} & \cdots & x_{k_n} \end{bmatrix}. \tag{4-12}$$

We will show that any $k \times k$ submatrix of $A(t)$ is nonsingular when evaluated at $t_0 \in \mathbb{Z}_+ \setminus \{1\}$. Without loss of generality, define the submatrix $A_k$ by the first $k$ columns of (4-12). We will apply elementary row operations on $A_k$ to arrive at a new matrix $M$. If we can prove that $\det M \neq 0$, then we can infer that $A_k$ will be nonsingular. We will use the notation $A_k \sim B$ to indicate that the matrix $B$ comes from using elementary row operations on $A_k$, and thus the matrices have equal rank.

Apply (4-1), and then multiply the $k$-th row by $1/b$ to get

$$A_k \sim \begin{bmatrix} t^{r_1} & t^{r_2} & \cdots & t^{r_k} \\ \frac{d}{dt} t^{r_1} & \frac{d}{dt} t^{r_2} & \cdots & \frac{d}{dt} t^{r_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{d^{k-2}}{dt^{k-2}} t^{r_1} & \frac{d^{k-2}}{dt^{k-2}} t^{r_2} & \cdots & \frac{d^{k-2}}{dt^{k-2}} t^{r_k} \\ 1 & 1 & \cdots & 1 \end{bmatrix}. \tag{4-13}$$

Swap rows until $[1 \ 1 \ \cdots \ 1]$ is the first row, and all following rows appear in their original order:

$$A_k \sim B = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ t^{r_1} & t^{r_2} & \cdots & t^{r_k} \\ \frac{d}{dt}t^{r_1} & \frac{d}{dt}t^{r_2} & \cdots & \frac{d}{dt}t^{r_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{d^{k-2}}{dt^{k-2}}t^{r_1} & \frac{d^{k-2}}{dt^{k-2}}t^{r_2} & \cdots & \frac{d^{k-2}}{dt^{k-2}}t^{r_k} \end{bmatrix}. \tag{4-14}$$

Perform the following row operations, where $B_{hi}$ represents the entry of $B$ on the $h$-th row, $i$-th column, and $B_{h*}$ represents the $h$-th row of $B$:

$$\begin{gathered} B_{k*} - (B_{k1})B_{1*} \to B_{k*}, \\ B_{(k-1)*} - (B_{(k-1)1})B_{1*} \to B_{(k-1)*}, \\ \vdots \\ B_{2*} - (B_{21})B_{1*} \to B_{2*}. \end{gathered}$$

Then $A_k \sim B \sim M$, where

$$M = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & t^{r_2}-t^{r_1} & t^{r_3}-t^{r_1} & \cdots & t^{r_k}-t^{r_1} \\ 0 & \frac{d}{dt}t^{r_2}-\frac{d}{dt}t^{r_1} & \frac{d}{dt}t^{r_3}-\frac{d}{dt}t^{r_1} & \cdots & \frac{d}{dt}t^{r_k}-\frac{d}{dt}t^{r_1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \frac{d^{k-2}}{dt^{k-2}}t^{r_2}-\frac{d^{k-2}}{dt^{k-2}}t^{r_1} & \frac{d^{k-2}}{dt^{k-2}}t^{r_3}-\frac{d^{k-2}}{dt^{k-2}}t^{r_1} & \cdots & \frac{d^{k-2}}{dt^{k-2}}t^{r_k}-\frac{d^{k-2}}{dt^{k-2}}t^{r_1} \end{bmatrix}. \tag{4-15}$$

Observe that

$$\det M = \det \begin{bmatrix} t^{r_2}-t^{r_1} & t^{r_3}-t^{r_1} & \cdots & t^{r_k}-t^{r_1} \\ \frac{d}{dt}t^{r_2}-\frac{d}{dt}t^{r_1} & \frac{d}{dt}t^{r_3}-\frac{d}{dt}t^{r_1} & \cdots & \frac{d}{dt}t^{r_k}-\frac{d}{dt}t^{r_1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{d^{k-2}}{dt^{k-2}}t^{r_2}-\frac{d^{k-2}}{dt^{k-2}}t^{r_1} & \frac{d^{k-2}}{dt^{k-2}}t^{r_3}-\frac{d^{k-2}}{dt^{k-2}}t^{r_1} & \cdots & \frac{d^{k-2}}{dt^{k-2}}t^{r_k}-\frac{d^{k-2}}{dt^{k-2}}t^{r_1} \end{bmatrix}. \tag{4-16}$$

Let $\beta_i$ for $i \in \{1, 2, \ldots, k\}$ represent constant coefficients. Consider a Cauchy–Euler differential equation whose solution is

$$\begin{aligned} y(t) &= \beta_2(t^{r_2} - t^{r_1}) + \beta_3(t^{r_3} - t^{r_1}) + \cdots + \beta_k(t^{r_k} - t^{r_1}) \\ &= -(\beta_2 + \beta_3 + \cdots + \beta_k)t^{r_1} + \beta_2 t^{r_2} + \beta_3 t^{r_3} + \cdots + \beta_k t^{r_k}. \end{aligned}$$

This is a solution to a Cauchy–Euler differential equation where

$$c_1 = -(c_2 + c_3 + \cdots + c_k).$$

We may then consider (4-16) as a Wronskian determinant for a set of solutions $\{t^{r_2}-t^{r_1}, t^{r_3}-t^{r_1}, \ldots, t^{r_k}-t^{r_1}\}$, which are linearly independent, since all the $r_i$ are distinct. Therefore, $\det M \neq 0$ when $t_0 \in \mathbb{Z}_+ \setminus \{1\}$ and hence $A_k$ is nonsingular.

## 5. Recovering the secret

When $k$ points on a hyperplane are known, one point is chosen to be the base point, $P_B$. Then $k-1$ vectors $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_{k-1}$ are calculated using $P_B$ as the initial point and each of the remaining $k-1$ points as terminal points. Once $k-1$ vectors are known, the participants will find the *generalized cross product* $\bigwedge$ of these linearly independent vectors, resulting in a normal vector $\vec{N}$. Here, $\hat{e}_1 \in \mathbb{R}^k$ represents a unit basis vector with $k$ components having 1 as the $i$-th component and 0 elsewhere:

$$\vec{N} = \bigwedge(\vec{v}_1, \ldots, \vec{v}_{k-1}) = \det \begin{bmatrix} & \vec{v}_1 & \\ & \vec{v}_2 & \\ & \vdots & \\ & \vec{v}_{k-1} & \\ \hat{e}_1 & \hat{e}_2 \cdots & \hat{e}_k \end{bmatrix}. \tag{5-1}$$

$\vec{N}$ is parallel to $\vec{n}$, given in (3-1). To find $\vec{n}$, the participants divide $\vec{N}$ by its last component. Next, $P_B$ is considered as a row vector $\vec{P}_B$. The equation of the hyperplane can be found:

$$[x_1 \;\; x_2 \;\; \cdots \;\; x_k]\,\vec{n} = \vec{P}_B\vec{n}. \tag{5-2}$$

Evaluating the products reveals (2-4). Finally, the participants set $x_2 = x_3 = \cdots = x_k = 0$ and solve the resulting equation for $x_1 = d$.

## 6. The weighted scheme

Instead of assigning each participant a single point, one could also assign a line, plane, or $m$-dimensional affine space that lies on the hyperplane in $\mathbb{R}^k$ with $m < k-1$. For example, someone who knows the equation of a line on the hyperplane has twice as much information as someone who only has a point. Generalizing this notion, the knowledge of an $m$-dimensional affine space is equivalent to knowing $m+1$ points that lie on the hyperplane. A point is regarded as a 0-dimensional affine space. These pieces of $d$, called *shadows*, [Ito et al. 1989] can be thought of in terms of geometric representations such as lines or affine spaces, but in practice will be assigned as a set $S$ of points. The number of points in $S$ is referred to as the weight of the shadow. Points that are already assigned cannot be reused.

In the weighted scheme, $n$ does not represent the number of people involved, but rather the total weight of all assigned shadows, while $k$ represents the threshold

sum of weights needed to recover the secret. The case where every participant has only a weight-1 shadow is equivalent to the fully democratic scheme.

Without a defined system of organizing the points, issues may arise as new participants are added to the weighted scheme. For example, it would be difficult to know whether an arbitrary point $P_i$ can be assigned to a new participant, or whether that point exists in $S$ for an existing participant. Each person must have an infinite set of points on reserve, sharing a property unique to that person, so that those points will never be assigned to a different participant.

$S_i$ will denote a shadow for the $i$-th participant. $P_{i,j}$ refers to the $j$-th point given to the $i$-th participant, $f_{i,j}$ refers to the solution of a Cauchy–Euler differential equation used to generate $P_{i,j}$, and $p_i$ denotes the $i$-th prime number. A shadow of weight $m+1$ is defined as

$$\begin{aligned} S_i &= \{P_{i,1}, P_{i,2}, \ldots, P_{i,m+1}\}, \quad && 1 \le i \le n, m+1 < k, \\ f_{i,j} &= t^{p_i^j}, && 1 \le j \le m+1. \end{aligned} \tag{6-1}$$

This will allow each participant to have a reserve of points that can be used as shadows for that participant only.

## 7. Weighted scheme example

Suppose six people with a total shadow weight of $n = 10$ are given partial information, and a total shadow weight of $k = 5$, or greater, is needed to recover the secret $d = 9$.

The dealer chooses $k - 1 = 4$ random integers from (0, 10]: 6, 7, 7, 4. This is used to construct (3-1).

$$\vec{n} = [6 \;\; 7 \;\; 7 \;\; 4 \;\; 1]^\top.$$

The dealer calculates $a_1 d = 6 \cdot 9 = 54$. An equation of form (2-4) is created:

$$6x_1 + 7x_2 + 7x_3 + 4x_4 + x_5 = 54. \tag{7-1}$$

Next, the dealer must create $n = 10$ vectors in $\mathbb{R}^{k-1} = \mathbb{R}^4$ such that any set of four vectors is guaranteed to be linearly independent. For this task, a Wronskian matrix of a tenth order Cauchy–Euler differential equation is used.

Suppose there are six people entrusted to hold partial information about the secret: Alice, Bob, Carlos, David, Elif, and Fahad. Furthermore, suppose Alice has a shadow of weight 3 (a 2-dimensional affine space). Bob and Carlos both have shadows of weight 2 (a line), while David, Elif, and Fahad each have shadows of weight 1 (a point). To stay organized and identify which point belongs to whom, (6-1) is applied:

$$\begin{aligned} &f_{1,1} = t^2, \quad f_{1,2} = t^4, \quad f_{1,3} = t^8, \quad f_{2,1} = t^3, \quad f_{2,2} = t^9, \\ &f_{3,1} = t^5, \quad f_{3,2} = t^{25}, \quad f_{4,1} = t^7, \quad f_{5,1} = t^{11}, \quad f_{6,1} = t^{13}. \end{aligned} \tag{7-2}$$

The dealer considers the fundamental set of solutions $F$ listed in (7-2). The Wronskian matrix is given by

$$W(F)(t) = \begin{bmatrix} t^2 & t^4 & t^8 & t^3 & t^9 & t^5 & t^{25} & \cdots & t^{13} \\ 2t & 4t^3 & 8t^7 & 3t^2 & 9t^8 & 5t^4 & 25t^{24} & \cdots & 13t^{12} \\ 2 & 12t^2 & 56t^6 & 6t & 72t^7 & 20t^3 & 600t^{23} & \cdots & 156t^{11} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 9! & 0 & P(25,9)t^{16} & \cdots & P(13,9)t^4 \end{bmatrix}. \quad (7\text{-}3)$$

The first four rows of (7-3) are evaluated at $t = 2$ to get the $4 \times 10$ submatrix

$$A' = \begin{bmatrix} 4 & 16 & 256 & 8 & 512 & 32 & 2^{25} & 128 & 2048 & 8192 \\ 4 & 32 & 1024 & 12 & 2304 & 80 & 25 \cdot 2^{24} & 448 & 11 \cdot 2^{10} & 13 \cdot 2^{12} \\ 2 & 48 & 3584 & 12 & 9216 & 160 & 600 \cdot 2^{23} & 1344 & 110 \cdot 2^9 & 156 \cdot 2^{11} \\ 0 & 48 & 336 \cdot 2^5 & 6 & 504 \cdot 2^6 & 240 & P(25,3)2^{22} & 3360 & 990 \cdot 2^8 & P(13,3)2^{10} \end{bmatrix}. \quad (7\text{-}4)$$

Next, the dealer determines the $x_5$-value that will be associated with each column of (7-4) using (3-2). The points are then assigned as follows.

Alice:

$$P_{1,1} = (4, 4, 2, 0, 54 - (4 \cdot 6 - 4 \cdot 7 - 2 \cdot 7 - 0 \cdot 4)) = (4, 4, 2, 0, -12),$$

$$P_{1,2} = (16, 32, 48, 48, 54 - (16 \cdot 6 - 32 \cdot 7 - 48 \cdot 7 - 48 \cdot 4)) = (16, 32, 48, 48, -794),$$

$$P_{1,3} = (256, 1024, 3584, 336 \cdot 2^5, -76746).$$

Bob:

$$P_{2,1} = (8, 12, 12, 6, 54 - (8 \cdot 6 - 12 \cdot 7 - 12 \cdot 7 - 6 \cdot 4)) = (8, 12, 12, 6, -186),$$

$$P_{2,2} = (512, 2304, 9216, 504 \cdot 2^6, -212682).$$

Carlos:

$$P_{3,1} = (32, 80, 160, 240, -2778),$$

$$P_{3,2} = (2^{25}, 25 \cdot 2^{24}, 600 \cdot 2^{23}, P(25,3)2^{22}, -269895073738).$$

David:

$$P_{4,1} = (128, 448, 1344, 990 \cdot 2^8, -1027018).$$

Elif:

$$P_{5,1} = (2048, 11 \cdot 2^{10}, 110 \cdot 2^9, 990 \cdot 2^8, -1499082).$$

Fahad:

$$P_{6,1} = (8192, 13 \cdot 2^{12}, 156 \cdot 2^{11}, P(13,3)2^{10}, -9686986).$$

Suppose that Alice, David, and Elif must collaborate to recover the secret. Their total shadow weight $3 + 1 + 1 = k$ is large enough to do so. The participants choose

a base point, such as Alice's point $P_{1,1}$. Vectors are expressed starting from this point:

$$\overrightarrow{P_{1,1}P_{1,2}} = P_{1,2} - P_{1,1} = \langle 12, 28, 46, 48, -782\rangle,$$
$$\overrightarrow{P_{1,1}P_{1,3}} = P_{1,3} - P_{1,1} = \langle 252, 1020, 3582, 10752, -76734\rangle,$$
$$\overrightarrow{P_{1,1}P_{4,1}} = P_{4,1} - P_{1,1} = \langle 124, 444, 1342, 253440, -1027006\rangle,$$
$$\overrightarrow{P_{1,1}P_{5,1}} = P_{5,1} - P_{1,1} = \langle 2044, 11260, 56318, 253440, -1499070\rangle.$$

The participants then use (5-1):

$$\vec{N} = \det \begin{bmatrix} 12 & 28 & 46 & 48 & -782 \\ 52 & 1020 & 3582 & 10752 & -76734 \\ 124 & 444 & 1342 & 253440 & -1027006 \\ 2044 & 11260 & 56318 & 253440 & -1499070 \\ \hat{e}_1 & \hat{e}_2 & \hat{e}_3 & \hat{e}_4 & \hat{e}_5 \end{bmatrix}. \tag{7-5}$$

Thus,

$$\vec{N} = \begin{bmatrix} -71428262068224 \\ -83332972412928 \\ -83332972412928 \\ -47618841378816 \\ -11904710344704 \end{bmatrix} = -11904710344704 \begin{bmatrix} 6 \\ 7 \\ 7 \\ 4 \\ 1 \end{bmatrix}. \tag{7-6}$$

After dividing by $-11904710344704$, the participants recover $\vec{n}$. $\vec{P_B}$ represents the row vector with the components of the base point $P_{1,1}$. Then, the participants apply (5-2) to find (7-1). They find the $x_1$-intercept by letting $x_2 = x_3 = x_4 = x_5 = 0$, which implies $6x_1 = 54$. Finally, they discover the secret $x_1 = 9$.

## 8. The multilevel hierarchy

Some organizations may wish to use a hierarchy of participants that require a specified number individuals from various levels of the hierarchy to be present. For example, consider a military organization which is structured so that *private officers* have the lowest level of security clearance, *corporals* are one level above private officers, *sergeants* are one level above corporals, and so on to the top security level. Suppose further that there is a military secret which can be recovered by $k$ participants, from which at least $\alpha_1$ must be corporals, $\alpha_2$ must be sergeants, and in general at least $\alpha_j$ participants must be $j$ levels above private officers. Additionally, someone at a higher level may substitute for someone lower on the hierarchy. For instance, a corporal may take the place of a private officer if needed and a sergeant may take the place of a corporal or a private officer. This model is equivalent to

the fully democratic scheme if there is only one security level, treated as the top security level, requiring $k$ participants to be present. This secret-sharing scheme may accommodate the multilevel hierarchy by making several modifications.

Firstly, instead of assigning a point to each individual on the lowest level of the hierarchy, each of these participants will instead receive a distinct vector parallel to the hyperplane. Then, each participant above the lowest level will be assigned a point on the lowest level's hyperplane, as well as a distinct vector parallel to it. In this way, higher level participants must be present to recover (2-4); without them, a group of private officers at best can find $\vec{n}$ but not $a_1 d$. As far as that group knows the hyperplane could pass through the origin; based on this information alone, $d$ is as likely to be 0 as it is any other value.

Secondly, it requires secrets to be *nested*; for any participant who is assigned a point on a hyperplane, its *last coordinate* is kept secret (with the exception of the top security level). An unknown last coordinate becomes the secret $x_1$-intercept of a *different* hyperplane associated with the next higher security level. Each hyperplane will exist in a different vector space associated with each level of the hierarchy. Since each hyperplane can have only *one* $x_1$-intercept, a point with an unknown coordinate must be *shared*; it must be the *same* point for each participant who receives it. Any participant below the top security level is assigned a distinct vector parallel to the hyperplane associated with their level, instead of a point on it. Participants above the lowest level are assigned a point accompanied by a vector on all the hyperplanes associated with security levels below them. Finally, each participant at the top security level would know all coordinates of a *distinct* point that lies on their hyperplane; assignment of shares on the top level works as it does for the fully democratic scheme with respect to participants at the top level.

Recovering the secret begins on the top level. Once a hyperplane equation in each level is solved for its $x_1$-intercept, it reveals the unknown coordinate of a point one level below to be solved, until eventually the hyperplane equation on the lowest level can be recovered. Finding the $x_1$-intercept of that hyperplane reveals $d$.

As a consequence of the shared point, hyperplanes created by the dealer (with the exception of that at the top level) must have *one more dimension* than the number of participants in that access group. For example, consider the case where *three* participants are required, one of which must be a corporal. Then, two private officers hold distinct vectors parallel to their hyperplane, along with one corporal who holds a point on this hyperplane and his own distinct vector parallel to it. This describes a hyperplane defined by three vectors along with one point. However, in $\mathbb{R}^3$ only two vectors along with one point are needed to recover the hyperplane. Therefore, the dealer should instead create a hyperplane in $\mathbb{R}^4$ at the lowest level to accommodate the three participants. To generalize this, suppose $k$ participants are required and the highest level in the hierarchy is $T$ levels above the lowest. Let $H_j$

denote the hyperplane $j$ security levels above the lowest level, with $H_0$ denoting the hyperplane at the lowest level:

$$H_0 \in \mathbb{R}^{k+1}, \quad H_1 \in \mathbb{R}^{\alpha_1+\alpha_2+\alpha_3+\cdots+\alpha_T+1}, \quad \cdots, \quad H_{T-1} \in \mathbb{R}^{\alpha_{T-1}+\alpha_T+1}, \quad H_T \in \mathbb{R}^{\alpha_T}.$$

A shared point will lie on $H_0, H_1, \ldots, H_{T-1}$ respectively. Since a point $P_i$ generated by $t^{p_i}$ is associated with the $i$-th participant, the shared points should be of a different form. It is sufficient to use the solution $f(t) = t$ to generate each shared point.

By $d_{H_j}$ we denote the secret $x_1$-intercept on hyperplane $H_j$ (the overall secret is $d = d_{H_0}$). The dealer can convert any point $P_i$ to a vector $\vec{v}_i$ parallel to a hyperplane by using the point $(0, 0, \ldots, a_1 d_{H_j})$ as the initial point for $\vec{v}_i$:

$$\vec{v}_i = P_i - (0, 0, \ldots, a_1 d_{H_j}). \tag{8-1}$$

Let the $i$-th column of $A'$ be denoted by $A'_{*i}$. Subtracting $a_1 d_{H_j}$ from (3-2) we get

$$x_{k_i} - a_1 d_{H_j} = (a_1 d_{H_j} - (A'_{*i})^T \vec{n}') - a_1 d_{H_j} = -(A'_{*i})^T \vec{n}'.$$

Therefore,

$$\vec{v}_i = \left[ t^{p_i} \quad \frac{d}{dt} t^{p_i} \quad \frac{d}{dt^2} t^{p_i} \quad \cdots \quad \frac{d}{dt^{k-1}} t^{p_i} \quad -(A'_{*i})^T \vec{n}' \right]^\top. \tag{8-2}$$

## 9. Multilevel hierarchy example

A secret $d = 331$ is shared with $n = 8$ people, such that $k = 6$ participants must collaborate to recover the secret. Furthermore, among these six participants there must be at least $\alpha_1 = 3$ corporals and $\alpha_2 = 2$ sergeants at the top security level. The remaining participant may be a private officer at the lowest security level. Suppose the eight people are ranked as follows, with their respective Cauchy–Euler solutions in parentheses: *Private* Alice ($t^2$), *Private* Bob ($t^3$), *Corporal* Carlos ($t^5$), *Corporal* David ($t^7$), *Corporal* Elif ($t^{11}$), *Sergeant* Fahad ($t^{13}$), *Sergeant* Garret ($t^{17}$), *Sergeant* Hanna ($t^{19}$).

The dealer considers three vector spaces, one for each security level, which will each contain a single hyperplane $H_0$, $H_1$, $H_2$ respectively:

$$H_0 \in \mathbb{R}^{k+1} = \mathbb{R}^7, \quad H_1 \in \mathbb{R}^{\alpha_1+\alpha_2+1} = \mathbb{R}^6, \quad H_2 \in \mathbb{R}^{\alpha_2} = \mathbb{R}^2.$$

In constructing $H_0 \in \mathbb{R}^7$, the dealer randomly chooses $7 - 1 = 6$ integers from (0,331]: 256, 154, 306, 207, 111, 151. By $\vec{n}_{H_0}$ we denote the normal vector for $H_0$. Then

$$\vec{n}'_{H_0} = [256 \;\; 154 \;\; 306 \;\; 207 \;\; 111 \;\; 151]^\top,$$
$$\vec{n}_{H_0} = [256 \;\; 154 \;\; 306 \;\; 207 \;\; 111 \;\; 151 \;\; 1]^\top.$$

After calculating $a_1 d = 256 \cdot 331 = 84736$, the dealer writes the equation for $H_0$:

$$256x_1 + 154x_2 + 306x_3 + 207x_4 + 111x_5 + 151x_6 + x_7 = 84736. \tag{9-1}$$

There are eight to receive information, along with an additional shared point generated by $f(t) = t$, so the dealer uses a Wronskian matrix of size $9 \times 9$. However, only the first $7 - 1 = 6$ rows will be used for $H_0$, the first $6 - 1 = 5$ rows for $H_1$, and the first $2 - 1 = 1$ row for $H_2$. The last column will be used for a shared point by the higher-level participants on $H_0$ and $H_1$. Each of the other columns corresponds with a distinct participant, generated by their Cauchy–Euler solution and its derivatives:

$$W(F)(t) = \begin{bmatrix} t^2 & t^3 & t^5 & t^7 & \cdots & t^{19} & t \\ 2t & 3t^2 & 5t^4 & 7t^6 & \cdots & 19t^{18} & 1 \\ 2 & 6t & 20t^3 & 42t^5 & \cdots & 342t^{17} & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & 0 & \cdots & P(19,8)t^{12} & 0 \end{bmatrix}. \tag{9-2}$$

$W(F)(2)$ is evaluated, and the first $\mathbf{7} - 1 = 6$ rows define the assignment matrix $A'$:

$$A' = \begin{bmatrix} 4 & 8 & 32 & 128 & \cdots & 2^{19} & 2 \\ 4 & 12 & 80 & 448 & \cdots & 19 \cdot 2^{18} & 1 \\ 2 & 12 & 160 & 1344 & \cdots & 342 \cdot 2^{17} & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 5! & 10080 & \cdots & P(19,5) \cdot 2^{14} & 0 \end{bmatrix}. \tag{9-3}$$

Using the last column of $A'$, each of the corporals and sergeants knows a point lies on $H_0$ of the form (2,1,0,0,0,0,$d_{H_1}$). The private officers have no knowledge about $H_1$. Applying (8-2), the dealer determines a vector $\vec{v}_{i,H_0}$ parallel to $H_0$, which is assigned to the $i$-th participant:

$$\begin{aligned} \vec{v}_{1,H_0} &= \langle 4, 4, 2, 0, 0, 0, -A'^{T}_{*1}\vec{n}'_{H_0} \rangle \\ &= \langle 4, 4, 2, \ldots, -(4{\cdot}256{+}4{\cdot}154{+}2{\cdot}306{+}0{\cdot}207{+}0{\cdot}111{+}0{\cdot}151) \rangle \\ &= \langle 4, 4, 2, 0, 0, 0, -2252 \rangle, \end{aligned}$$

$$\vec{v}_{2,H_0} = \langle 8, 12, 12, 6, 0, 0, -A'^{T}_{*2}\vec{n}'_{H_0} \rangle = \langle 8, 12, 12, 6, 0, 0, -8810 \rangle,$$

$$\begin{aligned} \vec{v}_{3,H_0} &= \langle 32, 80, 160, 240, 240, 120, -A'^{T}_{*3}\vec{n}'_{H_0} \rangle \\ &= \langle 32, 80, 160, 240, 240, 120, -163912 \rangle, \end{aligned}$$

$$\vec{v}_{4,H_0} = \langle 128, 448, 1344, 3360, 6720, 10080, -3476544 \rangle,$$

$$\vec{v}_{5,H_0} = \langle 2^{11}, 11{\cdot}2^{10}, 110{\cdot}2^{9}, 990{\cdot}2^{8}, 7920{\cdot}2^{7}, 55440{\cdot}2^{6}, -720254464 \rangle,$$

$$\vec{v}_{6,H_0} = \langle 2^{13}, P(13,1)2^{12}, P(13,2)2^{11}, P(13,3)2^{10}, P(13,4)2^{9}, P(13,5)2^{8}, -7417067520\rangle,$$

$$\vec{v}_{7,H_0} = \langle 2^{17}, P(17,1)2^{16}, P(17,2)2^{15}, P(17,3)2^{14}, P(17,4)2^{13}, P(17,5)2^{12}, -527980036096\rangle,$$

$$\vec{v}_{8,H_0} = \langle 2^{19}, P(19,1)2^{18}, P(19,2)2^{17}, P(19,3)2^{16}, P(19,4)2^{15}, P(19,5)2^{14}, -3883940315136\rangle.$$

The dealer creates an equation for $H_1 \in \mathbb{R}^6$ and calculates vectors parallel to this hyperplane in a similar manner. Each corporal and sergeant will be assigned a vector parallel to $H_1$. This hyperplane will have $d_{H_1}$ as its $x_1$-intercept.

The point $(2, 1, 0, 0, 0, 0, d_{H_1})$ lies on $H_0$. The dealer uses this fact to calculate its value:

$$d_{H_1} = 84736 - (256 \cdot 2 + 154 \cdot 1 + 306 \cdot 0 + 207 \cdot 0 + 111 \cdot 0 + 151 \cdot 0) = 84070.$$

By $\vec{n}_{H_1}$ we denote the normal vector for $H_1$. The dealer chooses $6-1=5$ random integers from (0,331] for its first five components, using 1 as its last component:

$$\vec{n}'_{H_1} = [86 \;\; 237 \;\; 156 \;\; 158 \;\; 72]^\top,$$
$$\vec{n}_{H_1} = [86 \;\; 237 \;\; 156 \;\; 158 \;\; 72 \;\; 1]^\top.$$

After calculating $a_1 d_{H_1} = 86 \cdot 84070 = 7230020$, the equation for $H_1$ can be written as

$$86x_1 + 237x_2 + 156x_3 + 158x_4 + 72x_5 + x_6 = 7230020. \tag{9-4}$$

Let $A''$ be defined by the first five rows of $A'$, given in (9-3). The dealer uses $\vec{n}'_{H_1}$ and $A''$ to determine vectors parallel to $H_1$ by applying (8-2). From the last column of $A''$, each sergeant knows a point lies on $H_1$ of the form $(2, 1, 0, 0, 0, d_{H_2})$. The corporals have no knowledge about $H_2$. Vectors $v_{i,H_1}$ parallel to $H_1$ are assigned to each corporal and sergeant:

$$\begin{aligned}\vec{v}_{3,H_1} &= \langle 32, 80, 160, 240, 240, -A''^{T}_{*3}\vec{n}'_{H_1}\rangle \\ &= \langle 32, 80, \ldots, -(32\cdot 86+80\cdot 237+160\cdot 156+240\cdot 158+240\cdot 72)\rangle, \\ &= \langle 32, 80, 160, 240, 240, -101872\rangle,\end{aligned}$$

$$\begin{aligned}\vec{v}_{4,H_1} &= \langle 128, 448, 1344, 3360, 6720, -A''^{T}_{*4}\vec{n}'_{H_1}\rangle \\ &= \langle 128, 448, 1344, 3360, 6720, -1341568\rangle,\end{aligned}$$

$$\vec{v}_{5,H_1} = \langle 2^{11}, 11\cdot 2^{10}, 110\cdot 2^{9}, 990\cdot 2^{8}, 7920\cdot 2^{7}, -124665856\rangle,$$

$$\vec{v}_{6,H_1} = \langle 2^{13}, P(13,1)2^{12}, P(13,2)2^{11}, P(13,3)2^{10}, P(13,4)2^{9}, -973385728\rangle,$$

$$\vec{v}_{7,H_1} = \langle 2^{17}, P(17,1)2^{16}, P(17,2)2^{15}, P(17,3)2^{14}, P(17,4)2^{13}, -45918257152\rangle,$$

$$\vec{v}_{8,H_1} = \langle 2^{19}, P(19,1)2^{18}, P(19,2)2^{17}, P(19,3)2^{16}, P(19,4)2^{15}, -287891783680\rangle.$$

The dealer creates an equation for $H_2 \in \mathbb{R}^2$ and calculates *points* on this top-security-level hyperplane. Each sergeant will be assigned a point on $H_2$.

The point $(2, 1, 0, 0, 0, d_{H_2})$ lies on $H_1$. The dealer uses this fact to calculate its value:

$$d_{H_2} = 7230020 - (86 \cdot 2 + 237 \cdot 1 + 156 \cdot 0 + 158 \cdot 0 + 72 \cdot 0) = 7229611.$$

The dealer chooses $2 - 1 = 1$ random integers from (0,331]:

$$\vec{n}'_{H_2} = [33], \quad \vec{n}_{H_2} = \begin{bmatrix} 33 \\ 1 \end{bmatrix}.$$

After calculating $a_1 d_{H_2} = 33 \cdot 7229611 = 238577163$, the equation for $H_2$ can be written as

$$33x_1 + x_2 = 238577163. \tag{9-5}$$

The $x_1$-value for each sergeant comes from each of their respective entries on the first row of $A'$. These values, along with (9-5), are used to determine their respective $x_2$-values. Points $P_{i,H_2}$ are assigned to each sergeant:

$$P_{6,H_2} = (2^{13}, 238577163 - 33 \cdot 2^{13}) = (2^{13}, 238306827),$$
$$P_{7,H_2} = (2^{17}, 238577163 - 33 \cdot 2^{17}) = (2^{17}, 234251787),$$
$$P_{8,H_2} = (2^{19}, 238577163 - 33 \cdot 2^{19}) = (2^{19}, 221275659).$$

Suppose *Private* Bob, *Corporal* Carlos, *Corporal* David, *Corporal* Elif, *Sergeant* Fahad, and *Sergeant* Garret must recover the secret $d = 331$. These participants meet the necessary conditions, in number and security level, to recover it.

First the two sergeants collaborate to find $d_{H_2}$. They start by finding a vector:

$$\overrightarrow{P_{6,H_2}P_{7,H_2}} = P_{7,H_2} - P_{6,H_2} = \langle 122880, -4055040\rangle.$$

They calculate a normal vector $\vec{N}_{H_2}$ for $H_2$:

$$\vec{N}_{H_2} = \det \begin{bmatrix} 122880 & -4055040 \\ \hat{e}_1 & \hat{e}_2 \end{bmatrix} = 4055040\hat{e}_1 + 122880\hat{e}_2 = 122880 \begin{bmatrix} 33 \\ 1 \end{bmatrix}. \tag{9-6}$$

Dividing by 122880 reveals $\vec{n}_{H_2}$. The sergeants define $\vec{P}_B$ as a row vector whose components are those of $P_{6,H_2}$, the base point of $\overrightarrow{P_{6,H_2}P_{7,H_2}}$. Then, using (5-2) they find the equation of $H_2$ (9-5). They find the $x_1$-intercept of $H_2$, which reveals $d_{H_2} = 7229611$. Now they know $P_{H_1} = (2, 1, 0, 0, 0, 7229611)$ lies on $H_1$.

Using their two vectors on $H_1$, they collaborate with the three corporals to consider five vectors parallel to $H_1$: $\vec{v}_{3,H_1}, \vec{v}_{4,H_1}, \vec{v}_{5,H_1}, \vec{v}_{6,H_1}, \vec{v}_{7,H_1}$. They apply (5-1) to calculate $\vec{N}_{H_1}$:

$$\vec{N}_{H_1} = \det \begin{bmatrix} & \vec{v}_{3,H_1} & & & & \\ & \vec{v}_{4,H_1} & & & & \\ & \vec{v}_{5,H_1} & & & & \\ & \vec{v}_{6,H_1} & & & & \\ & \vec{v}_{7,H_1} & & & & \\ \hat{e}_1 & \hat{e}_2 & \hat{e}_3 & \hat{e}_4 & \hat{e}_5 & \hat{e}_6 \end{bmatrix}$$

$$= \begin{bmatrix} 10039064001364120043520 \\ 27665792654922051747840 \\ 18210395165265147985920 \\ 18443861769948034498560 \\ 840479776858391445504 0 \\ 11673330234144325632 0 \end{bmatrix} = 116733302341443256320 \begin{bmatrix} 86 \\ 237 \\ 156 \\ 158 \\ 72 \\ 1 \end{bmatrix}.$$

Dividing by 116733302341443256320 reveals $\vec{n}_{H_1}$. Then, the sergeants and corporals collectively define $\vec{P}_B$ as a row vector with the components of $P_{H_1}$ and apply (5-2) to discover the equation for $H_1$ (9-4). Solving for its $x_1$-intercept reveals $d_{H_1} = 84070$. They now know that $P_{H_0} = (2, 1, 0, 0, 0, 0, 84070)$ lies on $H_0$.

They collaborate with *Private* Bob to consider six vectors parallel to $H_0$:

$$\vec{v}_{2,H_0}, \quad \vec{v}_{3,H_0}, \quad \vec{v}_{4,H_0}, \quad \vec{v}_{5,H_0}, \quad \vec{v}_{6,H_0}, \quad \vec{v}_{7,H_0}.$$

They apply (5-1) to find $\vec{N}_{H_0}$, a normal vector for $H_0$. They divide this vector by its last component to find $\vec{n}_{H_0}$. The participants define $\vec{P}_B$ as a row vector with the components of $P_{H_0}$ and apply (5-2) to find the equation for $H_0$ (9-1). Finally, they solve this for its $x_1$-intercept, revealing the secret $d = 331$.

## 10. Eliminating a potential problem

If two points were assigned on a hyperplane such that they are collinear with the $x_1$-intercept when $k > 2$, it would be possible for the security of the scheme to be compromised. A line through these points, intersecting the $x_1$-intercept, would reveal the secret preemptively. However, it can be proven algebraically that this is impossible using the method presented (4-8) for point assignment when $k > 3$. For the case $k = 3$, the dealer can make sure that this does not happen by a careful assignment of shares.

Case 1: Assume that $k = 3$. Note that for $p_i \in \mathbb{Z}_+$ we have $(d/dt)t^{p_i} = p_i t^{p_i - 1}$. Without loss of generality we assume $p_2 > p_1$ and that the dealer chooses these

values so that $p_1(b/a_1)$ is not divisible by $t$. Define three points as follows:

$$A = (t^{p_1}, p_1 t^{p_1-1}, b - a_1 t^{p_1} - a_2 p_1 t^{p_1-1}),$$
$$B = (t^{p_2}, p_2 t^{p_2-1}, b - a_1 t^{p_2} - a_2 p_2 t^{p_2-1}),$$
$$C = (b/a_1, 0, 0).$$

Assume $A$, $B$, $C$ are distinct, collinear points. Then $\overrightarrow{CA} = u\overrightarrow{CB}$ for some $u \in \mathbb{R}\setminus\{0\}$. Let:

(I) $t^{p_1} - b/a_1 = u(t^{p_2} - b/a_1)$.

(II) $p_1 t^{p_1-1} = u p_2 t^{p_2-1}$.

(III) $b - a_1 t^{p_1} - a_2 p_1 t^{p_1-1} = u(b - a_1 t^{p_2} - a_2 p_2 t^{p_2-1})$.

From (I) and (II) and assuming $t \neq 0$,

$$u(t^{p_2} - b/a_1)(p_2 t^{p_2-1})t = (t^{p_1} - b/a_1)(p_2 t^{p_2-1})t = (t^{p_2} - b/a_1)(p_1 t^{p_1-1})t,$$

which gives

$$(t^{p_1} - b/a_1)(p_2 t^{p_2}) = (t^{p_2} - b/a_1)(p_1 t^{p_1}),$$

implying

$$p_2 t^{p_1+p_2} - (b/a_1) p_2 t^{p_2} = p_1 t^{p_1+p_2} - (b/a_1) p_1 t^{p_1}.$$

Add $(b/a_1) p_2 t^{p_2} - p_1 t^{p_1+p_2}$ to both sides and factor out the common factors:

$$t^{p_1+p_2}(p_2 - p_1) = (b/a_1)(p_2 t^{p_2} - p_1 t^{p_1}).$$

Canceling out $t^{p_1}$ from both sides we get

$$t^{p_2}(p_2 - p_1) - (b/a_1) p_2 t^{p_2-p_1} = -(b/a_1) p_1.$$

This is a contradiction because the left-hand side is divisible by $t$, whereas the right-hand side is not.

Case 2: Assume that $k > 3$. Then $p_i \geq 2$ implies

$$\frac{d}{dt} t^{p_i} = p_i t^{p_i-1}, \quad \frac{d^2}{dt^2} t^{p_i} = p_i(p_i - 1) t^{p_i-2}.$$

Define three points as follows such that $p_1 \neq p_2$ and each $x_i \in \mathbb{R}$:

$$A = \left(t^{p_1}, p_1 t^{p_1-1}, p_1(p_1-1)t^{p_1-2}, \ldots, b - \left(a_1 t^{p_1} + \cdots + a_{k-1} \frac{d}{dt^{k-2}} t^{p_1}\right)\right),$$
$$B = \left(t^{p_2}, p_1 t^{p_2-1}, p_2(p_2-1)t^{p_2-2}, \ldots, b - \left(a_1 t^{p_2} + \cdots + a_{k-1} \frac{d}{dt^{k-2}} t^{p_2}\right)\right),$$
$$C = (x_1, 0, 0, x_4, \ldots, x_k).$$

Assume $A$, $B$, $C$ are distinct, collinear points. Note that the $x_1$-intercept is a point of the form $C$. Consider points on the line through $A$ and $B$. For some $u_0 \in \mathbb{R}$, since its second and third components are zero, point $C$ implies

$$p_1 t^{p_1-1} + (p_2 t^{p_2-1} - p_1 t^{p_1-1})u_0 = 0,$$
$$p_1(p_1-1)t^{p_1-2} + (p_2(p_2-1)t^{p_2-2} - p_1(p_1-1)t^{p_1-2})u_0 = 0,$$

from which we get

$$\begin{aligned} u_0(p_2 t^{p_2-1} - p_1 t^{p_1-1})(p_2(p_2-1)t^{p_2-2} - p_1(p_1-1)t^{p_1-2}) \\ = -p_1 t^{p_1-1}(p_2(p_2-1)t^{p_2-2} - p_1(p_1-1)t^{p_1-2}) \\ = -p_1(p_1-1)t^{p_1-2}(p_2 t^{p_2-1} - p_1 t^{p_1-1}). \end{aligned}$$

Thus,

$$\begin{aligned} & t^{p_1-1}(p_2(p_2-1)t^{p_2-2} - p_1(p_1-1)t^{p_1-2}) = (p_1-1)t^{p_1-2}(p_2 t^{p_2-1} - p_1 t^{p_1-1}) \\ \Longrightarrow\ & p_2(p_2-1)t^{p_1+p_2-3} - p_1(p_1-1)t^{2p_1-3} = (p_1-1)(p_2 t^{p_1+p_2-3} - p_1 t^{2p_1-3}) \\ \Longrightarrow\ & p_2(p_2-1)t^{p_1+p_2-3} - p_1(p_1-1)t^{2p_1-3} = p_2(p_1-1)t^{p_1+p_2-3} - p_1(p_1-1)t^{2p_1-3} \\ \Longrightarrow\ & p_2(p_2-1) = p_2(p_1-1) \\ \Longrightarrow\ & p_1 = p_2. \end{aligned}$$

This is a contradiction, since $p_1 \neq p_2$.

## 11. Conclusion

Secret-sharing schemes are great examples of how elementary mathematical ideas can be used to construct systems that can have profound application areas. Since their first appearance, they have been studied extensively from many directions, with the two main goals being constructing new schemes and finding application areas for the existing ones. In this paper we were able to use elementary mathematical ideas from an undergraduate curriculum to construct generalized secret-sharing schemes including the multilevel hierarchical ones. Our approach and the tools we have used are accessible to students and readers with limited technical background and they differ from the more complex approaches that have been used in the literature.

A detailed security cryptanalysis of the scheme can be done as part of a future research project. Due to the similarity of the tools used, we expect the security of the scheme to be similar to Shamir's scheme. Several attacks are possible, especially when the secret is chosen to be a small number. We propose a few countermeasures to the attacks. The first one is an obvious one, that is, choosing a large number and making the scheme more complex (with a complex hierarchy and distribution of shares if possible). Our second approach is inspired by code-based schemes. Instead of letting the secret be a single number, we could let $S = (S_1, S_2, \ldots, S_n)$

be the secret; i.e., the secret is an ordered $n$-tuple of *subsecrets*, where for each subsecret we apply the scheme to define shares. While this would increase the computational complexity of the scheme, it will be much more secure because to find the secret, every coordinate has to be found correctly. The probability of a successful attack will approach zero by increasing $n$. A possible direction for future research is to find practical areas in cryptography and information security in which our schemes can be implemented together with a detailed security and computational analysis.

## Acknowledgement

The authors would like to thank the anonymous referees for their valuable comments and suggestions that helped improve the paper profoundly.

## References

[Benaloh and Leichter 1990] J. Benaloh and J. Leichter, "Generalized secret sharing and monotone functions", pp. 27–35 in *Advances in cryptology—CRYPTO '88* (Santa Barbara, CA, 1988), edited by S. Goldwasser, Lecture Notes in Comput. Sci. **403**, Springer, 1990. MR Zbl

[Beutelspacher and Vedder 1989] A. Beutelspacher and K. Vedder, "Geometric structures as threshold schemes", pp. 255–268 in *Cryptography and coding* (Cirencester, 1986), edited by H. J. Beker and F. C. Piper, Inst. Math. Appl. Conf. Ser. New Ser. **20**, Oxford Univ. Press, 1989. MR

[Blakley 1979] G. Blakley, "Safeguarding cryptographic keys", pp. 313–318 in *International workshop on managing requirements knowledge* (New York, 1979), AFIPS Conference Proceedings **48**, AFIPS, Montvale, NJ, 1979.

[Brickell 1989] E. F. Brickell, "Some ideal secret sharing schemes", *J. Combin. Math. Combin. Comput.* **6** (1989), 105–113. MR Zbl

[Charnes et al. 1997] C. Charnes, K. Martin, J. Pieprzyk, and R. Safavi-Naini, "Secret sharing in hierarchical groups", pp. 81–86 in *ICICS 1997: information and communications security*, Lecture Notes in Computer Science **1334**, Springer, 1997. Zbl

[Dawson and Donovan 1994] E. Dawson and D. Donovan, "The breadth of Shamir's secret-sharing scheme", *Comput. and Secur.* **13**:1 (1994), 69–78.

[Ito et al. 1989] M. Ito, A. Saito, and T. Nishizeki, "Secret sharing scheme realizing general access structure", *Electron. Comm. Japan Part III Fund. Electron. Sci.* **72**:9 (1989), 56–63. MR

[Krusemeyer 1988] M. Krusemeyer, "The teaching of mathematics: why does the Wronskian work?", *Amer. Math. Monthly* **95**:1 (1988), 46–49. MR Zbl

[Lai and Ding 2004] C.-P. Lai and C. Ding, "Several generalizations of Shamir's secret sharing scheme", *Internat. J. Found. Comput. Sci.* **15**:2 (2004), 445–458. MR Zbl

[Nijenhuis 1980] A. Nijenhuis, "Complete solutions of linear difference equations", *Amer. Math. Monthly* **87**:8 (1980), 658–660. MR Zbl

[Pólya 1922] G. Pólya, "On the mean-value theorem corresponding to a given linear homogeneous differential equation", *Trans. Amer. Math. Soc.* **24**:4 (1922), 312–324. MR

[Rushanan 1989] J. J. Rushanan, "On the Vandermonde matrix", *Amer. Math. Monthly* **96**:10 (1989), 921–924. MR Zbl

[Shamir 1979] A. Shamir, "How to share a secret", *Comm. ACM* **22**:11 (1979), 612–613. MR Zbl

[Simmons 1990] G. J. Simmons, "How to (really) share a secret", pp. 390–448 in *Advances in cryptology—CRYPTO '88* (Santa Barbara, CA, 1988), edited by S. Goldwasser, Lecture Notes in Comput. Sci. **403**, Springer, 1990. MR

[Simmons 1992] G. J. Simmons, "An introduction to shared secret and/or shared control schemes and their application", pp. 441–497 in *Contemporary cryptology*, edited by G. J. Simmons, IEEE, New York, 1992. MR

[Tassa 2007] T. Tassa, "Hierarchical threshold secret sharing", *J. Cryptology* **20**:2 (2007), 237–264. MR Zbl

[Zill and Cullen 2006] D. G. Zill and M. R. Cullen, *Advanced engineering mathematics*, 3rd ed., Jones & Bartlett, Sudbury, MA, 2006.

Received: 2019-04-19 Revised: 2019-12-16 Accepted: 2020-02-24

wloucks@gmail.com *Department of Mathematics and Statistics, Northern Arizona University, Flagstaff, AZ, United States*

bahattin.yildiz@nau.edu *Department of Mathematics and Statistics, Northern Arizona University, Flagstaff, AZ, United States*

INVOLVE 13:2 (2020)
dx.doi.org/10.2140/involve.2020.13.281

# Generalized Cantor functions: random function iteration

Jordan Armstrong and Lisbeth Schaubroeck

(Communicated by Michael Dorff)

We provide a generalization of the classical Cantor function. One characterization of the Cantor function is generated by a sequence of real numbers that starts with a seed value and at each step randomly applies one of two different linear functions. The Cantor function is defined as the probability that this sequence approaches infinity. We generalize the Cantor function to instead use a set of any number of linear functions with integer coefficients. We completely describe the resulting probability function and give a full explanation of which intervals of seed values lead to a constant probability function value.

## Generalizing the Cantor function

Most analysis books include a discussion of the Cantor set (or Cantor middle-thirds set) as a closed subset of the interval $[0, 1]$ whose complement has total length 1 but has an uncountably infinite number of points; see, for example, [Royden 1988]. The classical Cantor function is defined on the complement of the Cantor set. There are many different descriptions of the Cantor function; for a very complete overview, see [Dovgoshey et al. 2006]. In its simplest form, it can be described iteratively as the function $f(x)$ that has height $\frac{1}{2}$ on the interval $\left(\frac{1}{3}, \frac{2}{3}\right)$, height $\frac{1}{4}$ on the interval $\left(\frac{1}{9}, \frac{2}{9}\right)$ $\left(\text{which is the middle third of the interval } \left(0, \frac{1}{3}\right)\right)$, height $\frac{3}{4}$ on the middle-third of the interval $\left(\frac{2}{3}, 1\right)$, or equivalently height $\frac{3}{4}$ on interval $\left(\frac{7}{9}, \frac{8}{9}\right)$. Continue in this manner, defining the height of $f(x)$ on the middle-third of a remaining interval to be the average of the heights on either side. So for example, since $f(x)$ has a height of $\frac{1}{2}$ on $\left(\frac{1}{3}, \frac{2}{3}\right)$, and a height of $\frac{3}{4}$ on $\left(\frac{7}{9}, \frac{8}{9}\right)$, it must have a height of $\frac{5}{8}$ on the middle third of the interval $\left(\frac{2}{3}, \frac{7}{9}\right)$, or the interval $\left(\frac{19}{27}, \frac{20}{27}\right)$. Continue this process indefinitely to get a function that has steps of height $\gamma/2^k$ for any integer $\gamma < 2^k$ and $k \in \mathbb{N}$. See Figure 1 for a graph of the Cantor function.

*MSC2010:* 26A18.
*Keywords:* iteration, cantor function, sequence, devil's staircase.
Public affairs release number: USAFA-DF-2020-3.

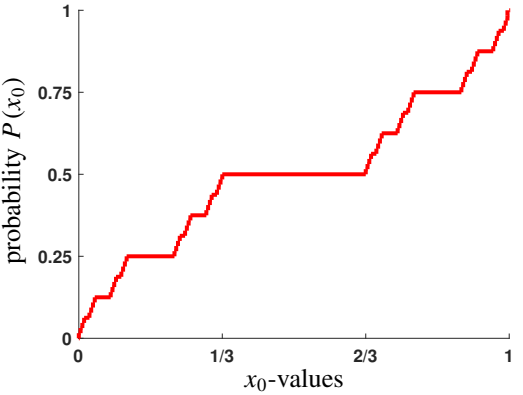
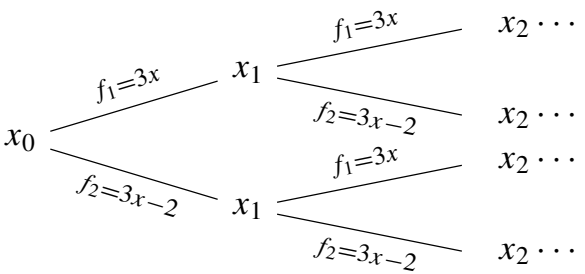



**Figure 1.** Original Cantor function from [Stankewitz and Rolf 2012] computed using $f_1(x) = 3x$ and $f_2(x) = 3x - 2$.

An alternative description of the Cantor function is found in [Stankewitz and Rolf 2012], in which the function arises from a random iterative process. We recall the fundamentals of single function iteration: Starting with a seed value $x_0$, compute $x_1 = f(x_0)$ and $x_2 = f(x_1) = f(f(x_0))$, and so forth, so that

$$x_n = \underbrace{f(f(\cdots f}_{n}(x_0))).$$

We can then examine the long-term behavior of the sequence $\{x_n\}$. To get to the description of the Cantor function found in [Stankewitz and Rolf 2012], we instead start with a seed value $x_0$, randomly choose a function from a finite set of functions, apply that function to $x_0$ to find $x_1$, and then repeat the process indefinitely. In particular, for the Cantor function, at each step we choose to compute either $f_1(x_j)$ or $f_2(x_j)$, with $f_1(x) = 3x$ and $f_2(x) = 3x - 2$. This is diagrammed in the figure below, where the tree would continue to the right:




Now we determine, out of all possible branches on the tree, what proportion of them will diverge to positive infinity. Equivalently, what is the probability of picking a sequence of functions such that $\{x_j\} \to \infty$? We describe this process formally.

**Definition 1** (Cantor function). Let $f_1(x) = 3x$ and $f_2(x) = 3x - 2$. Given a seed value $x_0$, let $x_1 = f_{k_1}(x_0)$, $x_2 = f_{k_2}(x_1)$, and so forth, so that $x_j = f_{k_j}(x_{j-1})$, where $k_j \in \{1, 2\}$ is chosen with equal probability. For each initial value $x_0$, we define $P(x_0)$ as the probability that the sequence $\{x_j\}$ diverges to $\infty$ using the seed value $x_0$.

In this work, we generalize Definition 1 to $n$ linear functions instead of the two functions $f_1(x) = 3x$ and $f_2(x) = 3x - 2$. We begin with a concrete example and the computation of a few values of $P(x_0)$.

**Example 2.** Consider the functions $f_1(x) = 20x - 1$, $f_2(x) = 6x - 3$, and $f_3(x) = 10x - 8$. We evaluate $P(x_0)$ for various values of $x_0$.

(a) Suppose $x_0 > 1$. Then $f_1(x_0) > 19$, $f_2(x_0) > 3$, and $f_3(x_0) > 2$. No matter which $f_k$ we choose, we have $x_1 = f_k(x_0) > 2$. At the next step, if $x_2 = f_1(x_1)$, then $x_2 > 20(2) - 1 = 39$, if $x_2 = f_2(x_1)$, then $x_2 > 6(2) - 3 = 9$, and if $x_2 = f_3(x_1)$, then $x_3 > 10(2) - 8 = 12$. Each successive $x_j$ is larger than $x_{j-1}$, thus $\{x_j\} \to \infty$ and $P(x_0) = 1$.

(b) Suppose $x_0 = 0.25$. If $x_1 = f_1(x_0) = 20x_0 - 1$, then $x_1 = 4 > 1$, and the rest of the sequence will diverge to infinity as in part (a) of this example. However, if $x_1 = f_2(x_0)$ then $x_1 = -1.5$ and if $x_1 = f_3(x_0)$ then $x_1 = -5.5$. We note that all of the functions $f_1$, $f_2$, and $f_3$ have a negative $y$-intercept and a positive slope, meaning that a negative input into each of these functions will result in a negative output. Thus if $x_1 = f_2(0.25)$ or if $x_1 = f_3(0.25)$, there is no way for the sequence $\{x_j\}$ to have any positive values after $x_0$, and thus the sequence cannot diverge to infinity. Combining the three different choices for $x_1$, we see that one of the choices leads to $\{x_j\} \to \infty$, while the other two do not, so $P(0.25) = \frac{1}{3}$.

For this set of functions, the graph of $P(x_0)$ is shown in Figure 2 (see next page). In the figure, we can see long intervals of constancy for $P(x_0)$ at heights of $\frac{1}{3}$ and $\frac{2}{3}$, and shorter intervals of constancy at heights of $\gamma/9$ for $\gamma \in \mathbb{N}$ and $\gamma < 9$.

We note that in Example 2, the three functions were chosen such that any $x_j > 1$ would result in that sequence diverging to infinity, and any $x_j < 0$ would result in the remainder of that sequence being negative. Thus the interval of interest for those three functions is the interval $[0, 1]$, since the probability function is $P(x_0) = 1$ for all $x_0 > 1$ and $P(x_0) = 0$ for all $x_0 < 0$. Furthermore, if the input of the function is $x_0 \in [0, 1]$, then the outputs are ordered such that $f_3(x_0) < f_2(x_0) < f_1(x_0)$, as can be seen in Example 2(b) with $x_0 = 0.25$. That is, there is some ordering to our functions over the interval $[0, 1]$. The following definition does not quite have the same consequence of having $f_3(x_0) < f_2(x_0) < f_1(x_0)$ for all $x_0 \in [0, 1]$, but applies a restriction on the functions that forces an order in which sequences involving the different functions diverge to infinity.

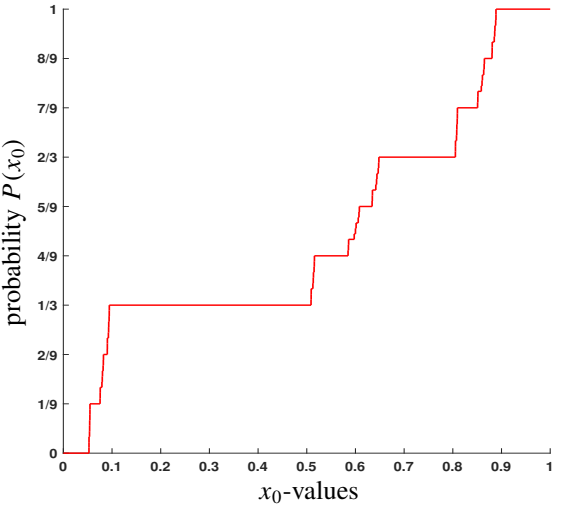



**Figure 2.** $P(x_0)$ over the interval $[0, 1]$ given $f_1(x) = 20x - 1$, $f_2(x) = 6x - 3$, and $f_3(x) = 10x - 8$.

**Definition 3.** Define the functions $f_k(x) = a_k x - b_k$ for $k \in \{1, 2, \ldots, n\}$, with $a_k, b_k \in \mathbb{N} \cup \{0\}$, $a_k > b_k$, and

$$\frac{b_k + 1}{a_k} < \frac{b_{k+1}}{a_{k+1}} \tag{1}$$

for all $k \in \{1, 2, \ldots, n-1\}$.

In much of the work that follows, the notions of preimages and of fixed points will be critical.

**Definition 4.** Let $f_k$ be a function as in Definition 3:

(a) The *preimage* of a number $y \in \mathbb{R}$ is the value $x$ such that $f_k(x) = y$, and is denoted by $x = f_k^{-1}(y)$.

(b) For a given function $f_k(x)$, a *fixed point* of the function is a point $p_k$ such that $f(p_k) = p_k$.

The following lemma formalizes the consequences of the restrictions on $a_k$ and $b_k$ in Definition 3.

**Lemma 5.** *Let $f_k$ be given as in Definition 3. Then $f_k^{-1}(1) < f_{k+1}^{-1}(0)$ and consequently, if $0 < f_k(x) < 1$ then $f_{k+1}(x) < 0 < f_k(x) < 1 < f_{k-1}(x)$.*

*Proof.* Given $f_k(x) = a_k x - b_k$, we compute $f_k^{-1}(0)$ by solving $0 = a_k x - b_k$ to get $x = b_k/a_k$. Thus $f_k^{-1}(0) = b_k/a_k$. A similar computation where we solve $1 = a_k x - b_k$ gives $f_k^{-1}(1) = (b_k + 1)/a_k$. Thus condition (1) ensures

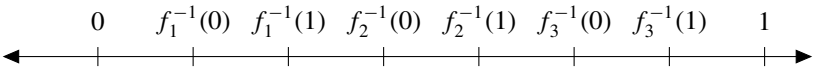



**Figure 3.** The ordering the preimages of 0 and 1 for three functions in Definition 3.

that $f_k^{-1}(1) < f_{k+1}^{-1}(0)$. Figure 3 illustrates these observations for three functions.

The condition that $0 < f_k(x) < 1$ is equivalent to $f_k^{-1}(0) < x < f_k^{-1}(1)$. Furthermore, the condition that $f_k^{-1}(1) < f_{k+1}^{-1}(0)$ ensures that

$$f_{k-1}^{-1}(1) < f_k^{-1}(0) < x < f_k^{-1}(1) < f_{k+1}^{-1}(0).$$

The left-hand part of the equation ensures that $f_{k-1}^{-1}(1) < x$, or equivalently, that $1 < f_{k-1}(x)$. The right-hand part of the equation ensures that $x < f_{k+1}^{-1}(0)$ or that $f_{k+1}(x) < 0$ as desired. □

The following definition mimics the definition of the Cantor function, but gives the definition of a sequence and a probability function $P(x_0)$ for the more general case of $n$ functions described in Definition 3.

**Definition 6.** Let $f_k$ be given as in Definition 3. Given a seed value $x_0$, let $x_1 = f_{k_1}(x_0)$, $x_2 = f_{k_2}(x_1)$, and so forth, so that $x_j = f_{k_j}(x_{j-1})$, where $k_j \in \{1, 2, \ldots, n\}$ is chosen with equal probability. For each initial value $x_0$, we define $P(x_0)$ as the probability that the sequence $\{x_j\}$ diverges to $\infty$ using the seed value $x_0$.

The following lemma formalizes the behavior of $P(x_0)$ as seen in the examples above. In particular, part (c) of Lemma 7 is particularly useful when we want to find $P(x_0)$, since it permits us to find the largest $m$ for which the sequence $\{f_m(x_{j-1})\}$ diverges to infinity, and then draw conclusions about the behavior of sequences where the index of the function is restricted to the set $\{1, 2, \ldots, m\}$.

**Lemma 7.** *Let $f_k$ be given as in Definition 3 and the sequence $\{x_j\}$ be given as in Definition 6. Then the following are true*:

(a) *If $x_0 \le 0$, then $P(x_0) = 0$.*

(b) *If $x_0 \ge 1$, then $P(x_0) = 1$.*

(c) *Suppose a seed value $x_0$ has the property that if the sequence $\{x_j\}$ is formed by $x_j = f_m(x_{j-1})$ for a fixed value of $m$, then $\lim_{j\to\infty} x_j = \infty$. Then any sequence $\{x_j\}$ formed by $x_j = f_{k_j}(x_{j-1})$, where $k_j \in \{1, 2, \ldots, m\}$, will also have $\lim_{j\to\infty} x_j = \infty$.*

*Proof.* We prove each part individually.

Proof of (a): Suppose $x_0 \le 0$. By Definition 3, $x_1 \le -b_k < 0$, so any sequence $\{x_j\}$ will have all negative terms. Thus $P(x_0) = 0$.

Proof of (b): Suppose $x_0 \geq 1$. By definition, $x_1 \geq a_{k_1} - b_{k_1} \geq 1$. Applying Definition 3 gives $x_j > 1$ for all $j$. We now show that $x_{j-1} < x_j$, or equivalently that $x_{j-1} < a_{k_j} x_{j-1} - b_{k_j}$. By way of contradiction, suppose instead that $x_{j-1} \geq a_{k_j} x_{j-1} - b_{k_j}$, which is equivalent to $x_{j-1} \leq b_{k_j}/(a_{k_j} - 1) < 1$, and this contradicts the assumption of $x_{j-1} \geq 1$. Thus the sequence $\{x_j\}$ is increasing. Since the functions are linear with positive integer slope, the sequence increases without bound, and $P(x_0) = 1$.

Proof of (c): Suppose that $x_0$ has the property that if the sequence $\{x_j\}$ is formed by $x_j = f_m(x_{j-1})$ for a fixed value of $m$, then $\lim_{j\to\infty} x_j = \infty$. By part (a), we know that all terms of the sequence are positive. Since $f_m(x_0) > 0$, we have $f_k(x_0) > 1$ for all $k \in \{1, 2, \ldots, m-1\}$ by Lemma 5. Thus any sequence $\{x_j\} = \{f_{k_j}(x_{j-1})\}$ with $k_j \in \{1, 2, \ldots, m\}$ would have $\lim_{j\to\infty} x_j = \infty$. □

Finally, the notion of a fixed point in Definition 4 will be important in determining for which values of $x_0$ the function $P(x_0)$ will have intervals of constancy, as seen in Figures 1 and 2. Since the fixed points of $f_k$ result in the output being equal to the input, if $x_{j-1} = p_k$, then $x_j = p_k$ as long as the same function has been chosen for the next iteration. Thus if we had only one function $f_k$ in Definition 6, then for $x_{j-1} > p_k$ we would have $x_j > x_{j-1}$ and thus $P(x_0) = 1$ (because the $x_j$ sequence relies on linear functions with integer coefficients, so if it is increasing it must be increasing without bound); similarly for $x_{j-1} < p_k$ we would have $x_j < x_{j-1}$ and thus $P(x_0) = 0$. Of course, our situation is more complicated, since we will randomly choose from among $n$ different functions at each $x_{j-1}$.

**Example 8.** Consider the same functions from Example 2:

$$f_1(x) = 20x - 1, \quad f_2(x) = 6x - 3, \quad f_3(x) = 3x - 2.$$

We calculate their fixed points. Since the general form of the fixed points for the functions in Definition 3 is $p_k = b_k/(a_k - 1)$, we have $p_1 = \frac{1}{19}$, $p_2 = \frac{3}{5}$, and $p_3 = \frac{8}{9}$. We also note that in Figure 2, the step at height 0 ends at the fixed point $p_1 = \frac{1}{19}$, and the step at height 1 begins at $p_3 = \frac{8}{9}$. As we shall see in Theorems 11 and 12, this is not a coincidence.

Figure 4 illustrates how the values of $p_k$ are related to $f_k^{-1}(0)$ and $f_k^{-1}(1)$ for three functions. Since each $f_k$ is an increasing function and $f_k^{-1}(p_k) = p_k$, we see that $f_k^{-1}(0) < p_k < f_k^{-1}(1)$.

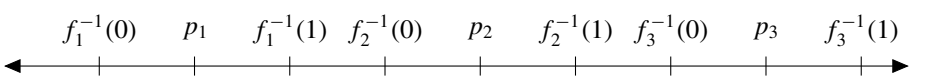



**Figure 4.** The ordering of the preimages of 0 and 1 and fixed points for three functions.

## Probabilities for three functions

In this section, we completely describe the function $P(x_0)$ when $n=3$ in Definition 3. We first prove that the function $P(x_0)$ is nondecreasing. This proof does not rely on the number of functions and is valid for any set of $n$ functions.

**Theorem 9.** *The function $P(x_0)$ given in Definition 6 is nondecreasing. That is, if $y_0 < z_0$, then $P(y_0) \le P(z_0)$.*

*Proof.* By Lemma 7, we only need to consider the case where $0 < y_0 < z_0 < 1$. Consider a particular sequence of functions $\{f_{k_j}\}$ for a fixed sequence $\{k_j\}$ with each $k_j \in \{1, 2, \ldots, n\}$. Let the sequence given in Definition 6 be identified as $\{y_j\}$ if the seed value is $y_0$ and $\{z_j\}$ if the seed value is $z_0$. If $\lim_{j\to\infty}\{y_j\} = \infty$, then the same is true for $\lim_{j\to\infty}\{z_j\}$ by Lemma 7. If $\lim_{j\to\infty}\{y_j\} \neq \infty$, we only need to consider the case where $\lim_{j\to\infty}\{z_j\} \neq \infty$ (because if $\lim_{j\to\infty}\{z_j\} = \infty$, then $\lim_{j\to\infty}\{y_j\} \le \lim_{j\to\infty}\{z_j\}$). In the case where both limits are finite, we know via Definition 3 that $f_k(y_0) < f_k(z_0)$ for all $f_k$. Thus in this case, we have that $\lim_{j\to\infty}\{y_j\} \le \lim_{j\to\infty}\{z_j\}$, because limits preserve inequalities.

For any one fixed sequence of functions $\{f_{k_j}\}$, we have that $\lim_{j\to\infty}\{y_j\} \le \lim_{j\to\infty}\{z_j\}$. This is true for all sequences of functions; thus $P(y_0) \le P(z_0)$. □

The following theorem provides the foundation for the results in Theorem 13.

**Theorem 10.** *Consider three linear functions $f_k$ defined by Definition 3 and $P(x_0)$ defined by Definition 6*:

(a) *For $x_0 \in \mathbb{R}$, $P(x_0) = \frac{1}{3}$ if and only if $x_0 \in I_1 = [f_1^{-1}(p_3),\ f_2^{-1}(p_1)]$.*

(b) *For $x_0 \in \mathbb{R}$, $P(x_0) = \frac{2}{3}$ if and only if $x_0 \in I_2 = [f_2^{-1}(p_3),\ f_3^{-1}(p_1)]$.*

We will prove part (a), as the proof of part (b) is similar.

*Proof.* Let $x_0 \in \mathbb{R}$. First, suppose $x_0 \in I_1 = [f_1^{-1}(p_3),\ f_2^{-1}(p_1)]$. We consider three cases.

Case 1: $x_1 = f_1(x_0)$. Consider the lower bound of $I_1$, which is $x_0 \ge f_1^{-1}(p_3)$. Thus,

$$x_1 = f_1(x_0) \ge f_1(f_1^{-1}(p_3)) = p_3.$$

Since all elements in $f_1(I_1)$ are greater than or equal to $p_3$, consider the sequence defined by $x_j = f_3(x_{j-1})$. When $x_{j-1} > p_3$, we have $x_j = f_3(x_{j-1}) > x_{j-1}$, and the sequence increases without bound due to the restrictions on $a_k$ and $b_k$. By Lemma 7, if the sequence is defined by $x_j = f_{k_j}(x_{j-1})$, with $x_j \in \{1, 2, 3\}$, this sequence also increases without bound and $P(x_0) = 1$. In the particular case where $x_0 = f_1^{-1}(p_3)$ and $x_1 = f_1(x_0) = p_3$, the only sequence that does not diverge to infinity is the one where $x_j = f_3(x_{j-1})$ for all $j \ge 2$. But if at any point one element of the sequence is $x_j = f_2(x_{j-1})$ or $x_j = f_1(x_{j-1})$, then $x_j > p_3$ and the sequence will diverge to

infinity. But the sequence where $x_1 = p_3$ and $x_j = f_3(x_{j-1})$ for all $j \geq 2$ occurs with 0 probability, so $P(x_0) = 1$ if $x_0 = f_1^{-1}(p_3)$ and $x_1 = f_1(x_0)$.

Case 2: $x_1 = f_2(x_0)$. Consider the upper bound of $I_1$, that is, $x_0 \leq f_2^{-1}(p_1)$. We have

$$f_2(x_0) \leq f_2(f_2^{-1}(p_1)) = p_1.$$

Since all elements of the interval $f_2(I_1)$ are less than or equal to $p_1$, we know that if $x_2 = f_1(x_1)$, the sequence $\{x_j\}$ cannot diverge to infinity. By Lemma 5, we know that $f_2(x_1)$ and $f_3(x_1)$ are negative, so no sequence of functions will cause $\{x_j\}$ to diverge to infinity. Thus $P(x_0) = 0$ for this case.

Case 3: $x_1 = f_3(x_0)$. Since the probability of the sequence $\{x_j\}$ diverging to infinity is 0 for any element of $I_1$ if $x_1 = f_2(x_0)$, the same is true for $x_1 = f_3(x_0)$ by Lemma 7. Thus $P(x_0) = 0$ in this case.

Since each of the cases has a probability of $\frac{1}{3}$ of occurring, the resultant probability is $\frac{1}{3}(1 + 0 + 0) = \frac{1}{3}$. We conclude that if $x_0 \in I_1 = [f_1^{-1}(p_3), f_2^{-1}(p_1)]$ then $P(x_0) = \frac{1}{3}$.

To show that if $P(x_0) = \frac{1}{3}$ then $x_0 \in I_1$, we instead show the contrapositive. To that end, we consider $x_0 \notin I_1$, so $x_0 < f_1^{-1}(p_3)$ or $x_0 > f_2^{-1}(p_1)$.

Case 1: $0 < x_0 < f_1^{-1}(p_3)$. For $0 < x_0 < f_1^{-1}(p_3)$, we know that if $x_1 = f_2(x_0)$ or $x_1 = f_3(x_0)$, then the limit of the sequence $\{x_j\}$ is finite, by Cases 2 and 3 above, and $P(x_0) = 0$ for those two cases. Now if $x_1 = f_1(x_0)$, we have that $f_1(0) < x_1 < f_1(f_1^{-1}(p_3)) = p_3$.

If the rest of the sequence is defined by $x_j = f_3(x_{j-1})$ for $j \geq 2$, then the sequence would be bounded above by $p_3$. Thus $P(x_0) < 1$ in this case. Putting together the three different scenarios for $x_1$, we see that $P(x_0) < \frac{1}{3}(1 + 0 + 0)$, so $P(x_0) < \frac{1}{3}$.

Case 2: $f_2^{-1}(p_1) < x_0 \leq 1$. We first recall that by Definitions 3 and 4 and Lemma 5, we have

$$f_1^{-1}(0) \leq p_1 \leq f_1^{-1}(1) < f_2^{-1}(0) \leq p_2 \leq f_2^{-1}(1) < f_3^{-1}(0) \leq p_3 \leq f_3^{-1}(1), \quad (2)$$

as illustrated in Figure 4.

If $f_2^{-1}(p_1) < x_0 \leq 1$, we consider the three different possibilities for $x_1$. If $x_1 = f_1(x_0)$, then

$$x_1 > f_1(f_2^{-1}(p_1)) \geq f_1(f_2^{-1}(0)) > f_1(f_1^{-1}(1)) = 1,$$

so by Lemma 7, $\{x_j\} \to \infty$. If $x_1 = f_2(x_0) > f_2(f_2^{-1}(p_1)) = p_1$, then $x_1 > p_1$. If $x_2 = f_1(x_1)$, then $x_2 > f_1(p_1) = p_1$ and the function is increasing; hence the sequence would diverge to $\infty$ if we continued to apply $f_1$. Thus $P(x_0)$ is strictly positive. So we have determined that for one of the cases $P(x_0) = 1$ and for one of the cases $P(x_0) > 0$. Regardless of the outcome of the third case (where $x_1 = f_3(x_0)$),

we know that $P(x_0) > \frac{1}{3}(1+0)$. Therefore if $x_0 > f_2^{-1}(p_1)$ then $P(x_0) > \frac{1}{3}$. We have shown the contrapositive is true, so if $P(x_0) = \frac{1}{3}$, then $x_0 \in I_1$, as desired.

The proof of part (b) of the theorem uses similar reasoning to that in part (a). □

The following two theorems combine to prove for which $x_0$-values $P(x_0)$ is either 0 or 1.

**Theorem 11.** *Consider three linear functions $f_k$ defined by Definition 3 and $P(x_0)$ defined by Definition 6. For $x_0 \in \mathbb{R}$, $P(x_0) = 0$ if and only if $x_0 \leq p_1$.*

*Proof.* First, assume that $x_0 \leq p_1$. We note that by (2),

$$f_2(x_0) \leq f_2(p_1) \leq f_2(f_1^{-1}(1)) < f_2(f_2^{-1}(0)) = 0.$$

The same would be true if we replaced $f_2$ with $f_3$, so by Lemma 7 we have $P(x_0) = 0$ in these cases. So we only need to consider what happens if $x_1 = f_1(x_0)$. Since $f_1$ is an increasing function, we have $f_1(x_0) \leq f_1(p_1) = p_1$. Thus all terms in the sequence $\{x_j\}$ are bounded above by $p_1$, so $P(x_0) = 0$.

To show that if $P(x_0) = 0$ then $x_0 \leq p_1$, we instead show the contrapositive. Suppose that $x_0 > p_1$. The sequence $\{f_1(x_{j-1})\}$ will have the property that $x_j > x_{j-1}$ for all $j$. Since $f_1(x)$ is a linear function with positive integer slope, the fact that $\{x_j\}$ is increasing and positive means that it is increasing without bound, so $P(x_0) > 0$. Thus if $P(x_0) = 0$, we know that $x_0 \leq p_1$. □

**Theorem 12.** *Consider three linear functions $f_k$ defined by Definition 3 and $P(x_0)$ defined by Definition 6. For $x_0 \in \mathbb{R}$, $P(x_0) = 1$ if and only if $x_0 \geq p_3$.*

*Proof.* We first assume that $x_0 \geq p_3$. Consider the sequence $\{x_j\}$ defined by $x_j = f_3(x_{j-1})$ for all $j \geq 1$. Since $f_3(x_0) \geq f_3(p_3) = p_3$, we know that each term of the sequence is greater than $p_3$, except in the situation where $x_0 = p_3$. Furthermore, by the restrictions of the coefficients of $f_3$, the sequence $\{x_j\}$ increases without bound. By Lemma 7, any sequence with initial value $x_0$ will increase without bound since $\{f_3(x_j)\}$ is unbounded so $P(x_0) = 1$. If $x_0 = p_3$, the sequence $\{f_3(x_{j-1})\}$ will remain as $p_3$, but if any one of the $f_{k_j}(x_{j-1})$ of the sequence is $f_2$ or $f_1$, the sequence will once again increase without bound. Since the probability of the sequence of functions having all $k_j = 3$ is 0, $P(x_0) = 1$ when $x_0 = p_3$.

Now we prove the other direction of the biconditional by proving the contrapositive. Suppose that $x_0 < p_3$ and write $x_0 = p_3 - \varepsilon$, where $\varepsilon > 0$. We consider the sequence of $\{x_j\}$ where the first several terms are found by applying $f_3$ to the previous term. Observe that

$$x_1 = f_3(x_0) = a_3(p_3 - \varepsilon) - b_3 = p_3 - a_3\varepsilon,$$
$$x_2 = a_3(p_3 - a_3\varepsilon) - b_3 = p_3 - a_3^2\varepsilon.$$

The general term is $x_k = p_3 - a_3^k \varepsilon$, which will be negative for a finite value of $k$. Such a sequence will be bounded above, and so will not diverge to infinity. These bounded sequences occur with nonzero probability; thus $P(x_0) < 1$. □

Now that we have established for which intervals the probability $P(x_0) \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}$, we provide a theorem which demonstrates how to recursively find intervals for where $P(x_0) = \gamma/3^k$ for all nonnegative integer choices of $\gamma$ and $k$.

**Theorem 13.** *Consider three linear functions $f_k$ defined by Definition 3 and $P(x_0)$ defined by Definition 6. Assume*

$$J = \left\{ x : P(x_0) = \frac{\gamma}{3^k} \right\},$$

*where $\gamma \in \mathbb{N} \cup \{0\}$ and $3 \nmid \gamma$. Then*

$$\left\{ y_0 : P(y_0) = \frac{\alpha}{3} + \frac{\gamma}{3^{k+1}} \right\} = \begin{cases} f_1^{-1}(J), & \alpha = 0, \\ f_2^{-1}(J), & \alpha = 1, \\ f_3^{-1}(J), & \alpha = 2. \end{cases}$$

**Example 14.** Suppose we wanted to find which $x_0$-values give rise to $P(x_0) = \frac{52}{81}$. We write

$$\begin{aligned} \tfrac{52}{81} &= \tfrac{1(27)}{81} + \tfrac{2(9)}{81} + \tfrac{2(3)}{81} + \tfrac{1(1)}{81} = \tfrac{1}{3} + \tfrac{25}{81} \\ &= \tfrac{1}{3} + \tfrac{1}{3}\left(\tfrac{25}{27}\right) = \tfrac{1}{3} + \tfrac{1}{3}\left(\tfrac{2}{3} + \tfrac{7}{27}\right) \\ &= \tfrac{1}{3} + \tfrac{1}{3}\left(\tfrac{2}{3} + \tfrac{1}{3}\left(\tfrac{2}{3} + \tfrac{1}{3} \cdot \tfrac{1}{3}\right)\right). \end{aligned}$$

Then $P(x_0) = \frac{52}{81}$ if and only if $x_0 \in f_2^{-1}(f_3^{-1}(f_3^{-1}(I_1)))$, where we have built the interval recursively from $I_1$ via Theorem 13.

*Proof.* We assume that $J$ is as stated in Theorem 13. We need to show

$$\left\{ y_0 : P(y_0) = \frac{\alpha}{3} + \frac{\gamma}{3^{k+1}} \right\} = \begin{cases} f_1^{-1}(J), & \alpha = 0, \\ f_2^{-1}(J), & \alpha = 1, \\ f_3^{-1}(J), & \alpha = 2. \end{cases}$$

We start with the set containment of

$$\left\{ y_0 : P(y_0) = \frac{\alpha}{3} + \frac{\gamma}{3^{k+1}} \right\} \supseteq \begin{cases} f_1^{-1}(J), & \alpha = 0, \\ f_2^{-1}(J), & \alpha = 1, \\ f_3^{-1}(J), & \alpha = 2. \end{cases}$$

We proceed with proof by cases.

Case 1: $x_0 \in f_1^{-1}(J)$. Since $x_0 \in f_1^{-1}(J)$ and $J \subseteq [0, 1]$, we have

$$x_0 \in f_1^{-1}(J) \subseteq f_1^{-1}([0, 1]) \subseteq [0, 1].$$

Now $x_1 = f_\beta(x_0)$ for $\beta \in \{1, 2, 3\}$. We examine each $\beta$-value in subcases.

Subcase 1.1: $x_1 = f_1(x_0)$. Since $x_0 \in f_1^{-1}(J)$, we apply $f_1$ to both sides to get $x_1 = f_1(x_0) \in f_1(f_1^{-1}(J)) = J$. Since $x_1 \in J$, we have $P(x_1) = P(y)$ for any $y \in J$; thus $P(x_1) = \gamma/3^k$.

Subcase 1.2: $x_1 = f_2(x_0)$. Since $x_0$ is bounded above by $f_1^{-1}(1)$, we have $x_1 = f_2(x_0) < f_2(f_1^{-1}(1))$. By Lemma 5, $f_1^{-1}(1) < f_2^{-1}(0)$, so we conclude that

$$x_1 = f_2(x_0) < f_2(f_1^{-1}(1)) < f_2(f_2^{-1}(0)) = 0.$$

Since $x_1 < 0$, we conclude by Lemma 7 that the probability of the sequence $\{x_j\}$ diverging to $\infty$ is 0.

Subcase 1.3: $x_1 = f_3(x_0)$. As in Subcase 1.2, we conclude that

$$x_1 = f_3(x_0) < f_3(f_1^{-1}(1)) < f_3(f_3^{-1}(0)) = 0,$$

and the probability that $\{x_j\} \to \infty$ is 0.

We determined the probability that $\{x_j\} \to \infty$ for each of the three possible values of $x_1$, and we know each of these values has a probability of $\frac{1}{3}$ of being chosen. The value of $P(x_0)$ when $x_0 \in f_1^{-1}(J)$ is

$$P(x_0) = \frac{1}{3}\left(\frac{\gamma}{3^k} + 0 + 0\right) = \frac{\gamma}{3^{k+1}}.$$

Case 2: $x_0 \in f_2^{-1}(J)$. Once again, we have $x_0 \in f_2^{-1}(J) \subseteq f_2^{-1}([0,1]) \subseteq [0,1]$. Again, $x_1$ could come from applying any of the three functions to $x_0$, so we proceed with examining each subcase individually.

Subcase 2.1: $x_1 = f_1(x_0)$. Since $x_0 \in f_2^{-1}(J)$, and $J$ is bounded below by 0, we know that in this case $x_0 > f_2^{-1}(0)$. Hence, by Lemma 5, $f_2^{-1}(0) > f_1^{-1}(1)$; thus $x_0 > f_1^{-1}(1)$. Therefore, $x_1 = f_1(x_0) > 1$ and we conclude the probability that $\{x_j\} \to \infty$ is 1.

Subcase 2.2: $x_1 = f_2(x_0)$. Using the same reasoning from Subcase 1.1, we see $x_1 = f_2(x_0) \in f_2(f_2^{-1}(J)) = J$. We conclude $P(x_1) = \gamma/3^k$.

Subcase 2.3: $x_1 = f_3(x_0)$. As demonstrated in Subcases 1.2 and 1.3, here $P(x_1) = 0$.

Combining the three subcases, we have

$$P(x_0) = \frac{1}{3}\left(1 + \frac{\gamma}{3^k} + 0\right) = \frac{1}{3} + \frac{\gamma}{3^{k+1}}.$$

Case 3: $x_0 \in f_3^{-1}(J)$. We use the methods of Cases 1 and 2 and get $P(x_1) = 1$ when $x_1 = f_1(x_0)$ and $x_1 = f_2(x_0)$, and $P(x_1) = \gamma/3^k$ when $x_1 = f_3(x_0)$. Thus

$$P(x_0) = \frac{1}{3}\left(1 + 1 + \frac{\gamma}{3^k}\right) = \frac{2}{3} + \frac{\gamma}{3^{k+1}}$$

when $x_0 \in f_3^{-1}(J)$.

Now we prove that

$$\left\{ y_0 : P(y_0) = \frac{\alpha}{3} + \frac{\gamma}{3^{k+1}} \right\} \subseteq \begin{cases} f_1^{-1}(J), & \alpha = 0, \\ f_2^{-1}(J), & \alpha = 1, \\ f_3^{-1}(J), & \alpha = 2. \end{cases}$$

We proceed with proof by contrapositive, which is if

$$x \notin f_1^{-1}(J) \cup f_2^{-1}(J) \cup f_3^{-1}(J)$$

then

$$P(x_0) \neq \frac{\alpha}{3} + \frac{\gamma}{3^{k+1}}. \tag{3}$$

We define $J = [r, s] \subseteq [0, 1]$, where $r, s \in \mathbb{R}$ and $0 \leq r < s \leq 1$. The assumption that $x \notin \bigcup_{i=1}^{3} f_i^{-1}(J)$ is equivalent to $x_0 \in (-\infty, f_1^{-1}(r))$ or $x_0 \in (f_1^{-1}(s), f_2^{-1}(r))$ or $x_0 \in (f_2^{-1}(s), f_3^{-1}(r))$ or $x_0 \in (f_3^{-1}(s), \infty)$. We examine each piece of this statement individually.

Case 1: $x_0 \in (-\infty, f_1^{-1}(r))$. We examine $x_1 = f_\beta(x_0)$ for $\beta \in \{1, 2, 3\}$ by subcases.

Subcase 1.1: $x_1 = f_1(x_0)$. We have

$$x_1 = f_1(x_0) < f_1(f_1^{-1}(r)) = r.$$

However, $J = [r, s]$ so the fact that $f_1(x_0) < r$ indicates that $P(x_0) < P(y)$, where $y \in J$ so $P(x_0) < \gamma/3^k$. As the probability is a nondecreasing function, we have for any $x_0 \in (-\infty, f_1^{-1}(r))$ that $0 \leq P(x_0) < \gamma/3^k$.

Subcase 1.2: $x_1 = f_2(x_0)$. Similarly to Subcase 1.1, we have

$$x_1 = f_2(x_0) < f_2(f_1^{-1}(r)) < f_2(f_2^{-1}(0)) = 0.$$

Thus, $x_1 < 0$ and we conclude $P(x_0) = 0$ for the upper bound of this interval. Since our probability function is a nondecreasing function, we conclude that $P(x_0) = 0$ for any $x_0$ in this interval by Lemma 7.

Subcase 1.3: $x_1 = f_3(x_0)$. The argument for this subcase is similar to that of Subcase 1.2, for

$$x_1 = f_3(x_0) < f_3(f_1^{-1}(r)) < f_3(f_3^{-1}(0)) = 0,$$

so $P(x_0) = 0$.

Now since each case represents a possible $x_1$ that all have an equal probability of $\frac{1}{3}$, we can multiply each case's probability by $\frac{1}{3}$ and sum them to get $P(x_0)$ for $x_0 \in (-\infty, f_1^{-1}(r))$. This gives

$$\frac{1}{3}(0+0+0) \leq P(x_0) < \frac{1}{3}\left(\frac{\gamma}{3^k} + 0 + 0\right) \quad \text{or} \quad 0 \leq P(x_0) < \frac{\gamma}{3^{k+1}}.$$

Thus for $x_0 \in (-\infty, f_1^{-1}(r))$, we have shown (3).

Case 2: $x_0 \in (f_1^{-1}(s), f_2^{-1}(r))$. We will examine $x_1 = f_\beta(x_0)$ for $\beta \in \{1, 2, 3\}$ by subcases.

Subcase 2.1: $x_1 = f_1(x_0)$. We examine the lower bound of our interval $f_1^{-1}(s) < x_0$. Since $x_1 = f_1(x_0)$, we apply the function $f_1$ to both sides of our inequality to get

$$x_1 = f_1(x_0) > f_1(f_1^{-1}(s)) = s.$$

However, $J = [r, s]$ so $f_1(x_0) > s$ indicates for any $y \in J$ that $P(x_0) > P(y) = \gamma/3^k$. As our probability function is nondecreasing, we have $\gamma/3^k < P(x_0) \le 1$.

Subcase 2.2: $x_1 = f_2(x_0)$. We examine the upper bound of our interval $x_0 < f_2^{-1}(r)$. Again, we apply $f_2$ to both sides of the inequality to get

$$f_2(x_0) < f_2(f_2^{-1}(r)) = r.$$

Since $J = [r, s]$ and $f_2(x_0) < r$, we know for any $y \in J$, $P(x_0) < P(y) = \gamma/3^k$.

Thus, for any $x_0 \in (f_1^{-1}(s), f_2^{-1}(r))$, we have $0 \le P(x_0) < \gamma/3^k$.

Subcase 2.3: $x_1 = f_3(x_0)$. We have $x_0 < f_2^{-1}(r)$, and we apply $f_3$ to both sides of the inequality to get $x_1 = f_3(x_0) < f_3(f_2^{-1}(r)) < f_3(f_3^{-1}(0)) = 0$. Thus $P(x_0) = 0$ by Lemma 7.

As in Case 1, we multiply each of the subcases by $\frac{1}{3}$ and sum them to see

$$\frac{1}{3}\left(\frac{\gamma}{3^k} + 0 + 0\right) < P(x_0) < \frac{1}{3}\left(\frac{\gamma}{3^k} + 1 + 0\right) \quad \text{or} \quad \frac{\gamma}{3^{k+1}} < P(x_0) < \frac{1}{3} + \frac{\gamma}{3^{k+1}}.$$

Thus for $x_0 \in (f_1^{-1}(s), f_2^{-1}(r))$, we have shown (3).

Case 3: $x_0 \in (f_2^{-1}(s), f_3^{-1}(r))$. The proof reduces to the subcases examined in Case 2, therefore again, we have

$$\frac{1}{3}\left(1 + \frac{\gamma}{3^k} + 0\right) < P(x_0) < \frac{1}{3}\left(1 + 1 + \frac{\gamma}{3^k}\right) \quad \text{or} \quad \frac{1}{3} + \frac{\gamma}{3^{k+1}} < P(x_0) < \frac{2}{3} + \frac{\gamma}{3^{k+1}}.$$

Thus for $x_0 \in (f_2^{-1}(s), f_3^{-1}(r))$, we have shown (3).

Case 4: $x_0 \in (f_3^{-1}(s), \infty)$. As before, we examine $x_1 = f_\beta(x_0)$ for $\beta \in \{1, 2, 3\}$ by subcases.

Subcase 4.1: $x_1 = f_1(x_0)$. We examine the lower bound of the interval: $x_0 > f_3^{-1}(s)$. Since $f_3^{-1}(s) > p_1$, we know $x_0 > p_1$; therefore $P(x_0) = 1$ for the lower bound.

Subcase 4.2: $x_1 = f_2(x_0)$. By the same reasoning as Subcase 4.1, since $f_3^{-1}(s) > p_2$, we know $P(x_0) = 1$ for the lower bound.

Subcase 4.3: $x_1 = f_3(x_0)$. We examine the lower bound of the interval: $x_0 > f_3^{-1}(s)$. Since $x_1 = f_3(x_0)$, we apply $f_3$ to both sides of the inequality to get $x_1 = f_3(x_0) > f_3(f_3^{-1}(s)) = s$. However, $J = [r, s]$ so the fact that $f_3(x_0) > s$

indicates for any $y \in J$ that $P(x_0) > P(y)$, which means $\gamma/3^k < P(x_0)$. Thus for this $x_1$, for $x_0 \in (f_3^{-1}(s), \infty)$, we have $\gamma/3^k < P(x_0)$.

Now since each case represents a possible $x_1$ that all have an equal probability of $\frac{1}{3}$,

$$\frac{1}{3}\left(1+1+\frac{\gamma}{3^k}\right) < P(x_0) \quad \text{or} \quad \frac{2}{3}+\frac{\gamma}{3^{k+1}} < P(x_0).$$

Thus for $x_0 \in (f_1^{-1}(s), \infty)$, we have shown (3).

Thus we have shown (3) for each of the four cases of $x_0 \notin \bigcup_{i=1}^{3} f_i^{-1}(J)$. □

We end this section with a corollary to Theorem 13 that helps describe the intervals of constant height, as seen in Figure 2.

**Corollary 15.** *Consider three linear functions $f_k$ defined by Definition 3 and $P(x_0)$ defined by Definition 6. For any value of $\gamma/3^k$, with $\gamma, k \in \mathbb{N}$ and $\gamma \leq 3^k$, there is some interval of $x_0$-values such that $P(x_0) = \gamma/3^k$ for all $x_0$ in that interval.*

## Probabilities for *n* functions

The situation for $n$ functions is quite similar to that of three functions. We begin this section by first generalizing Theorem 10 and then generalizing Theorem 13.

**Theorem 16.** *Consider the linear functions $f_k$ defined by Definition 3 and $P(x_0)$ as in Definition 6, and let $n$ be the number of functions. For $x_0 \in \mathbb{R}$, $P(x_0) = \beta/n$ if and only if $x_0 \in I_\beta = [f_\beta^{-1}(p_n), f_{\beta+1}^{-1}(p_1)]$, where $\beta \in \{1, 2, \ldots, n-1\}$.*

*Proof.* Let $\beta \in \{1, 2, \ldots, n-1\}$ and let $n$ be the number of functions. We assume $x_0 \in I_\beta = [f_\beta^{-1}(p_n), f_{\beta+1}^{-1}(p_1)]$. We examine $x_1 = f_\lambda(x_0)$ for $\lambda \in \{1, 2, \ldots, n\}$.

Case 1: $\lambda = \beta$. Consider the lower bound of $I_\beta$, which is $x_0 \geq f_\beta^{-1}(p_n)$. Since $\lambda = \beta$, we have $f_\lambda(x_0) \geq f_\beta(f_\beta^{-1}(p_n)) = p_n$. Since the lower bound of the interval is greater than or equal to $p_n$, we know that iteration through $f_n$ will result in a sequence where $\{x_j\} \to \infty$. Therefore, due to Lemma 7, iteration using the set of all possible functions results in $\{x_j\} \to \infty$. Thus, $P(x_0) = 1$ for this case.

Case 2: $\lambda < \beta$. By Lemma 7, since $P(x_0) = 1$ for Case 1 when $x_1 = f_\lambda(x_0)$, we know that $P(x_0) = 1$ for this case as well.

Case 3: $\lambda > \beta$. Consider the upper bound of $I_\beta$, which is $x_0 \leq f_{\beta+1}^{-1}(p_1)$. The smallest value for $\lambda$ in this case is $\beta + 1$. Let $\lambda = \beta + 1$. We apply $f_\lambda$ to the upper bound to get $f_\lambda(x_0) \leq f_{\beta+1}(f_{\beta+1}^{-1}(p_1)) = p_1$. Thus $x_1 \leq p_1$ and $\{x_j\}$ is bounded above by $p_1$. By Lemma 7, since $P(x_0) = 0$ for $\lambda = \beta + 1$, which is the smallest value for $\lambda$ in this case, $P(x_0) = 0$ for all $\lambda > \beta$. Therefore, for Case 3, we conclude that $P(x_0) = 0$.

Each function has a probability of $1/n$ of being chosen on each iteration; thus $P(x_0)$ for any $x_0 \in I_\beta$ is

$$P(x_0) = \frac{1}{n}(\underbrace{1+1+\cdots+1}_{\beta\text{-times}} + \underbrace{0+0+\cdots+0}_{n-\beta\text{-times}}) = \frac{\beta}{n}.$$

Assume $P(x_0) = \beta/n$. Since the probability is of the form $P(x_0) = \beta/n$, we know that only $\beta$ choices of $x_1$ can cause $\{x_j\} \to \infty$. This means that $n - \beta$ choices for $x_1$ will lead to $\{x_j\} \nrightarrow \infty$. Thus the lower bound for $x_0$ must be greater than $f_\beta^{-1}(p_n)$, to ensure that when $x_1 = f_\beta(x_0)$, we have $f_\beta(x_0) > p_n$. Since $x_1$ is greater than the fixed point of $f_n$, we know $\{x_j\} \to \infty$. By Lemma 7, since $x_1 = f_\beta(x_0)$ leads to $\{x_j\} \to \infty$, all choices of $x_1 = f_\lambda(x_0)$ where $\lambda < \beta$ will lead to $\{x_j\} \to \infty$. However, to ensure that $n - \beta$ choices for $x_1$ lead to the sequence $\{x_j\}$ not diverging to infinity, the upper bound for $x_0$ must be $f_{\beta+1}^{-1}(p_1)$. This is because for all $x_0 < f_{\beta+1}^{-1}(p_1)$, it is true that $f_{\beta+1}(f_{\beta+1}^{-1}(p_1)) = p_1$, and for any $\lambda > \beta + 1$, we will have $f_\lambda(f_{\beta+1}^{-1}(p_1)) < 0$, so all choices of $f_\lambda(x_0)$ will lead to $\{x_j\} \nrightarrow \infty$. Therefore, we conclude that if $P(x_0) = \beta/n$ then $x_0 \in I_\beta = [f_\beta^{-1}(p_n), f_{\beta+1}^{-1}(p_1)]$. □

**Theorem 17.** *Consider the linear functions $f_k$ defined by Definition 3 and $P(x_0)$ as in Definition 6, and let $n$ be the number of functions. Assume*

$$J = \left\{ x_0 : P(x_0) = \frac{\gamma}{n^k} \right\},$$

*where $\gamma \in \mathbb{N} \cup \{0\}$ and $n \nmid \gamma$. Then for $\alpha \in \{0, 1, \ldots, n-1\}$ we have*

$$\left\{ y_0 : P(y_0) = \frac{\alpha}{n} + \frac{\gamma}{n^{k+1}} \right\} = f_{\alpha+1}^{-1}(J).$$

**Example 18.** Assume that we have $n$ functions, and we want to know for which $x_0$-values $P(x_0) = v/n^k$, where $v/n^k$ is fully reduced. The value of $k - 1$ is the number of preimages of the $n$ functions you take to find this $P(x_0)$. We can express $v$ as $v = c_{k-1}n^{k-1} + c_{k-2}n^{k-2} + \cdots + c_0 n^0$, where $c_j \in \mathbb{Z} \cap [0, n]$. The coefficient $c_j$ is the number of functions that cause $\{x_j\} \to \infty$ under iteration using $j$ preimages. We next give a numerical example.

**Example 19.** Suppose we have five functions, and we wish to see for which $x_0$-values $P(x_0) = \frac{427}{625}$. We know that $625 = 5^4$. Thus, we would need to take three preimages using our five functions to find this $P(x_0)$. We rewrite $\frac{427}{625}$ as

$$\begin{aligned} \tfrac{427}{625} &= \tfrac{3(125)}{625} + \tfrac{2(25)}{625} + \tfrac{0(5)}{625} + \tfrac{2(1)}{625} = \tfrac{3}{5} + \tfrac{1}{5}\left(\tfrac{52}{125}\right) \\ &= \tfrac{3}{5} + \tfrac{1}{5}\left(\tfrac{2}{5} + \tfrac{2}{125}\right) = \tfrac{3}{5} + \tfrac{1}{5}\left(\tfrac{2}{5} + \tfrac{1}{5}\left(\tfrac{0}{5} + \tfrac{2}{25}\right)\right). \end{aligned}$$

Therefore, we have $P(x_0) = \frac{427}{625}$ if and only if $x_0 \in f_4^{-1}(f_3^{-1}(f_1^{-1}(I_2)))$. This is the same method for finding the interval of constant $P(x_0)$ as used in Example 14.

*Proof.* We assume

$$J = \left\{ x_0 : P(x_0) = \frac{\gamma}{n^k} \right\},$$

where $\gamma \in \{\mathbb{N} \cup 0\}$ and $n \nmid \gamma$. Let $\alpha \in \{0, 1, \ldots, n-1\}$. Since we are proving set equality, we start with proving that

$$f_{\alpha+1}^{-1}(J) \subseteq \left\{ y_0 : P(y_0) = \frac{\alpha}{n} + \frac{\gamma}{n^{k+1}} \right\}.$$

Let $x_0 \in f_{\alpha+1}^{-1}(J)$. Since $x_1 = f_\beta(x_0)$ for any $\beta \in \{1, 2, \ldots, n\}$, we proceed with proof by cases.

Case 1: $\beta < \alpha + 1$. Since $x_0 \in f_{\alpha+1}^{-1}(J)$, we know that $x_0 \geq f_{\alpha+1}^{-1}(0) > f_\beta^{-1}(1)$ by Lemma 5. Thus $x_1 > f_\beta(f_\beta^{-1}(1)) = 1$ and $\{x_j\} \to \infty$, so $P(x_0) = 1$.

Case 2: $\beta = \alpha + 1$. As $x_0 \in f_{\alpha+1}^{-1}(J)$, when $x_1 = f_\beta(x_0)$, we have $x_1 = f_\beta(x_0) \in f_\beta(f_{\alpha+1}^{-1}(J))$ but since $\alpha + 1 = \beta$, this is

$$x_1 = f_\beta(x_0) \in f_{\alpha+1}(f_{\alpha+1}^{-1}(J)) = J.$$

Thus, $P(x_0) = P(y)$ for any $y \in J$, so $P(x_0) = \gamma/n^k$.

Case 3: $\beta > \alpha + 1$. As $x_0 \in f_{\alpha+1}^{-1}(J)$ and $\beta > \alpha + 1$,

$$x_1 < f_\beta(f_{\alpha+1}^{-1}(1)) < f_\beta(f_\beta^{-1}(0)) = 0;$$

thus $\{x_j\}$ is bounded above by 0 and $P(x_0) = 0$ by Lemma 7. Since $\alpha \in \{0, 1 \ldots, n-1\}$ and $\beta \in \{1, 2, \ldots, n\}$, we know this case will occur $n - (\alpha + 1)$ times.

Since each function has the probability of $1/n$ of being chosen on the first iteration, we have

$$P(x_0) = \frac{1}{n}\left( \underbrace{1 + 1 + \cdots + 1 + 1}_{\alpha-\text{times}} + \frac{\gamma}{n^k} + \underbrace{0 + 0 + \cdots + 0}_{n-\alpha-1\text{-times}} \right) = \frac{\alpha}{n} + \frac{\gamma}{n^{k+1}}.$$

We have accounted for all $n$ cases because $\alpha + 1 + (n - \alpha - 1) = n$.

We now show the set containment of

$$\left\{ y_0 : P(y_0) = \frac{\alpha}{n} + \frac{\gamma}{n^{k+1}} \right\} \subseteq f_{\alpha+1}^{-1}(J).$$

We proceed with proof by contrapositive; that is, if $x_0 \notin \bigcup_{\alpha=0}^{n-1} f_{\alpha+1}^{-1}(J)$ then

$$P(x_0) \neq \frac{\alpha}{n} + \frac{\gamma}{n^{k+1}}. \tag{4}$$

We examine all cases, using the convention that $J = [r, s]$.

Case 1: $x_0 \in (-\infty, f_1^{-1}(r))$. There are two subcases.

Subcase 1.1: $x_1 = f_1(x_0)$. Consider the upper bound of the interval, $x_0 < f_1^{-1}(r)$. When $x_0 = f_1(x_0)$, we get $x_0 = f_1(x_0) < f_1(f_1^{-1}(r)) = r$. However, since $J = [r, s]$

and $x_1 < r$, we know that $P(x_0) \le P(y)$, where $y \in J$. Since the lower bound of $P(x_0)$ is 0, we have $0 \le P(x_0) < \gamma/n^k$.

Subcase 1.2: $x_1 = f_\tau(x_0)$, where $\tau \in \mathbb{Z} \cap [2, n]$. We know $x_0 < f_1^{-1}(r)$; thus since $f_1^{-1}(r) < f_\tau^{-1}(0)$, we have $f_\tau(x_0) < f_\tau(f_\tau^{-1}(0)) = 0$. Therefore, by Lemma 7, we conclude $P(x_0) = 0$ for this subcase.

Since each function has a probability of $1/n$ of being chosen, the probability for Case 1 is

$$\frac{1}{n}(0+0) \le P(x_0) < \frac{1}{n}\left(0 + \frac{\gamma}{n^k}\right) \quad \text{or} \quad 0 \le P(x_0) < \frac{\gamma}{n^{k+1}}.$$

Hence, we have shown (4).

Case 2: $x_0 \in (f_\alpha^{-1}(s), f_{\alpha+1}^{-1}(r))$. We examine $x_1 = f_\beta(x_0)$. Note $\beta \in \{1, 2, \ldots, n\}$ and $\alpha \in \{0, 1, \ldots, n-1\}$. We proceed with proof by cases, letting $x_0 \in (f_\alpha^{-1}(s), f_{\alpha+1}^{-1}(r))$.

Subcase 2.1: $\beta < \alpha$. When we have $\beta < \alpha$, the $x_0$-value is greater than $f_\beta^{-1}(1)$, so $x_1 > f_\beta(f_\beta^{-1}(1)) = 1$. Therefore, by Lemma 7, we have $P(x_0) = 1$ for these $\alpha - 1$ cases.

Subcase 2.2: $\beta = \alpha$. Since $x_1 = f_\alpha(x_0) > f_\alpha(f_\alpha^{-1}(s)) = s$, and $J = [r, s]$, we know that $P(x_0) > \gamma/n^k$. Thus for this one instance of $\beta = \alpha$, we have $\gamma/n^k < P(x_0) \le 1$.

Subcase 2.3: $\beta = \alpha + 1$. Let $x_1 = f_{\alpha+1}(x_0)$. Now

$$f_{\alpha+1}(x_0) < f_{\alpha+1}(f_{\alpha+1}^{-1}(r)) = r,$$

so $P(x_0) < \gamma/n^k$ for this case.

Subcase 2.4: $\beta > \alpha + 1$. Since $x_1 = f_\beta(x_0) < f_\beta(f_{\alpha+1}^{-1}(r))$ and $f_{\alpha+1}^{-1}(r) < f_\beta^{-1}(0)$ by Definition 3, we have $x_1 < 0$ and $P(x_0) = 0$. There are $n - (\alpha + 1) = n - \alpha - 1$ instances where $\beta > \alpha + 1$ would be true.

Since each function has a probability of $1/n$ of being chosen, $P(x_0)$ is bounded by

$$\frac{1}{n}\left(\underbrace{1 + \cdots + 1}_{\alpha-1\text{-times}} + \frac{\gamma}{n^k} + \underbrace{0 + \cdots + 0}_{n-\alpha\text{-times}}\right) < P(x_0) < \frac{1}{n}\left(\underbrace{1 + \cdots + 1}_{\alpha\text{-times}} + \frac{\gamma}{n^k} + \underbrace{0 + \cdots + 0}_{n-\alpha-1\text{-times}}\right),$$

which simplifies to

$$\frac{\alpha - 1}{n} + \frac{\gamma}{n^{k+1}} < P(x_0) < \frac{\alpha}{n} + \frac{\gamma}{n^{k+1}}.$$

Therefore, for Case 2 we have shown (4).

Case 3: $x_0 \in (f_n^{-1}(s), \infty)$. There are two subcases, where we consider the lower bound of $x_0 > f_n^{-1}(s)$.

Subcase 3.1: $\beta < n$. We know $f_n^{-1}(s) > f_\beta^{-1}(1)$ by Lemma 5. Thus $P(f_n^{-1}(s)) = 1$, and $P(x_0)$ is nondecreasing, so we conclude $P(x_0) = 1$ in this subcase.

Subcase 3.2: $\beta = n$. We apply $f_n$ to our $x_0$-value to get $f_n(x) > f_n(f_n^{-1}(s)) = s$. However, since $J = [r, s]$ and the lower bound of the interval is greater than the upper bound of $J$, we know that $P(x_0) > P(y)$ where $y \in J$. Thus $P(x_0) > \gamma/n^k$. Therefore, we conclude that $\gamma/n^k < P(x_0) \leq 1$.

From Subcase 3.1, we have shown that the probability of every function except for $f_n$ iterating our $x_0$-value toward $\infty$ is 1. Since the probability of each function being chosen on the first iteration is $1/n$, we have

$$\frac{1}{n}\left(\underbrace{1+1+\cdots+1}_{n-1\text{-times}}+\frac{\gamma}{n^k}\right) < P(x_0) \leq \frac{1}{n}(\underbrace{1+1+\cdots+1+1}_{n\text{-times}}),$$

which simplifies to

$$\frac{n-1}{n}+\frac{\gamma}{n^{k+1}} < P(x_0) \leq 1.$$

Therefore, for Case 3 we have shown (4). □

Finally, we can apply Theorem 17 to see that the function $P(x_0)$ can have intervals of constant height for any value of $\gamma/n^k$.

**Corollary 20.** *Consider the linear functions $f_k$ as in Definition 3 and $P(x_0)$ as in Definition 6, and let $n$ be the number of functions. For any value of $\gamma/n^k$, with $\gamma, k \in \mathbb{N}$ and $\gamma \leq n^k$, there is some interval of $x_0$-values such that $P(x_0) = \gamma/n^k$ for all $x_0$ in that interval.*

## Future directions

The functions studied here are sets of linear functions where we restrict the coefficients of $f_k(x) = a_k x - b_k$ such that $a_k, b_k$ are whole numbers with $a_k > b_k$. There are numerous natural extensions to this work. One such extension would be to remove the whole number restrictions on $a_k$ and $b_k$. Even further, one might explore the possibility of $a_k, b_k$ being complex numbers. Additionally, we can expand this research into higher-order functions, such as taking a set of functions of the form $F = \{f_k(x) = a_k x^2 + b_k x + c_k\}$. The interesting thing to note with quadratic functions is that each function can have zero, one, or two fixed points. Lastly, another natural extension is to consider what behavior $P(x_0)$ would have if the functions $f_k(x)$ were not chosen with equal probability at each iteration.

## References

[Dovgoshey et al. 2006] O. Dovgoshey, O. Martio, V. Ryazanov, and M. Vuorinen, "The Cantor function", *Expo. Math.* **24**:1 (2006), 1–37. MR Zbl

[Royden 1988] H. L. Royden, *Real analysis*, 3rd ed., Macmillan, New York, 1988. MR Zbl

[Stankewitz and Rolf 2012] R. L. Stankewitz and J. S. Rolf, "Complex dynamics: chaos, fractals, the Mandelbrot set, and more", pp. 1–83 in *Explorations in complex analysis*, Math. Assoc. America, Washington, DC, 2012. MR Zbl

Received: 2019-05-13 Revised: 2019-12-11 Accepted: 2019-12-23

jordanlarmstrong@aol.com *Department of Mathematical Sciences, U.S. Air Force Academy, Air Force Academy, CO, United States*

beth.schaubroeck@usafa.edu *Department of Mathematical Sciences, U.S. Air Force Academy, Air Force Academy, CO, United States*

INVOLVE 13:2 (2020)
dx.doi.org/10.2140/involve.2020.13.301

# Numerical semigroup tree of multiplicities 4 and 5

Abby Greco, Jesse Lansford and Michael Steward

(Communicated by Vadim Ponomarenko)

A numerical semigroup $S$ is a cofinite submonoid of the nonnegative integers under addition. The cardinality of the complement of $S$ in the nonnegative integers is called the genus. The smallest nonzero element of $S$ is the multiplicity of $S$. There is an extensive literature about the tree of numerical semigroups, which has been used to count numerical semigroups by genus, yet the structure of the tree itself has not been described in the literature. In this paper, we completely describe the structure of the subtrees of the numerical semigroup tree of multiplicities 4 and 5. We conclude with an application of these numerical semigroup trees' structure.

## 1. Introduction

Let $\mathbb{N}$ denote the nonnegative integers. In the study of numerical semigroups, the question of how many ways we can form a numerical semigroup by removing $g$ elements from $\mathbb{N}$ is particularly important. This question is equivalent to finding a formula for the number of numerical semigroups with genus $g$, denoted by $N(g)$. The tree of numerical semigroups has been a powerful tool for making conjectures and answering questions in this field, and it has been especially useful in analyzing the behavior of the function $N(g)$. Bras-Amorós [2008] conjectured and Zhai [2013] proved that asymptotically $N(g)$ exhibits Fibonacci-like behavior. That is,

$$\lim_{g\to\infty} \frac{N(g)}{\phi^g} = \beta,$$

where $\phi$ is the golden ratio $\frac{1}{2}(1+\sqrt{5})$ and $\beta$ is a constant greater than or equal to 3.78. In particular, this means that

$$\lim_{g\to\infty} \frac{N(g-1)+N(g-2)}{N(g)} = 1 \quad \text{and} \quad \lim_{g\to\infty} \frac{N(g)}{N(g-1)} = \phi.$$

*MSC2010:* primary 05A15, 20M14; secondary 11D07.
*Keywords:* numerical semigroup, tree of numerical semigroups.

Fromentin and Hivert [2016] have computed $N(g)$ for $g \leq 67$ and have extended this to $g \leq 70$ in unpublished computations.[1] This was further extended to $g \leq 71$ by Bras-Amorós and Fernández-González [2019]. Nevertheless, fundamental questions about $N(g)$ remain open. Two related open questions are the validity of the *strong genus conjecture* [Bras-Amorós 2008] and the *weak genus conjecture*.

**Conjecture 1.** *For* $g \geq 2$, $N(g) \geq N(g-1) + N(g-2)$.

**Conjecture 2.** *For* $g \geq 1$, $N(g) \geq N(g-1)$.

Several authors have approached the weak genus conjecture by considering the functions $N(m, g)$ which count the numerical semigroups of multiplicity $m$ and genus $g$. It is immediate to observe that $N(g) = \sum_{m=1}^{g+1} N(m, g)$, which led Kaplan [2012] to conjecture the following, which clearly implies Conjecture 2.

**Conjecture 3.** *For* $m \geq 2$, $N(m, g) \geq N(m, g-1)$.

In this paper, we give a complete description of the structure of the numerical semigroup tree in multiplicities 4 and 5. The paper is organized as follows. In Section 2 we review some preliminaries about numerical semigroups and the numerical semigroup tree. Section 3 contains the main results of the paper about the structure of the tree. In Section 4 we use the main result in multiplicity 4 to give an alternate proof of the quasipolynomial expression for $N(4, g)$.

## 2. Preliminaries

In this section we define terms and state prior results that will be used throughout the rest of this paper. Readers familiar with the fundamentals of numerical semigroups can continue to Section 3 and refer to this section as needed. A good general reference for numerical semigroups is [Rosales and García-Sánchez 2009].

A subset $S$ of $\mathbb{N}$ is called a *numerical semigroup* if it is cofinite in $\mathbb{N}$, contains 0, and is closed under addition. We sometimes refer to numerical semigroups as semigroups for brevity. The cardinality of $\mathbb{N} \setminus S$ is called the *genus*, $g(S)$, and the smallest nonzero element of $S$ is called the *multiplicity*, $m(S)$. The largest element of $\mathbb{N} \setminus S$ is called the *Frobenius number*, $F(S)$. It is well known that the number of numerical semigroups of any fixed genus is finite. We can thus define a function $N : \mathbb{N} \to \mathbb{N}$ where $N(g)$ is the number of numerical semigroups of genus $g$. The *minimal generating set* of $S$ is the unique smallest subset of $S$ such that every element of $S$ is an $\mathbb{N}$-linear combination of the numbers in the set. We write

$$S = \langle n_1, \ldots, n_t \rangle = \{a_1 n_1 + \cdots + a_t n_t \mid a_1, \ldots, a_t \in \mathbb{N}\}. \tag{1}$$

The multiplicity of $S$ is always the smallest element of the minimal generating

[1] See https://github.com/hivert/NumericMonoid.

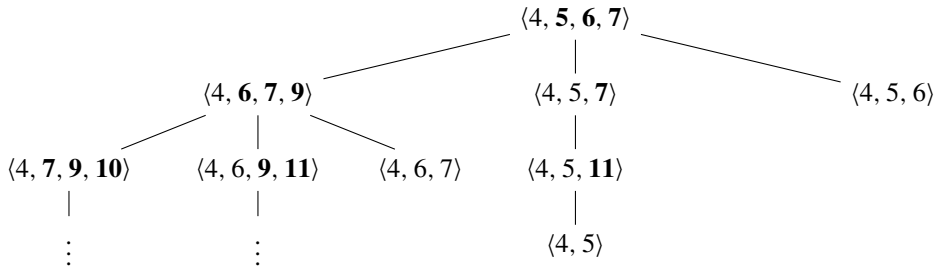



**Figure 1.** Portion of numerical semigroup tree under $\langle 4, 5, 6, 7\rangle$.

set. The generators in the minimal generating set that are larger than the Frobenius number are called *effective generators*. The *Apéry set* of a numerical semigroup with respect to its multiplicity is defined as

$$\mathrm{Ap}(m, S) = \{0, k_1 m + 1, k_2 m + 2, \ldots, k_{m-1} m + (m-1)\}, \tag{2}$$

where $\{k_1, \ldots, k_{m-1}\}$ are positive integers such that $k_i m + i$ is the smallest positive integer in $S$ congruent to $i$ modulo $m$. This uniquely associates to each numerical semigroup of multiplicity $m$ an $(m-1)$-tuple, $(k_1, \ldots, k_{m-1})$, called the *Apéry coordinates* of $S$.

Numerical semigroups can be organized into a tree by letting the vertices be the set of numerical semigroups, with root $\mathbb{N}$. The parent of a vertex $S$ is given by $S \cup \{F(S)\}$. This gives a unique path from any vertex to the root. A portion of the tree can be seen in Figure 1. The bold numbers in the figure indicate which generators are effective. Notice that by this construction, any child of $S$ has genus one more than $g(S)$. This form of construction has been explored extensively, and an example of it can be seen in [Bras-Amorós and Bulygin 2009].

Every numerical semigroup with genus $g(S)$ can be obtained from a numerical semigroup of genus $g(S)-1$ by removing an effective generator from the semigroup. The numerical semigroup tree is naturally sorted by multiplicity, with the numerical semigroups of multiplicity $m$ forming a subtree rooted at $(\{0\} \cup [m, \infty)) \cap (\mathbb{Z})$. All of the numerical semigroups of this form fall on the same branch of the tree, and this special branch is called the *ordinary* branch [Bras-Amorós and Bulygin 2009]. Some numerical semigroups contain no effective generators and hence no children in the tree. These numerical semigroups are called *leaves*.

## 3. Numerical semigroup tree structure

The tree structure of multiplicity $m = 3$ is known; see for instance [García-Sánchez et al. 2018]. In this section we give a complete description of the structure of the subtrees of multiplicities 4 and 5. Throughout our discussion, we use the

numerical semigroup and its corresponding vertex in the tree interchangeably. Throughout the section when we refer to a branch, we mean a connected subtree. When we refer to the length of a linear graph, we mean the number of edges in that graph. Note that whenever $k$ appears without a subscript in this section it is the smallest Apéry coordinate of the current numerical semigroup under discussion.

**3.1.** ***The multiplicity-4 subtree.*** The root of this subtree is the semigroup $\langle 4, 5, 6, 7\rangle$. By successively removing the smallest effective generator greater than 4, we create an infinite branch of this subtree, which we call the principal branch. It is easy to see that a multiplicity-4 semigroup lies on the principal branch if and only if it has three effective generators.

To organize the analysis of this subtree, we divide the branches rooted on the principal branch into three classes based on the congruence class of the first effective generator of the root vertex. We show in Lemma 13 that when the first effective generator is even, the branch is infinite. The other branches of the numerical semigroup tree are finite. In a sequence of lemmas, we describe the lengths of all finite branches of the subtree and the structure of the infinite branches. These results are compiled in Theorem 15.

We begin with an example to demonstrate the patterns that we state generally in the lemmas below. In Figure 2 we see the branch of the multiplicity-4 subtree rooted at $\langle 4, 15, 17, 18\rangle$. The bold numbers in the figure are effective generators. Observe that the child formed by removing the effective generator $e$ either has $e+4$ as a new generator or has one fewer generator than its parent.

There are a few observations to make about this branch. First notice that by successively removing the second-largest effective generator, we create the linear branch from the root to $\langle 4, 15, 29\rangle$. To describe the structure of the branch rooted at $\langle 4, 15, 17, 18\rangle$, it is sufficient to know the length of this main subbranch and the lengths of each linear branch descending from it. Each of these linear branches follows one of three patterns.

The pattern of lengths we observe in Figure 2 is $X,\ 3,\ X-1,\ 3,\ X-2,\ \ldots$. Starting at the first branch, rooted at $\langle 4, 15, 17, 18\rangle$, every second branch's length decreases by 1, and starting at the second branch, rooted at $\langle 4, 15, 18, 21\rangle$, every second branch has length 3. Finally the last linear branch, when the semigroup on the main branch has lost an effective generator, has a distinct length. In the case of Figure 2, this subbranch begins at $\langle 4, 15, 29\rangle$ and has length 4. Describing similar patterns for general branches of the multiplicity-4 subtree allows us to describe its full structure.

We consider three cases for the roots of the subtrees. First consider the case where the root of the subtree has the form $\langle 4, 4k+1, 4k+2, 4k+3\rangle$.

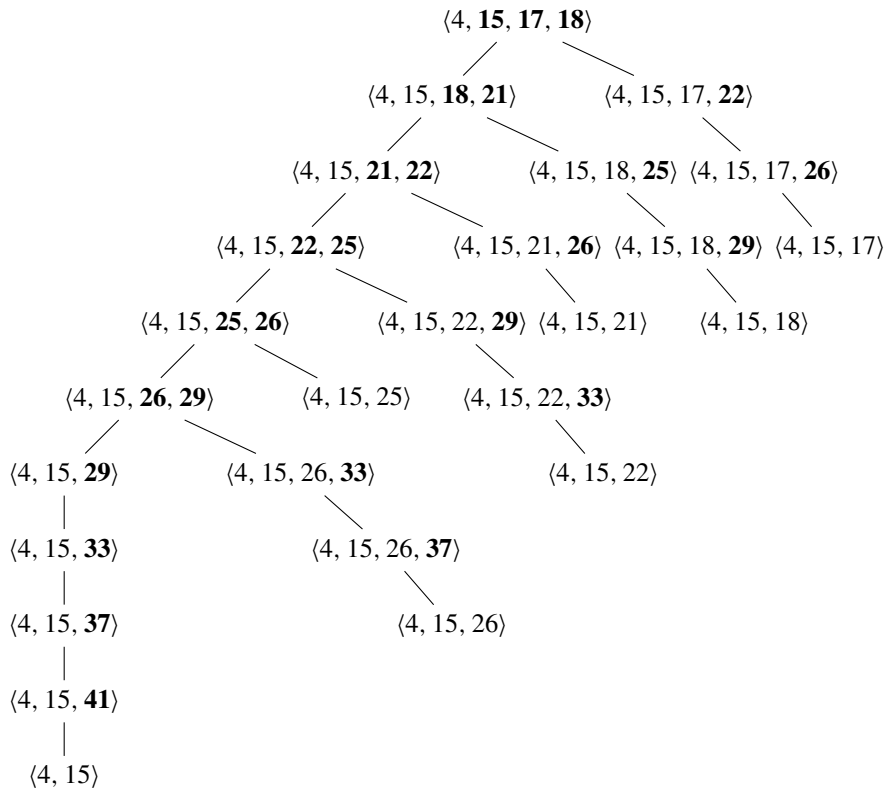



**Figure 2.** Numerical semigroup tree rooted at $\langle 4, 15, 17, 18\rangle$.

**Definition 4.** We define the main branch of any subtree rooted on the principal branch to be the set of semigroups formed by consecutively removing the second-largest effective generator.

Since we primarily care about the lengths of the branches, we can make a more abstract diagram that shows this information directly. We call such pictures *comb diagrams* for their shape, as seen in Figure 3. More precisely, a comb diagram is a depiction of some subtree of the numerical semigroup tree that shows the lengths of the subbranches. Typically, we take the root of the diagram to be a numerical semigroup from the principal branch. We use comb diagrams to illustrate patterns in the linear branch lengths. For instance, in Figure 3, we notice that the decreasing pattern begins with length $k-1$ and continues until the length is equal to 1. Note that this is a different case from the example in Figure 2, whose root was of the form $\langle 4, 4k+3, 4(k+1)+1, 4(k+1)+2\rangle$. We can also note that the constant-length branches are length $k$ and the final linear branch that occurs when the main branch has lost a generator is length $k+1$. Finally, observe that by asserting that the last

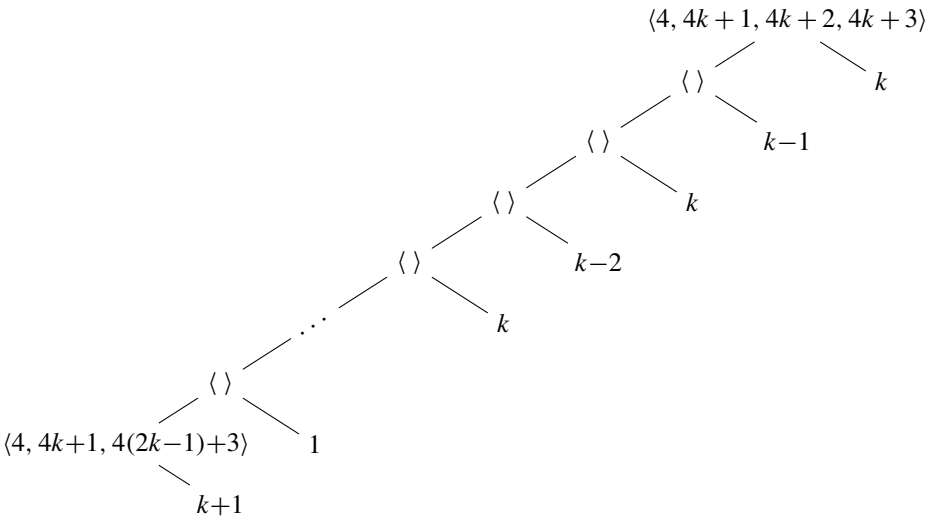



**Figure 3.** Comb diagram of the subtree rooted at $\langle 4, 4k+1, 4k+2, 4k+3\rangle$.

element of the main branch is $\langle 4, 4k+1, 4(2k-1)+3\rangle$, we have tacitly asserted that the length of the main branch is $2k-1$.

In the following lemmas we codify the information shown in Figure 3. In particular, we describe the structure of the subtree whose root lies on the principal branch of the multiplicity-4 subtree and has the form $\langle 4, 4k+1, 4k+2, 4k+3\rangle$. Every semigroup in this branch has 4 and $4k+1$ as generators. We consider several cases for the other minimal generators. First we describe when the linear branches have fixed length.

**Lemma 5.** *Let $S=\langle 4, \alpha, \beta, \gamma\rangle$ be on the main branch under $\langle 4, 4k+1, 4k+2, 4k+3\rangle$, where $\alpha = 4k+1$ and $\gamma = \beta + 1$. Then the linear branch rooted at $S$ has length $k$.*

*Proof.* First note that the hypotheses imply that $\beta \equiv 2 \pmod 4$ and $\gamma \equiv 3 \pmod 4$. Let the corresponding Apéry coordinates be $(k, k_2, k_3)$. Since $\beta$ and $\gamma$ are always effective, we know $k_2 = k_3$, and by hypothesis $k \leq k_2$. The branch terminates at the semigroup $\langle 4, \alpha, \beta\rangle$. Each semigroup in this branch has $\gamma + 4l$ as the element of its Apéry set congruent to 3 modulo 4, where $l$ is some nonnegative integer, and the other elements of its Apéry set are unchanged. Since $\alpha+\beta = 4k+4k_2+3 = \alpha+\gamma$, we have $\gamma+4l = \alpha+\beta$ when $l = k$. It is straightforward to show that $\alpha+\beta$ is the smallest element of $\langle 4, \alpha, \beta\rangle$ congruent to 3 modulo 4. Hence the length of the branch is $k$. □

In order to prove Lemma 5 we used relations between Apéry coordinates to show which semigroup in the branch was a leaf. We found a linear combination of other generators in the same congruence class of the effective generator we removed and

observed that this example was minimal. This technique is used throughout the rest of the paper, with deviations noted when applicable.

**Lemma 6.** *Let* $S=\langle 4, \alpha, \beta, \gamma\rangle$ *be on the main branch under* $\langle 4, 4k+1, 4k+2, 4k+3\rangle$, *where* $\alpha = 4k+1$ *and* $\gamma = \beta + 3 = 4k_2 + 2$. *Then the linear branch rooted at* $S$ *has length* $l = 2k - k_2$, *where* $k_2$ *is the largest Apéry coordinate.*

*Proof.* Note that $\beta \equiv 3 \pmod 4$ and $\gamma \equiv 2 \pmod 4$; thus $k_2 = k_3 + 1$. The linear branch terminates at $\langle 4, \alpha, \beta\rangle$. As in Lemma 5, observe that when $l = 2k - k_2$, we have $\gamma + 4l = 4(k_2) + 2 + 4(2k - k_2) = 2\alpha \in \langle 4, \alpha, \beta\rangle$. This is the smallest linear combination of $\langle 4, \alpha, \beta\rangle$ that is congruent to 2 modulo 4. □

**Lemma 7.** $S = \langle 4, 4k+1, 4(2k-1)+3\rangle$ *is the last element of the main branch. Hence, the length of the main branch is* $2k-1$.

*Proof.* For any semigroup $S$ on the main branch, the Apéry set is

$$\mathrm{Ap}(4, S) = \{0, 4k+1, 4(c)+2, 4(c)+3\}$$

or

$$\mathrm{Ap}(4, S) = \{0, 4k+1, 4(c+1)+2, 4(c)+3\}$$

for some $c \geq k$. Since the Apéry set elements congruent to 2 and 3 modulo 4 are within 4 of each other, we need only consider the span of 4 and $4k+1$ to see if a potential generator is redundant. Clearly the smallest element of $\langle 4, 4k+1\rangle$ congruent to 2 or 3 is $2(4k+1)$. This first occurs when

$$\mathrm{Ap}(4, S) = \{0, 4k+1, 4(c+1)+2, 4(c)+3\}$$

and $c = (2k-1)$. When $c = (2k-1)$, the generator equivalent to $4(c+1)+2$ is redundant, and we have $S = \langle 4, 4k+1, 4(2k-1)+3\rangle$. Since $c$ is initially $k$ and increases by 1 every other semigroup, the total length of the main branch is $2(c-k)+1$, where $c = 2k-1$. Simplifying, the total length of the main branch is $2(2k-1-k)+1 = 2k-1$. □

**Lemma 8.** *Let* $S = \langle 4, \alpha, \beta\rangle$, *where* $\alpha = 4k+1$ *and* $\beta = 4(2k-1)+3$. *Then the branch rooted at* $S$ *has length* $k+1$, *where* $k$ *is the smallest Apéry coordinate.*

*Proof.* The Apéry coordinates of $S$ are $(k, 2k, 2k-1)$. The leaf in this branch is $\langle 4, 4k+1\rangle$, and the smallest element of this set congruent to 3 modulo 4 is $3\alpha$. So $\beta + 4(k+1) = 4(2k-1)+3+4(k+1) = 3\alpha \in \langle 4, \alpha\rangle$, and the length of the branch is $k+1$. □

Lemmas 5–8 fully describe the subtrees rooted at numerical semigroups where the first effective generator is congruent to 1 (mod 4). The subtrees rooted at numerical semigroups with first effective generator congruent to 3 (mod 4) are similar. We omit the proofs, which are analogous to the proofs of Lemmas 5–8.

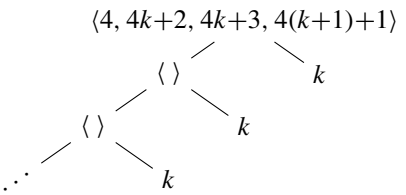



**Figure 4.** Comb diagram of the subtree rooted at $\langle 4, 4k+2, 4k+3, 4(k+1)+1\rangle$.

**Lemma 9.** *Let $S = \langle 4, \alpha, \beta, \gamma\rangle$ be on the main branch under $\langle 4, 4k+3, 4(k+1)+1, 4(k+1)+2\rangle$, where $\alpha = 4k+3$ and $\gamma = \beta + 1 = 4k_2 + 2$. Then the linear branch rooted at $S$ has length $1 + 2k - k_2$.*

**Lemma 10.** *Let $S = \langle 4, \alpha, \beta, \gamma\rangle$ be on the main branch under $\langle 4, 4k+3, 4(k+1)+1, 4(k+1)+2\rangle$, where $\alpha = 4k+3$ and $\gamma = \beta + 3 = 4k_1 + 1$. Then the linear branch rooted at $S$ has length $k$, where $k$ is the smallest Apéry coordinate.*

**Lemma 11.** *$S = \langle 4, 4k+3, 4(2k+1)+1\rangle$ is the last element of the main branch. Hence, the length of the main branch is $2k$.*

**Lemma 12.** *Let $S = \langle 4, \alpha, \beta\rangle$, where $\alpha = 4k+3$ and $\beta = 4(2k+1)+1$. Then the branch rooted at $S$ has length $k+1$, where $k$ is the smallest Apéry coordinate.*

The subtree rooted at numerical semigroups where the first effective generator is congruent to 2 (mod 4) are infinite, but they still exhibit regular behavior that we can describe. Figure 4 shows a comb diagram for this case.

**Lemma 13.** *Let $S = \langle 4, \alpha, \beta, \gamma\rangle$, where $\alpha = 4k+2$ and $\beta$ and $\gamma$ are effective generators. Then the main branch of the subtree rooted at $S$ is infinite.*

*Proof.* Each element on the main branch has the form $\langle 4, 4k+2, 4c+3, 4(c+1)+1\rangle$ or the form $\langle 4, 4k+2, 4(c+1)+1, 4(c+1)+3\rangle$, where $c \geq k$. In either case, every element of $S$ less than the conductor is even. Hence by [Bras-Amorós and Bulygin 2009, Theorem 12], $S$ lies on an infinite chain. Since this is true for every semigroup of these forms, the chain comprising them is infinite. □

**Lemma 14.** *Let $S = \langle 4, \alpha, \beta, \gamma\rangle$ be on the main branch under $\langle 4, 4k+2, 4k+3, 4(k+1)+1\rangle$. Then the linear branch rooted at $S$ has length $k$.*

*Proof.* This is analogous to Lemma 5. □

We compile these results in one theorem for ease of reference.

**Theorem 15.** *The numerical semigroup tree rooted at $S = \langle 4, 5, 6, 7\rangle$ has the following properties*:

(1) *For $S = \langle 4, \alpha, \beta, \gamma\rangle$, where $\alpha < \beta < \gamma$ are all effective minimal generators, $0 < \gamma - \alpha \leq 3$.*

(2) *Let $S$ be a numerical semigroup with multiplicity $m = 4$, and three effective generators $\alpha < \beta < \gamma$. Then, the descendant $S \setminus \{\alpha\}$ has three effective generators, the descendant $S \setminus \{\beta\}$ has at most two effective generators, and $S \setminus \{\gamma\}$ has at most one effective generator.*

(3) *The subbranches of the numerical semigroup tree rooted at numerical semigroups in the form of $S = \langle 4, 4k+1, 4k+2, 4k+3\rangle$ have the following properties*:

(a) *Let $S = \langle 4, \alpha, \beta, \gamma\rangle$ be on the main branch under $\langle 4, 4k+1, 4k+2, 4k+3\rangle$, where $\alpha = 4k+1$ and $\gamma = \beta + 1$. Then the linear branch rooted at $S$ has length $k$.*

(b) *Let $S = \langle 4, \alpha, \beta, \gamma\rangle$ be on the main branch under $\langle 4, 4k+1, 4k+2, 4k+3\rangle$, where $\alpha = 4k+1$ and $\gamma = \beta + 3 = 4k_2 + 2$. Then the linear branch rooted at $S$ has length $l = 2k - k_2$, where $k_2$ is the largest Apéry coordinate.*

(c) *$S = \langle 4, 4k+1, 4(2k-1)+3\rangle$ is the last element of the main branch. Hence, the length of the main branch is $2k-1$.*

(d) *Let $S = \langle 4, \alpha, \beta\rangle$, where $\alpha = 4k+1$ and $\beta = 4(2k-1)+3$. Then the branch rooted at $S$ has length $k+1$, where $k$ is the smallest Apéry coordinate.*

(4) *The subbranches of the numerical semigroup tree rooted at numerical semigroups in the form of $S = \langle 4, 4k+2, 4k+3, 4(k+1)+1\rangle$ have the following properties*:

(a) *Let $S = \langle 4, \alpha, \beta, \gamma\rangle$, where $\alpha = 4k+2$ and $\beta$ and $\gamma$ are effective generators. Then the main branch of the subtree rooted at $S$ is infinite.*

(b) *Let $S = \langle 4, \alpha, \beta, \gamma\rangle$ be on the main branch under $\langle 4, 4k+2, 4k+3, 4(k+1)+1\rangle$. Then the linear branch rooted at $S$ has length $k$.*

(5) *The subbranches of the numerical semigroup tree rooted at numerical semigroups in the form of $S = \langle 4, 4k+3, 4(k+1)+1, 4(k+1)+2\rangle$ have the following properties*:

(a) *Let $S = \langle 4, \alpha, \beta, \gamma\rangle$ be on the main branch under $\langle 4, 4k+3, 4(k+1)+1, 4(k+1)+2\rangle$, where $\alpha = 4k+3$ and $\gamma = \beta + 1 = 4k_2 + 2$. Then the linear branch rooted at $S$ has length $1 + 2k - k_2$.*

(b) *Let $S = \langle 4, \alpha, \beta, \gamma\rangle$ be on the main branch under $\langle 4, 4k+3, 4(k+1)+1, 4(k+1)+2\rangle$, where $\alpha = 4k+3$ and $\gamma = \beta + 3 = 4k_1 + 1$. Then the linear branch rooted at $S$ has length $k$, where $k$ is the smallest Apéry coordinate.*

(c) *$S = \langle 4, 4k+3, 4(2k+1)+1\rangle$ is the last element of the main branch. Hence, the length of the main branch is $2k$.*

(d) *Let $S = \langle 4, \alpha, \beta\rangle$, where $\alpha = 4k+3$ and $\beta = 4(2k+1)+1$. Then the branch rooted at $S$ has length $k+1$, where $k$ is the smallest Apéry coordinate.*

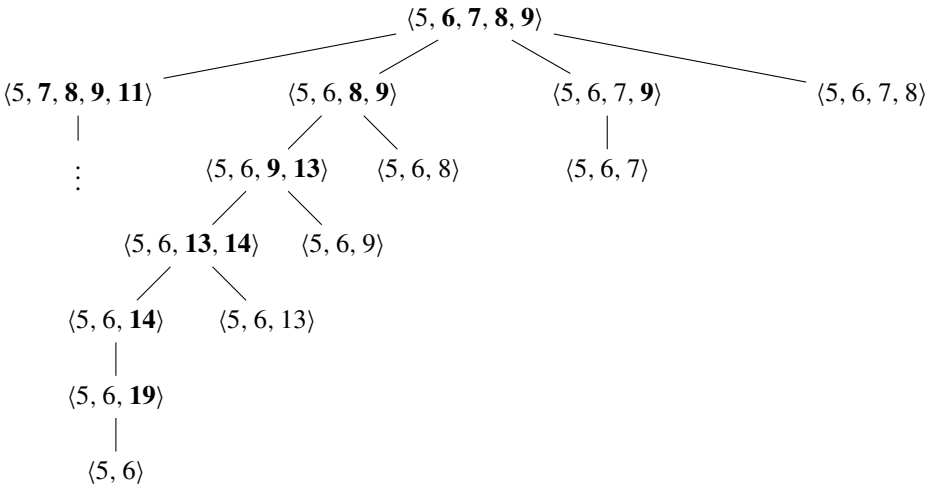



**Figure 5.** Portion of the numerical semigroup tree under $S = \langle 5, 6, 7, 8, 9\rangle$.

*Proof.* Parts (1) and (2) are straightforward. The rest of the theorem follows immediately from Lemmas 5–14. □

**3.2.** ***The subtree rooted at*** $S = \langle 5, 6, 7, 8, 9\rangle$**.** The numerical semigroup tree rooted at $\langle 5, 6, 7, 8, 9\rangle$ is similar to the subtrees rooted at $\langle 3, 4, 5\rangle$ and $\langle 4, 5, 6, 7\rangle$; see Figure 5. We again define an infinite principal branch by successively removing the smallest generator greater than 5. Besides the principal branch, there are no infinite branches. This is because 5 is prime.

The key insight that we use throughout this section is that when we ignore removing the largest effective generator, we get a pattern very similar to what we had in multiplicity 4. We break the branches of the tree into four classes based on the congruence class of the first effective generator. The patterns of lengths act similarly when the first effective generator is congruent to 1 (mod 5) and 2 (mod 5), and again when the first effective generator is congruent to 3 (mod 5) and 4 (mod 5). We omit redundant proofs in our analysis below. Once we describe these patterns we address the largest effective generator in each case.

Each class in multiplicity 5 behaves like the branches rooted at $\langle 4, 4k+1, 4k+2, 4k+3\rangle$ from the multiplicity-4 tree. There is a main branch, and from this branch descend subbranches whose lengths follow a simple pattern. We demonstrate each pattern separately, again using Apéry coordinates to find when a new generator would be redundant.

Here, in Figure 6, we see the comb diagram of the subtree rooted at the numerical semigroup $S = \langle 5, 12, 13, 14, 16\rangle$. Note that in multiplicity 5 the leaves in our comb diagrams are semigroups with exactly one effective generator. There is a

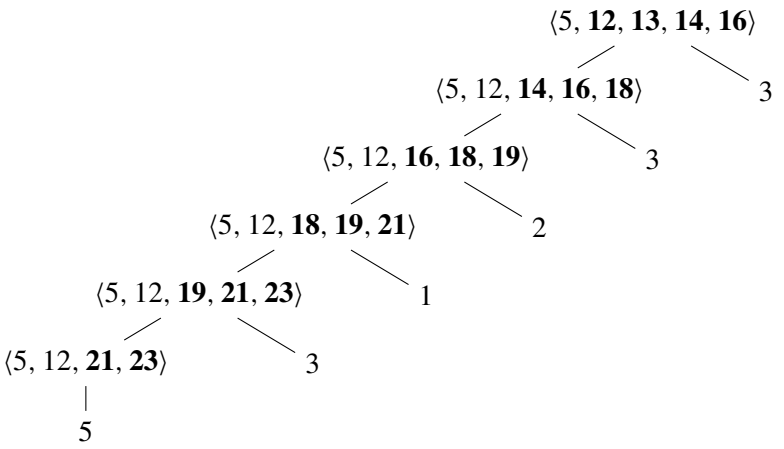



**Figure 6.** Portion of the comb diagram under $S = \langle 5, 12, 13, 14, 16\rangle$.

linear branch descending from each semigroup in the diagram, which the diagram does not address. We describe the patterns of lengths in these diagrams, just as we did in multiplicity 4. However, with multiplicity 5, the constant-length subbranches occur every third branch instead of every other branch as they did in multiplicity 4. The lengths of the other subbranches decrease by 1 until there is a subbranch of length 1. Finally, the subbranch that occurs when the main branch has just lost a generator has length $2k+1$, where $k$ is the smallest Apéry coordinate of $S$.

**Definition 16.** We define the main branch of any subtree rooted on the principal branch of the multiplicity 5 tree to be the set of semigroups formed by consecutively removing the third-largest effective generator.

**Lemma 17.** *Let* $S = \langle 5, \alpha, \beta, \gamma, \delta\rangle$ *be on the main branch under* $\langle 5, 5k+2, 5k+4, 5(k+1)+1, 5(k+1)+3\rangle$, *where* $\alpha = 5k+2,\ \beta = 5k_4+4,\ \gamma = 5(k_4+1)+1$, *and* $\delta = 5(k_4+1)+3$. *Then the linear branch rooted at* $S$ *generated by consecutively removing the second-largest generator loses an effective generator at length* $2k-1$.

*Proof.* The second-largest effective generator is always congruent to 1 or 3 modulo 5, alternating with each removal. The branch of the comb diagram terminates when the new semigroup has only one effective generator, which occurs the first time $\gamma + 5(h+1)$ or $\delta + 5(h+1)$ is in $\langle 5, \alpha, \beta\rangle$. It is easy to verify that if $\gamma$ becomes redundant first, the branch has length $2h+1$, and if $\delta$ becomes redundant first, it has length $2h+2$. We know that $\gamma = 5k_1 + 1$ and $\delta = 5k_3 + 3$. Thus, in order for $\gamma$ to become redundant, it must be equal to 3 copies of $\alpha$, or $\alpha+\beta$. In order for $\delta$ to become redundant, it must be equal to 4 copies of $\alpha$, or $\alpha+\gamma$. Since $\alpha < \beta < \gamma$ and $\beta < 2\alpha$, we know that $\alpha+\beta < 3\alpha$, also note $\alpha+\beta < \alpha+\gamma$, and

so $\gamma$ becomes redundant first when the new generator would be $\alpha+\beta$. We now set $\gamma+5(h+1)=\alpha+\beta=5k_2+5k_4+6$ and solve for $h$. Since $k_4=k_1-1$, we get $h=k-1$. Thus, the total length of the branch until we lose an effective generator is $2h+1=2(k-1)+1=2k-1$. □

The other proofs in this section are similar to this one. In each, we determine which generator becomes redundant first. We then express the length of the branch in terms of the Apéry coordinates of the numerical semigroup, as we did in multiplicity 4. In the following proofs, we omit the restatement of this framework.

**Lemma 18.** *Let $S=\langle 5,\alpha,\beta,\gamma,\delta\rangle$ be on the main branch under $\langle 5,5k+2,5k+4,5(k+1)+1,5(k+1)+3\rangle$, where $\alpha=5k_2+2$ and $k_3=k_4$. Then the branch rooted at $S$ generated by consecutively removing the second-largest generator will lose an effective generator at length $4k_2-k_4-k_1$, where $k_1,k_2,k_3,k_4$ are the Apéry coordinates.*

*Proof.* Let $S=\langle 5,\alpha,\beta,\gamma,\delta\rangle$ with at least two effective generators, $\gamma<\delta$, where $\alpha=5k_2+2$ and $k_3=k_4$. We consider two cases. The first is when $k_1=k_4+1$, and the second is when $k_1=k_4$.

Case 1: Let $k_1=k_4+1$. Then we know that $\gamma=5k_4+4$ and $\delta=5k_1+1$. Now $\gamma$ becomes redundant when $2\alpha$ would be the new generator, and $\delta$ becomes redundant when the smaller of $3\alpha$ and $2\beta$ would be the new generator. Thus, $\gamma$ will become redundant before $\delta$. We set $\gamma+5(h+1)=2\alpha$ and solve for $h$. We get $h=2k_2-k_4-1$. The total length of the branch until we lose an effective generator is $2h+1=4k_2-2k_4-1=4k_2-k_4-k_1$ since by hypothesis we know that $k_1=k_4+1$.

Case 2: Let $k_1=k_4$. Then we know that $\gamma=5k_3+3$, and $\delta=5k_4+4$. $\gamma$ becomes redundant when $\alpha+\beta$ would be the new generator. Now $\delta$ becomes redundant when $2\alpha$ would be the new generator. Thus, $\delta$ will become redundant before $\gamma$. We set $\delta+5(h+1)=2\alpha$ and solve for $h$. We get $h=2k_2-k_4-1$. The total length of the branch until we lose an effective generator is $2h+2=4k_2-2k_4=4k_2-k_4-k_1$ since by hypothesis we know that $k_1=k_4$. □

**Lemma 19.** *$S=\langle 5,5k+2,5(2k)+1,5(2k)+3\rangle$ is the last element of the main branch. Hence, the length of the main branch is $3k-1$.*

*Proof.* Using Definition 16, we define the Apéry set on the main branch as

$$\mathrm{Ap}(5,S)=\{0,5(c+1)+1,5k+2,5(c+1)+3,5c+4\},$$
$$\mathrm{Ap}(5,S)=\{0,5c+1,5k+2,5c+3,5c+4\},$$

or

$$\mathrm{Ap}(5,S)=\{0,5(c+1)+1,5k+2,5c+3,5c+4\}$$

for some $c\geq k$. Since $\alpha=5k+2$, we know the generator equivalent to $5c+4$ will become redundant first, because $2\alpha\equiv 4 \pmod 5$ is the smallest element of $\langle 5,\alpha\rangle$

congruent to 1, 3, or 4 (mod 5). The first value of $c$ for which $5c+4 = 2\alpha = 10k+4$ is when $c = 2k$. Since we increase $c$ by 1 every third semigroup, we know the total length of the main branch is $3(c-k)-1$ when the generator equivalent to $5c+4$ becomes redundant first. Thus, the total length of the main branch is $3(c-k)-1 = 3(2k-k)-1 = 3k-1$. □

**Lemma 20.** *Let $S = \langle 5, 5k+2, 5(2k)+1, 5(2k)+3\rangle$. Then the linear branch rooted at $S$ generated by consecutively removing the second-largest generator will lose an effective generator at length $2k+1$.*

*Proof.* Each semigroup on this branch has the form $\langle 5, \alpha, \beta, \gamma\rangle$, where $\alpha = 5k+2$, $k_3 \geq 2k$, and $k_1 = k_3$ or $k_3+1$. The generator congruent to 3 (mod 5) becomes redundant when $4\alpha$ would be the new generator. The generator congruent to 1 (mod 5) becomes redundant when $3\alpha$ would be the new generator, and hence will become redundant first. We set $5(2k)+1+5(h+1) = 3\alpha$ and solve for $h$. We get $h = k$. Since $h$ increases by 1 every second time we remove the second-largest generator, the total length of the branch until we lose an effective generator is $2k+1$. □

Lemmas 17–20 fully describe the patterns of length for the subbranches rooted at numerical semigroups with first effective generator congruent to 2 (mod 5). The remaining subsets of subbranches are shown in a similar manner.

**Lemma 21.** *Let $S = \langle 5, \alpha, \beta, \gamma, \delta\rangle$ be on the main branch under $\langle 5, 5k+1, 5k+2, 5k+3, 5k+4\rangle$, where $\alpha = 5k+1$:*

(1) *In the case, $k_2 = k_4$, the linear branch rooted at $S$ generated by consecutively removing the second-largest generator will lose an effective generator at length $2k-1$.*

(2) *In the case $k_4 = k_2 - 1$, the branch rooted at $S$ generated by consecutively removing the second-largest generator will lose an effective generator at length $4k_1 - k_2 - k_3 - 1$, where $k_1, k_2, k_3, k_4$ are the Apéry coordinates of $S$.*

(3) *$\langle 5, 5k+1, 5(2k-1)+3, 5(2k-1)+4\rangle$ is the last element of the main branch under $\langle 5, 5k+1, 5k+2, 5k+3, 5k+4\rangle$. Hence, the length of the main branch is $3k-2$.*

(4) *The linear branch under $\langle 5, 5k+1, 5(2k-1)+3, 5(2k-1)+4\rangle$ which is generated by consecutively removing the second-largest generator loses an effective generator at length $2k+1$.*

**Lemma 22.** *Let $S = \langle 5, \alpha, \beta, \gamma, \delta\rangle$ be on the main branch under $\langle 5, 5k+3, 5k+4, 5(k+1)+1, 5(k+1)+2\rangle$, where $\alpha = 5k+3$:*

(1) *In the case $k_1 = k_4$, the linear branch rooted at $S$ generated by consecutively removing the second-largest generator will lose an effective generator at length $2k$.*

(2) *In the case* $k_4 = k_1 - 1$, *the linear branch rooted at* $S$ *generated by consecutively removing the second-largest generator will lose an effective generator at length* $4k_3 - k_4 - k_2$, *where* $k_1, k_2, k_3, k_4$ *are the Apéry coordinates of* $S$.

(3) $\langle 5, 5k+3, 5(2k)+2, 5(2k)+4\rangle$ *is the last element of the main branch under* $\langle 5, 5k+3, 5k+4, 5(k+1)+1, 5(k+1)+2\rangle$. *Hence, the length of the main branch is* $3k-1$.

(4) *The linear branch under* $\langle 5, 5k+3, 5(2k)+2, 5(2k)+4\rangle$ *which is generated by consecutively removing the second-largest generator loses an effective generator at length* $2k+2$.

**Lemma 23.** *Let* $S = \langle 5, \alpha, \beta, \gamma, \delta\rangle$ *be on the main branch under* $\langle 5, 5k+4, 5(k+1)+1, 5(k+1)+2, 5(k+1)+3\rangle$, *where* $\alpha = 5k+4$:

(1) *In the case* $k_3 = k_2 - 1$, *the linear branch rooted at* $S$ *generated by consecutively removing the second-largest generator will lose an effective generator at length* $2k$.

(2) *In the case* $k_2 = k_3$, *the branch rooted at* $S$ *generated by consecutively removing the second-largest generator will lose an effective generator at length* $4k_4 - k_2 - k_1 + 2$, *where* $k_1, k_2, k_3, k_4$ *are the Apéry coordinates.*

(3) $\langle 5, 5k+4, 5(2k+1)+1, 5(2k+1)+2\rangle$ *is the last element of the main branch under* $\langle 5, 5k+4, 5(k+1)+1, 5(k+1)+2, 5(k+1)+3\rangle$. *Hence, the length of the main branch is* $3k$.

(4) *The linear branch under* $\langle 5, 5k+4, 5(2k+1)+1, 5(2k+1)+2\rangle$ *generated by consecutively removing the second-largest generator loses an effective generator at length* $2k+2$.

Until now we have disregarded the linear branches descending from each semigroup in the comb diagram. To fully describe the structure of the multiplicity-5 numerical semigroup tree, we must describe the lengths of these branches as well. We do this in a sequence of lemmas based on the congruence class modulo 5 of the largest generator.

**Lemma 24.** *For* $S = \langle 5, 5a+1, 5b+2, 5c+3, 5d+4\rangle$, *with Apéry coordinates* $(a, b, c, d)$ *and largest generator* $\delta \equiv 1 \pmod 5$, *the length of the branch generated by consecutively removing the largest generator is* $\min(b+d-a+1, 2c-a+1)$.

*Proof.* Let $S = \langle 5, 5a+1, 5b+2, 5c+3, 5d+4\rangle$, where $(a, b, c, d)$ are the Apéry coordinates of $S$, and let the largest generator be $\delta \equiv 1 \pmod 5$. The Apéry coordinates of each numerical semigroup on this branch are in the form $(a+k, b, c, d)$, where $\ell$ is the distance down the branch from $S$. A leaf occurs on this branch when

$$5(a+\ell)+1 \in \langle 5, 5b+2, 5c+3, 5d+4\rangle$$

or in other words when

$$5(a+\ell)+1 = 5t_1 + (5b+2)t_2 + (5c+3)t_3 + (5d+4)t_4 \quad \text{for } t_i \in \mathbb{N}.$$

If $t_2 = t_4 = 1$ and all $t_i = 0$, then we have $5b + 5d + 6 \equiv 1 \pmod 5$ with length $b + d - a + 1$. We can also have $t_3 = 2$ and all other $t_i = 0$; then we have $10c + 6 \equiv 1 \pmod 5$ with length $(2c - a + 1)$. A similar argument to those used in previous lemmas shows that one of these is the smallest element of $\langle 5, b, c, d\rangle$ that is congruent to 1 (mod 5). Thus, the length of the subbranch is $\min(b + d - a + 1, 2c - a + 1)$. □

**Lemma 25.** *Let $S = \langle 5, 5a+1, 5b+2, 5c+3, 5d+4\rangle$, with Apéry coordinates $(a, b, c, d)$. Denote the largest generator of $S$ by $\delta$:*

(1) *When $\delta \equiv 2 \pmod 5$, the length of the branch generated by consecutively removing the largest generator is* $\min(2a - b, c + d - b + 1)$.

(2) *When $\delta \equiv 3 \pmod 5$, the length of the branch generated by consecutively removing the largest generator is* $\min(a + b - c, 2d - c + 1)$.

(3) *When $\delta \equiv 4 \pmod 5$, the length of the branch generated by consecutively removing the largest generator is* $\min(a + c - d, 2b - d)$.

We compile these results in one theorem for ease of reference.

**Theorem 26.** *The numerical semigroup tree rooted at $S = \langle 5, 6, 7, 8, 9\rangle$ has the following properties*:

(1) *For $S = \langle 5, \alpha, \beta, \gamma, \delta\rangle$, where $\alpha < \beta < \gamma < \delta$ are effective, $0 < \delta - \alpha \leq 4$.*

(2) *For $S = \langle 5, \alpha, \beta, \gamma, \delta\rangle$, with effective generators, $\alpha < \beta < \gamma < \delta$, $S \setminus \{\beta\}$ has at most three effective generators, $S \setminus \{\gamma\}$ has at most two effective generators, and $S \setminus \{\delta\}$ has at most one effective generator.*

(3) *The subbranches of the numerical semigroup tree rooted at numerical semigroups in the form of $S = \langle 5, 5k+1, 5k+2, 5k+3, 5k+4\rangle$ have the following properties*:

(a) *Let $S = \langle 5, \alpha, \beta, \gamma, \delta\rangle$ be on the main branch under $\langle 5, 5k+1, 5k+2, 5k+3, 5k+4\rangle$, where $\alpha = 5k+1$ and $\beta = 5k+2$. Then the linear branch rooted at $S$ generated by consecutively removing the second-largest generator will lose an effective generator at length $2k - 1$.*

(b) *Let $S = \langle 5, \alpha, \beta, \gamma, \delta\rangle$ be on the main branch under $\langle 5, 5k+1, 5k+2, 5k+3, 5k+4\rangle$, where $\alpha = 5k+1$, and $k_4 = k_2 - 1$. Then the branch rooted at $S$ generated by consecutively removing the second-largest generator will lose an effective generator at length $4k_1 - k_2 - k_3 - 1$, where $k_1, k_2, k_3, k_4$ are the Apéry coordinates.*

(c) $S = \langle 5, 5k+1, 5(2k-1)+3, 5(2k-1)+4\rangle$ *is the last element of the main branch. Hence, the length of the main branch is* $3k-2$.

(d) *Let* $S = \langle 5, \alpha, \beta, \gamma\rangle$ *be on the main branch under* $\langle 5, 5k+1, 5(2k-1)+3, 5(2k-1)+4\rangle$, *where* $\alpha = 5k+1$. *Then the linear branch rooted at S generated by consecutively removing the second-largest generator loses an effective generator at length* $2k+1$.

(4) *The subbranches of the numerical semigroup tree rooted at numerical semigroups in the form of* $S = \langle 5, 5k+2, 5k+3, 5k+4, 5(k+1)+1\rangle$ *have the following properties*:

(a) *Let* $S = \langle 5, \alpha, \beta, \gamma, \delta\rangle$ *be on the main branch under* $\langle 5, 5k+2, 5k+4, 5(k+1)+1, 5(k+1)+3\rangle$, *where* $\alpha = 5k+2$, $\beta = 5k_4+4$, $\gamma = 5(k_4+1)+1$, *and* $\delta = 5(k_4+1)+3$. *Then the linear branch rooted at S generated by consecutively removing the second-largest generator loses an effective generator at length* $2k-1$.

(b) *Let* $S = \langle 5, \alpha, \beta, \gamma, \delta\rangle$ *be on the main branch under* $\langle 5, 5k+2, 5k+4, 5(k+1)+1, 5(k+1)+3\rangle$, *where* $\alpha = 5k_2+2$ *and* $k_3 = k_4$. *Then the branch rooted at S generated by consecutively removing the second-largest generator will lose an effective generator at length* $4k_2 - k_4 - k_1$, *where* $k_1, k_2, k_3, k_4$ *are the Apéry coordinates.*

(c) $S = \langle 5, 5k+2, 5(2k)+1, 5(2k)+3\rangle$ *is the last element of the main branch. Hence, the length of the main branch is* $3k-1$.

(d) *Let* $S = \langle 5, \alpha, \beta, \gamma\rangle$ *be on the main branch under* $\langle 5, 5k+2, 5(2k)+1, 5(2k)+3\rangle$, *where* $\alpha = 5k+2$. *Then the linear branch rooted at S generated by consecutively removing the second-largest generator will lose an effective generator at length* $2k+1$.

(5) *The subbranches of the numerical semigroup tree rooted at numerical semigroups in the form of* $S = \langle 5, 5k+3, 5k+4, 5(k+1)+1, 5(k+1)+2\rangle$ *have the following properties*:

(a) *Let* $S = \langle 5, \alpha, \beta, \gamma, \delta\rangle$ *be on the main branch under* $\langle 5, 5k+3, 5k+4, 5(k+1)+1, 5(k+1)+2\rangle$, *where* $\alpha = 5k+3$, $\beta = 5(k+1)+1$, $\gamma = 5(k+1)+2$, *and* $\delta = 5(k+1)+4$. *Then the linear branch rooted at S generated by consecutively removing the second-largest generator will lose an effective generator at length* $2k$.

(b) *Let* $S = \langle 5, \alpha, \beta, \gamma, \delta\rangle$ *be on the main branch under* $\langle 5, 5k+3, 5k+4, 5(k+1)+1, 5(k+1)+2\rangle$, *where* $\alpha = 5k+3$. *Then the branch rooted at S generated by consecutively removing the second-largest generator will lose an effective generator at length* $4k_3 - k_4 - k_2$, *where* $k_1, k_2, k_3, k_4$ *are the Apéry coordinates.*

(c) $S = \langle 5, 5k+3, 5(2k)+2, 5(2k)+4\rangle$ *is the last element of the main branch. Hence, the length of the main branch is* $3k-1$.

(d) *Let* $S = \langle 5, \alpha, \beta, \gamma\rangle$ *be on the main branch under* $\langle 5, 5k+3, 5(2k)+2, 5(2k)+4\rangle$, *where* $\alpha = 5k+3$. *Then the linear branch rooted at S generated by consecutively removing the second-largest generator will lose an effective generator at length* $2k+2$.

(6) *The subbranches of the numerical semigroup tree rooted at numerical semigroups in the form of* $S = \langle 5, 5k+4, 5(k+1)+1, 5(k+1)+2, 5(k+1)+3\rangle$ *have the following properties*:

(a) *Let* $S = \langle 5, \alpha, \beta, \gamma, \delta\rangle$ *be on the main branch under* $\langle 5, 5k+4, 5(k+1)+1, 5(k+1)+2, 5(k+1)+3\rangle$, *where* $\alpha = 5k+4$ *and* $k_3 = k_2 - 1$. *Then the linear branch rooted at S generated by consecutively removing the second-largest generator will lose an effective generator at length* $2k$.

(b) *Let* $S = \langle 5, \alpha, \beta, \gamma, \delta\rangle$ *be on the main branch under* $\langle 5, 5k+4, 5(k+1)+1, 5(k+1)+2, 5(k+1)+3\rangle$, *where* $\alpha = 5k+4$, *and* $k_2 = k_3$. *Then the branch rooted at S generated by consecutively removing the second-largest generator will lose an effective generator at length* $4k_4 - k_2 - k_1 + 2$, *where* $k_1, k_2, k_3, k_4$ *are the Apéry coordinates.*

(c) $S = \langle 5, 5k+4, 5(2k+1)+1, 5(2k+1)+2\rangle$ *is the last element of the main branch. Hence, the length of the main branch is* $3k$.

(d) *Let* $S = \langle 5, \alpha, \beta, \gamma\rangle$ *be on the main branch under* $\langle 5, 5k+4, 5(2k+1)+1, 5(2k+1)+2\rangle$, *where* $\alpha = 5k+4$. *Then the linear branch rooted at S generated by consecutively removing the second-largest generator will lose an effective generator at length* $2k+2$.

(7) *The linear branches formed by successively removing the largest generator from a numerical semigroup anywhere in the multiplicity* 5 *tree have the following properties*:

(a) *For* $S = \langle 5, 5a+1, 5b+2, 5c+3, 5d+4\rangle$, *with Apéry coordinates* $(a, b, c, d)$ *and largest generator* $\delta \equiv 1 \pmod 5$, *the length of the branch generated by consecutively removing the largest generator is* $\min(b+d-a+1, 2c-a+1)$.

(b) *For* $S = \langle 5, 5a+1, 5b+2, 5c+3, 5d+4\rangle$, *with Apéry coordinates* $(a, b, c, d)$ *and largest generator* $\delta \equiv 2 \pmod 5$, *the length of the branch generated by consecutively removing the largest generator is* $\min(2a-b, c+d-b+1)$.

(c) *For* $S = \langle 5, 5a+1, 5b+2, 5c+3, 5d+4\rangle$, *with Apéry coordinates* $(a, b, c, d)$ *and largest generator* $\delta \equiv 3 \pmod 5$, *the length of the branch generated by consecutively removing the largest generator is* $\min(a+b-c, 2d-c+1)$.

(d) *For* $S = \langle 5, 5a+1, 5b+2, 5c+3, 5d+4\rangle$, *with Apéry coordinates* $(a, b, c, d)$ *and largest generator* $\delta \equiv 4 \pmod 5$, *the length of the branch generated by consecutively removing the largest generator is* $\min(a+c-d, 2b-d)$.

*Proof.* Parts (1) and (2) are straightforward. The rest of the theorem is a direct consequence of Lemma 17–Lemma 25. □

## 4. Counting $N(4, g)$

Understanding the structure of the numerical semigroup tree allows us to count the number of semigroups by genus. In this section we provide an alternate proof for a known quasipolynomial expression for $N(4, g)$ when $g \geq 9$. Formulas for $N(4, g)$ and $N(5, g)$ are already known, but they were arrived at by computational methods. In Theorem 16 of [García-Sánchez et al. 2018], there is an equation for counting numerical semigroups of multiplicity $m = 4$ for genus greater than $g = 9$. A similar, much longer, formula for $N(5, g)$ is also shown in that work. Another noncomputational proof of the fact that $N(4, g)$ is nondecreasing was recently found by Eliahou and Fromentin [2019] using the related tool of gapset filtrations. Their approach is in some sense dual to considering the tree of numerical semigroups, but they establish their result without finding the quasipolynomial expression for $N(4, g)$.

**Theorem 27** [García-Sánchez et al. 2018, Theorem 16]. *Let g be an integer greater than* 9. *Then*

$$\begin{aligned}
N(4,g) = {} & \left\lfloor \tfrac{g}{4} \right\rfloor^2 - \tfrac{3}{2}\left\lfloor \tfrac{g}{3} \right\rfloor^2 + \left\lfloor \tfrac{g+1}{5} \right\rfloor^2 - \tfrac{3}{2}\left\lfloor \tfrac{g+1}{3} \right\rfloor^2 - \tfrac{3}{2}\left\lfloor \tfrac{g+2}{6} \right\rfloor^2 - \tfrac{3}{2}\left\lfloor \tfrac{g+2}{5} \right\rfloor^2 \\
& + \tfrac{3}{2}\left\lfloor \tfrac{g+2}{4} \right\rfloor^2 - \tfrac{3}{2}\left\lfloor \tfrac{g+2}{3} \right\rfloor^2 - \tfrac{3}{2}\left\lfloor \tfrac{g+5}{6} \right\rfloor^2 + \tfrac{3}{2}\left\lfloor \tfrac{2g+1}{5} \right\rfloor^2 - \tfrac{3}{2}\left\lfloor \tfrac{2g+4}{5} \right\rfloor^2 \\
& + \left(g - \tfrac{3}{2}\right)\left\lfloor \tfrac{g}{3} \right\rfloor - \left\lfloor \tfrac{g}{5} \right\rfloor \left\lfloor \tfrac{2g}{5} \right\rfloor + \left\lfloor \tfrac{g}{4} \right\rfloor \left(1 - \left\lfloor \tfrac{g}{2} \right\rfloor\right) + \left(\tfrac{1}{4} + \left\lfloor \tfrac{g}{2} \right\rfloor\right)\left\lfloor \tfrac{g}{2} \right\rfloor \\
& + \left\lfloor \tfrac{g+1}{5} \right\rfloor + \left(g - \tfrac{1}{2}\right)\left\lfloor \tfrac{g+1}{3} \right\rfloor + \left(\left\lfloor \tfrac{g}{2} \right\rfloor + \tfrac{1}{2}\right)\left\lfloor \tfrac{g+2}{6} \right\rfloor + \left\lfloor \tfrac{2g}{5} \right\rfloor \left\lfloor \tfrac{g+2}{5} \right\rfloor \\
& - \left\lfloor \tfrac{g}{2} \right\rfloor \left\lfloor \tfrac{g+2}{4} \right\rfloor + \left(g + \tfrac{1}{2}\right)\left\lfloor \tfrac{g+2}{3} \right\rfloor + \left(-\left\lfloor \tfrac{g}{2} \right\rfloor + g + \tfrac{1}{2}\right)\left\lfloor \tfrac{g+5}{6} \right\rfloor \\
& + \left(\left\lfloor \tfrac{g}{5} \right\rfloor - g + 1\right)\left\lfloor \tfrac{2g+1}{5} \right\rfloor + \left(-\left\lfloor \tfrac{g+1}{5} \right\rfloor + g + 1\right)\left\lfloor \tfrac{2g+4}{5} \right\rfloor - \tfrac{5g^2}{8} - \tfrac{3g}{8}.
\end{aligned}$$

In [Alhajjar et al. 2019], the authors found the following quasipolynomial expression for $N(4, g)$, when $g \geq 9$, using observations about Kunz polytopes:

$$N(4,g) = \begin{cases} \frac{g^2}{12} + \frac{g}{2}, & g \equiv 0 \pmod 6, \\ \frac{g^2}{12} + \frac{g}{2} - \frac{7}{12}, & g \equiv 1, 5 \pmod 6, \\ \frac{g^2}{12} + \frac{g}{2} - \frac{1}{3}, & g \equiv 2, 4 \pmod 6, \\ \frac{g^2}{12} + \frac{g}{2} - \frac{1}{4}, & g \equiv 3 \pmod 6. \end{cases} \tag{3}$$

We demonstrate an alternate proof of this formula using the structure of the semigroup tree directly.

**Theorem 28.** *$N(4,3)=1$, and for any $g\geq 4$,*

$$N(4,g)=\begin{cases}\frac{g^2}{12}+\frac{g}{2}, & g\equiv 0 \pmod 6,\\ \frac{g^2}{12}+\frac{g}{2}-\frac{7}{12}, & g\equiv 1,5 \pmod 6,\\ \frac{g^2}{12}+\frac{g}{2}-\frac{1}{3}, & g\equiv 2,4 \pmod 6,\\ \frac{g^2}{12}+\frac{g}{2}-\frac{1}{4}, & g\equiv 3 \pmod 6.\end{cases}$$

*Proof.* The only semigroup of multiplicity 4 and genus 3 is $\langle 4,5,6,7\rangle$. In order to count the semigroups of multiplicity 4 by genus, we consider the structure of the tree of semigroups of multiplicity 4. The structure of our argument is to first count the number of semigroups on each subtree rooted on the principal branch then add these counting functions together. As we observed in Theorem 15, the behavior of the branch is determined by the congruence class modulo 4 of the second generator of the root. Throughout the proof we use many facts about the structure of the tree. All of these are immediate consequences of Theorem 15.

First consider the branch rooted at $\langle 4, 4k+2, 4k+3, 4(k+1)+1\rangle$. This branch is composed of length $k$ linear subbranches descending from a main branch, with one new linear subbranch in each genus. Observe that the genus of the root is $3k+1$. The first linear subbranch terminates at genus $4k+1$. Until that point there is one new subbranch in each genus. From genus $4k+1$ onward, one new subbranch begins, and one subbranch ends. Thus the total number of semigroups is constant. The number of semigroups on this branch is

$$\#_{2,k}(g)=\begin{cases}\max\{0,g-3k\}, & g\leq 4k,\\ k+1, & g>4k.\end{cases} \tag{4}$$

The total number of semigroups on any branch of this form is $\#_2(g)=\sum_{k=1}^{\infty}\#_{2,k}(g)$. To count this, we need only consider how many branches have stabilized at each genus and how many are still growing.

At genus $g$, $\lfloor\frac{g-1}{4}\rfloor$ branches have stabilized, yielding $2+3+4+\cdots+\big(\lfloor\frac{g-1}{4}\rfloor+1\big)$, which sums to

$$\tfrac{1}{2}\big(\lfloor\tfrac{g-1}{4}\rfloor+1\big)\big(\lfloor\tfrac{g-1}{4}\rfloor+2\big)-1.$$

The branches that have not yet stabilized begin at each genus $3k+1$, $k\geq 1$. So at some genus $g$, all of the growing branches have the same number of vertices modulo 3. Hence the vertices on the growing branches are either

$$1+4+7+\cdots+(3n+1),\quad 2+5+8+\cdots+(3n+2),\quad \text{or}\quad 3+6+9+\cdots+3n$$

in number. Branches stabilize in genus $4k+1$, with $k+1$ vertices. So the first time there are $k$ vertices in a growing branch is at genus $4k$. Thus each sum becomes one term longer when the genus is increased by 12. Consider the case where the

genus is congruent to 0 modulo 3. In this case, the sum is

$$3+6+\cdots+3\left\lfloor\tfrac{g}{12}\right\rfloor=\tfrac{1}{2}\left(3\left\lfloor\tfrac{g}{12}\right\rfloor+3\right)\left(\left\lfloor\tfrac{g}{12}\right\rfloor\right).$$

The other cases work similarly. Since every branch is either stabilized or growing, this accounts for every vertex at each genus. So we have the sum

$$\#_2(g)=\begin{cases}\frac{1}{2}\left(\left\lfloor\frac{g-1}{4}\right\rfloor+1\right)\left(\left\lfloor\frac{g-1}{4}\right\rfloor+2\right)-1+\left(3\left\lfloor\frac{g+8}{12}\right\rfloor-1\right)\frac{1}{2}\left\lfloor\frac{g+8}{12}\right\rfloor, & g\equiv 1 \pmod 3,\\ \frac{1}{2}\left(\left\lfloor\frac{g-1}{4}\right\rfloor+1\right)\left(\left\lfloor\frac{g-1}{4}\right\rfloor+2\right)-1+\left(3\left\lfloor\frac{g+4}{12}\right\rfloor+1\right)\frac{1}{2}\left\lfloor\frac{g+4}{12}\right\rfloor, & g\equiv 2 \pmod 3,\\ \frac{1}{2}\left(\left\lfloor\frac{g-1}{4}\right\rfloor+1\right)\left(\left\lfloor\frac{g-1}{4}\right\rfloor+2\right)-1+\left(3\left\lfloor\frac{g}{12}\right\rfloor+3\right)\frac{1}{2}\left\lfloor\frac{g}{12}\right\rfloor, & g\equiv 0 \pmod 3.\end{cases}$$

Next consider the branch rooted at $\langle 4, 4k+1, 4k+2, 4k+3\rangle$, which has genus $3k$. We can observe from the structure of this branch that the first leaves occur in genus $4k$, and the function counting the number of leaves in each genus is

$$\ell_{1,k}(g)=\begin{cases}2, & 4k\le g\le 5k-2,\ g \text{ even},\\ 1, & 4k\le g\le 5k-2,\ g \text{ odd},\\ 1, & 5k-1\le g\le 6k,\ g \text{ even},\\ 0, & \text{otherwise}.\end{cases}$$

From genus $3k$ to genus $4k$, there is one additional semigroup in each genus. From genus $4k$ to genus $5k-2$, there are two leaves in each even genus and one in each odd genus. So the number of semigroups in the branch of genus $g$ decreases by 1 whenever $g$ is even. From genus $5k-1$ to genus $6k$, there are no new subbranches, and there is a leaf in each even genus, so the number of semigroups decreases by 1 in each even genus. The largest genus of any semigroup on this branch is $6k$, the genus of $\langle 4, 4k+1\rangle$. Hence we can write the counting function for this branch:

$$\#_{1,k}(g)=\begin{cases}\max\{g+1-3k, 0\}, & g\le 4k,\\ \left\lfloor\frac{6k+2-g}{2}\right\rfloor, & 4k<g\le 6k,\\ 0, & \text{otherwise}.\end{cases}$$

We compute the sum of these functions to find the total number of semigroups on any branch of this form, $\#_1(g)=\sum_{k=1}^{\infty}\#_{1,k}(g)$. At any genus each branch is either growing, declining, or empty. In particular $\langle 4, 4k+1, 4k+2, 4k+3\rangle$ is growing for $3k\le g\le 4k$ and declining for $4k+1\le g\le 6k$. Consider $g\equiv 0 \pmod 6$. At genus $g$, $\left\lfloor\frac{g}{12}\right\rfloor+1$ branches are growing, yielding

$$1+4+7+\cdots+\left(3\left\lfloor\tfrac{g}{12}\right\rfloor+1\right),$$

which sums to

$$\tfrac{1}{2}\left(3\left\lfloor\tfrac{g}{12}\right\rfloor+2\right)\left(\left\lfloor\tfrac{g}{12}\right\rfloor+1\right).$$

There are also $\left\lfloor\frac{g-5}{12}\right\rfloor+1$ declining branches at genus $g$, yielding

$$1+4+7+\cdots+\left(3\left\lfloor\tfrac{g-5}{12}\right\rfloor+1\right),$$

which sums to

$$\tfrac{1}{2}\left(3\left\lfloor\tfrac{g-5}{12}\right\rfloor+2\right)\left(\left\lfloor\tfrac{g-5}{12}\right\rfloor+1\right).$$

Since every branch is either empty, growing, or declining, the sum of these two values accounts for every vertex at this genus. The other cases are similar, yielding the counting function:

$$\#_1(g)=\begin{cases}\frac{1}{2}\left(\left(3\left\lfloor\frac{g}{12}\right\rfloor+2\right)\left(\left\lfloor\frac{g}{12}\right\rfloor+1\right)+\left(3\left\lfloor\frac{g-5}{12}\right\rfloor+2\right)\left(\left\lfloor\frac{g-5}{12}\right\rfloor+1\right)\right), & g\equiv 0 \pmod 6,\\ \frac{1}{2}\left(\left(3\left\lfloor\frac{g-4}{12}\right\rfloor+4\right)\left(\left\lfloor\frac{g-4}{12}\right\rfloor+1\right)+\left(3\left\lfloor\frac{g}{12}\right\rfloor+3\right)\left(\left\lfloor\frac{g}{12}\right\rfloor\right)\right), & g\equiv 1 \pmod 6,\\ \frac{1}{2}\left(\left(3\left\lfloor\frac{g+4}{12}\right\rfloor+3\right)\left(\left\lfloor\frac{g+4}{12}\right\rfloor\right)+\left(3\left\lfloor\frac{g}{12}\right\rfloor+3\right)\left(\left\lfloor\frac{g}{12}\right\rfloor\right)\right), & g\equiv 2 \pmod 6,\\ \frac{1}{2}\left(\left(3\left\lfloor\frac{g}{12}\right\rfloor+2\right)\left(\left\lfloor\frac{g}{12}\right\rfloor+1\right)+\left(3\left\lfloor\frac{g-9}{12}\right\rfloor+4\right)\left(\left\lfloor\frac{g-9}{12}\right\rfloor+1\right)\right), & g\equiv 3 \pmod 6,\\ \frac{1}{2}\left(\left(3\left\lfloor\frac{g-4}{12}\right\rfloor+4\right)\left(\left\lfloor\frac{g-4}{12}\right\rfloor+1\right)+\left(3\left\lfloor\frac{g-9}{12}\right\rfloor+4\right)\left(\left\lfloor\frac{g-9}{12}\right\rfloor+1\right)\right), & g\equiv 4 \pmod 6,\\ \frac{1}{2}\left(\left(3\left\lfloor\frac{g+4}{12}\right\rfloor+3\right)\left(\left\lfloor\frac{g+4}{12}\right\rfloor\right)+\left(3\left\lfloor\frac{g-5}{12}\right\rfloor+2\right)\left(\left\lfloor\frac{g-5}{12}\right\rfloor+1\right)\right), & g\equiv 5 \pmod 6.\end{cases}$$

The branches with root $\langle 4, 4k+3, 4(k+1)+1, 4(k+1)+2\rangle$ behave similarly to the previous case. Repeating the argument, we obtain the following counting function, valid for $g \geq 4$:

$$\#_3(g)=\begin{cases}\frac{1}{2}\left(\left(3\left\lfloor\frac{g-4}{12}\right\rfloor+4\right)\left(\left\lfloor\frac{g-4}{12}\right\rfloor+1\right)+\left(3\left\lfloor\frac{g-7}{12}\right\rfloor+4\right)\left(\left\lfloor\frac{g-7}{12}\right\rfloor+1\right)\right), & g\equiv 0 \pmod 6,\\ \frac{1}{2}\left(\left(3\left\lfloor\frac{g+4}{12}\right\rfloor+3\right)\left(\left\lfloor\frac{g+4}{12}\right\rfloor\right)+\left(3\left\lfloor\frac{g-7}{12}\right\rfloor+4\right)\left(\left\lfloor\frac{g-7}{12}\right\rfloor+1\right)\right), & g\equiv 1 \pmod 6,\\ \frac{1}{2}\left(\left(3\left\lfloor\frac{g}{12}\right\rfloor+2\right)\left(\left\lfloor\frac{g}{12}\right\rfloor+1\right)+\left(3\left\lfloor\frac{g-3}{12}\right\rfloor+2\right)\left(\left\lfloor\frac{g-3}{12}\right\rfloor+1\right)\right), & g\equiv 2 \pmod 6,\\ \frac{1}{2}\left(\left(3\left\lfloor\frac{g-4}{12}\right\rfloor+4\right)\left(\left\lfloor\frac{g-4}{12}\right\rfloor+1\right)+\left(3\left\lfloor\frac{g-3}{12}\right\rfloor+2\right)\left(\left\lfloor\frac{g-3}{12}\right\rfloor+1\right)\right), & g\equiv 3 \pmod 6,\\ \frac{1}{2}\left(\left(3\left\lfloor\frac{g+4}{12}\right\rfloor+3\right)\left(\left\lfloor\frac{g+4}{12}\right\rfloor\right)+\left(3\left\lfloor\frac{g+1}{12}\right\rfloor+3\right)\left(\left\lfloor\frac{g+1}{12}\right\rfloor\right)\right), & g\equiv 4 \pmod 6,\\ \frac{1}{2}\left(\left(3\left\lfloor\frac{g}{12}\right\rfloor+2\right)\left(\left\lfloor\frac{g}{12}\right\rfloor+1\right)+\left(3\left\lfloor\frac{g+1}{12}\right\rfloor+3\right)\left(\left\lfloor\frac{g+1}{12}\right\rfloor\right)\right), & g\equiv 5 \pmod 6\end{cases}$$

Finally we sum these functions. $N(4, g) = \#_1(g) + \#_2(g) + \#_3(g)$. By working modulo 12, the least common multiple of the denominators, we can eliminate the floor functions. Then a simple computation reduces the function to

$$N(4,g)=\begin{cases}\frac{g^2}{12}+\frac{g}{2}, & g\equiv 0 \pmod 6,\\ \frac{g^2}{12}+\frac{g}{2}-\frac{7}{12}, & g\equiv 1,5 \pmod 6,\\ \frac{g^2}{12}+\frac{g}{2}-\frac{1}{3}, & g\equiv 2,4 \pmod 6,\\ \frac{g^2}{12}+\frac{g}{2}-\frac{1}{4}, & g\equiv 3 \pmod 6.\end{cases}$$

□

Theorem 28 uses the structure of the numerical semigroup tree in multiplicity 4 to find a concise quasipolynomial expression for $N(4, g)$. We suspect that a similar approach could yield a relatively concise quasipolynomial expression for $N(5, g)$, but the computations are much more complicated.

## References

[Alhajjar et al. 2019] E. Alhajjar, T. Russell, and M. Steward, "Numerical semigroups and Kunz polytopes", *Semigroup Forum* **99**:1 (2019), 153–168. MR Zbl

[Bras-Amorós 2008] M. Bras-Amorós, "Fibonacci-like behavior of the number of numerical semigroups of a given genus", *Semigroup Forum* **76**:2 (2008), 379–384. MR Zbl

[Bras-Amorós and Bulygin 2009] M. Bras-Amorós and S. Bulygin, "Towards a better understanding of the semigroup tree", *Semigroup Forum* **79**:3 (2009), 561–574. MR Zbl

[Bras-Amorós and Fernández-González 2019] M. Bras-Amorós and J. Fernández-González, "The right-generators descendant of a numerical semigroup", preprint, 2019. arXiv

[Eliahou and Fromentin 2019] S. Eliahou and J. Fromentin, "Gapsets of small multiplicity", preprint, 2019. arXiv

[Fromentin and Hivert 2016] J. Fromentin and F. Hivert, "Exploring the tree of numerical semigroups", *Math. Comp.* **85**:301 (2016), 2553–2568. MR Zbl

[García-Sánchez et al. 2018] P. A. García-Sánchez, D. Marín-Aragón, and A. M. Robles-Pérez, "The tree of numerical semigroups with low multiplicity", preprint, 2018. arXiv

[Kaplan 2012] N. Kaplan, "Counting numerical semigroups by genus and some cases of a question of Wilf", *J. Pure Appl. Algebra* **216**:5 (2012), 1016–1032. MR Zbl

[Rosales and García-Sánchez 2009] J. C. Rosales and P. A. García-Sánchez, *Numerical semigroups*, Developments in Mathematics **20**, Springer, 2009. MR Zbl

[Zhai 2013] A. Zhai, "Fibonacci-like growth of numerical semigroups of a given genus", *Semigroup Forum* **86**:3 (2013), 634–662. MR Zbl

Received: 2019-08-06 Revised: 2019-12-16 Accepted: 2020-01-15

abbygreco1997@gmail.com *Department of Mathematical Sciences, United States Military Academy at West Point, West Point, NY, United States*

jesse.lansford@westpoint.edu *Department of Mathematical Sciences, United States Military Academy at West Point, West Point, NY, United States*

michael.steward@westpoint.edu *Department of Mathematical Sciences, United States Military Academy at West Point, West Point, NY, United States*

INVOLVE 13:2 (2020)
dx.doi.org/10.2140/involve.2020.13.323

# Enumerating diagonalizable matrices over $\mathbb{Z}_{p^k}$

Catherine Falvey, Heewon Hah, William Sheppard,
Brian Sittinger and Rico Vicente

(Communicated by Stephan Garcia)

Although a good portion of elementary linear algebra concerns itself with matrices over a field such as $\mathbb{R}$ or $\mathbb{C}$, many combinatorial problems naturally surface when we instead work with matrices over a finite field. As some recent work has been done in these areas, we turn our attention to the problem of enumerating the square matrices with entries in $\mathbb{Z}_{p^k}$ that are diagonalizable over $\mathbb{Z}_{p^k}$. This turns out to be significantly more nontrivial than its finite-field counterpart due to the presence of zero divisors in $\mathbb{Z}_{p^k}$.

## 1. Introduction

A classic problem in linear algebra concerns whether a matrix $A \in M_n(K)$ (where $K$ is a field) is diagonalizable; that is, there exists an invertible matrix $P \in \mathrm{GL}_n(K)$ and a diagonal matrix $D \in M_n(K)$ such that $A = PDP^{-1}$. It is known that if $A$ is diagonalizable, then $D$ is unique up to the order of its diagonal elements. Besides being useful for computing functions of matrices (and therefore often giving a solution to a system of linear differential equations), this problem has applications in the representation of quadratic forms.

If we consider $M_n(K)$ when $K$ is a finite field, one natural problem is to enumerate $\mathrm{Eig}_n(K)$, the set of $n \times n$ matrices over $K$ whose $n$ eigenvalues, counting multiplicity, are in $K$. Olšavský [2003] initiated this line of inquiry and determined that for any prime $p$

$$|\mathrm{Eig}_2(\mathbb{F}_p)| = \tfrac{1}{2}(p^4 + 2p^3 - p^2).$$

More recently, Kaylor and Offner [2014] gave a procedure to enumerate $\mathrm{Eig}_n(\mathbb{F}_q)$, thereby extending Olsavsky's work for any $n$ and any finite field $\mathbb{F}_q$.

Inspired by these works, we turn our attention to $n \times n$ matrices over $\mathbb{Z}_{p^k}$, where $p$ is a prime and $k$ is a positive integer. More specifically, we investigate the problem of enumerating $\mathrm{Diag}_n(\mathbb{Z}_{p^k})$, the set of $n \times n$ diagonalizable matrices over $\mathbb{Z}_{p^k}$. This is

*MSC2010:* 05A05, 05C22, 15A18, 15B33.
*Keywords:* eigenvalues, matrices, finite commutative rings.

significantly more involved when $k \geq 2$, and many of the difficulties arise from having to carefully consider the zero divisors of $\mathbb{Z}_{p^k}$, namely any integral multiple of $p$.

In Section 2, we review the pertinent definitions and notation for working with matrices over commutative rings. Most notably, we give a crucial theorem that essentially states that a diagonalizable matrix over $\mathbb{Z}_{p^k}$ is unique up to the ordering of its diagonal entries. In Section 3, we give the basic procedure for enumerating $\mathrm{Diag}_n(\mathbb{Z}_{p^k})$ and apply it to the case where $n = 2$ in Section 4. In order to deal with the cases where $n \geq 3$ in a systematic manner, we introduce to any diagonal matrix an associated weighted graph in Section 5 that allows us to find $|\mathrm{Diag}_3(\mathbb{Z}_{p^k})|$ and $|\mathrm{Diag}_4(\mathbb{Z}_{p^k})|$ in Sections 6 and 7, respectively. In the final sections, we use our work to find the proportion of matrices that are diagonalizable over $\mathbb{Z}_{p^k}$ and conclude by giving ideas for future research based on the ideas in this article. As far as we understand, all results and definitions from Proposition 3.4 in Section 3 onward are original.

## 2. Background

We give some definitions from matrix theory over rings that allow us to extend some notions of matrices from elementary linear algebra to those having entries in $\mathbb{Z}_{p^k}$. For the following definitions, we let $R$ denote a commutative ring with unity. For further details, we refer the interested reader to [Brown 1993].

To fix some notation, let $M_n(R)$ denote the set of $n \times n$ matrices with entries in $R$. The classic definitions of matrix addition and multiplication as well as determinants generalize in $M_n(R)$ in the expected manner. In general, $M_n(R)$ forms a noncommutative ring with unity $I_n$, the matrix with 1s on its main diagonal and 0s elsewhere.

Next, we let $\mathrm{GL}_n(R)$ denote the set of invertible matrices in $M_n(R)$; that is,

$$\mathrm{GL}_n(R) = \{A \in M_n(R) : AB = BA = I_n \text{ for some } B \in M_n(R)\}.$$

Note that $\mathrm{GL}_n(R)$ forms a group under matrix multiplication and has the alternative characterization

$$\mathrm{GL}_n(R) = \{A \in M_n(R) : \det A \in R^*\},$$

where $R^*$ denotes the group of units in $R$. Observe that when $R$ is a field $K$, we have $K^* = K \setminus \{0\}$; thus we retrieve the classic fact for invertible matrices over $K$. For this article, we are specifically interested in the case when $R = \mathbb{Z}_{p^k}$, where $p$ is prime and $k \in \mathbb{N}$. Then,

$$\mathrm{GL}_n(\mathbb{Z}_{p^k}) = \{A \in M_n(\mathbb{Z}_{p^k}) : \det A \not\equiv 0 \bmod p\};$$

in other words, we can think of an invertible matrix with entries in $\mathbb{Z}_{p^k}$ as having a determinant not divisible by $p$.

**Definition 2.1.** We say that $A \in M_n(R)$ is *diagonalizable over* $R$ if $A$ is similar to a diagonal matrix $D \in M_n(R)$; that is, $A = PDP^{-1}$ for some $P \in \mathrm{GL}_n(R)$.

Recall that any diagonalizable matrix over a field is similar to a distinct diagonal matrix that is unique up to ordering of its diagonal entries. Since $\mathbb{Z}_{p^k}$ is *not* a field whenever $k \geq 2$, we now give a generalization of this key result to matrices over $\mathbb{Z}_{p^k}$. This provides a foundational result that allows us to use the methods from [Kaylor and Offner 2014] to enumerate diagonalizable matrices over $\mathbb{Z}_{p^k}$. Although we originally came up for a proof for this result, the following elegant proof was suggested to the authors by an anonymous MathOverflow user.[1]

**Theorem 2.2.** *Any diagonalizable matrix over $\mathbb{Z}_{p^k}$ is similar to exactly one diagonal matrix that is unique up to the ordering of its diagonal entries.*

*Proof.* Suppose that $D, D' \in M_n(\mathbb{Z}_{p^k})$ are diagonal matrices such that $D' = PDP^{-1}$ for some $P \in \mathrm{GL}_n(\mathbb{Z}_{p^k})$. Writing $D = \mathrm{diag}(d_1, \ldots, d_n)$, $D' = \mathrm{diag}(d'_1, \ldots, d'_n)$, and $P = (p_{ij})$, we see that $D' = PDP^{-1}$ rewritten as $PD = D'P$ yields $p_{ij}d_i = p_{ij}d'_j$ for all $i, j$.

Since $P \in \mathrm{GL}_n(\mathbb{Z}_{p^k})$, we know that $\det P \in {\mathbb{Z}_{p^k}}^*$, and thus $\det P \not\equiv 0 \bmod p$. However, since

$$\det P = \sum_{\sigma \in S_n} (-1)^{\mathrm{sgn}(\sigma)} \prod_i p_{i,\sigma(i)},$$

and the set of nonunits in $\mathbb{Z}_{p^k}$ (which is precisely the subset of elements congruent to $0 \bmod p$) is additively closed, there exists $\sigma \in S_n$ such that $\prod_i p_{i,\sigma(i)} \in {\mathbb{Z}_{p^k}}^*$ and thus $p_{i,\sigma(i)} \in {\mathbb{Z}_{p^k}}^*$ for all $i$.

Then for this choice of $\sigma$, it follows that $p_{i,\sigma(i)}d_i = p_{i,\sigma(i)}d'_{\sigma(i)}$ for each $i$, and since $p_{i,\sigma(i)} \in {\mathbb{Z}_{p^k}}^*$, we deduce that $d_i = d'_{\sigma(i)}$ for each $i$. In other words, $\sigma$ is a permutation of the diagonal entries of $D$ and $D'$, giving us the desired result. □

**Remark.** Theorem 2.2 does not extend to $\mathbb{Z}_m$ for a modulus $m$ with more than one prime factor. As an example from [Brown 1993], the matrix

$$\begin{pmatrix} 2 & 3 \\ 4 & 3 \end{pmatrix} \in M_2(\mathbb{Z}_6)$$

has two distinct diagonalizations

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

The resulting diagonal matrices are thus similar over $\mathbb{Z}_6$ although their diagonal entries are not rearrangements of one another.

[1] See the response from user 44191 at https://mathoverflow.net/questions/303634/uniqueness-of-diagonalizing-a-matrix-over-mathbbz-pk.

## 3. How to determine $|\mathrm{Diag}_n(\mathbb{Z}_{p^k})|$

We give a procedure that allows us to determine $|\mathrm{Diag}_n(\mathbb{Z}_{p^k})|$, the number of matrices in $M_n(\mathbb{Z}_{p^k})$ that are diagonalizable over $\mathbb{Z}_{p^k}$. The main idea is to use a generalization of [Kaylor and Offner 2014, Lemma 3.1]. Before stating it, we first fix some notation in the following definition.

**Definition 3.1.** Let $R$ be a commutative ring with 1, and fix $A \in M_n(R)$.

- The *similarity* (*conjugacy*) *class* of $A$, denoted by $S(A)$, is the set of matrices similar to $A$:

$$S(A) = \{B \in M_n(R) : B = PAP^{-1} \text{ for some } P \in \mathrm{GL}_n(R)\}.$$

- The *centralizer* of $A$, denoted by $C(A)$, is the set of invertible matrices that commute with $A$:

$$C(A) = \{P \in \mathrm{GL}_n(R) : PA = AP\}.$$

Note that $P \in C(A)$ if and only if $A = PAP^{-1}$, and moreover $C(A)$ is a subgroup of $\mathrm{GL}_n(R)$.

**Lemma 3.2.** *Let $R$ be a finite commutative ring. For any $A \in M_n(R)$, we have*

$$|S(A)| = \frac{|\mathrm{GL}_n(R)|}{|C(A)|}.$$

*Proof.* This is proved verbatim as Lemma 3.1 in [Kaylor and Offner 2014] upon replacing a finite field with a finite commutative ring. Alternatively, this is a direct consequence of the orbit-stabilizer theorem where $\mathrm{GL}_n(R)$ is acting on $M_n(R)$ via conjugation. □

To see how this helps us in $M_n(\mathbb{Z}_{p^k})$, recall by Theorem 2.2 that the similarity class of a given diagonalizable matrix can be represented by a unique diagonal matrix (up to the ordering of diagonal entries). Therefore, we can enumerate $\mathrm{Diag}_n(\mathbb{Z}_{p^k})$ by first enumerating the diagonal matrices in $M_n(\mathbb{Z}_{p^k})$ and then counting how many matrices in $M_n(\mathbb{Z}_{p^k})$ are similar to a given diagonal matrix. Then, Lemma 3.2 yields

$$|\mathrm{Diag}_n(\mathbb{Z}_{p^k})| = \sum_{D \in M_n(\mathbb{Z}_{p^k})} |S(D)| = \sum_{D \in M_n(\mathbb{Z}_{p^k})} \frac{|\mathrm{GL}_n(\mathbb{Z}_{p^k})|}{|C(D)|}, \tag{1}$$

where it is understood that each diagonal matrix $D$ represents a distinct similarity class of diagonal matrices. Observe that diagonal matrices having the same diagonal entries up to order belong to the same similarity class and are counted as different matrices when computing the size of their similarity class.

First, we give a formula for $|\mathrm{GL}_n(\mathbb{Z}_{p^k})|$. As this seems to be surprisingly not well known, we state and give a self-contained proof of this result inspired by [Bollman and Ramírez 1969]; for a generalization, see [Han 2006].

**Lemma 3.3.** $$|\mathrm{GL}_n(\mathbb{Z}_{p^k})| = p^{n^2(k-1)} \prod_{l=1}^{n} (p^n - p^{l-1}).$$

*Proof.* First, we compute $|\mathrm{GL}_n(\mathbb{Z}_p)|$ by enumerating the possible columns of its matrices. For $A \in \mathrm{GL}_n(\mathbb{Z}_p)$, there are $p^n - 1$ choices for the first column of $A$, as the zero column vector is never linearly independent. Next, we fix $l \in \{2, 3, \ldots, n\}$. After having chosen the first $l-1$ columns, there are $(p^n - 1) - (p^{l-1} - 1) = p^n - p^{l-1}$ choices for the $l$-th column, because we want these $l$ columns to be linearly independent over $\mathbb{Z}_p$ (and there are $p$ multiples for each of the first $l-1$ columns). Therefore, we conclude that

$$|\mathrm{GL}_n(\mathbb{Z}_p)| = \prod_{l=1}^{n} (p^n - p^{l-1}).$$

Hereafter, we assume that $k \geq 2$. Consider the mapping $\psi : M_n(\mathbb{Z}_{p^k}) \to M_n(\mathbb{Z}_p)$ defined by $\psi(A) = A \mod p$; note that $\psi$ is a well-defined (due to $p \mid p^k$) surjective ring homomorphism. Moreover, since

$$\ker \psi = \{A \in M_n(\mathbb{Z}_{p^k}) : \psi(A) = 0 \mod p\}$$

(so that every entry in such a matrix is divisible by $p$), we deduce that

$$|\ker \psi| = \left(\frac{p^k}{p}\right)^{n^2} = p^{(k-1)n^2}.$$

Then, restricting $\psi$ to the respective groups of invertible matrices, the first isomorphism theorem yields

$$\frac{\mathrm{GL}_n(\mathbb{Z}_{p^k})}{\ker \psi} \cong \mathrm{GL}_n(\mathbb{Z}_p).$$

Therefore, we conclude that

$$|\mathrm{GL}_n(\mathbb{Z}_{p^k})| = |\ker \psi|\,|\mathrm{GL}_n(\mathbb{Z}_p)| = p^{n^2(k-1)} \prod_{l=1}^{n} (p^n - p^{l-1}). \qquad \square$$

We next turn our attention to the problem of enumerating the centralizer of a diagonal matrix in $\mathbb{Z}_{p^k}$.

**Proposition 3.4.** *Let $D \in M_n(\mathbb{Z}_{p^k})$ be a diagonal matrix whose distinct diagonal entries $\lambda_1, \ldots, \lambda_g$ have multiplicities $m_1, \ldots, m_g$, respectively. Then,*

$$|C(D)| = \left(\prod_{i=1}^{g} |\mathrm{GL}_{m_i}(\mathbb{Z}_{p^k})|\right)\left(\prod_{j=2}^{g} \prod_{i=1}^{j-1} p^{2m_i m_j l_{ij}}\right),$$

*where $l_{ij}$ is the nonnegative integer satisfying $p^{l_{ij}} \parallel (\lambda_i - \lambda_j)$ for each $i$ and $j$; that is,*

$$\lambda_i - \lambda_j = r p^{l_{ij}} \quad \text{for some } r \in \mathbb{Z}^*_{p^{k-l_{ij}}}.$$

*Proof.* Assume without loss of generality that all matching diagonal entries of $D$ are grouped together; that is, we can think of each $\lambda_i$ with multiplicity $m_i$ as having its own $m_i \times m_i$ diagonal block of the form $\lambda_i I_{m_i}$ within $D$.

To find the centralizer of $D$, we need to account for all $A \in \mathrm{GL}_n(\mathbb{Z}_{p^k})$ such that $AD = DA$. Writing $A = (A_{ij})$, where $A_{ij}$ is an $m_i \times m_j$ block, computing the necessary products and equating like entries yields

$$\lambda_i A_{ij} = \lambda_j A_{ij}.$$

If $i \neq j$, then $(\lambda_i - \lambda_j)A_{ij} \equiv 0 \bmod p^k$. Therefore, $A_{ij} \equiv 0 \bmod p^{k-l_{ij}}$, and thus $A_{ij} \equiv 0 \bmod p$. Observe that this gives $p^{l_{ij}}$ possible values for each entry in $A_{ij}$ (and similarly for those in $A_{ji}$).

Therefore, $A$ is congruent to a block diagonal matrix modulo $p$ with blocks $A_{ii}$ having dimensions $m_i \times m_i$ for each $i \in \{1, \ldots, g\}$. Finally since $A \in \mathrm{GL}_n(\mathbb{Z}_{p^k})$, this means that $A_{ii} \in \mathrm{GL}_{m_i}(\mathbb{Z}_{p^k})$ for all $i$. With this last observation, the formula for $|C(D)|$ now follows immediately. □

Proposition 3.4 motivates the following classification of diagonal matrices in $\mathbb{Z}_{p^k}$.

**Definition 3.5.** Let $D \in M_n(\mathbb{Z}_{p^k})$ be a diagonal matrix whose distinct diagonal entries $\lambda_1, \ldots, \lambda_g$ have multiplicities $m_1, \ldots, m_g$, respectively. The *type* of $D$ is given by the following two quantities:

- the partition $n = m_1 + \cdots + m_g$,
- the set $\{l_{ij}\}$ indexed over all $1 \leq i < j \leq g$, where $p^{l_{ij}} \parallel (\lambda_j - \lambda_i)$.

Then we say that two diagonal matrices $D, D' \in M_n(\mathbb{Z}_{p^k})$ have the *same type* if and only if $D$ and $D'$ share the same partition of $n$, and there exists a permutation $\sigma \in S_n$ such that $l_{ij} = l'_{\sigma(i)\sigma(j)}$ for all $1 \leq i < j \leq g$. We denote the set of all distinct types of diagonal $n \times n$ matrices by $\mathcal{T}(n)$.

**Example.** Consider the following three diagonal matrices from $M_3(\mathbb{Z}_8)$:

$$D_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}, \quad D_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 5 \end{pmatrix}, \quad D_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix}, \quad D_4 = \begin{pmatrix} 7 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 7 \end{pmatrix}.$$

Since $D_1$ has partition $1+1+1$, while $D_2$, $D_3$, and $D_4$ have the partition $2+1$, $D_1$ does not have the same type as any of $D_2$, $D_3$, and $D_4$. Moreover, $D_2$ and $D_3$ do not have the same type, because $2^2 \parallel (5-1)$, while $2^1 \parallel (3-1)$. However, $D_3$ and $D_4$ have the same type, because they share the same partition $2+1$ and $2^1$ exactly divides both $3-1$ and $7-5$.

It is easy to verify that if $D$ and $D'$ are two $n \times n$ diagonal matrices of the same type, then $|C(D)| = |C(D')|$ and thus $|S(D)| = |S(D')|$. Consequently for any type $T$, define $c(T)$ and $s(T)$ by $c(T) = |C(D)|$ and $s(T) = |S(D)|$, where $D$ is

any matrix of type $T$. Then, letting $t(T)$ denote the number of diagonal matrices (up to permutations of the diagonal entries) having type $T$, we can rewrite (1) as

$$|\mathrm{Diag}_n(\mathbb{Z}_{p^k})| = \sum_{T \in \mathcal{T}(n)} t(T) \frac{|\mathrm{GL}_n(\mathbb{Z}_{p^k})|}{c(T)}. \tag{2}$$

## 4. Enumerating the 2 × 2 diagonalizable matrices

We now illustrate our procedure for determining the value of $|\mathrm{Diag}_2(\mathbb{Z}_{p^k})|$.

**Theorem 4.1.** *The number of $2 \times 2$ matrices with entries in $\mathbb{Z}_{p^k}$ that are diagonalizable over $\mathbb{Z}_{p^k}$ is*

$$|\mathrm{Diag}_2(\mathbb{Z}_{p^k})| = p^k + \frac{p^{k+1}(p^2-1)(p^{3k}-1)}{2(p^3-1)}.$$

*Proof.* In order to find $|\mathrm{Diag}_2(\mathbb{Z}_{p^k})|$, we need to enumerate all of the $2 \times 2$ diagonal matrix types. First of all, there are two possible partitions of 2, namely 2 and $1+1$. The trivial partition yields one distinct type of diagonal matrix

$$T_1 = \left\{ \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} : \lambda \in \mathbb{Z}_{p^k} \right\},$$

which consists of the $2 \times 2$ scalar matrices. Since there are $p^k$ choices for $\lambda$, we have $t(T_1) = p^k$. Moreover $c(T_1) = |\mathrm{GL}_2(\mathbb{Z}_{p^k})|$, because any invertible matrix commutes with a scalar matrix.

The nontrivial partition $2 = 1+1$ yields the remaining $k$ distinct types of matrices that we index by $i \in \{0, 1, \ldots, k-1\}$:

$$T_2^{(i)} = \left\{ \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} : p^i \parallel (\lambda_1 - \lambda_2) \right\}.$$

Fix $i \in \{0, 1, \ldots, k-1\}$; we now enumerate $t(T_2^{(i)})$ and $c(T_2^{(i)})$. For $t(T_2^{(i)})$, we first observe that there are $p^k$ choices for $\lambda_1$. To find the number of choices for $\lambda_2$, observe that $\lambda_1 - \lambda_2 \equiv r p^i \mod p^k$ for some unique $r \in (\mathbb{Z}_{p^{k-i}})^*$. Hence, there are $\phi(p^{k-i})$ choices for $r$ and thus for $\lambda_2$. (As a reminder, $\phi$ denotes the Euler phi function, and $\phi(p^l) = p^{l-1}(p-1)$.) Since swapping $\lambda_1$ and $\lambda_2$ does not change the similarity class of the diagonal matrix, we conclude that

$$t(T_2^{(i)}) = \frac{p^k \phi(p^{k-i})}{2!}.$$

Next, applying Proposition 3.4 yields

$$c(T_2^{(i)}) = p^{2i} \phi(p^k)^2.$$

Finally, we use (2) to enumerate the $2 \times 2$ diagonal matrices and conclude that

$$\begin{aligned}
|\mathrm{Diag}_2(\mathbb{Z}_{p^k})| &= t(T_1)\frac{|\mathrm{GL}_n(\mathbb{Z}_{p^k})|}{c(T_1)} + \sum_{i=0}^{k-1} t(T_2^{(i)})\frac{|\mathrm{GL}_n(\mathbb{Z}_{p^k})|}{c(T_2^{(i)})} \\
&= p^k + \frac{p^k}{2}\frac{p^{4(k-1)}(p^2-1)(p^2-p)}{\phi(p^k)^2}\sum_{i=0}^{k-1}\frac{\phi(p^{k-i})}{p^{2i}} \\
&= p^k + \frac{p^k}{2}\frac{p^{4(k-1)}(p^2-1)(p^2-p)}{(p^{k-1}(p-1))^2}\sum_{i=0}^{k-1}\frac{p^{k-i-1}(p-1)}{p^{2i}} \\
&= p^k + \frac{p^{4k-2}(p^2-1)}{2}\sum_{i=0}^{k-1}\frac{1}{p^{3i}} \\
&= p^k + \frac{p^{4k-2}(p^2-1)}{2}\frac{1-p^{-3k}}{1-p^{-3}} \quad \text{(using the geometric series)} \\
&= p^k + \frac{p^{k+1}(p^2-1)(p^{3k}-1)}{2(p^3-1)}.
\end{aligned}$$

□

**Remarks.** In the case where $k = 1$, the formula reduces to $\frac{1}{2}(p^4 - p^2 + p)$, which can be found at the end of Section 3 in [Kaylor and Offner 2014] after you remove the contributions from the $2 \times 2$ Jordan block case. Moreover, for the diagonal matrix types corresponding to the nontrivial partition and $i \geq 1$, we are dealing with differences of diagonal entries yielding zero divisors in $\mathbb{Z}_{p^k}$; these scenarios never occur when $k = 1$ because $\mathbb{Z}_p$ is a field.

## 5. Enumerating $n \times n$ diagonal matrices of a given type

***Representing a diagonal matrix with a valuation graph.*** As we increase the value of $n$, the enumeration of $n \times n$ diagonalizable matrices over $\mathbb{Z}_{p^k}$ becomes more involved, because the number of distinct types becomes increasingly difficult to catalog. The difficulties come both from the powers of $p$ dividing the differences of the diagonal entries of the matrix as well as the increasing number of partitions of $n$. In order to aid us in classifying diagonal matrices into distinct types, we introduce an associated graph to help visualize these scenarios.

Let $D \in M_n(\mathbb{Z}_{p^k})$ be diagonal with distinct diagonal entries $\lambda_1, \ldots, \lambda_g \in \mathbb{Z}_{p^k}$. Ordering the elements in $\mathbb{Z}_{p^k}$ by $0 < 1 < 2 < \cdots < p^k - 1$, we can assume without loss of generality that $\lambda_1 < \lambda_2 < \cdots < \lambda_g$ (since $D$ is similar to such a matrix by using a suitable permutation matrix as the change of basis matrix). Associated to $D$, we define its associated weighted complete graph $G_D$ (abbreviated as $G$ when no ambiguity can arise) as follows: we label its $g$ vertices with the diagonal entries $\lambda_1, \lambda_2, \ldots, \lambda_g$, and given the edge between the vertices $\lambda_i$ and $\lambda_j$, we define its weight $l_{ij}$ as the unique nonnegative integer satisfying $p^{l_{ij}} \parallel (\lambda_i - \lambda_j)$.

**Definition 5.1.** Let $D \in M_n(\mathbb{Z}_{p^k})$ be diagonal. We call the weighted complete graph $G$ associated to $D$ as constructed above the *valuation graph* of $D$.

The following fundamental property of such graphs justifies why we call these valuation graphs.

**Proposition 5.2** (triangle inequality). *Let $G$ be a valuation graph. Given vertices $\lambda_a$, $\lambda_b$, and $\lambda_c$ in $G$ and edges $E_{ab}$, $E_{ac}$, and $E_{bc}$, the weights satisfy $l_{bc} \geq \min\{l_{ab}, l_{ac}\}$. In particular, $l_{bc} = \min\{l_{ab}, l_{ac}\}$ if $l_{ab} \neq l_{ac}$.*

*Proof.* By hypothesis, we know that $l_{ab}$ and $l_{ac}$ are the biggest nonnegative integers satisfying

$$\lambda_a - \lambda_b = rp^{l_{ab}} \quad \text{and} \quad \lambda_a - \lambda_c = sp^{l_{ac}} \quad \text{for some } r, s \in \mathbb{Z}_{p^k}{}^*.$$

Without loss of generality, assume that $l_{ab} \geq l_{ac}$. Then, we obtain

$$\lambda_b - \lambda_c = (\lambda_a - \lambda_c) - (\lambda_a - \lambda_b) = p^{l_{ac}}(s - rp^{l_{ab}-l_{ac}}).$$

If $l_{ab} > l_{ac}$, then $(s - rp^{l_{ab}-l_{ac}}) \in \mathbb{Z}_{p^k}{}^*$, and if $l_{ab} = l_{ac}$ then $s - r$ may or may not be a zero divisor in $\mathbb{Z}_{p^k}$. The claim now immediately follows. □

Observe that since the valuation graph arises from a diagonal matrix in $M_n(\mathbb{Z}_{p^k})$, it is clear that its weights can only attain integral values between 0 and $k-1$ inclusive. In fact, we can give another restriction on the possible values of its weights.

**Lemma 5.3.** *A valuation graph $G$ on $g$ vertices has no more than $g-1$ weights.*

*Proof.* We prove this by induction on the number of vertices $g$. This claim is true for $g = 2$, because such a graph has exactly one weight. Next, we assume that the claim is true for any valuation graph on $g$ vertices, and consider a valuation graph $G$ with vertices $\lambda_1, \ldots, \lambda_{g+1}$. By the inductive hypothesis, the valuation subgraph $H$ of $G$ with vertices $\lambda_1, \ldots, \lambda_g$ has no more than $g-1$ weights. It remains to consider the weights of the edges from these vertices to the remaining vertex $\lambda_{g+1}$. If none of these edges have any of the $g-1$ weights of $H$, then we are done. Otherwise, suppose that one of these edges (call it $E$) has an additional weight. Then for any edge $E'$ other than $E$ that has $\lambda_{g+1}$ as a vertex, the triangle inequality (Proposition 5.2) implies that $E'$ has no new weight. Hence, $G$ has no more than $(g-1)+1 = g$ weights as required, and this completes the inductive step. □

We know that for any diagonal matrix $D \in M_n(\mathbb{Z}_{p^k})$, its valuation graph $G$ satisfies the triangle inequality. Moreover, any complete graph on $n$ vertices satisfying the triangle inequality necessarily corresponds to a collection of diagonal matrices with distinct diagonal entries in $M_n(\mathbb{Z}_{p^k})$ as long as there are at most $n-1$ weights and the maximal weight is at most $k-1$. Moreover, such a graph also corresponds to a collection of diagonal matrices with nondistinct diagonal entries in $M_N(\mathbb{Z}_{p^k})$, where $N$ is the sum of these multiplicities.

***Enumerating diagonalizable matrices with a given valuation graph.*** Throughout this section, we assume that the diagonal matrix in $M_n(\mathbb{Z}_{p^k})$ has distinct diagonal entries. Given its valuation graph $G$, we construct a specific kind of spanning tree that will aid us in enumerating the diagonal matrices in $M_n(\mathbb{Z}_{p^k})$ having valuation graph $G$. In a sense, such a spanning tree concisely shows the dependencies among the diagonal entries of a given diagonal matrix.

**Proposition 5.4.** *Given a diagonal matrix $D \in M_n(\mathbb{Z}_{p^k})$ with distinct diagonal entries having valuation graph $G$, there exists a spanning tree $T \subset G$ from which we can uniquely reconstruct $G$. We call $T$ a permissible spanning tree of $G$.*

*Proof.* Suppose that $G$ is a valuation graph on $n$ vertices with $r$ distinct weights $a_1, a_2, \ldots, a_r$ listed in increasing order. In order to construct a permissible spanning tree for $G$, we consider the following construction.

For each weight $a_i$ with $1 \leq i \leq r$, define $G_{a_i}$ to be the subgraph of $G$ consisting of the edges with weight *at most* $a_i$ along with their respective vertices. From the definition of a weight, we immediately see that $G_{a_1} \supseteq G_{a_2} \supseteq \cdots \supseteq G_{a_r}$. Moreover, Proposition 5.2 implies that each connected component of $G_{a_i}$ is a complete subgraph of $G$.

To use these subgraphs to construct a permissible spanning tree for $G$, we start with the edges in $G_{a_r}$. For each connected component of $G_{a_r}$, we select a spanning tree and include all of their edges into the edge set $E$. Next, we consider the edges in $G_{a_{r-1}}$. For each connected component of $G_{a_{r-1}}$, we select a spanning tree that includes the spanning tree from the previous step. We inductively repeat this process until we have added any pertinent edges from $G_{a_1}$. (Note that since $G_{a_1}$ contains only one connected component, $T$ must also be connected.) The result is a desired permissible spanning tree $T$ for our valuation graph $G$.

Next, we show how to uniquely reconstruct the valuation graph $G$ from $T$. To aid in this procedure, we say that the *completing edge* of two edges $e_1, e_2$ in $G$ that share a vertex is the edge $e_3$ which forms a complete graph $K_3$ with $e_1$ and $e_2$.

Start by looking at the edges having the largest weight $a_r$ in $T$. If two edges with weight $a_r$ share a vertex, then their completing edge in $G$ must also have weight $a_r$ by the maximality of $a_r$. Upon completing this procedure, there can be no other edges in $G$ of weight $a_r$, as this would violate the construction of $T$.

Next consider the edges having weight $a_{r-1}$ (if they exist). For any two edges of weight $a_{r-1}$ that share a vertex, their completing edge must have weight $a_{r-1}$ or $a_r$ by the triangle inequality. If the completing edge has weight $a_r$, then we have already included this edge in the previous step. Otherwise, we conclude that the completing edge must have weight $a_{r-1}$.

Continuing this process to the lowest edge coloring $a_1$, we reconstruct $G$ as desired. □

We now return to the problem of enumerating diagonal $n \times n$ matrices over $\mathbb{Z}_{p^k}$ of a given type. We begin with the case that $A \in M_n(\mathbb{Z}_{p^k})$ is a diagonal matrix over $\mathbb{Z}_{p^k}$ with distinct diagonal entries. Let $G$ be its associated valuation graph with $r$ distinct weights $a_1, a_2, \ldots, a_r$.

**Definition 5.5.** Let $T$ be a permissible spanning tree of a valuation graph $G$. We say that a subset of edges in $T$ all with weight $a_t$ are *linked* if there exists a subtree $S$ of $T$ containing these edges such that each edge in $S$ has weight at least $a_t$.

We use the notion of linked edges to partition the set of edges from our permissible tree $T$ beyond their weights as follows. Let $L^t$ denote the set of edges in $T$ with weight $a_t$. Then, $L^t$ decomposes into pairwise disjoint sets $L^t_1, \ldots, L^t_{\ell(t)}$ for some positive integer $\ell(t)$, where each $L^t_j$ is a maximal subset of linked edges from $L^t$.

**Definition 5.6.** Let $T$ be a permissible spanning tree for a given valuation graph $G$. For a given weight $a_t$, we say that $L^t_1, \ldots, L^t_{\ell(t)}$ are the *linked cells* of the weight $a_t$.

**Theorem 5.7.** *Let $G$ be a valuation graph having $r$ distinct weights $a_1, a_2, \ldots, a_r$ listed in increasing order, and let $T$ be a permissible spanning tree of $G$ with linked cells $L^t_j$. Then, the total number of diagonal matrix classes having distinct diagonal entries in $M_n(\mathbb{Z}_{p^k})$ with an associated valuation graph isomorphic to $G$ equals*

$$\frac{p^k}{|\mathrm{Aut}(G)|} \prod_{t=1}^{r} \prod_{j=1}^{\ell(t)} \prod_{i=1}^{|L^t_j|} \phi_i(p^{k-a_t}),$$

*where $\phi_i(p^j) = p^j - ip^{j-1}$, and $\mathrm{Aut}(G)$ denotes the set of weighted graph automorphisms of $G$.*

*Proof.* Fix a valuation graph $G$. The key idea is to consider the edges of its permissible spanning tree via linked cells, one weight at a time in descending order. Throughout the proof, we use the following convention: if an edge $E$ has vertices $\lambda_1, \lambda_2$ with $\lambda_2 > \lambda_1$, we refer to the value $\lambda_2 - \lambda_1$ as the *edge difference* associated with $E$.

First consider the edges in the linked cell of the maximal weight $a_r$. Without loss of generality, we start with the edges in $L^r_1$. Since $a_r$ is maximal, we know that $L^r_1$ is itself a tree. For brevity, we let $m = |L^r_1|$. Then, $L^r_1$ has $m$ edges connecting its $m+1$ vertices. We claim that there are $\prod_{i=1}^{m} \phi_i(p^{k-a_r})$ ways to label the values of the edge differences.

To show this, we start by picking an edge in $L^r_1$, and let $\lambda_1$ and $\lambda_2$ denote its vertices. Since $\lambda_2 - \lambda_1 = s_1 p^{a_r}$ for some $s_1 \in \mathbb{Z}^*_{p^{k-a_r}}$, we see that $\lambda_2 - \lambda_1$ can attain $\phi(p^{k-a_r}) = \phi_1(p^{k-a_r})$ distinct values. Next, we pick a second edge in $L^r_1$ that connects to either $\lambda_1$ or $\lambda_2$; without loss of generality (relabeling vertices as needed), suppose it is $\lambda_2$. Letting $\lambda_3$ denote the other vertex of this edge, $\lambda_3 - \lambda_2 = s_2 p^{a_r}$

for some $s_2 \in \mathbb{Z}^*_{p^{k-a_r}}$. However because $a_r$ is the maximal weight in $G$, the edge connecting $\lambda_1$ and $\lambda_3$ also has weight $a_r$. On the other hand, we have

$$\lambda_3 - \lambda_1 = (\lambda_3 - \lambda_2) + (\lambda_2 - \lambda_1) = (s_2 + s_1)p^{a_r}, \quad \text{where } s_2 + s_1 \in \mathbb{Z}^*_{p^{k-a_r}}.$$

Hence, $s_2 \not\equiv -s_1 \mod p^{k-a_r}$, and therefore there are $\phi_1(p^{k-a_r}) - p^{k-a_r-1} = \phi_2(p^{k-a_r})$ possible values for $s_2$. Repeating this procedure, we can assign $\phi_i(p^{k-a_r})$ values to the difference of the vertices from the $i$-th edge in $L_1^r$. Now the claim immediately follows.

The preceding discussion applies to any of the linked cells of weight $a_r$, because edges in distinct linked cells never share a common vertex. Hence, we conclude that the number of possible values of edge differences in $L^r$ equals

$$\prod_{j=1}^{\ell(r)} \prod_{i=1}^{|L_j^r|} \phi_i(p^{k-a_r}).$$

Next, suppose that we have enumerated all edge differences from all linked cells having weights $a_{t+1}, \ldots, a_r$ for some fixed $t$. We now consider linked cells for the weight $a_t$. The procedure proceeds just as before, with the only difference being that two edges of any weight lower than $a_r$ may be linked via some subtree of $T$ containing other higher weights. However this presents no new difficulties.

Fix a linked cell with weight $a_t$ and choose a first edge with vertices $\lambda_{c_1}$ and $\lambda_{c_2}$. As above, this edge corresponds to one of $\phi_1(p^{k-a_t})$ possible differences between values $\lambda_{c_1}$ and $\lambda_{c_2}$. Given another edge linked to the aforementioned edge in this linked cell, it either shares or does not share a vertex with the first edge. We consider these cases separately.

First, suppose the two edges share a common vertex $\lambda_{c_2}$. Then as in the previous case, the connecting edge between $\lambda_{c_1}$ and $\lambda_{c_3}$ must have weight at least $a_t$ (as this edge otherwise has weight greater than $a_t$ and such vertices have been previously considered), and thus we can choose the value for $\lambda_{c_3} - \lambda_{c_2}$ in $\phi_2(p^{k-a_t})$ ways.

Alternatively, suppose that the two edges are connected through already established edges of higher weights on the vertices $\lambda_{d_1}, \lambda_{d_2}, \ldots, \lambda_{d_s}$. Without loss of generality, assume that the vertices $\lambda_{c_1}$ and $\lambda_{c_4}$ are the initial and terminal vertices, respectively, in this second edge. We know that $\lambda_{c_2} - \lambda_{c_1} = rp^{k-a_t}$ and $\lambda_{c_4} - \lambda_{c_3} = r'p^{a_t}$ for some $r, r' \in \mathbb{Z}^*_{p^{k-a_t}}$. Also since the edges connecting $\lambda_{c_2}$ to $\lambda_{d_1}$, $\lambda_{d_s}$ to $\lambda_{c_3}$, and $\lambda_{d_i}$ to $\lambda_{d_j}$ for all $1 \le i < j \le s$ have weights higher than $a_t$, it follows that

$$0 \equiv \lambda_{d_1} - \lambda_{c_2} \equiv \lambda_{c_3} - \lambda_{d_s} \equiv \lambda_{d_j} - \lambda_{d_i} \mod p^{a_t+1}$$

and these observations give us

$$\begin{aligned} \lambda_{c_4} - \lambda_{c_1} &\equiv (\lambda_{c_2} - \lambda_{c_1}) + (\lambda_{d_1} - \lambda_{c_2}) + (\lambda_{d_2} - \lambda_{d_1}) + \cdots + (\lambda_{c_3} - \lambda_{d_s}) + (\lambda_{c_4} - \lambda_{c_3}) \\ &\equiv (r + r')p^{a_t} \mod p^{a_t+1}. \end{aligned}$$

However, by an inductive use of the triangle inequality, we see that the edge directly connecting $c_1$ and $c_4$ must have weight $a_t$. Thus, $r + r' \not\equiv 0 \bmod p$, and the number of permissible choices for $r'$ is therefore

$$p^{k-a_t} - 2p^{k-a_t-1} = \phi_2(p^{k-a_t}).$$

Continuing this process, we can see that when we add the $i$-th edge in this linked cell (if it exists), we can find a path between it and the previous $i-1$ edges in $T$ sharing the same linked cell, giving $\phi_i(p^{k-a_t})$ choices for the corresponding edge differences.

At this point we have considered every edge in $T$. The number of possible edge differences among all of the edges in $T$ equals

$$\prod_{t=1}^{r} \prod_{j=1}^{\ell(t)} \prod_{i=1}^{|L_j^t|} \phi_i(p^{k-a_t}).$$

In summary, we have specified the number of values that the differences of the vertices to each of the edges in our permissible tree can attain. Consequently, as soon as we specify the value of one vertex, in which there are $p^k$ possible choices, we have uniquely determined (by our work above) the values of the remaining vertices through their differences. Therefore, the number of possible diagonal matrices with the given valuation graph equals

$$p^k \prod_{t=1}^{r} \prod_{j=1}^{\ell(t)} \prod_{i=1}^{|L_j^t|} \phi_i(p^{k-a_t}).$$

Finally, we note that permuting the order of the diagonal entries of any diagonal matrix associated with $G$ yields a valuation graph isomorphic to $G$. Since these correspond to the weighted graph automorphisms of $G$, dividing our last formula by $|\mathrm{Aut}(G)|$ yields the desired enumeration formula. □

**Remark.** Note that the group of weighted automorphisms of $G$ is a subgroup of all automorphisms (under composition of isomorphisms) of the corresponding unweighted graph version of $G$. Since $G$ is a complete graph with $n$ vertices, we know that there are $|S_n| = n!$ unweighted graph automorphisms of $G$ (which can be represented by $n \times n$ permutation matrices). Then, Lagrange's theorem for groups implies that $|\mathrm{Aut}(G)| = n!/\sigma(G)$, where $\sigma(G) = [S_n : \mathrm{Aut}(G)]$ denotes the number of vertex permutations yielding nonisomorphic valuation graphs from $G$. In this manner, one can determine the value of $|\mathrm{Aut}(G)|$ by directly computing $\sigma(G)$.

So far, Theorem 5.7 allows us to enumerate diagonal matrices with distinct diagonal entries with an associated valuation graph. The following proposition addresses how to extend this theorem to also enumerate diagonal matrices whose diagonal entries are not distinct.

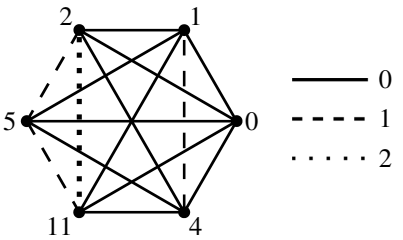



**Figure 1.** The valuation graph $G$ corresponding to $D$.

**Proposition 5.8.** *Let $D \in M_n(\mathbb{Z}_{p^k})$ be a diagonal matrix with distinct diagonal entries $\lambda_1, \ldots, \lambda_g$, and let $D' \in M_g(\mathbb{Z}_{p^k})$ be the corresponding diagonal matrix with (distinct) diagonal entries $\lambda_1, \ldots, \lambda_g$. If $D$ has exactly $n_m$ distinct $m \times m$ diagonal blocks for each $m \in \{1, 2, \ldots, g\}$, then*

$$t(T) = \frac{g!}{n_1! \cdots n_g!} t(T'),$$

*where $T$ and $T'$ are the types of $D$ and $D'$, respectively.*

*Proof.* Since we know by hypothesis that $D$ and $D'$ share the same number of distinct diagonal entries, it suffices to count the number of ways to arrange the diagonal blocks (each of which is distinguished by a different scalar on their respective diagonals) in $D$. Since the number of ways of arranging these diagonal blocks in $D$ equals $g!/(n_1! \cdots n_g!)$, the conclusion of this theorem is now an immediate consequence. □

Now that we have Theorem 5.7 and Proposition 5.8 at our disposal, we are more than ready to enumerate the diagonalizable $n \times n$ matrices in the cases where $n = 3$ and 4; this we address in the next two sections. Before doing this, we would like to put our theory of valuation graphs into perspective by giving an example that illustrates the theory we have developed for the valuation graph.

**Example.** Consider the diagonal matrix $D \in M_6(\mathbb{Z}_{3^3})$ whose diagonal entries are 0, 1, 2, 4, 5, and 11. Then, its corresponding valuation graph $G$ is depicted in Figure 1. Observe the number of distinct weights in $G$ is 3, consistent with Lemma 5.3, and that the highest edge weight is 2.

Next, we give examples of permissible spanning trees for $G$ and partition their edges into linked cells. Figure 2 shows three permissible spanning trees $T_1, T_2, T_3$ for $G$ and their linked cells $L_1^1$, $L_1^2$, $L_2^2$, and $L_1^3$.

Although each of these spanning trees have different degrees, they all have the same edge decomposition into linked cells. Thus, we can use any of these permissible spanning trees to enumerate the number of similarity classes of diagonal matrices sharing $G$ as their valuation graph. To this end, it remains to compute

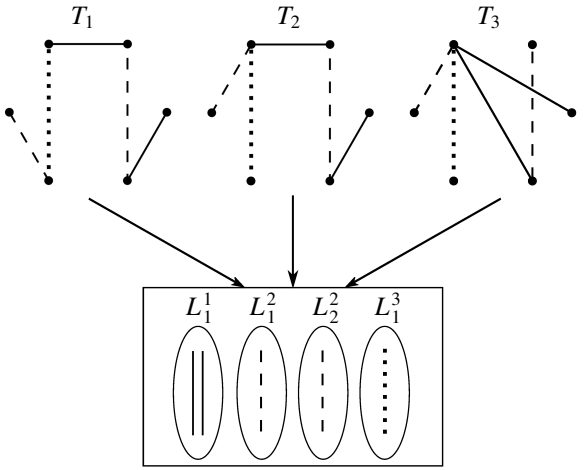



**Figure 2.** Three permissible spanning trees for $G$ and their linked cells.

$|\mathrm{Aut}(G)|$. Since we can permute the vertices 2 and 11, as well as the vertices 1 and 4 without altering $G$, this implies $|\mathrm{Aut}(G)| = 2!\,2!$. Therefore by Theorem 5.7, the number of similarity classes of diagonal matrices with valuation graph $G$ equals

$$\frac{3^3}{2!\,2!}\prod_{t=0}^{2}\prod_{j=1}^{\ell(t)}\prod_{i=1}^{|L_j^t|}\phi_i(3^{3-t}) = \frac{27}{4}\,\phi_1(3^3)\,\phi_2(3^3)\,\phi_1(3^2)\,\phi_1(3^2)\,\phi_1(3^1) = 78732.$$

## 6. Enumerating the 3 × 3 diagonalizable matrices

**Theorem 6.1.** *The number of* $3\times 3$ *matrices with entries in* $\mathbb{Z}_{p^k}$ *that are diagonalizable over* $\mathbb{Z}_{p^k}$ *is*

$$|\mathrm{Diag}_3(\mathbb{Z}_{p^k})| = p^k + \frac{p^{k+2}(p^3-1)(p^{5k}-1)}{p^5-1} + \frac{p^{k+3}(p^3-1)(p-2)(p+1)(p^{8k}-1)}{6(p^8-1)} + \frac{p^{k+3}(p^2-1)}{2}\left(\frac{p^{8k}-p^8}{p^8-1} - \frac{p^{5k}-p^5}{p^5-1}\right).$$

*Proof.* We first enumerate all of the $3\times 3$ diagonal matrix types. There are three partitions of 3, namely 3, $2+1$, and $1+1+1$. The trivial partition yields the type of scalar matrices

$$T_1 = \left\{\begin{pmatrix}\lambda & & \\ & \lambda & \\ & & \lambda\end{pmatrix} : \lambda \in \mathbb{Z}_{p^k}\right\}.$$

As with this type of $2\times 2$ scalar diagonal matrices, we have $t(T_1) = p^k$ and $c(T_1) = |\mathrm{GL}_3(\mathbb{Z}_{p^k})|$.

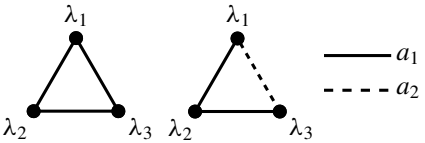



**Figure 3.** Two valuation graph classes in the $3 \times 3$ case.

The partition $3 = 2 + 1$ comprises $k$ distinct types as $i \in \{0, 1, \ldots, k-1\}$:

$$T_2^{(i)} = \left\{ \begin{pmatrix} \lambda_1 & & \\ & \lambda_1 & \\ & & \lambda_2 \end{pmatrix} : p^i \parallel (\lambda_1 - \lambda_2) \right\}.$$

Proposition 5.8 relates these types to the nonscalar types of $2 \times 2$ diagonal matrices, and thus

$$t(T_2^{(i)}) = \frac{2!}{1!\,1!} \frac{p^k \phi(p^{k-i})}{2!} = p^k \phi(p^{k-i}).$$

Next, Proposition 3.4 gives us $c(T_2^{(i)}) = \phi(p^k) \cdot |\mathrm{GL}_2(\mathbb{Z}_{p^k})| \cdot p^{4i}$.

Finally, the partition $3 = 1 + 1 + 1$ comprises two distinct classes of diagonal matrix types that we concisely give by their respective valuation graphs in Figure 3.

For the first valuation graph, let $i \in \{0, 1, \ldots, k-1\}$ denote the common weight of the three edges on the first valuation graph given above. Letting $T_{3a}^{(i)}$ denote this type, Theorem 5.7 yields

$$t(T_{3a}^{(i)}) = \frac{p^k \phi(p^{k-i}) \phi_2(p^{k-i})}{3!},$$

and Proposition 3.4 gives us $c(T_{3a}^{(i)}) = \phi(p^k)^3 p^{6i}$.

For the second valuation graph, let $i$ and $j$ denote the weights in the second valuation graph given above; note that $i \in \{0, \ldots, k-2\}$ and $j \in \{i+1, \ldots, k-1\}$. Letting $T_{3b}^{(i,j)}$ denote this type, Theorem 5.7, gives us

$$t(T_{3b}^{(i,j)}) = \frac{p^k \phi(p^{k-i}) \phi(p^{k-j})}{2!},$$

and Proposition 3.4 yields $c(T_{3b}^{(i,j)}) = \phi(p^k)^3 p^{4i+2j}$.

Finally, we use (2) to enumerate the $3 \times 3$ diagonal matrices and conclude that

$$|\mathrm{Diag}_3(\mathbb{Z}_{p^k})| = p^k + \frac{p^{k+2}(p^3-1)(p^{5k}-1)}{p^5-1} + \frac{p^{k+3}(p^3-1)(p-2)(p+1)(p^{8k}-1)}{6(p^8-1)} + \frac{p^{k+3}(p^2-1)}{2} \left( \frac{p^{8k}-p^8}{p^8-1} - \frac{p^{5k}-p^5}{p^5-1} \right). \quad \square$$

## 7. Enumerating the 4 × 4 diagonalizable matrices

We first address the $4 \times 4$ diagonal matrices with repeated diagonal entries. By using Propositions 3.4 and 5.8, we obtain the results in the following tables. Table 1 deals with the cases where there are at most two distinct diagonal entries.

In Table 2, we consider the more involved case where a given diagonal matrix has three distinct diagonal entries.

| type $T$ | valuation graph | $t(T)$ | $c(T)$ |
|---|---|---|---|
| $\begin{pmatrix} \lambda & & & \\ & \lambda & & \\ & & \lambda & \\ & & & \lambda \end{pmatrix}$ | $\bullet^{\lambda}$ | $p^k$ | $\lvert \mathrm{GL}_4(\mathbb{Z}_{p^k}) \rvert$ |
| $\begin{pmatrix} \lambda_1 & & & \\ & \lambda_1 & & \\ & & \lambda_1 & \\ & & & \lambda_2 \end{pmatrix}$ | $\lambda_1 \overset{a_1}{\text{—}} \lambda_2$ | $p^k \phi(p^{k-i})$ | $p^{6i} \phi(p^k) \lvert \mathrm{GL}_3(\mathbb{Z}_{p^k}) \rvert$ |
| $\begin{pmatrix} \lambda_1 & & & \\ & \lambda_1 & & \\ & & \lambda_2 & \\ & & & \lambda_2 \end{pmatrix}$ | $\lambda_1 \overset{a_1}{\text{—}} \lambda_2$ | $\dfrac{p^k \phi(p^{k-i})}{2}$ | $p^{8i} \lvert \mathrm{GL}_2(\mathbb{Z}_{p^k}) \rvert^2$ |

**Table 1.** $4 \times 4$ diagonal matrix types with at most two distinct diagonal entries.

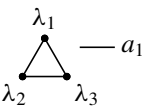
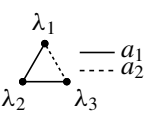
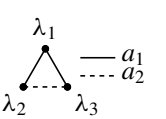
| type $T$ | valuation graph | $t(T)$ | $c(T)$ |
|---|---|---|---|
| $\begin{pmatrix} \lambda_1 & & & \\ & \lambda_1 & & \\ & & \lambda_2 & \\ & & & \lambda_3 \end{pmatrix}$ |   | $\dfrac{p^k \phi(p^{k-i}) \phi_2(p^{k-i})}{2}$ | $p^{10i} \phi(p^k)^2 \lvert \mathrm{GL}_2(\mathbb{Z}_{p^k}) \rvert$ |
| $\begin{pmatrix} \lambda_1 & & & \\ & \lambda_1 & & \\ & & \lambda_2 & \\ & & & \lambda_3 \end{pmatrix}$ |   | $\dfrac{3p^k \phi(p^{k-i}) \phi(p^{k-j})}{2}$ | $p^{6i+4j} \phi(p^k)^2 \lvert \mathrm{GL}_2(\mathbb{Z}_{p^k}) \rvert$ |
| $\begin{pmatrix} \lambda_1 & & & \\ & \lambda_1 & & \\ & & \lambda_2 & \\ & & & \lambda_3 \end{pmatrix}$ |   | $\dfrac{3p^k \phi(p^{k-i}) \phi(p^{k-j})}{2}$ | $p^{8i+2j} \phi(p^k)^2 \lvert \mathrm{GL}_2(\mathbb{Z}_{p^k}) \rvert$ |

**Table 2.** $4 \times 4$ diagonal matrix types with three distinct diagonal entries.

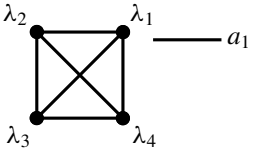
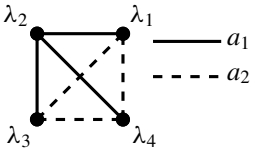
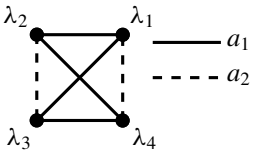
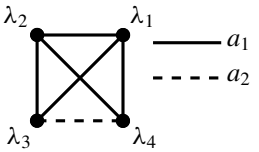
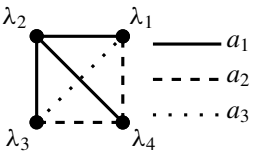
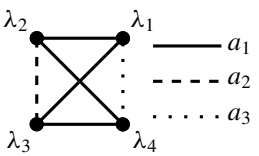
| valuation graph | $t(T)$ | $c(T)$ |
|---|---|---|
|  | $\frac{p^k\phi(p^{k-i})\phi_2(p^{k-i})\phi_3(p^{k-i})}{4!}$ | $p^{12i}\phi(p^k)^4$ |
|  | $\binom{4}{3}\frac{p^k\phi(p^{k-i})\phi(p^{k-j})\phi_2(p^{k-j})}{4!}$ | $p^{6i+6j}\phi(p^k)^4$ |
|  | $\frac{1}{2}\binom{4}{2}\frac{p^k\phi(p^{k-i})\phi(p^{k-j})^2}{4!}$ | $p^{8i+4j}\phi(p^k)^4$ |
|  | $\binom{4}{2}\frac{p^k\phi(p^{k-i})\phi_2(p^{k-i})\phi(p^{k-j})}{4!}$ | $p^{10i+2j}\phi(p^k)^4$ |
|  | $\binom{4}{3}\binom{3}{1}\frac{p^k\phi(p^{k-i})\phi(p^{k-j})\phi(p^{k-m})}{4!}$ | $p^{6i+4j+2m}\phi(p^k)^4$ |
|  | $\binom{4}{4}\binom{4}{2}\frac{p^k\phi(p^{k-i})\phi(p^{k-j})\phi(p^{k-m})}{4!}$ | $p^{8i+2j+2m}\phi(p^k)^4$ |

**Table 3.** $4 \times 4$ diagonal matrix types with distinct diagonal entries.

It remains to enumerate the diagonal matrix types where the diagonal entries are distinct. By inspection, we find that there are six distinct classes of valuation graphs if we disregard the actual weights of their edges. We summarize the pertinent information for each of these six valuation graphs in Table 3.

By using (2), one can now find the number of $4 \times 4$ diagonalizable matrices over $\mathbb{Z}_{p^k}$. In light of the many cases from the three tables above, the final formula will be quite long and messy to explicitly write out, and we therefore have chosen not to include it here (although the curious reader should have no problem constructing it if necessary).

## 8. The proportion of diagonalizable matrices over $\mathbb{Z}_{p^k}$

Kaylor and Offner [2014] noted that as the size of the field $\mathbb{F}_q$ increases, the proportion of matrices in $M_n(\mathbb{F}_q)$ with all eigenvalues in $\mathbb{F}_q$ approaches $1/n!$; that is,

$$\lim_{q\to\infty} \frac{|\mathrm{Eig}_n(\mathbb{F}_q)|}{|M_n(\mathbb{F}_q)|} = \frac{1}{n!}.$$

In particular, [Kaylor and Offner 2014] also implies that as the size of $\mathbb{F}_q$ increases, the proportion of matrices in $M_n(\mathbb{F}_q)$ that are diagonalizable over $\mathbb{F}_q$ approaches $1/n!$ as well. We generalize this latter result by replacing $\mathbb{F}_q$ in the case of $q = p$ with $\mathbb{Z}_{p^k}$.

**Theorem 8.1.** *Fix positive integers $n$ and $k$, and let $p$ be a prime number. Then,*

$$\lim_{p\to\infty} \frac{|\mathrm{Diag}_n(\mathbb{Z}_{p^k})|}{|M_n(\mathbb{Z}_{p^k})|} = \frac{1}{n!}.$$

*Proof.* Letting $i$ index the distinct types of diagonal matrices, we let $T_{n,i}$ denote the $i$-th distinct type of a diagonal matrix in $M_n(\mathbb{Z}_{p^k})$. Note that we can view $|\mathrm{Diag}_n(\mathbb{Z}_{p^k})|$ as a polynomial in powers of $p$. Since we are taking a limit as $p \to \infty$, it suffices to determine which diagonal matrix types contributes to the leading term of $|\mathrm{Diag}_n(\mathbb{Z}_{p^k})|$. We accomplish this by first computing its degree:

$$\begin{aligned}
\deg|\mathrm{Diag}_n(\mathbb{Z}_{p^k})| &= \deg\left(\sum_{i=1}^{|\mathcal{T}(n)|} t(T_{n,i})s(T_{n,i})\right)\\
&= \max_{1\le i\le|\mathcal{T}(n)|} \deg(t(T_{n,i})s(T_{n,i}))\\
&= \max_{1\le i\le|\mathcal{T}(n)|} \deg\left(t(T_{n,i})\frac{|\mathrm{GL}_n(\mathbb{Z}_{p^k})|}{c(T_{n,i})}\right)\\
&= \max_{1\le i\le|\mathcal{T}(n)|} (\deg|\mathrm{GL}_n(\mathbb{Z}_{p^k})| + \deg t(T_{n,i}) - \deg c(T_{n,i}))\\
&= \max_{1\le i\le|\mathcal{T}(n)|} (kn^2 + \deg t(T_{n,i}) - \deg c(T_{n,i})).
\end{aligned}$$

By Proposition 3.4, we find that

$$\begin{aligned}
\deg c(T_{u,i}) &= \sum_{i=1}^{r} \deg|\mathrm{GL}_{m_i}(\mathbb{Z}_{p^k})| + \sum_{1\le i<j\le k} \deg p^{2m_im_jl_{ij}}\\
&= k\sum_{i=1}^{r} m_i^2 + \sum_{1\le i<j\le k} 2m_im_jl_{ij}\\
&\ge k\sum_{i=1}^{r} m_i^2 && \text{(since each } l_{ij}\ge 0)\\
&\ge kn && \text{(since each } m_i\ge 1).
\end{aligned}$$

Moreover, Theorem 5.7 yields

$$\deg t(T_{n,i}) = k + \sum_{t=1}^{r} \sum_{j=1}^{J_t} \sum_{i=1}^{|L_j^{(a_t)}|} (k - a_t) \leq k + \sum_{t=1}^{r} \sum_{j=1}^{J_t} k|L_j^{(a_t)}| = k + k(n-1) = kn.$$

Therefore $\deg t(T_{n,i}) \leq \deg c(T_{n,i})$, with equality occurring if and only if the diagonal matrix is of the type in which its diagonal entries are distinct and their differences are units in $\mathbb{Z}_{p^k}$. Hence, $\deg |\mathrm{Diag}_n(\mathbb{Z}_{p^k})| = kn^2$, and using the aforementioned diagonal matrix type, the leading coefficient of $|\mathrm{Diag}_n(\mathbb{Z}_{p^k})|$ equals $1/n!$ by Theorem 5.7. Thus, we have

$$\frac{|\mathrm{Diag}_n(\mathbb{Z}_{p^k})|}{|M_n(\mathbb{Z}_{p^k})|} = \frac{(1/n!)p^{kn^2} + O(p^{kn^2-1})}{p^{kn^2}}.$$

The desired limit immediately follows by letting $p \to \infty$. □

## 9. Future research

As we have seen, given a ring of the form $\mathbb{Z}_{p^k}$ and a positive integer $n$, we have given a procedure to compute $|\mathrm{Diag}_n(\mathbb{Z}_{p^k})|$. The main difficulty that remains is enumerating the possible valuation graph classes (up to automorphism and disregarding the actual values of the weights) corresponding to $n \times n$ diagonal matrices. As demonstrated in the previous sections, it suffices to enumerate such classes corresponding to $n \times n$ diagonal matrices with distinct diagonal entries; let $a_n$ denote this quantity. We have seen that $a_2 = 3$ and $a_4 = 6$, and it turns out that $a_5 = 20$ (see Figure 4 for these classes). It would be of interest to find at least a recursive formula that determines $a_n$ for a given value of $n$.

In addition to this, it would be of interest to extend our work to include matrices with Jordan canonical forms (JCFs) over $\mathbb{Z}_{p^k}$, that is, matrices similar to a block diagonal matrix comprised of the Jordan matrices

$$\begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & 1 & \cdots & 0 & 0 \\ 0 & 0 & \lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{pmatrix}$$

for some $\lambda \in \mathbb{Z}_{p^k}$. One would have to be careful performing such an enumeration, because it is possible for a given matrix to have more than one distinct JCF over $\mathbb{Z}_{p^k}$. For instance in $\mathbb{Z}_4$, we have

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}^{-1}.$$

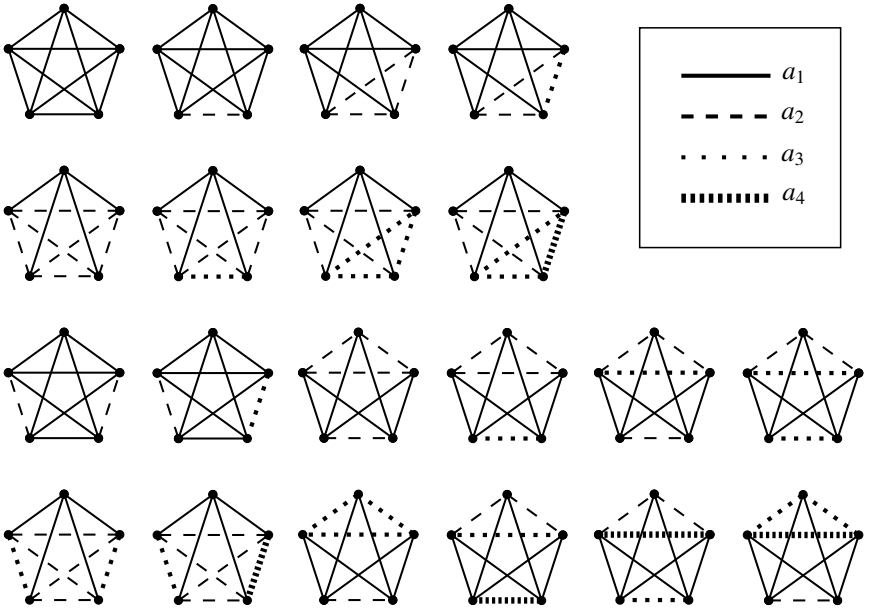



**Figure 4.** The twenty $5 \times 5$ valuation graph classes.

Although finding an enumeration formula for the centralizer of a Jordan matrix should be straightforward, this is not expected to be the case for an arbitrarily chosen JCF.

As a final remark, besides the potential nonuniqueness of a JCF, there is a reason why we have not enumerated $|\mathrm{Eig}_n(\mathbb{Z}_{p^k})|$. Unlike in the finite-field case where any matrix in $\mathrm{Eig}_n(\mathbb{F}_q)$ has a JCF (see [Kaylor and Offner 2014] for more details), this is not even necessarily the case in $\mathrm{Eig}_n(\mathbb{Z}_{p^k})$. For example in $\mathbb{Z}_{p^2}$, any matrix of the form $\left(\begin{smallmatrix} \lambda & p \\ 0 & \lambda \end{smallmatrix}\right)$ has double eigenvalue $\lambda$, but lacks a Jordan canonical form over $\mathbb{Z}_{p^2}$. Determining all similarity classes of such matrices is in general still an open question.

## References

[Bollman and Ramírez 1969] D. Bollman and H. Ramírez, "On the enumeration of matrices over finite commutative rings", *Amer. Math. Monthly* **76** (1969), 1019–1023. MR Zbl

[Brown 1993] W. C. Brown, *Matrices over commutative rings*, Monographs and Textbooks in Pure and Applied Mathematics **169**, Marcel Dekker, New York, 1993. MR Zbl

[Han 2006] J. Han, "The general linear group over a ring", *Bull. Korean Math. Soc.* **43**:3 (2006), 619–626. MR Zbl

[Kaylor and Offner 2014] L. Kaylor and D. Offner, "Counting matrices over a finite field with all eigenvalues in the field", *Involve* **7**:5 (2014), 627–645. MR Zbl

[Olšavský 2003] G. Olšavský, "The number of 2 by 2 matrices over $\mathbb{Z}/p\mathbb{Z}$ with eigenvalues in the same field", *Math. Mag.* **76**:4 (2003), 314–317. MR Zbl

Received: 2019-08-12 Revised: 2019-11-27 Accepted: 2019-12-23

c.falvey24@gmail.com *American University, Washington, D.C., United States*

heewonhah@gmail.com *University of North Carolina, Charlotte, NC, United States*

sheppard@math.ucsb.edu *University of California, Santa Barbara, CA, United States*

bdsittinger@gmail.com *California State University Channel Islands, Camarillo, CA, United States*

ricoevicente@gmail.com *California State University, Long Beach, CA, United States*

INVOLVE 13:2 (2020)
dx.doi.org/10.2140/involve.2020.13.345

# On arithmetical structures on complete graphs

Zachary Harris and Joel Louwsma

(Communicated by Joshua Cooper)

An arithmetical structure on the complete graph $K_n$ with $n$ vertices is given by a collection of $n$ positive integers with no common factor, each of which divides their sum. We show that, for all positive integers $c$ less than a certain bound depending on $n$, there is an arithmetical structure on $K_n$ with largest value $c$. We also show that, if each prime factor of $c$ is greater than $(n+1)^2/4$, there is no arithmetical structure on $K_n$ with largest value $c$. We apply these results to study which prime numbers can occur as the largest value of an arithmetical structure on $K_n$.

## 1. Introduction

How can one have a collection of positive integers, with no common factor, each of which divides their sum? For example, 105, 70, 15, 14, and 6 sum to 210, which is divisible by each of these numbers. Introducing notation, we seek positive integers $r_1, r_2, \ldots, r_n$ with no common factor such that

$$r_j \,\Big|\, \sum_{i=1}^{n} r_i \quad \text{for all } j. \tag{1}$$

It is well known that finding such $r_i$ is equivalent to finding positive integer solutions of the Diophantine equation

$$\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n} = 1. \tag{2}$$

Indeed, given $r_1, r_2, \ldots, r_n$ satisfying (1), dividing both sides of the equation $r_1 + r_2 + \cdots + r_n = \sum_{i=1}^{n} r_i$ by $\sum_{i=1}^{n} r_i$ gives a solution to (2), and, given a solution of (2), the numbers $\operatorname{lcm}(x_1, x_2, \ldots, x_n)/x_i$ satisfy (1) and have no common factor.

Our interest in this question stems from an interest in arithmetical structures. An *arithmetical structure* on a finite, connected graph is an assignment of positive integers to the vertices such that:

*MSC2010:* primary 11D68; secondary 05C50, 11A41.
*Keywords:* arithmetical structure, complete graph, Diophantine equation, Laplacian matrix, prime number.

(a) At each vertex, the integer there is a divisor of the sum of the integers at adjacent vertices (counted with multiplicity if the graph is not simple).

(b) The integers used have no nontrivial common factor.

Arithmetical structures were introduced in [Lorenzini 1989] to study intersections of degenerating curves in algebraic geometry. The usual definition, easily seen to be equivalent to the one given here, is formulated in terms of matrices. From that perspective, an arithmetical structure may be regarded as a generalization of the *Laplacian matrix*, which encodes many important properties of a graph. Notions in this direction that have received a significant amount of attention include the sandpile group and the chip-firing game; for details, see [Corry and Perkinson 2018; Klivans 2019; Glass and Kaplan 2019].

On the *complete graph* $K_n$ with $n$ vertices, positive integers $r_1, r_2, \ldots, r_n$ with no common factor give an arithmetical structure if and only if

$$r_j \;\Big|\; \sum_{\substack{i=1\\ i\neq j}}^{n} r_i \quad \text{for all } j;$$

it is immediate that this condition is equivalent to (1). Therefore, in this language, the opening question of this paper seeks arithmetical structures on complete graphs.

Lemma 1.6 of [Lorenzini 1989] shows that there are finitely many arithmetical structures on any finite, connected graph, but this result does not give a bound on the number of structures. Several recent papers [Braun et al. 2018; Archer et al. 2020; Glass and Wagner 2019] count arithmetical structures on various families of graphs, including path graphs, cycle graphs, bidents, and certain path graphs with doubled edges. However, counting arithmetical structures on complete graphs is a difficult problem. The number of arithmetical structures on $K_n$ for $n \leq 8$ is given in [Sloane 1991]. For general $n$, bounds have been obtained in [Erdős and Graham 1980; Sándor 2003; Browning and Elsholtz 2011; Konyagin 2014], which work from the perspective of the Diophantine equation (2). Other papers such as [Burshtein 2007; 2008; Arce-Nazario et al. 2013] determine, for specific $n$, the number of solutions of (2) satisfying certain additional conditions on the $x_i$.

It is conjectured in [Corrales and Valencia 2018, Conjecture 6.10] that, for any connected, simple graph $G$ with $n$ vertices, the number of arithmetical structures on $G$ is at most the number of arithmetical structures on $K_n$. To approach this conjecture, one would like a better understanding of the types of arithmetical structures that occur on complete graphs. In this direction, this paper studies which positive integers can occur as the largest value of an arithmetical structure on $K_n$. Clearly the $r_i$ of an arithmetical structure can be permuted; in the following we make the assumption $r_1 \geq r_2 \geq \cdots \geq r_n$. We construct arithmetical structures to show that $r_1$ can take certain values and give obstructions to show that it cannot take other values.

Our primary construction theorem (Theorem 1) shows that $r_1$ can take any value up to a certain bound depending on $n$. More specifically, $r_1$ can be any positive integer less than or equal to $\max_{k\in\mathbb{Z}_{>0}}(2^k n - (k+2^k-2)2^k - 1)$. This bound improves somewhat if we restrict attention to prime numbers; $r_1$ can be any prime number less than or equal to $\max_{k\in\mathbb{Z}_{>0}}(2^k n - (k+2^k-3)2^k - 3)$.

We also prove an obstruction theorem (Theorem 7) that shows $r_1$ cannot take any value all of whose prime factors are greater than $(n+1)^2/4$. Restricting attention to prime numbers, this bound improves to show that $r_1$ cannot be any prime number greater than $n^2/4+1$ (Theorem 8).

The final section focuses on the possible prime values $r_1$ can take. We explicitly check prime numbers in the gap between the bound of Theorem 1 and the bound of Theorem 8, showing that $r_1$ can take some of these values but not others. In particular, we observe that there can be prime numbers $p_1$ and $p_2$ with $p_1 < p_2$ such that there is an arithmetical structure on $K_n$ with $r_1 = p_2$ but no arithmetical structure on $K_n$ with $r_1 = p_1$.

## 2. Construction

In this section, we show how to construct arithmetical structures on complete graphs with certain values of $r_1$. Our main construction theorem is the following.

**Theorem 1.** (a) *For any positive integer $c \le \max_{k\in\mathbb{Z}_{>0}}(2^k n - (k+2^k-2)2^k - 1)$, there is an arithmetical structure on $K_n$ with $r_1 = c$.*

(b) *For any prime number $p \le \max_{k\in\mathbb{Z}_{>0}}(2^k n - (k+2^k-3)2^k - 3)$, there is an arithmetical structure on $K_n$ with $r_1 = p$.*

We establish Propositions 2, 4, and 5 on the way to proving Theorem 1.

**Proposition 2.** *For any positive integer $c \le n-1$, there is an arithmetical structure on $K_n$ with $r_1 = c$.*

*Proof.* Let

$$r_i = \begin{cases} c & \text{for } i \in \{1, 2, \ldots, n-c\}, \\ 1 & \text{for } i \in \{n-c+1, n-c+2, \ldots, n\}. \end{cases}$$

Then

$$\sum_{i=1}^{n} r_i = c(n-c) + c = c(n-c+1).$$

Since this is divisible by both $c$ and 1, we have thus produced an arithmetical structure on $K_n$. □

Before turning to Propositions 4 and 5, we establish the following lemma.

**Lemma 3.** (a) *Let $k \in \mathbb{Z}_{\ge 0}$ and $\ell \in \mathbb{Z}_{>0}$ with $k \le \ell$. Every integer $c$ satisfying $\ell \le c \le (\ell-k+1)2^k - 1$ can be expressed as $\sum_{j=1}^{\ell} 2^{k_j}$ for some $k_j \in \{0, 1, \ldots, k\}$, where $k_j = 0$ for some $j \in \{1, 2, \ldots, \ell\}$.*

(b) *Let $k \in \mathbb{Z}_{\geq 0}$ and $\ell \in \mathbb{Z}_{>0}$ with $k \leq \ell$. Every odd integer $c$ satisfying $\ell \leq c \leq (\ell - k + 2)2^k - 3$ can be expressed as $\sum_{j=1}^{\ell} 2^{k_j}$ for some $k_j \in \{0, 1, \ldots, k\}$, where $k_j = 0$ for some $j \in \{1, 2, \ldots, \ell\}$.*

*Proof.* To show (a), we proceed by induction on $c$. In the base case, $c = \ell$, we can let $k_j = 0$ for all $j$ and have $c = \sum_{j=1}^{\ell} 2^{k_j}$. Now suppose $c = \sum_{j=1}^{\ell} 2^{k_j}$ for $c \leq (\ell - k + 1)2^k - 2$. Then $k_j = k$ for at most $\ell - k$ values of $j$, meaning $k_j < k$ for at least $k$ values of $j$. If among these values we had each of $0, 1, \ldots, k-1$ only once, we would then have $\sum_{j=1}^{\ell} 2^{k_j} = (\ell - k + 1)2^k - 1 > (\ell - k + 1)2^k - 2$. Therefore there is some $b < k$ for which $k_{j_1} = b = k_{j_2}$ for some $j_1 \neq j_2$. Define

$$k'_j = \begin{cases} k_j + 1 & \text{for } j = j_1, \\ 0 & \text{for } j = j_2, \\ k_j & \text{otherwise.} \end{cases}$$

Then $\sum_{j=1}^{\ell} 2^{k'_j} = \sum_{j=1}^{\ell} 2^{k_j} + 1 = c + 1$. The result follows.

For (b), first note that if $c \leq (\ell - k + 1)2^k - 1$ then the result follows from (a). Therefore, assume $c > (\ell - k + 1)2^k - 1$ and let $c' = c - ((\ell - k + 1)2^k - 1)$, noting that $c' \leq 2^k - 2$. Since $c$ is odd, $c'$ must be even. Therefore $c'$ can be written in the form $\sum_{j=1}^{k-1} s_j 2^j$, where each $s_j$ is either 0 or 1; the $s_j$ are iteratively determined in reverse by letting $s_j = 1$ if $c' - \sum_{i=j+1}^{k-1} s_i 2^i \geq 2^j$ and letting $s_j = 0$ otherwise. Define

$$k_j = \begin{cases} 0 & \text{for } j = 1, \\ j - 1 + s_{j-1} & \text{for } j \in \{2, 3, \ldots, k\}, \\ k & \text{for } j \in \{k+1, k+2, \ldots, \ell\}. \end{cases}$$

Then

$$\sum_{j=1}^{\ell} 2^{k_j} = \sum_{j=1}^{k} 2^{j-1} + \sum_{j=2}^{k} s_{j-1} 2^{j-1} + \sum_{j=k+1}^{\ell} 2^k = 2^k - 1 + c' + (\ell - k)2^k = c. \quad \square$$

We use Lemma 3 to prove Propositions 4 and 5.

**Proposition 4.** *Fix $n \geq 2$. For any positive integer $k$ satisfying $k + 2^k - 1 \leq n$ and any positive integer $c$ satisfying $n - 2^k + 1 \leq c \leq (n - k - 2^k + 2)2^k - 1$, there is an arithmetical structure on $K_n$ with $r_1 = c$.*

*Proof.* Let $r_i = c$ for all $i \in \{1, 2, \ldots, 2^k - 1\}$. Let $\ell = n - 2^k + 1$, noting that our assumptions guarantee that $k \leq \ell$ and $\ell \leq c \leq (\ell - k + 1)2^k - 1$. Lemma 3(a) then shows how to write $c = \sum_{j=1}^{\ell} 2^{k_j}$. We use the values $2^{k_j}$, in decreasing order, to define $r_i$ for $i \in \{2^k, 2^k + 1, \ldots, n\}$, noting that $r_n = 1$. Then

$$\sum_{i=1}^{n} r_i = (2^k - 1)c + c = 2^k c.$$

Since $2^k c$ is divisible by $c$ and by $2^{k'}$ for all $k' \in \{0, 1, \ldots, k\}$, we have thus produced an arithmetical structure on $K_n$. □

Although we imposed the condition $k + 2^k - 1 \leq n$ here to ensure that $k \leq \ell$ in the proof, this does not restrict possible values of $r_1$, as we show in the proof of Theorem 1. Together with Proposition 2, Proposition 4 with $k = 1$ allows us to construct arithmetical structures on $K_n$ with $r_1$ taking any value up to $2n - 3$. When $k = 2$, the bound is $4n - 17$; when $k = 3$, the bound is $8n - 73$; and when $k = 4$, the bound is $16n - 289$. If we restrict attention to prime $r_1$, these bounds can be improved slightly, as the following proposition shows.

**Proposition 5.** *Fix $n \geq 2$. For any positive integer $k$ satisfying $k + 2^k - 1 \leq n$ and any prime number $p$ satisfying $n - 2^k + 1 \leq p \leq (n - k - 2^k + 3)2^k - 3$, there is an arithmetical structure on $K_n$ with $r_1 = p$.*

*Proof.* If $p = 2$ (and $n \geq 3$), Proposition 2 gives an arithmetical structure on $K_n$ with $r_1 = p$. Therefore suppose $p$ is odd. Let $r_i = p$ for all $i \in \{1, 2, \ldots, 2^k - 1\}$. Let $\ell = n - 2^k + 1$, noting that our assumptions guarantee $k \leq \ell$ and $\ell \leq p \leq (\ell - k + 2)2^k - 3$. Lemma 3(b) then shows how to write $p = \sum_{j=1}^{\ell} 2^{k_j}$. As in the proof of Proposition 4, we use the values $2^{k_j}$, in decreasing order, to define $r_i$ for $i \in \{2^k, 2^k + 1, \ldots, n\}$, noting that $r_n = 1$. Then

$$\sum_{i=1}^{n} r_i = (2^k - 1)p + p = 2^k p,$$

which is divisible by $p$ and by $2^{k'}$ for all $k' \in \{0, 1, \ldots, k\}$. Therefore we have produced an arithmetical structure on $K_n$. □

For example, when $k = 1$, Proposition 5 allows us to construct arithmetical structures on $K_n$ with $r_1$ taking prime values as large as $2n - 3$. When $k = 2$, the bound is $4n - 15$; when $k = 3$, the bound is $8n - 67$; and when $k = 4$, the bound is $16n - 275$.

We are now prepared to complete the proof of Theorem 1.

*Proof of Theorem 1.* The necessary constructions are given in Propositions 2, 4, and 5. It remains only to show that, for each $n$, values of $k$ that maximize the upper bounds in Propositions 4 and 5 satisfy $k + 2^k - 1 \leq n$.

The upper bound $(n - k - 2^k + 2)2^k - 1$ in Proposition 4 is linear in $n$ with slope $2^k$. A straightforward calculation shows that the bound with slope $2^{k-1}$ coincides with the bound with slope $2^k$ when $n = k + 3 \cdot 2^{k-1} - 1$ and that the bound with slope $2^k$ coincides with the bound with slope $2^{k+1}$ when $n = k + 3 \cdot 2^k$. Therefore the bound with slope $2^k$ is maximal exactly when $n$ is between $k + 3 \cdot 2^{k-1} - 1$ and $k + 3 \cdot 2^k$. When the bound is maximized, we therefore have that $n \geq k + 3 \cdot 2^{k-1} - 1 \geq k + 2^k - 1$, meaning the condition of Proposition 4 is satisfied. This proves (a).

The argument for (b) is very similar. The upper bound $(n-k-2^k+3)2^k-3$ in Proposition 5 is maximal for $n$ between $k+3\cdot 2^{k-1}-2$ and $k+3\cdot 2^k-1$. When the bound is maximized, we then have that $n \geq k+3\cdot 2^{k-1}-2 \geq k+2^k-1$, meaning the condition of Proposition 5 is satisfied. □

We conclude this section by giving another construction that allows us to produce some arithmetical structures with values of $r_1$ other than those guaranteed by Theorem 1.

**Proposition 6.** *For any positive integer $k \leq n-1$, there is an arithmetical structure on $K_n$ with $r_1 = k(n-k)+1$.*

*Proof.* Let

$$r_i = \begin{cases} k(n-k)+1 & \text{for } i \in \{1,2,\ldots,k-1\}, \\ k & \text{for } i \in \{k,k+1,\ldots,n-1\}, \\ 1 & \text{for } i = n. \end{cases}$$

Then

$$\sum_{i=1}^{n} r_i = (k-1)(k(n-k)+1)+k(n-k)+1 = k(k(n-k)+1).$$

Since this is divisible by $k(n-k)+1$, $k$, and 1, we have thus produced an arithmetical structure on $K_n$. □

For example, when $n=13$, Theorem 1 guarantees that we can find an arithmetical structure on $K_n$ with $r_1 = p$ for all prime $p \leq 37$. By taking $k=5$ in Proposition 6, we can also produce an arithmetical structure with $r_1 = 41$. By taking $k=6$, we can produce an arithmetical structure with $r_1 = 43$. The results of this section cannot be extended too much further, as we show in the following section.

## 3. Obstruction

We next prove obstruction results that complement our constructions in the previous section. Our first result shows that $r_1$ cannot be a product of primes all of which are too large.

**Theorem 7.** *Suppose $c \geq 2$ is an integer with prime factorization $p_1^{a_1}p_2^{a_2}\cdots p_k^{a_k}$, where $p_1 < p_2 < \cdots < p_k$ and $a_i \geq 1$ for all $i$. If $p_1 > (n+1)^2/4$, then there is no arithmetical structure on $K_n$ with $r_1 = c$.*

*Proof.* Suppose we have an arithmetical structure on $K_n$ with $r_1 = c$. Knowing that $r_1 \mid \sum_{i=1}^n r_i$, we define $b = \sum_{i=1}^n r_i/r_1$. Then $\sum_{i=1}^n r_i = bc$, meaning that $r_i \mid bp_1^{a_1}p_2^{a_2}\cdots p_k^{a_k}$ for all $i$. Let $m$ be the largest value of $i$ for which $r_i = c$. For all $i \in \{m+1, m+2, \ldots, n\}$, we have $r_i < c$, which implies that $r_i \leq bp_1^{a_1-1}p_2^{a_2}\cdots p_k^{a_k}$.

This means $\sum_{i=m+1}^{n} r_i \le (n-m)bp_1^{a_1-1}p_2^{a_2}\cdots p_k^{a_k}$. We also have that $\sum_{i=m+1}^{n} r_i = (b-m)p_1^{a_1}p_2^{a_2}\cdots p_k^{a_k}$. Therefore

$$(b-m)p_1^{a_1}p_2^{a_2}\cdots p_k^{a_k} \le (n-m)bp_1^{a_1-1}p_2^{a_2}\cdots p_k^{a_k},$$

and hence $(b-m)p_1 \le (n-m)b$. When $b=m$, there is only one arithmetical structure on $K_n$, namely that with $r_i=1$ for all $i$, so the desired structure cannot arise in this case. Therefore we assume $b>m$, in which case we have

$$p_1 \le \frac{(n-m)b}{b-m}.$$

When $b=m+1$, this gives $p_1 \le (n-b+1)b$. It is a simple calculus exercise to show this bound is maximized when $b=(n+1)/2$. It follows that $p_1 \le (n+1)^2/4$.

When $b \ge m+2$, we have

$$p_1 \le \frac{(n-m)b}{b-m} = \frac{nb-mb+nm-nm}{b-m} = n + \frac{m(n-b)}{b-m} \le n + \frac{m(n-m-2)}{2}.$$

It is a simple calculus exercise to show this bound is maximized when $m=n/2-1$, so therefore

$$p_1 \le n + \frac{(n/2-1)(n/2-1)}{2} = \frac{n^2}{8} + \frac{n}{2} + \frac{1}{2} = \frac{(n+1)^2}{4} - \frac{n^2-1}{4} \le \frac{(n+1)^2}{4}.$$

The result follows. □

If we restrict attention to arithmetical structures where $r_1$ is a prime number, then Theorem 7 can be improved to Theorem 8. The general outline of the proof is similar, with some of the bounds improved.

**Theorem 8.** *If $p$ is a prime number with $p > n^2/4+1$, then there is no arithmetical structure on $K_n$ with $r_1 = p$.*

*Proof.* If $p=2$, the hypothesis of the theorem is only satisfied for $n=1$, and there is no arithmetical structure on $K_1$ with $r_1=2$. Suppose we have an arithmetical structure on $K_n$ with $r_1=p$, where $p \ge 3$. Knowing that $r_1 \mid \sum_{i=1}^{n} r_i$, we define $b = \sum_{i=1}^{n} r_i/r_1$. Then $\sum_{i=1}^{n} r_i = bp$, meaning that $r_i \mid bp$ for all $i$. Let $m$ be the largest value of $i$ for which $r_i = p$. We can only have $b=m$ if $r_i=1$ for all $i$, but then $r_1$ is not prime. We consider two cases: when $b=m+1$ and when $b \ge m+2$.

Case I: $b=m+1$. For all $i \in \{m+1, m+2, \ldots, n\}$, we have that $r_i \mid bp$ and $r_i < p$, so therefore $r_i \mid b$. Whenever $r_i < b$, this means $r_i \le b/2$. If $r_{n-1}, r_n < b$, we would then have that $\sum_{i=m+1}^{n} r_i \le (n-m-1)b$. If instead $r_i = b$ for all $i \in \{m+1, m+2, \ldots, n-1\}$, we would have that $r_n \mid r_i$ for all $i \in \{m+1, m+2, \ldots, n\}$. Since $\sum_{i=m+1}^{n} r_i = (b-m)p = p$, this would also mean $r_n \mid p$, and hence that $r_n \mid r_i$ for all $i$. Therefore we would need to have $r_n = 1$, meaning that $\sum_{i=m+1}^{n} r_i \le$

$(n-m-1)b+1$. Regardless of the value of $r_{n-1}$, we thus have

$$p = \sum_{i=m+1}^{n} r_i \leq (n-m-1)b+1 = (n-b)b+1 = nb-b^2+1.$$

It is a simple calculus exercise to show that this bound is maximized when $b = n/2$. Hence we have that $p \leq n^2/4+1$.

Case II: $b \geq m+2$. We have that $r_i \leq b$ for all $i \in \{m+1, m+2, \ldots, n\}$ and $\sum_{i=m+1}^{n} r_i = (b-m)p$, so therefore $(b-m)p \leq (n-m)b$. As in the proof of Theorem 7, this yields

$$p \leq \frac{(n-m)b}{b-m} \leq n + \frac{m(n-m-2)}{2},$$

and this bound is maximized when $m = n/2-1$. Therefore

$$p \leq n + \frac{(n/2-1)(n/2-1)}{2} = \frac{n^2}{8} + \frac{n}{2} + \frac{1}{2} = \frac{n^2}{4} + 1 - \frac{(n-2)^2}{8} < \frac{n^2}{4} + 1.$$

We have thus shown that in all cases we must have $p \leq n^2/4+1$. □

For even $n$, we can choose $k = n/2$ in Proposition 6 and get an arithmetical structure on $K_n$ with $r_1 = n^2/4+1$. For odd $n$, we can choose $k = (n-1)/2$ and get an arithmetical structure on $K_n$ with $r_1 = (n^2-1)/4+1$. As some of these values of $r_1$ are prime, the bound in Theorem 8 therefore cannot be improved.

There are arithmetical structures for which $r_1$ takes composite values larger than the bound given in Theorem 8. For instance, the example in the opening paragraph of this paper gives an arithmetical structure on $K_5$ with $r_1 = 105$.

## 4. Prime $r_1$

This section considers the possible prime values $r_1$ can take in an arithmetical structure on $K_n$. Theorem 1(b) guarantees that $r_1$ can take any prime value up to $2^k n - (k+2^k-3)2^k - 3$ for any $k$. Theorem 8 says that $r_1$ cannot take any prime value larger than $n^2/4+1$. These bounds are not too far from each other. The function $n^2/4+1$ has linear approximations of the form $2^k n - 2^{2k} + 1$. When $k$ is 1 or 2, this linear approximation coincides with the bound from Theorem 1(b). In general, it differs from this bound by $(k-3)2^k + 4$.

Proposition 6 shows that $r_1$ can take some of the prime values in the gap between the bound of Theorem 1(b) and the bound of Theorem 8. We can check by hand whether it can take other prime values; to illustrate how to do this, we explain why there is no arithmetical structure on $K_{18}$ with $r_1 = 79$. Suppose there were such a structure, and let $b = \sum_{i=1}^{18} r_i/r_1$, so that $\sum_{i=1}^{18} r_i = 79b$. Let $m$ be the largest value of $i$ for which $r_i = 79$. Then $\sum_{i=m+1}^{18} r_i = 79(b-m)$. For all $i \in \{m+1, m+2, \ldots, 18\}$, we have that $r_i \,|\, 79b$ and $r_i < 79$. Therefore $r_i \,|\, b$,

| $n$ | yes, Theorem 1(b) | no, Theorem 8 | yes, Proposition 6 | yes, other | no, other |
|---|---|---|---|---|---|
| 3 | $p \leq 3$ | $p > 3.25$ | | | |
| 4 | $p \leq 5$ | $p > 5$ | | | |
| 5 | $p \leq 7$ | $p > 7.25$ | | | |
| 6 | $p \leq 9$ | $p > 10$ | | | |
| 7 | $p \leq 13$ | $p > 13.25$ | | | |
| 8 | $p \leq 17$ | $p > 17$ | | | |
| 9 | $p \leq 21$ | $p > 21.25$ | | | |
| 10 | $p \leq 25$ | $p > 26$ | | | |
| 11 | $p \leq 29$ | $p > 31.25$ | 31 | | |
| 12 | $p \leq 33$ | $p > 37$ | 37 | | |
| 13 | $p \leq 37$ | $p > 43.25$ | 41, 43 | | |
| 14 | $p \leq 45$ | $p > 50$ | | 47 | |
| 15 | $p \leq 53$ | $p > 57.25$ | 57 | | |
| 16 | $p \leq 61$ | $p > 65$ | | | |
| 17 | $p \leq 69$ | $p > 73.25$ | 71, 73 | | |
| 18 | $p \leq 77$ | $p > 82$ | | | 79 |
| 19 | $p \leq 85$ | $p > 91.25$ | 89 | | |
| 20 | $p \leq 93$ | $p > 101$ | 97, 101 | | |
| 21 | $p \leq 101$ | $p > 111.25$ | 109 | 103, 107 | |
| 22 | $p \leq 109$ | $p > 122$ | 113 | | |
| 23 | $p \leq 117$ | $p > 133.25$ | 127, 131 | | |
| 24 | $p \leq 125$ | $p > 145$ | | 127, 131, 137, 139 | |
| 25 | $p \leq 133$ | $p > 157.25$ | 137, 151, 157 | 139, 149 | |
| 26 | $p \leq 141$ | $p > 170$ | | 149, 151, 157, 163 | 167 |
| 27 | $p \leq 149$ | $p > 183.25$ | 163, 181 | 151, 157, 167, 173 | 179 |

**Table 1.** Possible prime $r_1$ in arithmetical structures on $K_n$ for $n \leq 27$.

and hence $r_i \leq b$. This means that $\sum_{i=m+1}^{18} r_i \leq (18-m)b$, so we must have $79(b-m) \leq (18-m)b$. If $b \geq m+2$, we would have

$$61b - 79m + mb \geq 61(m+2) - 79m + m(m+2) = (m-8)^2 + 58 > 0,$$

which would imply that $79(b-m) > (18-m)b$. Therefore we cannot have $b \geq m+2$. Since $b = m$ is only possible if $r_1 = 1$, it therefore remains to consider whether we can have $b = m+1$. In this case, we would have $\sum_{i=m+1}^{18} r_i \leq (18-m)(m+1)$. This bound is less than 79 except when $m$ satisfies $6 \leq m \leq 11$. If $m = 6$, we would need to have 12 divisors of 7 that sum to 79, but this is not possible. If $m = 7$, we would need to have 11 divisors of 8 that sum to 79, but this is not possible. If $m = 8$, we would need to have 10 divisors of 9 that sum to 79, but this is not

possible. If $m = 9$, we would need to have 9 divisors of 10 that sum to 79, but this is not possible. If $m = 10$, we would need to have 8 divisors of 11 that sum to 79, but this is not possible. If $m = 11$, we would need to have 7 divisors of 12 that sum to 79, but this is not possible. Therefore there is no arithmetical structure on $K_{18}$ with $r_1 = 79$. A similar approach can be used to either find arithmetical structures with other prime values of $r_1$ or to show that they do not exist. We have done this for all $n \leq 27$; the results are shown in Table 1.

We conclude by noting that, on $K_{27}$, there is no arithmetical structure with $r_1 = 179$, whereas there is an arithmetical structure with $r_1 = 181$. This shows that there is not a cutoff function $f(n)$ such that, for each $n$, there is an arithmetical structure on $K_n$ with $r_1 = p$ for all primes $p \leq f(n)$ and no such structure for any $p > f(n)$. Therefore, while one could attempt to improve the bound of Theorem 1(b), the possible prime values of $r_1$ cannot be fully explained by a result of this form.

## Acknowledgments

We would like to thank Nathan Kaplan for a helpful conversation. Harris was supported by a Niagara University Undergraduate Student Summer Support Grant.

## References

[Arce-Nazario et al. 2013] R. Arce-Nazario, F. Castro, and R. Figueroa, "On the number of solutions of $\sum_{i=1}^{11} 1/x_i = 1$ in distinct odd natural numbers", *J. Number Theory* **133**:6 (2013), 2036–2046. MR Zbl

[Archer et al. 2020] K. Archer, A. C. Bishop, A. Diaz-Lopez, L. D. García Puente, D. Glass, and J. Louwsma, "Arithmetical structures on bidents", *Discrete Math.* **343**:7 (2020), 111850. MR

[Braun et al. 2018] B. Braun, H. Corrales, S. Corry, L. D. García Puente, D. Glass, N. Kaplan, J. L. Martin, G. Musiker, and C. E. Valencia, "Counting arithmetical structures on paths and cycles", *Discrete Math.* **341**:10 (2018), 2949–2963. MR Zbl

[Browning and Elsholtz 2011] T. D. Browning and C. Elsholtz, "The number of representations of rationals as a sum of unit fractions", *Illinois J. Math.* **55**:2 (2011), 685–696. MR Zbl

[Burshtein 2007] N. Burshtein, "The equation $\sum_{i=1}^{9} 1/x_i = 1$ in distinct odd integers has only the five known solutions", *J. Number Theory* **127**:1 (2007), 136–144. MR Zbl

[Burshtein 2008] N. Burshtein, "All the solutions of the equation $\sum_{i=1}^{11} 1/x_i = 1$ in distinct integers of the form $x_i \in 3^{\alpha}5^{\beta}7^{\gamma}$", *Discrete Math.* **308**:18 (2008), 4286–4292. MR Zbl

[Corrales and Valencia 2018] H. Corrales and C. E. Valencia, "Arithmetical structures on graphs", *Linear Algebra Appl.* **536** (2018), 120–151. MR Zbl

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

[Erdős and Graham 1980] P. Erdős and R. L. Graham, *Old and new problems and results in combinatorial number theory*, Monographies de L'Enseignement Mathématique **28**, Université de Genève, L'Enseignement Mathématique, Geneva, 1980. MR Zbl

[Glass and Kaplan 2019] D. Glass and N. Kaplan, "Chip-firing games and critical groups", preprint, 2019. arXiv

[Glass and Wagner 2019] D. Glass and J. Wagner, "Arithmetical structures on paths with a doubled edge", preprint, 2019. arXiv

[Klivans 2019] C. J. Klivans, *The mathematics of chip-firing*, CRC Press, Boca Raton, FL, 2019. MR Zbl

[Konyagin 2014] S. V. Konyagin, "Double exponential lower bound for the number of representations of unity by Egyptian fractions", *Mat. Zametki* **95**:2 (2014), 312–316. In Russian; translated in *Math. Notes* **95**:1-2 (2014), 277–281. MR Zbl

[Lorenzini 1989] D. J. Lorenzini, "Arithmetical graphs", *Math. Ann.* **285**:3 (1989), 481–501. MR Zbl

[Sándor 2003] C. Sándor, "On the number of solutions of the Diophantine equation $\sum_{i=1}^{n} 1/x_i = 1$", *Period. Math. Hungar.* **47**:1-2 (2003), 215–219. MR Zbl

[Sloane 1991] N. J. A. Sloane, "Egyptian fractions: number of solutions of $1 = 1/x_1 + \cdots + 1/x_n$ in positive integers", entry A002967 in *The On-Line Encyclopedia of Integer Sequences* (http://oeis.org), 1991.

Received: 2019-09-15 Revised: 2020-01-01 Accepted: 2020-01-06

zharris@mail.niagara.edu *Department of Mathematics, Niagara University, Niagara University, NY, United States*

jlouwsma@niagara.edu *Department of Mathematics, Niagara University, Niagara University, NY, United States*

INVOLVE 13:2 (2020)
dx.doi.org/10.2140/involve.2020.13.357

# Connectedness of digraphs from quadratic polynomials

Siji Chen and Sheng Chen

(Communicated by Vadim Ponomarenko)

Suppose that $f(x) = x(x-k)$, where $k$ is an odd positive integer. First, an infinite digraph $G_k = (V, E)$ is defined, where the vertex set is $V = \mathbb{Z}$ and the edge set is $E = \{(x, y) \mid x, y \in \mathbb{Z},\ f(x) = f(2y)\}$. Then the following results are proved: if $k = 1$, then the digraph $G_k$ is weakly connected; if $p$ is a safe prime, i.e., both $p$ and $q = (p-1)/2$ are primes, then the number $w_p$ of weakly connected components of the digraph $G_p$ is 2. Finally, a conjecture that there are infinitely many primes $p$ such that $w_p = 2$ is presented.

## 1. Introduction

Denote by $\mathrm{Int}(\mathbb{Z}) = \{f(x) \in \mathbb{Q}[x] \mid f(\mathbb{Z}) \subseteq \mathbb{Z}\}$ the ring of integer-valued polynomials on $\mathbb{Z}$. Chapman and Ponomarenko [2011] completely characterized the pairs of polynomials $(f, g)$ in $\mathrm{Int}(\mathbb{Z})$ such that $f(\mathbb{Z}) = g(\mathbb{Z})$. Motivated by this work, in this paper we study the digraph $G_k = (V, E)$ defined by

$$V = \mathbb{Z}, \quad E = \{(x, y) | x, y \in \mathbb{Z},\ f(x) = g(y)\},$$

where $f(x) = x(x-k)$ and $g(x) = f(2x)$. If $k$ is odd, then by Lemma 1 in [Chapman and Ponomarenko 2011], it is easy to see that the digraph $G_k$ is an infinite digraph in which each vertex has out-degree 1 and in-degree 2. In fact, on one hand, since $f(x) = f(k-x)$, for every vertex $x \in \mathbb{Z}$ we have $(x, y)$ is the unique edge from $x$ where

$$y = \begin{cases} x/2 & \text{if } x \text{ is even,} \\ (k-x)/2 & \text{if } x \text{ is odd,} \end{cases}$$

and on the other hand, for every vertex $y$, there are two edges $(2y, y)$ and $(k-2y, y)$ into $y$.

For general definitions about digraphs, we can refer to [Bang-Jensen and Gutin 2009]. Recall that a digraph is weakly connected if the underlying graph is connected.

*MSC2010:* primary 05C25; secondary 05C40, 05C20.
*Keywords:* connectivity, digraph, prime.

## 2. Results

**Theorem 1.** *If $k = 1$, then the digraph $G_k$ is weakly connected.*

*Proof.* To prove the result, it suffices to prove the following statement by induction on positive integer $n$: every vertex $x \in \mathbb{Z}$ such that $|x/2| \le n$ is weakly connected to the vertex 0.

If $n = 1$, then $|x/2| \le n$ implies $x = -2,\ x = -1,\ x = 0,\ x = 1$, or $x = 2$. Since $(-2, -1),\ (-1, 1),\ (2, 1),\ (1, 0)$, and $(0, 0)$ are edges in $G_1$, the statement is true.

Suppose the statement is true for some $n \ge 1$, that is, every vertex $x \in \mathbb{Z}$ such that $|x/2| \le n$ is weakly connected to the vertex 0. Then for a vertex $x \in \mathbb{Z}$ such that $n < |x/2| \le n + 1$, we have $x = 2n+2,\ 2n+1,\ -2n-2$, or $-2n-1$. Note that the following are edges in $G_1$:

$$(2n + 2, n + 1), \quad (2n + 1, -n), \quad (-2n - 2, -n - 1), \quad (-2n - 1, n + 1).$$

Since $|n/2| \le n$ and $|(n+1)/2| \le n$, we know that vertex $x$ such that $n < |x/2| \le n+1$ is connected to some vertex $y$ such that $|y/2| \le n$. Therefore it is weakly connected to the vertex 0. □

**Theorem 2.** *If $p$ is a safe prime, that is, both $p$ and $q = (p-1)/2$ are primes, then the digraph $G_p$ has precisely two weakly connected components.*

*Proof.* Suppose that $p = 2q + 1$ is a safe prime. We prove the result in four steps.

Step 1: Take

$$U_1 = \{x \in \mathbb{Z} \mid p \text{ divides } x\},$$
$$U_2 = \{x \in \mathbb{Z} \mid p \text{ does not divide } x\}.$$

We use $G_{p,U_1}$ and $G_{p,U_2}$ to denote the two subdigraphs induced by the vertex subsets $U_1$ and $U_2$ respectively. Since there is no edge $(x, y)$ such that $x \in U_1$ and $y \in U_2$, or $x \in U_2$ and $y \in U_1$, we know $G_p$ is a disjoint union of the subdigraphs $G_{p,U_1}$ and $G_{p,U_2}$.

Step 2: There is an isomorphism of digraphs from $G_{p,U_1}$ to $G_1$ defined by the mapping of vertex sets: $U_1 \to \mathbb{Z} : x \mapsto x/p$. By Theorem 1, we know that $G_{p,U_1}$ is weakly connected.

Step 3: Take $W = \{1, 2, \ldots, p-1\}$. In this step, we aim to prove the following statement by induction on positive integers $n$:

Every vertex $x \in U_2$ with $|x/p| < n$ can be weakly connected to some vertex $y \in W$.

First, if $n = 1$, then the condition $x \in U_2$ with $|x/p| < 1$ is equivalent to $x \in \mathbb{Z}$ with $-p < x < 0$ or $0 < x < p$. If $x \in U_2$ is odd, then $x$ is connected to $y = (p-x)/2 \in W$. Otherwise, $x \in U_2$ is even, say $x = 2^r s$, where $r$ is a nonnegative integer and $s$ is

an odd integer. Then $x$ is connected to $s$ and thus connected to some vertex in $W$, and thus the statement is true for $n = 1$.

Next, suppose the statement is true for some positive integer $n$, that is, every $x \in U_2$ with $|x/p| < n$ is weakly connected to some vertex $y \in W$. For even $x \in U_2$ with $n < |x/p| < n+1$, the terminal vertex $y = x/2$ of the unique edge from $x$ satisfies $|y/p| < (n+1)/2 \le n$. For odd $x \in U_2$ with $n < |x/p| < n+1$, both $(x, z)$ and $(z, y)$ are edges, where $z = (p - x)/2$, and one of the following holds: $y = (p - z)/2$ or $y = z/2$. So $y = (p \pm x)/4$, which satisfies $|y/p| < (1 + (n+1))/4 < n$. Therefore $x \in U_2$ is always weakly connected to some vertex in $W$. Thus the statement is true for $n + 1$.

Step 4: Denote by $G_{p,W}$ the subdigraph induced by the vertex set

$$W = \{1, 2, \ldots, p-1\} = \{1, 2, \ldots, 2q\}.$$

The edge set of $G_{p,W}$ is

$$\{(2y, y) \mid y = 1, 2, \ldots, q\} \cup \{(p - 2y, y) \mid y = 1, 2, \ldots, q\}.$$

For any $x \in W$, we have some $y \in \{1, 2, \ldots, q\}$ such that both $(x, y)$ and $(p-x, y)$ are edges in $G_{p,W}$. So $x \in W$ is weakly connected to $p - x \in W$. For any $y \in \{1, 2, \ldots, q\}$, we have $(2y, y)$ is an edge in $G_{p,W}$ and thus $y$ is weakly connected to $2y$. For any $y \in \{q+1, q+2, \ldots, q+q\}$, we know $(2p - 2y, p - y)$ is an edge in $G_{p,W}$ and thus $y,\ p - y,\ 2p - 2y,\ p - (2p - 2y) = 2y - p$ are weakly connected. That is to say, any $y \in W$ is weakly connected to $2y \in W$ or $2y - p \in W$.

Denote by $U(\mathbb{F}_p)$ the multiplicative group of the invertible elements of the finite field $\mathbb{F}_p = (\mathbb{Z}, +, \cdot)/p\mathbb{Z}$. Since $p$ is a safe prime, the unit group $U(\mathbb{F}_p)$ is a cyclic group of order $p - 1 = 2q$, which can be generated by the set $S = \{-1, 2\}$. Therefore the digraph $C(U(\mathbb{F}_p), S)$ with vertex set $U(\mathbb{F}_p)$ and edge set

$$\{(a, b) \mid a, b \in U(\mathbb{F}_p),\ a = -b \text{ or } a = 2b\}$$

is connected.

Using elementary number theory, it follows that $G_{p,W}$ is weakly connected. □

## 3. A conjecture

Safe primes appear in many places of mathematical theory and applications; see, e.g., [OEIS; von zur Gathen and Shparlinski 2013]. Although a folklore conjecture states that infinitely many Sophie Germain primes $q$ exist, i.e., $q$ and $p = 2q + 1$ are prime, a proof of this conjecture is not yet in sight; see [von zur Gathen and Shparlinski 2013]. Since we do not know whether the conjecture that there are infinitely many safe primes is true or not, we present the following slightly weaker *conjecture*:

**Conjecture.** There are infinitely many primes $p$ such that the digraph $G_p$ has exactly two weakly connected components.

**Remark.** For primes $p = 13$, $p = 17$, $p = 31$, which are not safe primes, the number of weakly connected components of $G_p$ are $2, 3, 4$ respectively. We may use the result in [Chen and Chen 2015] to study the number of weakly connected components of $G_k$ for general nonprime odd integers $k$. However, an explicit formula for $w_k$ seems out of reach.

## Acknowledgement

The authors would like to thank the editor and referee for valuable comments and suggestions which greatly improve the presentation the paper.

## References

[Bang-Jensen and Gutin 2009] J. Bang-Jensen and G. Gutin, *Digraphs: theory, algorithms and applications*, 2nd ed., Springer, 2009. MR Zbl

[Chapman and Ponomarenko 2011] S. T. Chapman and V. Ponomarenko, "On image sets of integer-valued polynomials", *J. Algebra* **348** (2011), 350–353. MR Zbl

[Chen and Chen 2015] S. Chen and X. Chen, "Weak connectedness of tensor product of digraphs", *Discrete Appl. Math.* **185** (2015), 52–58. MR Zbl

[von zur Gathen and Shparlinski 2013] J. von zur Gathen and I. E. Shparlinski, "Generating safe primes", *J. Math. Cryptol.* **7**:4 (2013), 333–365. MR Zbl

[OEIS] N. Sloane, "Safe primes $p$: $(p-1)/2$ is also prime", entry A005385 in *The online encyclopedia of integer sequences*.

Received: 2020-02-02 Revised: 2020-03-10 Accepted: 2020-03-14

1756474535@qq.com *RDFZ Xishan AP Center, RDFZ Xishan School, Beijing, China*

schen@hit.edu.cn *Department of Mathematics, Harbin Institute of Technology, Harbin, China*

## Guidelines for Authors

Submissions in all mathematical areas are encouraged. All manuscripts accepted for publication in *Involve* are considered publishable in quality journals in their respective fields, and include a minimum of one-third student authorship. Submissions should include substantial faculty input; faculty co-authorship is strongly encouraged. Authors may submit manuscripts in PDF format on-line at the Submission page at the Involve website.

**Originality**. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

**Language**. Articles in *Involve* are usually in English, but articles written in other languages are welcome.

**Required items**. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

**Format**. Authors are encouraged to use LaTeX but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

**References**. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

**Figures**. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@msp.org with details about how your graphics were generated.

**White space**. Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

**Proofs**. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

involve
2020 vol. 13 no. 2

Arithmetic functions of higher-order primes 181
Kyle Czarnecki and Andrew Giddings
Spherical half-designs of high order 193
Daniel Hughes and Shayne Waldron
A series of series topologies on $\mathbb{N}$ 205
Jason DeVito and Zachary Parker
Discrete Morse functions, vector fields, and homological sequences on trees 219
Ian Rand and Nicholas A. Scoville
An explicit third-order one-step method for autonomous scalar initial value problems of first order based on quadratic Taylor approximation 231
Thomas Krainer and Chenzhang Zhou
New generalized secret-sharing schemes with points on a hyperplane using a Wronskian matrix 257
Weston Loucks and Bahattin Yildiz
Generalized Cantor functions: random function iteration 281
Jordan Armstrong and Lisbeth Schaubroeck
Numerical semigroup tree of multiplicities 4 and 5 301
Abby Greco, Jesse Lansford and Michael Steward
Enumerating diagonalizable matrices over $\mathbb{Z}_{p^k}$ 323
Catherine Falvey, Heewon Hah, William Sheppard, Brian Sittinger and Rico Vicente
On arithmetical structures on complete graphs 345
Zachary Harris and Joel Louwsma
Connectedness of digraphs from quadratic polynomials 357
Siji Chen and Sheng Chen