involve

a journal of mathematics

involve

pjm.math.berkeley.edu/involve

Generating and zeta functions, structure, spectral and analytic properties of the moments of the Minkowski question mark function

Giedrius Alkauskas

(Communicated by Ken Ono)

In this paper we are interested in moments of the Minkowski question mark function $?(x)$. It appears that, to some extent, the results are analogous to results obtained for objects associated with Maass wave forms: period functions, L-series, distributions. These objects can be naturally defined for $?(x)$ as well. Various previous investigations of $?(x)$ are mainly motivated from the perspective of metric number theory, Hausdorff dimension, singularity and generalizations. In this work it is shown that analytic and spectral properties of various integral transforms of $?(x)$ do reveal significant information about the question mark function. We prove asymptotic and structural results about the moments, calculate certain integrals which involve $?(x)$, define an associated zeta function, generating functions, Fourier series, and establish intrinsic relations among these objects.

1. Introduction

The aim of this paper is to continue investigations on the moments of the Minkowski question mark function, begun in [Alkauskas ≥ 2009]. The function $F(x)$, the *question mark function*, was introduced by Minkowski in 1904 as an example of a monotone and continuous function $F : [0, \infty) \cup \{\infty\} \to [0, 1]$, which maps rationals to dyadic rationals, and quadratic irrationals to nondyadic rationals. For nonnegative real x it is defined by the expression

$$F([a_0, a_1, a_2, a_3, \ldots]) = 1 - 2^{-a_0} + 2^{-(a_0+a_1)} - 2^{-(a_0+a_1+a_2)} + \cdots, \quad (1)$$

where $x = [a_0, a_1, a_2, a_3, \ldots]$ stands for the representation of x by a (regular) continued fraction [Khinchin 1964]. Figure 1 shows the image of $F(x)$ for $x \in [0, 2]$. More often this function is investigated in the interval $[0, 1]$; in this case we use a standard notation $?(x) = 2F(x)$ for $x \in [0, 1]$. For rational x, the series terminates
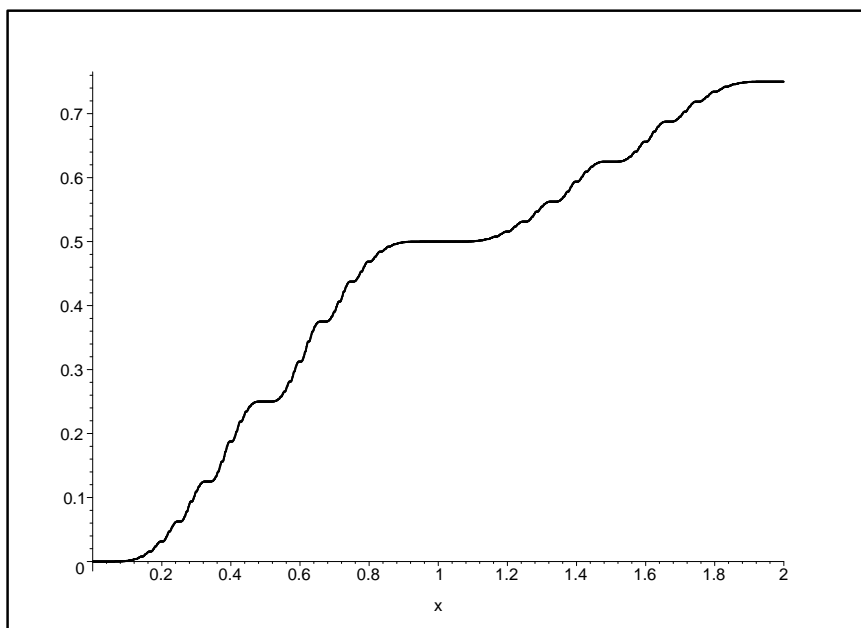
Figure 1. The Minkowski question mark function $F(x)$, $x \in [0, 2]$.

at the last nonzero partial quotient a_n of the continued fraction. This function was investigated by many authors. In particular, Denjoy [1938] showed that $?(x)$ is singular, and that the derivative vanishes almost everywhere. In fact, singularity of $?(x)$ follows from Khinchin's average value theorem on continued fractions [Khinchin 1964, chapter III]. The nature of singularity of $?(x)$ was clarified by Paradís et al. [2001]. In particular, the existence of the derivative $?'(x)$ in $\mathbb{R}$ for fixed x forces it to vanish. Salem [1943] proved (see also [Kinney 1960]) that $?(x)$ satisfies the Lipschitz condition of order $(\log 2)/(2 \log \gamma)$, where $\gamma = (1 + \sqrt{5})/2$, and this is in fact the best possible exponent for the Lipschitz condition. The Fourier–Stieltjes coefficients of $?(x)$, defined as $\int_0^1 e^{2\pi i n x} \, d?(x)$, were also investigated in [Salem 1943]. It is worth noting that in Section 8 we will encounter analogous coefficients (see Proposition 3). Meanwhile, [Grabner et al. 2002], out of all papers in the bibliography list, is the closest in spirit to the current article. In order to derive precise error bounds for the so-called Garcia entropy of a certain measure, the authors consider the moments of the monotone, continuous singular function

$$F_2([a_1, a_2, \ldots]) = \sum_{n=1}^{\infty} (-1)^{n-1} 3^{-(a_1 + \cdots + a_n - 1)} (q_n + q_{n-1}),$$

where $q_\star$ stand for a corresponding denominator of the convergent to $[a_1, a_2, \ldots]$. The moments of $F(x)$ itself were never considered before. Lamberger [2006] has shown that $F(x)$ and $F_2(x)$ are the first two members of a family (indexed by natural numbers) of mutually singular measures, derived from the subtractive Euclidean algorithm. From a number-theoretic point of view this generalization is extremely interesting and natural, and it deserves much wider attention.

We confine ourselves to a cursory overview of the properties of $?(x)$, and refer the reader to [Alkauskas $\geq$ 2009] for a short survey on available literature. These works include [Beaver and Garrity 2004; Bonanno et al. 2008; Calkin and Wilf 2000; Denjoy 1938; 1956a; 1956b; 1956c; Dushistova and Moshchevitin $\geq$ 2009; Esposti et al. $\geq$ 2009; Finch 2003; Girgensohn 1996; Grabner et al. 2002; Isola 2002; Kesseböhmer and Stratmann 2007; 2008; Kinney 1960; Lagarias 1991; Lagarias and Tresser 1995; Lamberger 2006; Moshchevitin and Vielhaber $\geq$ 2009; Okamoto and Wunsch 2007; Panti 2008; Paradís et al. 2001;1998; Ramharter 1987; Reese 1989; Reznick $\geq$ 2009; Ryde 1922; 1983; Salem 1943; Tichy and Uitz 1995; Vepštas 2004; Wirsing 2006.]

Recently, Calkin and Wilf [2000] (re)defined a binary tree which is generated by the iteration

$$\frac{a}{b} \quad \mapsto \quad \frac{a}{a+b}, \quad \frac{a+b}{b},$$

starting from the root $1/1$. Elementary considerations show that this tree contains every positive rational number once and only once, each being represented in lowest terms. The first four iterations lead to
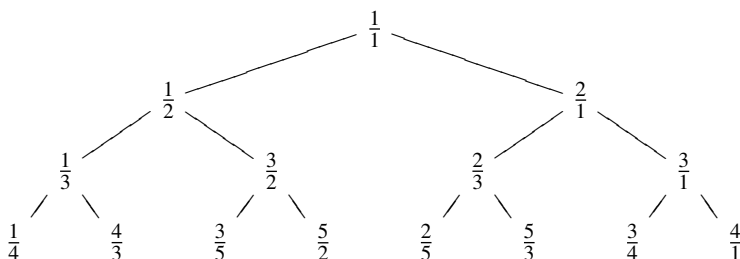
This tree is in fact a permutation (inside each generation) of the Stern–Brocot tree. Its limitation to $[0, 1]$ is a permutation of the Farey tree. Thus, the n-th generation consists of 2^{n-1} positive rationals. It is surprising that the iteration discovered by Newman [2003],

$$x_1 = 1, \quad x_{n+1} = 1/(2[x_n] + 1 - x_n),$$

produces exactly rationals of this tree, reading them line-by-line, and thus gives an example of a simple recurrence which produces all positive rationals (here, as usual, $[\star]$ stands for the integer part function). Recently, Dilcher and Stolarsky [2007] produced a natural analogue of this tree, replacing integers r with polynomials

$r \in (\mathbb{Z}/2\mathbb{Z})[x]$. One of the results is that these polynomials also satisfy analogous recurrence (following the proper definition of an integral part of a rational function, which comes from the Euclidean algorithm). It is important to note that the n-th generation of the Calkin–Wilf binary tree consists of exactly those rational numbers, whose elements of the continued fraction sum up to n. This fact can be easily inherited directly from the definition. First, if a rational number a/b is represented as a continued fraction $[a_0, a_1, \ldots, a_r]$, then the map $a/b \to (a+b)/b$ maps a/b to $[a_0 + 1, a_1 \ldots, a_r]$. Second, the map $a/b \to a/(a + b)$ maps a/b to $[0, a_1 + 1, \ldots, a_r]$ if $a/b < 1$, and to $[1, a_0, a_1, \ldots, a_r]$ if $a/b > 1$. This is an important fact which makes the investigations of rational numbers according to their position in the Calkin–Wilf tree highly motivated from the perspective of metric number theory and dynamics of continued fractions. The sequence of numerators

$$0, 1, 1, 2, 1, 3, 2, 3, 1, 4, 3, 5, 2, 5, 3, 4, 1, \ldots$$

is called the Stern diatomic sequence and was introduced in [Stern 1858]. It satisfies the recurrence relations

$$s(0) = 0, \quad s(1) = 1, \quad s(2n) = s(n), \quad s(2n + 1) = s(n) + s(n + 1).$$

This sequence and the pairs $(s(n), s(n+1))$ have also been investigated by Reznick [$\geq$ 2009]. It is not surprising (bearing in mind the relation to the Farey tree) that the *distribution* of numerators, which are defined via the moments

$$Q_N^{(\tau)} = \sum_{n=2^N+1}^{2^{N+1}} s^{2\tau}(n), \quad \text{for } \tau > 0,$$

has an interesting application in thermodynamics and spin physics [Contucci and Knauf 1997; Cvitanović et al. 1998].

In [Alkauskas $\geq$ 2009] it was shown that each generation of the Calkin–Wilf tree possesses a distribution function $F_n(x)$, and that $F_n(x)$ converges uniformly to $F(x)$. This is, of course, a well known fact about the Farey tree. The function $F(x)$ as a distribution function is uniquely determined by the functional equation [Alkauskas $\geq$ 2009]

$$2F(x) = \begin{cases} F(x - 1) + 1 & \text{if } x \geq 1, \\ F\left(\frac{x}{1-x}\right) & \text{if } 0 \leq x < 1. \end{cases} \tag{2}$$

This implies $F(x) + F(1/x) = 1$. The mean value of $F(x)$ has been investigated by several authors, and was proved to be 3/2 [Alkauskas $\geq$ 2009; Reznick $\geq$ 2009; Steuding 2006; Wirsing 2006].

On the other hand, almost all the results mentioned reveal the properties of the Minkowski question mark function as a function itself. Nevertheless, the final goal and motivation of [Alkauskas ≥ 2009] and this work is to show that in fact there exist several unique and very interesting analytic objects associated with $F(x)$ which encode a great deal of essential information about it. These objects will be introduced in Section 2.

Lastly, and most importantly, let us point out that, surprisingly, there are striking similarities between the results proved here and in [Alkauskas ≥ 2009] with the results on period functions for Maass wave forms in [Lewis and Zagier 2001] . That work is an expanded and clarified exposition of an earlier paper by Lewis [1997]. The concise exposition of these objects, their properties and relations to the Selberg zeta function can be found in [Zagier 2001]. The reader who is not indifferent to the beauty of the Minkowski question mark function is strongly urged to compare results in this work with those in [Lewis and Zagier 2001]. Thus, instead of making quite numerous references to [Lewis and Zagier 2001] at various stages of the work (mainly in Sections 2, 3, 8 and 9), it is more useful to give a table of most important functions encountered there, juxtaposed with analogous objects in this work. Here is the summary (the notations on the right will be explained in Sections 2 and 9).

Maass wave form	$u(z)$	$\Psi(x)$	Periodic function on the real line
Period function	$\psi(z)$	$G(z)$	Dyadic period function
Distribution	$U(x)\,dx$	$dF(x)$	Minkowski's "question mark"
L-functions	$L_0(\rho), L_1(\rho)$	$\zeta_{\mathcal{M}}(s)$	Dyadic zeta function
Entire function	$g(w)$	$\mathfrak{m}(t)$	Generating function of moments
Entire function	$\phi(w)$	$M(t)$	Generating function of moments
Spectral parameter	s	$1/2; 1$	Analogue of a spectral parameter

As a matter of fact, the first entry is the only one where the analogy is not precise. Indeed, the distribution $U(x)$ is the limit value of the Maass wave form $u(x + iy)$ on the real line (as $y \rightarrow +0$, in the sense that $u(x + iy) \sim y^{1-s}U(x) + y^s U(x)$, whereas $\Psi(x)$ is the same $F(x)$ made periodic. As far as the last entry of the table in concerned, the *analogue* of a spectral parameter, sometimes this role is played by 1, sometimes by 1/2. This occurs, obviously, because the relation between the Maass forms and $F(x)$ is just an analogy which is not strictly defined.

This work is organized as follows. In Section 2 we give a summary of the previous results obtained in [Alkauskas ≥ 2009]. In Section 3 we give a short proof of the three-term functional (13), and prove the existence of certain distributions, which can be thought of as close relatives of $F(x)$. In Section 4 we demonstrate that there are linear relations among moments M_L, and they are presented in an explicit manner. Moreover, we formulate a conjecture, based on the analogy with periods, that these are the only possible relations. In Section 5, the estimate for the

moments m_L is proved. As a consequence, $\lim_{L\to\infty}(\log m_L)/(\sqrt{L}) = -2\sqrt{\log 2}$. In Section 6 we prove the exactness of a certain sequence of functional vector spaces and linear maps related to $F(x)$ in an essential way. Section 7 is devoted to the calculation of a number of integrals, giving a rare example of a Stieltjes integral, involving the question mark function, that *can* be calculated. In Section 8 we compute the Fourier expansion of $F(x)$. It is shown that this establishes yet another relation among $\mathfrak{m}(t)$, $G(z)$ and $F(x)$ via Taylor coefficients and special values. In Section 9, the associated Dirichlet series $\zeta_{\mathcal{M}}(s)$ is introduced. In Section 10, some concluding remarks are presented, regarding future research; relations between $F(x)$ and the Calkin–Wilf tree (and the Farey tree as well) to the known objects are established. Note also that we use the word *distribution* to describe a monotone function on $[0, \infty)$ with variation 1, and also for a continuous linear functional on some space of analytic functions. In each case the meaning should be clear from the context.

2. Summary of previous results

This section provides a summary of previous results. For $L \in \mathbb{N}_0$, let

$$
\begin{aligned}
M_L &= \int_0^\infty x^L \, dF(x), \\
m_L &= \int_0^\infty \left(\frac{x}{x+1}\right)^L dF(x) = 2 \int_0^1 x^L \, dF(x) = \int_0^1 x^L \, d?(x).
\end{aligned}
\tag{3}
$$

Both sequences are of definite number-theoretic significance because

$$
\begin{aligned}
M_L &= \lim_{n\to\infty} 2^{1-n} \sum_{a_0+a_1+\cdots+a_s=n} [a_0, a_1, .., a_s]^L, \\
m_L &= \lim_{n\to\infty} 2^{2-n} \sum_{a_1+\cdots+a_s=n} [0, a_1, .., a_s]^L,
\end{aligned}
\tag{4}
$$

(the summation takes place over rational numbers presented as continued fractions; thus, $a_0 \geq 0$, $a_i \geq 1$ for $i \geq 1$, and $a_s \geq 2$. In fact, clarification of their nature was the initial main motivation for our work. We define the exponential generating functions

$$
M(t) = \sum_{L=0}^\infty \frac{M_L}{L!} t^L, \qquad \mathfrak{m}(t) = \sum_{L=0}^\infty \frac{m_L}{L!} t^L.
$$

Thus,

$$
M(t) = \int_0^\infty e^{xt} \, dF(x), \quad \mathfrak{m}(t) = \int_0^\infty \exp\left(\frac{xt}{x+1}\right) dF(x) = 2 \int_0^1 e^{xt} \, dF(x).
$$

One easily verifies that $\mathfrak{m}(t)$ is an entire function and that the Taylor series at the origin for $M(t)$ has a radius of convergence $\log 2$. There are natural relations among values M_L and m_L, independent of a specific distribution, like $F(x)$. They encode the relations among functions x^L, $L \in \mathbb{N}_0$, and functions $\left(x/(x+1)\right)^L$, $L \in \mathbb{N}_0$, given by

$$x^L = \sum_{s \geq L} \binom{s-1}{L-1} \left(\frac{x}{x+1}\right)^s.$$

Therefore,

$$M_L = \sum_{s \geq L} \binom{s-1}{L-1} m_s. \tag{5}$$

On the other hand, the intrinsic information about $F(x)$ is encoded in the relations

$$m_L = M_L - \sum_{s=0}^{L-1} M_s \binom{L}{s}, \quad L \geq 0. \tag{6}$$

Further, we have

$$M(t) = \frac{1}{2-e^t} \mathfrak{m}(t), \quad \mathfrak{m}(t) = e^t \mathfrak{m}(-t). \tag{7}$$

The first relation is equivalent to the system (6), and it encodes all the information about $F(x)$ (provided we take into account the natural relations just mentioned). The second one represents only the symmetry property, given by

$$F(x) + F(1/x) = 1.$$

One of the main results about $\mathfrak{m}(t)$ is that it is uniquely determined by the regularity condition $\mathfrak{m}(-t) \ll e^{-\sqrt{t \log 2}}$, as $t \to \infty$, the boundary condition $\mathfrak{m}(0) = 1$, and the integral equation

$$\mathfrak{m}(-s) = (2e^s - 1) \int_0^\infty \mathfrak{m}'(-t) J_0(2\sqrt{st}) \, dt, \quad s \in \mathbb{R}_+. \tag{8}$$

(Here $J_0(*)$ stands for the Bessel function $J_0(z) = 1/\pi \int_0^\pi \cos(z \sin x) \, dx$). This equation can be rewritten as a second type Fredholm integral equation [Kolmogorov and Fomin 1989, chapter 9]. In fact, if we denote

$$\psi(s) = \sqrt{2e^s - 1}, \quad \frac{J_1(2\sqrt{st})}{\psi(s)\psi(t)} = K(s,t), \quad \frac{\mathfrak{m}(-s) - 1}{\sqrt{s}\,\psi(s)} = \mathsf{Y}(s),$$

then one has

$$\mathsf{Y}(s) = \ell(s) - \int_0^\infty \mathsf{Y}(t) K(s,t) \, dt, \tag{9}$$

where

$$\ell(s) = -\frac{1}{\psi(s)} \int_0^\infty \frac{J_1(2\sqrt{st})}{\sqrt{t}(2e^t - 1)} \, dt = \frac{1}{\sqrt{s}\,\psi(s)} \Big(\sum_{n=1}^\infty e^{-s/n} 2^{-n} - 1 \Big).$$

Even more importantly, all the results about the exponential generating function can be restated in terms of a generating function of moments. Let

$$G(z) = \sum_{L=1}^\infty m_L z^{L-1} \quad \text{for } |z| \leq 1 \tag{10}$$

(the series converge absolutely on the boundary of a unit disc as well, as is clear from Equation (5), or Theorem 3.) Then the integral

$$G(z) = \int_0^\infty \frac{\frac{x}{x+1}}{1 - \frac{x}{x+1}z} \, dF(x) = 2 \int_0^1 \frac{x}{1 - xz} \, dF(x) \tag{11}$$

extends $G(z)$ to the cut plane $\mathbb{C} \setminus (1, \infty)$. The generating function of moments M_L does not exists due to the factorial growth of M_L, but the generating function can still be defined in the cut plane $\mathbb{C}' = \mathbb{C} \setminus (0, \infty)$ by $\int_0^\infty (x/(1 - xz)) \, dF(x)$. In fact, this integral equals $G(z + 1)$, which is the consequence of an algebraic identity

$$\frac{x}{1 - xz} = \frac{\frac{x}{x+1}}{1 - \frac{x}{x+1}(z+1)}.$$

The following result was proved in [Alkauskas $\geq$ 2009].

Theorem 1. *The function $G(z)$, defined initially as a power series, has an analytic continuation to the cut plane $\mathbb{C} \setminus (1, \infty)$ via Equation* (11). *It satisfies the functional equation*

$$-\frac{1}{1-z} - \frac{1}{(1-z)^2} G\Big(\frac{1}{1-z}\Big) + 2G(z+1) = G(z), \tag{12}$$

and also the symmetry property

$$G(z+1) = -\frac{1}{z^2} G\Big(\frac{1}{z} + 1\Big) - \frac{1}{z}.$$

Moreover, $G(z) \to 0$, if $z \to \infty$ and the distance from z to a half line $[0, \infty)$ tends to infinity.

Conversely, the function having these properties is unique.

Note that two functional equations for $G(z)$ can be merged into a single one. It is easy to check that the equation

$$\frac{1}{z} + \frac{1}{z^2} G\Big(\frac{1}{z}\Big) + 2G(z+1) = G(z) \tag{13}$$

is equivalent to both of them together. In fact, the change $z \mapsto 1/z$ in the last equation gives the symmetry property, and application of it to the term $G(1/z)$ in Equation (13) gives the functional equation in Theorem 1. Nevertheless, it is sometimes convenient to separate Equation (13) into two equations. The reason for this is that in (12) all arguments belong simultaneously to $\mathbb{H}$ (the upper half plane $\Re z > 0$), $\mathbb{R}$, or $\mathbb{H}^-$ (the lower half plane), whereas in (13) they are mixed. This will become crucial later (see the Section 10).

The transition $\mathfrak{m}(t) \to G(z)$ is given by the Laplace transform:

$$1 + zG(z) = \int_0^\infty \mathfrak{m}(zt)e^{-t}\,dt.$$

The same transform applied to the eigenfunctions of the Fredholm operator (9) yields the following result [Alkauskas $\geq$ 2009].

Theorem 2. *For every eigenvalue λ of the integral operator associated with the kernel $K(s,t)$, there exists at least one holomorphic function $G_\lambda(z)$ (defined for $z \in \mathbb{C} \setminus (1,\infty)$), such that*

$$2G_\lambda(z+1) = G_\lambda(z) + \frac{1}{\lambda z^2}\,G_\lambda\Big(\frac{1}{z}\Big). \tag{14}$$

Moreover, $G_\lambda(z)$ for $\Re z < 0$ satisfies all regularity conditions, imposed by it being an image under the Laplace transform [Lavrentjev and Shabat 1987, page 468].

Conversely, for every λ such that there exists a function which satisfies (14) and these conditions, λ is the eigenvalue of this operator. The set of all possible λ is countable, and $\lambda_n \to 0$, as $n \to \infty$.

Figure 2 shows the functions $G_\lambda(z)$ (for the first six eigenvalues) for real z in the interval $[-1, -0.2]$. The choice of this interval is motivated by Theorem 2. Note also that the functional equation implies $G_\lambda(0) = \big(1/2 + 1/(2\lambda)\big)G_\lambda(-1)$. Thus, one has $G_\lambda(0)/G_\lambda(-1) \to \infty$, as $\lambda \to 0$. This can also be seen empirically from Figure 2.

Summarizing, there are three objects associated with the Minkowski question mark function.

- The distribution $F(x)$ = functional equations (2) + continuity.
- The dyadic period function $G(z)$ = three-term functional Equation (13) + mild growth condition (as in Theorem 1).
- The exponential generating function $\mathfrak{m}(t)$ = the integral Equation (8) + the boundary value and diminishing condition on the negative real line.

Each of these objects is characterized by the functional equation, and subject to some regularity conditions, is unique, and thus arises exactly from $F(x)$. The objects are described via the "equality" *Function = Equation + Condition*. This
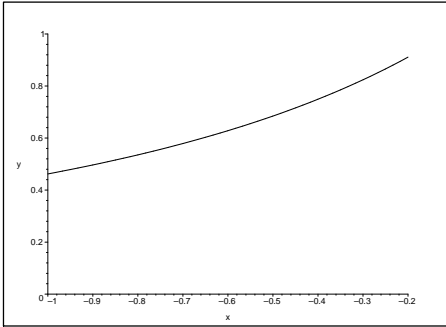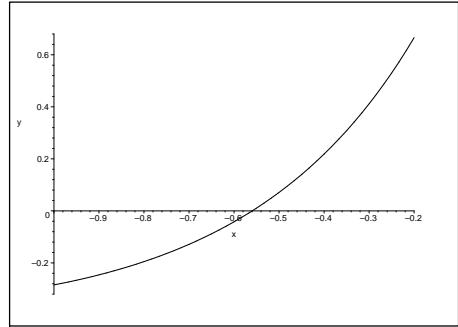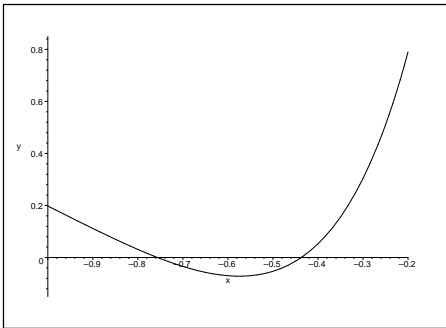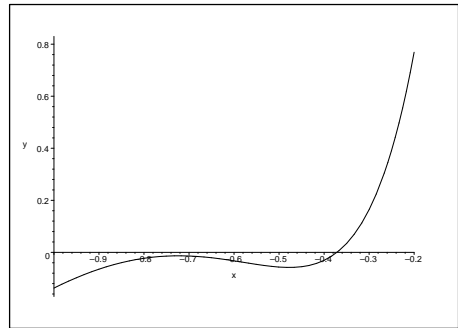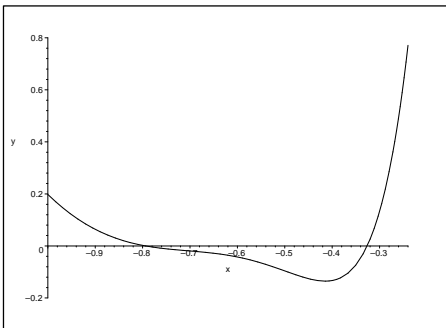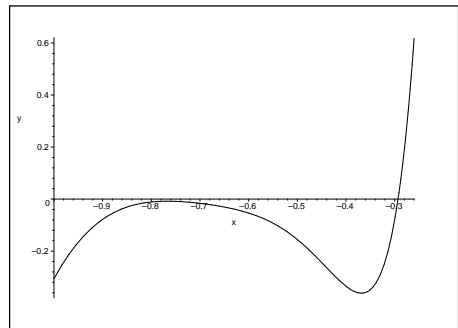
Figure 2. Eigenfunctions $G_\lambda(z)$ for $z \in [-1, -0.2]$.

means that the object on the left possesses both features; conversely, any object with these properties is necessarily the object on the left.

As expected, here we encounter the phenomenon of *bootstrapping*: in all cases, regularity conditions can be significantly relaxed, and they are sufficient for the uniqueness, which automatically implies stronger regularity conditions. Here we

show the rough picture of this phenomenon. In each case, we suppose that the object satisfies the corresponding functional equation. For the details, see [Alkauskas $\geq$ 2009].

(i) $F(x)$ is continuous at one point $\Rightarrow F(x)$ is continuous.

(ii) For every z with $\Re z < 0$, $G(z - x) = O(2^{x/2})$ as

$$x \to \infty \Rightarrow \quad G(z) = O(|z|^{-1}) \text{ as } \mathtt{dist}(z, \mathbb{R}_+) \to \infty.$$

(iii) $\mathfrak{m}'(-t) = O(t^{-1})$ as $t \to \infty \Rightarrow |\mathfrak{m}(-t)| \ll e^{-\sqrt{t \log 2}}$ as $t \to \infty$.

Corresponding converse results were proved in [Alkauskas $\geq$ 2009]. As far as $F(x)$ is concerned, this was in fact the starting point of these investigations, since the distribution of rationals in the Calkin–Wilf tree is a certain continuous function satisfying Equation (2); thus, it is exactly $F(x)$. The converse result for $\mathfrak{m}(t)$ follows from Fredholm alternative, since all eigenvalues of the operator (9) are strictly less than 1 in an absolute value. Finally, the converse theorem for $G(z)$ follows from a technical detail in the proof, which is the numerical estimate $0 < (\pi^2/12) - (\log^2 2/2) < 1$; as a matter of fact, it appears that this is essentially the same argument as in the case of $\mathfrak{m}(t)$, since this constant gives the upper bound for the moduli of eigenvalues.

One of the aims of this paper is to clarify the connections among these three objects, and to add the final fourth satellite, associated with $F(x)$. Henceforth, we have the complete list:

- The dyadic zeta function $\zeta_{\mathcal{M}}(s)$ (see Definition 1 below) = the functional equation with symmetry $s \to -s$ (27) + the regularity behavior in vertical strips.

In this case, we do not present a proof of a converse result. Indeed, the converse result for $G(z)$ is strongly motivated by its relation to the Eisenstein series $G_1(z)$ (see [Alkauskas $\geq$ 2009] and Section 10). In the case of $\zeta_{\mathcal{M}}(s)$, this question is of small importance, and we rather concentrate on the direct result and its consequences.

3. Three term functional equation, distributions $F_\lambda(x)$

In this section, we give a proof of (13) different from the one presented in [Alkauskas $\geq$ 2009], since it is considerably shorter. For our purposes, it is convenient to work in slightly greater generality. Suppose that $\lambda \in \mathbb{R}$ has the property that there exists a function $F_\lambda(x)$, $x \in [0, \infty)$, such that

$$\mathrm{d}F_\lambda(x + 1) = \frac{1}{2}\,\mathrm{d}F_\lambda(x), \ \mathrm{d}F_\lambda\left(\frac{1}{x}\right) = \frac{1}{\lambda}\,\mathrm{d}F_\lambda(x). \tag{15}$$

We omitted the word *continuous* in the description of the function intentionally. For a moment, consider $F_\lambda(x) = F(x)$ with $\lambda = -1$. Then $F_{-1}(x)$ is certainly continuous. The reason for introducing λ will be apparent later. Let

$$G_\lambda(z) = \int_0^\infty \frac{1}{x+1-z} \, dF_\lambda(x).$$

Since $F(x) + F(1/x) = 1$, we see that for $\lambda = -1$ the above definition of $G_\lambda(z)$ agrees with that of (11). This integral converges to an analytic function in the cut plane $\mathbb{C} \setminus (1, \infty)$. We have

$$
\begin{aligned}
2G_\lambda(z+1) &= 2\int_0^1 \frac{1}{x-z} \, dF_\lambda(x) + 2\int_1^\infty \frac{1}{x-z} \, dF(x) \\
&= 2\int_0^\infty \frac{1}{\frac{x}{x+1}-z} \, dF_\lambda\left(\frac{x}{x+1}\right) + 2\int_0^\infty \frac{1}{x+1-z} \, dF_\lambda(x+1) \\
&= \frac{2}{z} \int_0^\infty \left(\frac{x+1}{x+1-\frac{1}{z}} - 1 + 1\right) dF_\lambda\left(\frac{1}{x+1}\right) + G_\lambda(z) \\
&= \frac{\alpha}{\lambda z} + \frac{1}{\lambda z^2} G_\lambda\left(\frac{1}{z}\right) + G_\lambda(z), \text{ where } \alpha = \int_0^\infty dF_\lambda(x).
\end{aligned}
$$

For $\lambda = -1$ and $F_{-1}(x) = F(x)$, this gives Theorem 1. Further, suppose $\lambda \neq -1$. Then

$$\alpha = \int_0^\infty dF_\lambda(x) = \int_1^\infty dF_\lambda(x) + \int_0^1 dF_\lambda(x) = \frac{\alpha}{2} - \frac{\alpha}{2\lambda} \Rightarrow \alpha = 0.$$

Therefore, the last functional equation reads as

$$2G_\lambda(z+1) = \frac{1}{\lambda z^2} G_\lambda\left(\frac{1}{z}\right) + G_\lambda(z).$$

As a matter of fact, there cannot be any reasonable function $F_\lambda(x)$ which satisfies (15). Nevertheless, the last functional equation is identical to (14). Thus, Theorem 2 gives a description of all such possible λ. This suggests that we can still find certain distributions $F_\lambda(x)$. Further, as it was mentioned, -1 is not an eigenvalue of the operator (9). Due to the minus sign in front of the operator, this is exactly the exceptional eigenvalue, which is essential in the Fredholm alternative. The above proof (rigorous at least in case $\lambda = -1$), surprisingly, proves that the next tautological sentence has a certain point: "-1 is not an eigenvalue because it is -1". Indeed, we obtain a nonhomogeneous part of the three-term functional equation only because $\lambda = -1$, since otherwise $\alpha = 0$ and the equation is homogenic.

Distributions $F_\lambda(x)$ can indeed be strictly defined, at least in the space of functions, which are analytic in the disk $\mathbf{D} = \{z : |z - (1/2)| \leq (1/2)\}$, including its boundary. This space is equipped with a topology of uniform convergence, and a

distribution on this space is any continuous linear functional. Denote this space by C^ω. Now, since

$$\int_0^1 \frac{x}{1-xz}\,\mathrm{d}F_\lambda(x) = -\frac{\lambda}{2}\,G_\lambda(z) := \sum_{L=1}^\infty m_L^{(\lambda)} z^{L-1},$$

define a distribution F_λ on the space C^ω by $\langle z^L, F_\lambda \rangle = m_L^{(\lambda)}$, $L \geq 1$, $\langle 1, F_\lambda \rangle = 0$, and for any analytic function $B(z) \in C^\omega$, $B(z) = \sum_{L=0}^\infty b_L z^L$, by

$$\langle B, F_\lambda \rangle = \sum_{L=0}^\infty b_L \langle z^L, F_\lambda \rangle.$$

First, $\langle *, F_\lambda \rangle$ is certainly a linear functional and is properly defined, since the functional Equation (14) implies that $G_\lambda(z)$ possesses all left derivatives at $z = 1$; as a consequence, the series $\sum_{L=1}^\infty L^p |m_L^{(\lambda)}|$ converges for any $p \in \mathbb{N}$ (see Theorem 3 for the estimates on moments m_L). Second, let

$$B_n(z) = \sum_{L=0}^\infty b_L^{(n)} z^L, \qquad n \geq 1,$$

converge uniformly to $B(z)$ in the circle $|z| \leq 1$. Thus,

$$\sup_{|z| \leq 1} |B_n(z) - B(z)| = r_n \to 0.$$

Then by Cauchy formula,

$$b_L^{(n)} = \frac{1}{2\pi i} \oint_{|z|=1} \frac{B_n(z)}{z^{L+1}}\,\mathrm{d}z.$$

This obviously implies that $|b_L^{(n)} - b_L| \leq r_n$, $L \geq 0$, and therefore $\langle *, F_\lambda \rangle$ is continuous, and hence it is a distribution. Using the condition $\mathrm{d}F_\lambda(x+1) = (1/2)\,\mathrm{d}F_\lambda(x)$, these distributions can be extended to other spaces. Summarizing, we have shown that the Minkowski question mark function has an infinite sequence of "peers" $F_\lambda(x)$ which are also related to continued fraction expansion, in somewhat similar manner. $F(x)$ is the only "nonhomogeneous" one among them.

4. Linear relations among moments M_L

In this section we clarify the nature of linear relations among the moments M_L. This was mentioned in [Alkauskas $\geq$ 2009], but not done in explicit form. Note that the second identity of Equation (7) gives linear relations among moments m_L:

$$m_L = \sum_{s=0}^L \binom{L}{s} (-1)^s m_s, \qquad L \geq 0.$$

These linear relations can be written in terms of M_L. Despite the fact that these relations form a general phenomena for symmetric distributions, in conjunction with the first identity in (7) they give an essential information about $F(x)$. Let us denote

$$q(x,t) = (2 - e^t)e^{xt} - (2e^t - 1)e^{-xt} = \sum_{n=1}^{\infty} Q_n(x)\frac{t^n}{n!}.$$

We see that $Q_n(x)$ are polynomials with integer coefficients and they are given by

$$Q_n(x) = 2x^n - (x+1)^n - 2(1-x)^n + (-x)^n. \tag{16}$$

The following table gives the first few polynomials.

n	$Q_n(x)$	n	$Q_n(x)$
1	$2x - 3$	5	$2x^5 - 15x^4 + 10x^3 - 30x^2 + 5x - 3$
2	$2x - 3$	6	$6x^5 - 45x^4 + 20x^3 - 45x^2 + 6x - 3$
3	$2x^3 - 9x^2 + 3x - 3$	7	$2x^7 - 21x^6 + 21x^5 - 105x^4 + 35x^3 - 63x^2 + 7x - 3$
4	$4x^3 - 18x^2 + 4x - 3$	8	$8x^7 - 84x^6 + 56x^5 - 210x^4 + 56x^3 - 84x^2 + 8x - 3$

Moreover, the following statement holds.

Proposition 1. *Polynomials $Q_n(x)$ have the following properties*:

(i) $Q_{2n}(x) \in L_{\mathbb{Q}}(Q_1(x), Q_3(x), \ldots, Q_{2n-1}(x))$, $\quad n \geq 1$;

(ii) $\deg Q_{2n} = 2n - 1$, $\quad \deg Q_{2n-1} = 2n - 1$, $\quad n \geq 1$;

(iii) $\widehat{Q}_{2n}(x) := (Q_{2n}(x) + 3)/x$ *is reciprocal*: $\widehat{Q}_{2n}(x) = x^{2n-2}\widehat{Q}_{2n}(1/x)$;

(iv) $\int_0^{\infty} Q_n(x)\,dF(x) = 0$.

Naturally, it is property (iv) which makes these polynomials very important in the study of the Minkowski question mark function. Here $L_{\mathbb{Q}}(*)$ denotes the $\mathbb{Q}$-linear space spanned by the specified polynomials.

Proof. (i) Let $q_e(x,t) = (1/2)(q(x,t) + q(x,-t))$, and $q_o(x,t) = (1/2)(q(x,t) - q(x,-t))$. Direct calculation shows that, if $e^t = T$, then

$$2q_e = e^{xt}(3 - T - \frac{2}{T}) + e^{-xt}(3 - \frac{1}{T} - 2T),$$

$$2q_o = e^{xt}(1 - T + \frac{2}{T}) - e^{-xt}(1 - \frac{1}{T} + 2T).$$

This yields

$$\sum_{n=1}^{\infty} Q_{2n}(x)\frac{t^{2n}}{(2n)!} = q_e(x,t) = \frac{T-1}{T+1}q_o(x,t) = \frac{e^t-1}{e^t+1}\sum_{n=0}^{\infty} Q_{2n+1}(x)\frac{t^{2n+1}}{(2n+1)!}.$$

The multiplier on the right, $(e^t - 1)/(e^t + 1) = \tanh(t/2)$, is independent of x, and this obviously proves the part (i). Also, part (ii) follows easily from Equation (16).

(iii) Since $\widehat{Q}_{2n}(x) = (1/x)(3x^{2n} - (x+1)^{2n} - 2(x-1)^{2n} + 3)$, the proof is immediate.

(iv) In fact, Equation (7) gives $(2 - e^t)M(t) = (2e^t - 1)M(-t)$. For real $|t| < \log 2$, we have $M(t) = \int_0^\infty e^{xt} \, dF(x)$. This implies

$$\int_0^\infty q(x, t) \, dF(x) = \sum_{n=0}^\infty \frac{t^n}{n!} \int_0^\infty Q_n(x) \, dF(x) \equiv 0, \quad \text{for } |t| < \log 2,$$

and this completes the proof. $\qquad\square$

Consequently, there exist linear relations among the moments M_L. Thus, for example, part (iv) (in case $n = 1$ and $n = 3$) implies $2M_1 - 3 = 0$ and $2M_3 - 9M_2 + 3M_1 = 3$ respectively. The exact values of M_L belong to the class of constants, which can be thought as emerging from arithmetic-geometric chaos. This resembles the situation concerning polynomial relations among various periods. We will not present the definition of a period (it can be found in [Kontsevich and Zagier 2001]). In particular, the authors conjecture (and there is no support for possibility that it can be proved wrong) that *"if a period has two integral representations, then one can pass from one formula to another using only additivity, change of variables, and Newton–Leibniz formula, in which all functions and domains of integration are algebraic with coefficients in $\overline{\mathbb{Q}}$"*. Thus, for example, the conjecture predicts the possibility to prove directly that

$$\iint_{\frac{x^2}{4} + 3y^2 \leq 1} dx \, dy = \int_{-1}^1 \frac{dx}{\sqrt[3]{(1-x)(1+x)^2}},$$

without knowing that they both are equal to $\frac{2\pi}{\sqrt{3}}$, and this indeed can be done. Similarly, returning to the topic of this paper, we believe that any finite $\mathbb{Q}$-linear relation among the constants M_L can be proved simply by applying the functional equation of $F(x)$, by means of integration by parts and change of variables. The last proposition supports this claim. In other words, we believe that there cannot be any other miraculous coincidences regarding the values of M_L. More precisely, we formulate

Conjecture 1. Suppose, $r_k \in \mathbb{Q}, 0 \leq k \leq L$, are rational numbers such that

$$\sum_{k=0}^L r_k M_k = 0.$$

Let $\ell = \left[\frac{L-1}{2}\right]$. Then

$$\sum_{k=0}^{L} r_k x^k \in L_{\mathbb{Q}}\big(Q_1(x), Q_3(x), \ldots, Q_{2\ell+1}(x)\big).$$

This conjecture, if true, should be difficult to prove. It would imply, for example, that M_L for $L \geq 2$ are irrational. On the other hand, this conjecture seems to be much more natural and approachable, compared to similar conjectures regarding arithmetic nature of constants emerging from geometric chaos, e.g. spectral values s for Maass wave forms (say, for $\mathrm{PSL}_2(\mathbb{Z})$), or those coming from arithmetic chaos, like nontrivial zeros of Riemann's $\zeta(s)$. We cannot give any other evidence, save the last proposition, to support this conjecture.

5. Estimate for the moments m_L

This section deals with an asymptotic estimate for the moments m_L. This result was not obtained before, and in view of the expression in Equation (4), it is of certain number-theoretic interest. This result should be compared with the asymptotic formula for M_L, obtained in [Alkauskas $\geq$ 2009]:

$$M_L \sim \frac{\mathfrak{m}(\log 2)}{2 \log 2}\left(\frac{1}{\log 2}\right)^L L!, \text{ for } L \in \mathbb{N}. \tag{17}$$

A priori, as it is implied by the fact that the radius of convergence of $G(z)$ at $z = 0$ is 1, and by Equation (5), for every $\varepsilon > 0$ and $p > 1$, one has

$$\frac{1}{L^p} \gg m_L \gg (1 - \varepsilon)^L,$$

as $L \to \infty$. More precisely, we have

Theorem 3. *Let* $C = e^{-2\sqrt{\log 2}} = 0.18917\ldots$. *Then the following estimate holds, as* $L \to \infty$:

$$C^{\sqrt{L}} \ll m_L \ll L^{1/4} C^{\sqrt{L}}.$$

Both implied constants are absolute.

Proof. Fix $J \in \mathbb{N}$, and choose an increasing sequence of positive real numbers $\mu_j < 1$, $1 \leq j \leq J$. We will soon specify μ_j in such a way that $\mu_j \to 0$ uniformly as $L \to \infty$. An estimate for m_L is obtained via the defining integral (recall that

$F(x) + F(1/x) = 1)$:

$$m_L = \left(\int_0^{\mu_1} + \sum_{j=1}^{J-1} \int_{\mu_j}^{\mu_{j+1}} + \int_{\mu_J}^{\infty} \right) \left(\frac{1}{x+1} \right)^L \, dF(x)$$

$$< F(\mu_1) + \sum_{j=1}^{J-1} \left(\frac{1}{\mu_j+1} \right)^L F(\mu_{j+1}) + \left(\frac{1}{\mu_J+1} \right)^L.$$

Indeed, in the first integral, the integrand is bounded by 1. In the middle integrals, we choose the largest value of integrand, and change bounds of integration to $[0, \mu_{j+1}]$. The same is done with the last integral, with bounds changed to $[0, \infty)$. Now choose $\mu_j = 1/(c_j \sqrt{L})$ for some decreasing sequence of constants c_j. The functional equation for $F(x)$ implies

$$F(x+n) = 1 - 2^{-n} + 2^{-n} F(x), \quad x \geq 0.$$

Thus, $1 - F(x) \asymp 2^{-x}$, as $x \to \infty$ (the implied constants being min and max of the function $\Psi(x)$; see Figure 3 and Section 8). Using the identity $F(x) + F(1/x) = 1$,
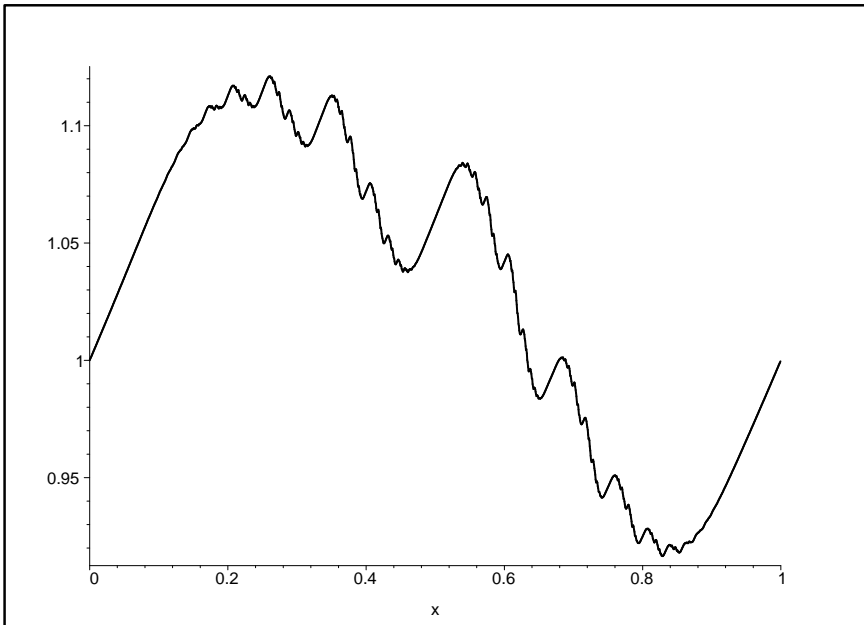
Figure 3. Periodic function $\Psi(x)$.

we therefore obtain

$$
m_L \ll 2^{-c_1\sqrt{L}} + \sum_{j=1}^{J-1} \left(\frac{1}{\frac{1}{c_j\sqrt{L}}+1}\right)^L 2^{-c_{j+1}\sqrt{L}} + \left(\frac{1}{\frac{1}{c_J\sqrt{L}}+1}\right)^L
$$

$$
\ll e^{-\sqrt{L}c_1\log 2} + \sum_{j=1}^{J-1} e^{-\sqrt{L}(\frac{1}{c_j}+c_{j+1}\log 2)} + e^{-\sqrt{L}\frac{1}{c_J}}. \tag{18}
$$

Here we need an elementary lemma.

Lemma 1. *For given $J \in \mathbb{N}$, there exists a unique sequence of positive real numbers $c_1^*, \ldots, c_J^*$, such that*

$$
c_1^* = \frac{1}{c_1^*} + c_2^* = \frac{1}{c_2^*} + c_3^* = \cdots = \frac{1}{c_{J-1}^*} + c_J^* = \frac{1}{c_J^*}.
$$

Moreover, this sequence $\{c_j^*, 1 \le j \le J\}$ is decreasing, and it is given by

$$
c_j^* = \frac{\sin\frac{(j+1)\pi}{J+2}}{\sin\frac{j\pi}{J+2}}, \quad j = 1, 2, \ldots, J \Rightarrow c_1^* = 2\cos\frac{\pi}{J+2}.
$$

Proof. Indeed, we see that $c_1^* = x$ determines the sequence c_j^* uniquely. First, $c_2 = x - 1/x = (x^2 - 1)/x$. Let $F_1(x) = x$, $F_2(x) = x^2 - 1$. Suppose we have shown that $c_j = F_j(x)/F_{j-1}(x)$ for certain sequences of polynomials. Then from the above equations one obtains

$$
c_{j+1} = c_1 - \frac{F_{j-1}(x)}{F_j(x)} = \frac{xF_j(x) - F_{j-1}(x)}{F_j(x)}.
$$

Thus, using induction we see that $c_j = F_j(x)/F_{j-1}(x)$, where polynomials $F_j(x)$ are given by the initial values $F_0(x) = 1$, $F_1(x) = x$ and then for $j \ge 1$ recurrently by $F_{j+1}(x) = xF_j(x) - F_{j-1}(x)$. This shows that $F_j(2x) = U_j(x)$, where $U(x)$ stand for the classical Chebyshev U-polynomials, given by

$$
U_j(\cos\theta) = \frac{\sin(j+1)\theta}{\sin\theta}.
$$

The last equation $c_1^* = 1/c_J^*$ implies $F_{J+1}(x) = 0$. Thus, $U_{J+1}(x/2) = 0$, and all possible values of c_1^* are given by $c_1^* = x = 2\cos\big((k\pi)/(J+2)\big), k = 1, 2, \ldots, J+1$. Thus,

$$
c_j^* = \frac{F_j(x)}{F_{j-1}(x)} = \frac{U_j(x/2)}{U_{j-1}(x/2)} = \frac{\sin\frac{k(j+1)\pi}{J+2}}{\sin\frac{kj\pi}{J+2}}.
$$

Since our concern is only positive solutions, this gives the last statement of lemma. Finally, monotonicity is easily verifiable. Indeed, system of equations imply $c_2^* < c_1^*$, and then we act by induction. ▢

Thus, $c_1^* > 2 - b/J^2$ for some constant $b > 0$. Returning to the proof of the Theorem 3, for given J, let c_j^* be the sequence in the lemma, and let $c_i^* = c_i \sqrt{\log 2}$. Thus,

$$c_1 \log 2 = \frac{1}{c_1} + c_2 \log 2 = \frac{1}{c_2} + c_3 \log 2 = \cdots = \frac{1}{c_{J-1}} + c_J \log 2 = \frac{1}{c_J}.$$

Choosing exactly this sequence for the estimate (18), and using the bound for c_1^*, we get:

$$m_L \ll (J+1)e^{-\sqrt{L}c_1 \log 2} < (J+1)C^{\sqrt{L}}e^{\frac{b\sqrt{\log 2}}{J^2}\sqrt{L}}.$$

Finally, the choice $J = [L^{1/4}]$ establishes the upper bound.

The lower estimate is immediate. In fact, let $\mu = 1/(c\sqrt{L})$. Then

$$m_L > \int_0^\mu \left(\frac{1}{x+1}\right)^L dF(x) > \left(\frac{1}{\mu+1}\right)^L F(\mu) \gg 2^{-c\sqrt{L}} \cdot e^{-\sqrt{L}\frac{1}{c}}$$

The choice $c = \log^{-1/2} 2$ gives the desired bound. ▢

The constants in Theorem 3 can also be calculated without great effort, but this is astray of the main topic of the paper.

It should be noted that, if we start directly from the last definition (3) of m_L, then in the course of the proof of Theorem 3, we use both equalities $F(x) + F(1/x) = 1$ and $2F(x/(x+1)) = F(x)$. Since these two determine $F(x)$ uniquely, generally speaking, our estimate for m_L is characteristic only to $F(x)$. A direct inspection of the proof also reveals that the true asymptotic "action" in the second definition (3) of m_L takes place in the neighborhood of 1. This, obviously, is a general fact for probabilistic distributions with proper support on the interval $[0, 1]$. Additionally, calculations show that the sequence $m_L/(L^{1/4}C^{\sqrt{L}})$ is monotonically decreasing. This is indeed the case, and there exists $\lim_{L\to\infty}(m_L/L^{1/4}C^{\sqrt{L}})$ [Alkauskas 2008].

As a final remark, we note that the result of Theorem 3 must be considered in conjunction with the linear relations $m_L = \sum_{s=0}^L \binom{L}{s}(-1)^s m_s$, $L \geq 0$, and the natural inequalities, imposed by the fact that m_L is a sequence of moments of probabilistic distribution with support on the interval $[0, 1]$. We thus have Hausdorff conditions, which state that for all nonnegative integers m and n, one has

$$2\int_0^1 x^n(1-x)^m \, dF(x) = \sum_{i=0}^m \binom{m}{i}(-1)^i m_{i+n} > 0.$$

This is, of course, the consequence of monotonicity of $F(x)$.

6. Exact sequence

In this section we prove the exactness of a sequence of continuous linear maps, intricately related to the Minkowski question mark function $F(x)$. Let $\mathsf{C}[0, 1]$ denote the space of continuous, complex-valued functions on the interval $[0, 1]$ with supremum norm. For $f \in \mathsf{C}[0, 1]$, one has the identity

$$\int_0^1 f(x)\, dF(x) = \sum_{n=1}^{\infty} \int_0^1 f\left(\frac{1}{x+n}\right) 2^{-n}\, dF(x), \qquad (19)$$

Indeed, using the functional Equation (2), we have

$$\int_0^1 f(x)\, dF(x) = \int_1^{\infty} f\left(\frac{1}{x}\right) dF(x) = \sum_{n=1}^{\infty} \int_0^1 f\left(\frac{1}{x+n}\right) dF(x+n),$$

which is exactly (19). Let C^{ω} denote, as before, the space of analytic functions in the disk $\mathbf{D} = |z - 1/2| \leq 1/2$, including its boundary. We equip this space with the topology of uniform convergence (as a matter of fact, we have a wider choice of spaces; this one is chosen as an important example). Now, consider a continuous functional on C^{ω} given by $T(f) = \int_0^1 f(x)\, dF(x)$, and a continuous noncompact linear operator $[\mathscr{L}f](x) = f(x) - \sum_{n=1}^{\infty} f\left(\frac{1}{x+n}\right) 2^{-n}$. Finally, let i stand for the natural inclusion $i : \mathbb{C} \to \mathsf{C}^{\omega}$.

Theorem 4. *The following sequence of maps is exact*:

$$0 \to \mathbb{C} \xrightarrow{i} \underset{(*)}{\mathsf{C}^{\omega}} \xrightarrow{\mathscr{L}} \underset{(**)}{\mathsf{C}^{\omega}} \xrightarrow{T} \mathbb{C} \to 0. \qquad (20)$$

Proof. First, i is obviously a monomorphism. Let $f \in \mathrm{Ker}(\mathscr{L})$. This means that

$$f(x) = \sum_{n=1}^{\infty} f\big(1/(x+n)\big) 2^{-n}.$$

Let $x_0 \in [0, 1]$ be such that $|f(x_0)| = \sup_{x \in [0,1]} |f(x)|$. Since $\sum_{n=1}^{\infty} 2^{-n} = 1$, this yields

$$f(1/(x_0 + n)) = f(x_0) \quad \text{for} \quad n \in \mathbb{N}.$$

By induction,

$$f([0, n_1, n_2, \dots, n_I + x_0]) = f(x_0)$$

for all $I \in \mathbb{N}$, and all $n_i \in \mathbb{N}$, $1 \leq i \leq I$; here $[\star]$ stands for the (regular) continued fraction. Since this set is everywhere dense in $[0, 1]$ and f is continuous, this forces $f(x) \equiv \text{const}$ for $x \in [0, 1]$. Due to the analytic continuation, this is valid for $x \in \mathbf{D}$ as well. Hence, we have the exactness at the term $(*)$.

Next, T is obviously an epimorphism. Further, the identity in Equation (19) implies that $\text{Im}(\mathscr{L}) \subset \text{Ker}(T)$. The task is to show that indeed we have an equality. At this stage, we need the following lemma. Denote

$$[\mathscr{G} f](x) = \sum_{n=1}^{\infty} f\big(1/(x+n)\big) 2^{-n}.$$

Lemma 2. *Let* $f \in C^{\omega}$. *Then* $[\mathscr{G}^n f](x) = 2T(f) + O(\gamma^{-2n})$ *for* $x \in \mathbf{D}$; *here* $T(f)$ *stands for the constant function,* $\gamma = \big((1+\sqrt{5})/2\big)$ *is the golden section, and the bound implied by* O *is uniform for* $x \in \mathbf{D}$.

Proof. In fact, the lemma is true for any function with continuous derivative. Let $x \in \mathbf{D}$. We have

$$[\mathscr{G}^r f](x) = \sum_{n_1,n_2,\ldots,n_r=1}^{\infty} 2^{-(n_1+n_2+\cdots+n_r)} f([0, n_1, n_2, \ldots, n_r + x]).$$

The direct inspection of this expression and Equation (1) shows that this is exactly twice the Riemann sum for the integral $\int_0^1 f(x)\, dF(x)$, corresponding to the division of unit interval into intervals with endpoints being $[0, n_1, n_2, \ldots, n_r]$, $n_i \in \mathbb{N}$. From the basic properties of Möbius transformations we inherit that the set $[0, n_1, n_2, \ldots, n_r + x]$ for $x \in \mathbf{D}$ is a circle $\mathbf{D}_r$ whose diagonal is one of these intervals, say I_r. For fixed r, the largest of these intervals has endpoints F_{r-1}/F_r and F_r/F_{r+1}, where F_r stands for the usual Fibonacci sequence. Thus, its length is $1/(F_r F_{r+1}) \sim c\gamma^{-2r}$. Let $x_0, x_1 \in \mathbf{D}_r$, and $\sup_{x \in \mathbf{D}} |f'(x)| = A$. We have

$$\sup_{x_0,x_1 \in \mathbf{D}_r} |f(x_0) - f(x_1)| \leq Ac\gamma^{-2r}.$$

Thus, the Riemann sum deviates from the Riemann integral no more than

$$|[\mathscr{G}^r f](x) - 2T(f)| \leq Ac\gamma^{-2r} \sum_{n_1,n_2,\ldots,n_r=1}^{\infty} 2^{-(n_1+n_2+\cdots+n_r)} = Ac\gamma^{-2r}.$$

This proves the Lemma. ☐

Thus, let $f \in \text{Ker}(T)$. All we need is to show that the equation $f = g - \mathscr{G}g$ has a solution $g \in C^{\omega}$. Indeed, let $g = f + \sum_{n=1}^{\infty} \mathscr{G}^n f$. By the above lemma, $\|\mathscr{G}^n f\| = O(\gamma^{-2n})$. Thus, the series defining g converges uniformly and hence g is an analytic function. Finally, $g - \mathscr{G}g = f$; this shows that $\text{Ker}(T) \subset \text{Im}(\mathscr{L})$ and the exactness at the term $(\ast\ast)$ is proved. ☐

These results imply that, for example, $Q := \text{Im}(\mathscr{L})$ is a linear subspace of C^{ω} of codimension 1. Further research proves that $\mathscr{L}|_Q$ is an isomorphism.

The eigenfunctions of $\mathscr{S}$ acting on the space C^ω are given by

$$G^\star(-x) = \int_0^{-x} G_\lambda(z)\,dz + \int_{-1}^0 G_\lambda(z)\,dz$$

(see Equations (22) and (23) in Section 7). Thus, the problem of convergence of $\mathscr{S}^n f$ is completely analogous to the problem of convergence for the iterates of the Gauss–Kuzmin–Wirsing operator. Let us remind that if $f \in C[0, 1]$, it is given by

$$[\mathbf{W}f](x) = \sum_{n=1}^\infty \frac{1}{(x+n)^2} f\left(\frac{1}{x+n}\right).$$

Dominant eigenvalue 1 correspond to an eigenfunction $1/(1+x)$. As it was proved by Kuzmin, provided that $f(x)$ has a continuous derivative, there exists $c > 0$, such that

$$[\mathbf{W}^n f](x) = \frac{A}{1+x} + O(e^{-c\sqrt{n}}), \text{ as } n \to \infty; \quad A = \frac{1}{\log 2} \int_0^1 f(x)\,dx.$$

The proof can be found in [Khinchin 1964]. Note that this was already conjectured by Gauss, but he did not give the proof nor for the main neither for the error term. For the most important case, when $f(x) = 1$, Lévy established the error term of the form $O(C^n)$ for $C = 0.7$. Finally, Wirsing [1973/74] gave the exact result in terms of eigenfunctions of $\mathbf{W}$, establishing the error term of the form $c^n \Psi(x) + O(x(1-x)\mu^n)$, where $c = -0.303663\ldots$ is the subdominant eigenvalue (the Gauss–Kuzmin–Wirsing constant), $\Psi(x)$ is a corresponding eigenfunction, and $\mu < |c|$. Returning to our case, we have completely analogous situation: operator $\mathbf{W}$ is replaced by $\mathscr{S}$, and the measure dx is replaced by $dF(x)$. The leading eigenvalue 1 corresponds to the constant function. However strange, Wirsing did not notice that eigenvalues of $\mathbf{W}$ are in fact eigenvalues of certain Hilbert–Schmidt operator. This was later clarified by Babenko [1978]. Recently, the Gauss–Kuzmin–Lévy theorem was generalized by Manin and Marcolli [2002]. The paper is very rich in ideas and results; in particular, it sheds a new light on the theorem just mentioned.

Concerning spaces for which Theorem 4 holds, we can investigate the space $C[0, 1]$ as well. However, if $f \in C[0, 1]$ and $f \in \text{Ker}(T)$, the significant difficulty arises in proving uniform convergence of the series $\sum_{n=0}^\infty \mathscr{S}^n f$. Moreover, operator $\mathscr{S}$, acting on the space $C[0, 1]$, has additional point spectra apart from λ. Indeed, let

$$P_n(y) = y^n + \sum_{i=0}^{n-1} a_i y^i$$

be a polynomial of degree n which satisfies yet another variation of three-term functional equation

$$2P_n(1-2y) - P_n(1-y) = \frac{1}{\delta_n} P_n(y)$$

for certain δ_n. The comparison of leading terms shows that

$$\delta_n = \frac{(-1)^n}{2^{n+1}-1},$$

and that indeed for this δ_n there exists a unique polynomial, since each coefficient a_j can be uniquely determined with the knowledge of coefficients a_i for $i > j$. Thus,

$$P_1(y) = y - \frac{1}{4}, \qquad\qquad P_2(y) = y^2 - \frac{3}{5}y + \frac{1}{15},$$

$$P_3(y) = y^3 - \frac{21}{22}y^2 + \frac{3}{11}y - \frac{7}{352}, \quad P_4(y) = y^4 - \frac{30}{23}y^3 + \frac{14}{23}y^2 - \frac{45}{391}y + \frac{37}{5865}.$$

The equation for $P_n(y)$ implies that (after a substitution $y \mapsto 2^{-\ell}y$ and division by 2^ℓ)

$$2^{1-\ell}P_n(1-2^{1-\ell}y) - 2^{-\ell}P_n(1-2^{-\ell}y) = \delta_n^{-1}2^{-\ell}P_n(2^{-\ell}y).$$

Now, sum this over $\ell \in \mathbb{N}$, and finally substitute $y \mapsto 1-y$. This gives

$$\delta_n P_n(y) = \sum_{\ell=1}^{\infty} \frac{1}{2^\ell} P_n\left(\frac{1-y}{2^\ell}\right). \tag{21}$$

Then we have:

Proposition 2. *The function $P_n(F(x))$ is the eigenfunction of $\mathcal{S}$, acting on the space $C[0,1]$, and eigenvalue $(-1)^n/(2^{n+1}-1)$ belongs to the point spectra of $\mathcal{S}$.*

Proof. Indeed,

$$[\mathcal{S}(P_n \circ F)](x) = \sum_{\ell=1}^{\infty} \frac{1}{2^\ell} P_n \circ F\left(\frac{1}{x+\ell}\right) \overset{(2)}{=} \sum_{\ell=1}^{\infty} \frac{1}{2^\ell} P_n\left(1 - F(x+\ell)\right) \overset{(2)}{=}$$

$$= \sum_{\ell=1}^{\infty} \frac{1}{2^\ell} P_n\left(2^{-\ell} - 2^{-\ell}F(x)\right) \overset{(21)}{=} \delta_n P_n(F(x)). \qquad \square$$

Thus, the operator $\mathcal{S}$ behaves differently in spaces $C[0,1]$ and C^ω. We postpone the analysis of this operator in various spaces for the future.

7. Integrals involving $F(x)$

In this section we calculate certain integrals. Only rarely it is possible to express an integral involving $F(x)$ in closed form. In fact, all results we possess come from the identity $M_1 = 3/2$, and any iteration of identities similar to (19). The following theorem adds identities of quite a different sort.

Theorem 5. *Let $G_\lambda(z)$ be any function that satisfies the hypotheses of Theorem 2. Then*

(i) $\frac{\lambda}{\lambda+1} \int_0^1 G_\lambda(-x)\,dx = \int_0^1 G_\lambda(-x)F(x)\,dx;$

(ii) $-\int_0^1 \log x\,dF(x) = 2\int_0^1 \log(1+x)\,dF(x) = \int_0^1 G(-x)\,dx;$

(iii) $\int_0^1 G(-x)(1+x^2)\,dF(x) = \frac{1}{4};$

(iv) $\int_0^1 G_\lambda(-x)\left(1-\frac{x^2}{\lambda}\right)\,dF(x) = 0.$

Proof. We first prove identity (i). By (14), for every integer $n \geq 1$, we have

$$2G_\lambda(-z-n+1) - G_\lambda(-z-n) = \frac{1}{\lambda(z+n)^2}\, G_\lambda\left(-\frac{1}{z+n}\right).$$

Divide this by 2^n and sum over $n \geq 1$. By Theorem 1, the sum on the left is absolutely convergent. Thus,

$$G_\lambda(-z) = \sum_{n=1}^{\infty} \frac{1}{\lambda 2^n (z+n)^2}\, G_\lambda\left(-\frac{1}{z+n}\right).$$

Let $G_\lambda^\star(x) = \int_0^x G_\lambda(z)\,dz$. In terms of $G_\lambda^\star(x)$, the last identity reads as

$$-G_\lambda^\star(-x) = \sum_{n=1}^{\infty} \frac{1}{\lambda 2^n} G_\lambda^\star\left(-\frac{1}{x+n}\right) - \sum_{n=1}^{\infty} \frac{1}{\lambda 2^n} G_\lambda^\star\left(-\frac{1}{n}\right). \qquad (22)$$

In particular, setting $x = 1$, one obtains

$$\sum_{n=1}^{\infty} \frac{1}{\lambda 2^n} G_\lambda^\star\left(-\frac{1}{n}\right) = \left(\frac{1}{\lambda}-1\right)G_\lambda^\star(-1). \qquad (23)$$

Now we are able to calculate the following integral (we use integration by parts in Stieltjes integral twice).

$$\int_0^1 G_\lambda(-x)F(x)\,dx$$

$$= -\int_0^1 \frac{d}{dx}G_\lambda^\star(-x)F(x)\,dx = -\frac{1}{2}G_\lambda^\star(-1) + \int_0^1 G_\lambda^\star(-x)\,dF(x)$$

$$\overset{(22)}{=} -\frac{1}{2}G_\lambda^\star(-1) + \frac{1}{2}\sum_{n=1}^\infty \frac{1}{\lambda 2^n}G_\lambda^\star\left(-\frac{1}{n}\right) - \frac{1}{\lambda}\sum_{n=1}^\infty \int_0^1 G_\lambda^\star\left(-\frac{1}{x+n}\right)2^{-n}\,dF(x)$$

$$\overset{(19),(23)}{=} -\frac{1}{2}G^\star(-1) + \frac{1}{2}\left(\frac{1}{\lambda}-1\right)G_\lambda^\star(-1) - \frac{1}{\lambda}\int_0^1 G_\lambda^\star(-x)\,dF(x)$$

$$= -G^\star(-1) - \frac{1}{\lambda}\int_0^1 G_\lambda(-x)F(x)\,dx.$$

Thus, the same integral is on the both sides, and this gives

$$\int_0^1 G_\lambda(-x)F(x)\,dx = -\frac{\lambda}{\lambda+1}G_\lambda^\star(-1).$$

This establishes the statement (i).

Now we proceed with the second identity. Integral (11) and the Fubini theorem imply

$$\int_0^1 G(-z)\,dz = 2\int_0^1\int_0^1 \frac{x}{1+xz}\,dz\,dF(x) = 2\int_0^1 \log(1+x)\,dF(x).$$

Lastly, we apply (19) twice to obtain the needed equality. Indeed,

$$I = \int_0^1 \log(1+x)\,dF(x) \overset{(19)}{=} \sum_{n=1}^\infty \frac{1}{2^n}\int_0^1 \log\left(1+\frac{1}{x+n}\right)dF(x)$$

$$= \sum_{n=1}^\infty \frac{1}{2^n}\int_0^1 \log(x+n)\,dF(x) - I \overset{(19)}{=} -\int_0^1 \log x\,dF(x) - I.$$

This finishes the proof of (ii).

In proving (iii), we can be more concise, since the pattern of the proof goes along the same line. One has

$$G(-z) = -\sum_{n=1}^\infty \frac{1}{2^n(z+n)^2}G\left(-\frac{1}{z+n}\right) + \sum_{n=1}^\infty \frac{1}{2^n(z+n)}.$$

Thus,

$$\int_0^1 G(-x)\,dF(x) = -\sum_{n=1}^{\infty} \int_0^1 \frac{1}{2^n(x+n)^2}\,G\left(-\frac{1}{x+n}\right)\,dF(x) +$$

$$\sum_{n=1}^{\infty} \int_0^1 \frac{1}{2^n(x+n)}\,dF(x) \overset{(19)}{=} -\int_0^1 x^2\,G(-x)\,dF(x) + \int_0^1 x\,dF(x).$$

Since $\int_0^1 x\,dF(x) = \frac{m_1}{2} = \frac{1}{4}$, this finishes the proof of (iii). Part (iv) is completely analogous. $\square$

Part (iii), unfortunately, gives no new information about the sequence m_L. Indeed, the identity can be rewritten as

$$\sum_{L=1}^{\infty} m_L(-1)^{L-1}(m_{L-1}+m_{L+1}) = 1/2,$$

which, after regrouping, turns into the identity $m_0 m_1 = 1/2$.

Concerning part (iv), and taking into account Theorem 4, one could expect that in fact $\mathrm{Ker}(T)$ is equal to the closure of vector space spanned by functions $G_\lambda(-x)(1-x^2/\lambda)$. If this is the case, then these functions, along with $G(z)(1+x^2)$, produce a Schauder basis for C^ω. Thus, if

$$x^L = \sum_\lambda a_L^{(\lambda)}\,G_\lambda(-x)(1-x^2/\lambda),$$

then $a_L^{(-1)} = 2m_L$. We hope to return to this point in the future.

Concerning (i), note that the values of both integrals depend on the normalization of G_λ, since it is an eigenfunction. Replacing $G_\lambda(z)$ by $cG_\lambda(z)$ for some $c \in \mathbb{R}$, we deduce that the left integral is equal to 1 or 0. Then (i) states that $\int_0^1 F(x)G_\lambda(-x)\,dx = \lambda/(\lambda+1)$ or 0 (apparently, it is never equal to 0). The presence of $\lambda+1$ in the denominator should come as no surprise, minding that λ is the eigenvalue of the Hilbert–Schimdt operator. The Fredholm alternative gives us a way of solving the integral equation in terms of eigenfunctions. Since $|\lambda| \le \lambda_1 = 0.25553210\ldots < 1$, the integral equation is *a posteriori* solvable, and $\lambda+1$ appears in the denominators. Curiously, it is possible to approach this identity numerically. One of the motivations is to check its validity, since the result heavily depends on the validity of almost all the preceding results in [Alkauskas $\ge$ 2009]. The left integral causes no problems, since Taylor coefficients of $G_\lambda(z)$ can be obtained at high precision as an eigenvector of a finite matrix, which is the truncation of an infinite one. On the other hand, the right integral can be evaluated with less precision, since it involves $F(x)$, and thus requires more time and space

consuming continued fractions algorithm. Nevertheless, the author of this paper has checked it with a completely satisfactory outcome, confirming the validity.

Just as interestingly, results (i) and (iv) can be considered a reflection of a "pair-correlation" between eigenvalues λ and eigenvalue -1 (see Section 3 for some remarks on this topic). Moreover, minding properties of the distributions $F_\mu(x)$ (here μ simply means another eigenvalue), the following formal result can be obtained. Given the conditions enforced on F_μ by (15), the identity (19) is replaced by (for $f \in C^\omega$)

$$\int_0^1 f(x)\, dF_\mu(x) = -\frac{1}{\mu} \sum_{n=1}^\infty \int_0^1 f\left(\frac{1}{x+n}\right) 2^{-n}\, dF_\mu(x).$$

Then our trick works smoothly again, and this yields an identity

$$\int_0^1 G_\lambda(-x)(\lambda + \mu x^2)\, dF_\mu(x) = 0.$$

This fact is an interesting example of pair-correlation between eigenvalues of the Hilbert–Schmidt operator in (9). Using a definition of the distribution F_μ, the last identity is equivalent to

$$\sum_{L=1}^\infty (-1)^L \left(m_L^{(\mu)} m_{L+1}^{(\lambda)} \lambda - m_L^{(\lambda)} m_{L+1}^{(\mu)} \mu\right) = 0,$$

and thus is a property of "orthogonality" of $G_\lambda(z)$. This expression is symmetric regarding μ and λ. As could be expected, it is void in case $\mu = \lambda$. As a matter of fact, the proof of the above identity is fallacious, since the definition of distributions F_λ does not imply properties (15) (these simply have no meaning). Nevertheless, numerical calculations suggest that the last identity truly holds. We also hope to return to this topic in the future.

8. Fourier series

The Minkowski question mark function $F(x)$, originally defined for $x \geq 0$ by Equation (1), can be extended naturally to $\mathbb{R}$ simply by the functional equation

$$F(x+1) = 1/2 + 1/2F(x).$$

Such an extension is still given by the expression (1), with the difference that a_0 can be negative integer. Naturally, the second functional equation is not preserved for negative x. Thus, we have

$$2^{x+1}(F(x+1) - 1) = 2^x(F(x) - 1) \text{ for } x \in \mathbb{R}.$$

So, $2^x(F(x)-1)$ is a periodic function, which we will denote by $-\Psi(x)$. Figure 3 gives the graph of $\Psi(x)$ for $x \in [0, 1]$. Thus,

$$F(x) = -2^{-x}\Psi(x) + 1.$$

Since $F(x)$ is singular, the same is true for $\Psi(x)$: it is differentiable almost everywhere, and for these regular points one has $\Psi'(x) = \log 2 \cdot \Psi(x)$. As a periodic function, it has an associated Fourier series expansion

$$\Psi(x) \sim \sum_{n=-\infty}^{\infty} c_n e^{2\pi i n x}.$$

Since $F(x)$ is real function, this gives $c_{-n} = \overline{c_n}$, $n \in \mathbb{Z}$. Let for $n \geq 1$, $c_n = a_n + i b_n$, and $a_0 = c_0/2$, $b_0 = 0$. Here we list initial numerical values for

$$c_n^\star = c_n(2\log 2 - 4\pi i n)$$

(see Proposition 3 for the reason of this normalization).

$c_0^\star = 1.428159,$ $\qquad\qquad$ $c_3^\star = +0.128533 - 0.026840i,$ $c_6^\star = -0.262601 + 0.004128i,$

$c_1^\star = -0.521907 + 0.148754i,$ $c_4^\star = -0.140524 - 0.021886i,$ $c_7^\star = +0.198742 - 0.013703i,$

$c_2^\star = -0.334910 - 0.017869i,$ $c_5^\star = +0.285790 + 0.003744i,$ $c_8^\star = -0.008479 + 0.024012i.$

It is important to note that we do not pose the question about the convergence of this Fourier series. For instance, Salem [1943] and Reese [1989] give examples of singular monotone increasing functions $f(x)$, whose Fourier–Stieltjes coefficients $\int_0^1 e^{2\pi i n x} \, df(x)$ do not vanish, as $n \to \infty$. Salem [1943] even investigated $f(x) = {?}(x)$. In our case, the convergence problem is far from clear. Nevertheless, in all cases we substitute $-2^{-x}\Psi(x)$ instead of $(F(x)-1)$ under an integral. Let, for example, $W(x)$ be a continuous function of at most polynomial growth, as $x \to \infty$, and let $\Psi_N(x) = \sum_{n=-N}^{N} c_n e^{2\pi i n x}$. Then

$$\left| \int_0^\infty W(x)\Big((F(x)-1) + 2^{-x}\Psi_N(x)\Big) \, dx \right|$$

$$\ll \sum_{r=0}^{\infty} |W(r)| 2^{-r} \cdot \int_0^1 |2^x(F(x)-1) + \Psi_N(x)| \, dx.$$

Since $2^x(F(x)-1) \in \mathcal{L}_2[0,1]$, the last integral tends to 0, as $N \to \infty$. As it was said, this makes the change of $(F(x)-1)$ into $-2^{-x}\Psi(x)$ under integral legitimate, and this also justifies term-by-term integration. Henceforth, we will omit a step of changing $\Psi(x)$ into $\Psi_N(x)$, and taking a limit $N \to \infty$.

A general formula for the Fourier coefficients is given by

Proposition 3. *Fourier coefficients c_n are related to special values of exponential generating function $\mathfrak{m}(t)$ through the equality*

$$c_n = \frac{\mathfrak{m}(\log 2 - 2\pi i n)}{2\log 2 - 4\pi i n}, \text{ and } c_n = O(n^{-1}).$$

Proof. We have (note that $F(1) = 1/2$):

$$c_n = -\int_0^1 2^x (F(x) - 1) e^{-2\pi i n x} \, \mathrm{d}x = -\frac{1}{\log 2 - 2\pi i n} \int_0^1 (F(x) - 1) \, \mathrm{d}e^{x(\log 2 - 2\pi i n)}$$

$$= \frac{1}{\log 2 - 2\pi i n} \int_0^1 e^{x(\log 2 - 2\pi i n)} \, \mathrm{d}F(x) = \frac{\mathfrak{m}(\log 2 - 2\pi i n)}{2\log 2 - 4\pi i n}.$$

The last assertion of the proposition is obvious. $\qquad\square$

This proposition is a good example of intrinsic relations among the three functions $F(x)$, $G(z)$ and $\mathfrak{m}(t)$. Indeed, the moments m_L of $F(x)$ give Taylor coefficients of $G(z)$, which are proportional (up to the factorial multiplier) to Taylor coefficients of $\mathfrak{m}(t)$. Finally, special values of $\mathfrak{m}(t)$ on a discrete set of vertical line produce Fourier coefficients of $F(x)$.

Proposition 4 describes explicit relations among Fourier coefficients and the moments. Additionally, in the course of the proof we obtain the expansion of $G(z)$ for negative real z in terms of incomplete gamma integrals.

Proposition 4. *For $L \geq 1$, one has*

$$M_L = L! \sum_{n \in \mathbb{Z}} \frac{c_n}{(\log 2 - 2\pi i n)^L}. \tag{24}$$

Proof. Let $z < 0$ be fixed negative real. Then integration by parts gives

$$G(z+1) = \int_0^\infty \frac{x}{1 - xz} \, \mathrm{d}(F(x) - 1) = \int_0^\infty \frac{1}{(1 - xz)^2} 2^{-x} \Psi(x) \, \mathrm{d}x$$

$$= \sum_{n=-\infty}^\infty c_n \int_0^\infty \frac{1}{(1 - xz)^2} 2^{-x} e^{2\pi i n x} \, \mathrm{d}x = \sum_{n=-\infty}^\infty c_n V_n(z),$$

where

$$V_n(z) = \int_0^\infty \frac{1}{(1 - xz)^2} e^{-x(\log 2 - 2\pi i n)} \, \mathrm{d}x$$

$$= \frac{1}{\log 2 - 2\pi i n} \int_0^{\infty(\log 2 - 2\pi i n)} \frac{1}{(1 - \frac{yz}{\log 2 - 2\pi i n})^2} e^{-y} \, \mathrm{d}y.$$

Since by our convention $z < 0$, the function under integral does not have poles for $\Re y > 0$, and Jordan's Lemma gives

$$V_n(z) = \frac{1}{\log 2 - 2\pi in} \int_0^\infty \frac{1}{(1 - \frac{yz}{\log 2 - 2\pi in})} e^{-y} \, dy$$

$$= \frac{1}{\log 2 - 2\pi in} \cdot V\left(\frac{z}{\log 2 - 2\pi in}\right), \text{ where } V(z) = \int_0^\infty \frac{1}{(1 - yz)^2} e^{-y} \, dy.$$

The function $V(z)$ is defined for the same values of z as $G(z+1)$ and therefore is defined in the cut plane $\mathbb{C} \setminus (0, \infty)$. Consequently, this implies

$$G(z+1) = \sum_{n \in \mathbb{Z}} \frac{c_n}{\log 2 - 2\pi in} \cdot V\left(\frac{z}{\log 2 - 2\pi in}\right). \qquad (25)$$

The formula is only valid for real $z < 0$. The obtained series converges uniformly, since $|1 - y\frac{z}{\log 2 - 2\pi in}| \geq 1$ for $n \in \mathbb{Z}$ and $z < 0$. Since

$$V\left(\frac{1}{z}\right) = -ze^{-z} \int_1^\infty \frac{1}{y^2} e^{yz} \, dy,$$

this gives us the expansion of $G(z+1)$ on a negative real line in terms of incomplete gamma integrals. As noted before, and this can be seen from Equation (5), the function $G(z)$ has all left derivatives at $z = 1$. Further, the $(L-1)$-fold differentiation of $V(z)$ gives

$$V^{(L-1)}(z) = L! \int_0^\infty \frac{y^{L-1}}{(1 - yz)^{L+1}} e^{-y} \, dy \Rightarrow V^{(L-1)}(0) = L!(L-1)!.$$

Comparing (25) with (5) and (10), this gives the desired relation among moments M_L and Fourier coefficients, as stated in the proposition. $\qquad \square$

It is important to compare this expression with the first equality of (7). Indeed, since $\mathfrak{m}(t)$ is entire, that equality via the Cauchy residue formula implies (17). It is exactly the leading term in (24), corresponding to $n = 0$.

9. Associated zeta function

Recall that for complex c and s, c^s is a multivalued complex function, defined as $e^{s \log c} = e^{s(\log |c| + i \arg(c))}$. Henceforth, we fix the branch of the logarithm by requiring that the value of $\arg c$ for c in the right half plane $\Re c > 0$ be in the range $(-\pi/2, \pi/2)$. Thus, if $s = \sigma + it$, and if we denote $r_n = \log 2 + 2\pi in$, then $|r_n^{-s}| = |r_n|^{-\sigma} e^{t \arg r_n} \sim |r_n|^{-\sigma} e^{\pm \pi t/2}$ as $n \to \pm\infty$. Minding this convention and the identity (24), we introduce the zeta function, associated with the Minkowski question mark function.

Definition 1. The dyadic zeta function $\zeta_M(s)$ is defined in the half plane $\Re s > 0$ by the series

$$\zeta_M(s) = \sum_{n \in \mathbb{Z}} \frac{c_n}{(\log 2 - 2\pi i n)^s}, \tag{26}$$

where c_n are Fourier coefficients of $\Psi(x)$, and for each n, $(\log 2 - 2\pi i n)^s$ is understood in the meaning just described.

Then we have

Theorem 6. $\zeta_M(s)$ *has an analytic continuation as an entire function to the whole plane* $\mathbb{C}$, *and satisfies the functional equation*

$$\zeta_M(s)\Gamma(s) = -\zeta_M(-s)\Gamma(-s). \tag{27}$$

Further, $\zeta_M(L) = M_L/L!$ *for* $L \geq 0$. $\zeta_M(s)$ *has trivial zer s for negative integers*: $\zeta_M(-L) = 0$ *for* $L \geq 1$ *and* $\zeta_M'(-L) = (L-1)!(-1)^L M_L$. *Additionally*, $\zeta_M(s)$ *is real on the real line, and thus* $\zeta_M(\bar{s}) = \overline{\zeta_M(s)}$. *The behavior of* $\zeta_M(s)$ *in vertical strips is given by estimate*

$$|\zeta_M(\sigma + it)| \ll t^{-\sigma - 1/2} \cdot e^{\pi |t|/2}$$

uniformly for $a \leq \sigma \leq b$, $|t| \to \infty$.

As we will see, these properties are immediate (subject to certain regularity conditions) for any distribution $f(x)$ with a symmetry property $f(x) + f(1/x) = 1$. Nevertheless, it is a unique characteristic of $F(x)$ that the corresponding zeta function can be given a Dirichlet series expansion, like Equation (26). We do not give the proof of the converse result, since there is no motivation for this. But empirically, we see that this functional equation is equivalent exactly to the symmetry property. Additionally, the presence of a Dirichlet series expansion yields a functional equation of the kind $f(x+1) = 1/2 f(x) + 1/2$. Generally speaking, these two together are unique for $F(x)$. Note also that the functional equation implies that $\zeta_M(it)\Gamma(1 + it) = \int_0^\infty x^{it} \, dF(x)$ is real for real t. Figures 4 and 5 shows its graph for $1.5 \leq t \leq 180$. Further calculations support the claim that this function has infinitely many zeros on the critical line $\Re s = 0$. On the other hand, numerical calculations of contour integrals reveal that there exist many more zeros apart from these. We need one classical integral.

Lemma 3. *Let* A *be real number*, $\arctan(A) = \phi \in (-\pi/2, \pi/2)$, *and* $\Re s > 0$. *Then*

$$\int_0^\infty x^{s-1} e^{-x} \cos(Ax) \, dx = \frac{1}{(1 + A^2)^{s/2}} \cos(\phi s)\Gamma(s).$$

The same is valid with cos *replaced by* sin *on both sides.*
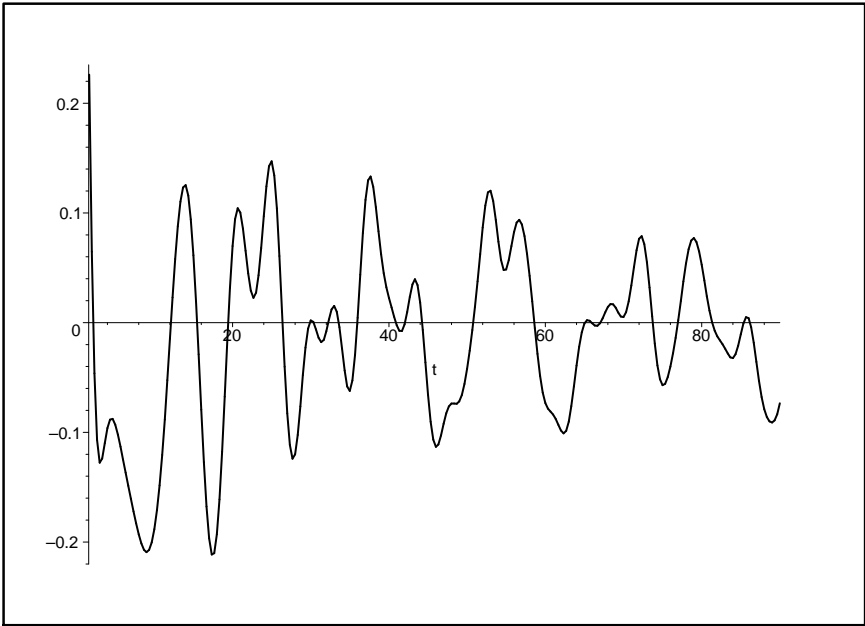
Figure 4. $\zeta_{\mathcal{M}}(it)\Gamma(1+it)$, $1.5 \leq t \leq 90$.
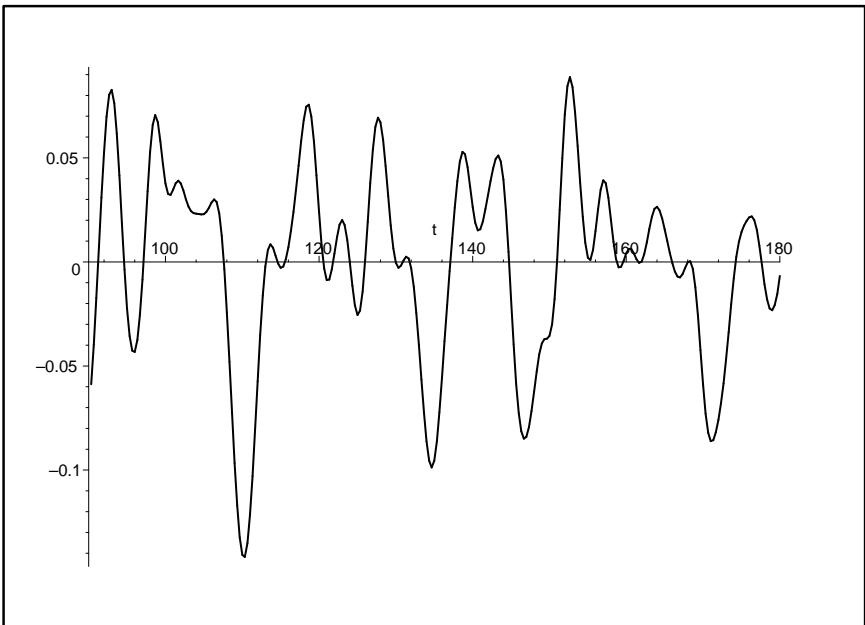
Figure 5. $\zeta_{\mathcal{M}}(it)\Gamma(1+it)$, $90 \leq t \leq 180$.

This can be found in any extensive table of gamma integrals or tables of Mellin transforms.

Proof of Theorem 6. Let for $n \geq 0$, arctan $(2\pi n / \log 2) = \phi_n$. We will calculate the following integral. Let $\Re s > 0$. Then integrating by parts and using Lemma 3, one obtains

$$\int_0^\infty x^s \, \mathrm{d}(F(x) - 1) = s \int_0^\infty x^{s-1} 2^{-x} \Psi(x) \, \mathrm{d}x = s \sum_{n \in \mathbb{Z}} c_n \int_0^\infty x^{s-1} 2^{-x} e^{2\pi i n x} \, \mathrm{d}x$$

$$= s \sum_{n=0}^\infty \int_0^\infty x^{s-1} \Big(2a_n \cos(2\pi n x) - 2b_n \sin(2\pi n x) \Big) 2^{-x} \, \mathrm{d}x$$

$$= 2s\Gamma(s) \sum_{n=0}^\infty | \log 2 + 2\pi n i |^{-s} \Big(a_n \cos(\phi_n s) - b_n \sin(\phi_n s) \Big)$$

$$= s\Gamma(s) \sum_{n \in \mathbb{Z}} \frac{c_n}{(\log 2 - 2\pi i n)^s}.$$

Note that the function $\int_0^\infty x^s \, \mathrm{d}F(x)$ is clearly analytic and entire. Thus, $s\Gamma(s)\,\zeta_{\mathcal{M}}(s)$ is an entire function, and this proves the first statement of the theorem. Since $F(x) + F(1/x) = 1$, this gives $\int_0^\infty x^s \, \mathrm{d}F(x) = \int_0^\infty x^{-s} \, \mathrm{d}F(x)$, and this, in turn, implies the functional equation. All other statements follow easily from this, our previous results, and known properties of the Γ function. In particular, if $s = \sigma + it$,

$$| \zeta_{\mathcal{M}}(s)\Gamma(s+1) | \leq \int_0^\infty |x^s| \, \mathrm{d}F(x) = \zeta_{\mathcal{M}}(\sigma)\Gamma(\sigma+1),$$

and the last statement of the theorem follows from the Stirling's formula for the Γ function: $|\Gamma(\sigma + it)| \sim \sqrt{2\pi}\, t^{\sigma - 1/2} e^{-\pi |t|/2}$ uniformly for $a \leq \sigma \leq b$, as $|t| \to \infty$. $\square$

At this stage, we remark on the similarity and differences with classical results known for the Riemann zeta function $\zeta(s) = \sum_{n=1}^\infty (1/n^s)$. Let $\theta(x)$ denote the usual theta function $\theta(x) = \sum_{n \in \mathbb{Z}} e^{\pi i n^2 x}$, $\Im x > 0$. The following table summarizes all the ingredients which eventually produce the functional equation both for $\zeta(s)$ and $\zeta_{\mathcal{M}}(s)$.

Function	$\zeta(s)$	$\zeta_{\mathcal{M}}(s)$
Dirichlet series exp. Functional equation	Periodicity: $\theta(x+2) = \theta$ $\theta(ix) = (1/\sqrt{x})\theta(i/x)$	Periodicity: $F'(x+1) = (1/2)F'(x)$ $F'(x) = -F'(1/x)$

Since $F(x)$ is a singular function, its derivative should be considered as a distribution on the real line. For this purpose, it is sufficient to consider a distribution $U(x)$ as a derivative of a continuous function $V(x)$, for which the scalar

product $\langle U, f \rangle$, defined for functions $f \in C^\infty(\mathbb{R})$ with compact support, equals $-\langle V, f' \rangle = -\int_{\mathbb{R}} f'(x) V(x) \, dx$. Thus, both $\theta(x)$ and $2^x F'(x)$ are periodic distributions. This guarantees that the appropriate Mellin transform can be factored into the product of Dirichlet series and gamma factors. Finally, the functional equation for the distribution produces the functional equation for the Mellin transform. The difference arises from the fact that for $\theta(x)$ the functional equation is symmetry property on the imaginary line, whereas for $F'(x)$ we have the symmetry on the real line instead. This explains the unusual fact that in Equation (26) we have the summation over the discrete set of the vertical line, instead of the summation over integers.

We will finish by proving another result, which links $\zeta_{\mathcal{M}}(s)$ to the Mellin transform of $G(-z + 1)$. This can be done using expansion (25), but we rather chose a direct way. Let $\int_0^\infty G(-z + 1) z^{s-1} \, dz = G^*(s)$. The symmetry property for Theorem 1 implies that $G(-z + 1)$ has a simple zero, as $z \to \infty$ along the positive real line. Thus, basic properties of Mellin transform imply that $G^*(s)$ is defined for $0 < \Re s < 1$. For these values of s, we have the following classical integral:

$$\int_0^\infty \frac{z^{s-1}}{1+z} \, dz \overset{\frac{z}{1+z} \to x}{=} \int_0^1 x^{s-1}(1-x)^{-s} \, dx = \Gamma(s)\Gamma(1-s) = \frac{\pi}{\sin \pi s}.$$

Thus, using Equation (11), we get

$$\begin{aligned}
G^*(s) &= \int_0^\infty \int_0^\infty \frac{x z^{s-1}}{1 + xz} \, dF(x) \, dz \\
&= \int_0^\infty \int_0^\infty \frac{z^{s-1}}{1+z} x^{1-s} \, dz \, dF(x) = \frac{\pi}{\sin \pi s} \int_0^\infty x^{1-s} \, dF(x).
\end{aligned}$$

This holds for $0 < \Re s < 1$. Due to the analytic continuation, this gives

Proposition 5. *For $s \in \mathbb{C} \setminus \mathbb{Z}$, we have an identity $G^*(s) = \zeta_{\mathcal{M}}(s - 1)\Gamma(s) \cdot \pi/(\sin \pi s)$.*

Therefore, $G^*(s)$ is a meromorphic function, $G^*(s + 1) = -G^*(-s + 1)$, and $\mathrm{res}_{s=L} G^*(s) = (-1)^L M_{L-1}$. This is the general property of the Mellin transform, since formally $G(z+1) = \sum_{L=0}^\infty M_L z^{L-1}$. Thus, $G(z+1) \sim \sum_{L=0}^M M_L z^{L-1}$ in the left neighborhood of $z = 0$.

10. Concluding remarks

Dyadic period functions in $\mathbb{H}$. As noted in [Alkauskas $\geq$ 2009], one encounters the surprising fact that in the upper half plane $\mathbb{H}$, Equation (12) is also satisfied by $(i/2\pi) G_1(z)$, where $G_1(z)$ stands for the Eisenstein series [Serre 1973]. Let $f_0(z) = G(z) - (i/2\pi) G_1(z)$, where $G(z)$ is the function in Theorem 1. Then for

$z \in \mathbb{H}$, $f_0(z)$ satisfies the homogeneous form of the three-term functional Equation (12); moreover, $f_0(z)$ is bounded, when $\Im z \to \infty$. Thus, if $f(z) = f_0(z)$,

$$-\frac{1}{(1-z)^2} f\left(\frac{1}{1-z}\right) + 2f(z+1) = f(z).$$

Therefore, denote by DPF^0 the $\mathbb{C}$-linear vector space of solutions of this three-term functional equation, which are holomorphic in $\mathbb{H}$ and are bounded at infinity, and call it *the space of dyadic period functions in the upper half-plane*. Consequently, this space is at least one-dimensional. If we abandon the growth condition, then the corresponding space DPF is infinite-dimensional. This is already true for periodic solutions. Indeed, if $f(z)$ is a periodic solution, then

$$f(z) = (1/z^2) f(-1/z).$$

Let $P(z) \in \mathbb{C}[z]$, and suppose that $j(z)$ stands, as usually, for the j-invariant. Then any modular function of the form $j'(z) P(j(z))$ satisfies this equation. Additionally, there are nonperiodic solutions, given by $f_0(z) P(j(z))$. Therefore, $G(z)$ surprisingly enters the profound domain of classical modular forms and functions for $\mathrm{PSL}_2(\mathbb{Z})$. Hence, it is greatly desirable to give the full description and structure of spaces DPF^0 and DPF.

Where should the true arithmetic zeta function come from? Here we present some remarks, concerning the zeta function $\zeta_\mathcal{M}(s)$. This object is natural for the question mark function — its Dirichlet coefficients are the Fourier coefficients of $F(x)$, and its special values at integers are proportional to the moments M_L. Moreover, its relation to $G(z)$, $\mathfrak{m}(t)$ and $F(x)$ is the same as that of the L-series of Maass wave forms to analogous objects [Zagier 2001]. Nevertheless, one expects a richer arithmetic object associated with the Calkin–Wilf tree, since the latter consists or rational numbers, and therefore can be canonically embedded into the group of idèles $\mathbb{A}_\mathbb{Q}$. The p-adic distribution of rationals in the n-th generation of Calkin–Wilf tree was investigated in [Alkauskas $\geq$ 2009]. Surprisingly, the Eisenstein series $G_1(z)$ yet again manifests itself, as in case of $\mathbb{R}$ (see previous subsection). Nevertheless, there is no direct way of normalizing moments of the n-th generation in order for them to converge in the p-adic norm. There is an exception. As can easily be seen,

$$\sum_{a_0+a_1+\cdots+a_s=n} [a_0, a_1, .., a_s] = 3 \cdot 2^{n-2} - 1/2,$$

and thus we have a convergence only in the 2-adic topology, namely to the value $-1/2$. The investigation of p-adic values of moments is relevant for the following reason. Let us apply $F(x)$ to each rational number in the Calkin–Wilf tree. What

we obtain is the following:
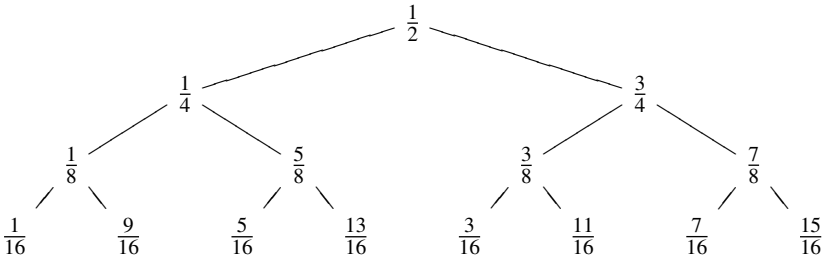
Using Equation (2), we deduce that this tree starts from the root $1/2$, and then inductively each rational r produces two offsprings: $r/2$ and $r/2 + 1/2$. One is therefore led to the following.

Task. Produce a natural algorithm, which takes into account p-adic and real properties of the above tree, and generates Riemann zeta function $\zeta(s)$.

We emphasize that the choice of $\zeta(s)$ is not accidental. In fact, the $\mathbb{R}$-distribution of the above tree is a uniform one with support $[0, 1]$. Further, there is a natural algorithm to produce a *characteristic function of ring of integers of $\mathbb{R}$* (that is, $e^{-\pi x^2}$) from the uniform distribution via the central limit theorem through the expression

$$\int_{\mathbb{R}} f(x) e^{-\pi x^2} \, dx = \lim_{N \to \infty} \frac{1}{2^N} \int_{-1}^{1} dx_1 \dots \int_{-1}^{1} dx_N \, f\left(\frac{x_1 + \dots + x_N}{\sqrt{\frac{2}{3}\pi N}} \right).$$

(For clarity, here we take the uniform distribution in the interval $[-1, 1]$). This formula and this explanation and treatment of $e^{-\pi x^2}$ as a *characteristic function of the ring of integer of $\mathbb{R}$* is borrowed from [Haran 2001, page 7]. Further, the operator which is invariant under uniform measure has the form

$$[\mathcal{U}f](x) = \frac{1}{2} f\left(\frac{x}{2} \right) + \frac{1}{2} f\left(\frac{x}{2} + \frac{1}{2} \right).$$

Indeed, for every $f \in C[0, 1]$, one has $\int_0^1 [\mathcal{U}f](x) \, dx = \int_0^1 f(x) \, dx$. The spectral analysis of $\mathcal{U}$ shows that its eigenvalues are 2^{-n}, $n \geq 0$, with corresponding eigenfunctions being Bernoulli polynomials $B_n(x)$ [Flajolet and Vallée 1998]. These, as is well known from the time of Euler, are intricately related with $\zeta(s)$. Moreover, the partial moments of the above tree can be defined as $\sum_{i=1}^{2^N} \left((2i - 1)/2^N \right)^L$. These values are also expressed in term of Bernoulli polynomials. As we know, there are famous Kummer congruences among Bernoulli numbers, which later led to the introduction of the p-adic zeta function $\zeta_p(s)$. Thus, the real distribution of the above tree and its spectral decomposition is deeply related to the p-adic properties. This justifies the choice in the task of $\zeta(s)$. Therefore, returning to the

Calkin–Wilf tree, one expects that moments can be *p*-adically interpolated, and some natural arithmetic zeta function can be introduced, as a *preimage* of $\zeta(s)$ under the map F.

Acknowledgement

References

[Alkauskas 2008] G. Alkauskas, "An asymptotic formula for the moments of Minkowski question mark function in the interval [0, 1]", *Lithuanian Math. J.* **48**:4 (2008), 357–367.

[Alkauskas ≥ 2009] G. Alkauskas, "The moments of Minkowski question mark function: the dyadic period function", preprint. submitted.

[Babenko 1978] K. I. Babenko, "On a problem of Gauss", *Dokl. Akad. Nauk SSSR* **238**:5 (1978), 1021–1024. MR 57 #12436 Zbl 0389.10036

[Beaver and Garrity 2004] O. R. Beaver and T. Garrity, "A two-dimensional Minkowski ?(x) function", *J. Number Theory* **107**:1 (2004), 105–134. MR 2005g:11125 Zbl 1064.11051

[Bonanno et al. 2008] C. Bonanno, S. Graffi, and S. Isola, "Spectral analysis of transfer operators associated to Farey fractions", *Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl.* **19**:1 (2008), 1–23. MR 2383559 Zbl 1142.37021

[Calkin and Wilf 2000] N. Calkin and H. S. Wilf, "Recounting the rationals", *Amer. Math. Monthly* **107**:4 (2000), 360–363. MR 2001d:11024 Zbl 0983.11009

[Contucci and Knauf 1997] P. Contucci and A. Knauf, "The phase transition of the number-theoretical spin chain", *Forum Math.* **9**:4 (1997), 547–567. MR 98j:82011 Zbl 0896.11033

[Cvitanović et al. 1998] P. Cvitanović, K. Hansen, J. Rolf, and G. Vattay, "Beyond the periodic orbit theory", *Nonlinearity* **11**:5 (1998), 1209–1232. MR 99g:58102

[Denjoy 1938] A. Denjoy, "Sur une fonction réelle de Minkowski", *J. Math. Pures Appl.* **17** (1938), 105–151.

[Denjoy 1956a] A. Denjoy, "La fonction minkowskienne complexe uniformisée détermine les intervalles de validité des transformations de la fonction reélle", *C. R. Acad. Sci. Paris* **242** (1956), 1924–1930. MR 19,401i Zbl 0070.28903

[Denjoy 1956b] A. Denjoy, "La fonction minkowskienne complexe uniformisée éclaire la genèse des fractions continues canoniques réelles", *C. R. Acad. Sci. Paris* **242** (1956), 1817–1823. MR 19,401h Zbl 0070.28902

[Denjoy 1956c] A. Denjoy, "Propriétés différentielles de la fonction minkowskienne réelle. Statistique des fractions continues", *C. R. Acad. Sci. Paris* **242** (1956), 2075–2079. MR 19,402a Zbl 0070.29001

[Dilcher and Stolarsky 2007] K. Dilcher and K. B. Stolarsky, "A polynomial analogue to the Stern sequence", *Int. J. Number Theory* **3**:1 (2007), 85–103. MR 2008d:11023 Zbl 1117.11017

[Dushistova and Moshchevitin ≥ 2009] A. Dushistova and N. G. Moshchevitin, "On the derivative of the Minkowski question mark function ?(x)", preprint.

[Esposti et al. ≥ 2009] M. D. Esposti, S. Isola, and A. Knauf, "Generalized Farey trees, transfer operators and phase transitions", preprint.

[Finch 2003] S. R. Finch, *Mathematical constants*, vol. 94, Encyclopedia of Mathematics and its Applications, Cambridge University Press, Cambridge, 2003. MR 2004i:00001 Zbl 1054.00001

[Flajolet and Vallée 1998] P. Flajolet and B. Vallée, "Continued fraction algorithms, functional operators, and structure constants", *Theoret. Comput. Sci.* **194**:1-2 (1998), 1–34. MR 98j:11061 Zbl 0981.11044

[Girgensohn 1996] R. Girgensohn, "Constructing singular functions via Farey fractions", *J. Math. Anal. Appl.* **203**:1 (1996), 127–141. MR 97f:26006 Zbl 0866.26003

[Grabner et al. 2002] P. J. Grabner, P. Kirschenhofer, and R. F. Tichy, "Combinatorial and arithmetical properties of linear numeration systems", *Combinatorica* **22**:2 (2002), 245–267. Special issue: Paul Erdős and his mathematics. MR 2003f:11113 Zbl 1012.11070

[Haran 2001] M. J. S. Haran, *The mysteries of the real prime*, vol. 25, London Mathematical Society Monographs. New Series, The Clarendon Press Oxford University Press, New York, 2001. MR 2003b:11085 Zbl 1014.11001

[Isola 2002] S. Isola, "On the spectrum of Farey and Gauss maps", *Nonlinearity* **15**:5 (2002), 1521–1539. MR 2003h:37027 Zbl 1018.37019

[Kesseböhmer and Stratmann 2007] M. Kesseböhmer and B. O. Stratmann, "A multifractal analysis for Stern-Brocot intervals, continued fractions and Diophantine growth rates", *J. Reine Angew. Math.* **605** (2007), 133–163. MR 2008f:37018

[Kesseböhmer and Stratmann 2008] M. Kesseböhmer and B. O. Stratmann, "Fractal analysis for sets of non-differentiability of Minkowski's question mark function", *J. Number Theory* **128**:9 (2008), 2663–2686. MR 2444218

[Khinchin 1964] A. Y. Khinchin, *Continued fractions*, The University of Chicago Press, Chicago, Ill.-London, 1964. MR 28 #5037 Zbl 0117.28601

[Kinney 1960] J. R. Kinney, "Note on a singular function of Minkowski", *Proc. Amer. Math. Soc.* **11** (1960), 788–794. MR 24 #A194 Zbl 0109.28101

[Kolmogorov and Fomin 1989] A. N. Kolmogorov and S. V. Fomin, *Elementy teorii funktsii i funktsionalnogo analiza*, Sixth ed., "Nauka", Moscow, 1989. With a supplement, "Banach algebras", by V. M. Tikhomirov. MR 90k:46001

[Kontsevich and Zagier 2001] M. Kontsevich and D. Zagier, "Periods", pp. 771–808 in *Mathematics unlimited—2001 and beyond*, Springer, Berlin, 2001. MR 2002i:11002 Zbl 1039.11002

[Lagarias 1991] J. C. Lagarias, "The Farey shift and the Minkowski ?-function", unpublished manuscript, 1991.

[Lagarias and Tresser 1995] J. C. Lagarias and C. P. Tresser, "A walk along the branches of the extended Farey tree", *IBM J. Res. Develop.* **39**:3 (1995), 283–294. MR 2361372

[Lamberger 2006] M. Lamberger, "On a family of singular measures related to Minkowski's ?(x) function", *Indag. Math. (N.S.)* **17**:1 (2006), 45–63. MR 2008g:60116 Zbl 1124.11011

[Lavrentjev and Shabat 1987] M. A. Lavrentjev and B. V. Shabat, *Methods in the theory of functions of complex variable*, Nauka, Moscow, 1987.

[Lewis 1997] J. B. Lewis, "Spaces of holomorphic functions equivalent to the even Maass cusp forms", *Invent. Math.* **127**:2 (1997), 271–306. MR 98f:11036 Zbl 0922.11043

[Lewis and Zagier 2001] J. Lewis and D. Zagier, "Period functions for Maass wave forms. I", *Ann. of Math. (2)* **153**:1 (2001), 191–258. MR 2003d:11068 Zbl 1061.11021

[Manin and Marcolli 2002] Y. I. Manin and M. Marcolli, "Continued fractions, modular symbols, and noncommutative geometry", *Selecta Math. (N.S.)* **8**:3 (2002), 475–521. MR 2004a:11039 Zbl 1116.11033

[Moshchevitin and Vielhaber ≥ 2009] N. Moshchevitin and M. Vielhaber, "Moments for generalized Farey-Brocot partitions", preprint.

[Newman 2003] M. Newman, "Recounting the rationals, continued", *Amer. Math. Monthly* **110** (2003), 642–643.

[Okamoto and Wunsch 2007] H. Okamoto and M. Wunsch, "A geometric construction of continuous, strictly increasing singular functions", *Proc. Japan Acad. Ser. A Math. Sci.* **83**:7 (2007), 114–118. MR 2008k:26008

[Panti 2008] G. Panti, "Multidimensional continued fractions and a Minkowski function", *Monatsh. Math.* **154**:3 (2008), 247–264. MR 2413304

[Paradís et al. 2001] J. Paradís, P. Viader, and L. Bibiloni, "The derivative of Minkowski's $?(x)$ function", *J. Math. Anal. Appl.* **253**:1 (2001), 107–125. MR 2002c:11092 Zbl 0995.26005

[Ramharter 1987] G. Ramharter, "On Minkowski's singular function", *Proc. Amer. Math. Soc.* **99**:3 (1987), 596–597. MR 88c:11013

[Reese 1989] S. Reese, "Some Fourier-Stieltjes coefficients revisited", *Proc. Amer. Math. Soc.* **105**:2 (1989), 384–386. MR 89i:42020 Zbl 0682.42005

[Reznick ≥ 2009] B. Reznick, "Regularity property of the Stern enumeration of the rationals", preprint.

[Ryde 1922] F. Ryde, "Arithmetical continued fractions", *Lunds universitets arsskrift N. F. Avd. 2.* **22**:2 (1922), 1–182.

[Ryde 1983] F. Ryde, "On the relation between two Minkowski functions", *J. Number Theory* **17**:1 (1983), 47–51. MR 85b:11008 Zbl 0519.10007

[Salem 1943] R. Salem, "On some singular monotonic functions which are strictly increasing", *Trans. Amer. Math. Soc.* **53** (1943), 427–439. MR 4,217b Zbl 0060.13709

[Serre 1973] J.-P. Serre, *A course in arithmetic*, Springer-Verlag, New York, 1973. Translated from the French, Graduate Texts in Mathematics, No. 7. MR 49 #8956 Zbl 0256.12001

[Stern 1858] M. A. Stern, "Über eine zahlentheoretische Funktion", *J. Reine Angew. Math.* **55** (1858), 193–220.

[Steuding 2006] J. Steuding, 2006. personal communication.

[Tichy and Uitz 1995] R. F. Tichy and J. Uitz, "An extension of Minkowski's singular function", *Appl. Math. Lett.* **8**:5 (1995), 39–46. MR 96i:26005 Zbl 0871.26008

[Vepštas 2004] L. Vepštas, "The Minkowski question mark, $GL(2, \mathbb{Z})$ and the modular monoid", preprint, 2004, Available at http://www.linas.org/math/chap-minkowski.pdf.

[Viader et al. 1998] P. Viader, J. Paradís, and L. Bibiloni, "A new light on Minkowski's $?(x)$ function", *J. Number Theory* **73**:2 (1998), 212–227. MR 2000a:11104 Zbl 0928.11006

[Wirsing 1973/74] E. Wirsing, "On the theorem of Gauss-Kusmin-Lévy and a Frobenius-type theorem for function spaces", *Acta Arith.* **24** (1973/74), 507–528. Collection of articles dedicated to Carl Ludwig Siegel on the occasion of his seventy-fifth birthday, V. MR 49 #2637

[Wirsing 2006] E. Wirsing, "J. Steuding's Problem", unpublished manuscript, 2006.

[Zagier 2001] D. Zagier, "New points of view on the Selberg zeta function", unpublished manuscript, 2001.

giedrius.alkauskas@gmail.com   *Vilnius University, The Department of Mathematics and Informatics, Naugarduko 24, Vilnius, Lithuania*

The School of Mathematical Sciences, The University of Nottingham, University Park, Nottingham NG7 2RD, United Kingdom
http://alkauskas.ten.lt

The index of a vector field on an orbifold with boundary

Elliot Paquette and Christopher Seaton

(Communicated by Michael Dorff)

A Poincaré–Hopf theorem in the spirit of Pugh is proven for compact orbifolds with boundary. The theorem relates the index sum of a smooth vector field in generic contact with the boundary orbifold to the Euler–Satake characteristic of the orbifold and a boundary term. The boundary term is expressed as a sum of Euler characteristics of tangency and exit-region orbifolds. As a corollary, we express the index sum of the vector field induced on the inertia orbifold to the Euler characteristics of the associated underlying topological spaces.

1. Introduction

The Poincaré–Hopf Theorem states that if M is a smooth, compact n-manifold and X is a vector field on M that points outwards everywhere on ∂M, then $\mathfrak{Ind}(X)$, the index of X, is equal to the Euler characteristic $\chi(M)$ of M. Pugh [1968] gave a generalization of this theorem for such manifolds where the vector field X on M has generic contact with ∂M. This means that the subset Γ^1 of ∂M on which X is tangent to ∂M is a codimension-1 submanifold of ∂M, the subset Γ^2 of Γ^1 on which X is tangent to Γ^1 is a codimension-1 submanifold of Γ^1, etc. This generalization bears the elegance of associating the index sum with a sum of Euler characteristics only. Here we show that in the case of a compact orbifold with boundary and a smooth vector field in generic contact with the boundary, Pugh's result extends naturally. A proper introduction to orbifolds and the precise definition we use are available as an appendix in [Chen and Ruan 2002]. Note that this definition of an orbifold requires group actions to have fixed-point sets of codimension at least 2 as opposed to other definitions which do not (see, for example, [Thurston 1978]); we make this requirement as well. By *smooth*, we always mean $\mathscr{C}^\infty$.

The main result we prove is as follows.

Theorem 1.1. *Let Q be an n-dimensional smooth*, *compact orbifold with boundary. Let Y be a smooth vector field on Q that is in* generic contact *with ∂Q, and then*

$$\mathfrak{Ind}^{\mathrm{orb}}(Y; Q) = \chi_{\mathrm{orb}}(Q, \partial Q) + \sum_{i=1}^{n} \chi_{\mathrm{orb}}(R_-^i, \Gamma^i). \tag{1-1}$$

The expressions $\mathfrak{Ind}^{\mathrm{orb}}$ and χ_{orb} are orbifold analogues of the manifold notions of the topological index of a vector field and the Euler characteristic, respectively. The definitions of both of these, along with R_-^i, Γ^i, and generic contact, are reviewed in Section 2.

In this paper, we follow a procedure resembling Pugh's original technique, and we show that many of the same techniques applicable to manifolds can be applied to orbifolds as well. In Section 2, we explain our notation and review the result of Satake which relates the orbifold index to the Euler–Satake characteristic for closed orbifolds. We give the definition of each of these terms. In Section 3, we show that a neighborhood of the boundary of an orbifold may be decomposed as a product $\partial Q \times [0, \epsilon)$. We then construct the double of Q and charts near the boundary respecting this product structure. This generalizes well-known results and constructions for manifolds with boundary. Section 4 provides elementary results relating the topological index of an orbifold vector field to an orbifold Morse Index. The orbifold Morse Index is defined in terms of the Morse Index on a manifold in a manner analogous to Satake's definition of the topological vector field index. These results generalize corresponding results for manifolds. In Section 5, we use the above constructions to show that a smooth vector field on Q may be perturbed near the boundary to form a smooth vector field on the double whose index can be computed in terms of the data given by the original vector field. We use this to prove Theorem 1.1. We also prove Corollary 5.2, which gives a similar formula where the left side is the orbifold index of the induced vector field on the inertia orbifold and on the right side, the Euler–Satake characteristics are replaced with the Euler characteristics of the underlying topological spaces.

Another generalization of the Poincaré–Hopf Theorem to orbifolds with boundary is explored in [Seaton 2008] and follows as a corollary to Satake's Gauss–Bonnet Theorem for orbifolds with boundary [Satake 1957]. In each of these cases, the boundary term is expressed by evaluation of an auxiliary differential form representing a global topological invariant of the boundary pulled back via the vector field. The generalization given in our paper expresses the boundary term in terms of Euler–Satake characteristics of suborbifolds determined by the vector field.

2. Preliminaries and definitions

Satake proved a Poincaré–Hopf Theorem for closed orbifolds; however, he worked with a slightly different definition of orbifold, the so-called *V*-manifold [Satake 1956; 1957]. A *V*-manifold corresponds to an *effective* or *reduced* (codimension-2) orbifold, an orbifold such that the group in each chart acts effectively [Chen and Ruan 2002]. That is, the only group element that acts trivially is the identity element. We adopt the language of his result and use it here.

Theorem 2.1 (Satake's Poincaré–Hopf Theorem for Closed Orbifolds). *Let Q be an effective, closed orbifold, and let X be a vector field on Q that has isolated zeros. Then*

$$\mathfrak{Ind}^{\mathrm{orb}}(X; Q) = \chi_{\mathrm{orb}}(Q).$$

Note that the requirement that Q is effective is unnecessary; as mentioned in [Chen and Ruan 2002], an ineffective orbifold can be replaced with an effective orbifold Q_{red}, and the differential geometry of the tangent bundle (or any other *good* orbifold vector bundle) is unchanged.

The *orbifold index* $\mathfrak{Ind}^{\mathrm{orb}}(X; p)$ at a zero p of the vector field X is defined in terms of the topological index of a vector field on a manifold. Let a neighborhood of p be uniformized by the chart $\{V, G, \pi\}$ and choose $x \in V$ with $\pi(x) = p$. Let $G_x \leq G$ denote the isotropy group of x. Then $\pi^* X$ is a G-invariant vector field on V with a zero at x. The orbifold index at p is defined by

$$\mathfrak{Ind}^{\mathrm{orb}}(X; p) = \frac{1}{|G_x|} \mathfrak{Ind}(\pi^* X; x),$$

where $\mathfrak{Ind}(\pi^* X; x)$ is the usual topological index of the vector field $\pi^* X$ on the manifold V at x [Guillemin and Pollack 1974; Milnor 1965]. Note that this definition does not depend on the chart, nor on the choice of x. We use the notation

$$\mathfrak{Ind}^{\mathrm{orb}}(X; Q) = \sum_{p \in Q, X(p)=0} \mathfrak{Ind}^{\mathrm{orb}}(X; p).$$

The *Euler–Satake characteristic* $\chi_{\mathrm{orb}}(Q)$ is most easily defined in terms of an appropriate simplicial decomposition of Q. In particular, let $\mathcal{T}$ be a simplicial decomposition of Q such that the isomorphism class of the isotropy group is constant on the interior of each simplex (such a simplicial decomposition always exists; see [Moerdijk and Pronk 1999]). For each simplex $\sigma \in \mathcal{T}$, the (isomorphism class of the) isotropy group on the interior of σ is denoted G_σ. The Euler–Satake characteristic of Q is then defined by

$$\chi_{\mathrm{orb}}(Q) = \sum_{\sigma \in \mathcal{T}} (-1)^{\dim \sigma} \frac{1}{|G_\sigma|}.$$

This coincides with Satake's *Euler characteristic of Q as a V-manifold*. Note that it follows from this definition that if $Q = Q_1 \cup Q_2$ for orbifolds Q_1 and Q_2 with $Q_1 \cap Q_2$ a suborbifold, then

$$\chi_{\text{orb}}(Q) = \chi_{\text{orb}}(Q_1) + \chi_{\text{orb}}(Q_2) - \chi_{\text{orb}}(Q_1 \cap Q_2). \tag{2-1}$$

In the case that Q has boundary, $\chi_{\text{orb}}(Q)$ is defined in the same way. We let

$$\chi_{\text{orb}}(Q, \partial Q) = \chi_{\text{orb}}(Q) - \chi_{\text{orb}}(\partial Q).$$

This coincides with Satake's *inner Euler characteristic of Q as a V-manifold with boundaries*. The reader is warned that there are many different Euler characteristics defined for orbifolds; both the topological index of a vector field and the Euler–Satake characteristic used here are generally rational numbers.

Vector fields in *generic contact* with the boundary have orbifold exit regions, which we now describe. Let Q be a compact n-dimensional orbifold with boundary and X a smooth vector field on Q. In Lemma 3.1, we show that, as with the case of manifolds, there is a neighborhood of ∂Q in Q diffeomorphic to $\partial Q \times [0, \epsilon)$. Given a metric, the tangent bundle of Q on the boundary decomposes with respect to this product so that there is a well-defined normal direction to the boundary. Let R_-^1 be the closure of the subset of ∂Q where X points out of Q. Analogously, let R_+^1 be the closure of the subset of ∂Q where X points into Q. We require that R_-^1 and R_+^1 are $(n-1)$-dimensional orbifolds with boundary. The subset of ∂Q where the vector field is tangent to ∂Q is denoted Γ^1; we require that Γ^1 be a suborbifold of ∂Q of codimension 1. Note that, by the continuity of X, the component of the vector field pointing outward must approach zero near the boundary of R_-^1 and R_+^1. Hence $\Gamma^1 = \partial R_-^1 = \partial R_+^1$.

The vector field X is tangent to Γ^1, and so it may be considered a vector field on the orbifold Γ^1. We again require this vector field to have orbifold exit regions. Call R_-^2 the closure of the subset of Γ^1 where the vector field points out of R_-^1, and R_+^2 the closure of the subset where it points into R_-^1. The subset of Γ^1 where the vector field is tangent to Γ^1 is denoted Γ^2, and is required to be a codimension-1 suborbifold of Γ^1.

In the same way, we define Γ^i, R_-^i, R_+^i, requiring that these sets form a chain of closed suborbifolds $\{\Gamma^i\}_{i=1}^n$ and compact orbifolds with boundary $\{R_-^i\}_{i=1}^n$. We require that $\dim R_-^i = \dim R_+^i = n-i$ and $\dim \Gamma^i = n-i-1$. Since each successive Γ^i has strictly smaller dimension, we eventually run out of space, and so both of these sequences terminate. The last entry in the sequence of Γ^i is Γ^n, which is necessarily the empty set.

3. Formation of the double orbifold

In the proof of Theorem 1.1, we pass from an orbifold with boundary to a closed orbifold in order to employ Theorem 2.1. In this section, we construct the double of an orbifold with boundary. In the process, we develop charts near the boundary of a specific form which are required in the sequel. The construction of the double is similar to the case of a manifold [Munkres 1963].

Let $\mathbf{B}_x(r)$ denote the ball of radius r about x in $\mathbb{R}^n$ where $\mathbb{R}^n$ has basis $\{e_i\}_{i=1}^n$. For convenience, $\mathbf{B}_0$ denotes the ball of radius 1 centered at the origin in $\mathbb{R}^n$. We let $\mathbb{R}_+^n = \{x_1, \ldots, x_n : x_n \geq 0\}$ where the x_i are the coordinates with respect to the basis $\{e_i\}$, $\mathbf{B}_x^+(r) = \mathbf{B}_x(r) \cap \mathbb{R}_+^n$, and $\mathbf{B}_0^+ = \mathbf{B}_0 \cap \mathbb{R}_+^n$. Also, $\mathbf{B}_0^k$ denotes the ball of radius 1 about the origin in $\mathbb{R}^k$.

Let Q be a compact orbifold with boundary. For each point $p \in Q$, we choose an orbifold chart $\{V_p, G_p, \pi_p\}$ where V_p is $\mathbf{B}_0$ or $\mathbf{B}_0^+$ and $\pi_p(0) = p$. Let U_p denote $\pi_p(V_p) \subseteq Q$ for each p, and then the U_p form an open cover of Q. Choose a finite subcover of the U_p, and on each V_p corresponding to a U_p in the subcover, we put the standard Riemannian structure on V_p so that the $\{\partial/\partial x_i\}$ form an orthonormal basis. Endow Q with a Riemannian structure by patching these Riemannian metrics together using a partition of unity subordinate to the finite subcover of Q chosen above.

Now, let $p \in Q$, and then there is a geodesic neighborhood U_p about p uniformized by $\{V_p, G_p, \pi_p\}$ where $V_p = \mathbf{B}_0(r)$ or $\mathbf{B}_0^+(r)$ for some $r > 0$, and G_p acts as a subgroup of $O(n)$ [Chen and Ruan 2002]. Identifying V_p with a subset of $T_0 V_p$ via the exponential map, we can assume as above that $\{e_i\}$ forms an orthonormal basis with respect to which coordinates are denoted $\{x_i\}$. In the case with boundary, $\mathbf{B}_0^+(r)$ corresponds to points with $x_n \geq 0$. We call such a chart a *geodesic chart of radius r at p*. Note that in such charts, the action of $\gamma \in G_p$ on V_p and the action of $d\gamma = D(\gamma)_0$ on a neighborhood of 0 in $T_0 V_p$ (or in half-space in the case with boundary) are identified via the exponential map.

Lemma 3.1. *At every point p in ∂Q, there is a geodesic chart at p of the form $\{V_p, G_p, \pi_p\}$ where G_p fixes e_n. On the boundary, the tangent space $TQ|_{\partial Q}$ is decomposed orthogonally into $(T\partial Q) \oplus \nu$ where ν is a trivial 1-bundle on which each group acts trivially.*

Proof. Let $p \in \partial Q$, and let a neighborhood of p be uniformized by the geodesic chart $\{V_p, G_p, \pi_p\}$ so that $V_p = \mathbf{B}_0^+(r)$. Let $\langle \cdot, \cdot \rangle_0$ denote the inner product on $T_0 V_p$. Let T_0^+ correspond to the half-space in $T_0 V_p$ corresponding to vectors with non-negative $(\partial/\partial x_n)$-component. The exponential map identifies an open ball about $0 \in T_0^+$ with V_p.

Suppose γ is an arbitrary element of G_p so that $d\gamma$ acts on $T_0 V_p$. Any $v \in T_0^+$ satisfies

$$\left\langle v, \frac{\partial}{\partial x_n} \Big|_0 \right\rangle \geq 0.$$

Furthermore, $(d\gamma)v \in T_0^+$, so

$$\left\langle (d\gamma)v, \frac{\partial}{\partial x_n} \Big|_0 \right\rangle \geq 0 \quad \text{or equivalently} \quad \left\langle v, d\gamma^{-1} \frac{\partial}{\partial x_n} \Big|_0 \right\rangle \geq 0,$$

for all $v \in T_0^+$.

We claim that G_p fixes $\partial/(\partial x_n)$. Pick $j \neq n$; since $\frac{\partial}{\partial x_j} \in T_0^+$,

$$\left\langle \frac{\partial}{\partial x_j}, d\gamma^{-1} \frac{\partial}{\partial x_n} \Big|_0 \right\rangle \geq 0.$$

However, $-\frac{\partial}{\partial x_j}$ is also a vector in T_0^+, and so

$$\left\langle -\frac{\partial}{\partial x_j}, d\gamma^{-1} \frac{\partial}{\partial x_n} \Big|_0 \right\rangle \geq 0.$$

By the linearity of the inner product, this is only possible if

$$\left\langle \frac{\partial}{\partial x_j}, d\gamma^{-1} \frac{\partial}{\partial x_n} \Big|_0 \right\rangle = 0.$$

Furthermore, since $j \neq n$ was arbitrary, this implies that $d\gamma^{-1}(\partial/\partial x_n)$ has no component in the direction of any $(\partial/\partial x_j)$, $j \neq n$. Since $d\gamma^{-1}$ is an isometry,

$$d\gamma^{-1} \frac{\partial}{\partial x_n} = \pm \frac{\partial}{\partial x_n},$$

but because $d\gamma^{-1} T_0^+ = T_0^+$, it must be the case that

$$d\gamma^{-1} \frac{\partial}{\partial x_n} = \frac{\partial}{\partial x_n}.$$

As $\gamma \in G_p$ was arbitrary, this implies G_p fixes $\partial/(\partial x_n)$.

Now, for each $p \in \partial Q$, pick a geodesic chart $\{V_p, G_p, \pi_p\}$ at p and let N_p denote the constant vector field $\partial/(\partial x_n)$ on V_p. Recall from [Satake 1957] that $\tilde{T}_0 V_p$ denotes the dG_p-invariant tangent space of $T_0 V_p$ on which the differential of π_p is invertible. If $q \in \pi_p(V_p) \subset Q$ with geodesic chart $\{V_q, G_q, \pi_q\}$ at q, then the fact that $D(\pi_q)_p^{-1} \circ D(\pi_p)_0 : \tilde{T}_0 V_p \to \tilde{T}_0 V_q$ maps $\tilde{T}_0 \partial V_p$ to $\tilde{T}_0 \partial V_q$ and preserves the metric ensures that the value of $N_q(0)$ coincides with that of $D(\pi_q)_p^{-1} \circ D(\pi_p)_0 [N_p(0)]$ up to a sign. The sign is characterized by the property that for any curve $c : (-1, 1) \to V_p$ with derivative $c'(t) = N_p$, there is an $\epsilon > 0$ such that $c(t)$ is in the interior of V_p for $t \in (0, \epsilon)$; a curve in V_q with derivative

$D(\pi_q)_p^{-1} \circ D(\pi_p)_0[N_p(0)]$ has the same property. With this, we see that the N_p patch together to form a nonvanishing section of $TQ|_{\partial Q}$ that is orthogonal to $T\partial Q$ at every point; hence, it defines a trivial subbundle ν orthogonal to $T\partial Q$. Clearly, $TQ = (T\partial Q) \oplus \nu$. $\square$

Let Q' be an identical copy of Q. In order to form a closed orbifold from the two, the boundaries of these two orbifolds are identified via

$$\partial Q \ni x \longleftrightarrow x' \in \partial Q'. \tag{3-1}$$

The resulting space inherits the structure of a smooth orbifold from Q as is demonstrated below.

Note by Lemma 3.1 that for each point $p \in \partial Q$, a geodesic chart $\{V_p, G_p, \pi_p\}$ can be restricted to a chart $\{C_p^+, G_p, \phi_p\}$ where $C_p^+ = \mathbf{B}_0^{n-1}(r/2) \times [0, \epsilon_p)$, ϕ_p is the restriction of π_p to C_p^+, and $\phi_p(\mathbf{B}_0^{n-1} \times \{0\}) = \partial\phi_p(C_p^+)$. We refer to such a chart as a *boundary product chart* for Q.

It follows, in particular, that there is a neighborhood of ∂Q in Q that is diffeomorphic to $\partial Q \times [0, \epsilon]$ for some $\epsilon > 0$ and that the metric respects the product structure. This can be shown by forming a cover of ∂Q of sets uniformized by charts of the form $\{C^+, G_p, \phi_p\}$, choosing a finite subcover, and setting $\epsilon = \min\{\epsilon_p/2\}$.

Lemma 3.2. *The glued set $\hat{Q}$, that is, the set of equivalence classes under the identification made by Equation (3-1), may be made into a smooth orbifold containing diffeomorphic copies of both Q and Q' such that $Q \cap Q' = \partial Q = \partial Q'$.*

Proof. For each point $p \in \partial Q$, form a boundary product chart $\{C_p^+, G_p, \phi_p\}$. Then glue each chart of the boundary of Q to its corresponding chart of Q' in the following way. Let $\alpha : \mathbb{R}^n \to \mathbb{R}^n$ be the reflection that sends $e_n \mapsto -e_n$ and fixes all other coordinates. A point p in the boundary is uniformized by two corresponding boundary product charts on either side of ∂Q, $\{C_p^+, G_p, \phi_p\}$ and $\{C_p^{+'}, G_p', \phi_p'\}$. From these two charts, a new chart $\{C_p, G_p, \psi_p\}$ for a neighborhood of p in $\hat{Q}$ is constructed where $C_p = \mathbf{B}_0^{n-1}(r/2) \times (-\epsilon_p, \epsilon_p)$, and

$$\psi_p(x) = \begin{cases} \phi_p(x), & x_n \geq 0, \\ \phi_p' \circ \alpha(x), & x_n < 0. \end{cases}$$

These charts cover a neighborhood of $\partial Q = \partial Q'$ in $\hat{Q}$. By taking a geodesic chart at each point on the interiors of Q and Q' together with these new charts, the entire set $\hat{Q}$ is covered. Injections of charts at points in the interior of Q or Q' into charts of the form $\{C_p^+, G_p, \phi_p\}$ induce injections into $\{C_p, G_p, \psi_p\}$. Hence, $\hat{Q}$ is given the structure of a smooth orbifold with the desired properties. $\square$

Again, it follows that a neighborhood of $\partial Q \subset \hat{Q}$ admits a tubular neighborhood diffeomorphic to $\partial Q \times [-\epsilon, \epsilon]$ such that the metric respects this product structure.

4. The Morse Index of a vector field on an orbifold

The definition of the Morse Index and its relation to the topological index of a
vector field extend readily to orbifolds.

Let Q be a compact orbifold with or without boundary, and let X be a vector field
on Q that does not vanish on the boundary. Suppose $X(p) = 0$ for $p \in Q$. We say
that p is a *nondegenerate* zero of X if there is a chart $\{V, G, \pi\}$ for a neighborhood
U_p of p and an $x \in V$ with $\pi(x) = p$ such that π^*X has a nondegenerate zero at x;
that is, $D(\pi^*X)_x$ has trivial kernel. As in the manifold case, nondegenerate zeros
are isolated in charts and hence isolated on Q. The Morse Index $\lambda(\pi^*X; x)$ of
π^*X at x is defined to be the number of negative eigenvalues of $D(\pi^*X)_x$ [Milnor
1963]. Since the Morse Index is a diffeomorphism invariant, this index does not
depend on the choice of chart nor on the choice of x. Since the isomorphism-
class of the isotropy group does not depend on the choice of x, the expression
$|G_p|$ is well-defined. Hence, for simplicity, we may restrict to charts of the form
$\{V_p, G_p, \pi_p\}$ where $\pi_p(0) = p$ and G_p acts linearly. We define the *orbifold Morse
Index* of X at p to be

$$\lambda^{\mathrm{orb}}(X; p) = \frac{1}{|G_p|} \lambda(\pi_p^*X; 0).$$

Note that this index differs from that recently defined in [Hepworth 2007]; however,
it is sufficient for our purposes. We have

$$\begin{aligned}
\mathfrak{Ind}^{\mathrm{orb}}(X; p) &= \tfrac{1}{|G_p|} \mathfrak{Ind}(\pi_p^*X; 0) \\
&= \tfrac{1}{|G_p|} (-1)^{\lambda(\pi_p^*X; 0)}.
\end{aligned}$$

Suppose X has only nondegenerate zeros on Q. For each $\lambda \in \{0, 1, \ldots, n\}$, we
let $\{p_i : i = 1, \ldots, k_\lambda\}$ denote the points in Q at which the pullback of X in a chart
has Morse Index λ. Then we let

$$C_\lambda = \sum_{i=1}^{k_\lambda} \frac{1}{|G_{p_i}|}$$

count these points, where the orbifold-contribution of each zero p_i is $1/|G_{p_i}|$. Note
that as nondegenerate zeros are isolated, there is a finite number on Q.

As in the manifold case, we define

$$\Sigma^{\mathrm{orb}}(X; Q) = \sum_{\lambda=0}^{n} (-1)^\lambda C_\lambda,$$

and we have

$$
\begin{aligned}
\Sigma^{\mathrm{orb}}(X; Q) &= \sum_{\lambda=0}^{n}(-1)^{\lambda}\sum_{i=1}^{k_{\lambda}}\frac{1}{|G_{p_i}|} \\
&= \sum_{p \in Q, X(p)=0}\frac{1}{|G_p|}(-1)^{\lambda(\pi_p^*X;0)} \\
&= \sum_{p \in Q, X(p)=0}\mathfrak{Ind}^{\mathrm{orb}}(X; p) \\
&= \mathfrak{Ind}^{\mathrm{orb}}(X; Q).
\end{aligned}
$$

In the case that Q is closed, this quantity is equal to $\chi_{\mathrm{orb}}(Q)$ by Theorem 2.1.

We summarize these results as follows.

Proposition 4.1. *Let X be a smooth vector field on the compact orbifold Q that has nondegenerate zeros only, none of which occurring on ∂Q. Then*

$$
\Sigma^{\mathrm{orb}}(X; Q) = \mathfrak{Ind}^{\mathrm{orb}}(X; Q).
$$

If $\partial Q = \varnothing$, then

$$
\Sigma^{\mathrm{orb}}(X; Q) = \chi_{\mathrm{orb}}(Q).
$$

Remark 4.2. If Q is a compact orbifold (with or without boundary) and X a smooth vector field on Q that is nonzero on some compact subset Γ of the interior of Q, then X may be perturbed smoothly outside of a neighborhood of Γ so that it has only isolated, nondegenerate zeros. This is shown in [Waner and Wu 1986] for the case of a smooth global quotient M/G using local arguments, and so it extends readily to the case of a general orbifold by working in charts.

5. Proof of Theorem 1.1

Proof. Let Y be a vector field in generic contact with ∂Q that has isolated zeros on the interior of Q. Define $\hat{Y}$ on $\hat{Q}$ by letting $\hat{Y}$ coincide with Y on each copy of Q. Unfortunately, $\hat{Y}$ has conflicting definitions on ∂Q. However, as in the manifold case treated in [Pugh 1968], the vector field may be perturbed near the boundary to form a well-defined vector field using the product structure. We give an adaptation of Pugh's result to orbifolds.

Proposition 5.1. *Given a smooth vector field Y in generic contact with ∂Q with isolated zeros, none of which on ∂Q, there is a smooth vector field X on the double $\hat{Q}$ such that*

(i) *outside of a tubular neighborhood P_{ϵ} of ∂Q containing none of the zeros of Y, X coincides with Y on Q and Q';*

(ii) *$X|_{\partial Q}$ is tangent to ∂Q;*

(iii) *on Γ^1, X coincides with Y and in particular defines the same Γ^i, R_-^i, and R_+^i for $i > 1$;*

(iv) *the zeros of X are those of Y on the interior of Q and Q′ and a collection of isolated zeros on ∂Q which are nondegenerate as zeros of $X|_{\partial Q}$.*

Proof. As above, $\hat{Y}$ is defined everywhere on $\hat{Q}$ except on ∂Q. Let P_ϵ be a normal tubular ϵ-neighborhood of ∂Q in $\hat{Q}$ of the form $\partial Q \times [-\epsilon, \epsilon]$ which we parameterize as $\{(x, v) : x \in \partial Q, v \in [-\epsilon, \epsilon]\}$. We assume that P_ϵ is small enough so that it does not contain any of the zeros of $\hat{Y}$. On P_ϵ, decompose $\hat{Y}$ respecting the product structure of P_ϵ into

$$\hat{Y} = \hat{Y}_h + \hat{Y}_v.$$

These are the horizontal and vertical components of $\hat{Y}$, respectively. The horizontal component $\hat{Y}_h$ is well-defined, continuous, and smooth when restricted to the boundary. However, $\hat{Y}_v$ has conflicting definitions on the boundary, although they only differ by a sign. Note that the restriction of $\hat{Y}_h$ to ∂Q may not have isolated zeros. However, as Y does not have zeros on ∂Q and $\hat{Y}_h \equiv Y$ on Γ^1, none of the zeros of $\hat{Y}_h|_{\partial Q}$ occur on Γ^1.

Define Z_h to be a smooth vector field on ∂Q that coincides with $\hat{Y}_h$ on an open subset of ∂Q containing Γ^1 and has only nondegenerate zeros (see Remark 4.2). Let $f(x, v)$ be the parallel transport of $Z_h(x, 0)$ along the geodesic from $(x, 0)$ to (x, v), and then Z_h is a horizontal vector field on P_ϵ. For $s \in (0, \epsilon)$, let $\phi_s : \mathbb{R} \to [0, 1]$ be a smooth bump function which is one on $[-s/2, s/2]$ and zero outside of $[s, s]$.

Define the vector field X_s to be $\hat{Y}$ outside of P_ϵ and

$$X_s(x, v) = \phi_s(v)\big(f(x, v) + |v|\hat{Y}_v(x, v)\big) + (1 - \phi_s(v))\hat{Y}(x, v)$$

on P_ϵ. Note that X_s is smooth. By picking s sufficiently small, it may be ensured that the zeros of X are the zeros of $\hat{Y}$ and the zeros of $Z_h|_{\partial Q}$ only. We prove this as follows.

On points (x, v) where $x \in \Gamma^1$ and $|v| \le s$, the horizontal component of X is $\phi_s(v)f(x, v) + (1 - \phi_s(v))\hat{Y}_h(x, v)$. Note that $f(x, 0) = \tilde{Y}_h(x, 0)$ for $x \in \Gamma^1$ and $f(x, 0) \ne 0$ on Γ^1. Let $m > 0$ be the minimum value of $\|f(x, 0)\|$ on the compact set Γ^1, and then as $\Gamma^1 \times [-\epsilon, \epsilon]$ is compact and $\tilde{Y}_h(x, v)$ continuous, there is an s_0 such that

$$\|\hat{Y}_h(x, 0) - \hat{Y}_h(x, v)\| = \|f(x, 0) - \hat{Y}_h(x, v)\| < m/2$$

whenever $|v| < s_0$. Hence, for such v and for any $t \in [0, 1]$,

$$\begin{aligned}
\|tf(x, v) + (1 - t)\hat{Y}_h(x, v)\| &= \|\hat{Y}_h(x, v) + t[f(x, v) - \hat{Y}_h(x, v)]\| \\
&\ge \|\hat{Y}_h(x, v)\| - t\|f(x, v) - \hat{Y}_h(x, v)\| \\
&> m - \frac{tm}{2} \\
&\ge \frac{m}{2} > 0.
\end{aligned}$$

Therefore, the horizontal component is nonvanishing, implying that $X_s(v, h)$ does not vanish here.

Now let $\{x_i : i = 1, \ldots, k\}$ be the zeros of Z_h on ∂Q. Each x_i is contained in a ball $B_{\epsilon_i} \subset \partial Q$ whose closure does not intersect Γ^1. Hence, $\hat{Y}_v(x, 0) \neq 0$ on each B_{ϵ_i}. Therefore, for each i, there is an s_i such that $\hat{Y}_v(x, v) \neq 0$ on $B_{\epsilon_i} \times (-s_i, s_i)$. This implies that the vertical component of $X_s(x, v)$, and hence $X_s(x, v)$ itself, does not vanish on $B_{\epsilon_i} \times (-s_i, s_i)$ except where $v = 0$; i.e. on ∂Q.

Letting s be less than the minimum of $\{s_0, s_1, \ldots, s_k\}$, we see that X_s does not vanish on P_ϵ except on ∂Q, where it coincides with Z_h. Therefore, $X = X_s$ is the vector field which was to be constructed. $\qquad\square$

It follows that the index of the vector field X constructed in the proof of Proposition 5.1 is

$$\mathfrak{Ind}^{\mathrm{orb}}(X; \hat{Q}) = 2\mathfrak{Ind}^{\mathrm{orb}}(Y; Q) + \sum_{p \in \partial Q} \mathfrak{Ind}^{\mathrm{orb}}(X; p). \tag{5-1}$$

Let p be a zero of X on ∂Q, i.e. it is a zero of Z_h. We express the index of X at p in terms of the index of Z_h.

Because of Lemma 3.1, the isotropy group of p as an element of Q is the same as the isotropy group of p as an element of ∂Q, and so we may refer to G_p without ambiguity. For a neighborhood of p in Q small enough to contain no other zeros of X, choose a boundary product chart $\{C_p^+, G_p, \phi_p\}$. Then, as in Lemma 3.2, $\{C_p, G_p, \psi_p\}$ forms a chart about p in $\hat{Q}$. The product structure (y, w) of

$$C_p = \mathbf{B}_0^{n-1}(r/2) \times (-\epsilon_p, \epsilon_p)$$

coincides with that of P_ϵ near the boundary, so within the preimage of

$$\partial Q \times [-s/2, s/2]$$

by ψ_p, we have that

$$\psi_p^* X = \psi_p^* f + |w| \psi_p^* \hat{Y}_v.$$

Note that $\psi_p(0, 0) = p$, and then

$$D(\psi_p^* X)_{(0,0)} = \begin{pmatrix} D(\psi_p^* Z_h)_0 & \left(\frac{\partial \psi_p^* f}{\partial w}\right)_0 \\ D\left((|w|\psi_p^* \hat{Y}_v)|_{\partial C_p}\right)_0 & \left(\frac{\partial}{\partial w} |w| \psi_p^* \hat{Y}_v\right)_0 \end{pmatrix}$$

$$= \begin{pmatrix} D(\psi_p^* Z_h)_0 & 0 \\ 0 & \psi_p^* \hat{Y}_v(0, 0) \end{pmatrix}.$$

As $\psi_p^* \hat{Y}_v(0,0)$ is positive if $p \in R_+^1$ and negative if $p \in R_-^1$, we see that

$$\lambda\big(\psi_p^* X; (0,0)\big) = \begin{cases} \lambda\big(\psi_p^* X|_{\partial C_p}; 0\big), & p \in R_+^1, \\ \lambda\big(\psi_p^* X|_{\partial C_p}; 0\big) + 1, & p \in R_-^1. \end{cases}$$

Hence

$$\mathfrak{Ind}\big(\psi_p^* X; (0,0)\big) = \begin{cases} \mathfrak{Ind}\big(\psi_p^* Z_h|_{\partial C_p}; 0\big), & p \in R_+^1, \\ -\mathfrak{Ind}\big(\psi_p^* Z_h|_{\partial C_p}; 0\big), & p \in R_-^1. \end{cases}$$

Therefore, for $p \in R_+^1$,

$$\begin{aligned} \mathfrak{Ind}^{\mathrm{orb}}(X, p) &= \tfrac{1}{|G_p|}\mathfrak{Ind}\big(\psi_p^* X; 0\big) \\ &= \tfrac{1}{|G_p|}\mathfrak{Ind}\big(\psi_p^* Z_h|_{\partial C^+}; 0\big) \\ &= \mathfrak{Ind}^{\mathrm{orb}}\big(Z_h; p\big), \end{aligned}$$

and similarly

$$\mathfrak{Ind}^{\mathrm{orb}}\big(X; p\big) = -\mathfrak{Ind}^{\mathrm{orb}}\big(Z_h; p\big),$$

for $p \in R_-^1$.

With this, Equation (5-1) becomes

$$\mathfrak{Ind}^{\mathrm{orb}}(X; \hat{Q}) = 2\mathfrak{Ind}^{\mathrm{orb}}(Y; Q) + \mathfrak{Ind}^{\mathrm{orb}}(Z_h; R_+^1) - \mathfrak{Ind}^{\mathrm{orb}}(Z_h; R_-^1).$$

By Theorem 2.1 and Equation (2-1), $\mathfrak{Ind}^{\mathrm{orb}}(X; \hat{Q}) = 2\chi_{\mathrm{orb}}(Q) - \chi_{\mathrm{orb}}(\partial Q)$, with the result that

$$2\chi_{\mathrm{orb}}(Q) - \chi_{\mathrm{orb}}(\partial Q) = 2\mathfrak{Ind}^{\mathrm{orb}}(Y; Q) + \mathfrak{Ind}^{\mathrm{orb}}(Z_h; R_+^1) - \mathfrak{Ind}^{\mathrm{orb}}(Z_h; R_-^1).$$

Note that ∂Q is also a closed orbifold, so

$$\chi_{\mathrm{orb}}(\partial Q) = \mathfrak{Ind}^{\mathrm{orb}}(X; \partial Q) = \mathfrak{Ind}^{\mathrm{orb}}(X; R_+^1) - \mathfrak{Ind}^{\mathrm{orb}}(X; R_-^1).$$

Hence, restricting X to ∂Q,

$$\begin{aligned} \mathfrak{Ind}^{\mathrm{orb}}(Y; Q) &= \chi_{\mathrm{orb}}(Q) + 1/2\big(-\chi_{\mathrm{orb}}(\partial Q) + \mathfrak{Ind}^{\mathrm{orb}}(X; R_-^1) - \mathfrak{Ind}^{\mathrm{orb}}(X; R_+^1)\big) \\ &= \chi_{\mathrm{orb}}(Q) + 1/2\big[-\chi_{\mathrm{orb}}(\partial Q) + 2\mathfrak{Ind}^{\mathrm{orb}}(X; R_-^1) \\ &\qquad - \big(\mathfrak{Ind}^{\mathrm{orb}}(X; R_+^1) + \mathfrak{Ind}^{\mathrm{orb}}(X; R_-^1)\big)\big] \\ &= \chi_{\mathrm{orb}}(Q) + 1/2\big(-2\chi_{\mathrm{orb}}(\partial Q) + 2\mathfrak{Ind}^{\mathrm{orb}}(X; R_-^1)\big) \\ &= \chi_{\mathrm{orb}}(Q) - \chi_{\mathrm{orb}}(\partial Q) + \mathfrak{Ind}^{\mathrm{orb}}(X; R_-^1) \\ &= \chi_{\mathrm{orb}}(Q, \partial Q) + \mathfrak{Ind}^{\mathrm{orb}}(X; R_-^1). \end{aligned} \qquad (5\text{-}2)$$

Because X coincides with Y on Γ^1, it defines the same Γ^i that Y does. Since X is a smooth vector field defined on R_-^1 that does not vanish on $\partial R_-^1 = \Gamma^1$, we

may recursively apply this formula to higher and higher orders of R_-^i until R_-^i is empty, and there is no longer an index sum term. Hence,

$$\mathfrak{Ind}^{\mathrm{orb}}(X; R_-^1) = \sum_{i=1}^{n} \chi_{\mathrm{orb}}(R_-^i, \Gamma^i).$$

Along with Equation (5-2), this completes the proof of Theorem 1.1. $\qquad\square$

Let $\tilde{Q}$ denote the inertia orbifold of Q and $\pi : \tilde{Q} \to Q$ the projection (see [Chen and Ruan 2004]). It is shown in [Seaton 2008] that a vector field Y on Q induces a vector field $\tilde{Y}$ on $\tilde{Q}$, and that $\tilde{Y}(p, (g)) = 0$ if and only if $Y(p) = 0$.

For each point $p \in Q$ and $g \in G_p$, a chart $\{V_p, G_p, \pi_p\}$ induces a chart

$$\{V_p^g, C(g), \pi_{p,g}\} \quad \text{at} \quad (p, (g)) \in \tilde{Q},$$

where V_p^g denotes the points in V_p fixed by g and $C(g)$ denotes the centralizer of g in G_p. Clearly, $\partial V_p^g = (\partial V_p) \cap V_p^g$. An atlas for $\tilde{Q}$ can be taken consisting of charts of this form, so it is clear that $\partial \tilde{Q} = \widetilde{\partial Q}$.

Let $p \in \partial Q$ and pick a boundary product chart $\{C_p^+, G_p, \phi_p\}$. Then for $g \in G_p$, there is a chart $\{(C_p^+)^g, C(g), \phi_{p,g}\}$ for $(p, (g)) \in \tilde{Q}$. As the normal component to the boundary of C_p^+ is G_p-invariant,

$$\begin{aligned} (C_p^+)^g &= \left(\mathbf{B}_0^{n-1}(r/2) \times [0, \epsilon_p)\right)^g \\ &= \left(\mathbf{B}_0^{n-1}(r/2)\right)^g \times [0, \epsilon_p), \end{aligned}$$

and so

$$T_0(C_p^+)^g = T_0\left(\mathbf{B}_0^{n-1}(r/2)\right)^g \times \mathbb{R}.$$

It follows that $\tilde{Y}$ points out of $\partial \tilde{Q}$ at $(p, (g))$ if and only if Y points out of ∂Q at p. With this, applying Theorem 1.1 to $\tilde{Y}$ yields

$$\begin{aligned} \mathfrak{Ind}^{\mathrm{orb}}\left(\tilde{Y}; \tilde{Q}\right) &= \chi_{\mathrm{orb}}\left(\tilde{Q}, \partial \tilde{Q}\right) + \sum_{i=1}^{n} \chi_{\mathrm{orb}}\left(\widetilde{R_-^i}, \widetilde{\Gamma^i}\right) \\ &= \chi_{\mathrm{orb}}\left(\tilde{Q}\right) - \chi_{\mathrm{orb}}(\partial \tilde{Q}) + \sum_{i=1}^{n} \chi_{\mathrm{orb}}\left(\widetilde{R_-^i}\right) - \chi_{\mathrm{orb}}\left(\widetilde{\Gamma^i}\right). \end{aligned} \tag{5-3}$$

Each of the Γ^i and ∂Q are closed orbifolds, so it follows from the proof of Theorem 3.2 in [Seaton 2008] (note that the assumption of orientability is not used to establish this result) that

$$\chi_{\mathrm{orb}}(\widetilde{\Gamma^i}) = \chi(\mathbb{X}_{\Gamma^i}),$$

and

$$\chi_{\mathrm{orb}}(\partial \tilde{Q}) = \chi_{\mathrm{orb}}(\widetilde{\partial Q}) = \chi(\mathbb{X}_{\partial Q}), \tag{5-4}$$

where $\mathbb{X}_{\Gamma^i}$ and $\mathbb{X}_{\partial Q}$ denote the underlying topological spaces of Γ^i and ∂Q, respectively, and χ the usual Euler characteristic.

Letting $\hat{Q}$ denote, as above, the double of Q, it is easy to see that $\tilde{\hat{Q}} = \hat{\tilde{Q}}$. Hence, applying the same result to $\tilde{Q}$ yields

$$
\begin{aligned}
\chi(\mathbb{X}_{\hat{Q}}) &= \chi_{\mathrm{orb}}(\tilde{\hat{Q}}) \\
&= \chi_{\mathrm{orb}}(\hat{\tilde{Q}}) \\
&= 2\chi_{\mathrm{orb}}(\tilde{Q}) - \chi_{\mathrm{orb}}(\partial\tilde{Q}).
\end{aligned}
\tag{5-5}
$$

However, as

$$
\begin{aligned}
\chi(\mathbb{X}_{\hat{Q}}) &= 2\chi(\mathbb{X}_{Q}) - \chi(\mathbb{X}_{\partial Q}) \\
&= 2\chi(\mathbb{X}_{Q}) - \chi_{\mathrm{orb}}(\partial\tilde{Q}),
\end{aligned}
$$

it follows from Equation (5-5) that $\chi_{\mathrm{orb}}(\tilde{Q}) = \chi(\mathbb{X}_{Q})$. The same holds for each R_{-}^{i} so that Equation (5-3) becomes the following.

Corollary 5.2. *Let Q be an n-dimensional, smooth, compact orbifold with boundary, and let Y be a smooth vector field on Q. If $\tilde{Y}$ denotes the induced vector field on $\tilde{Q}$, then*

$$
\mathfrak{Ind}^{\mathrm{orb}}(\tilde{Y}; \tilde{Q}) = \chi(\mathbb{X}_{Q}, \mathbb{X}_{\partial Q}) + \sum_{i=1}^{n} \chi(\mathbb{X}_{R_{-}^{i}}, \mathbb{X}_{\Gamma^{i}}).
$$

Acknowledgments

References

[Chen and Ruan 2002] W. Chen and Y. Ruan, "Orbifold Gromov-Witten theory", pp. 25–85 in *Orbifolds in mathematics and physics (Madison, WI, 2001)*, Contemp. Math. **310**, Amer. Math. Soc., Providence, RI, 2002. MR 2004k:53145

[Chen and Ruan 2004] W. Chen and Y. Ruan, "A new cohomology theory of orbifold", *Comm. Math. Phys.* **248**:1 (2004), 1–31. MR 2005j:57036

[Guillemin and Pollack 1974] V. Guillemin and A. Pollack, *Differential topology*, Prentice-Hall, Englewood Cliffs, N.J., 1974. MR 50 #1276

[Hepworth 2007] R. Hepworth, "Morse inequalities for orbifold cohomology", preprint, 2007, Available at http://arxiv.org/pdf/0712.2432.

[Milnor 1963] J. Milnor, *Morse theory*, Based on lecture notes by M. Spivak and R. Wells. Annals of Mathematics Studies, No. 51, Princeton University Press, Princeton, N.J., 1963. MR 29 #634

[Milnor 1965] J. W. Milnor, *Topology from the differentiable viewpoint*, Based on notes by David W. Weaver, The University Press of Virginia, Charlottesville, Va., 1965. MR 37 #2239

[Moerdijk and Pronk 1999] I. Moerdijk and D. A. Pronk, "Simplicial cohomology of orbifolds", *Indag. Math. (N.S.)* **10**:2 (1999), 269–293. MR 2002b:55012

[Munkres 1963] J. R. Munkres, *Elementary differential topology*, Lectures given at Massachusetts Institute of Technology, Fall **1961**, Princeton University Press, Princeton, N.J., 1963. MR 29 #623

[Pugh 1968] C. C. Pugh, "A generalized Poincaré index formula", *Topology* **7** (1968), 217–226. MR 37 #4828

[Satake 1956] I. Satake, "On a generalization of the notion of manifold", *Proc. Nat. Acad. Sci. U.S.A.* **42** (1956), 359–363. MR 18,144a

[Satake 1957] I. Satake, "The Gauss-Bonnet theorem for *V*-manifolds", *J. Math. Soc. Japan* **9** (1957), 464–492. MR 20 #2022

[Seaton 2008] C. Seaton, "Two Gauss-Bonnet and Poincaré-Hopf theorems for orbifolds with boundary", *Differential Geom. Appl.* **26**:1 (2008), 42–51. MR 2009b:53135

[Thurston 1978] W. Thurston, "The geometry and topology of 3-manifolds", Lecture notes, Princeton University, Mathematics Department, 1978, Available at http://www.msri.org/publications/books/gt3m.

[Waner and Wu 1986] S. Waner and Y. Wu, "The local structure of tangent *G*-vector fields", *Topology Appl.* **23**:2 (1986), 129–143. MR 88c:55012

elliot.paquette@gmail.com   *University of Washington, Department of Mathematics, Box 354350, Seattle, WA 98195-4350, United States*

seatonc@rhodes.edu   *Department of Mathematics and Computer Science, Rhodes College, 2000 N. Parkway, Memphis, TN 38112, United States*
http://faculty.rhodes.edu/seaton/

On distances and self-dual codes over $F_q[u]/(u^t)$

Ricardo Alfaro, Stephen Bennett, Joshua Harvey and Celeste Thornburg

(Communicated by Nigel Boston)

New metrics and distances for linear codes over the ring $\mathbb{F}_q[u]/(u^t)$ are defined, which generalize the Gray map, Lee weight, and Bachoc weight; and new bounds on distances are given. Two characterizations of self-dual codes over $\mathbb{F}_q[u]/(u^t)$ are determined in terms of linear codes over $\mathbb{F}_q$. An algorithm to produce such self-dual codes is also established.

1. Introduction

Many optimal codes have been obtained by studying codes over general rings rather than fields. Lately, codes over finite chain rings (of which $\mathbb{F}_q[u]/(u^t)$ is an example) have been a source of many interesting properties [Norton and Salagean 2000a; Ozbudak and Sole 2007; Dougherty et al. 2007]. Gulliver and Harada [2001] found good examples of ternary codes over $\mathbb{F}_3$ using a particular type of *Gray map*. Siap and Ray-Chaudhuri [2000] established a relation between codes over $\mathbb{F}_q[u]/(u^2 - a)$ and codes over $\mathbb{F}_q$ which was used to obtain new codes over $\mathbb{F}_3$ and $\mathbb{F}_5$. In this paper we present a certain generalization of the method used in [Gulliver and Harada 2001] and [Siap and Ray-Chaudhuri 2000], defining a family of metrics for linear codes over $\mathbb{F}_q[u]/(u^t)$ and obtaining as particular examples the *Gray map, the Gray weight, the Lee weight* and the *Bachoc weight*. For the latter, we give a new bound on the distance of those codes. It also shows that the Gray images of codes over $\mathbb{F}_2 + u\mathbb{F}_2$ are more powerful than codes obtained by the so-called u-$(u+v)$ condition.

With these tools in hand, we study conditions for self-duality of codes over $\mathbb{F}_q[u]/(u^t)$. Norton and Salagean [2000b] studied the case of self-dual cyclic codes in terms of the generator polynomials. In this paper we study self-dual codes in terms of linear codes over $\mathbb{F}_q$ that are obtained as images under the maps defined on the first part of the paper. We provide a way to construct many self-dual codes over $\mathbb{F}_q$ starting from a self-dual code over $\mathbb{F}_q[u]/(u^t)$. We also study self-dual codes

in terms of the torsion codes, and provide a way to construct many self-dual codes over $\mathbb{F}_q[u]/(u^t)$ starting from a self-orthogonal code over $\mathbb{F}_q$. Our results contain many of the properties studied by Bachoc [1997] for self-dual codes over $\mathbb{F}_3 + u\mathbb{F}_3$.

2. Metric for codes over $\mathbb{F}_q[u]/(u^t)$

We will use $R(q, t)$ to denote the commutative ring $\mathbb{F}_q[u]/(u^t)$. The q^t elements of this ring can be represented in two different forms, and we will use the most appropriate in each case. First, we can use the polynomial representation with indeterminate u of degree less than or equal to $(t-1)$ with coefficients in $\mathbb{F}_q$, using the notation $R(q, t) = \mathbb{F}_q + u\mathbb{F}_q + u^2\mathbb{F}_q + \cdots + u^{t-1}\mathbb{F}_q$. We also use the u-ary coefficient representation as an $\mathbb{F}_q$-vector space.

Let $B \in M_t(\mathbb{F}_q)$ be an invertible $t \times t$ matrix, and let B act as right multiplication on $R(q, t)$ (seen as $\mathbb{F}_q$-vector space). We extend this action linearly to the $\mathbb{F}_q$-module $(R(q, t))^n$ by concatenation of the images $\phi_B : (R(q, t))^n \to (\mathbb{F}_q)^{tn}$ given by

$$\phi_B(x_1, x_2, \ldots, x_n) = (x_1 B, x_2 B, \ldots, x_n B)$$

An easy counting argument shows that ϕ_B is an $\mathbb{F}_q$-module isomorphism and if C is a linear code over $R(q, t)$ of length n, then $\phi_B(C)$ is a linear q-ary code of length tn.

Example 1. Consider the ring $R(3, 2) = \mathbb{F}_3 + u\mathbb{F}_3$ with $u^2 = 0$. Choosing

$$B = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix},$$

we obtain the *Gray map* $\phi_B : (\mathbb{F}_3 + u\mathbb{F}_3)^n \to \mathbb{F}_3^{2n}$ with

$$(a + ub)B = \begin{pmatrix} a & b \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} b & a+b \end{pmatrix}$$

used by Gulliver and Harada [2001].

Each such matrix B induces a new metric in the code C.

Definition 1. *Let C be a linear code over $R(q, t)$. Let B be an invertible matrix in $M_t(\mathbb{F}_q)$, and let ϕ_B be the corresponding map. The B-weight of an element $x \in R(q, t)$, $w_B(x)$, is defined as the Hamming weight of xB in $(\mathbb{F}_q)^t$. Also, the B-weight of a codeword $(x_1, \cdots, x_n) \in C$ is defined as:*

$$w_B(x_1, \cdots, x_n) = \sum_{i=1}^n w_B(x_i).$$

Similarly, the B-distance between two codewords in C is defined as the B-weight of their difference, and the B-distance, d_B, of the code C is defined as the minimal B-distance between any two distinct codewords.

Example 2. In the example above, the corresponding *B-weight* of an element of $\mathbb{F}_3 + u\mathbb{F}_3$ is given by

$$w_B(x) = w_B(a + ub) = w_H((a + ub)B)$$

$$= w_H(b, a + b) = \begin{cases} 0 & \text{if } x = 0, \\ 1 & \text{if } x = 1, 2, 2 + u, 1 + 2u, \\ 2 & \text{otherwise,} \end{cases}$$

which coincides with the *Gray weight* given in [Gulliver and Harada 2001].

Example 3. Consider the matrix

$$B = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix};$$

the corresponding *B-weight* of an element of $\mathbb{F}_2 + u\mathbb{F}_2$ is given by

$$w_B(x) = w_B(a + ub) = w_H((a + ub)B) = w_H(a + b, b) = \begin{cases} 0 & \text{if } x = 0, \\ 1 & \text{if } x = 1, 1 + u, \\ 2 & \text{if } x = u, \end{cases}$$

which produces the *Lee weight w_L* for codes over $\mathbb{F}_2 + u\mathbb{F}_2$.

Example 4. Consider the matrix

$$B = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix};$$

the corresponding *B-weight* of an element of $\mathbb{F}_q + u\mathbb{F}_q$ is given by

$$w_B(x) = w_B(a + ub) = w_H((a + ub)B)$$

$$= w_H(b, a) = \begin{cases} 0 & \text{if } x = 0, \\ 1 & \text{if exactly one of } a \text{ or } b \text{ is nonzero,} \\ 2 & \text{if both } a \text{ and } b \text{ are nonzero,} \end{cases}$$

which produces the Gray weight for codes in [Siap and Ray-Chaudhuri 2000].

The case $B = I_t$ corresponds to the special weight studied in [Ozbudak and Sole 2007] with regards to Gilbert–Varshamov bounds. A theorem similar to [Ozbudak and Sole 2007, Theorem 3] can be obtained using special families of matrices B. The definition leads immediately to the fact that ϕ_B preserves weights and distances between codewords.

When the generator matrix of a code C is of the form $G = (I \ M)$, C is called a *free code* over $R(q, t)$. In this case, we can establish the correspondence between

the parameters of the codes; see [Siap and Ray-Chaudhuri 2000, Section 2.2]. The case of nonfree codes will be considered later in Proposition 4.

Proposition 1. *Let B be an invertible matrix over $M_t(\mathbb{F}_q)$, let C be a linear free code over $R(q, t)$ of length n with B-distance d_B, and let ϕ_B be the corresponding map. Then $\phi_B(C)$ is a linear $[tn, tk, d_B]$-code over $\mathbb{F}_q$. Furthermore, the Hamming weight enumerator polynomial of the linear code $\phi_B(C)$ over $\mathbb{F}_q$ is the same as the B-weight enumerator polynomial of the code C over $R(q, t)$.*

Proof. Since B is nonsingular, $\phi_B(C)$ is a linear code over $\mathbb{F}_q$, with the same number of codewords. A basis for $\phi_B(C)$ can be obtained from a (minimal) set of generators for C, say, $y_1, y_2, \ldots, y_k$. The set $\{u^i y_j \mid i = 0..(t-1), j = 1..k\}$ forms a set of generators for C as an $\mathbb{F}_q$-submodule. Since C is free and B is invertible, it follows that $\{\phi_B(u^i y_j) \mid i = 0..(t-1), j = 1..k\}$ are linearly independent over $\mathbb{F}_q$ and form a basis for the linear code $\phi_B(C)$. Hence the dimension of the code $\phi_B(C)$ is tk. The equality of distance follows from the definition. □

In matrix form, we can construct a generator matrix for the linear code $\phi_B(C)$ as follows. Let G be a matrix of generators for C. For each row $(x_1, x_2, \ldots, x_n)$ of G consider the matrix representation $(X_1, X_2, \ldots, X_n)$ of the elements of $R(q, t)$ given by

$$X_i = \begin{pmatrix} a_0 & a_1 & a_2 & \cdots & a_{t-1} \\ 0 & a_0 & a_1 & \cdots & a_{t-2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & a_0 \end{pmatrix}.$$

For a free code, the rows of the matrix $(X_1 B, X_2 B, \ldots, X_n B)$ produce t linearly independent generators for the linear code $\phi_B(C)$. Repeating this process for each row of G, we will obtain the tk generators for $\phi_B(C)$. We denote this matrix by $\phi_B(G)$. For the case of nonfree linear codes, several rows will become zero and need to be deleted from the matrix. A counting of these rows will be given in Section 3.

Some choices of B can produce some optimal ternary and quintic codes as we now illustrate.

Example 5. Consider a linear code C over $\mathbb{F}_3 + u\mathbb{F}_3$ of length 9 with generator matrix:

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & u & 2+u & 1+u & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & u & 2+u & 1+u & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & u & 2+u & 1+u \\ 0 & 0 & 0 & 1 & 1+u & 1 & 0 & u & 2+u \end{pmatrix}$$

Let

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

The *B-weight* enumerator polynomial is given by

$$1 + 98x^7 + 206x^8 + 412x^9 + 780x^{10} + 1032x^{11} + 1308x^{12} + 1224x^{13}$$
$$+ 828x^{14} + 462x^{15} + 166x^{16} + 40x^{17} + 4x^{18}.$$

The corresponding linear ternary code $\phi_B(C)$ is an optimal ternary $[18, 8, 7]$-code.

Notice that if we take

$$B = \begin{pmatrix} 1 & 2 \\ 1 & 0 \end{pmatrix},$$

we get a linear ternary code $\phi_B(C)$ of length 18, dimension 8, but now, with minimal distance 4. The challenge now is to look for matrices B that produce optimal codes.

Example 6. Consider a linear code C over $\mathbb{F}_5 + u\mathbb{F}_5$ of length 5 with a generator matrix:

$$G = \begin{pmatrix} 1 & 0 & 2u & 3+3u & 4 \\ 0 & 1 & 4 & 2u & 3+3u \end{pmatrix}.$$

Let

$$B = \begin{pmatrix} 3 & 0 \\ 2 & 3 \end{pmatrix}.$$

The linear $\mathbb{F}_5$-code $\phi_B(C)$ is an optimal $[10, 4, 6]$-code, with generator matrix given by

$$\phi_B(G) = \begin{pmatrix} 1 & 0 & 0 & 0 & 2 & 3 & 2 & 2 & 0 & 3 \\ 0 & 1 & 0 & 0 & 2 & 3 & 3 & 1 & 2 & 1 \\ 0 & 0 & 1 & 0 & 0 & 3 & 2 & 3 & 2 & 2 \\ 0 & 0 & 0 & 1 & 2 & 1 & 2 & 3 & 3 & 1 \end{pmatrix}.$$

Example 7. Consider a linear code C over $R(5, 3) = \mathbb{F}_5 + u\mathbb{F}_5 + u^2\mathbb{F}_5$ of length 14 with generator matrix obtained by cyclic shifts of the first 5 components and cyclic shift of the last 9 components of the vector:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & u & 3+3u & 2+4u & 4u & 0 & 4 & 3+u^2 & 2+u+u^2 & u+u^2 \end{pmatrix}.$$

Let

$$B = \begin{pmatrix} 0 & 3 & 3 \\ 0 & 0 & 4 \\ 3 & 3 & 2 \end{pmatrix}.$$

The *B-weight* enumerator polynomial is given by

$$1 + 24x^{16} + 32x^{17} + 80x^{18} + 150x^{19} + 158x^{20} + 140x^{21} + 82x^{22} + 44x^{23} + 14x^{24} + 4x^{25}$$

and the linear $\mathbb{F}_5$-code $\phi_B(C)$ is an optimal $[42, 15, 16]$-code over $\mathbb{F}_5$.

3. Metrics using the torsion codes

A generalization of the residue and torsion codes for $\mathbb{F}_2 + u\mathbb{F}_2$ has been studied in [Norton and Salagean 2000b] where a *generator matrix* for a code C over $R(q, t)$ is defined as a matrix G over $R(q, t)$ whose rows span C and none of them can be written as a linear combination of the other rows of G. Recalling that two codes over $R(q, t)$ are *equivalent* if one can be obtained from the other by permuting the coordinates or by multiplying all entries in a specified coordinate by an invertible element of $R(q, t)$, and performing Gauss elimination (remembering not to multiply by nonunits) we can always obtain a generator matrix for a code (or equivalent code) which is in *standard form*, that is, in the form

$$G = \begin{pmatrix} I_{k_1} & B_{1,2} & B_{1,3} & B_{1,4} & \cdots & B_{1,t} & B_{1,t+1} \\ 0 & uI_{k_2} & uB_{2,3} & uB_{2,4} & \cdots & uB_{2,t} & uB_{2,t+1} \\ 0 & 0 & u^2I_{k_3} & u^2B_{3,4} & \cdots & u^2B_{3,t} & u^2B_{3,t+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & u^{t-1}I_{k_t} & u^{t-1}B_{t,t+1} \end{pmatrix},$$

where $B_{i,j}$ is a matrix of polynomials in $\mathbb{F}_q[u]/(u^t)$ of degrees at most $j - i - 1$. In fact, we can think of $B_{i,j}$ as a matrix of the form

$$B_{i,j} = A_{i,j,0} + A_{i,j,1}u + \cdots + A_{i,j,j-i-1}u^{j-i-1},$$

where the matrices $A_{i,j,r}$ are matrices over the field $\mathbb{F}_q$.

We define the following *torsion codes* over $\mathbb{F}_q$:

$$C_i = \{X \in (\mathbb{F}_q)^n \mid \exists Y \in (\langle u^i \rangle)^n \; with \; Xu^{i-1} + Y \in C\},$$

for $i = 1 \ldots t$. It is then easy to see that these are linear q-ary codes, and we have:

Proposition 2. *Let C be a linear $R(q, t)$ code of length n, and let C_i, $i = 1 \ldots t$ be the torsion codes defined above. Then*

(1) $C_1 \subseteq C_2 \subseteq \cdots \subseteq C_t$;

(2) *a generator matrix for the code C_1 is given by*

$$G_1 = \begin{pmatrix} I_{k_1} & A_{1,2,0} & A_{1,3,0} & \cdots & A_{1,t+1,0} \end{pmatrix};$$

(3) *if G_i is a generator matrix for the code C_i, then a generator matrix G_{i+1} for the code C_{i+1} is given by*

$$G_{i+1} = \begin{pmatrix} & & G_i & & \\ 0 & \cdots & 0 & I_{k_{i+1}} & A_{i+1,i+2,0} & \cdots & A_{i+1,t+1,0} \end{pmatrix}.$$

Proof. Let $X \in C_i$, then there exists $Y \in (\langle u^i \rangle)^n \mid z := Xu^{i-1} + Y \in C$. Then $uz \in C$. But $uz = Xu^i + uY \in C$. Hence $X \in C_{i+1}$. Now, let $X \in C_1$. Then there

exist vectors $Y_i, i = 1..t-1$ over $(\mathbb{F}_q)^n$ such that $X + Y_1 u + \cdots + Y_{t-1} u^{t-1} \in C$. Thus, the coefficients of X must come from independent coefficients of elements on the first row-group of the generator matrix G. A similar reasoning indicates that at each stage, the remaining generators come from the independent coefficients of elements in the next row-group of the matrix G.   □

Note that the code C_i has dimension $k_1 + \cdots + k_i$. The code C then contains all products $[v_1, v_2, \ldots, v_t]G$ where the components of the vectors $v_i \in (R(q, t))^{k_i}$ have degree at most $t-i$. The number of codewords in C is then $q^{(t)k_1+(t-1)k_2+\cdots+k_t}$, which can also be seen as $q^{k_1} q^{k_1+k_2} \ldots q^{k_1+k_2+\cdots k_t}$. For the case $\mathbb{F}_2 + u\mathbb{F}_2$, the code C_1 is called the *residue* code, and the code $C_t = C_2$ is called the *torsion* code.

For $X \in C_i$, we know there exists $Y \in (\langle u^i \rangle)^n$ such that $Xu^{i-1} + Y \in C$. Y can be written as

$$Y = u^i \overline{Y} + \text{hot}, \quad \text{with } \overline{Y} \in \mathbb{F}_q^n,$$

where 'hot' designates *higher order terms*. With this notation, define the map

$$F_i : C_i \to \mathbb{F}_q^n / C_{i+1}$$

by $F_i(X) = \overline{Y} + C_{i+1}$. If two such vectors $Y_1, Y_2 \in (\langle u^i \rangle)^n$ exist, we have

$$Y_1 = u^i \overline{Y}_1 + \text{hot} \quad \text{and} \quad Y_2 = u^i \overline{Y}_2 + \text{hot}.$$

Then,

$$Y_2 - Y_1 = u^i(\overline{Y}_2 - \overline{Y}_1) + \text{hot} \in C.$$

Therefore $\overline{Y}_2 - \overline{Y}_1 \in C_{i+1}$ and F_i is well defined. It is easy to see that the maps F_i are $\mathbb{F}_q$-morphisms. By its very definition, it can be seen that the image of these maps consist of direct sums of the matrices $A_{i,j,r}$ in a generator matrix G for C in standard form. We then have:

Theorem 1. *Let C be a code over $R(q, t)$ with a generator matrix G in standard form. C is determined uniquely by a chain of linear codes C_i over $\mathbb{F}_q$ and $\mathbb{F}_q$-module homomorphisms $F_i : C_i \to \mathbb{F}_q^n / C_{i+1}$.*

Example 8. If $G = (I_{k_1} A)$ then $C_1 = C_2 = \cdots = C_t$. Also $k_i = 0$ for all $i \geq 2$ and hence the code C has $(q^t)^{k_1}$ elements. These are called *free codes* since they are free $R(q, t)$-modules. Furthermore, if $A = A_0 + u B_1 + u^2 B_2 + \cdots + u^{t-1} B_{t-1}$, where B_i is a matrix over $\mathbb{F}_q$, then C_1 determines A_0 and $F_i(C_i)$ determines B_i.

Example 9. Let

$$G = \begin{pmatrix} 1 & 0 & 2 & 2+u & 1+u+u^2 \\ 0 & 1 & 1 & 1+2u & u+u^2 \\ 0 & 0 & u & 2u & u+u^2 \\ 0 & 0 & 0 & u^2 & 2u^2 \end{pmatrix}$$

be a generator matrix for a code C over $R(3, 3)$. The corresponding generator matrices for the linear codes are:

$$C_1 = \begin{pmatrix} 1 & 0 & 2 & 2 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix}, \qquad \text{a } [5, 2, 3]\text{-code over } \mathbb{F}_3,$$

$$C_2 = \begin{pmatrix} 1 & 0 & 2 & 2 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 2 & 1 \end{pmatrix}, \qquad \text{a } [5, 3, 2]\text{-code over } \mathbb{F}_3,$$

$$C_3 = \begin{pmatrix} 1 & 0 & 2 & 2 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix}, \qquad \text{a } [5, 4, 1]\text{-code over } \mathbb{F}_3,$$

and the code C has $(3^3)^2 (3^2)^1 (3)^1 = 27^3$ codewords.

Utilizing the torsion codes of C we can define a new weight on C and obtain a bound for their minimum distance.

Definition 2. *Let $x \in R(q, t)$ and let p be the characteristic of the field $\mathbb{F}_q$. Let $i_0 = \max\{i \mid x \in \langle u^i \rangle\}$. Define the p-weight of x as $wt_p(x) = p^{i_0}$, if $x \neq 0$ and $wt_p(0) = 0$. For an element of $(R(q, t))^n$ define the p-weight as the sum of the p-weights of its coordinates.*

Note. For the case $R(2, 2) = \mathbb{F}_2 + u\mathbb{F}_2$, the p-weight coincides with the Lee weight, and for $R(p, 2) = \mathbb{F}_p + u\mathbb{F}_p$, the p-weight coincides with the Bachoc weight defined in [Bachoc 1997].

Theorem 2. *Let C be a linear code over $R(q, t)$, and let $C_1, C_2, \ldots, C_t$ be the associated torsion codes over $\mathbb{F}_q$. Let d_i be the Hamming distance of the codes C_i, then the minimum weight d of the code C with respect to the p-weight satisfies*

$$\min\{p^{i-1}d_i \mid i = 1, .., t\} \leq d \leq p^{t-1}d_t.$$

Proof. Let $W = (y_1, y_2, \ldots, y_n) \in C$ with minimum weight. Then for some i, $W = u^i X + Y$ with $Y \in \langle u^{i+1} \rangle$. Thus $X \in C_{i+1}$ and $wt_p(W) \geq p^i \cdot wt_H(X) \geq p^i d_{i+1}$. Now take $X_1 \in C_t$ to be a word of minimum weight d_t, then $u^{t-1}X_1 \in C$, and, by the minimality of W, we have $wt_p(W) \leq wt_p(u^{t-1}X_1) = p^{t-1}d_t$. $\qquad \square$

It is well known [Bonnecaze and Udaya 1999; Ling and Sole 2001], that the Lee weight for a cyclic code C over $\mathbb{F}_2 + u\mathbb{F}_2$ is the lower bound above. Here we show an example over $\mathbb{F}_2 + u\mathbb{F}_2$ that attains the upper bound.

Example 10. Let C be the linear code over $\mathbb{F}_2 + u\mathbb{F}_2$, with generator matrix

$$G = \begin{pmatrix} 1 & 0 & u & 1 \\ 0 & 1 & 1+u & u \end{pmatrix}.$$

The codeword (u, u, u, u) has Lee (or 2-) weight 8, while all the other nonzero codewords have weight 4. On the other hand C_1 and C_2 are equal with generator matrix

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

Hence $d_1 = d_2 = 2$, and $\min\{d_1, 2d_2\} = 2 \neq d$.

Since the p-weight coincides with the Lee weight for codes over $\mathbb{F}_2 + u\mathbb{F}_2$, we obtain the general version for the Lee weight of those codes as a corollary of Theorem 2.

Corollary 1. *The minimum Lee weight of a code C over $\mathbb{F}_2 + u\mathbb{F}_2$, satisfies*

$$\min\{d_1, 2d_2\} \leq d \leq 2d_2$$

where d_1, d_2 are respectively the Hamming distance of the residue code C_1 and the torsion code C_2.

Example 11. Return to Example 9 over $R(3, 3)$ with $d_1 = 3, d_2 = 2, d_3 = 1$. Hence $3 \leq d \leq 9$. The first and second generators combine to form a codeword of p-weight 3. Hence $d = 3$, and in this example the minimum weight attains the lower bound.

Example 12. Let C be the linear code over $\mathbb{F}_3 + u\mathbb{F}_3$, with generator matrix

$$G = \begin{pmatrix} 1 & 0 & u & 2 \\ 0 & 1 & 1+u & u \end{pmatrix}.$$

There are only 4 codewords with 2 zero entries, and they have Bachoc weight (and hence p-weight) 6. There are no codewords with Bachoc weight 3, and the Bachoc distance d of the code is 4. On the other hand the associated ternary codes are

$$C_1 = C_2 = \begin{pmatrix} 1 & 0 & 0 & 2 \\ 0 & 1 & 1 & 0 \end{pmatrix}.$$

Thus $d_1 = d_2 = 2$ and the Bachoc weight d lies strictly between the bounds given above.

Corollary 2. *For free codes the p-weight d satisfies: $d_1 \leq d \leq p^{t-1}d_1$.*

We can also use the torsion codes to study the Hamming weight of the code C. The results given here use a straightforward proof in comparison with the proof given in [Norton and Salagean 2000a].

For a code C over $R(q, t)$ and $w \in C$, we denote $w_H(w)$ the usual Hamming weight of w. Accordingly, the minimum Hamming distance of the code will be denoted by $d_H(C)$.

Proposition 3. *Let C be a linear code over $R(q,t)$, and let $C_1, C_2, \ldots, C_t$ be the associated torsion codes over $\mathbb{F}_q$. Let d_i be the Hamming distance of the codes C_i, then the minimal Hamming weight d_H of the code C satisfies*

$$dd_H = d_t \leq d_{t-1} \leq \cdots \leq d_1.$$

Proof. Since $C_i \subseteq C_{i+1}$, it follows that $d_{i+1} \leq d_i$, for $i = 1..t$ Now let $X \in C_t$. Then $Xu^{t-1} \in C$ and hence $d_H \leq d_t$. Conversely, let w^* be a codeword in C with minimum Hamming weight d_H. Let j be the maximum integer such that u^j divides w^*. Then $w^* = u^j v$ and $z = u^{t-j-1} w^* = u^{t-1} v \in C$. Thus $\hat{v} \in C_t$, where $\hat{v}$ denotes the canonical projection from $R(q,t)^n$ into $\mathbb{F}_q^n$. We then have $w_H(w^*) \geq w_H(\hat{v}) \geq d_t$, and therefore $d_H \geq d_t$. □

From the above proof, the Singleton bound for C_t, and the comment after [Proposition 2](#), we have:

Corollary 3. *Let C be a linear code over $R(q,t)$, and let $C_1, C_2, \ldots, C_t$ be the associated torsion codes. Then:*

$$d_H \leq n - (k_1 + k_2 + \cdots k_t) + 1.$$

Proposition 4. $\phi_B(C)$ *is a $[\, nt, \; \sum_{i=1}^t k_i(t-i+1) \,, d^*]$ linear code over $\mathbb{F}_q$, with $d^* \leq t d_t$.*

Proof. Since u^{i-1} divides y_j for each y_j in the i-th row-block of G, $u^s y_j = 0$ for $s \geq t - i + 1$. Furthermore, the generators $u^s y_j \neq 0$ for $s < t - i + 1$ are linearly independent. Since there are k_i such y_j, we have

$$\dim(\phi_B(C)) = \sum_{i=1}^t k_i\big(t - (i-1)\big). □$$

4. Self-dual codes over $\mathbb{F}_q[u]/(u^t)$ using torsion codes

Duality for codes over $\mathbb{F}_q[u]/(u^t)$ is understood with respect to the inner product $x \cdot y = \sum x_i y_i$, where $x_i, y_i \in R(q,t)$. As usual, a code is called *self-dual* if $C = C^\perp$, and is called *self-orthogonal* is $C \subseteq C^\perp$.

First, we give an examples of self-dual codes over $R(q,t)$ of length n when t is even and n is a multiple of p (the characteristic of the field $\mathbb{F}_q$.) The construction mimics the C_n codes studied by [Bachoc [1997]](#) for the case $t = 2$.

Example 13. For t even, let $I = \langle u^{t/2} \rangle \subseteq R(q,t)$. Define the set:

$$D_n := \{(x_1, x_2, \ldots, x_n) \in R(q,t)^n \mid \sum_{i=1}^n x_i = 0 \text{ and } x_i - x_j \in I \text{ for all } i \neq j\}.$$

Let $X, Y \in D_n$,

$$X \cdot Y = \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n}(x_i - x_1)(y_i - y_1) + \sum_{i=1}^{n} x_i y_1 + \sum_{i=1}^{n} x_1 y_i - n x_1 y_1.$$

The first term is in $I^2 = 0$, the next two terms are zero by definition and the third term is zero since $p|n$. Thus $D_n \subseteq D_n^\perp$. Now, for each $i = 1 \ldots n$, we can write $x_i = a + b_i$ where a is a common polynomial of degree less than $t/2$, and $b_i \in I$ with $\sum b_i = 0$. There are $q^{t/2}$ choices for a, and $(q^{n-1})^{t/2}$ choices for the b_i's, thus

$$|D_n| = q^{t/2}(q^{n-1})^{t/2} = q^{nt/2},$$

and hence D_n is self-dual.

The torsion q-ary codes are as follows: for $i = 1, \ldots t/2$, C_i is the code generated by the $\mathbf{1}$ word, with $d_i = n$; and for $i = t/2 + 1 \ldots t$, C_i is the parity check code of length n and dimension $n - 1$, thus $d_i = 2$. Applying Theorem 2, we obtain

$$\min\{n, 2p^{t/2}\} \le d \le 2p^{t-1}.$$

But 1 and $(0, 0, \ldots, 0, u^{t/2}, -u^{t/2}, 0, \ldots, 0) \in D_n$, hence $d = \min\{n, 2p^{t/2}\}$.

We study self-orthogonal and self-dual codes over $R(q, t)$ taking two different approaches. We look at the linear codes $\phi_B(C)$, and also look at the torsion codes corresponding to C.

To study the latter we need some results on the parity check matrix of these codes, which can be defined in terms of block matrices using the recurrence relation

$$D_{i,j} = \sum_{k=i+1}^{t+2-j} -B_{i,k} D_{k,j}$$

for blocks, such that $i + j \le t + 1$. For blocks such that $i + j = t + 2$, $D_{i,j} = u^{t-j+1} I_{k_j}$ for $i = 2, \ldots, t$ and $D_{t+1,1} = I_{n-(k_1+k_2+\ldots k_t)}$. All remaining blocks are 0. From here a generator matrix for the dual code can be obtained and we easily observe the following relations: $k_1(C^\perp) = n - (k_1 + \ldots + k_t)$ and $k_h(C^\perp) = k_{t-h+2}(C)$ for $h = 2, \ldots, t$.

A different recurrence relation for the definition of the parity check matrix is given in [Norton and Salagean 2000a].

Proposition 5. *Let C be an $R(q, t)$ code, and let C_i's be its corresponding torsion codes. Then*

$$(C^\perp)_i = (C_{t-i+1})^\perp, \; i = 1..t.$$

Proof. Let $w \in (C^\perp)_i$ and $v \in C_{t-i+1}$. Then there exists $z \in (\langle u^i \rangle)^n$ with $a := wu^{i-1} + z \in C^\perp$, and $y \in (\langle u^{t-i+1} \rangle)^n$ with $b := vu^{t-i} + y \in C$. Since $a \cdot b = 0$, we

have

$$0 = (wu^{i-1} + z) \cdot (vu^{t-i} + y) = (w \cdot v)u^{t-1},$$

which implies $w \cdot v = 0$, and $w \in (C_{t-i+1})^{\perp}$. So $(C^{\perp})_i \subseteq (C_{t-i+1})^{\perp}$. Looking at dimensions

$$\dim((C^{\perp})_i) = \sum_{j=1}^{i} k_j(C^{\perp}) = n - (k_1 + \ldots + k_t) + \sum_{j=2}^{i} k_{t-j+2}(C)$$

$$= n - \sum_{j=1}^{t-i+1} k_j(C) = n - \dim(C_{t-i+1}) = \dim((C_{t-i+1})^{\perp}). \qquad \square$$

Using the generator in standard form of a code C and forming the inner products of its row-blocks we obtain:

Proposition 6. *Let C be an $R(q, t)$ code with a generator matrix in standard form. C is self-orthogonal if and only if*

$$\sum_{h=0}^{k} \sum_{j=\max\{i,k\}}^{t+1} A_{i,j,h} A_{l,j,k-h}^t = 0, \qquad \text{for each } k = 0, \ldots, t - (i + l - 2) - 1.$$

This gives us the first characterization of self-dual codes:

Theorem 3. *Let C be an $R(q, t)$ code; and let C_i's be its corresponding torsion codes. The code C is self-orthogonal and $C_i = C_{t-i+1}^{\perp}$ if and only if C is self-dual.*

Proof. By Proposition 5 we have $(C^{\perp})_i = C_{t-i+1}^{\perp} = C_i$ for all $i = 1 \ldots t$. Furthermore, $\text{rk}(C) = \dim(C_t) = \dim((C^{\perp})_t) = \text{rk}(C^{\perp})$; but C is self-orthogonal, hence $C = C^{\perp}$. Similarly, the converse follows immediately from Proposition 5. $\qquad \square$

As an immediate consequence we have:

Corollary 4. *If C is self-dual, then C_i is self-orthogonal for all $i \leq (t + 1)/2$.*

Note that when t is odd, $C_{\lfloor (t+1)/2 \rfloor}$ is self-dual and hence n must be even. For the case t even, we can contruct self-dual codes of even or odd length.

Proposition 6 and Theorem 3 provide us with an algorithm to produce self-dual codes over $R(q, t)$ starting from self-orthogonal codes over $\mathbb{F}_q$.

(1) Take a self-orthogonal code C_1 over $\mathbb{F}_q$.

(2) Define $C_t := C_1^{\perp}$.

(3) Choose a set of self-orthogonal words $\{R_1, R_2, ..., R_l\}$ in C_t that are linearly independent from C_1. Define

$$C_2 := \langle C_1 \cup \{R_1, R_2, \ldots, R_l\} \rangle \quad \text{and} \quad C_{t-1} = C_2^{\perp}.$$

(4) Repeat, if possible, the step above defining C_i and $C_{t-i+1} = C_i^{\perp}$ until you produce $C_{\lfloor (t+1)/2 \rfloor}$.

(5) For each $i = 1..t$, multiply the generators of $\{C_{i+1} - C_i\}$ by u^i. This will produce a self-dual code.

Additional self-dual codes are obtained as follows:

(6) Form a generator matrix G in standard form, adding, where appropriate, variables to represent higher powers of u.

(7) Now we find the system of equations on the defined variables arising from Proposition 6. Note that for fixed $i, l = 1 \ldots t$ each k will produce a matrix equation, which in turn produces several nonlinear equations.

(8) Write this system of equations in terms of the independent variables. There will be

$$\sum_{i=1}^{\lfloor t/2 \rfloor} \sum_{j=i}^{t-i} (t - i - j + 1)k_i k_j$$

equations on

$$\sum_{i=1}^{t-1} \sum_{j=i+2}^{t+1} (j - i - 1)k_i k_j \text{ total variables.}$$

(9) By Theorem 3 every solution to this system of equations will produce a self-dual code (some may be equivalent).

We now provide an example of this construction.

Example 14. Self-dual codes in $R(3, 4)$:

Consider the self-orthogonal code

$$C_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 2 \\ 0 & 1 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

Define

$$C_4 := C_1^{\perp} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 2 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

Since there are no more self-orthogonal words in C_4 to append to C_1, we let $C_2 := C_1$, and since $C_2^{\perp} = C_4$ we let $C_3 := C_4$. Multiplying the rows in $C_3 - C_2$ by u^2 we obtain a generator matrix for a self-dual code over $R(3, 4)$:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 2 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & u^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & u^2 & 0 & 0 \end{pmatrix}$$

Now we can form a generator matrix using variables to represent higher powers of u obtaining

$$\begin{pmatrix} 1 & 0 & au & bu & 1+cu+du^2+eu^3 & 2+fu+gu^2+hu^3 \\ 0 & 1 & iu & ju & 1+ku+lu^2+mu^3 & 1+nu+pu^2+qu^3 \\ 0 & 0 & u^2 & 0 & ru^3 & su^3 \\ 0 & 0 & 0 & u^2 & tu^3 & vu^3 \end{pmatrix}.$$

The equation

$$\sum_{h=0}^{k} \sum_{j=\max\{i,k\}}^{t+1} A_{i,j,h} A^t_{l,j,k-h} = 0$$

produces a system of equations over $\mathbb{F}_q$. For example, for $i = 1, l = 2, k = 3$ we obtain the equation

$$a + r + 2s = 0,$$
$$b + t + 2v = 0,$$
$$i + r + s = 0,$$
$$j + t + v = 0.$$

Likewise, the remaining equations can be obtained, and we solve in terms of a set of independent variables $\{a, b, h, i, j, n, p\}$:

$$c = n,$$
$$d = ai + bj + i^2 + j^2 + p + 2a^2 + 2b^2,$$
$$e = n(ai + bj + i^2 + j^2 + 2n^2 + p) + h,$$
$$f = n,$$
$$g = a^2 + b^2 + ai + bj + i^2 + j^2 + n^2 + p,$$
$$k = 2n,$$
$$l = i^2 + j^2 + 2p + 2n^2,$$
$$m = n(i^2 + j^2 + 2a^2 + 2b^2 + ai + bj) + 2h,$$
$$q = n(a^2 + b^2 + p + 2ai + 2bj + 2n^2) + h,$$
$$r = a - 2i,$$
$$s = i - a,$$
$$t = b - 2j,$$
$$v = j - b.$$

These equations allow us to generate up to 3^7 self-dual codes over $R(3,4)$. As an example, letting all the independent variables take the value 1 except for $b = 0$, we obtain the self-dual code

$$
\begin{pmatrix}
1 & 0 & u & 0 & 1+u+u^3 & 2+u+u^3 \\
0 & 1 & u & u & 1+2u+u^3 & 1+u+u^2+u^3 \\
0 & 0 & u^2 & 0 & 2u^3 & 0 \\
0 & 0 & 0 & u^2 & u^3 & u^3
\end{pmatrix}.
$$

5. Self-dual codes over $\mathbb{F}_q[u]/(u^t)$ using linear images

As discussed in Section 2, given a code C over $R(q,t)$ of length n and a nonsingular $t \times t$ matrix B over $\mathbb{F}_q$, we can define a linear code $\phi_B(C)$ over $\mathbb{F}_q$ of length nt. In this section, we will consider an element $x \in R(q,t)$ in its polynomial representation, and will use $\overline{x}$ for its vector representation.

Let $w = (w_1, w_2, \ldots, w_n)$ be a codeword in C. Recall that

$$
\phi_B(w) = (\overline{w}_1 B, \overline{w}_2 B, \ldots, \overline{w}_n B).
$$

Let E denote the square matrix

$$
\begin{pmatrix}
1 & 1 & \cdots & 1 \\
1 & 1 & & \iddots \\
\vdots & & \iddots & \\
1 & & & 0
\end{pmatrix}
\quad \text{over } \mathbb{F}_q.
$$

Theorem 4. *If C is self-orthogonal and $BB^T = cE$ where $c \neq 0 \in \mathbb{F}_q$, then $\phi_B(C)$ is self-orthogonal.*

Proof. Let R_j denote the j-th row of B. Then $R_j R_k^T = c$, for all $j + k < t + 2$ and $R_j R_k^T = 0$, for all $j + k \geq t + 2$. If $w, v \in C$, then

$$
\phi_B(w)\phi_B(v)
$$

$$
= \sum_{i=1}^{n} \overline{w}_i B (\overline{v}_i B)^T = \sum_{i=1}^{n} \overline{w}_i B B^T \overline{v}_i
$$

$$
= \sum_{i=1}^{n} \sum_{j,k=0}^{t-1} w_{i,j} R_{j+1} R_{k+1}^T v_{i,k} = c \sum_{i=1}^{n} \sum_{j+k<t}^{t-1} w_{i,j} v_{i,k} + 0 \sum_{i=1}^{n} \sum_{j+k\geq t}^{2t-2} w_{i,j} v_{i,k},
$$

but since C is self-orthogonal, the sum in the first term is 0. Therefore,

$$
\phi_B(w)\phi_B(v) = 0,
$$

and thus $\phi_B(C)$ is self-orthogonal. $\qquad\square$

Corollary 5. *If C is self-dual, $BB^T = cE$, and*

$$\sum_{i=2}^{t} k_i(t - 2i + 2) = 0,$$

then $\phi_B(C)$ is self-dual.

Proof. Splitting the equation from the hypothesis we have

$$\sum_{i=2}^{t} k_i(t - i + 1) = \sum_{i=2}^{t} k_i(i - 1),$$

$$2\sum_{i=2}^{t} k_i(t - i + 1) = \sum_{i=2}^{t} k_i(i - 1) + \sum_{i=2}^{t} k_i(t - i + 1) = \sum_{i=2}^{t} tk_i,$$

$$2\sum_{i=1}^{t} k_i(t - i + 1) = 2k_1 t + \sum_{i=2}^{t} tk_i.$$

Since C is self-dual, we know

$$C_1^{\perp} = C_t \quad \text{and} \quad \dim(C_t) = \text{rk}(C).$$

Thus,

$$\dim(C_1^{\perp}) = \text{rk}(C) \quad \text{and} \quad n - k_1 = \sum_{i=1}^{t} k_i.$$

Therefore,

$$2\sum_{i=1}^{t} k_i(t - i + 1) = nt,$$

making the length of $\phi_B(C)$ twice its dimension. By Theorem 4, $\phi_B(C)$ is self-orthogonal and hence $\phi_B(C)$ is self-dual. $\square$

Let M, N be two matrices over $\mathbb{F}_q$. We say they are *root-equivalent* ($M \sim N$) if M can be obtained from N by a column permutation, or a column multiplication by an element $\alpha \in \mathbb{F}_q$ such that $\alpha^2 = 1$. This implies $MM^T = NN^T$, and by the definition of ϕ_B, we obtain the following

Corollary 6. *If $B \sim D$ in the hypothesis of Corollary 5 then $\phi_B(C)$ and $\phi_D(C)$ are equivalent self-dual codes.*

Example 15. For R(3,3), all matrices B that satisfy $BB^t = cE$ are root-equivalent, and therefore produce equivalent codes. Hence we can restrict ourselves to just one such matrix, for example,

$$B = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

The cases of $R(2,2)$ and $R(3,3)$ are singular. For $R(3,4)$ we have 6 different classes of root-equivalent matrices.

In general, note that there exist self-dual codes A and matrices B with $BB^T \neq cE$ whose image $\phi_B(A)$ is self-dual. For example, consider the self-dual code A over $R(3,4)$ with a generator matrix

$$G = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1+2u+u^2 & 1+2u & 1+2u+u^2 & 1+2u & 1+2u+u^2 & 1+2u \\ 1+u^2 & 1+u^2 & 1+u^2 & 1 & 1 & 1 \\ u+u^2 & u & u & u+u^2 & u+u^2 & u \\ 0 & 0 & 0 & 0 & u^2 & 2u^2 \end{pmatrix}.$$

Passing to standard form,

$$G_1 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & u^2 & 0 & 0 & 0 & 2u^2 \\ 0 & 0 & u^2 & 0 & 0 & 2u^2 \\ 0 & 0 & 0 & u^2 & 0 & 2u^2 \\ 0 & 0 & 0 & 0 & u^2 & 2u^2 \end{pmatrix}.$$

Consider the matrix

$$B = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 1 \\ 0 & 1 & 2 & 1 \\ 2 & 1 & 1 & 0 \end{pmatrix},$$

for which $BB^T \neq cE$ for any c. The image code $\phi_B(A)$ is a self-dual code:

$$\phi_B(A) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 0 & 0 & 2 & 2 & 0 & 0 & 2 & 2 & 0 & 0 & 2 & 2 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 2 & 0 & 0 & 1 & 2 & 0 & 0 & 1 & 2 & 0 & 0 & 1 & 2 & 0 & 2 & 2 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 1 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 2 & 0 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 1 & 0 & 0 & 0 & 0 & 2 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 2 & 0 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 1 & 0 & 2 & 1 & 2 \end{pmatrix}.$$

References

[Bachoc 1997] C. Bachoc, "Applications of coding theory to the construction of modular lattices", *J. Combin. Theory Ser. A* (**78**) (1997), pp. 92–119.

[Bonnecaze and Udaya 1999] A. Bonnecaze and P. Udaya, "Cyclic codes and self-dual codes over $F_2 + uF_2$", *IEEE Trans. Inform. Theory* **(45)** (1999), pp. 1250–1255.

[Dougherty et al. 2007] S. Dougherty, A. Gulliver, Y. Park, and J. Wong, "Optimal linear codes over $\mathbb{Z}_m$", *J. Korean Math. Soc.* **(44)** (2007), pp. 1139–1162.

[Gulliver and Harada 2001] T. A. Gulliver and M. Harada, "Codes over $F_3 + uF_3$ and improvements to the bounds on ternary linear codes", *Design, Codes and Cryptography* **(22)** (2001), pp. 89–96.

[Ling and Sole 2001] S. Ling and P. Sole, "Duadic codes over $F_2 + uF_2$", *AAECC* **(12)** (2001), pp. 365–379.

[Norton and Salagean 2000a] G. Norton and A. Salagean, "On the Hamming distance of linear codes over a finite chain ring", *IEEE Transactions on Information Theory* **(46)** (2000), pp. 1060–1067.

[Norton and Salagean 2000b] G. Norton and A. Salagean, "On the structure of linear and cyclic codes over a finite chain ring", *AAECC* **(10)** (2000), pp. 489–506.

[Ozbudak and Sole 2007] F. Ozbudak and P. Sole, "Gilbert-Varshamov type bounds for linear codes over finite chain rings", *Advances in Mathematics of Communications* **(1)** (2007), pp. 99–109.

[Siap and Ray-Chaudhuri 2000] I. Siap and D. Ray-Chaudhuri, "New linear codes over F_3 and F_5 and improvements on bounds.", *Design, Codes and Cryptography.* **(21)** (2000), pp. 223–233.

ralfaro@umflint.edu   *Mathematics Department, University of Michigan–Flint, Flint, MI 48502, United States*

stbennet@umflint.edu   *Mathematics Department, University of Michigan–Flint, Flint, MI 48502, United States*

joshuaha@umflint.edu   *Mathematics Department, University of Michigan–Flint, Flint, MI 48502, United States*

cthornbu@umflint.edu   *Mathematics Department, University of Michigan–Flint, Flint, MI 48502, United States*

Bounds for Fibonacci period growth

Chuya Guo and Alan Koch

(Communicated by Arthur T. Benjamin)

We study the Fibonacci sequence mod n for some positive integer n. Such a sequence is necessarily periodic; we introduce a function $Q(n)$ which gives the ratio of the length of this period to n itself. We compute $Q(n)$ in certain cases and provide bounds for it which depend on the nature of the prime divisors of n.

1. Introduction

Any sequence of integers which satisfy a recurrence relation becomes periodic when reduced modulo n for any positive integer n. Here we investigate the behavior of the length of the period of the Fibonacci sequence modulo n, $n \in \mathbb{Z}^+$. We shall denote this length by $k(n)$, and call it the Fibonacci period mod n. This sequence was first studied by Wall [1960], and since that time many interesting properties of $k(n)$ have been discovered. For example, if m and n are relatively prime then $k(mn) = \mathrm{lcm}(k(m), k(n))$. Thus it seems reasonable to compute $k(n)$ using its prime power factorization: if $n = p_1^{e_1} \cdots p_t^{e_t}$ then $k(n) = \mathrm{lcm}(k(p_1^{e_1}), \ldots k(p_t^{e_t}))$. Assuming one can find the prime power factorization of n in a reasonable amount of time, two problems remain: there is no known formula for $k(p)$ when p is a prime, and though it is generally believed that $k(p^e) = p^{e-1}k(p)$ it has not been proven.

While $\{k(n)\}$ is not an increasing sequence it tends to grow as n does, in the sense that for $\{a_r\}$ an infinite sequence of positive integers we have that $\{k(a_r)\}$ is unbounded. A study of $k(n)$ makes it clear that certain such $\{a_r\}$ lead to period lengths which blow up much faster that others — two extreme examples being $\{2 \cdot 5^r\}$ and $\{F_r\}$.

In order to study the growth rates of such sequences we introduce the Fibonacci Q-function. The notion of this ratio is implicit in previous words — often n is compared to $k(n)$; for example, see [Coleman et al. 2006, Fig. 1] for a plot of $k(p)$

versus p, p prime. For all n we define $Q(n)$ to be the ratio of period length to modulus. We will see that for all $r \in \mathbb{Z}^+$ we have $Q(2 \cdot 5^r) = 6$ and $Q(F_r) \to 0$, where F_r is the rth Fibonacci number. Using new and known results about Fibonacci periods we will compute $Q(n)$ for many classes of integers, including some classes of prime numbers as well as Fibonacci and Lucas numbers. Additionally, viewing Q as a function on positive integers, we will show that the image of Q, denoted $\mathcal{Q}$, is contained in but not equal to $[0, 6] \cap \mathbb{Q}$. It turns out that $\mathcal{Q}$ is infinite, as is its complement in $[0, 6] \cap \mathbb{Q}$. It is interesting as well to note that $1, 2, 3, 4,$ and 6 are all in $\mathcal{Q}$ – $Q(24) = 1$, $Q(60) = 2$, $Q(20) = 3$, $Q(5) = 4$, and $Q(10) = 6$ – but $5 \notin \mathcal{Q}$.

We also establish bounds for $Q(n)$ which depend on the number t_1 of prime factors of n whose last digits are either 3 or 7. The bounds are given explicitly and depend on $\gcd(10, n)$. These bounds are useful when n has a small number of such factors, but the bound increases with t_1 and eventually exceeds 6. We also examine the *unit disk preimage* $U = \{n \in \mathbb{Z}^+ \mid Q(n) < 1\}$, showing it is closed under multiplication by relatively prime numbers, and give a sufficient (but not necessary) criterion for a number to be in U.

We finish with a number of open questions concerning values of $Q(n)$ and topological properties of $\mathcal{Q}$. Perhaps the most famous conjecture concerning k is that $k(p^e) = p^{e-1}k(p)$ when p is a prime. Using the Q-function the conjecture becomes $Q(p^e) = Q(p)$, and it is no surprise that our conjecture is equivalent the one on k. We will show that this conjecture holds in the case where p is a Fibonacci prime. We will see that the answers to many of our other open questions on values of $Q(n)$ will follow immediately from other famous conjectures. For example, we conjecture that there are infinitely many primes p such that $Q(p) = 2(1 + 1/p)$ and infinitely many primes p such that $Q(p) = 1 - 1/p$. With the exception of $p = 5$ we have $Q(p) \leq 2(1 + 1/p)$, and if there are an infinite number of Mersenne Primes whose last digit is 3 or 7 then there are an infinite number of points where we have equality. Also, if $p \equiv \pm 1 \pmod{10}$ then $Q(p) \leq 1 - 1/p$, and if there are an infinite number of Sophie Germaine primes then U contains infinitely many primes (primes which, in fact, are safe rather than Sophie Germaine).

2. Preliminaries

Consider the recurrence relation

$$a_n = a_{n-1} + a_{n-2}, \ n \geq 2.$$

If we set $a_0 = 0$ and $a_1 = 1$ we obtain the Fibonacci sequence, which we denote by $\{F_n\}$. Each F_n is a Fibonacci number, and if F_n is prime then it is called a *Fibonacci prime*. Examples of Fibonacci primes include $2, 3, 5, 13,$ and 89. It is

well known (see, for example, [Hardy and Wright 1954, Theorem 179 (iv)]) that if $m \mid n$ then $F_m \mid F_n$, hence if F_q is a Fibonacci prime, $q > 4$ then q is also prime. It is conjectured that there are an infinite number of Fibonacci primes. We may extend the Fibonacci sequence to negative indices. If we define $F_{-n} = (-1)^{n-1} F_n$, $n \geq 1$ then $F_n = F_{n-1} + F_{n-2}$ for all integers n.

While the Fibonacci sequence is the focus here, we will also need to consider the Lucas sequence $\{L_n\}$, obtained by $L_0 = 2$, $L_1 = 1$, and the recurrence relation above. Note that $L_n = F_{n+1} - F_{n-1}$. In a manner similar to F_n we may define $L_{-n} = (-1)^n L_n$, and we may then extend the identity $L_n = L_{n-1} + L_{n-2}$ to all $n \in \mathbb{Z}$.

For any $n \geq 0$ we define $k(n)$ to be the smallest positive integer such that

$$F_{k(n)} \equiv 0 \text{ and } F_{k(n)+1} \equiv 1 \pmod{n}.$$

The number $k(n)$ is called the *Fibonacci period* mod n. Notice that this term is appropriate because, mod n, the sequence of Fibonacci numbers is necessarily a periodic sequence mod n, i.e. $F_{k(n)+i} \equiv F_i \pmod{n}$. Periodicity is guaranteed since there are only n^2 possibilities for F_i and F_{i+1}, and if

$$F_i \equiv F_j \pmod{n} \qquad \text{and} \qquad F_{i+1} \equiv F_{j+1} \pmod{n},$$

then it is easy to show (by repeated subtraction) that

$$F_{i-j} \equiv 0 \pmod{n} \qquad \text{and} \qquad F_{i-j+1} \equiv 1 \pmod{n}.$$

Example 2.1. Modulo 2 the Fibonacci sequence is $0, 1, 1, 0, 1, 1, \ldots$ and thus $k(2) = 3$. Modulo 3 we have the sequence

$$0, 1, 1, 2, 0, 2, 2, 1, 0, 1, \ldots$$

hence $k(3) = 8$. The sequence mod 4 is

$$0, 1, 1, 2, 3, 1, 0, 1, \ldots$$

and $k(4) = 6$. If we take the sequence mod 5 we get

$$0, 1, 1, 2, 3, 0, 3, 3, 1, 4, 0, 4, 3, 2, 0, 2, 2, 4, 1, 0, 1, \ldots$$

and thus $k(5) = 20$.

There is no known formula for $k(n)$, however many of its properties are known. The result below summarizes the facts that we will need. Proofs of each can be found in [Renault 1996], although many are first described in Wall's original paper.

Lemma 2.2. *Let $p, n \in \mathbb{Z}^+$, p prime. The length of a Fibonacci period satisfies all of the following.*

(1) *For $n > 2$, $k(n)$ is even.*

(2) *If $p \equiv \pm 1 \pmod{10}$ then $k(p) \mid p - 1$.*

(3) *If $p \equiv \pm 3 \pmod{10}$ then $k(p) \mid 2(p+1)$ and $k(p) \nmid (p+1)$.*

(4) *Let t be the largest integer such that $k(p^t) = k(p)$. Then for all $e \geq t$, $k(p^e) = p^{e-t}k(p)$.*

(5) *For $\gcd(m, n) = 1$ we have $k(mn) = \operatorname{lcm}(k(m), k(n))$.*

(6) *Suppose $n \geq L_t$. Then $k(n) \geq 2t$.*

(7) *Let $a(n)$ be the smallest positive integer such that $F_{a(n)} \equiv 0 \pmod{n}$, and let $b(n)$ be the number of indices $1 \leq i \leq k(n)$ with $F_i \equiv 0 \pmod{n}$. Then $k(n) = a(n)b(n)$ and $b(n) = 1, 2$, or 4.*

(8) *For all $e \geq 1$, $b(p^e) = b(p)$.*

Throughout the paper, the numbers $a(n)$ and $b(n)$ will be as described above.

It was conjectured by Wall that $k(p^2) = pk(p)$, and it is only a slight generalization to conjecture $k(p^e) = p^{e-1}k(p)$, i.e. that $t = 1$ in the fourth statement above. This is the most famous conjecture related to the study of $k(n)$, and we will refer to this as Wall's Conjecture. As mentioned in the introduction, we will show it is true for Fibonacci primes.

Notice that, for any given prime, Lemma 2.2 (6) implies that one need only check a finite number of exponents to establish Wall's Conjecture for a given prime. For example, one can directly compute $k(17) = 36$. To show

$$k(17^e) = 17^{e-1}k(17) = 36 \cdot 17^{e-1},$$

notice that $L_{19} = 9349$. Then for each $n \geq 9349$ we have $k(n) \geq 2 \cdot 19 = 38$, hence if $17^t \geq 9349$ then $k(17^t) \geq 38 > 36 = k(17)$. Thus since $17^2 = 289$, $17^3 = 4913$, and $17^4 = 83521 > 9349$ one needs to check that $k(17^2)$ and $k(17^3)$ are not 36 — they are, in fact, 612 and 10404, as expected.

We are now ready to formally introduce the tool we will use to study Fibonacci periods mod n.

Definition 2.3. *For any $n \in \mathbb{Z}^+$ let*

$$Q(n) = \frac{k(n)}{n}.$$

Q is called the Fibonacci Q-Function.

Example 2.4. $Q(2) = 3/2$, $Q(3) = 8/3$, $Q(4) = 3/2$, and $Q(5) = 4$. Notice that $Q(2) = Q(4)$ and that $Q(5)$ is much larger than the others – we will discuss both of these observations later.

Example 2.5. For all e, $Q(17^e) = Q(17) = 36/17$.

3. Points

The rest of the paper is an investigation of the properties of $Q(n)$. In the previous section we computed $Q(n)$ for numbers at most 5 (note that $Q(1) = 1$). Here we will look at certain classes of numbers for which we can compute $Q(n)$ exactly. Perhaps the easiest numbers to study are the Fibonacci numbers themselves.

Example 3.1. The numbers 8 and 11 are both Fibonacci numbers: $8 = F_6$ and $13 = F_7$. Clearly $a(8) = 6$ since $F_6 \equiv 8 \equiv 0 \pmod 8$ and since $F_n < 8$ for $1 \leq n \leq 5$ this is the smallest Fibonacci number for which we get zero mod 8. By Lemma 2.2 (7)–(8) we know that $k(8) = 6$, 12, or 24. Since $F_7 = F_6 + F_5$ we have

$$F_6 + 1 < F_7 < 2 \cdot F_6$$

and hence F_7 is not 1 mod 8, i.e. $k(8) \neq 6$. Finally, note

$$F_{-6} = (-1)^5 F_6 = -8 \equiv 0 \pmod 8$$
$$F_{-5} = (-1)^4 F_5 = F_5 \equiv F_7 \pmod 8$$

since $F_7 = F_5 + F_6 \equiv F_5 \pmod{F_6}$. Thus the sequence is periodic, period 12, so $k(8) = 12$. Similarly, $a(13) = 7$ and thus $k(13) = 7$, 14, or 28. Since

$$F_7 + 1 < F_8 < 2 \cdot F_9$$

we know $k(13) \neq 7$. In this case note that $F_{-7} = F_7$ and $F_{-6} = -F_6$ which is not congruent to $F_8 \pmod{11}$ since then $F_6 \equiv F_8 \equiv -F_6 \pmod{11}$ and this implies $2 \cdot F_6 = F_7$ which cannot occur. Thus $k(13) = 28$.

We have $Q(8) = 3/2$ and $Q(13) = 28/13$. We can use the results on these Fibonacci numbers to help us compute $Q(n)$ for other numbers. For example,

$$Q(104) = \frac{k(104)}{104} = \frac{\mathrm{lcm}(k(8), k(13))}{104} = \frac{\mathrm{lcm}(2 \cdot 6, 4 \cdot 7)}{104} = \frac{21}{26}.$$

Note that the Q-function has an interesting property on this product:

$$Q(104) = Q(8)Q(13)/4.$$

More generally, we have

Theorem 3.2. *Let $n \geq 3$ be an integer.*

(1) *If n is even, then $Q(F_n) = 2n/F_n$. If n is odd, $Q(F_n) = 4n/F_n$.*

(2) *If n is even, then $Q(L_n) = 4n/F_n$. If n is odd, $Q(L_n) = 2n/F_n$.*

(3) *If $F_q = p$ is an odd prime, then $Q(p^e) = Q(p) = 4q/p$. Also, $Q(2^e) = 3/2$.*

(4) *Let* $\{n_1, n_2, \ldots, n_t\}$ *be a sequence of positive integers such that* $\gcd(n_i, n_j) \leq 2, t > 1$. *Then*

$$Q(F_{n_1} F_{n_2} \cdots F_{n_t}) = 4^{1-t} Q(F_{n_1}) \cdots Q(F_{n_t}).$$

Proof. A proof of Theorem 3.2 (1) is omitted, however it may be obtained by generalizing the previous example. Also, (2) is similar to (1) and is also omitted.

For Theorem 3.2 (3), by Lemma 2.2 (4) we know that $k(p^e) = p^{e-t} k(p)$, where t is the largest positive integer such that $k(p^t) = k(p)$. Since $p^e > p = F_q$ for all $e > 1$ it is clear that $a(p^e) > a(p)$. Since $b(p^e) = b(p)$ we see that $k(p^t) > k(p)$ unless $t = 1$. Thus

$$Q(p^e) = \frac{p^{e-1} k(p)}{p^e} = \frac{k(p)}{p} = \frac{4q}{p}.$$

To show that $Q(2^e) = 3/2$, it suffices to show that $k(2^e) > 3$ for all $e > 1$. But since $2^e > 3 = L_2$ we know that $k(2^e) \geq 2 \cdot 3$ and we are done.

We now prove Theorem 3.2 (4). Note that $\gcd(n_i, n_j) \leq 2$ implies that

$$\gcd(F_{n_i}, F_{n_j}) = 1$$

since

$$\gcd(F_{n_i}, F_{n_j}) = F_{\gcd(n_i, n_j)}.$$

Suppose $n_1, n_2, \ldots, n_s$ are all even and $n_{s+1}, n_{s+2}, \ldots, n_t$ are all odd. Then we have

$$
\begin{aligned}
Q(F_{n_1} F_{n_2} \cdots F_{n_t}) &= \frac{\operatorname{lcm}(k(F_{n_1}), \ldots, k(F_{n_t}))}{F_{n_1} F_{n_2} \cdots F_{n_t}} \\
&= \frac{\operatorname{lcm}(2n_1, 2n_2, \ldots, 2n_s, 4n_{s+1}, \ldots, 4n_t)}{F_{n_1} F_{n_2} \cdots F_{n_t}} \\
&= 4\left(\frac{\operatorname{lcm}(n_1/2, n_2/2, \ldots, n_s/2, n_{s+1}, \ldots, n_t)}{F_{n_1} F_{n_2} \cdots F_{n_t}}\right) \\
&= 4\frac{n_1 n_2 \cdots n_t}{2^s F_{n_1} F_{n_2} \cdots F_{n_t}},
\end{aligned}
$$

the last equality since the set $\{n_1/2, n_2/2, \ldots, n_s/2, n_{s+1}, \ldots, n_t\}$ is pairwise relatively prime. Note that $n_i/F_{n_i} = Q(F_{n_i})/2$ for $i \leq s$ and $n_i/F_{n_i} = Q(F_{n_i})/4$ otherwise. Thus

$$
\begin{aligned}
Q(F_{n_1} F_{n_2} \cdots F_{n_t}) &= \frac{4}{2^s}((Q(F_{n_1})/2) \cdots (Q(F_{n_s})/2))((Q(F_{n_{s+1}})/4) \cdots (Q(F_{n_t})/4)) \\
&= \frac{4}{2^s} \frac{1}{2^s}(Q(F_{n_1}) \cdots Q(F_{n_s})) \frac{1}{4^{t-s}}(Q(F_{n_{s+1}}) \cdots Q(F_{n_t})) \\
&= \frac{4}{4^t} Q(F_{n_1}) \cdots Q(F_{n_t}) = 4^{1-t} Q(F_{n_1}) \cdots Q(F_{n_t}). \qquad \square
\end{aligned}
$$

Remark 3.3. Note that [Theorem 3.2](4) does not extend to powers of Fibonacci numbers since $Q(45) = 8/3$ and $(1/4)Q(3)Q(5) = 8$. Nonetheless, it remains easy to compute $Q(n)$ whenever n is a product of powers of relatively prime Fibonacci numbers.

We will now try to compute $Q(p)$ when p is in one of a few well-known classes of primes. Let $p \equiv \pm 3 (\mathrm{mod}\ 10)$. From [Lemma 2.2](3) we see that $k(p) \mid 2(p+1)$. To find a class of primes where we can explicitly compute $Q(p)$ we can consider primes such that $p + 1$ has few divisors. Thus it is natural to consider Mersenne primes, primes of the form $2^q - 1$ for some q. It is well known that q must be prime for $2^q - 1$ to be prime: see, for example, [Rosen 2000, Theorem 7.11]. On the other hand, if $p \equiv \pm 1 (\mathrm{mod}\ 10)$ then $k(p) \mid (p-1)$, so any prime of this form that also satisfies $p = 2q + 1$, q prime will have few divisors.

Theorem 3.4. *Let p be prime.*

(1) *If p is a Fibonacci prime, say $p = F_q$ with $q > 4$ then $Q(p) = 4q/p$.*

(2) *If p is a Mersenne prime, $p = 2^q - 1$, such that $q \equiv 3 (\mathrm{mod}\ 4)$ then $Q(p) = 2(1 + 1/p) = 2^{q+1}/p$.*

(3) *If p is a safe prime i.e. $p = 2q + 1$ for some Sophie Germaine prime q such that $q \equiv -1 (\mathrm{mod}\ 10)$ then $Q(p) = 1 - 1/p = 2q/p$.*

Proof. Notice that [Theorem 3.4](1) is a special case of [Theorem 3.2](1) and (3).

We now prove [Theorem 3.4](2). If $q = 4s + 3$ then since the last digit of 6^s is 6 we get
$$p = 2^{4s+3} - 1 = 16^s \cdot 8 - 1 \equiv 6^s \cdot 8 - 1 \equiv 47 \equiv 7 (\mathrm{mod}\ 10)$$
and so $k(p) \mid 2(p+1)$, i.e. $k(p) \mid 2^{q+1}$. But $k(p) \nmid (p+1)$, i.e. $k(p) \nmid 2^q$, thus $k(p) = 2^{q+1}$ and $Q(p) = 2(1 + 1/p) = 2^{q+1}/p$.

Finally, if $p = 2q + 1$ with $q \equiv -1 (\mathrm{mod}\ 10)$ then $p \equiv 2(-1) + 1 \equiv -1 (\mathrm{mod}\ 10)$, thus $k(p) \mid p - 1$. But $p - 1 = 2q$, so $k(p) = 1, 2, q,$ or $2q$. Since $k(p)$ is even and $k(p) > 2$ we must have $k(p) = 2q$, hence $Q(p) = 1 - 1/p = 2q/p$ and [Theorem 3.4](3) is proven. □

Example 3.5. The largest known Mersenne prime of the form above is
$$M_{42} = 2^{25964951} - 1.$$

The Fibonacci Q-function of this 7,816,230-digit number is
$$Q(M_{42}) = \frac{2^{25964952}}{2^{25964951} - 1} \approx 2 + 10^{-7816230}.$$

This is the largest prime for which we know $Q(p)$, and we do not know a prime p such that $Q(p)$ is closer to 2 than $Q(M_{42})$. (We do, however conjecture that for all $\varepsilon > 0$ there is a prime p within ε of 2.)

We now investigate certain values of the Q function.

Theorem 3.6. *Let $r \geq 1$. Then*

(1) $Q(n) = 1$ *if and only if $n = 24 \cdot 5^{r-1}$.*

(2) $Q(n) = 3/2$ *if $n = 10^{r+2}$.*

(3) $Q(n) = 6$ *if and only if $n = 2 \cdot 5^r$.*

Proof. The "if" portions of each of these can be determined by direct calculations since each prime factor is a Fibonacci prime. For example,

$$Q(10^n) = \frac{\text{lcm}(k(2^n), k(5^n))}{10^n} = \frac{\text{lcm}(2^{n-1} \cdot 3, 5^{n-1} \cdot 20)}{10^n} = \frac{2^{n-1} \cdot 3 \cdot 5^n}{2^n 5^n} = \frac{3}{2}$$

establishes Theorem 3.6 (2). The "only if" portions of Theorem 3.6 (1) and (3) follow from the results presented in [Fulton and Morris 1969/1970] and [Brown 1992] respectively. $\square$

In fact, [Brown 1992] proves something stronger: the author shows that $k(n) \leq 6n$ with equality if and only if $n = 24 \cdot 5^{r-1}$. This will be useful when we construct bounds for $Q(n)$ in the next section.

Note that there is no "only if" in Theorem 3.6 (2) since, for example, $Q(2) = 3/2$. It would be interesting to be able to describe the set $Q^{-1}(3/2)$.

Having determined some of the values of Q, it is worth describing certain rational numbers in $[0, 6]$ which are not values of Q. Clearly $Q(n) \neq 0$ for all n since $k(n) \geq 1$ for all $n \in \mathbb{Z}^+$. The following gives an infinite number of other rationals in this interval which are not values of Q.

Theorem 3.7. *Let n be a positive integer. Then*

(1) $Q(n) \neq 5$.

(2) *For each Fibonacci prime p, $Q(n) \neq \frac{t}{p^j u}$ for any t, u relatively prime to p and $j \geq 2$.*

Proof. Let n be the smallest positive integer so that $Q(n) = 5$. Then $k(n) = 5n$, and since $n > 2$ we know that $k(n)$ is even, hence $5n$ is be even. Write $n = 2^i s$, s odd, $i \in \mathbb{Z}^+$. Then

$$5 \cdot 2^i s = k(n) = \text{lcm}(2^{i-1} \cdot 3, k(s)).$$

Since 3 divides the right-hand side we see that $3 \mid s$. Write $s = 3^j t$, t odd, $j \in \mathbb{Z}^+$, $3 \nmid t$. Then

$$5(2^i \cdot 3^j \cdot t) = k(n) = \text{lcm}(2^{i-1} \cdot 3, 3^{j-1} \cdot 8, k(t)).$$

From this we see that $5t \mid k(t)$. Note that if $5t < k(t)$ then $k(t) \geq 10t > 6t$ which cannot occur by [Brown 1992]. Thus $k(t) = 5t$. But $t < n$, contradicting the minimality of n. Thus Theorem 3.7 (1) is proved.

For Theorem 3.7 (2), the trick is that the denominator, before cancellation, is always n. Suppose $Q(n) = t/p^j u$. Then $p^j u \mid n$, so we can write $n = p^i um$ for some $i \geq j$ and $\gcd(p, um) = 1$. Then

$$\frac{t}{p^j u} = Q(n) = \frac{\operatorname{lcm}(k(p^i), k(um))}{p^i um} = \frac{\operatorname{lcm}(p^{i-1} k(p), k(um))}{p^i um}.$$

Cross-multiplying gives

$$t p^i um = p^j u \operatorname{lcm}(p^{i-1} k(p), k(um)).$$

The right-hand side is clearly divisible by p^{i+j-1}, however since $i + j - 1 \geq i + 1$ and $p \nmid mtu$ we see that the left hand side is not divisible by p^{i+j-1}, a contradiction. Thus such an n cannot occur and we are done. □

4. Bounds

In general, it is no easier to compute $Q(n)$ than $k(n)$. However, it is more natural to describe bounds on the Q-function than it is on the period. For example, the statement $k(n) \leq 6n$ can be stated more naturally as $Q(n) \leq 6$. This fact, together with Lemma 2.2 (6) gives

Proposition 4.1. $L_t/(2n) \leq Q(n) \leq 6$, where $n \geq L_t$.

To show that these are the best bounds possible in general we have

Corollary 4.2. $\sup\{Q(n)\} = 6$ and $\inf\{Q(n)\} = 0$.

Proof. The sequence $\{Q(2 \cdot 5^{r-1})\}$ is a sequence of 6's and hence converge to 6. The sequence $\{Q(F_n)\}$ converges to 0. □

We can get a better upper bound if we restrict ourselves to certain classes of integers. The natural place to start is to find an upper bound for Q restricted to primes. We already know the result in this case.

Lemma 4.3. *Let p be prime.*

(1) *$Q(2) = 3/2$ and $Q(5) = 4$.*

(2) *Suppose $p \equiv \pm 1 \pmod{10}$. Then $Q(p) \leq 1 - 1/p$.*

(3) *Suppose $p \equiv \pm 3 \pmod{10}$. Then $Q(p) \leq 2(1 + 1/p)$.*

Proof. Of course, Lemma 4.3 (1) was stated before — is included here only for completeness. Lemma 4.3 (2) and (3) follow immediately from Lemma 2.2 (2) and (3). □

Note that $p = 3$ and 11 give the largest possible value for $Q(p)$ under the conditions in Lemma 4.3 (3) and (2), respectively. Combining those two parts gives

Corollary 4.4. *For $p \neq 5$ a prime $Q(p) \leq 2(1 + 1/p)$.*

We now consider powers of primes. With the exception of $p = 5$ there is a universal bound for such numbers.

Lemma 4.5. *For any prime $p \neq 5$ we have $Q(p^e) \leq Q(p) \leq 8/3$; furthermore $Q(5^e) = Q(5) = 4$.*

Proof. If $p = 2$ then $Q(p) = 3/2 \leq 8/3$. Furthermore, since 2 is a Fibonacci prime we know $Q(2^e) = Q(2) = 3/2$. Likewise, $Q(3^e) = Q(3) = 8/3$ and $Q(5^e) = Q(5) = 4$. We shall assume $p \geq 7$. Then $Q(p) \leq 2(1 + 1/p) \leq 2(1 + 1/7) = 16/7 < 8/3$, so it remains to show $Q(p^e) \leq Q(p)$ for $e > 1$.

Suppose $a = a(p)$ and $a' = a(p^e)$. Since $F_{a'} \equiv 0 \pmod{p^e}$ we know $F_{a'} \equiv 0 \pmod{p}$ and hence a' is a multiple of a. We claim that $F_{p^{e-1}a} \equiv 0 \pmod{p^e}$, and hence $a' \leq p^{e-1}a$. Applying the well-known identity

$$F_{mn} = \sum_{i=1}^{m} \binom{m}{i} F_i F_n^i F_{n-1}^{m-i}$$

we have

$$F_{p^{e-1}a} = \sum_{i=1}^{p^{e-1}} \binom{p^{e-1}}{i} F_i F_a^i F_{a-1}^{p^{e-1}-i}.$$

For $1 \leq i \leq p^{e-1}$ we clearly have $p^i \mid F_a^i$. If we write $i = p^f j$, $p \nmid j$ it can be shown that $p^{e-f-1} \mid \binom{p^{e-1}}{i}$. Thus $p^{e-f-1+i}$ divides the ith term in the series above. Since $i > 0$ we have $i \geq f + 1$ and so each term is divisible by p^e, hence $F_{p^{e-1}a} \equiv 0 \pmod{p^e}$.

Thus,

$$Q(p^e) = \frac{a'b(p^e)}{p^e} \leq \frac{(p^{e-1}a)b(p)}{p^e} = \frac{k(p)}{p} = Q(p). \qquad \square$$

Next, we look at how Q behaves with relatively prime numbers.

Lemma 4.6. *For $\gcd(m, n) = 1$ we have $Q(mn) \leq Q(m)Q(n)$. If furthermore $m, n > 2$ then $Q(mn) \leq \frac{1}{2}Q(m)Q(n)$.*

Proof. For m and n relatively prime we have

$$Q(mn) = \frac{\text{lcm}(k(m), k(n))}{mn} = \frac{k(m)k(n)}{mn \gcd(k(m), k(n))}$$

$$= \frac{1}{\gcd(k(m), k(n))} Q(m)Q(n) \leq Q(m)Q(n).$$

If $m, n > 2$ then $k(m)$ and $k(n)$ are both even, hence $\gcd(k(m), k(n)) \geq 2$ and we are done. $\qquad \square$

Before continuing to generalize n, we note that this lemma gives us insight into the structure of the *unit disk preimage*.

Corollary 4.7. *Let* $U = \{n \in \mathbb{Z}^+ \mid Q(n) < 1\}$. *Then* U *is infinite, and is closed under multiplication by relatively prime elements.*

Proof. Certainly $p \in U$ for all primes $p \equiv \pm1 \pmod{10}$, and by Lemma 4.5 we have $p^e \in U$ for all e hence U is infinite. (One could obtain a different proof using Dirichlet's Theorem on Primes in Arithmetic Progressions.) That U is closed under multiplication by relatively prime elements is clear from the previous result. $\square$

Finally, we are ready to consider arbitrary n. For the remainder of the section we write

$$n = 2^r 5^s p_1^{r_1} \cdots p_t^{r_t}, \ m = p_1^{r_1} \cdots p_t^{r_t}, \ \gcd(10, m) = 1$$

and we let

$$t_0 = \#\{i \mid p_i \equiv \pm1 \pmod{10}\} \text{ and } t_1 = \#\{i \mid p_i \equiv \pm3 \pmod{10}\}.$$

Proposition 4.8. *We have* $Q(m) \leq Q(p_1) \cdots Q(p_t)/2^{t-1}$.

Proof. Immediate from Lemmas 4.5 and 4.6. $\square$

Theorem 4.9. *We have* $Q(m) \leq 2^{2t_1 - t_0 + 1}/3^{t_1}$. *Furthermore,*

$$Q(5^s m) \leq \frac{2^{2t_1 - t_0 + 2}}{3^{t_1}}, \ Q(2^r m) \leq \frac{2^{2t_1 - t_0}}{3^{t_1 - 1}}, \ and \ Q(2^r 5^s m) \leq \frac{2^{2t_1 - t_0 + 1}}{3^{t_1 - 1}}.$$

Proof. Since $Q(p_i) \leq 8/3$ when $p_i \equiv \pm3 \pmod{10}$ and $Q(p_i) < 1$ when $p_i \equiv \pm1 \pmod{10}$ we have

$$Q(m) \leq \frac{1^{t_0} \left(\frac{8}{3}\right)^{t_1}}{2^{t-1}} = \left(\frac{8}{3}\right)^{t_1} \frac{1}{2^{t-1}} = \frac{2^{3t_1}}{3^{t_1} \cdot 2^{t_0 + t_1 - 1}} = \frac{2^{2t_1 - t_0 + 1}}{3^{t_1}}.$$

Similarly we have

$$Q(5^s m) \leq \frac{1}{2} Q(5^s) Q(m) = 2 Q(m) \leq \frac{2^{2t_1 - t_0 + 2}}{3^{t_1}}.$$

and

$$Q(2^r m) \leq \frac{3}{2} Q(m) \leq \frac{2^{2t_1 - t_0}}{3^{t_1 - 1}},$$

and

$$Q(2^r 5^s m) \leq 3 Q(m) \leq \frac{2^{2t_1 - t_0 + 1}}{3^{t_1 - 1}}. \qquad \square$$

One can obtain an overall bound by taking the largest of the four expressions.

Corollary 4.10. *For any n we have*

$$Q(n) \leq \frac{2^{2t_1 - t_0 + 1}}{3^{t_1 - 1}}.$$

Notice that this bound is quite significant if n has a lot of primes of the form $p \equiv \pm 1 \pmod{10}$, however there will also be cases where the bound does not provide useful information. For example, if $t_0 = 0$ and $t_1 = 7$ we have $Q(m) \leq 7.5$, which we already knew.

We can use Theorem 4.9 to obtain a sufficient, though not necessary, criterion for a number to be in the *unit disk preimage*.

Corollary 4.11. *If*

$$t_0 \geq t_1 \frac{\ln 4/3}{\ln 2} + 1,$$

then $m \in U$. If

$$t_0 \geq t_1 \frac{\ln 4/3}{\ln 2} + 2,$$

then $5^s m \in U$. If

$$t_0 \geq t_1 \frac{\ln 4/3}{\ln 2} + \ln 3,$$

then $2^r m \in U$. Finally, if

$$t_0 \geq t_1 \frac{\ln 4/3}{\ln 2} + \ln 3 + 1,$$

then $n \in U$.

Proof. Obtained by setting each of the bounds equal to 1 and solving for t_0. For example, if we set

$$Q(m) \leq \frac{2^{2t_1 - t_0 + 1}}{3^{t_1}} \leq 1,$$

we get

$$2^{2t_1 - t_0 + 1} \leq 3^{t_1}$$
$$(2t_1 - t_0 + 1) \ln 2 \leq t_1 \ln 3$$
$$(2 \ln 2 - \ln 3) t_1 + \ln 2 = \ln(4/3) t_1 + \ln 2 \leq t_0 \ln 2$$
$$t_0 \geq t_1 \frac{\ln 4/3}{\ln 2} + 1.$$

The others are similar.   □

Notice that if $t_0 = 0$ and $t_1 = 2$, Theorem 4.9 gives $Q(m) \leq 32/9$. However, it is clear we can do better since $Q(p) = 8/3$ only when $p = 3$. In fact, we have

$$Q(m) \leq \frac{1}{2} \frac{8}{3} \frac{16}{7} = \frac{64}{21} < 3.048$$

This leads to a stronger bound.

Theorem 4.12. *For $t_1 > 2$,*

$$Q(m) \leq \frac{2^{6-t_0}}{21} \prod_{i=3}^{t_1} \left(\frac{5i - 12}{5i - 13} \right).$$

Proof. Assume that $3 \leq p_1 < p_2 < \cdots < p_t$. Again if $p_i \equiv \pm 1 \pmod{10}$ then $Q(p_i) \leq 1$. If $p_i \equiv \pm 3 \pmod{10}$ then $Q(p_i) \leq 2(1 + 1/p_i)$. If we write $p_i = 10h \pm 3$ then $h \geq (i/2 - 1)$ since there are at most two such primes between $10z$ and $10(z + 1)$. Thus

$$p_i = 10h \pm 3 \geq 10h - 3 \geq 10(i/2 - 1) - 3 = 5i - 13$$

and hence

$$Q(p_i) \leq 2\left(1 + \frac{1}{5i - 13}\right) = 2\left(\left(\frac{5i - 12}{5i - 13}\right)\right)$$

for $i \geq 3$. Since $Q(p_1) \leq 8/3$, $Q(p_2) \leq 16/7$ we have

$$Q(m) \leq \frac{1}{2^{t-1}} \frac{8}{3} \frac{16}{7} \prod_{i=3}^{t_1} 2\left(\frac{5i - 12}{5i - 13}\right) = \frac{2^7}{2^{t_0 - 1 + t_1} \cdot 21} 2^{t_1 - 2} \prod_{i=3}^{t_1} \left(\frac{5i - 12}{5i - 13}\right)$$

$$= \frac{2^{6-t_0}}{21} \prod_{i=3}^{t_1} \left(\frac{5i - 12}{5i - 13}\right). \qquad \square$$

Here is a table of bounds for m when $t_0 = 0$; similar tables for $2^r 5^s m$ can be constructed. For the second bound, we use $8/3$ and $64/21$ when $t = 1$ and 2 respectively.

t_1	1	2	3	4	5	6	7
First Bound	2.667	3.556	4.741	(over 6)			
Second Bound	2.667	3.048	4.572	5.225	5.660	5.993	(over 6)

We could, with much work, obtain progressively sharper bounds for large t_1 by noticing that our bounds constructed above use the fact that there are at most two primes whose last digit is 3 or 7 between $10z$ and $10(z + 1)$; there may be fewer, e.g. when $z = 2$ or 3.

5. Questions

We conclude with several conjectures and questions. Many of the these relate directly to Wall's conjecture or other well-known questions. We start with the obvious

Conjecture 5.1. $Q(p^e) = Q(p)$ for all primes p.

Notice that this is equivalent to Wall's conjecture since if

$$Q(p^e) = k(p^e)/p^e = k(p)/p = Q(p)$$

then $k(p^e) = p^{e-1}k(p)$.

Theorem 3.7 established that the image of Q , viewed as a function $\mathbb{Z}^+ \to \mathbb{Q}$, does not include numbers which cannot be expressed with denominators divisible by a Fibonacci prime power greater than one. It seems likely that this result extends.

Conjecture 5.2. For any prime p, $Q(n) \neq \frac{t}{p^j u}$ for all n and any t relatively prime to p, and $j \geq 2$.

If Conjecture 5.1 is true, then by Theorem 3.7 (2) so is this one. There is a partial converse.

Proposition 5.3. *If Conjecture 5.2 is true, then Wall's Conjecture is true when either $e \neq 2$ or $p \equiv \pm 1 \pmod{10}$.*

Proof. Suppose Conjecture 5.2 holds. We know Wall's Conjecture holds when $p = 2$, so hereafter we assume $p \neq 2$. We have

$$Q(p^e) = \frac{k(p^e)}{p^e}.$$

Since the denominator, when reduced, can have at most one power of p we see that p^{e-1} must divide the numerator. Thus $k(p^e) \geq p^{e-1}$. If $e \geq 3$ then

$$k(p^e) \geq p^{e-1} \geq p^2 > 2(p+1) \geq k(p)$$

since $p \geq 3$. Thus if $k(p^t) = k(p)$ then $t = 1$ or 2. Finally, if $e \geq 2$ and $p \equiv \pm 1 \pmod{10}$ then

$$p^{e-1} \geq p > p - 1 \geq k(p)$$

and hence $k(p^t) = k(p)$ can only occur if $t = 1$.   □

We saw that the "unit disk preimage" U is closed under multiplication by relatively prime numbers, and Lemma 4.5 can be used to show that U is closed under powers, i.e. $u \in U$ implies $u^i \in U$ for all $i \geq 1$. This suggests that the following may be true.

Conjecture 5.4. If $m, n \in U$, then $mn \in U$.

Note that the converse to this is not true: $Q(3) = 8/3$ and $Q(7) = 16/7$, so $3, 7 \notin U$; however $Q(21) = 16/21$ and hence $21 \in U$. If this conjecture is true then U is a semigroup; furthermore $V := \left\{ n \in \mathbb{Z}^+ \mid Q(n) \leq 1 \right\}$ is a monoid since $Q(1) = 1$.

The final two conjectures are motivated by the empirical observation that $k(p)$ is often $p - 1$ or $2(p+1)$.

Conjecture 5.5. There are an infinite number of primes with $Q(p) = 2(1 + 1/p)$.

If there are an infinite number of Mersenne primes $2^q - 1$ with $q \equiv 3 \pmod 4$ then this conjecture is true.

Conjecture 5.6. There are an infinite number of primes with $Q(p) = (1 - 1/p)$.

If there are an infinite number of Sophie Germaine primes with $q \equiv -1 \pmod{10}$ then this conjecture is true. Alternatively, if one could show that there are an infinite number of length four Cunningham chains of the first kind then the conjecture would be proved.

Finally, viewing Q once again as a function $\mathbb{Z}^+ \to \mathbb{Q}$ we can ask a variety of questions about the image. Let $\mathfrak{Q}$ be the image of Q, and let $I = [0, 6] \cap \mathbb{Q}$. What are the topological properties of $\mathfrak{Q}$ as a subset of I? Is it dense? What are its limit points? We know that 0 is an accumulation point since $\{Q(F_{2k+1})\}$ is a strictly decreasing sequence in $\mathfrak{Q}$ converging to 0. (This also establishes that $\mathfrak{Q}$ is infinite.) Thus $\mathfrak{Q}$ is certainly not a closed set – what is its closure in I? If there are an infinite number of Fibonacci primes then 0 would be a boundary point since $\frac{1}{p^2} \notin \mathfrak{Q}$ for all Fibonacci primes. (In fact we have that every point in $\mathfrak{Q}$ is a boundary point since for each $q \in \mathfrak{Q}$ any each $\varepsilon > 0$ there exists a $t/2^i$, t odd, $i \geq 2$ such that $\left| q - t/2^i \right| < \varepsilon$.) The two previous conjectures would also imply that 1 and 2 are accumulation points. Are there others? Is 6 an isolated point? What about 4? If these points are isolated than $\mathfrak{Q}$ cannot be open in I. A topological study of $\mathfrak{Q}$ seems to be interesting in its own right, as well as a useful way to gain more insight into $k(n)$.

References

[Brown 1992] K. Brown, "The period of Fibonacci sequences modulo m", *American Mathematical Monthly* **99**:3 (1992), 278. problem E3410.

[Coleman et al. 2006] D. A. Coleman, C. J. Dugan, R. A. McEwen, C. A. Reiter, and T. T. Tang, "Periods of (q, r)-Fibonacci sequences and elliptic curves", *Fibonacci Quart.* **44**:1 (2006), 59–70. MR 2006j:11019 Zbl 1133.11009

[Fulton and Morris 1969/1970] J. D. Fulton and W. L. Morris, "On arithmetical functions related to the Fibonacci numbers", *Acta Arith.* **16** (1969/1970), 105–110. MR 40 #4193

[Hardy and Wright 1954] G. H. Hardy and E. M. Wright, *An introduction to the theory of numbers*, Oxford, at the Clarendon Press, 1954. 3rd ed. MR 16,673c Zbl 0058.03301

[Renault 1996] M. Renault, *Properties of the Fibonacci sequence under various moduli*, Master's thesis, Wake Forest University, 1996.

[Rosen 2000] K. H. Rosen, *Elementary number theory and its applications*, Fourth ed., Addison-Wesley, Reading, MA, 2000. MR 2000i:11001 Zbl 0964.11002

[Wall 1960] D. D. Wall, "Fibonacci series modulo m", *Amer. Math. Monthly* **67** (1960), 525–532. MR 22 #10945 Zbl 0101.03201

cguo@agnesscott.edu   *Agnes Scott College, Department of Mathematics,*
                             141 E. College Ave., Decatur, GA 30030, United States

akoch@agnesscott.edu   *Agnes Scott College, Department of Mathematics,*
                             141 E. College Ave., Decatur, GA 30030, United States

Ordering BS(1, 3) using the Magnus transformation

Patrick Bahls, Voula Collins and Elizabeth Heron

(Communicated by Nigel Boston)

Following a similar treatment of the Baumslag–Solitar group BS(1, 2) by Bahls, we modify a transformation developed by Magnus to linearly order the group BS(1, 3) given by the presentation $\langle a, b \mid ab = ba^3 \rangle$. We demonstrate how this same method will fail to admit such a treatment of the groups BS(1, n), $n \geq 4$.

1. Introduction

This paper is heavily based on the work [Bahls 2007] on ordering the Baumslag–Solitar group BS(1, 2). The purpose of this paper is to modify the method of [Bahls 2007] to linearly order the Baumslag–Solitar group BS(1, 3) with presentation

$$P = \langle a, b \mid ab = ba^3 \rangle.$$

Theorem 1.1. *The positive monoid* $\mathrm{BS}^+(1, 3)$ *can be linearly ordered by an order* $\leq$ *compatible with multiplication on the left*:

$$u \leq v \Rightarrow w \cdot u \leq w \cdot v,$$

for all $u, v, w \in \mathrm{BS}^+(1, 3)$. *This order passes to an order on the corresponding group* BS(1, 3) *which is also compatible with multiplication on the left.*

We will also indicate why our method does not apply to any BS(1, n), $n \geq 4$.

We would like to emphasize that it *is* possible to construct an order on BS(1, n) by other means; the significance of this current work lies in its exploration of the methods developed first in [Duchamp and Krob 1990; 1993], and [Duchamp and Thibon 1992], and later modified by Bahls [2007]. Our results here highlight both the potential and the limitations of these methods.

Before proceeding further we briefly motivate our study of orderability.

As in our theorem above, the group G is said to be *left orderable* if it admits a linear ordering $\leq$ satisfying $g_1 \leq g_2 \Rightarrow g \cdot g_1 \leq g \cdot g_2$ for all $g, g_1, g_2 \in G$. *Right*

orderable and *biorderable* groups are defined in a similar fashion; biorderability clearly implies both left and right orderability.

Orderability works well with other algebraic conditions. For instance, it is known that if G is left orderable then it satisfies the *Zero Divisor Conjecture*: the integral group ring $\mathbb{Z}G$ has no nontrivial divisors of zero. Local indicability is another closely related property: G is said to be *locally indicable* if every nontrivial finitely generated subgroup surjects onto $\mathbb{Z}$. Such groups are known to be orderable on one side, but conversely there are examples of groups which are right orderable and not locally indicable. (See [Bergman 1991] for examples; more details can be found in [Rhemtulla 2002].)

The *braid groups* B_n are one such class: they are orderable on one side but not on both, and they are not locally indicable. Dehornoy et al. [2002] give an in-depth treatment of B_n. The braid groups are one of many topologically and geometrically significant classes of groups whose orderability has recently drawn attention. Other examples include various mapping class groups of punctured surfaces with boundary [Short and Wiest 2000] and fundamental groups of 3-manifolds [Boyer et al. 2005].

A sketch of our argument is as follows. As in [Bahls 2007], we will first define the *Magnus transformation*, μ, which maps a generator x_i of a monoid M to $1+x_i$, an element of the algebra $A_k(M)$ of formal power series freely generated by M with coefficients in the integral domain k. A more extensive discussion of the Magnus transformation in various settings can be found in the second section of [Bahls 2007], in Magnus's own classical work with Karrass and Solitar [Magnus et al. 1976], or in [Duchamp and Krob 1990; 1993; Duchamp and Thibon 1992].

Due to the simplicity of the relations governing right-angled Artin groups, Duchamp and his collaborators were able to work with μ without passing to a quotient algebra. In our present case, as in [Bahls 2007], μ is not inherently a homomorphism, so we must force it to be one by introducing a relation on the algebra, thereby passing to a quotient. After defining a normal form for the elements in BS(1, 3) we will apply the new relation to determine a normal form for elements in the algebra. This will allow us to prove that the modified mapping μ is injective and to define a linear order on the elements of BS(1, 3) by linearly ordering their images under μ.

2. The mapping μ and normal forms in BS(1, 3)

Let M be the noncancellative positive monoid $BS^+(1,3)$ given by the presentation

$$\langle a, b \mid ab = ba^3 \rangle.$$

It is known that an element of $\mathrm{BS}^+(1, 3)$ has normal form $b^m a^l$ where m and l are nonnegative integers. Similarly, an element in $\mathrm{BS}(1, 3)$ has normal form $b^m a^l b^{-k}$ where k and m are nonnegative integers and l is any integer not divisible by 3.

Let $A_{\mathbb{Q}}(M)$ be the associative algebra of formal series freely generated by the elements of M with coefficients in $\mathbb{Q} \cup \{\pm\infty\}$.

Note. For reasons that will become clear in the next section we will require infinite coefficients. For any $q \in \mathbb{Q}$ we define $\infty + q = \infty$ and $\infty \cdot q = \pm\infty$, depending on the sign of q, and similarly for $-\infty$. Though strictly speaking addition is not defined on all pairs of elements in our algebra, the computation $-\infty + \infty$ will not arise.

We define $\mu : M \to A_{\mathbb{Q}}(M)$ by $x \mapsto 1 + x$ for $x \in \{a, b\}$ and extend it in the natural fashion:

$$x_1^{\epsilon_1} x_2^{\epsilon_2} \cdots x_k^{\epsilon_k} \overset{\mu}{\mapsto} (1 + x_1)^{\epsilon_1} (1 + x_2)^{\epsilon_2} \cdots (1 + x_k)^{\epsilon_k}.$$

We then make use of the existence of formal inverses in $A_{\mathbb{Q}}(M)$: given any $x \in M$,

$$(1 + x)^{-1} = 1 - x + x^2 - x^3 + \cdots,$$

and thus we may extend μ to a mapping on $\mathrm{BS}(1, 3)$ by extending the natural mapping

$$\mu(a^{-1}) = 1 - a + a^2 - a^3 + \cdots \quad \text{and} \quad \mu(b^{-1}) = 1 - b + b^2 - b^3 + \cdots.$$

As defined, the map μ is not a homomorphism on $\mathrm{BS}(1, 3)$. In order to ensure that μ preserves the structure of the group we pass to the quotient of $A_{\mathbb{Q}}(M)$ by the image of the relation $ab = ba^3$ under μ. That is, we define

$$\mathscr{A} = A_{\mathbb{Q}}(M)/I,$$

where

$$I = \langle (1+a)(1+b) = (1+b)(1+a)^3 \rangle = \langle ba^2 = -(1/3)a^3 - a^2 - (2/3)a - ba \rangle.$$

Abusing notation, we let μ refer to the composition of the original mapping with this quotient map.

Every element in $\mathscr{A}$ can be placed in *normal form* by successive applications of the two relations

$$ab = ba^3 \quad \text{and} \quad ba^2 = -(1/3)a^3 - a^2 - (2/3)a - ba.$$

Such a normal form will admit additive terms in one of three forms: b^h, a^i, or $b^j a$. (The group relation allows us to move b to the left past a, and the quotient

relation allows us to reduce powers of a that follow at least one b.) Images of group elements under μ take the form

$$(1+b)^m (1+a)^l (1+b)^{-k},$$

where m and k are nonnegative integers and l is an arbitrary integer. Upon expanding these binomials and applying the two above relations, we obtain a normal form

$$1 + \sum_{h=1}^{\infty} \beta_h b^h + \sum_{i=1}^{\infty} \alpha_i a^i + \sum_{j=1}^{\infty} \gamma_j b^j a,$$

for rational numbers β_h, α_i, and γ_j.

3. Injectivity of μ

By passing to $\mathcal{A}$ we have ensured that μ is a homomorphism. However, before we will be able to linearly order the group by ordering its image under μ, we must prove that μ is an embedding of $\mathrm{BS}(1,3)$ into $\mathcal{A}$. To prove that μ is injective we must show that if $g \in \mathrm{BS}(1,3)$ satisfies $g \neq 1$, then $\mu(g) \neq 1$ in $\mathcal{A}$.

We will need the following:

Lemma 3.1. *For $k \in \mathbb{N}$, and $x \in (1, \infty)$, then*

$$\sum_{i=0}^{\infty} \binom{i+k-1}{k-1} \left(1 - \frac{1}{x}\right)^i = x^k.$$

To prove Lemma 3.1 we use the following obvious fact:

Lemma 3.2. *Let $g_k(x) = \left(1/(1-x)\right)^k$. Then*

$$\frac{d^n g_k}{dx^n} = \frac{(n+k-1)!}{(k+1)!} \left(\frac{1}{1-x}\right)^{n+k},$$

so

$$g_k^{(n)}(0) = \frac{(n+k-1)!}{(k+1)!}.$$

Proof of Lemma 3.1. Using the binomial theorem, we see

$$\left(\tfrac{1}{1-x}\right)^k = \sum_{i=0}^{\infty} \binom{i+k-1}{k-1} x^i.$$

But Taylor expansion gives

$$\sum_{i=0}^{\infty} \binom{i+k-1}{k-1} \left(1 - \frac{1}{x}\right)^i = g_k\left(1 - \frac{1}{x}\right)$$

$$= \left(\frac{1}{1 - (1 - (1/x))}\right)^k$$

$$= \left(\frac{1}{(1/x)}\right)^k$$

$$= x^k,$$

and we are done. □

Let $c(z, m)$ be the coefficient of the monoid element m as an additive term in $z \in \mathscr{A}$ written in normal form, and define H to be the image of BS(1, 3) under μ. To prove injectivity we will derive formulas for $c(z, a)$ for any arbitrary $z \in H$.

Proposition 3.3. *The mapping μ is injective.*

Proof. Let $g = b^m a^l b^{-k} \neq 1$. Clearly $c(\mu(g), b) \neq 0$ if $l = 0$ and $m \neq k$. Thus we may assume $l \neq 0$.

Suppose at least one of m or k is equal to 0 and $l \neq 0$. If $m = k = 0$, then $c(\mu(g), a) = l$, so $\mu(g) \neq 1$. If $m = 0$ and $k > 0$, then $c(\mu(g), b) = -k$; if $k = 0$ and $m > 0$, then $c(\mu(g), b) = m$. Therefore no such group elements can be mapped by μ to the identity.

Now let $g = b^m a^l b^{-k}$ where $m, k > 0$ and $l > 0$. We show that $c(\mu(g), a) \neq 0$. Expanding $\mu(b^m a^l b^{-k})$ gives a formal series with additive terms $b^h a^j b^i$ ($h \leq m$, $j \leq l$, and i arbitrarily large), before reducing to normal form. It is not difficult to compute inductively the coefficient on a in such a term once it is reduced to normal form:

Lemma 3.4. *For any $z = b^h a^j b^i$ as above,*

$$c(z, a) = (-1)^{h+i+j} (2/3)^{h+i} .$$

Proof. First apply the group relation $ab = ba^3$ to move all powers of b to the left, resulting in $b^{h+i} a^{j3^i}$.

We now show by induction on s and t that

$$c(b^s a^t, a) = (-1)^{s+t} (2/3)^s.$$

First consider $s = 1$. In the base case $t = 2$,

$$c(ba^2, a) = -2/3,$$

as desired. Suppose inductively we have shown

$$c(ba^t, a) = (-1)^{s+t} (2/3).$$

Then

$$ba^{t+1} = (ba^2)a^{t-1} = -(1/3)a^{t+2} - a^{t+1} - (2/3)a^t - ba^t.$$

Our inductive hypothesis thus gives us coefficient

$$(-1)(-1)^{t+1}(2/3) = (-1)^{[(t+1)+1]}(2/3)$$

on a, as needed.

Now suppose inductively we have shown

$$c(b^s a^t, a) = (-1)^{s+t}\big((2/3)\big)^s,$$

for any $t \geq 2$ and for some fixed s. In the base case (for $s + 1$) $t = 2$, we have

$$b^{s+1}a^2 = b^s(ba^2) = -(1/3)b^s a^3 - b^s a^2 - b^s a - b^{s+1}a.$$

The last two terms contribute no a's, while inductively the first two contribute $(-1/3)\,(-1)^{s+3}\,(2/3)^s$ and $(-1)(-1)^{s+2}\,(2/3)^s$ a's, respectively. Adding these and simplifying yields the desired sum:

$$
\begin{aligned}
(-1/3)\,(-1)^{s+3}\,(2/3)^s + (-1)(-1)^{s+2}\,(2/3)^s &= (-1)^{s+2}\,(1/3 - 1)\,(2/3)^s \\
&= (-1)^{s+3}\,(2/3)\,(2/3)^s \\
&= (-1)^{(s+1)+2}\,(2/3)^{s+1},
\end{aligned}
$$

as needed.

Thus

$$c(z, a) = c(b^{h+i}a^{j3^i}, a) = (-1)^{h+i+j3^i}\,(2/3)^{h+i} = (-1)^{h+i+j}\,(2/3)^{h+i},$$

where the last equality holds since j and $j3^i$ have the same parity.   □

We now claim that

$$c(\mu(b^m a^l b^{-k}), a) = 1 + l + \sum_{h=0}^{m}\sum_{j=1}^{l}\sum_{i=0}^{\infty}(-1)^{h+j}\left(\frac{2}{3}\right)^{h+i}\binom{m}{h}\binom{l}{j}\binom{i+k-1}{k-1}.$$

Indeed, the innermost sum, involving i, considers the contribution made by the terms in $\mu(b^{-k})$ (the formal inverse makes this sum infinite). The next sum, involving j, considers the contribution made by each term from $\mu(a^l)$. The outermost sum, involving h, considers the contribution made by each term from $\mu(b^m)$.

The binomial coefficients represent the coefficients appearing on the terms of the expanded binomials. We obtain $(-1)^{h+j}$ from $(-1)^{h+i+j} \cdot (-1)^i$, the first term arising from Lemma 3.4 and the second from the sign on the term b^i in the infinite formal inverse $(1 + b)^{-k}$. Finally, the term $(2/3)^{h+i}$ appears courtesy of

Lemma 3.4. We now compute, first rearranging and then applying Lemma 3.1 to the innermost sum:

$$c(\mu(b^m a^l b^{-k}), a)$$

$$= 1 + l + \sum_{h=0}^{m} \sum_{j=1}^{l} \sum_{i=0}^{\infty} (-1)^{h+j} \left(\frac{2}{3}\right)^{h+i} \binom{m}{h} \binom{l}{j} \binom{i+k-1}{k-1}$$

$$= 1 + l + \sum_{h=0}^{m} (-1)^h \left(\frac{2}{3}\right)^h \binom{m}{h} \sum_{j=1}^{l} (-1)^j \binom{l}{j} \sum_{i=0}^{\infty} \left(\frac{2}{3}\right)^i \binom{i+k-1}{k-1}$$

$$= 1 + l + \sum_{h=0}^{m} \left(-\frac{2}{3}\right)^h \binom{m}{h} \sum_{j=1}^{l} (-1)^j \binom{l}{j} 3^k$$

$$= 1 + l + \sum_{h=0}^{m} \left(-\frac{2}{3}\right)^h \binom{m}{h} (-3^k)$$

$$= 1 + l + \left(-\frac{2}{3} + 1\right)^m (-3^k)$$

$$= 1 + l + \left(\frac{1}{3}\right)^m (-3^k) = \qquad 1 + l - 3^{k-m}.$$

If either $m > k$ or $m = k$, this yields a nonzero quantity. If $m < k$, there are two situations to consider. If m, l, k do not satisfy $3^{k-m} = l + 1$, then $c(\mu(g), a) \neq 0$. If m, l, k satisfy $3^{k-m} = l + 1$, then $c(\mu(g), a) = 0$, but since $k > m$ we know that $c(\mu(g), b) = -k + m \neq 0$, and thus $\mu(g)$ is still not the identity.

Finally, we compute $c(\mu(b^m a^l b^{-k}), a)$ when $m, k > 0$ and $l < 0$. Arguing as in the case $l > 0$, we obtain a similar formula for this coefficient, which reduces nicely by applying Lemma 3.1 once more:

$$c(\mu(b^m a^l b^{-k}), a)$$

$$= -l + \sum_{h=0}^{m} \sum_{j=1}^{\infty} \sum_{i=0}^{\infty} (-1)^{h+i+j} (-1)^i (-1)^j \left(\frac{2}{3}\right)^{h+i} \binom{m}{h} \binom{j+l-1}{l-1} \binom{i+k-1}{k-1}$$

$$= -l + \sum_{h=0}^{m} \sum_{j=1}^{\infty} \sum_{i=0}^{\infty} (-1)^{h+2i+2j} \left(\frac{2}{3}\right)^{h+i} \binom{m}{h} \binom{j+l-1}{l-1} \binom{i+k-1}{k-1}$$

$$= -l + \sum_{h=0}^{m} (-1)^h \left(\frac{2}{3}\right)^h \binom{m}{h} \sum_{j=1}^{\infty} \binom{j+l-1}{l-1} \sum_{i=0}^{\infty} \left(\frac{2}{3}\right)^i \binom{i+k-1}{k-1} - \sum_{j=1}^{\infty} \binom{j+l-1}{l-1}$$

$$= -l + \sum_{j=1}^{\infty} \binom{j+l-1}{l-1} \left(\sum_{h=0}^{m} \left(-\frac{2}{3}\right)^h \binom{m}{h} \sum_{i=0}^{\infty} \left(\frac{2}{3}\right)^i \binom{i+k-1}{k-1} - 1\right)$$

$$= -l + \sum_{j=1}^{\infty} \binom{j+l-1}{l-1} \left(\left(-\frac{2}{3} + 1 \right)^m (-(3^k)) - 1 \right)$$

$$= -l + (3^{k-m} - 1) \sum_{j=1}^{\infty} \binom{j+l-1}{l-1}.$$

Since $\sum_{j=1}^{\infty} \binom{j+l-1}{l-1} = \infty$, the coefficient on a is either ∞ or $-\infty$ if $m \neq k$. In this case $c(\mu(g), a) \neq 0$, so $\mu(g)$ is not the identity in $\mathcal{A}$. Finally, if $m = k$, then $c(\mu(g), a) = l \neq 0$ so that in this case too $\mu(g)$ is not the identity.

As we have now shown that $\mu(g) \neq 1$ for all $1 \neq g \in \mathrm{BS}(1, 3)$, our mapping μ is injective. $\square$

4. Ordering BS(1, 3)

Using a method like that in [Bahls 2007], we will define an order on our group H by defining a *strict positive cone C* of the algebra which satisfies the following four properties:

(C1) $C \cdot C \subseteq C$,

(C2) $hCh^{-1} \subseteq C$ for all $h \subseteq H$,

(C3) $C \cap C^{-1} = \varnothing$, and

(C4) $C \cup C^{-1} \cup \{1\} = H$.

Once we know that a set C in H satisfies the above properties, then we may define an order on H that is compatible with multiplication on the left, by demanding $h_1 < h_2$ in $H \Leftrightarrow h_1^{-1}h_2 \in C$ (as in [Bahls 2007] or [Duchamp and Thibon 1992], for example).

Let

$$x = \sum_{i=1}^{\infty} \beta_i b^i + \sum_{j=1}^{\infty} \alpha_j a^j + \sum_{h=1}^{\infty} \gamma_h b^h a \in \mathcal{A}.$$

If $c(x, b) \neq 0$, then we will define $\tau(x) = b$, otherwise $\tau(x) = a$. (We may think of τ as indicating the "dominant" term of x.) Let $\lambda(x) = c(x, \tau(x))$ and define the positive cone C by

$$C = \{1 + x \in H \mid \lambda(x) > 0\}.$$

We require a few simple technical results.

Lemma 4.1. *For positive integers i, j, i',*

$$c(b^i a^j, b^{i'} a) = \begin{cases} (-1)^{i+1} & \text{if } i = i', \\ 0 & \text{otherwise.} \end{cases}$$

Proof. Clearly if $i < i'$ then $c(b^i a^j, b^{i'} a) = 0$ since reducing to normal form never increases the exponent on the b. We must then consider the cases where $i = i'$ and $i > i'$. However, from this point on the proof consists of a pair of nested inductions (one for $i = i'$ and one for $i > i'$), each nearly identical to those in the proof of Lemma 3.4. The details are left to the reader.   □

Lemma 4.2. *Let* $y \in H$. *If* $c(y, b) = 0$, *then* $c(y, b^x) = 0$ *for any positive integer* x.

Proof. Indeed, $c(y, b) = 0$ implies that $y = \mu(b^m a^l b^{-m})$ for some $m \geq 0$. The only way to obtain terms of the form b^x from the product

$$(1 + b)^m (1 + a)^l (1 - b + b^2 - b^3 + \cdots)^m$$

is to avoid terms with as in them, *i.e.* extracting terms b^x from

$$(1 + b)^m \cdot 1 \cdot (1 - b + b^2 - b^3 + \cdots)^m = 1,$$

which clearly cannot be done.   □

Lemma 4.3. *Let* $y \in H$. *If* $c(y, b) = 0$, *then* $c(y, b^x a) = 0$ *for any positive integer* x.

Proof. As before, $c(y, b) = 0 \Rightarrow m = k$. Moreover, we have just shown that the only nonzero contribution to $c(y, b^x a)$ will come from reduction of terms $b^i a^j$ satisfying $i = x$. We therefore consider terms $b^h a^j b^i$ obtained from expanding

$$y = (1 + b)^m (1 + a)^l (1 - b + b^2 - b^3 + \cdots)^m$$

that satisfy $i + h = x$. (Note that moving the bs to the left past as does not change the exponent on the bs.)

The contribution to $c(y, b^x a)$ coming from such unreduced terms $b^h a^j b^i$ takes the form

$$\sum_{i=0}^{x} \binom{i + m - 1}{m - 1} \binom{m}{x - i} (-1)^i,$$

in which $\binom{m}{x-i}$ accounts for the contribution from $(1 + b)^m$ for a fixed i and $\binom{i+m-1}{m-1}(-1)^i$ accounts for the contribution from $(1 - b + b^2 - b^3 + \cdots)^m$ for the same i. It is not hard to show that the contribution from $(1 + a)^j$ is 1 as a consequence of basic combinatorics of binomial coefficients.

Thus

$$c(y, b^x a) = \sum_{i=0}^{x} \binom{i + m - 1}{m - 1} \binom{m}{x - i} (-1)^i = \frac{\Gamma(1)}{\Gamma(1 - x)\Gamma(1 + x)} = \frac{\sin(\pi x)}{\pi x}$$

by basic properties of the Gamma function. Since x is assumed to be a nonzero integer, this last quantity is 0, and we are done.   □

Thus if $c(y, b) = 0$ then the normal form of y consists only of powers of a.

Proposition 4.4. *The set C defined as above satisfies* (C1)–(C4).

Proof. Property (C4) is obvious.

For (C1), let $1 + x, 1 + y \in C$ be in normal form. If b does not appear as a term in these normals forms (and thus by the preceding lemmas neither do b^i or $b^i a$, $i \geq 1$) then no term of the form b^i or $b^i a$ will appear in the normal form of $(1+x)(1+y)$. Since $\tau(1+x) = a$ and $\tau(1+y) = a$, $c(1+x, a) > 0$ and $c(1+y, a) > 0$. As a result, $c((1+x)(1+y), a) > 0$ also.

If one of $1+x, 1+y \in C$ contains b as a term, by definition it will have a positive coefficient, and thus $c((1+x)(1+y), b) > 0$ as well.

For (C3), let $1 + x \in C$. Assume that $1 + x$ contains no terms b^i, and therefore contains only powers of a. Then

$$c((1+x)^{-1}, a) = c(1 - x + x^2 - x^3 + \cdots, a) = -c(1+x, a)$$

and thus $(1+x)^{-1} \notin C$. A similar argument may be used if terms b^i do appear in $1 + x$.

For (C2), let $1 + x = 1 + \beta b + x'$ where $\beta > 0$ is rational and all of the terms in x' have a form in

$$B = \{b^i \mid i \geq 2\} \cup \{a^i \mid i \geq 1\} \cup \{b^i a \mid i \geq 1\}.$$

Then

$$
\begin{aligned}
(1 + y)(1 + x)(1 &+ y)^{-1} \\
&= (1+y)(1 + \beta b + x')(1 - y + y^2 - y^3 + \cdots) \\
&= (1 + \beta b + x' + y + \beta y b + yx')(1 - y + y^2 - y^3 + \cdots) \\
&= (1 + \beta b + y + z)(1 - y + y^2 - y^3 + \cdots),
\end{aligned}
$$

where $z = x' + \beta y b + yx'$ consists of terms in B. Since terms that are not in the form γb (where $\gamma \neq 0$ is rational) do not contribute to $c(1 + x, b)$ in the reduced form, none of these terms will contribute to $c(1 + x, b)$ when reduced to normal form. Continuing, this becomes

$$
\begin{aligned}
(1 + \beta b + y + z)(1 - y &+ y^2 - y^3 + \cdots) \\
&= 1 + \beta b + y + z - y - \beta b y - y^2 - zy + y^2 + \beta b y^2 + \cdots \\
&= 1 + \beta b + \beta b(-y + y^2 - \cdots) + z(-y + y^2 - \cdots).
\end{aligned}
$$

The only term that will contribute to $c(1 + x, b)$ in this equation is βb. Thus $c(1 + x, b) = \beta$, and $(1 + y)(1 + x)(1 + y)^{-1} \in C$.

Next, assume that $1 + x \in C$ contains no bs. Then $1 + x = (1 + a)^l$ for some positive integer l. Consider $1 + y \in C$. As $1 + y$ is a mapping of a group element

into the algebra, it will be of the form $(1+b)^i (1+a)^j (1+b)^{-k}$ for some integers i, j, k, where $i, k \geq 0$. Therefore we can rewrite $(1+y)(1+x)(1+y)^{-1}$ as

$$(1+b)^i (1+a)^j (1+b)^{-k} (1+a)^l (1+b)^k (1+a)^{-j} (1+b)^{-i}.$$

This is $\mu(b^i a^j b^{-k} a^l b^k a^{-j} b^{-i}) = \mu(b^i a^{3^k l} b^{-i})$. However, it is easily shown that $c(\mu(b^i a^{3^k l} b^{-i}), a) = 3^k l$, which is positive because $k > 0$ is nonnegative and l is positive. We can also see that $\mu(b^i a^{3^k l} b^{-i})$ will contain no bs. Hence,

$$\lambda((1+y)(1+x)(1+y)^{-1}) > 0,$$

and $(1+y)(1+x)(1+y)^{-1} \in C$. $\square$

As discussed above, we have the following consequence:

Corollary 4.5. *The group* BS(1, 3) *is linearly orderable by an order that is compatible with multiplication on the left.*

5. BS(1, *n*) for *n* ≥ 4

Applying the method of this article to other groups was considered briefly in the final section of [Bahls 2007]. Although analysis of other classes of groups has not been performed, we conclude this article by indicating why the method we have pursued above will fail to admit a workable mapping μ when applied analogously to $BS(1, n) = \langle a, b \mid ab = ba^n \rangle$, $n \geq 4$.

As before, we may define the positive monoid M and the algebra $A_{\mathbb{Q}}(M)$ freely generated by M with coefficients in $\mathbb{Q} \cup \{\pm\infty\}$. The initial map μ taking a to $1+a$ and b to $1+b$ is still defined, and in fact we may even define $\mathcal{A}$ as before by forming the quotient of $A_{\mathbb{Q}}(M)$ by the ideal $I = \langle (1+a)(1+b) = (1+b)(1+a)^n \rangle$. This leads to a modified μ, as before.

The difficulty comes when we attempt to define a normal form for elements in $\mathcal{A}$. Expanding the relation $(1+a)(1+b) = (1+b)(1+a)^n$ yields

$$1+a+b+ab = \sum_{i=0}^{n} \binom{n}{i} a^i + \sum_{i=0}^{n} \binom{n}{i} ba^i \Rightarrow 0 = (n-1)a + \sum_{i=2}^{n} \binom{n}{i} a^i + \sum_{i=1}^{n-1} \binom{n}{i} ba^i$$

after canceling and applying the single group relation $ab = ba^n$.

What rule of reduction should we derive from this? In order that our replacement rule remain somewhat "context free" we ought to replace a single term by a sum containing $2n - 2$ terms. In order that our replacement rule give a terminating sequence of reductions, the single term must be one of either a^n or ba^{n-1}, since any other choice will give rise to an infinite sequence of rewritings in which "longer" strings continually replace "shorter" ones.

Choosing the reduction rule

$$ba^{n-1} \longrightarrow \frac{1}{n}\left[(n-1)a + \sum_{i=2}^{n}\binom{n}{i}a^i + \sum_{i=1}^{n-2}\binom{n}{i}ba^i\right],$$

as before, we obtain divergent alternating series as coefficients in certain reductions. For instance, if $n = 4$, the equation

$$ba^3 = -(3/4)a - (3/2)a^2 - a^3 - (1/4)a^4 - ba - (3/2)ba^2,$$

when applied to $\mu(ba^{-1})$ gives

$$
\begin{aligned}
\mu(ba^{-1}) \\
&= (1+b)(1 - a + a^2 - a^3 + a^4 - \cdots) \\
&= 1 - a + a^2 - a^3 + a^4 - \cdots + b - ba + ba^2 - ba^3 + ba^4 - \cdots \\
&= 1 - a + a^2 - a^3 + \cdots + b - ba + ba^2 + ((3/4)a \\
&\qquad + (3/2)a^2 + a^3 + (1/4)a^4 + ba + (3/2)ba^2) + ba^4 - \cdots \\
&= \cdots
\end{aligned}
$$

Already we see a trend that will continue: the coefficient $c(\mu(ba^{-1}), a)$ will receive contributions from each term of the form ba^i, and these contributions will continually alternate in sign and grow without bound, giving a divergent alternating sum. So long as $n \geq 4$, we will have this problem.

A similar difficulty arises if we attempt the only other feasible reduction rule, replacing a^n by the remaining $2n-2$ terms. Thus our method runs aground before we even have a chance to test μ's injectivity.

Acknowledgement

References

[Bahls 2007] P. Bahls, "The group BS(1, 2) and a construction of Magnus", *Comm. Algebra* **35**:12 (2007), 4088–4095. MR 2372320

[Bergman 1991] G. M. Bergman, "Right orderable groups that are not locally indicable", *Pacific J. Math.* **147**:2 (1991), 243–248. MR 92e:20030 Zbl 0677.06007

[Boyer et al. 2005] S. Boyer, D. Rolfsen, and B. Wiest, "Orderable 3-manifold groups", *Ann. Inst. Fourier (Grenoble)* **55**:1 (2005), 243–288. MR 2006a:57001 Zbl 1068.57001

[Dehornoy et al. 2002] P. Dehornoy, I. Dynnikov, D. Rolfsen, and B. Wiest, *Why are braids orderable?*, vol. 14, Panoramas et Synthèses [Panoramas and Syntheses], Société Mathématique de France, Paris, 2002. MR 2004e:20062 Zbl 1048.20021

[Duchamp and Krob 1990] G. Duchamp and D. Krob, "Partially commutative formal power series", pp. 256–276 in *Semantics of systems of concurrent processes (La Roche Posay, 1990)*, vol. 469, Lecture Notes in Comput. Sci., Springer, Berlin, 1990. MR 92e:68047

[Duchamp and Krob 1993] G. Duchamp and D. Krob, "Partially commutative Magnus transformations", *Internat. J. Algebra Comput.* **3**:1 (1993), 15–41. MR 94e:20041 Zbl 0820.20037

[Duchamp and Thibon 1992] G. Duchamp and J.-Y. Thibon, "Simple orderings for free partially commutative groups", *Internat. J. Algebra Comput.* **2**:3 (1992), 351–355. MR 93j:06017 Zbl 0772.20017

[Magnus et al. 1976] W. Magnus, A. Karrass, and D. Solitar, *Combinatorial group theory*, revised ed., Dover Publications Inc., New York, 1976. Presentations of groups in terms of generators and relations. MR 54 #10423 Zbl 0362.20023

[Rhemtulla 2002] A. H. Rhemtulla, "Orderable groups", pp. 47–55 in *Proceedings of the International Conference on Algebra and its Applications (ICAA 2002) (Bangkok)*, Chulalongkorn Univ., Bangkok, 2002. MR 2004g:20055 Zbl 1071.20508

[Short and Wiest 2000] H. Short and B. Wiest, "Orderings of mapping class groups after Thurston", *Enseign. Math. (2)* **46**:3-4 (2000), 279–312. MR 2003b:57003 Zbl 1023.57013

pbahls@unca.edu   *University of North Carolina, Asheville, Asheville, NC 28806, United States*

vcollins@wellesley.edu   *Department of Mathematics, Wellesley College, Wellesley, MA 02481, United States*

heroel8@wfu.edu   *Department of Mathematics, Wake Forest University, Winston-Salem, NC 27109, United States*

Congruences for Han's generating function

Dan Collins and Sally Wolfe

(Communicated by Kenneth S. Berenhaut)

For an integer $t \geq 1$ and a partition λ, we let $\mathcal{H}_t(\lambda)$ be the multiset of hook lengths of λ which are divisible by t. Then, define $a_t^{\text{even}}(n)$ and $a_t^{\text{odd}}(n)$ to be the number of partitions of n such that $|\mathcal{H}_t(\lambda)|$ is even or odd, respectively. In a recent paper, Han generalized the Nekrasov–Okounkov formula to obtain a generating function for $a_t(n) = a_t^{even}(n) - a_t^{odd}(n)$. We use this generating function to prove congruences for the coefficients $a_t(n)$.

1. Introduction and statement of results

Let $p(n)$ denote the number of integer partitions of n. Ramanujan proved the following important congruence relations for the partition function, which hold for all nonnegative n:

$$\begin{aligned} p(5n+4) &\equiv 0 \quad (\text{mod } 5), \\ p(7n+5) &\equiv 0 \quad (\text{mod } 7), \\ p(11n+6) &\equiv 0 \quad (\text{mod } 11). \end{aligned} \quad (1\text{-}1)$$

These congruences can be proven through q-series identities or with the theory of modular forms; both methods rely on the following generating function for $p(n)$:

$$\sum_{n=0}^{\infty} p(n)q^n = \prod_{n=1}^{\infty} \frac{1}{(1-q^n)} = 1 + q + 2q^2 + 3q^3 + 5q^4 + 7q^5 + 11q^6 + \cdots. \quad (1\text{-}2)$$

Recently, Nekrasov and Okounkov [2006] generalized Equation (1-2) by discovering a combinatorial interpretation of $\prod_{n=1}^{\infty}(1-q^n)^b$, for $b \in \mathbb{C}$, in terms of partition hook lengths. Here we briefly recall their results, beginning by introducing the necessary notation.

A Ferrers diagram, a pictorial representation of a partition, allows us to define the *hook length* of a box in the partition. The hook length of a box in the Ferrers diagram is the sum of the number of boxes in the same column below it, the number of boxes in the same row and to the right, and one for the box. For example, consider the partition $5 + 3 + 2$ of 10. Its Ferrers diagram, with hook lengths filled in, is:

7	6	4	2	1
4	3	1		
2	1			

We define $\mathcal{H}(\lambda)$, the *hook length multiset* of a partition λ, to be the multiset of hook lengths in each box in the Ferrers diagram of λ. We can then define $\mathcal{H}_t(\lambda) \subseteq \mathcal{H}(\lambda)$ to be the multiset of hook lengths of boxes in the partition that are multiples of t.

Nekrasov and Okounkov proved the following formula which uses these combinatorial objects, and holds for any complex b:

$$\sum_{\lambda \in P} q^{|\lambda|} \prod_{h \in \mathcal{H}(\lambda)} \left(1 - \frac{b}{h^2}\right) = \prod_{n \geq 1} (1 - q^n)^{b-1}.$$

More recently, Han obtained a generalization of this formula. A specialization of it gives useful infinite-product generating functions for the series

$$\sum_{n=0}^{\infty} a_t(n) q^n = \sum_{\lambda \in P} q^{|\lambda|} (-1)^{\#\mathcal{H}_t(\lambda)} = \prod_{n=1}^{\infty} \frac{(1 - q^{4tn})^t (1 - q^{tn})^{2t}}{(1 - q^{2tn})^{3t} (1 - q^n)}. \tag{1-3}$$

Remark 1.1. For $t = 1$, the sum over partitions is easy to understand directly: we have $\mathcal{H}_1(\lambda) = \mathcal{H}(\lambda)$, so $a_1(n) = (-1)^n p(n)$.

A number of congruences of the coefficients $a_t(n)$ are a direct consequence of the Ramanujan congruences in (1-1) combined with Han's generating function (1-3). Namely, for all $n \geq 0$, one has

$$\begin{aligned}
a_t(5n + 4) &\equiv 0 \pmod{5} && \text{if } t = 1 \text{ or } 5|t, \\
a_t(7n + 5) &\equiv 0 \pmod{7} && \text{if } t = 1 \text{ or } 7|t, \\
a_t(11n + 6) &\equiv 0 \pmod{11} && \text{if } t = 1 \text{ or } 11|t.
\end{aligned}$$

Here, we are interested in further congruences of the form $a_t(An + B)$.

Our search for such congruences over small arithmetic progressions and with small prime moduli yielded just one Ramanujan-type congruence that was not of the above form.

Theorem 1.2. *If $n \geq 0$, then $a_2(5n + 4) \equiv 0 \pmod{5}$.*

This congruence can be proven through q-series identities, which can in turn be proven using the theory of modular forms. More generally, we have:

Theorem 1.3. *If $3 \nmid t$ and $\ell > 3t^3 + t$ is prime, a positive proportion of primes p satisfy*

$$a_t\left(\frac{p^3 \ell n + 1}{24}\right) \equiv 0 \pmod{\ell},$$

for all n coprime to p.

Remark 1.4. Saying that *a positive proportion of primes* satisfy a condition means that the limit

$$\lim_{n \to \infty} \frac{\#\{p \leq n : p \text{ prime, and } p \text{ satisfies the condition}\}}{\#\{p \leq n : p \text{ prime}\}}$$

exists and is strictly positive.

Corollary 1.5. *For t, ℓ, p satisfying the previous theorem, there are linear congruences*

$$a_t(p^4 \ell n + b_p) \equiv 0 \pmod{\ell},$$

for a fixed $b_p < p^4 \ell$ and all nonnegative integers n.

To prove this, we also use methods of the theory of modular forms. Such methods were first employed by Ono [2000] and Ahlgren and Ono [2001] to prove the existence of classes of congruences for the partition function. We apply similar arguments to the generating functions for the $a_t(n)$.

Remark 1.6. Treneer [2006] extended the arguments in [Ono 2000] and [Ahlgren and Ono 2001] in a general way, to prove congruences for the coefficients of all weakly holomorphic modular forms. This result can be applied to our generating function to obtain similar conclusions. However, we proceed by other methods to obtain explicit constructions.

In Section 2, we discuss Han's generating function, and relate it to the theory of modular forms. In Section 3, we prove Theorem 1.2, and in Sections 4 and 5, we prove Theorem 1.3.

2. Han's generating function and modular forms

Han's Generating Function. Han [2008] proved the Nekrasov–Okounkov formula using combinatorial methods, and obtained the following generalization:

Theorem 2.1. *For any positive integer t, and complex numbers b, y, we have*

$$\sum_{\lambda \in P} q^{|\lambda|} \prod_{h \in \mathcal{H}_t(\lambda)} \left(y - \frac{tyb}{h^2}\right) = \prod_{n=1}^{\infty} \frac{(1 - q^{tn})}{(1 - (yq^t)^n)^{t-b}(1 - q^n)}.$$

Taking $b=0$ and $y=-1$, the left side reduces to the generating functions $\sum a_t(n)q^n$ that we are interested in:

$$\sum_{n=0}^{\infty} a_t(n)q^n = \prod_{n=1}^{\infty} \frac{(1-q^{tn})}{(1-(-q^t)^n)^t(1-q^n)}.$$

Manipulating the terms of the infinite product gives the following formula:

$$\sum_{n=0}^{\infty} a_t(n)q^n = \prod_{n=1}^{\infty} \frac{(1-q^{4tn})^t(1-q^{tn})^{2t}}{(1-q^{2tn})^{3t}(1-q^n)}.$$

Our results depend on the modularity of these series, which we explain in the next section.

Modularity of $\sum a_t(n)q^n$. Recall the definition of Dedekind's η-function, where we let $q = e^{2\pi i z}$:

$$\eta(z) = q^{1/24} \prod_{n=1}^{\infty} (1-q^n).$$

Using this definition and the infinite product generating function we can write

$$\sum_{n=0}^{\infty} a_t(n)q^n = q^{1/24} \frac{\eta(4tz)^t \eta(tz)^{2t}}{\eta(2tz)^{3t} \eta(z)}.$$

Replacing z by $24z$, we have

$$\sum_{n=0}^{\infty} a_t(n)q^{24n-1} = \frac{\eta(96tz)^t \eta(24tz)^{2t}}{\eta(48tz)^{3t} \eta(24z)}.$$

Combining Theorem 1.65 in [Ono 2004] about integer-weight η-quotients with the transformation properties for $\eta(24z)$, we have:

Theorem 2.2. $\sum_{n=0}^{\infty} a_t(n)q^{24n-1}$ is a weakly holomorphic modular form of weight $-1/2$ on the congruence subgroup $\Gamma_0(2304t)$, with character $\chi(d) = \left((2^t 3)/d\right)$.

3. Proof of Theorem 1.2

A q-series identity. The key step to the proof that $a_2(5n+4) \equiv 0 \pmod 5$ is the following q-series identity, which can be proven from the theory of modular forms:

Theorem 3.1. *The following identity is true*:

$$\prod_{n=1}^{\infty} \frac{(1-q^{4n})^2(1-q^n)^2}{1-q^{2n}} = \sum_{n\in\mathbb{Z}} (1-3n)q^{3n^2-2n}.$$

Proof. Let $q = e^{2\pi i z}$, and define

$$f(z) = \prod_{n \geq 1} \frac{(1 - q^{4n})^2 (1 - q^n)^2}{1 - q^{2n}}, \qquad g(z) = \sum_{n \in \mathbb{Z}} (1 - 3n) q^{3n^2 - 2n}.$$

We will prove that $q f(3z)$ and $q g(3z)$ are both modular forms in the same finite-dimensional space. Thus, to show equality it suffices to show that a finite number of terms in the q-expansion of $q(f(3z) - g(3z))$ are zero; this implies $f(z) = g(z)$.

We can write $q f(3z)$ as a quotient of Dedekind's eta-functions:

$$q f(3z) = q \prod_{n=1}^{\infty} \frac{(1 - q^{12n})^2 (1 - q^{3n})^2}{1 - q^{6n}} = \frac{\eta(12z)^2 \eta(3z)^2}{\eta(6z)}.$$

By the standard theory of eta-quotients (as in [Ono 2004, Section 1.4]), this is a cusp form of weight $3/2$, level 144, and character $\chi(d) = (3/d)$.

On the other hand, $q g(3z)$ can be expressed as a Jacobi theta function. Define $\psi(n)$ to be the Dirichlet character $(n/3)$. As in [Ono 2004, Section 1.3.1], define

$$\theta(\psi, 1, z) = \sum_{n=1}^{\infty} \psi(n) n q^{n^2},$$

which is a cusp form of weight $3/2$, level 36, and character $\chi(d) = (3/d)$. By periodicity of ψ, we have

$$\theta(\psi, 1, z) = \psi(0) \sum_{n=1}^{\infty} 3n q^{(3n)^2} + \psi(1) \sum_{n=0}^{\infty} (3n+1) q^{(3n+1)^2} + \psi(2) \sum_{n=0}^{\infty} (3n+2) q^{(3n+2)^2}$$

$$= \sum_{n=0}^{\infty} (3n+1) q^{(3n+1)^2} - \sum_{n=0}^{\infty} (3n+2) q^{(3n+2)^2}$$

$$= - \sum_{n=-\infty}^{\infty} (1 - 3n) q^{(3n-1)^2} = -q \sum_{n \in \mathbb{Z}} (1 - 3n) q^{3(3n^2 - 2n)} = -q\, g(3z).$$

Therefore, both $q f(3z)$ and $q g(3z)$ are in $S_{3/2}(\Gamma_0(144), \chi)$, so to check equality it suffices [Sturm 1987, Theorem 1] to check equality of the first $k/24[\Gamma_0(1) : \Gamma_0(144)] = 18$ coefficients. Thus, we have

$$f(z) = g(z) = 1 - 2q + 4q^5 - 5q^8 + 7q^{16} - 8q^{21} \pm \cdots. \qquad \square$$

Proof of the congruence. We can now prove that $a_2(5n + 4) \equiv 0 \pmod{5}$. First, note that we can formally factor the generating function for $a_2(n)$. By doing so

and applying the binomial theorem mod 5, we have

$$
\sum_{n=0}^{\infty} a_2(n)q^n = \prod_{n=1}^{\infty} \frac{1}{(1-q^{4n})^5} \prod_{n=1}^{\infty} \frac{(1-q^{8n})^2(1-q^{2n})^2}{(1-q^{4n})} \prod_{n=1}^{\infty} \frac{(1-q^{2n})^2}{(1-q^n)}
$$
$$
\equiv \prod_{n=1}^{\infty} \frac{1}{(1-q^{20n})} \prod_{n=1}^{\infty} \frac{(1-q^{8n})^2(1-q^{2n})^2}{(1-q^{4n})} \prod_{n=1}^{\infty} \frac{(1-q^{2n})^2}{(1-q^n)} \quad (\text{mod } 5).
$$

Using the partition generating function (1-2), Theorem 3.1, and an identity of Jacobi, we can write

$$
\sum_{n=0}^{\infty} a_2(n)q^n = \Big(\sum_{i=0}^{\infty} p(i)q^{20i}\Big)\Big(\sum_{k\in\mathbb{Z}}(1-3k)(q^2)^{3k^2-2k}\Big)\Big(\sum_{m=0}^{\infty} q^{(m^2+m)/2}\Big).
$$

A coefficient $a_2(5n+4)$ will thus be a sum of terms $p(i) \cdot (1-3k) \cdot 1$ where

$$
20i + 6k^2 - 4k + \frac{m^2+m}{2} \equiv 4 \quad (\text{mod } 5).
$$

We can check that this only holds when $m \equiv k \equiv 2 \pmod 5$. For such terms, $1 - 3k \equiv 0 \pmod 5$, so $a_2(5n+4) \equiv 0 \pmod 5$. $\qquad\square$

4. Sieved generating functions and cusp forms mod ℓ

To prove the existence of an infinite class of congruences, we follow similar arguments to those used by Ono [2000] and Ahlgren and Ono [2001] to prove congruences for the partition function. We first construct a cusp form congruent to a sieved version of our original generating function

$$
\sum_{n=0}^{\infty} a_t(n)q^n.
$$

Theorem 4.1. *If $3 \nmid t$ and $\ell > 3t^3 + t$ is prime, there exists a half-integer weight cusp form $g_{t,\ell}(z)$ with a q-series expansion satisfying the congruence*

$$
g_{t,\ell}(z) \equiv \sum_{n=0}^{\infty} a_t(\ell n + \beta_\ell)q^{24n+\frac{24\beta_\ell-1}{\ell}} \quad (\text{mod } \ell),
$$

where β_ℓ satisfies $24\beta_\ell \equiv 1 \pmod \ell$ and $0 < \beta_\ell < \ell$.

We can rewrite the sieved generating function as

$$
\sum_{n=0}^{\infty} a_t(\ell n + \beta_\ell)q^{24n+\frac{24\beta_\ell-1}{\ell}} = \sum_{\substack{n\geq 0 \\ \ell n \equiv -1 \,(\text{mod } 24)}} a_t\Big(\frac{\ell n+1}{24}\Big)q^n.
$$

If we take $a_t(m) = 0$ for any noninteger m, then the conclusion of Theorem 4.1 can be written as

$$g_{t,\ell}(z) \equiv \sum_{n=0}^{\infty} a_t\left(\frac{\ell n + 1}{24}\right) q^n \pmod{\ell}. \tag{4-1}$$

Preliminaries for Proof. We define the functions

$$F_t(z) = \frac{\eta(4tz)^t \eta(tz)^{2t}}{\eta(2tz)^{3t} \eta(z)} = q^{-1/24} \sum_{n=0}^{\infty} a_t(n) q^n,$$

$$H_{t,\ell}(z) = \eta(z)^\ell \eta(2tz)^{5t\ell} \eta(4tz)^{3t\ell} \eta(tz)^{2t\ell},$$

$$G_{t,\ell}(z) = F_t(z) H_{t,\ell}(z)^\ell.$$

By standard facts about eta-quotients (as in [Ono 2004, Section 1.4]), $G_{t,\ell}(z)$ is an integer-weight cusp form on $\Gamma_0(4t)$ and $H_{t,\ell}(24z)$ is a half-integer weight cusp form on $\Gamma_0(2304t)$. We relate the sieved generating function from Theorem 4.1 to these functions by:

Lemma 4.2. *The following congruence between q-series expansions holds*:

$$G_{t,\ell}(z)|T(\ell) \equiv H_{t,\ell}(z) \sum_{n=0}^{\infty} a_t(\ell n + \beta_\ell) q^{n + \frac{24\beta_\ell - 1}{24\ell}} \pmod{\ell}.$$

Here we let $T(\ell)$ denote the ℓ-th Hecke operator

$$\left(\sum_{n=0}^{\infty} b(n) q^n\right)|T(\ell) = \sum_{n=0}^{\infty} \left(b(\ell n) + \chi(\ell)\ell^{k-1} b(n/\ell)\right) q^n,$$

where k and χ are the weight and character of the form $\sum_{n=0}^{\infty} b(n) q^n$.

Proof. Define $\delta_\ell = (\ell^2 - 1)/24$, which is an integer for all primes $\ell \geq 5$. By definition of $G_{t,\ell}(z)$ and η, we have

$$G_{t,\ell}(z) = \left(\sum_{n=0}^{\infty} a_t(n) q^{n+\delta_l}\right) q^{t^2\ell^2} \prod_{n=1}^{\infty} (1-q^n)^{\ell^2} (1-q^{2tn})^{5t\ell^2} (1-q^{4tn})^{3t\ell^2} (1-q^{tn})^{2t\ell^2}.$$

Applying the binomial theorem mod ℓ gives

$$G_{t,\ell}(z) \equiv q^{t^2\ell^2} \prod_{n=1}^{\infty} (1 - q^{n\ell^2})(1 - q^{2t\ell^2 n})^{5t}(1 - q^{4t\ell^2 n})^{3t}(1 - q^{t\ell^2 n})^{2t}$$

$$\cdot \left(\sum_{n=0}^{\infty} a_t(n) q^{n+\delta_l}\right) \pmod{\ell}.$$

The $T(\ell)$ operator is equivalent mod ℓ to the $U(\ell)$ operator, which is defined as

$$\left(\sum_{n=0}^{\infty} b(n)q^n\right)\Big|U(\ell) = \sum_{n=0}^{\infty} b(\ell n)q^n.$$

We apply $T(\ell)$ to the left side and $U(\ell)$ to the right side to obtain

$$G_{t,\ell}(z)|T(\ell) \equiv q^{t^2\ell} \prod_{n=1}^{\infty}(1-q^{n\ell})(1-q^{2t\ell n})^{5t}(1-q^{4t\ell n})^{3t}(1-q^{t\ell n})^{2t}$$

$$\cdot \left(\sum_{n=0}^{\infty} a_t(\ell n + \beta_\ell)q^{n+\frac{\beta_\ell+\delta_\ell}{\ell}}\right)$$

$$\equiv H_{t,\ell}(z)\left(\sum_{n=0}^{\infty} a_t(\ell n + \beta_\ell)q^{n+\frac{24\beta_\ell-1}{24\ell}}\right) \quad (\text{mod } \ell). \qquad \square$$

We define

$$g_{t,\ell}(z) = \frac{G_{t,\ell}(z)|T(\ell)|V(24)}{H_{t,\ell}(24z)}.$$

Lemma 4.2 tells us that $g_{t,\ell}(z)$ is congruent mod ℓ to our sieved generating function. Thus, to prove Theorem 4.1 we need to show that $g_{t,\ell}(z)$ is a cusp form.

Proof of Theorem 4.1. It suffices to prove that $(G_{t,\ell}(z)|T(\ell))/H_{t,\ell}(z)$ vanishes at all of the cusps, since applying $V(24)$ will preserve cuspidality. Since the Hecke operator preserves the level, $G_{t,\ell}(z)|T(\ell)$ is a form on $\Gamma_0(4t)$; standard facts about eta-quotients show that $H_{t,\ell}(z)^{24}$ is a form on $\Gamma_0(4t)$ as well.

Since the order of vanishing of $H_{t,\ell}(z)$ at any cusp is $1/24$-th of the order of vanishing of $H_{t,\ell}(z)^{24}$ at that cusp, it suffices to consider orders of vanishing on a set of cusps containing a representative for each equivalence class on $\Gamma_0(4t)$. The cusps of the form c/d, where $d|4t$ and $(c,d)=1$, form such a set. We can divide the allowed values of d into three classes: $d=T$, $d=2T$, and $d=4T$, where $T|t$ and, for the latter two cases, $2T \nmid t$.

Let $\text{ord}_{c/d} f$ denote the invariant order of vanishing of a function f at a cusp c/d. We can compute:

$$d=T: \quad \text{ord}_{c/d} G_{t,\ell} = \frac{3T^2 - 4 + 21T^2\ell^2 + 4\ell^2}{96}, \quad \text{ord}_{c/d} H_{t,\ell} = \frac{21T^2\ell + 4\ell}{96},$$

$$d=2T: \quad \text{ord}_{c/d} G_{t,\ell} = \frac{-3T^2 - 1 + 15T^2\ell^2 + \ell^2}{24}, \quad \text{ord}_{c/d} H_{t,\ell} = \frac{15T^2\ell + \ell}{24},$$

$$d=4T: \quad \text{ord}_{c/d} G_{t,\ell} = \frac{-1 + 24T^2\ell^2 + \ell^2}{24}, \quad \text{ord}_{c/d} H_{t,\ell} = \frac{24T^2\ell + \ell}{24}.$$

Applying a Hecke operator $T(\ell)$ to a function takes the q-series expansion at a cusp c/d to a linear combination of q-series expansions around cusps of the form

c'/d, with q replaced by $q^{1/\ell}$. Because the order of vanishing depends only on the denominator, we have

$$\text{ord}_{c/d}\, G_{t,\ell}(z)|T(\ell) \geq \frac{1}{\ell}\, \text{ord}_{c/d}\, G_{t,\ell}(z).$$

Since $G_{t,\ell}(z)|T(\ell)$ is a form on $\Gamma_0(4t)$, we know that its order of vanishing must be of the form $A/4t$, where A is an integer. Using this fact, we can analyze the behavior at each cusp, and show that $\text{ord}_{c/d}\, G_{t,\ell}(z)|T(\ell) > \text{ord}_{c/d}\, H_{t,\ell}(z)$. For instance, at cusps c/d where $d = 2T$, we have

$$\text{ord}_{c/d}\, G_{t,\ell}(z)|T(\ell) = \frac{A}{4t} \geq \frac{1}{24}\left(\frac{-3T^2 - 1}{\ell} + 15T^2\ell + \ell\right).$$

This gives

$$6A \geq \frac{-3T^2 t - t}{\ell} + 15T^2 t\ell + t\ell.$$

By hypothesis, $\ell > 3t^3 + t \geq 3T^2 t + t$; hence

$$0 > \frac{-3T^2 t - t}{\ell} > -1.$$

Since the other terms in the inequality are integers, we must have

$$6A \geq 15T^2 t\ell + t\ell.$$

If equality held, the equation would reduce to $0 \equiv t\ell \pmod{3}$; since t, ℓ are coprime to 3, we must have the strict inequality $6A > 15T^2 t\ell + t\ell$. We therefore obtain the desired inequality

$$\text{ord}_{c/d}\, G_{t,\ell}(z)|T(\ell) = \frac{A}{4t} > \frac{15T^2\ell + \ell}{24} = \text{ord}_{c/d}\, H_{t,\ell}.$$

A similar analysis at cusps c/d where $d = T$ and $d = 4T$ shows that

$$\text{ord}_{c/d}\, G_{t,\ell}(z)|T(\ell) > \text{ord}_{c/d}\, H_{t,\ell}(z),$$

as well. Therefore,

$$\text{ord}_{c/d}\, \frac{G_{t,\ell}(z)|T(\ell)}{H_{t,\ell}(z)} > 0$$

at all cusps. $\qquad\square$

5. Proof of Theorem 1.3

We can now consider applying Hecke operators to function $g_{t,\ell}(z)$; modulo ℓ, this is equivalent to applying them to the sieved generating function

$$\sum_{n=0}^{\infty} a_t(\ell n + \beta_\ell)q^{24n+\frac{24\beta_\ell-1}{\ell}}.$$

For a half-integral weight modular form

$$f(z) = \sum_{n=0}^{\infty} b(n)q^n \in S_{\lambda+\frac{1}{2}}(\Gamma_0(N), \chi)$$

and a prime p, the Hecke operator $T(p^2)$ is defined by

$$f(z)|T(p^2) = \sum_{n=0}^{\infty}\left(b(p^2n) + \chi(p)\left(\frac{(-1)^\lambda n}{p}\right)p^{\lambda-1}b(n) + \chi(p^2)p^{2\lambda-1}b(n/p^2)\right)q^n.$$

Following the methods of Ono [2000], we will prove the following theorem, from which we can obtain congruences of the desired type.

Theorem 5.1. *If $(t, 3) = 1$ and $\ell > 3t^3 + 3$ is prime, then for a positive proportion of primes p,*

$$\sum_{n=0}^{\infty} a_t(\ell n + \beta_\ell)q^{24n+\frac{24\beta_\ell-1}{\ell}} \equiv 0 \pmod{\ell}.$$

A Theorem of Serre and the Shimura Correspondence. The proof of Theorem 5.1 relies on two important theorems, one of Serre and one of Shimura. Serre [1976] proves that many Hecke operators annihilate modulo ℓ an integer weight space of cusp forms.

Theorem 5.2 (Serre). *Consider a fixed space of cusp forms $S_k(\Gamma_0(N), \chi)$, where k is an integer. The set of primes $p \equiv -1 \pmod{N}$ such that $f|T(p) \equiv 0 \pmod{\ell}$ for all $f \in S_k(\Gamma_0(N), \chi)$ has positive density.*

To apply this to the half-integer weight case, we use the *Shimura correspondence* [Shimura 1973] to relate integer weight and half-integer weight forms.

Theorem 5.3 (Shimura). *Let $f = \sum_{n=1}^{\infty} b(n)q^n$ be a half-integer weight cusp form in $S_{\lambda+1/2}(\Gamma_0(4N), \psi)$. For a positive integer r, define $S_r(f)$ by*

$$S_r(f)(z) = \sum_{n=1}^{\infty} A_r(n)q^n, \qquad \sum_{n=1}^{\infty} \frac{A_r(n)}{n^s} = L(s - \lambda + 1, \psi\chi_{-1}^\lambda\chi_t)\sum_{n=1}^{\infty}\frac{b(rn^2)}{n^s},$$

where χ_{-1} and χ_t are the Kronecker characters for $\mathbb{Q}(i)$ and $\mathbb{Q}(\sqrt{t})$. Then

$$S_r(f) \in S_{2\lambda}(\Gamma_0(4N), \psi^2).$$

Moreover, if $p \nmid 4N$ is prime, then $S_r\left(f | T(p^2)\right) = S_r(f) | T(p)$.

Combining these two theorems will give us an analogue to Serre's theorem for half-integer weight modular forms, which proves the existence of primes that annihilate our sieved generating function.

Proof of Theorem 5.1. Let $\mathcal{P}_{t,\ell}$ be the set of primes $p \equiv -1 \pmod{2304t}$ such that $f | T(p^2) \equiv 0 \pmod{\ell}$ for all $f \in S_{2\lambda}(\Gamma_0(2304t), \chi_0)$, where χ_0 is the trivial Dirichlet character, and $\lambda + 1/2$ is the weight of the form $g_{t,\ell}(z)$ constructed in Section 4. By Serre's Theorem, $\mathcal{P}_{t,\ell}$ has positive density in the set of primes.

Furthermore, $S_r(g_{t,\ell})$, the image of g under the t-th Shimura correspondence, is in $S_{2\lambda}(\Gamma_0(2304t), \chi_0)$. So, for any $p \in \mathcal{P}_{t,\ell}$,

$$S_r(g_{t,\ell}) | T(p) = S_r(g_{t,\ell} | T(p^2)) \equiv 0 \pmod{\ell}.$$

By construction of the Shimura correspondence, if $S_r(f) \equiv 0 \pmod{\ell}$, then $f \equiv 0 \pmod{\ell}$. So, for all $p \in \mathcal{P}_{t,\ell}$, $g_{t,\ell} | T(p^2) \equiv 0 \pmod{\ell}$. $\qquad\square$

Proof of Theorem 1.3. From Theorem 5.1, for a positive proportion of primes p and all m,

$$b_{t,\ell}(p^2 m) + \chi(p)\left(\frac{(-1)^\lambda m}{p}\right) p^{\lambda-1} b_{t,\ell}(m) + \chi(p^2) p^{2\lambda-1} b_{t,\ell}(m/p^2) \equiv 0 \pmod{\ell},$$

where $b_{t,\ell}(n)$ is the coefficient of q^n in the Fourier expansion of $g_{t,\ell}(z)$.

In particular, consider $m = pn$ for some n coprime to p. Then m/p^2 is not an integer, and $b_{t,\ell}(m/p^2) = 0$; furthermore the Legendre symbol $\left(((-1)^\lambda m)/p\right)$ is zero. Recalling Equation (4-1), we have

$$b_{t,\ell}(p^3 n) \equiv a_t\left(\frac{p^3 \ell n + 1}{24}\right) \equiv 0 \pmod{\ell},$$

which proves Theorem 1.3. $\qquad\square$

Proof of Corollary 1.5. Let $0 \le r \le 24$ satisfy $rp\ell \equiv -1 \pmod{24}$. Replacing n by $24pn + r$, we obtain

$$a_t\left(\frac{24 p^4 \ell n - r p^3 \ell + 1}{24}\right) = a_t(p^4 \ell n + b_p) \equiv 0 \pmod{\ell},$$

where $b_p = (rp^3\ell + 1)/24$ is an integer. $\qquad\square$

Acknowledgements

References

[Ahlgren and Ono 2001] S. Ahlgren and K. Ono, "Congruence properties for the partition function", *Proceedings of the National Academy of Sciences* **98**:23 (2001), 12882–12884.

[Han 2008] G.-N. Han, "The Nekrasov-Okounkov hook length formula: refinement, elementary proof, extension, and applications", 2008. Preprint: arXiv:0805.1398v1 [math.CO].

[Nekrasov and Okounkov 2006] N. A. Nekrasov and A. Okounkov, "Seiberg-Witten theory and random partitions", pp. 525–596 in *The Unity of Mathematics*, vol. 244, Progress in Mathematics, Birkhäuser Boston, 2006.

[Ono 2000] K. Ono, "Distribution of the partition function modulo m", *The Annals of Mathematics* **151**:1 (2000), 293–307.

[Ono 2004] K. Ono, *The Web of modularity: arithmetic of the coefficients of modular forms and q-series*, American Mathematical Society, 2004.

[Serre 1976] J.-P. Serre, "Divisibilité de certaines fonctions arithmétiques", *L'Enseignement Mathématique* **22** (1976), 227–260.

[Shimura 1973] G. Shimura, "On modular forms of half integral weight", *The Annals of Mathematics* **97**:3 (1973), 440–481.

[Sturm 1987] J. Sturm, "On The congruence of modular forms", pp. 275–280 in *Number Theory*, vol. 1240, Lecture Notes in Mathematics, Springer Berlin / Heidelberg, 1987.

[Treneer 2006] S. Treneer, "Congruences for the coefficients of weakly holomorphic modular forms", *Proc. London Math. Soc.* **93**:2 (2006), 304–324.

djc224@cornell.edu　　　　　　　*Department of Mathematics, Cornell University, 310 Malott Hall, Ithaca, NY 14853, United States*

swolfe2@wisc.edu　　　　　　　*Department of Mathematics, University of Wisconsin, Madison, WI 53706, United States*

On the existence of unbounded solutions for some rational equations

Gabriel Lugo

(Communicated by Kenneth S. Berenhaut)

We resolve several conjectures regarding the boundedness character of the rational difference equation

$$x_n = \frac{\alpha + \delta x_{n-3}}{A + B x_{n-1} + C x_{n-2} + E x_{n-4}}, \quad n \in \mathbb{N}.$$

We show that whenever parameters are nonnegative, $A < \delta$, and $C, E > 0$, unbounded solutions exist for some choice of nonnegative initial conditions. We also partly resolve a conjecture regarding the boundedness character of the rational difference equation

$$x_n = \frac{x_{n-3}}{B x_{n-1} + x_{n-4}}, \quad n \in \mathbb{N}.$$

We show that whenever $B > 2^5$, unbounded solutions exist for some choice of nonnegative initial conditions.

1. Introduction

Palladino [2009a] studies a trichotomy behavior of the k-th order rational difference equation with nonnegative parameters and nonnegative initial conditions,

$$x_n = \frac{\alpha + \sum_{i=1}^{k} \beta_i x_{n-i}}{A + \sum_{j=1}^{k} B_j x_{n-j}}, \quad n \in \mathbb{N}.$$

Palladino established that there is a trichotomy behavior which is dependent on the relation between A and $\sum_{i=1}^{k} \beta_i$. In particular, in this paper, it was established that, under certain conditions, when $A < \sum_{i=1}^{k} \beta_i$ unbounded solutions exist. Here we will broaden that proof of unboundedness and show that when $A < \sum_{i=1}^{k} \beta_i$ unbounded solutions exist under different conditions. In Section 2 we present a

proof based on [Palladino 2009a, Section 5] which serves to generalize this work. An immediate consequence of this, as discussed later, will be to show that whenever parameters are nonnegative, $A < \delta$, and $C, E > 0$, unbounded solutions exist for some choice of nonnegative initial conditions for the rational difference equation,

$$x_n = \frac{\alpha + \delta x_{n-3}}{A + B x_{n-1} + C x_{n-2} + E x_{n-4}}, \quad n \in \mathbb{N}.$$

This resolves the conjectures regarding boundedness character for equations 609, 611, 617, and 619 presented in [Camouzis and Ladas 2008].

In Section 3, we partially resolve Conjecture 2 in [Palladino 2009a]. We show that the rational difference equation

$$x_n = \frac{x_{n-3}}{B x_{n-1} + x_{n-4}}, \quad n \in \mathbb{N},$$

has unbounded solutions whenever $B > 2^5$. In the process, we resolve the conjecture in [Camouzis and Ladas 2008] regarding the boundedness character of equation 584. The proof here will use similar techniques to those presented in [Lugo and Palladino $\geq$ 2009].

2. Preliminary results

During this section we use the ideas of modulo classes. Let us introduce these ideas in the following remark.

Remark 1. We say that a is congruent to b with modulus c and write $a \equiv b \bmod c$ if $c \mid a - b$. It is well known that given $z \in \mathbb{Z}$, there exists $a \in \{0, \ldots, c-1\}$ so that $z \equiv a \bmod c$. We call such a the residue of z with respect to the modulus c, and write $a = z \bmod c$.

Here we introduce a condition which allows us to construct unbounded solutions, namely Condition 1. Before doing so let us first introduce some notation. Let us define the following sets of indices:

$$I_\beta = \{i \in \{1, 2, \ldots, k\} \mid \beta_i > 0\} \quad \text{and} \quad I_B = \{j \in \{1, 2, \ldots, k\} \mid B_j > 0\}.$$

These sets are used extensively in [Palladino 2009b] when referring to the k-th order rational difference equation. Similarly we shall make extensive use of this notation.

Condition 1. We say that Condition 1 is satisfied if, for some $p \in \mathbb{N}$, $p \mid \gcd I_\beta$. We also must have disjoint sets $B, L \subset \{0, \ldots, p-1\}$ with $B \neq \varnothing$ and with the following properties.

(1) For all $b \in B$, $\{(b - j) \bmod p : j \in I_B\} \subset L$.

(2) For all $\ell \in L$, there exists $j \in I_B$ so that $(\ell - j) \bmod p \in B$.

We now present Theorem 1 which makes use of Condition 1. In the remainder of this section we will verify Condition 1 for a number of special cases of the fourth-order rational difference equation, thereby confirming several conjectures in [Camouzis and Ladas 2008].

Theorem 1. *Consider the k-th order rational difference equation*

$$x_n = \frac{\alpha + \sum_{i=1}^{k} \beta_i x_{n-i}}{A + \sum_{j=1}^{k} B_j x_{n-j}}, \quad n \in \mathbb{N}. \tag{1}$$

Assume nonnegative parameters and nonnegative initial conditions. Further assume that $A < \sum_{i=1}^{k} \beta_i$ and $\sum_{i=1}^{k} \beta_i > 0$, and that Condition 1 is satisfied for Equation (1). Then unbounded solutions of Equation (1) exist for some initial conditions.

Proof. By assumption, we may choose $p \in \mathbb{N}$ and $B, L \subset \{0, \ldots, p-1\}$ so that Condition 1 is satisfied. Choose initial conditions x_{-m} where $m \in \{0, \ldots, k-1\}$ so that the following holds. If $(-m \bmod p) \in B$, then

$$x_{-m} > \frac{2\alpha \sum_{j=1}^{k} B_j}{(\min_{j \in I_B} B_j)((\sum_{i=1}^{k} \beta_i) - A)} + \frac{\sum_{i=1}^{k} \beta_i}{\min_{j \in I_B} B_j}.$$

If $(-m \bmod p) \in L$, then

$$x_{-m} < \frac{(\sum_{i=1}^{k} \beta_i) - A}{2 \sum_{j=1}^{k} B_j}.$$

Also assume $x_{-m} > 0$ for all $m \in \{0, \ldots, k-1\}$.

Under this choice of initial conditions our solution $\{x_n\}$ has the following properties.

(a) $x_n > \dfrac{2\alpha \sum_{j=1}^{k} B_j}{(\min_{j \in I_B} B_j)((\sum_{i=1}^{k} \beta_i) - A)} + \dfrac{\sum_{i=1}^{k} \beta_i}{\min_{j \in I_B} B_j}$ whenever $(n \bmod p) \in B$.

(b) $x_n < \dfrac{(\sum_{i=1}^{k} \beta_i) - A}{2 \sum_{j=1}^{k} B_j}$ whenever $(n \bmod p) \in L$.

(c) $x_n > 0$ for all $n \in \mathbb{N}$.

We prove this using induction on n; our initial conditions provide the base case. Assume that the statement is true for all $n \leq N - 1$. We show the statement for $n = N$.

This induction proof has three cases. Let us begin by assuming $(N \bmod p) \in B$.

<u>Case (a).</u> Condition 1(1) tells us that in this case $\{(N-j) \bmod p : j \in I_B\} \subset L$. Hence

$$x_{N-j} < \frac{(\sum_{i=1}^{k} \beta_i) - A}{2 \sum_{j=1}^{k} B_j} \quad \text{for all } j \in I_B.$$

Since $p \mid \gcd(I_\beta)$, $N \bmod p = (N-i) \bmod p$ for all $i \in I_\beta$. Thus for all $i \in I_\beta$,

$$x_{N-i} > \frac{2\alpha \sum_{j=1}^{k} B_j}{(\min_{j \in I_B} B_j)((\sum_{i=1}^{k} \beta_i) - A)} + \frac{\sum_{i=1}^{k} \beta_i}{\min_{j \in I_B} B_j}.$$

Hence

$$x_N = \frac{\alpha + \sum_{i=1}^{k} \beta_i x_{N-i}}{A + \sum_{j=1}^{k} B_j x_{N-j}}$$

$$\geq \frac{\sum_{i=1}^{k} \beta_i}{A + (\sum_{j=1}^{k} B_j) \frac{(\sum_{i=1}^{k} \beta_i) - A}{2 \sum_{j=1}^{k} B_j}} \left(\frac{2\alpha \sum_{j=1}^{k} B_j}{(\min_{j \in I_B} B_j)((\sum_{i=1}^{k} \beta_i) - A)} + \frac{\sum_{i=1}^{k} \beta_i}{\min_{j \in I_B} B_j} \right)$$

$$\geq \frac{2\alpha \sum_{j=1}^{k} B_j}{(\min_{j \in I_B} B_j)((\sum_{i=1}^{k} \beta_i) - A)} + \frac{\sum_{i=1}^{k} \beta_i}{\min_{j \in I_B} B_j}.$$

This inequality is obtained by simply replacing the terms in the denominator with their upper bound, and replacing the terms in the numerator with their lower bound. This finishes case (a).

<u>Case (b).</u> We now assume $(N \bmod p) \in L$. Since $p \mid \gcd(I_\beta)$, we have $N \bmod p = (N-i) \bmod p$ for all $i \in I_\beta$. Hence

$$x_{N-i} < \frac{(\sum_{i=1}^{k} \beta_i) - A}{2 \sum_{j=1}^{k} B_j} \quad \text{for all } i \in I_\beta.$$

Condition 1(2) guarantees that there exists $j \in I_B$ so that

$$x_{N-j} > \frac{2\alpha \sum_{j=1}^{k} B_j}{(\min_{j \in I_B} B_j)((\sum_{i=1}^{k} \beta_i) - A)} + \frac{\sum_{i=1}^{k} \beta_i}{\min_{j \in I_B} B_j}.$$

Hence

$$x_N = \frac{\alpha + \sum_{i=1}^{k} \beta_i x_{N-i}}{A + \sum_{j=1}^{k} B_j x_{N-j}} < \frac{\alpha + (\sum_{i=1}^{k} \beta_i) \dfrac{(\sum_{i=1}^{k} \beta_i) - A}{2 \sum_{j=1}^{k} B_j}}{(\min_{j \in I_B} B_j) \left(\dfrac{2\alpha \sum_{j=1}^{k} B_j}{(\min_{j \in I_B} B_j)((\sum_{i=1}^{k} \beta_i) - A)} + \dfrac{\sum_{i=1}^{k} \beta_i}{\min_{j \in I_B} B_j} \right)}$$

$$
= \frac{\alpha + (\sum_{i=1}^{k} \beta_i) \dfrac{(\sum_{i=1}^{k} \beta_i) - A}{2 \sum_{j=1}^{k} B_j}}{\dfrac{2\alpha \sum_{j=1}^{k} B_j}{(\sum_{i=1}^{k} \beta_i) - A} + \sum_{i=1}^{k} \beta_i}
$$

$$
= \frac{\dfrac{2\alpha \sum_{j=1}^{k} B_j}{(\sum_{i=1}^{k} \beta_i) - A} + \sum_{i=1}^{k} \beta_i}{\dfrac{2\alpha \sum_{j=1}^{k} B_j}{(\sum_{i=1}^{k} \beta_i) - A} + \sum_{i=1}^{k} \beta_i} \left(\frac{(\sum_{i=1}^{k} \beta_i) - A}{2 \sum_{j=1}^{k} B_j} \right) = \frac{(\sum_{i=1}^{k} \beta_i) - A}{2 \sum_{j=1}^{k} B_j}. \qquad (2)
$$

This finishes case (b).

<u>Case (c).</u> It is clear that if $x_n > 0$ for $n < N$. Then $x_N > 0$ so case (c) is trivial.

We now use the facts we obtained from our induction to prove that a particular subsequence is unbounded. Take $b \in B$. We now show that $\{x_{mp+b}\}_{m=1}^{\infty}$ diverges to ∞. We explained earlier that

$$
x_{mp+b-j} < \frac{(\sum_{i=1}^{k} \beta_i) - A}{2 \sum_{j=1}^{k} B_j},
$$

since $\{(mp + b - j) \bmod p : j \in I_B\} \subset L$. Hence,

$$
x_{mp+b} = \frac{\alpha + \sum_{i=1}^{k} \beta_i x_{mp+b-i}}{A + \sum_{j=1}^{k} B_j x_{mp+b-j}} > \frac{\sum_{i=1}^{k} \beta_i x_{mp+b-i}}{A + (\sum_{j=1}^{k} B_j) \dfrac{(\sum_{i=1}^{k} \beta_i) - A}{2 \sum_{j=1}^{k} B_j}}
$$

$$
\geq \frac{(\sum_{i=1}^{k} \beta_i)(\min_{i \in \{1,\dots,\lfloor k/p \rfloor\}} (x_{mp+b-ip}))}{A + (\sum_{j=1}^{k} B_j) \dfrac{(\sum_{i=1}^{k} \beta_i) - A}{2 \sum_{j=1}^{k} B_j}}
$$

$$
\geq \frac{2 \sum_{i=1}^{k} \beta_i}{A + \sum_{i=1}^{k} \beta_i} \min_{i \in \{1,\dots,\lfloor k/p \rfloor\}} (x_{mp+b-ip}), \quad m \geq k.
$$

This is a difference inequality which holds for the subsequence $\{x_{mp+b}\}$ for $m \geq k$. We now rename this subsequence and apply the methods used in [Palladino 2008]. We set $z_m = x_{mp+b}$ for $m \in \mathbb{N}$. Since we have just shown that $\{z_m\}$ satisfies the difference inequality

$$
z_m \geq \frac{2 \sum_{i=1}^{k} \beta_i}{A + \sum_{i=1}^{k} \beta_i} \min_{i \in \{1,\dots,\lfloor k/p \rfloor\}} (z_{m-i}), \quad m \geq k,
$$

we can use the results of [Palladino 2008], particularly Theorem 3, to conclude that for $m \geq k$,

$$\min(z_{m-1}, \ldots, z_{m-\lfloor k/p \rfloor}) \geq \min(y_{\lfloor \frac{m-k}{\lfloor k/p \rfloor} \rfloor}, \ldots, y_{m-k}),$$

where $\{y_m\}_{m=0}^{\infty}$ is a solution of the difference equation

$$y_m = \frac{2 \sum_{i=1}^{k} \beta_i}{A + \sum_{i=1}^{k} \beta_i} y_{m-1}, \quad m \in \mathbb{N}, \tag{3}$$

with $y_0 = \min(z_{k-1}, \ldots, z_{k-\lfloor k/p \rfloor})$. Clearly every positive solution diverges to ∞ for the simple difference equation (3), since $A < \sum_{i=1}^{k} \beta_i$. Hence using the inequality we have obtained, $\{z_m\}_{m=1}^{\infty}$ diverges to ∞. Hence with given initial conditions, there is a subsequence of our solution $\{x_n\}_{n=1}^{\infty}$, namely $\{x_{mp+b}\}_{m=1}^{\infty}$, which diverges to ∞. Hence our solution $\{x_n\}_{n=1}^{\infty}$ is unbounded. So we have exhibited an unbounded solution whenever $A < \sum_{i=1}^{k} \beta_i$. □

Corollary 1. *Consider the fourth-order order rational difference equation*

$$x_n = \frac{\alpha + \delta x_{n-3}}{A + B x_{n-1} + C x_{n-2} + E x_{n-4}}, \quad n \in \mathbb{N}. \tag{4}$$

Assume nonnegative parameters and nonnegative initial conditions so that the denominator is nonvanishing. Further assume that $\delta, C, E > 0$.

  (i) *Whenever $A > \delta$, the unique equilibrium is globally asymptotically stable.*

 (ii) *Whenever $A = \delta$ and $\alpha > 0$, the unique equilibrium is globally asymptotically stable.*

(iii) *Whenever $A = \delta$ and $\alpha = 0$, every solution of Equation (4) converges to a periodic solution of period 3.*

(iv) *Whenever $A < \delta$, then Equation (4) has unbounded solutions for some choice of initial conditions.*

Proof. Cases (i), (ii), and (iii) were shown in [Palladino 2009b].

We now prove case (iv). Let us check Condition 1. Choose $B = \{0\}$ and $L = \{1, 2\}$. Condition 1(1) is satisfied since for all $b \in B$, namely $b = 0$, $\{(0 - j) \mod 3 : j \in \{2, 4\}\} = \{(0 - j) \mod 3 : j \in \{1, 2, 4\}\} = \{1, 2\}$. Condition 1(2) is satisfied since for $1 \in L$, there exists $4 \in I_B$ so that $(1 - 4) \mod 3 = -3 \mod 3 = 0 \in \{0\}$. Also for $2 \in L$, there exists $2 \in I_B$ so that $(2 - 2) \mod 3 = 0 \mod 3 = 0 \in \{0\}$. Furthermore

$$A < \delta = \sum_{i=1}^{k} \beta_i \quad \text{and} \quad \sum_{i=1}^{k} \beta_i = \delta > 0.$$

Thus Theorem 1 applies and so in case (iv) Equation (4) has unbounded solutions for some choice of initial conditions. □

Notice that Corollary 1 resolves conjectures 609, 611, 617, and 619 in [Camouzis and Ladas 2008] regarding boundedness character.

3. The equation $x_n = \dfrac{x_{n-3}}{Bx_{n-1} + x_{n-4}}$

In [Palladino 2009a] it is conjectured that the difference equation

$$x_n = \frac{x_{n-3}}{Bx_{n-1} + x_{n-4}}, \quad n \in \mathbb{N},$$

has unbounded solutions whenever $B > 0$. We show that whenever $B > 2^5$ unbounded solutions exist for some choice of nonnegative initial conditions. This does not fully establish the conjecture in [Palladino 2009a]. It does however establish the Conjecture 584 in [Camouzis and Ladas 2008]. We make use of the argument structure presented in Lemma 1 of [Lugo and Palladino ≥ 2009]. Let us repeat this lemma for the sake of the reader.

Lemma 1. *Let $\{x_n\}_{n=1}^{\infty}$ be a sequence in $[0, \infty)$. Suppose that there exists $D > 1$ and hypotheses $H_1, \ldots, H_k$ so that for all $n \in \mathbb{N}$ there exists $p_n \in \mathbb{N}$ so that the following holds. Whenever x_{n-i} satisfies H_i for all $i \in \{1, \ldots, k\}$, then x_{n+p_n-i} satisfies H_i for all $i \in \{1, \ldots, k\}$ and $x_{n+p_n-1} \geq Dx_{n-1}$. Further assume that for some $N \in \mathbb{N}$, x_{N-i} satisfies H_i for all $i \in \{1, \ldots, k\}$ and $x_{N-1} > 0$. Then $\{x_n\}_{n=1}^{\infty}$ is unbounded. Particularly $\{x_{z_m-1}\}_{m=1}^{\infty}$ is a subsequence of $\{x_n\}_{n=1}^{\infty}$ which diverges to ∞, where $z_m = z_{m-1} + p_{z_{m-1}}$ and $z_0 = N$.*

Proof. Let $z_m = z_{m-1} + p_{z_{m-1}}$ and $z_0 = N$. Using induction, we prove that given $m \in \mathbb{N}$ the following holds. $x_{z_m-1} \geq D^m x_{N-1}$ and x_{z_m-i} satisfies H_i for all $i \in \{1, \ldots, k\}$. By assumption, x_{N-i} satisfies H_i for all $i \in \{1, \ldots, k\}$ and $x_{N-1} \geq D^0 x_{N-1}$. This provides the base case. Assume $x_{z_{m-1}-i}$ satisfies H_i for all $i \in \{1, \ldots, k\}$ and $x_{z_{m-1}-1} \geq D^{m-1}x_{N-1}$. Using our earlier assumption this implies that there exists $p_{z_{m-1}}$ so that $x_{z_{m-1}+p_{z_{m-1}}-i}$ satisfies H_i for all $i \in \{1, \ldots, k\}$ and $x_{z_{m-1}+p_{z_{m-1}}-1} \geq Dx_{z_{m-1}-1} \geq (D)D^{m-1}x_{N-1} = D^m x_{N-1}$.

So we have shown that $x_{z_m-1} \geq D^m x_{N-1}$ for all $m \in \mathbb{N}$. Hence the subsequence $\{x_{z_m-1}\}_{m=1}^{\infty}$ of $\{x_n\}_{n=1}^{\infty}$ clearly diverges to ∞ since $D > 1$. □

Theorem 2. *Consider the fourth order rational difference equation,*

$$x_n = \frac{x_{n-3}}{Bx_{n-1} + x_{n-4}}, \quad n \in \mathbb{N}. \tag{5}$$

Suppose $B > 2^5$. Then Equation (5) has unbounded solutions for some initial conditions.

Proof. We choose initial conditions so that

$$x_{-2} > B, \qquad x_{-3} < \tfrac{1}{4},$$

and one of the following holds:

(1) $x_0 < \dfrac{1}{4B}$ and $x_{-1} < \dfrac{1}{B}$;

(2) $\dfrac{1}{4B} \leq x_0 \leq 2x_{-2}$ and $x_{-1} < \dfrac{1}{B^2 x_{-2}}$;

(3) $x_0 > 2x_{-2}$ and $x_{-1} < \dfrac{1}{B^2 x_{-2}}$.

We show that there exists $D = 2$ so that for all $n \in \mathbb{N}$ there exists $p_n \in \{2, 3, 5\}$ so that the following holds.

Whenever

$$x_{n-3} > B, \qquad x_{n-4} < \tfrac{1}{4},$$

and one of the following holds:

(1) $x_{n-1} < \dfrac{1}{4B}$ and $x_{n-2} < \dfrac{1}{B}$;

(2) $\dfrac{1}{4B} \leq x_{n-1} \leq 2x_{n-3}$ and $x_{n-2} < \dfrac{1}{B^2 x_{n-3}}$;

(3) $x_{n-1} > 2x_{n-3}$ and $x_{n-2} < \dfrac{1}{B^2 x_{n-3}}$;

then we have

$$x_{n+p_n-3} > Dx_{n-3} > B, \qquad x_{n+p_n-4} < \tfrac{1}{4},$$

and one of the following holds:

(1) $x_{n+p_n-1} < \dfrac{1}{4B}$ and $x_{n+p_n-2} < \dfrac{1}{B}$;

(2) $\dfrac{1}{4B} \leq x_{n+p_n-1} \leq 2x_{n+p_n-3}$ and $x_{n+p_n-2} < \dfrac{1}{B^2 x_{n+p_n-3}}$;

(3) $x_{n+p_n-1} > 2x_{n+p_n-3}$ and $x_{n+p_n-2} < \dfrac{1}{B^2 x_{n+p_n-3}}$.

First assume

$$x_{n-1} < \dfrac{1}{4B}, \qquad x_{n-2} < \dfrac{1}{B}, \qquad x_{n-3} > B, \qquad x_{n-4} < \tfrac{1}{4}.$$

In this case $p_n = 3$. Since $B > 2^5$ we have

$$x_{n+p_n-4} = x_{n-1} < \dfrac{1}{4B} < \tfrac{1}{4}.$$

Since $x_{n-4} < \tfrac{1}{4}$ and $x_{n-1} < \tfrac{1}{4B}$ we have

$$x_{n+p_n-3} = x_n = \dfrac{x_{n-3}}{Bx_{n-1} + x_{n-4}} \geq \dfrac{x_{n-3}}{2 \max\left(Bx_{n-1}, x_{n-4}\right)} > 2x_{n-3} > B.$$

Since $x_{n-2} < \frac{1}{B}$,

$$x_{n+p_n-2} = x_{n+1} = \frac{x_{n-2}}{Bx_n + x_{n-3}} \le \frac{x_{n-2}}{Bx_n} < \frac{1}{B^2 x_n} < \frac{1}{B^3} < \frac{1}{B}.$$

Hence regardless of the value of x_{n+p_n-1} one of our requirements is satisfied. If $x_{n+p_n-1} < \frac{1}{4B}$ then requirement (1) is satisfied. If $\frac{1}{4B} \le x_{n+p_n-1} \le 2x_{n+p_n-3}$ then requirement (2) is satisfied. If $x_{n+p_n-1} > 2x_{n+p_n-3}$ then requirement (3) is satisfied.

Next assume

$$\frac{1}{4B} \le x_{n-1} \le 2x_{n-3}, \qquad x_{n-2} < \frac{1}{B^2 x_{n-3}}, \qquad x_{n-3} > B, \qquad x_{n-4} < \frac{1}{4}.$$

In this case $p_n = 5$. Since $B > 2^5$ we have

$$x_{n+p_n-4} = x_{n+1} = \frac{x_{n-2}}{Bx_n + x_{n-3}} < \frac{x_{n-2}}{x_{n-3}} < \frac{1}{B^2 x_{n-3}^2} < \frac{1}{4}.$$

Since $x_{n-2} < \frac{1}{B^2 x_{n-3}}$ and $B > 2^5$ we have

$$x_{n+p_n-3} = x_{n+2} = \frac{x_{n-1}}{Bx_{n+1} + x_{n-2}} \ge \frac{x_{n-1}}{2 \max (Bx_{n+1}, x_{n-2})}$$

$$> \frac{x_{n-1}}{2 \max \left(\frac{1}{Bx_{n-3}^2}, \frac{1}{B^2 x_{n-3}}\right)} \ge \frac{B^2 x_{n-3}}{8B} > 2x_{n-3} > B.$$

Also notice that

$$x_{n+p_n-2} = x_{n+3} = \frac{x_n}{Bx_{n+2} + x_{n-1}} = \frac{x_{n-3}}{(Bx_{n+2} + x_{n-1})(Bx_{n-1} + x_{n-4})}$$

$$< \frac{x_{n-3}}{Bx_{n+2}(Bx_{n-1} + x_{n-4})} < \frac{8x_{n-3}}{B^2 x_{n-3}(Bx_{n-1} + x_{n-4})} < \frac{8}{B^3 x_{n-1}} < \frac{2^5}{B^2} < \frac{1}{B}.$$

Notice that

$$x_{n+p_n-1} = x_{n+4} = \frac{x_{n+1}}{Bx_{n+3} + x_n} < \frac{1}{(B^2 x_{n-3}^2)(Bx_{n+3} + x_n)} < \frac{1}{B^2 x_{n-3}^2 x_n}$$

$$= \frac{Bx_{n-1} + x_{n-4}}{B^2 x_{n-3}^3} < \frac{2Bx_{n-3} + .25}{B^2 x_{n-3}^3} < \frac{3}{Bx_{n-3}^2} < \frac{1}{4B}.$$

Hence requirement (1) is satisfied in this case. Finally assume

$$x_{n-1} > 2x_{n-3}, \qquad x_{n-2} < \frac{1}{B^2 x_{n-3}}, \qquad x_{n-3} > B, \qquad x_{n-4} < \frac{1}{4}.$$

In this case $p_n = 2$. Immediately we have

$$x_{n+p_n-4} = x_{n-2} < \frac{1}{B^2 x_{n-3}} < \frac{1}{4}.$$

Also by assumption,

$$x_{n+p_n-3} = x_{n-1} > 2x_{n-3} > B.$$

Further since $x_{n-1} > 2x_{n-3}$,

$$x_{n+p_n-2} = x_n = \frac{x_{n-3}}{Bx_{n-1} + x_{n-4}} < \frac{x_{n-3}}{Bx_{n-1}} < \frac{1}{2B} < \frac{1}{B}.$$

Furthermore

$$x_{n+p_n-1} = x_{n+1} = \frac{x_{n-2}}{Bx_n + x_{n-3}} < \frac{x_{n-2}}{x_{n-3}} < \frac{1}{B^2 x_{n-3}^2} < \frac{1}{4B}.$$

Hence requirement (1) is satisfied in this case, so after an application of Lemma 1 the proof is complete. □

4. Conclusion

As noted in the introduction, Theorem 2 partly resolves Conjecture 2 in [Palladino 2009a]; the latter, however, is only part of a larger conjecture, namely Conjecture 1 in the same reference. For convenience we restate this conjecture.

Conjecture 1. Consider the k-th order rational difference equation

$$x_n = \frac{\sum_{i=1}^{k} \beta_i x_{n-i}}{\sum_{j=1}^{k} B_j x_{n-j}}, \quad n \in \mathbb{N}. \tag{6}$$

Assume nonnegative parameters and nonnegative initial conditions so that the denominator is nonvanishing. Further assume that $\sum_{i=1}^{k} \beta_i > 0$ and that there does not exist $j \in I_B$ such that $\gcd(I_\beta) \mid j$. Then unbounded solutions of Equation (6) exist for some initial conditions.

It would be interesting to study this conjecture further utilizing techniques similar to that used in Theorem 2.

5. Acknowledgements

References

[Camouzis and Ladas 2008] E. Camouzis and G. Ladas, *Dynamics of third-order rational difference equations with open problems and conjectures*, Adv. Disc. Math. Appl. **5**, Chapman & Hall/CRC, Boca Raton, FL, 2008. MR 2008h:39001 Zbl 1129.39002

[Lugo and Palladino ≥ 2009] G. Lugo and F. J. Palladino, "Unboundedness for some classes of rational difference equations", *Int. J. Difference Equ*. To appear.

[Palladino 2008] F. J. Palladino, "Difference inequalities, comparison tests, and some consequences", *Involve* **1**:1 (2008), 91–100. MR 2403068

[Palladino 2009a] F. J. Palladino, "On periodic trichotomies", *J. Difference Equa. Appl.* **15** (2009). To appear.

[Palladino 2009b] F. J. Palladino, "On the characterization of rational difference equations", *J. Difference Equa. Appl.* **15** (2009), 253–260.

glugo@mail.uri.edu   *Department of Mathematics, University of Rhode Island,
                                    5 Lippitt Road, Kingston, RI 02881, United States*

Guidelines for Authors

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, Corel Draw, 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@mathscipub.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