involve

a journal of mathematics

mathematical sciences publishers

involve

pjm.math.berkeley.edu/inv

PRODUCTION

inv@mathscipub.org

Paulo Ney de Souza : Production Manager   Silvio Levy : Production Editor

Cover design: Alex Scorpan

See inside back cover or http://pjm.math.berkeley.edu/inv for submission instructions.
Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to Mathematical Sciences Publishers, Department of Mathematics, University of California, Berkeley, CA 94704-3840, USA.

Involve, at Mathematical Sciences Publisher, Department of Mathematics, University of California, Berkeley, CA 94720-3840 is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

PUBLISHED BY
mathematical sciences publishers
http://www.mathscipub.org
A NON-PROFIT CORPORATION
Typeset in LATEX

INVOLVE 1:1(2008)

Five-point boundary value problems for n-th order differential equations by solution matching

Johnny Henderson, John Ehrke and Curtis Kunkel

(Communicated by Kenneth S. Berenhaut)

For the ordinary differential equation

$$y^{(n)} = f(x, y, y', y'', \ldots, y^{(n-1)}), \qquad n \geq 3,$$

solutions of three-point boundary value problems on $[a, b]$ are matched with solutions of three-point boundary value problems on $[b, c]$ to obtain solutions satisfying five-point boundary conditions on $[a, c]$.

1. Introduction

We are concerned with the existence and uniqueness of solutions of boundary value problems on an interval $[a, c]$ for the n-th order ordinary differential equation

$$y^{(n)} = f(x, y, y', y'', \ldots, y^{(n-1)}), \tag{1}$$

satisfying the five-point boundary conditions

$$\begin{gathered} y(a) - y(x_1) = y_1, \qquad y^{(i-1)}(b) = y_{i+1}, \quad 1 \leq i \leq n-2, \\ y(x_2) - y(c) = y_n, \end{gathered} \tag{2}$$

where $a < x_1 < b < x_2 < c$ and $y_1, \ldots, y_n \in \mathbb{R}$.

It is assumed throughout that $f : [a, c] \times \mathbb{R}^n \to \mathbb{R}$ is continuous and that solutions of initial value problems for (1) are unique and exist on all of $[a, c]$. Moreover, the points $a < x_1 < b < x_2 < c$ are fixed throughout.

Nonlocal boundary value problems, for which the number of boundary points is possibly greater than the order of the ordinary differential equation, have received considerable interest. For a small sample of such works, we refer the reader to the papers by Bai and Fang [2003], Gupta [1997], Gupta and Trofimchuk [1998], Infante [2005], Ma [1997; 2002] and Webb [2005].

MSC2000: primary 34B15; secondary 34B10.

Keywords: boundary value problem, ordinary differential equation, solution matching.

Monotonicity conditions will be imposed on f. Sufficient conditions will be given such that, if $y_1(x)$ is a solution of a three-point boundary value problem on $[a, b]$, and if $y_2(x)$ is a solution of another three-point boundary value problem on $[b, c]$, then $y(x)$ defined by

$$y(x) = \begin{cases} y_1(x), & a \le x \le b, \\ y_2(x), & b \le x \le c, \end{cases}$$

will be a desired unique solution of (1), (2). In particular, a monotonicity condition is imposed on $f(x, r_1, \ldots, r_n)$ insuring that each three-point boundary value problem for (1) satisfying any one of the following conditions:

$$\begin{gathered} y(a) - y(x_1) = y_1, \quad y^{(i-1)}(b) = y_{i+1}, \quad 1 \le i \le n-2, \\ y^{(n-2)}(b) = m, \quad m \in \mathbb{R}, \end{gathered} \tag{3}$$

$$\begin{gathered} y(a) - y(x_1) = y_1, \quad y^{(i-1)}(b) = y_{i+1}, \quad 1 \le i \le n-2, \\ y^{(n-1)}(b) = m, \quad m \in \mathbb{R}, \end{gathered} \tag{4}$$

$$\begin{gathered} y^{(i-1)}(b) = y_{i+1}, \quad 1 \le i \le n-2, \quad y^{(n-2)}(b) = m, \\ y(x_2) - y(c) = y_n, \quad m \in \mathbb{R} \end{gathered} \tag{5}$$

or

$$\begin{gathered} y^{(i-1)}(b) = y_{i+1}, \quad 1 \le i \le n-2, \quad y^{(n-1)}(b) = m, \\ y(x_2) - y(c) = y_n, \quad m \in \mathbb{R} \end{gathered} \tag{6}$$

has at most one solution.

We will impose an additional hypothesis that solutions for (1) satisfying any of (3), (4), (5) or (6) exist. Then we will construct a unique solution of (1), (2).

Solution matching techniques were first used by Bailey et al. [1968]. They considered solutions of two-point boundary value problems for the second order equation $y''(x) = f(x, y(x), y'(x))$ by matching solution of initial value problems. Since then, there have been numerous papers in which solutions of two-point boundary value problems on $[a, b]$ were matched with solutions of two-point boundary value problems on $[b, c]$ to obtain solutions of three-point boundary value problems on $[a, c]$. See, for example [Barr and Miletta 1974; Das and Lalli 1981; Henderson 1983; Moorti and Garner 1978; Rao et al. 1981]. In 1973, Barr and Sherman [1973] used solution matching techniques to obtain solutions of three-point boundary value problems for third order differential equations from solutions of two-point problems. They also generalized their results to equations of arbitrary order by obtaining solutions of n-th equations. More recently, Henderson and Prasad [2001] and Eggensperger et al. [2004] used matching methods for solutions of multipoint boundary value problems on time scales. Finally, Henderson and

Tisdale [2005] adapted the matching methods to obtain solutions of five-point problems for third order equations. The present work extends the results of Henderson and Tisdale [2005] to n-th order five-point boundary value problems (1), (2) on $[a, c]$.

The monotonicity hypothesis on f which will play a fundamental role in uniqueness of solutions (and later existence as well), is given by:

(A) For all $w \in \mathbb{R}$,

$$f(x, v_1, \dots, v_{n-2}, v_{n-1}, w) > f(x, u_1, \dots, u_{n-2}, u_{n-1}, w),$$

(a) when $x \in (a, b]$, $(-1)^{n-i}u_i \geq (-1)^{n-i}v_i$, $1 \leq i \leq n-2$, and $v_{n-1} > u_{n-1}$, or

(b) when $x \in [b, c)$, $v_i \geq u_i$, $1 \leq i \leq n-2$, and $v_{n-1} > u_{n-1}$.

2. Uniqueness of solutions

In this section, we establish that under condition (A) solutions of the three-point boundary value problems, as well as the five-point problem are unique when they exist.

Theorem 2.1. *Let $y_1, \dots, y_n \in \mathbb{R}$ be given and assume condition* (A) *is satisfied. Then, given $m \in \mathbb{R}$, each of the boundary value problems for* (1) *satisfying any of conditions* (3), (4), (5) *or* (6) *has at most one solution.*

Proof. We will establish the result only for (1), (3). Arguments for the other boundary value problems are very similar.

In order to reach a contradiction, we assume that for some $m \in \mathbb{R}$, there are distinct solutions, α and β, of (1), (3), and set $w = \alpha - \beta$. Then

$$w(a) - w(x_1) = w^{(i-1)}(b) = 0, \qquad 1 \leq i \leq n-1.$$

By the uniqueness of solutions of initial value problems for (1), we may assume with no loss of generality that $w^{(n-1)}(b) < 0$. It follows from the boundary conditions satisfied by w that there exists $a < r < b$ such that

$$w^{(n-1)}(r) = 0 \quad \text{and} \quad w^{(n-1)}(x) < 0 \text{ on } (r, b].$$

Since $w^{(i-1)}(b) = 0,\ 1 \leq i \leq n-1$, it follows in turn that

$$(-1)^{n-j}w^{(j)}(x) > 0, \quad 0 \leq j \leq n-2, \text{ on } [r, b).$$

This leads to

$$w^{(n)}(r) = \lim_{x \to r^+} \frac{w^{(n-1)}(x)}{x - r} \leq 0.$$

However, from condition (A), we have

$$\begin{aligned} w^{(n)}(r) &= \alpha^{(n)}(r) - \beta^{(n)}(r) \\ &= f(r, \alpha(r), \alpha'(r), \dots, \alpha^{(n-2)}(r), \alpha^{(n-1)}(r)) \\ &\qquad - f(r, \beta(r), \beta'(r), \dots, \beta^{(n-2)}(r), \beta^{(n-1)}(r)) \\ &= f(r, \alpha(r), \alpha'(r), \dots, \alpha^{(n-2)}(r), \alpha^{(n-1)}(r)) \\ &\qquad - f(r, \beta(r), \beta'(r), \dots, \beta^{(n-2)}(r), \alpha^{(n-1)}(r)) \\ &> 0, \end{aligned}$$

which is a contradiction. Thus, (1), (3) has at most one solution. The proof is complete. □

Theorem 2.2. *Let $y_1, \dots, y_n \in \mathbb{R}$ be given. Assume condition* (A) *is satisfied. Then, the boundary value problem* (1), (2) *has at most one solution.*

Proof. Again, we argue by contradiction. Assume for some values $y_1, \dots, y_n \in \mathbb{R}$, there exist distinct solutions α and β of (1) and (2). Also, let $w = \alpha - \beta$. Then

$$w(a) - w(x_1) = w^{(i-1)}(b) = w(x_2) - w(c) = 0, \quad 1 \le i \le n-2.$$

By Theorem 2.1, $w^{(n-2)}(b) \neq 0$ and $w^{(n-1)}(b) \neq 0$. We assume with no loss of generality that $w^{(n-2)}(b) > 0$. Then, from the boundary conditions satisfied by w, there are points $a < r_1 < b < r_2 < c$ so that

$$w^{(n-2)}(r_1) = w^{(n-2)}(r_2) = 0, \quad \text{and} \quad w^{(n-2)}(x) > 0 \text{ on } (r_1, r_2).$$

There are two cases to analyze, that is, $w^{(n-1)}(b) > 0$ and $w^{(n-1)}(b) < 0$. The arguments for the two cases are completely analagous, therefore we will treat only the first case $w^{(n-1)}(b) > 0$. In view of the fact that $w^{(n-2)}(b) > 0$ and $w^{(n-2)}(r_2) = 0$, there exists $b < r < r_2$ so that

$$w^{(n-1)}(r) = 0, \quad \text{and} \quad w^{(n-1)}(x) > 0 \text{ on } [b, r).$$

Then

$$w^{(j)}(x) > 0, \quad 0 \le j \le n-2, \text{ on } (b, r].$$

This leads to

$$w^{(n)}(r) = \lim_{x \to r^-} \frac{w^{(n-1)}(x)}{x - r} \le 0.$$

However, again from condition (A), we have

$$\begin{aligned} w^{(n)}(r) &= \alpha^{(n)}(r) - \beta^{(n)}(r) \\ &= f(r, \alpha(r), \alpha'(r), \ldots, \alpha^{(n-2)}(r), \alpha^{(n-1)}(r)) \\ &\qquad - f(r, \beta(r), \beta'(r), \ldots, \beta^{(n-2)}(r), \beta^{(n-1)}(r)) \\ &= f(r, \alpha(r), \alpha'(r), \ldots, \alpha^{(n-2)}(r), \alpha^{(n-1)}(r)) \\ &\qquad - f(r, \beta(r), \beta'(r), \ldots, \beta^{(n-2)}(r), \alpha^{(n-1)}(r)) \\ &> 0, \end{aligned}$$

which contradicts the initial assumption. Thus, (1), (2) has at most one solution, and the proof is complete. □

3. Existence of solutions

In this section, we show that solutions of (1) satisfying each of (3), (4), (5) and (6) are monotone functions of m. Then, we use these monotonicity properties to obtain solutions of (1), (2).

For notation purposes, given $m \in \mathbb{R}$, let $\alpha(x, m)$, $u(x, m)$, $\beta(x, m)$ and $v(x, m)$ denote the solutions, when they exist, of the boundary value problems for (1) satisfying, respectively, (3), (4), (5) and (6).

Theorem 3.1. *Suppose that the monotonicity hypothesis* (A) *is satisfied and that, for each $m \in \mathbb{R}$, there exist solutions of* (1) *satisfying each of the conditions* (3), (4), (5) *and* (6). *Then, $\alpha^{(n-1)}(b, m)$ and $\beta^{(n-1)}(b, m)$ are, respectively, strictly increasing and decreasing functions of m with ranges all of $\mathbb{R}$.*

Proof. The "strictness" of the conclusion arises from Theorem 2.1. Let $m_1 > m_2$ and let $w(x) = \alpha(x, m_1) - \alpha(x, m_2)$. Then,

$$w(x_1) - w(a) = w^{(i-1)}(b) = 0, \ 1 \le i \le n-2, \ w^{(n-2)}(b) = m_1 - m_2 > 0$$

and $w^{(n-1)}(b) \neq 0$.

Contrary to the conclusion, assume $w^{(n-1)}(b) < 0$. Since there exists $a < r_1 < b$ so that $w^{(n-2)}(r_1) = 0$ and $w^{(n-2)}(x) > 0$ on $(r_1, b]$, it follows that there exists $r_1 < r_2 < b$ such that

$$w^{(n-1)}(r_2) = 0 \quad \text{and} \quad w^{(n-1)}(x) < 0 \text{ on } (r_2, b].$$

We also have

$$(-1)^{n-j} w^{(j)}(x) > 0, \quad 0 \le j \le n-2 \text{ on } [r_2, b).$$

As in the other proofs above, we arrive at the same contradiction, that is, $w^{(n)}(r_2) \le 0$ and $w^{(n)}(r_2) > 0$. Thus, $w^{(n-1)}(b) > 0$ and, as a consequence, $\alpha^{(n-1)}(b, m)$ is a strictly increasing function of m.

We next argue that $\{\alpha^{(n-1)}(b,m) \mid m \in \mathbb{R}\} = \mathbb{R}$. Let $k \in \mathbb{R}$ and consider the solution $u(x,k)$ of (1), (4) with u as defined above. Consider also the solution $\alpha(x, u^{(n-2)}(b,k))$ of (1), (3). Then $\alpha(x, u^{(n-2)}(b,k))$ and $u(x,k)$ are solutions of the same type boundary value problems (1), (3). Hence by Theorem 2.1, the functions are identical. Therefore,

$$\alpha^{(n-1)}(b, u^{(n-2)}(b,k)) = u^{(n-1)}(b,k) = k,$$

and the range of $\alpha^{(n-1)}(b,m)$, as a function of m, is the set of real numbers.

The argument for $\beta^{(n-1)}(b,m)$ is quite similar. This completes the proof. □

In a similar way, we also have a monotonicity result on $(n-2)$-derivatives of $u(x,m)$ and $v(x,m)$.

Theorem 3.2. *Assume the hypotheses of Theorem 3.1. Then, $u^{(n-2)}(b,m)$ and $v^{(n-2)}(b,m)$ are, respectively, strictly increasing and decreasing functions of m with ranges all of $\mathbb{R}$.*

We now provide our existence result.

Theorem 3.3. *Assume the hypotheses of Theorem 3.1. Then (1), (2) has a unique solution.*

Proof. The existence is immediate from either Theorem 3.1 or Theorem 3.2. Making use of Theorem 3.2, there exists a unique $m_0 \in \mathbb{R}$ such that $u^{(n-2)}(b,m_0) = v^{(n-2)}(b,m_0)$. Then

$$y(x) = \begin{cases} u(x,m_0), & a \le x \le b, \\ v(x,m_0), & b \le x \le c, \end{cases}$$

is a solution of (1), (2). By Theorem 2.2, $y(x)$ is the unique solution. The proof is complete. □

References

[Bai and Fang 2003] C. Bai and J. Fang, "Existence of multiple positive solutions for nonlinear m-point boundary value problems", *J. Math. Anal. Appl.* **281**:1 (2003), 76–85. MR 2004b:34035 Zbl 1030.34026

[Bailey et al. 1968] P. B. Bailey, L. F. Shampine, and P. E. Waltman, *Nonlinear two point boundary value problems*, Mathematics in Science and Engineering, Vol. 44, Academic Press, New York, 1968. MR 37 #6524

[Barr and Miletta 1974] D. Barr and P. Miletta, "An existence and uniqueness criterion for solutions of boundary value problems", *J. Differential Equations* **16**:3 (1974), 460–471. MR 54 #7933 Zbl 0289.34020

[Barr and Sherman 1973] D. Barr and T. Sherman, "Existence and uniqueness of solutions of three-point boundary value problems", *J. Differential Equations* **13** (1973), 197–212. MR 48 #11651 Zbl 0261.34014

[Das and Lalli 1981] K. M. Das and B. S. Lalli, "Boundary value problems for $y'''=f(x, y, y', y'')$", *J. Math. Anal. Appl.* **81**:2 (1981), 300–307. MR 82i:34018 Zbl 0465.34012

[Eggensperger et al. 2004] M. Eggensperger, E. R. Kaufmann, and N. Kosmatov, "Solution matching for a three-point boundary-value problem on a time scale", *Electron. J. Differential Equations* (2004), No. 91, 7 pp. (electronic). MR 2005b:34033

[Gupta 1997] C. P. Gupta, "A nonlocal multipoint boundary-value problem at resonance", pp. 253–259 in *Advances in nonlinear dynamics*, Stability Control Theory Methods Appl. **5**, Gordon and Breach, Amsterdam, 1997. MR 98g:34034 Zbl 0922.34013

[Gupta and Trofimchuk 1998] C. P. Gupta and S. I. Trofimchuk, "Solvability of a multi-point boundary value problem and related a priori estimates", *Canad. Appl. Math. Quart.* **6**:1 (1998), 45–60. Geoffrey J. Butler Memorial Conference in Differential Equations and Mathematical Biology (Edmonton, AB, 1996). MR 99f:34020

[Henderson 1983] J. Henderson, "Three-point boundary value problems for ordinary differential equations by matching solutions", *Nonlinear Anal.* **7**:4 (1983), 411–417. MR 84j:34014

[Henderson and Prasad 2001] J. Henderson and K. R. Prasad, "Existence and uniqueness of solutions of three-point boundary value problems on time scales by solution matching", *Nonlinear Stud.* **8**:1 (2001), 1–12. MR 2002f:34031

[Henderson and Tisdale 2005] J. Henderson and C. C. Tisdale, "Five-point boundary value problems for third-order differential equations by solution matching", *Math. Comput. Modelling* **42**:1-2 (2005), 133–137. MR 2006e:34033

[Infante 2005] G. Infante, "Positive solutions of some three-point boundary value problems via fixed point index for weakly inward A-proper maps", *Fixed Point Theory Appl.* **2005**:2 (2005), 177–184. MR 2006j:34045 Zbl 05038342

[Ma 1997] R. Ma, "Existence theorems for a second order three-point boundary value problem", *J. Math. Anal. Appl.* **212**:2 (1997), 430–442. MR 98h:34041

[Ma 2002] R. Ma, "Existence of positive solutions for second order m-point boundary value problems", *Ann. Polon. Math.* **79**:3 (2002), 265–276. MR 2004a:34037

[Moorti and Garner 1978] V. R. G. Moorti and J. B. Garner, "Existence-uniqueness theorems for three-point boundary value problems for nth-order nonlinear differential equations", *J. Differential Equations* **29**:2 (1978), 205–213. MR 58 #11598

[Rao et al. 1981] D. R. K. S. Rao, K. N. Murthy, and A. S. Rao, "Three-point boundary value problems associated with third order differential equations", *Nonlinear Anal.* **5**:6 (1981), 669–673. MR 82f:34016

[Webb 2005] J. R. L. Webb, "Optimal constants in a nonlocal boundary value problem", *Nonlinear Anal.* **63**:5-7 (2005), 672–685. MR 2006j:34060

Received: 2007-04-02 Revised: Accepted: 2007-10-27

Johnny_Henderson@baylor.edu
Department of Mathematics, Baylor University,
Waco, TX 76798-7328, United States
http://www.baylor.edu/math/index.php?id=22228

john.ehrke@acu.edu
Department of Mathematics, Abilene Christian University,
Abilene, TX 79699-8012, United States

ckunkel@utm.edu
Department of Mathematics and Statistics, 424
Humanities Building, University of Tennessee at Martin,
Martin, TN 38238, United States

Parity of the partition function and the modular discriminant

Sally Wolfe

(Communicated by Ken Ono)

We relate the parity of the partition function to the parity of the q-series coefficients of certain powers of the modular discriminant using their generating functions. This allows us to make statements about the parity of the initial values of the partition function and to obtain a modified Euler recurrence for its parity.

1. Introduction and statement of results

We begin by defining two power series in q, the power series of the modular discriminant, and the generating function of the partition function, $p(n)$. The q-series expansion of the modular discriminant $\Delta(q)$ defines the Ramanujan τ-function. Namely, we have that

$$\begin{aligned}\Delta(q) &= q\prod_{n=1}^{\infty}(1-q^n)^{24} = \sum_{n=0}^{\infty}\tau(n)q^n \\ &= q-24q^2+252q^3-1472q^4+4830q^5-6048q^6-16744q^7\cdots .\end{aligned} \tag{1.1}$$

Ramanujan investigated $\tau(n)$ and observed that $\tau(nm)=\tau(n)\tau(m)$ for $(n,m)=1$, as well as congruences like $\tau(n)\equiv\sum_{d|n}d^{11} \pmod{691}$.

The partition function counts the number of distinct partitions of integers n. Like $\Delta(q)$,the generating function for $p(n)$ is an infinite product. More precisely, we have

$$P(q)=\sum_{n=0}^{\infty}p(n)q^n=\frac{1}{\prod_{n=1}^{\infty}(1-q^n)}=1+q+2q^2+3q^3+5q^4+7q^5+11q^6\cdots . \tag{1.2}$$

MSC2000: 05A17.

Keywords: partition, congruences.

Ramanujan proved that for all nonnegative integers n

$$p(5n+4) \equiv 0 \pmod{5}, \tag{1.3}$$
$$p(7n+5) \equiv 0 \pmod{7}, \tag{1.4}$$
$$p(11n+6) \equiv 0 \pmod{11}. \tag{1.5}$$

However, much less is known about $p(n) \pmod{2}$. For example, it is conjectured that as x approaches infinity, the number of even and odd values of $p(n)$ with $n \le x$ approaches $\frac{1}{2}x$. Nicolas et al. [1998] prove that as $x \to \infty$,

$$\#\{n \le x : p(n) \equiv 0 \pmod{2}\} \gg \sqrt{x}$$
$$\#\{n \le x : p(n) \equiv 1 \pmod{2}\} \gg \sqrt{x} \cdot e^{\frac{-(\log 2+\epsilon)\log x}{\log\log x}}.$$

Ahlgren [1999] proves a slightly better bound for the number of odd values of $p(n)$: for sufficiently large x,

$$\#\{n \le x : p(n) \equiv 1 \pmod{2}\} \gg \frac{\sqrt{x}}{\log x}.$$

Nicolas [2006] proves that there exists a constant $\kappa > 0$ such that for sufficiently large x,

$$\#\{n \le x : p(n) \equiv 1 \pmod{2}\} \gg \frac{\sqrt{x}(\log\log x)^{\kappa}}{\log x}. \tag{1.6}$$

He proves this bound for all $\kappa > 0$ and sufficiently large x [Nicolas 2008], as well as proving a bound for the number of even values of $p(n)$ up to x:

$$\#\{n \le x : p(n) \equiv 0 \pmod{2}\} \gg 0.28\sqrt{x \log\log x} \tag{1.7}$$

The purpose of this paper is to investigate the parity of $p(n)$. We first recall Euler's recurrence for $p(n)$ [Andrews 1971]. If n is a positive integer, then

$$p(n) = \sum_{k\ge 1}(-1)^{k+1} p\Big(n - \frac{3k^2+k}{2}\Big) + \sum_{k\ge 1}(-1)^{k+1} p\Big(n - \frac{3k^2-k}{2}\Big).$$

We deform this to obtain many recurrences for $p(n) \pmod{2}$.

Theorem 1.1. *For integers $s \ge 2$, we have*:

$$\Delta(q)^{\frac{4^s-1}{3}} \equiv \Big(\sum_{n=0}^{\infty} p(n) q^{8n+\frac{4^s-1}{3}}\Big)\Big(\sum_{n=-\infty}^{\infty} q^{4^{s+1}(3n^2-n)}\Big) \pmod{2}.$$

To state the next theorem, we let $\tau_m(n)$ denote the n^{th} coefficient of $\Delta(q)^m$.

Theorem 1.2. *If $s \geq 2$ is an integer, then for any positive integer n we have*

$$p(n) \equiv \tau_{\frac{4^s-1}{3}}\left(8n+\frac{4^s-1}{3}\right) + \sum_{m=1}^{\lfloor \frac{1}{6}+\frac{1}{3\cdot 2^s}\sqrt{4^{s-1}+6n}\rfloor} p(n-2^{2s-1}(3m^2-m))$$
$$+ \sum_{m=1}^{\lfloor -\frac{1}{6}+\frac{1}{3\cdot 2^s}\sqrt{4^{s-1}+6n}\rfloor} p(n-2^{2s-1}(3m^2+m)) \pmod 2.$$

Remark 1. For n such that $\tau_{(4^s-1)/3}(n) \equiv 0 \pmod 2$, this gives an *Euler-type* recurrence. We note that it is known [Serre 1974] that

$$\lim_{x\to\infty} \frac{\#\{n \leq x : \tau_{(4^s-1)/3}(n) \equiv 0 \pmod 2\}}{x} = 1.$$

Therefore, for almost all n, we have

$$p(n) \equiv \sum_{m=1}^{\lfloor \frac{1}{6}+\frac{1}{3\cdot 2^s}\sqrt{4^{s-1}+6n}\rfloor} p(n-2^{2s-1}(3m^2-m)) + \sum_{m=1}^{\lfloor -\frac{1}{6}+\frac{1}{3\cdot 2^s}\sqrt{4^{s-1}+6n}\rfloor} p(n-2^{2s-1}(3m^2+m)) \pmod 2.$$

In order to state the next theorem, we define a function which counts the number of representations of an integer n by certain t-ary quadratic forms:

$$r_t(n) = \#\{n = x_1^2 + 4x_2^2 + \cdots 4^{t-1}x_t^2 : x_i \text{ are positive odd integers}\}.$$

Theorem 1.3. *If n is a positive integer, then for $s \geq 2$, we have*

$$\tau_{(4^s-1)/3}(n) \equiv r_s(n) \pmod 2.$$

Now we turn to some applications of Theorem 1.1. In particular, we study the case of $s = 2$ where we can determine $\tau_5(8n+5) \pmod 2$.

Theorem 1.4. *If n is an integer, then*

$$\tau_5(8n+5) \equiv$$
$$\begin{cases} 1 \pmod 2 & \textit{if } 8n+5 = k\cdot l^2, \textit{ where } k \equiv 5 \pmod 8 \textit{ is prime and } l \equiv 1 \pmod 2, \\ 0 \pmod 2 & \textit{otherwise.} \end{cases}$$

Corollary 1.5. *If $8n+5 = k \cdot l^2$, where $k \equiv 5 \pmod 8$ is prime and $l \equiv 1 \pmod 2$, then*

$$p(n) \equiv 1 + \sum_{m=1}^{\lfloor \frac{1}{6}+\frac{1}{12}\sqrt{4+6n}\rfloor} p(n-8(3m^2-m)) + \sum_{m=1}^{\lfloor -\frac{1}{6}+\frac{1}{12}\sqrt{4+6n}\rfloor} p(n-8(3m^2+m)) \pmod 2.$$

If $8n+5$ *cannot be written in such a form, then*

$$p(n) \equiv \sum_{m=1}^{\lfloor \frac{1}{6}+\frac{1}{12}\sqrt{4+6n}\rfloor} p(n-8(3m^2-m)) + \sum_{m=1}^{\lfloor -\frac{1}{6}+\frac{1}{12}\sqrt{4+6n}\rfloor} p(n-8(3m^2+m)) \pmod 2.$$

Using these results, we obtain estimates for the parity of $p(n)$ which fall just short of (1.7) and (1.6).

Corollary 1.6. *For all sufficiently large positive integers* x, *we have*

$$\#\{n \le x : p(n) \equiv 1 \pmod 2\} \gg \frac{\sqrt{x}}{\log x}.$$

Corollary 1.7. *For all sufficiently large positive integers* x, *we have*

$$\#\{n \le x : p(n) \equiv 0 \pmod 2\} \gg \sqrt{x}.$$

2. Proof of Theorems 1.1 and 1.2

Proof of Theorem 1.1. We recall the definition of $\Delta(q)$ as in (1.1),

$$\Delta(q) = q\prod_{n=1}^{\infty}(1-q^n)^{24}. \tag{2.8}$$

Raising the series to the $\frac{4^s-1}{3}$ power, we find

$$\begin{aligned}\Delta(q)^{\frac{4^s-1}{3}} &= \left(q\prod_{n=1}^{\infty}(1-q^n)^{24}\right)^{\frac{4^s-1}{3}}\\ &\equiv q^{\frac{4^s-1}{3}}\prod_{n=1}^{\infty}(1-q^{8n})^{4^s-1}\\ &\equiv q^{\frac{4^s-1}{3}}\prod_{n=1}^{\infty}(1-q^{8n\cdot 4^s})\frac{1}{\prod_{n=1}^{\infty}(1-q^{8n})} \pmod 2.\end{aligned} \tag{2.9}$$

Using the fact that $P(q) = \frac{1}{\prod_{k=1}^{\infty}(1-q^k)}$, and replacing q by q^8, we have

$$\Delta(q)^{\frac{4^s-1}{3}} \equiv q^{\frac{4^s-1}{3}}\left(\sum_{n=0}^{\infty}p(n)q^{8n}\right)\left(\prod_{n=1}^{\infty}(1-q^{8n\cdot 4^s})\right) \pmod 2.$$

Using Euler's identity,

$$\prod_{k=1}^{\infty}(1-q^k) = \sum_{n=-\infty}^{\infty}(-1)^n q^{\frac{3n^2-n}{2}},$$

and replacing q by $q^{8\cdot 4^s}$, we find

$$\begin{aligned}\Delta(q)^{\frac{4^s-1}{3}} &\equiv q^{\frac{4^s-1}{3}}\Big(\sum_{k=0}^{\infty} p(n)q^{8n}\Big)\Big(\sum_{n=-\infty}^{\infty} q^{\frac{8\cdot 4^s(3n^2-n)}{2}}\Big)\\ &\equiv \Big(\sum_{n=0}^{\infty} p(n)q^{8n+\frac{4^s-1}{3}}\Big)\Big(\sum_{n=-\infty}^{\infty} q^{4^{s+1}(3n^2-n)}\Big) \pmod 2. \qquad \square\end{aligned}$$

Proof of Theorem 1.2.. By Theorem 1.1, we have

$$\begin{aligned}\Delta(q)^{\frac{4^s-1}{3}} \equiv&\Big(\sum_{n=0}^{\infty} p(n)q^{8n+\frac{4^s-1}{3}}\Big)\Big(\sum_{n=-\infty}^{\infty} q^{4^{s+1}(3n^2-n)}\Big) \pmod 2\\ \equiv&\Big(\sum_{k=0}^{\infty} p(k)q^{8k+\frac{4^s-1}{3}}\Big)\Big(1+\sum_{m=1}^{\infty} q^{4^{s+1}(3m^2+m)}+\sum_{m=1}^{\infty} q^{4^{s+1}(3m^2-m)}\Big)\\ \equiv&\sum_{k=0}^{\infty} p(k)q^{8k+\frac{4^s-1}{3}}+\sum_{m=1}^{\infty}\Big(\sum_{k=0}^{\infty} p(k)q^{8k+\frac{4^s-1}{3}+4^{s+1}(3m^2+m)}\Big)\\ &+\sum_{m=1}^{\infty}\Big(\sum_{k=0}^{\infty} p(k)q^{8k+\frac{4^s-1}{3}+4^{s+1}(3m^2-m)}\Big) \pmod 2. \end{aligned}\tag{2.10}$$

We now examine the coefficient of q^r, where r is of the form $8n+\frac{4^s-1}{3}$. The left side of (2.10) becomes $\tau_{(4^s-1)/3}(8n+\frac{4^s-1}{3})$. The right side becomes the sum of $p(k)$ for all k such that there exists an integral m such that

$$8n+\frac{4^s-1}{3}=8k+\frac{4^s-1}{3}+4^{s+1}(3m^2-m)$$

or

$$8n+\frac{4^s-1}{3}=8k+\frac{4^s-1}{3}+4^{s+1}(3m^2+m).$$

Solving for k, we obtain

$$k=n-2^{2s-1}(3m^2\pm m).$$

Because $k\geq 0$, the limits on the sums must be chosen so that $n-2^{2s-1}(3m^2\pm m)\geq 0$. Thus, we have

$$\begin{aligned}\tau_{\frac{4^s-1}{3}}(8n+\frac{4^s-1}{3})\equiv\ & p(n)+\sum_{m=1}^{\lfloor\frac{1}{6}+\frac{1}{3\cdot 2^s}\sqrt{4^{s-1}+6n}\rfloor} p(n-2^{2s-1}(3m^2-m))\\ &+\sum_{m=1}^{\lfloor-\frac{1}{6}+\frac{1}{3\cdot 2^s}\sqrt{4^{s-1}+6n}\rfloor} p(n-2^{2s-1}(3m^2+m))) \pmod 2.\end{aligned}$$

Solving for $p(n)$, we obtain a recurrence formula,

$$p(n) \equiv \tau_{\frac{4^s-1}{3}}(8n+\frac{4^s-1}{3}) + \sum_{m=1}^{\lfloor \frac{1}{6}+\frac{1}{3\cdot 2^s}\sqrt{4^{s-1}+6n}\rfloor} p(n-2^{2s-1}(3m^2-m))$$
$$+ \sum_{m=1}^{\lfloor -\frac{1}{6}+\frac{1}{3\cdot 2^s}\sqrt{4^{s-1}+6n}\rfloor} p(n-2^{2s-1}(3m^2+m)) \pmod 2. \qquad \square$$

3. Proof of Theorems 1.3 and 1.4 and Corollary 1.5

Lemma 3.1. *If n is a positive integer, then*

$$\tau(n) \equiv \begin{cases} 1 & \text{if } n=(2k+1)^2, \\ 0 & \text{otherwise.} \end{cases}$$

Proof. By the definition of $\Delta(q)$, we have

$$\begin{aligned} \Delta(q) &= q\prod_{n=1}^{\infty}(1-q^n)^{24} \\ &= q\Big(\sum_{k=0}^{\infty}(-1)^k(2k+1)q^{k(k+1)/2}\Big)^8 \\ &\equiv \sum_{k=0}^{\infty} q\big(q^{4k(k+1)}\big) \\ &\equiv \sum_{k=0}^{\infty} q^{(2k+1)^2} \pmod 2. \end{aligned} \qquad \square$$

Lemma 3.2. *For integers $s \geq 2$, we have*

$$\Delta(q)^{\frac{4^s-1}{3}} \equiv \Delta(q)\Delta(4q)\cdots\Delta(4^{s-1}q) \pmod 2.$$

Proof. We can write $\frac{4^s-1}{3}$ as $1+4+\cdots 4^{s-1}$. Substituting this into the expression $\Delta(q)^{\frac{4^s-1}{3}}$, we find

$$\begin{aligned} \Delta(q)^{\frac{4^s-1}{3}} =& \Delta(q)^{1+4+\cdots 4^{s-1}} \\ =& \Delta(q)\Delta(q)^4\cdots\Delta(q)^{4^{s-1}} \\ \equiv& \Delta(q)\Delta(4q)\cdots\Delta(4^{s-1}q) \pmod 2. \end{aligned} \qquad \square$$

Proof of Theorem 1.3. Combining Lemmas 3.1 and 3.2, we find that $\tau_{(4^s-1)/3}(n)$ (mod 2) is equivalent to the number of representations of n as

$$x_1^2+4x_2^2+\cdots 4^{s-1}x_{s-1}^2,$$

where x_i are positive odd integers. We can write this as

$$r_s(n) = \#\{x_1^2 + 4x_2^2 + \cdots 4^{s-1}x_s : x_i \text{ are positive odd integers}\}.$$

Thus, we have $\tau_{(4^s-1)/3}(n) \equiv r_s(n) \pmod 2$. □

We examine the number of representations of n as x^2+y^2 for any integers x, y in order to find a formula for the number of representations of the form k^2+4l^2 for positive, odd integers k, l.

We define $F(q)$, a power series in q whose coefficients give the number of representations of n as the sum x^2+y^2 for integers x, y. This function is generated by summing $q^{x^2+y^2}$ over all integers x and y:

$$F(q) = \sum_{x=-\infty}^{\infty} \sum_{y=-\infty}^{\infty} q^{x^2+y^2} = \sum_{n=0}^{\infty} f(n)q^n. \tag{3.11}$$

We find a factorization for the coefficients of $F(q)$.

Theorem 3.3. *Let n be a positive integer such that the factorization of n contains no odd powers of primes which are* $3 \pmod 4$. *Then $f(n)$ has the factorization*

$$f(n) = \Big(4 \cdot \prod (m_p - 1)\Big),$$

where the product is taken over all primes $p \equiv 1 \pmod 4$ which divide n and where m_p is the largest integer such that $p^{m_p} | n$. If the factorization of n contains an odd power of a prime which is $3 \pmod 4$, *then $f(n) = 0$.*

This follows from the unique factorization of n in $\mathbb{Z}[i]$ [Hardy and Wright 1979].

If we restrict our function to count only the representations of n of the form k^2+4l^2 for positive odd k, l, we can create a similar power series in q, denoted by $G(q)$, such that the coefficients of $G(q)$ give the number of these representations. We write

$$G(q) = \sum_{x=0}^{\infty}\sum_{y=0}^{\infty} q^{(2x+1)^2+4(2y+1)^2} = \sum_{n=0}^{\infty} g(n)q^n. \tag{3.12}$$

We again find a factorization for these coefficients.

Theorem 3.4. *For integers $n \equiv 5 \pmod 8$, we have*

$$g(n) = \tfrac{1}{8} f(n).$$

Proof. Because the only quadratic residues of 8 are 0, 1 and 4, and $n \equiv 5 \pmod 8$, the only representations of n as the sum of two squares are of the form $k^2 + (2l)^2$, where k, l are positive odd integers. Therefore, the theorem states that for every representation of n as k^2+4l^2 for positive, odd k, l, there are 8 representations of n as x^2+y^2 for integers x, y. For each k, l, we can choose x and y to be either positive or negative. This gives us four new representations. Additionally, although switching l and k produces a different n, switching x and y yields two different representations of n.

Combining both above methods of generating multiple representations in x and y, we find that for each representation of n as the k^2+4l^2 for k and l nonnegative odd integers, there exists 8 representations of n as x^2+y^2, for integers x, y. □

Proof of Theorem 1.4. We now investigate the parity of $\tau_5(8n+5)$.

By Theorem 1.3,

$$\tau_5(8n+5) \equiv r_2(8n+5),$$

and by Theorem 3.4,

$$r_2(8n+5) = g(8n+5) = \tfrac{1}{8}(8n+5).$$

Combining these facts with the formula for $f(n)$ from Theorem 3.3, we have

$$\tau_5(8n+5) \equiv \tfrac{1}{2}\prod(m_p+1) \pmod 2.$$

The odd values of $\tau_5(8n+5)$ are those for which the factorization of $\prod(m_p+1)$ has exactly one power of 2. This occurs when exactly one m_{p_1} is odd, in which case we can write

$$8n+5 = p_1^{m_{p1}}(p_2^{m_{p2}}\cdots p_n^{m_{pn}})(r),$$

where r is the product of even powers of primes which are $3 \pmod 4$. In the factorization of $8n+5$, there are an even number of factors of every prime except p_1, so we can write

$$8n+5 = p_1^{m_{p1}}s^2$$

where s is odd. Because $p_1^{m_{p1}}s^2 \equiv 5 \pmod 8$, and the only quadratic residues of 8 are 0, 1 and 4, $p_1 \equiv 5 \pmod 8$.

If we cannot write $8n+5$ in this form, then

$$\tfrac{1}{2}\prod(m_p+1) \equiv 0 \pmod 2,$$

so $\tau_{(4^2-1)/3}(8n+5)$ is even. □

Remark 2. We have proven the additional result that $m_{p_1} = 4m+1$ for some nonnegative integer m, so $8n+5 = p_1^{4m+1}s^2$ with $p_1 \equiv 5 \pmod 8$ prime, $m \geq 0$, s odd, and $p_1 \nmid s$. This stronger version of Theorem 1.4 first appeared as Exercise 6.7 in [Serre 1976], and also appears in [Nicolas 2006].

Proof of Corollary 1.5. By Theorem 1.2 we have

$$p(n) \equiv \tau_5(8n+5) + \sum_{m=1}^{\lfloor \frac{1}{6}+\frac{1}{12}\sqrt{4+6n}\rfloor} p(n-8(3m^2-m)) + \sum_{m=1}^{\lfloor -\frac{1}{6}+\frac{1}{12}\sqrt{4+6n}\rfloor} p(n-8(3m^2+m)) \pmod 2.$$

By Theorem 1.4, we find for n such that $8n+5 = k\cdot l^2$, where $k \equiv 5 \pmod 8$ is prime and $l \equiv 1 \pmod 2$,

$$p(n) \equiv 1 + \sum_{m=1}^{\lfloor \frac{1}{6}+\frac{1}{12}\sqrt{4+6n}\rfloor} p(n-8(3m^2-m)) + \sum_{m=1}^{\lfloor -\frac{1}{6}+\frac{1}{12}\sqrt{4+6n}\rfloor} p(n-8(3m^2+m)) \pmod 2.$$

If $8n+5$ cannot be written in this form, then

$$p(n) \equiv \sum_{m=1}^{\lfloor \frac{1}{6}+\frac{1}{12}\sqrt{4+6n}\rfloor} p(n-8(3m^2-m)) + \sum_{m=1}^{\lfloor -\frac{1}{6}+\frac{1}{12}\sqrt{4+6n}\rfloor} p(n-8(3m^2+m)) \pmod 2. \qquad \square$$

Lemma 3.5. *For all sufficiently large positive integers x,*

$$\#\{n \le x : \tau_5(n) \equiv 1 \pmod 2\} \gg \frac{x}{\log x}.$$

Proof. By Theorem 1.4, $\tau_5(n) \equiv 1 \pmod 2$ if n can be written in the form kl^2, where $k \equiv 5 \pmod 8$ is prime, and $l \equiv 1 \pmod 2$. We look at the case where $n \equiv 5 \pmod 8$ is prime, and $k = n$ and $l = 1$. For sufficiently large x, we have

$$\frac{x}{4\log x}.$$

such that $n \le x$ [Apostol 1976]. This gives us a lower bound for the number of odd values of $\tau_5(n)$ where $n \le x$. □

Proof of Corollary 1.6. We rewrite Theorem 1.2 with $s = 2$:

$$\tau_5(8n+5) \equiv p(n) + \sum_{m=1}^{\lfloor \frac{1}{6}+\frac{1}{12}\sqrt{4+6n}\rfloor} p(n-8(3m^2-m)) + \sum_{m=1}^{\lfloor -\frac{1}{6}+\frac{1}{12}\sqrt{4+6n}\rfloor} p(n-8(3m^2+m))) \pmod 2. \tag{3.13}$$

By the proof of Lemma 3.5, we have

$$\#\{n \le x : \tau_5(8n+5) \equiv 1 \pmod 2\} \gg \frac{x}{\log x}. \tag{3.14}$$

For each of these n, there exists a nonnegative integer r such that

$$p(n-8(3r^2-r)) \equiv 1 \pmod 2$$

or

$$p(n-8(3r^2+r)) \equiv 1 \pmod 2.$$

Because the number of possible r is

$$\left\lfloor \tfrac{1}{6}+\tfrac{1}{12}\sqrt{4+6x}\right\rfloor + \left\lfloor -\tfrac{1}{6}+\tfrac{1}{12}\sqrt{4+6x}\right\rfloor + 1 \sim \sqrt{x},$$

there must be at least

$$c \cdot \tfrac{x}{\log x}\left(\tfrac{1}{\sqrt{x}}\right)$$

distinct values of $n \leq x$ such that $p(n)$ is odd.

Therefore, we have

$$\#\{n \leq x : p(n) \equiv 1 \pmod 2\} \gg \frac{\sqrt{x}}{\log x}. \qquad \square$$

Proof of Corollary 1.7. We rewrite Theorem 1.2 in the case of $s = 2$:

$$\tau_5(8n+5) \equiv p(n) + \sum_{m=1}^{\lfloor \frac{1}{6}+\frac{1}{12}\sqrt{4+6n}\rfloor} p(n-8(3m^2-m)) + \sum_{m=1}^{\lfloor -\frac{1}{6}+\frac{1}{12}\sqrt{4+6n}\rfloor} p(n-8(3m^2+m)) \pmod 2. \tag{3.15}$$

We note that the number of terms on the right hand side of (3.15) is

$$1 + \left\lfloor \tfrac{1}{6} + \tfrac{1}{12}\sqrt{4+6n} \right\rfloor + \left\lfloor -\tfrac{1}{6} + \tfrac{1}{12}\sqrt{4+6n} \right\rfloor,$$

which is odd only if, for some positive integer z,

$$24z^2 + 8z \leq n < 24z^2 + 40z + 16. \tag{3.16}$$

We also note, by the remark following Theorem 1.2, that

$$\lim_{x\to\infty} \frac{\#\{n \leq x : \tau_5(n) \equiv 0 \pmod 2\}}{x} = 1. \tag{3.17}$$

When an odd number of integers add up to an even number, at least one of the integers must be even. Thus, when $\tau_5(8n+5)$ is even, and (3.16) is satisfied, one of the terms on the right side of (3.15) must be even. We now count the number of intervals such that (3.16) holds and all values in the interval are $\leq x$. This yields

$$\left\lfloor -\tfrac{1}{6} + \tfrac{1}{12}\sqrt{4+6n} \right\rfloor$$

invervals, each of which contains $32z+16$ integers. Therefore, the number of $n \leq x$ for which the right side of (3.15) has an odd number of terms is at least

$$\tfrac{2}{3}n + c_1\sqrt{4+6n} + c_2 \tag{3.18}$$

for some constants $c_1, c_2 > 0$.

Combining (3.17) and (3.18), we find that, as $x \to \infty$, the number of $n \leq x$ for which $\tau_5(8n+5)$ is even and there are an odd number of terms on the right side of (3.15) approaches

$$\tfrac{2}{3}x. \tag{3.19}$$

For each of these n, there must be an even term on the right hand side of (3.15). However, (3.19) does not give the total number distinct n for which $p(n)$ is even;

we may be counting an integer w for each n, m such that $n-8(3m^2-m)$ or $n-8(3m^2+m)=w$.

We can put an upper bound on the number of m for which we are counting w because there are only $c\sqrt{x}$ values of m for which $n-8(3m\pm m)$ is positive for some n, for some constant $c>0$. We divide (3.19) by the number of m in order to compensate for the possibility of counting any w multiple times. Thus, we have, as $x\to\infty$,

$$\#\{n\leq x : p(n)\equiv 0 \pmod 2\}\gg\sqrt{x} \qquad \square$$

Acknowledgments

The author would to thank the National Science Foundation for their generous support of this research through the University of Wisconsin VIGRE program. The author would also like to thank Ken Ono and Amanda Folsom for their guidance.

References

[Ahlgren 1999] S. Ahlgren, "Distribution of parity of the partition function in arithmetic progressions", *Indag. Math. (N.S.)* **10**:2 (1999), 173–181. MR 2002i:11102 Zbl 1027.11079

[Andrews 1971] G. E. Andrews, *Number theory*, W. B. Saunders Co., Philadelphia, PA, 1971. MR 46 #8943

[Apostol 1976] T. M. Apostol, *Introduction to analytic number theory*, Springer, New York, 1976. Undergraduate Texts in Mathematics. MR 55 #7892 Zbl 0335.10001

[Hardy and Wright 1979] G. H. Hardy and E. M. Wright, *An introduction to the theory of numbers*, Fifth ed., The Clarendon Press Oxford University Press, New York, 1979. MR 81i:10002

[Nicolas 2006] J.-L. Nicolas, "Valeurs impaires de la fonction de partition $p(n)$", *Int. J. Number Theory* **2**:4 (2006), 469–487. MR 2281859

[Nicolas 2008] J.-L. Nicolas, "Parité des valeurs de la fonction de partition p(n) et anatomie des entiers", preprint, 2008, Available at http://math.univ-lyon1.fr/~nicolas/anatomie.pdf.

[Nicolas et al. 1998] J.-L. Nicolas, I. Z. Ruzsa, and A. Sárközy, "On the parity of additive representation functions", *J. Number Theory* **73**:2 (1998), 292–317. With an appendix in French by J.-P. Serre. MR 2000a:11151 Zbl 0921.11050

[Serre 1974] J.-P. Serre, "Divisibilité des coefficients des formes modulaires de poids entier", *C. R. Acad. Sci. Paris Sér. A* **279** (1974), 679–682. MR 52 #3060 Zbl 0304.10017

[Serre 1976] J.-P. Serre, "Divisibilité de certaines fonctions arithmétiques", *Enseignement Math.* **22**:2 (1976), 227–260. With an appendix in French by J.-P. Serre.

Received: 2007-08-03 Revised: 2007-10-28 Accepted: 2007-10-28

swolfe2@wisc.edu

Department of Mathematics, University of Wisconsin,
Madison, Wisconsin 53706, United States

Qualitative behavior and computation of multiple solutions of singular nonlinear boundary value problems

Grey Ballard and John Baxley

(Communicated by Kenneth S. Berenhaut)

We consider boundary value problems of the form

$$y'' = -f(t, y), \quad y(0) = 0, \quad y(1) = 0,$$

motivated by examples where $f(t, y) = \phi(t)g(y)$ and $g(y)$ behave like $y^{-\lambda}$ ($\lambda > 0$) as $y \to 0^+$. We explore conditions under which such problems have multiple positive solutions, investigate qualitative behavior of these solutions, and discuss computational methods for approximating the solutions.

1. Introduction

The present work is a first attempt to understand singular boundary value problems with multiple solutions. As such, it seeks to combine research on singular boundary value problems having unique solutions that began with the paper of Taliaferro [1979] with work on nonsingular boundary value problems having multiple solutions that received impetus from the paper by Henderson and Thompson [2000] but dates back at least to work by Parter [1974]. The majority of later papers dealt with theoretical questions of existence, but a few, such as [Baxley 1995; Baxley and Thompson 2000; Ballard et al. 2006], have dealt with computational questions.

We shall focus here on two examples, which have the form

$$y'' = -f(t, y), \qquad 0 < t < 1, \tag{1}$$

$$y(0) = 0, \quad y(1) = 0, \tag{2}$$

where the nonlinear function $f(t, y)$ is positive and singular as $y \to 0^+$ and may also be singular as $t \to 0^+$ or $t \to 1^-$.

MSC2000: 34B16, 65L10.

Keywords: singular nonlinear boundary value problems, multiple solutions, shooting, computation, qualitative behavior.

Taliaferro [1979] considered the case

$$f(t,y)=\frac{\phi(t)}{y^{\lambda}},$$

where $\lambda > 0$ and ϕ is continuous on $(0, 1)$. He proved the existence of a unique positive solution if

$$\int_0^1 t(1-t)\phi(t)\,dt<\infty.$$

He then described the asymptotic behavior at the endpoints of this solution $y(t)$. For example, if

$$\int_0^{1/2} \phi(t)t^{-\lambda}\,dt<\infty,$$

then the slope of the solution $y(t)$ is finite at $t = 0$. If this integral is infinite and, for example, $\phi(t) \sim t^{\alpha}$, as $t \to 0^{+}$, where $-2 < \alpha \le \lambda - 1$, then the slope of the solution is infinite at $t = 0$ and Taliaferro [1979] provides the detailed asymptotic behavior. Note that for these results, the function $f(t, y)$ is decreasing in y for fixed t and tends to ∞ as $t \to 0^{+}$.

To compute the positive solution to such a problem, the papers [Baxley 1995; Baxley and Thompson 2000] took advantage of the known asymptotic behavior of the solution at the endpoints to design a shooting method. Basically, the interval [0, 1] was replaced by a slightly smaller interval [a, b] and the asymptotic knowledge was used to design an initial value problem at a and a terminal value problem at b, each depending on a parameter. These problems were solved using an initial value method such as that of Runge–Kutta–Fehlburg and parameters were adjusted by a modified Newton method until the solutions met at $t = 1/2$ with essentially the same slope and altitude.

Henderson and Thompson [2000] dealt with the problem (Equation (1), (2)) in the autonomous case $f(t, y) = f(y)$ with $f(y)$ continuous for $y \ge 0$. They gave conditions under which the problem has at least three positive solutions, and the behavior of $f(y)$ which triggered the multiple solutions was, in contrast to Taliaferro [1979], a tendency for $f(y)$ to increase. Specifically, they required that there be numbers $0 < a < b < 2b$ so that $f(y)$ is much larger on the interval [b, 2b] than on the interval [0, a].

Henderson and Thompson [2000] also provided qualitative information about the size of the three positive solutions, and this knowledge was used in [Ballard et al. 2006] to compute solutions to such nonsingular problems. Since this qualitative knowledge has a global character and gives no information about the behavior near endpoints, the problem was discretized on the interval [0, 1] and an iterative method was used to obtain rough approximations to the solutions. The values of these approximations near the endpoints were then used to estimate slopes at the endpoints and these estimates were used to seed a shooting method similar to that used earlier on the Taliaferro problems.

The last example discussed in [Ballard et al. 2006] is singular and was designed by modifying an example in [Baxley 1995] so that the singular nonlinearity $f(t, y)$ exhibited also the behavior required in [Henderson and Thompson 2000]. The solution of the original example has finite slope at both endpoints. The computational work indicates that the problem has three solutions, each having finite slopes at the endpoints.

2. Solutions with finite slopes at endpoints

We begin with a synopsis of the last example considered in [Ballard et al. 2006].

Example 2.1. For $0 < t < 1$, let

$$f(t, y) = \begin{cases} \frac{2\sqrt{t(1-t)}}{\sqrt{y}}, & 0 < y \leq 1, \\ 2\sqrt{(2-y)t(1-t)+400(y-1)}, & 1 < y < 2, \\ 40, & 2 \leq y. \end{cases}$$

According to [Taliaferro 1979] (or see the generalization in [Baxley 1991]), we would expect solutions to exist and have finite slopes at the endpoints $t = 0$ and $t = 1$ since $f(t, \theta t)$ is integrable in a neighborhood of $t = 0$ and $f(t, \theta(1-t))$ is integrable in a neighborhood of $t = 1$, for each constant $\theta > 0$. Further, one easily verifies that $f(t, y)$ satisfies the Henderson–Thompson type estimates

$$\begin{aligned} f(t, y) &< 8a, & \alpha &\leq y \leq \alpha + a, \\ f(t, y) &> 16b, & b &\leq y \leq 2b, \\ f(t, y) &< 8c, & \alpha &\leq y \leq \alpha + c, \end{aligned}$$

where $\alpha = 1/32$, $a = 1$, $b = 2$, $c = 6$, so one might hope that Equations (1) and (2) will have three positive solutions. Note that the theory in [Taliaferro 1979] and [Baxley and Thompson 2000; Henderson and Thompson 2000] cannot actually be applied to this example, but work in progress will extend the results of Taliaferro [1979] and Henderson and Thompson [2000] to such problems.

Our method, used in [Ballard et al. 2006], is basically a two-step procedure. Step 2 is a shooting method and for each solution $y(t)$, we need approximate values of $y'(0)$ and $y'(1)$ to seed the method. Then we can choose a slightly smaller subinterval [a, b] of [0, 1] and use the asymptotic formulas of Taliaferro [1979] to estimate the values of $y(a)$, $y'(a)$ and $y(b)$, $y'(b)$. Employing any dependable initial value solver, such as RKF45, we can then solve the resulting initial value problem on [0, 1/2] and the terminal value problem on [1/2, 1]. The initial approximations of $y'(0)$, $y'(1)$ can then be adjusted by a modified Newton method until these two solutions meet at $t = 1/2$ with essentially the same altitude and slope. Details of such a shooting method appear in [Baxley 1995; Baxley and Thompson 2000; Ballard et al. 2006].

Thus step 1 of our method is designed to produce reasonably good approximations for $y'(0)$ and $y'(1)$ for each of the three solutions. For this purpose, we discretize the problem by dividing the interval [0, 1] into $n+1$ equal parts at the mesh points $t_i = i/(n+1)$ and seek to approximate $y(t_i)$, for $i = 1, 2, \dots, n$. We approximate the second derivative as usual with the central divided difference quotient

$$y''(t_k) \approx \frac{1}{h^2}\Big(y(t_{k+1}) - 2y(t_k) + y(t_{k-1})\Big),$$

where $h = 1/(n+1)$. Letting y_i be our approximation for $y(t_i)$ and Y be the n-dimensional column vector with components y_i, our discrete problem is

$$\frac{1}{h^2}AY = F(T, Y), \tag{3}$$

where T is the n-vector with components t_i, $F(T, Y)$ is the n-vector with components $f(t_i, y_i)$, and A is the matrix

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

Note that the boundary conditions $y(0) = 0$, $y(1) = 0$ have been used to obtain this formulation. We rewrite Equation (3) in the fixed point form

$$h^2 A^{-1} F(T, Y) = Y, \tag{4}$$

and then use iteration to obtain three solutions Y_1, Y_2, Y_3 of this problem that are viewed as crude approximations for y_1, y_2, y_3 at the mesh points.

Based on qualitative estimates in [Henderson and Thompson 2000], we expect Y_1 (the "small" solution) to have a maximum less than 1, Y_2, to have a maximum greater than 1, but a value at 1/4 less than 2,and Y_3 (the "large" solution) to have a value at 1/4 greater than 2. So we seed the iteration with initial vectors which satisfy these requirements. It turns out that Y_1 and Y_3 are attractors for this discrete problem, but Y_2 is a repeller. So, we "back" into an approximation for Y_2 by using averages of approximations for Y_1 and Y_3; details can be found in [Ballard et al. 2006], where one can find numerical results of the full computation. The same method will be used below.

3. Solutions with infinite slopes at endpoints

Example 3.1. For $0 < t < 1$, we now let

$$f(t, y) = \begin{cases} \frac{3(1-t)^2+4t(1-t)+4t^2}{16t^{3/2}(1-t)y}, & 0 < y \le 1, \\ (2-y)\frac{3(1-t)^2+4t(1-t)+4t^2}{16t^{3/2}(1-t)} + 40(y-1), & 1 < y < 2, \\ 40, & 2 \le y. \end{cases}$$

The asymptotic formulas in [Taliaferro 1979] (see also [Baxley and Thompson 2000, Theorem 10], [Baxley and Martin 2000, Lemma 12]) suggest that solutions to Equation (1) and (2) should now have infinite slope at both endpoints. Also the behavior of $f(t, y)$ resembles that of the first example, so it seems likely that there will be three solutions.

If $y(t)$ is any solution of Equation (1) and (2), then $y(t)$ is near zero in a neighborhood of the endpoints. Thus to examine asymptotic behavior near the endpoints, we let

$$\phi(t) = \frac{3(1-t)^2+4t(1-t)+4t^2}{16t^{3/2}(1-t)},$$

and we see that

$$\phi(t) \sim \frac{3}{16}t^{-3/2}, \quad \text{as} \quad t \to 0^+; \qquad \phi(t) \sim \frac{1}{4}(1-t)^{-1}, \quad \text{as} \quad t \to 1^-.$$

Thus the asymptotic formulas in Taliaferro [1979], Baxley [1995], and Baxley and Thompson [2000] indicate that any solution $y(t)$ of Equation (1) and (2) will exhibit the asymptotic behavior

$$y(t) \sim Qt^{(\alpha+2)/(\lambda+1)} = Qt^{1/4}, \qquad \text{as} \quad t \to 0^+, \tag{5}$$

where $\alpha = -1.5$, $\lambda = 1.0$, and

$$Q = \left(\frac{3(\lambda+1)^2}{16(\alpha+2)(\lambda-\alpha-1)}\right)^{1/(\lambda+1)} = 1.$$

A similar analysis leads to

$$y(t) \sim (1-t)^{1/2}, \quad \text{as} \quad t \to 1^-. \tag{6}$$

To compute approximations for these three solutions, the overall strategy is the same as before. We wish to use shooting, taking advantage of the asymptotic formulas (5) and (6) as we did in [Baxley 1995; Baxley and Thompson 2000], but as before we need a first step to find crude approximations for the three solutions.

In our first effort, we used the same iteration scheme as in Example 2.1, but found that it gave poor accuracy. After some confusion, we discovered that the problem lay in the matrix A, and the correction comes from a careful analysis of our approximation for $y''(t_k)$.

Approximating $y'(t_k)$ with a backward difference quotient

$$y'(t_k) \approx \frac{y(t_k) - y(t_{k-1})}{h},$$

where $h = t_k - t_{k-1}$, we then approximate $y''(t_k)$ with a forward difference quotient

$$y''(t_k) = \frac{y'(t_{k+1}) - y'(t_k)}{h}.$$

We combine these to get the usual second order divided difference quotient. But if we focus on an endpoint, say t_1, we are led, in the approximation for $y''(t_1)$, to replace $y'(t_1)$ with $y(t_1)/h$. This, it turns out, is a blunder. To see why, we apply Equation (5) to conclude

$$y'(t) \sim \frac{1}{4} t^{-3/4}, \quad \text{as} \quad t \to 0^+, \quad \text{and} \quad \frac{y(t)}{t} \sim t^{-3/4}, \quad \text{as} \quad t \to 0^+.$$

Thus

$$y'(t) \sim \frac{1}{4} \frac{y(t)}{t}, \quad \text{as} \quad t \to 0^+.$$

Therefore, a better approximation for $y'(t_1)$ is $\frac{1}{4}\frac{y(t_1)}{h}$, which leads to the approximation

$$y''(t_1) \approx \frac{-\frac{5}{4} y(t_1) + y(t_2)}{h^2}.$$

A similar analysis shows that a better approximation for $y''(t_n)$ is

$$y''(t_n) \approx \frac{y(t_{n-1}) - \frac{3}{2} y(t_n)}{h^2}.$$

So we modify the earlier matrix A to obtain instead

$$A = \begin{bmatrix} -\frac{5}{4} & 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -2 & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -\frac{3}{2} \end{bmatrix}.$$

4. Numerical calculations

We now discuss numerical calculations for Example 3.1, beginning with the approximation of the larger solution y_3. For the results reported here, we began by dividing our interval [0, 1] into 8 equal parts, seeking approximations of $y_3(t_k)$,

	$y_3(1/32)$	$y_3(1/4)$	$y_3(1/2)$	$y_3(31/32)$
unmodified A	0.5853	3.6660	4.8982	0.5347
modified A	2.2798	5.1757	6.1415	1.2206

Table 1. Results of discrete iteration for large solution.

where $t_k = k/8$, $k = 1, 2, \ldots, 7$. We seeded the iteration with the vector $Y = (1.8, 2.4, 2.7, 3.0, 2.7, 2.4, 1.8)$. Of course, we do not expect the solution to be symmetric about $t = 0.5$, but otherwise this approximation has roughly the right shape, with the value at $t = 1/4$ greater than 2, and the value at $t = 1/2$ less than 6. We iterated the fixed point form Equation (4) (with the 7×7 matrix A) 7 times, then extended this approximation, dividing the interval into 16 equal parts, by linear interpolation (we approximated $y(1/16) = 0.6y(1/8)$ and $y(15/16) = 0.6y(7/8)$). We then iterated the fixed point form (with the 15×15 matrix A) 7 times. Finally, we doubled the number of subintervals again by the same procedure and iterated 7 times with the 31×31 matrix A. The final iterate is then our approximation for Y_3, which in turn approximates y_3. The first component of Y_3 is then our approximation for $y_3(1/32)$, the 15th component is our approximation for $y_3(1/2)$ and the last component is our approximation for $y_3(31/32)$. In Table 1, we report the numerical results, not only using the modified matrix A above but also, for comparison, of the unmodified matrix A used for Example 2.1. Note how significant is the effect of the modification of A. All calculations for this iterative procedure were done using MATLAB. The large solution Y_3 is an asymptotically stable attractor and we obtained the same result with a variety of initial seeds.

The value of our approximation for $y_3(1/2)$ using the modified A is disconcerting, since it exceeds 6.0. We would expect from the qualitative Henderson–Thompson type estimates that the maximum value of our solution would be less than 6.0. This is actually the case, but our initial approximation is too rough to confirm this expectation.

Using the asymptotic estimate $y(t) \sim t^{1/4}$, we expect that $(32)^{1/4} y(1/32) = Q \approx 1.0$. Computing this value from Table 1, we get the value $Q = 5.422$ (resp. $Q = 1.392$) for the modified (resp. unmodified) matrix A. The effect of the modification is clearly large. The asymptotic estimate $y(t) \sim (1-t)^{1/2}$ and the corresponding expectation $(32)^{1/2} y(31/32) = P \approx 1.0$ leads to the value $P = 6.905$ (resp. $P = 3.025$) for the modified (resp. unmodified) matrix A. The import of the difference is fully realized only in passing to step 2 of our procedure and solving Equation (1) and (2) by shooting.

The shooting procedure, using the computed values of Q and P (with the modified A) to seed the shooting, and essentially Newton's method to adjust values of

interval	Q	$y_3(1/4)$	$y_3(1/2)$	P
[.03, .97]	5.118	5.067	6.055	6.650
[.02, .98]	3.916	4.647	5.698	5.265
[.01, .99]	2.491	4.209	5.334	3.661
[.002, .998]	1.298	3.918	5.091	1.890

Table 2. Shooting results for large solution.

Q and P so that the solutions to the appropriate initial and terminal value problems meet at $t = 0.5$ with altitudes and slopes agreeing to two decimal places, led to the results reported in Table 2. Since $1/32 \approx 0.03$, we first solved the boundary value problem, replacing the interval [0, 1] with [.03, .97]. Note that the values of $Q = 5.118$ and $P = 6.650$ reported in Table 2 are quite close to the seed values predicted by the modified A, but quite far from the seed values predicted by the unmodified A. (In fact, using the seed values from the unmodified A caused our computer program to terminate before completion.) We then enlarged this interval in steps of 0.004 by subtracting 0.002 from the left endpoint and adding 0.002 to the right endpoint and using the final values of Q and P from the previous step as seeds for Q and P on the current step. We report only a few of the results in Table 2, where it is seen that these values of Q and P are indeed moving (slowly) toward 1.0 and the value of the solution at $t = 0.5$ is falling significantly below $y = 6.0$ as expected from the Henderson–Thompson estimates. We also report the value of the solution at $t = 0.25$, where the Henderson–Thompson estimates would expect $y > 2.0$.

All computations involving shooting were done using the FORTRAN subroutine RKF45 [Forsythe et al. 1977] of Shampine and Watts. For these calculations, we only asked RKF45 for three decimal place accuracy and consequently that the altitude and slope of the solution agree to two decimal places at $t = 0.5$. Thus, the results of the shooting procedure should only be trusted to two decimal places. Of course, the difference in the computed solutions and the true solution depends also on replacing the interval [1, 0] by a smaller interval and using the asymptotic formulas to generate initial and terminal conditions. The computation in [Baxley 1991; Baxley and Thompson 2000] suggests that good approximation demands using an interval as large as [.001, .999].

We now discuss the smaller solution y_1. (This solution $y_1(t) = t^{1/4}(1-t)^{1/2}$ is known in closed form [Baxley 1995]; it can be quickly verified by direct substitution.) The iterative procedure using Equation (4) now has interesting features. The iteration exhibits characteristics of a two cycle, but also the two cycle to which the sequence of iterates converges appears to depend on the initial seed, a characteristic

$y_1(1/32)$	$y_1(1/4)$	$y_1(1/2)$	$y_1(31/32)$
0.3903	0.5168	0.4913	0.1623

Table 3. Results of discrete iteration for small solution.

interval	Q	$y_1(1/4)$	$y_1(1/2)$	P
[.03, .97]	.951	.6045	.5898	.981
[.02, .98]	.967	.6079	.5918	.987
[.01, .99]	.983	.6105	.5933	.994
[.002, .998]	.997	.6121	.5942	.999

Table 4. Shooting results for small solution.

of chaotic behavior. This iteration is taking place in a space of dimension 7, then 15, and finally 31. We tried a variety of initial seeds for the iteration and in every case the smaller of the two cycle was a reasonable approximation for the solution y_1. Some initial seeds provided very good approximations. These approximations were reasonable in the sense that the value of y at $1/32$, $31/32$ provided values of Q, P close enough to give convergence of the Newton iterates in the shooting method. In Table 3, we provide results of this iteration, only for the modified matrix A. We seeded the discrete iteration with the vector $Y = (.3, .4, .45, .5, .45, .4, .3)$ and show the results for the smaller member of the resulting two cycle.

$y_2(1/32)$	$y_2(1/4)$	$y_2(1/2)$	$y_2(31/32)$
0.6006	1.1068	1.2415	0.2438

Table 5. Results of discrete iteration for middle solution.

interval	Q	$y_2(1/4)$	$y_2(1/2)$	P
[.03, .97]	1.270	1.122	1.306	1.312
[.02, .98]	1.195	1.116	1.322	1.249
[.01, .99]	1.107	1.110	1.338	1.161
[.002, .998]	1.025	1.107	1.348	1.053

Table 6. Shooting results for middle solution.

Using the values in Table 3 for $y_1(1/32)$ and $y_1(31/32)$, we compute the seed values $Q = 0.93$ and $P = 0.92$ for the shooting method. We report the results of shooting in Table 4. Note that the values of Q and P are now quite close to 1. Also note that the closed form solution gives $y_1(1/4) = 0.6124$ and $y_1(1/2) = 0.5946$, so that our final approximation for y_1 is actually correct to three decimal places.

Finally, we pass to the middle solution y_2. As indicated earlier, the approximation Y_2 is a repeller for the discrete iteration. We proceed as we did in [Ballard et al. 2006]. We begin with Z_1 and Z_3, the 31 dimensional vectors which emerge from the discrete iteration as approximations for the solutions Y_1 and Y_3. We average these two vectors to get a vector Z_2. We then iterate one time to see if this seed vector is moving toward Z_1 or Z_3. If it is moving toward Z_1, we replace Z_1 by Z_2; otherwise we replace Z_3 by Z_2. We repeat this process until Z_1 and Z_3 differ by less than 0.001 in the 16th component. At this point, Z_1 and Z_3 are viewed as both close to the repeller Y_2, but on opposite sides. However, they are formed from averages of the original Z_1 and Z_3 and as such do not have the appropriate shape for Y_2. Thus we iterate once beginning with the final Z_1 and once beginning with the final Z_3. This iteration reshapes Z_1 and Z_3 without serious movement. We then take the average of these reshaped versions as our approximation for Y_2. In Table 5, we provide the result of this computation.

The values of $y_2(1/32)$ and $y_2(31/32)$ in Table 5 give us the seed values $Q = 1.43$ and $P = 1.38$ for shooting. These results are given in Table 6. Note again that the values of Q and P are moving towards 1.

References

[Ballard et al. 2006] G. Ballard, J. Baxley, and N. Libbus, “Qualitative behavior and computation of nonlinear boundary value problems”, *Comm. Pure Appl. Analysis* **5** (2006), 251–259.

[Baxley 1991] J. V. Baxley, “Some singular nonlinear boundary value problems”, *SIAM J. Math. Anal.* **22**:2 (1991), 463–479. MR 92f:34017 Zbl 0719.34038

[Baxley 1995] J. V. Baxley, “Numerical solution of singular nonlinear boundary value problems”, pp. 15–24 in *Proceedings of the Third International Colloquium on Numerical Analysis (Plovdiv, 1994)*, VSP, Utrecht, 1995. MR 1455945 Zbl 0843.65055

[Baxley and Martin 2000] J. V. Baxley and J. C. Martin, “Positive solutions of singular nonlinear boundary value problems”, *J. Comput. Appl. Math.* **113**:1-2 (2000), 381–399. Fixed point theory with applications in nonlinear analysis. MR 2001k:34045

[Baxley and Thompson 2000] J. V. Baxley and H. B. Thompson, “Boundary behavior and computation of solutions of singular nonlinear boundary value problems”, *Commun. Appl. Anal.* **4**:2 (2000), 207–226. MR 2001a:34027

[Forsythe et al. 1977] G. E. Forsythe, M. A. Malcolm, and C. B. Moler, *Computer methods for mathematical computations*, Prentice-Hall Inc., Englewood Cliffs, N.J., 1977. Prentice-Hall Series in Automatic Computation. MR 56 #16983

[Henderson and Thompson 2000] J. Henderson and H. B. Thompson, “Multiple symmetric positive solutions for a second order boundary value problem”, *Proc. Amer. Math. Soc.* **128**:8 (2000), 2373–2379. MR 2000k:34042 Zbl 0949.34016

[Parter 1974] S. V. Parter, "Solutions of a differential equation arising in chemical reactor processes", *SIAM J. Appl. Math.* **26** (1974), 687–716. MR 52 #3654 Zbl 0285.34013

[Taliaferro 1979] S. D. Taliaferro, "A nonlinear singular boundary value problem", *Nonlinear Anal.* **3**:6 (1979), 897–904. MR 81i:34011 Zbl 0421.34021

Received: 2007-10-28 Accepted: 2007-10-28

ballgm2@wfu.edu
Department of Mathematics, Wake Forest University,
Winston-Salem, NC 27106, United States

baxley@wfu.edu
Department of Mathematics, Wake Forest University,
Winston-Salem, NC 27106, United States

Maximal subgroups of the semigroup of partial symmetries of a regular polygon

Thomas L. Shelly and Janet E. Mills

(Communicated by Scott Chapman)

The semigroup of partial symmetries of a polygon P is the collection of all distance-preserving bijections between subpolygons of P, with composition as the operation. Around every idempotent of the semigroup there is a maximal subgroup that is the group of symmetries of a subpolygon of P. In this paper we construct all of the maximal subgroups that can occur for any regular polygon P, and determine for which P there exist nontrivial cyclic maximal subgroups, and for which there are only dihedral maximal subgroups.

1. Introduction and basic properties

The semigroup of partial symmetries of a polygon is a natural generalization of the group of symmetries of a polygon. In this paper we will assume knowledge of an undergraduate abstract algebra course and will generally use the terminology of Gallian [2002]. The group of symmetries of a polygon P is the set of distance-preserving mappings of P onto P, with composition as the operation. In particular, for any $n > 2$, the group of symmetries of a regular n-gon is a group, called the dihedral group, with $2n$ elements, and is denoted by D_n. The elements of this group are completely determined by the movement of the vertices, and D_n can be considered as a subgroup of the group of all permutations of the vertices of the polygon, under composition.

To generalize the notion of symmetries of a polygon, we first need to describe what a subpolygon should be. Let P be a convex polygon with set of vertices $V(P) = \{v_1, v_2, \ldots, v_n\}$, listed clockwise, where there exists an edge between v_i and v_{i+1} for $i = 1, 2, \ldots, n-1$, and an edge between v_n and v_1. Let

$$A = \left\{v_{i_1}, v_{i_2}, \ldots, v_{i_m}\right\}$$

be a subset of $V(P)$, where $i_1 < i_2 < \ldots < i_m$ and let P_A be the polygon with edges between v_{i_j} and $v_{i_{j+1}}$ for $j = 1, 2, \ldots, m-1$, and between v_{i_m} and v_{i_1}. The polygon

MSC2000: 20M18.

Keywords: semigroup, polygon, symmetries.

P_A is called a *subpolygon* of P, and in particular, P_A is said to be the subpolygon formed by A. Note that if v_{i_j} and $v_{i_{j+1}}$ are adjacent in P then the edge between them is still an edge in P_A; otherwise, the edge between them is new. Note also that the subpolygons include those with no vertices (the empty polygon), exactly one vertex (a point), or two vertices (a line segment).

For each subset of $V(P)$, there is a unique subpolygon described since the indices on the vertices must be increasing and each subpolygon is a convex polygon. The set of all subpolygons of P will be denoted by Π. Now we must describe the semigroup of partial symmetries of a convex polygon P. This class of semigroups was first defined in [Mills 1990b], and some of its properties explored in [Mills 1990a; 1993]. The domain and range of a function α will be denoted by $\operatorname{dom}\alpha$ and $\operatorname{ran}\alpha$ respectively.

Definition. Let P be a convex polygon. On the set

$$S = S(P) = \{\alpha : A \to B \mid P_A, P_B \in \Pi, \text{ and } \alpha \text{ is a distance-preserving bijection}\},$$

define composition by $x\alpha\beta = (x\alpha)\beta$ for all $x \in \operatorname{dom}\alpha$ such that $x\alpha \in \operatorname{dom}\beta$. Then under this operation, S is a semigroup, called the *semigroup of partial symmetries* of the polygon P.

Note that if α is in S and maps A onto B, then α^{-1} is also a distance-preserving bijection of B onto A, so α^{-1} is in S. The semigroup S is an example of an inverse semigroup. That is, for each $\alpha \in S$, there is a unique $\beta \in S$ such that $\alpha\beta\alpha = \alpha$ and $\beta\alpha\beta = \beta$. In our semigroup, for $\alpha \in S$, the mapping α^{-1} serves as the needed β.

An *idempotent* of S is any α such that $\alpha^2 = \alpha$. It is easy to see that because the mappings are one-to-one, α is an idempotent if and only if α is the identity on its domain A, denoted by ι_A. In any inverse semigroup the idempotents form a skeleton of the semigroup, and around every idempotent there is a maximal subgroup with that idempotent as its identity. In S, if A is a subset of $V(P)$, then the maximal subgroup with ι_A as its identity is

$$\begin{aligned} H_A &= \{\alpha \in S \mid \text{ there exists a } \beta \in S \text{ such that } \alpha\beta = \beta\alpha = \iota_A\} \\ &= \{\alpha \in S \mid \operatorname{dom}\alpha = \operatorname{ran}\alpha = A\}. \end{aligned}$$

This is the largest subgroup of S having ι_A as its identity.

It is the purpose of this paper to determine, for a regular polygon P, exactly which groups can occur as a maximal subgroup of $S(P)$. It is clear from the description above that the maximal subgroup around ι_A is the group of symmetries of the subpolygon P_A. Therefore, the effort to find all maximal subgroups of S reduces to describing the group of symmetries of each subpolygon of P. As is well known, the group of symmetries of any polygon is either a dihedral group or a cyclic group [Gallian 2002, Theorem 27.1]. The problem here is that we have a particular regular polygon P, and need to determine exactly which dihedral and cyclic groups can occur as symmetry groups of subpolygons of that polygon P.

From now on, we will assume that P is a regular polygon with n vertices. The semigroup $S(P)$ has an identity ι_P, and the maximal subgroup around ι_P is just the group of symmetries of P, or D_n. As we shall see, this subgroup plays an important part in determining the other maximal subgroups. Therefore, we need to recall some information about the group D_n. In particular, D_n is a group with $2n$ elements, having n reflections and n rotations. All reflections are about some line of symmetry of P that passes through the center of P. If n is even, every line of symmetry passes through two vertices or through the midpoint between two vertices, whereas if n is odd, every line of symmetry passes through exactly one vertex. Since P is regular, the rotations form a subgroup generated by ρ, which is a rotation about the center of P through $2\pi/n$ radians. This subgroup is often written as $\langle\rho\rangle$, the cyclic subgroup generated by ρ, which is isomorphic to $\mathbb{Z}_n$, the group of integers modulo n. In this paper, ρ will always denote this rotation.

2. Maximal subgroups

In this section we find all maximal subgroups of S for any regular polygon P. In addition, we provide a description of all subpolygons with rotational symmetry and we give a method for constructing subpolygons with cyclic symmetry groups. For the remainder of the paper we use the following notation: Greek letters are used to represent elements of S, and the letters v and w are used to represent vertices. Thus any expression of the form $\alpha\beta$ denotes composition, whereas the expression $v\alpha = w$ says that the vertex v is mapped to w under α (we always write the argument of the function to the left of the function, as in the definition of S in Section 1). The letter d is always used to represent an arbitrary element of D_n.

It was shown in [Mills 1993] that for a regular polygon P, every element $\alpha \in S$ can be extended to an element in the group of symmetries of P. That is, if $\operatorname{dom}\alpha = A$ then $\alpha = \iota_A d$ for some $d \in D_n$. Further, it was shown that if A has at least 3 elements, then d is unique. For the remainder of the paper, we always take any subset A of $V(P)$ to have more than two elements to ensure every element in the maximal subgroup H_A extends uniquely to D_n. Not much is lost by this restriction, since if $|A| \le 2$ then P_A is either a point or a line segment, and H_A is either $\mathbb{Z}_1$ or $\mathbb{Z}_2$. This unique extension guarantees that rotations in H_A are extended to rotations in D_n and reflections in H_A are extended to reflections in D_n. More specifically, we can connect elements in H_A to those in D_n as follows.

Lemma 2.1. *Let α and β be elements of a maximal subgroup H_A, with $|A| > 2$, such that $\alpha = \iota_A d_1$ and $\beta = \iota_A d_2$ for $d_1, d_2 \in D_n$. Then the following are true*:

(a) $\alpha\beta = \iota_A d_1 d_2$.

(b) $\alpha^j = \iota_A d_1^j$ *for all* $j \in \mathbb{Z}$.

(c) $|\alpha| = |d_1|$, *where* $|\alpha|$ *and* $|d_1|$ *are the orders of* α *and* d_1 *in* H_A *and* D_n *respectively.*

Proof. To prove Lemma 2.1a, let α and β be defined as above. Then since ι_A is the identity in H_A, $\alpha\beta = (\iota_A d_1)(\iota_A d_2) = ((\iota_A d_1)\iota_A) d_2 = (\iota_A d_1) d_2 = \iota_A d_1 d_2$. The proof of Lemma 2.1b is a simple application of Lemma 2.1a using induction and the fact that $\alpha^{-1} = \iota_A d_1^{-1}$. To prove Lemma 2.1c, suppose that $|\alpha| = m$. Then m is the smallest positive integer such that $\alpha^m = \iota_A$. From Lemma 2.1b, $\alpha^m = \iota_A d_1^m$, so $\iota_A = \iota_A d_1^m$. We have assumed that $|A| > 2$, so ι_A can be extended to a unique element of D_n. Since $\iota_A \iota_P = \iota_A$, the uniqueness of extension gives $d_1^m = \iota_P$. If $d_1^\ell = \iota_P$ for some $\ell < m$, then $\alpha^\ell = \iota_A d_1^\ell = \iota_A \iota_P = \iota_A$, which contradicts the minimality of m. Thus m is the smallest positive integer such that $d_1^m = \iota_P$. Therefore $|d_1| = m$. □

It should be noted that in general, for $d \in D_n$, $\iota_A d$ is not necessarily an element of H_A. For any $d \in D_n$, let $d|_A$ denote the function d with the domain of d restricted to A, and let $d|_A(A)$ denote the image of A under this mapping. Then $\iota_A d \in S$ is an element of H_A if and only if $d|_A(A) = A$. In this light, we can express H_A as

$$H_A = \{\iota_A d \mid d \in D_n \text{ and } d|_A(A) = A\}. \tag{1}$$

There is a subtlety in this notation that is worth mentioning. Equation (1) for H_A is guaranteed to be valid if $|A| > 2$, but may fail otherwise. For example, suppose that $|A| = 1$. Then P_A is a point, so clearly H_A contains only the identity ι_A. But the set $\{\iota_A d \mid d \in D_n \text{ and } d|_A(A) = A\}$ contains two elements, the identity in D_n and the reflection of P about the line through the vertex in A. We use the useful notation of Equation (1) freely, since we have assumed that $|A| > 2$.

It is evident from Lemma 2.1a that composition within maximal subgroups is essentially the same as composition in D_n. As groups then, it is not surprising that many properties of the maximal subgroups of S are consequences of properties of D_n (with Lemma 2.1c as just one example). Another important property of D_n that is reflected in maximal subgroups of S is the structure of cyclic subgroups. Such subgroups are important to this discussion since both dihedral and cyclic groups contain them. As mentioned in Section 1, the subgroup of all rotations in D_n is the cyclic group of order n generated by a rotation, ρ, of $2\pi/n$ radians. As a result, the subgroup of all rotations in any maximal subgroup of S is also a cyclic group generated by a rotation.

Lemma 2.2. *Let* H_A *be a maximal subgroup with a nontrivial rotation. Then the subgroup of all rotations in* H_A *is a cyclic group generated by some rotation* $\alpha \in H_A$ *such that* $\alpha = \iota_A \rho^k$, *where* k *divides* n. *In particular, the subgroup of all rotations in* H_A *is isomorphic to* $\mathbb{Z}_{n/k}$, *where* k *is the smallest positive integer such that* $\iota_A \rho^k \in H_A$.

Proof. Assume H_A has a nontrivial rotation γ. Then $\gamma = \iota_A \rho^j$ for some j. Since the set $\{m \mid \iota_A \rho^m \in H_A, m \geq 1\}$ is thus nonempty, by the well-ordering principle it has a least element k. Hence there exists $\alpha \in H_A$ such that $\alpha = \iota_A \rho^k$. By the division algorithm, there exist unique $q, r \in \mathbb{N}$ such that $n = kq + r$ with $0 \leq r < k$. So

$$\iota_A \rho^r = \iota_A \rho^{n-kq} = \iota_A \rho^n \rho^{-kq} = \iota_A \iota_P \rho^{-kq} = \iota_A \rho^{-kq} = \left(\iota_A \rho^k\right)^{-q} = \alpha^{-q},$$

and $\alpha^{-q} \in H_A$ by closure. Thus $r = 0$ by minimality of k. Therefore k divides n.

It remains to be shown that the subgroup of all rotations in H_A is exactly $\langle \alpha \rangle$. To this end, let β be a nontrivial rotation in H_A. Then $\beta = \iota_A \rho^m$ for some m. Applying the division algorithm again, there exist unique $s, t \in \mathbb{N}$ such that $m = ks + t$, with $0 \leq t < k$. Then

$$\iota_A \rho^t = \iota_A \rho^{m-ks} = \left(\iota_A \rho^m\right) \rho^{-ks} = \beta \rho^{-ks} = \beta \iota_A \left(\rho^k\right)^{-s} = \beta \left(\iota_A \rho^k\right)^{-s} = \beta \alpha^{-s} \in H_A.$$

So t must be zero, by minimality of k. Thus $\beta = \iota_A \rho^{ks} = \left(\iota_A \rho^k\right)^s = \alpha^s \in \langle \alpha \rangle$.

From Lemma 2.1c, $|\langle \alpha \rangle| = \left|\iota_A \rho^k\right| = \left|\rho^k\right|$. And $\left|\rho^k\right| = n/k$ since k divides n. So $\langle \alpha \rangle \approx \mathbb{Z}_{n/k}$. □

Though Lemma 2.2 is useful for describing the cyclic subgroups of maximal subgroups as groups, it says nothing about what the subpolygons look like that have such subgroups. To aid in the description of these subpolygons, a distance function is defined on $V(P)$ that exploits the regularity of P and is independent of the actual size of the regular polygon.

Definition. The *polygonal distance* between two vertices v and w is the fewest number of edges between v and w in the regular polygon P. The polygonal distance is denoted $P(v, w)$.

The polygonal distance is equivalent to the usual Euclidean distance in the sense that for any vertices v_1, v_2, w_1, w_2, $P(v_1, w_1) = P(v_2, w_2)$ if and only if $E(v_1, w_1) = E(v_2, w_2)$, where $E(v, w)$ denotes the Euclidean distance between v and w. This follows immediately from the fact that P is a regular polygon. In particular, since the elements of D_n are isometries, we know $E(v, w) = E(vd, wd)$ for all $d \in D_n$. This implies

$$P(v, w) = P(vd, wd), \quad \text{for all } d \in D_n. \tag{2}$$

Further, since P is a regular polygon, a subpolygon P_A is regular if and only if there exists an $\ell \in \mathbb{N}$ such that for all $v, w \in A$ that are connected by an edge in P_A, the polygonal distance $P(v, w) = \ell$. So the vertices of any subpolygon which is itself a regular polygon must be evenly spaced around P with respect to the polygonal distance. The vertex sets of regular subpolygons are fundamental to describing the maximal subgroups of S, so we give these sets of vertices their own notation.

Definition. Let k divide n such that $k \neq n/2$. The k class of the vertex v_i, denoted by $[v_i]_k$, is defined as $[v_i]_k = \{v_j \mid P(v_i, v_j) = mk \text{ for some } m \in \mathbb{N}\}$.

From the discussion above, these k classes are precisely the vertex sets of regular subpolygons. Since $P(v_i, v_i\rho^{\ell k})$ is a multiple of k for all $\ell \in \mathbb{N}$, it is apparent that $v_j \in [v_i]_k$ if and only if $v_i\rho^{mk} = v_j$ for some $m \in \mathbb{N}$. So $[v_i]_k = \{v_i\rho^{mk} \mid m \in \mathbb{N}\}$. The set $\{v_i\rho^{mk} \mid m \in \mathbb{N}\}$ is the set of all vertices that v_i is mapped to under elements of $\langle\rho^k\rangle$, called the *orbit* of v_i under $\langle\rho^k\rangle$. No two distinct elements of $\langle\rho^k\rangle$ map v_i to the same vertex, so $|[v_i]_k| = |\langle\rho^k\rangle| = n/k$. So, each k class forms a regular subpolygon with n/k vertices (we have disallowed $k = n/2$ to ensure each k class has more than two vertices). Moreover, it is well known that the set of all orbits of any set under some group is a partition of that set, so $P_{[v]_k}$ is the unique regular subpolygon with n/k vertices containing v. Since the symmetry group of a regular polygon is dihedral, we have proven the following result:

Proposition 2.1. *Let P be a regular polygon with n sides. Then S has a maximal subgroup isomorphic to the dihedral group $D_{n/k}$ for every k which divides n.*

Dihedral maximal subgroups are thus abundant in the sense that as long as n is not prime (and P is not a square), S contains at least one nontrivial dihedral maximal subgroup (the trivial case being the group of symmetries of P). In contrast, the restriction that n be not prime is not sufficient to show that S contains a nontrivial cyclic maximal subgroup. As we will show, stronger restrictions must be placed on the divisors of n to guarantee nontrivial cyclic maximal subgroups of S exist.

We will use k classes to describe all subpolygons that have rotational symmetry.

Lemma 2.3. *A maximal subgroup H_A has a nontrivial subgroup of rotations if and only if*

$$A = \bigcup_{v \in A_0} [v]_k \text{ for some } k \text{ and some } A_0 \subseteq V(P).$$

Moreover, any nontrivial subgroup of rotations in H_A is isomorphic to $\mathbb{Z}_{n/h}$ for some h.

Proof. First, assume H_A has a nontrivial rotation. Then there exists $\alpha \in H_A$ such that $\alpha = \iota_A\rho^k$ for some k. Since $\iota_A\rho^k \in H_A$, we have $[v]_k \subseteq A$ for all $v \in A$. Let A_0 be a subset of A consisting of a representative of each k class. Then $A = \bigcup_{v \in A_0} [v]_k$.

For the other direction, assume $A = \bigcup_{v \in A_0} [v]_k$ for some k and some $A_0 \subseteq V(P)$. Since for any $v \in A$ the set $[v]_k$ is the orbit of v under ρ^k, we know $v\rho^k \in A$. Thus $\alpha = \iota_A\rho^k$ is an element of H_A. Since ρ^k is a nontrivial rotation in D_n, α is a nontrivial rotation in H_A. Therefore $\langle\alpha\rangle$ is a nontrivial subgroup of rotations of H_A.

From Lemma 2.1c, $|\alpha| = \left|\rho^k\right| = n/k$. Since $|\alpha| = |\langle\alpha\rangle|$, the cyclic group $\langle\alpha\rangle$ is of order n/k, and is therefore isomorphic to $\mathbb{Z}_{n/k}$. □

These k classes can then be viewed as the building blocks for all cyclic and dihedral maximal subgroups since both contain subgroups of rotations. However, at this point we have no way of knowing whether a subpolygon formed by a union of more than one k class will have a cyclic or a dihedral symmetry group. The following lemma gives one method of proving or disproving that a maximal subgroup is cyclic:

Lemma 2.4. *Let A be a collection of vertices of P that can be written as a union of k classes, with k the smallest positive integer such that $A = \bigcup_{v\in A_0}[v]_k$ for some set of vertices A_0. Then $H_A \approx \mathbb{Z}_{n/k}$ if and only if H_A contains no reflections. Otherwise, $H_A \approx D_{n/k}$.*

Proof. First note that since k is the smallest positive integer such that A can be written as the union of k classes, k is also the smallest positive integer such that $\iota_A\rho^k \in H_A$. So, by Lemma 2.2, the subgroup of all rotations in H_A is isomorphic to $\mathbb{Z}_{n/k}$. So clearly if $H_A \approx \mathbb{Z}_{n/k}$ then every element in H_A is a rotation. Conversely, if H_A contains no reflections, then it must contain only rotations. Since the set of all rotations in H_A is $\left\langle\iota_A\rho^k\right\rangle$, we have $H_A = \left\langle\iota_A\rho^k\right\rangle \approx \mathbb{Z}_{n/k}$.

If $H_A \not\approx \mathbb{Z}_{n/k}$, then $\mathbb{Z}_{n/k}$ is still the subgroup of all rotations in H_A. So H_A must contain some element $\gamma \notin \mathbb{Z}_{n/k}$. This implies that H_A is not cyclic. Since all finite plane symmetry groups are either cyclic or dihedral, H_A is thus isomorphic to a dihedral group. The only dihedral group that contains $\mathbb{Z}_{n/k}$ as its largest cyclic subgroup is $D_{n/k}$. Therefore $H_A \approx D_{n/k}$. □

So, we can now state the problem of finding cyclic maximal subgroups as follows: for which values of n can we find a collection of k classes, A, such that H_A contains no reflections?

We begin this search by constructing some subpolygons with cyclic symmetry groups. In order to construct such subpolygons we make use of the concept of an integer partition. A partition of an integer m is a way of expressing m as the sum of positive integers. Each summand in the expression of m is called a part of the partition. The usual convention dictates that the parts of a partition be written in nonincreasing order, but for our purposes this requirement is irrelevant. For our construction we will need the following definitions:

Definition. Let A and B be collections of vertices of P.

(a) Two vertices of A are said to be adjacent in P_A if they are connected by an edge in the subpolygon P_A.

(b) Let $B \subseteq A$. If for all $v \in B$ there exists $w \in B$ such that v is adjacent to w in P_A, then B is said to be a consecutive subset of A.

(c) The set $\mathscr{E}_A$ is defined by $\mathscr{E}_A = \{P(v_i, v_j) \mid v_i, v_j \in A$ and v_i is adjacent to v_j in $P_A\}$.

Note that the polygonal distance always refers to the minimum number of edges of the original polygon P between two vertices, even in reference to two vertices of a subpolygon. Note also that vertices labeled with consecutive indices, for example w_1 and w_2, are always meant to be adjacent in P_A.

Theorem 2.1. *Let k divide n. Then for every partition of k into m distinct parts, with $m \geq 3$, there exists a cyclic maximal subgroup of S isomorphic to $\mathbb{Z}_{n/k}$.*

Proof. Let k divide n. Suppose $k = a_1 + a_2 + \ldots + a_m$, for some $m \geq 3$, where all the a_i are distinct. Using this partition of k, we wish to construct a set of vertices A that will give rise to a subpolygon P_A with only rotations in its symmetry group H_A. To do this, let $A_0 = \{v_{i_1}, v_{i_2}, \ldots, v_{i_m}\}$ be a set of vertices such that $P(v_{i_j}, v_{i_{j+1}}) = a_j$ for all $j = 1, 2, \ldots, m-1$. Consider the set $A = \bigcup_{v \in A_0} [v]_k$. For convenience, let $A = \{w_1, w_2, \ldots, w_q\}$, where the first m vertices in A are precisely the elements of A_0, and $q = mn/k$. We break up the proof that $H_A \approx \mathbb{Z}_{n/k}$ into three parts.

(a) The set A has the following properties:

(i) $\mathscr{E}_A = \{a_i \mid i = 1, 2, \ldots, m\}$.

(ii) If B is any consecutive subset of A, and $|B| \leq m+1$, then all elements of $\mathscr{E}_B$ are distinct.

We know from the construction of A that

$$P(w_i, w_{i+1}) = a_i \quad \text{for } i = 1, 2, \ldots, m-1.$$

Also, since $w_{m+1} = w_1 \rho^k$, we have $P(w_1, w_{m+1}) = k$. Since the set

$$\{w_1, w_2, \ldots, w_{m+1}\}$$

is a consecutive subset of A we may write

$$P(w_1, w_{m+1}) = P(w_1, w_m) + P(w_m, w_{m+1}).$$

So

$$\begin{aligned} P(w_m, w_{m+1}) &= P(w_1, w_{m+1}) - P(w_1, w_m) \\ &= k - (a_1 + a_2 + \ldots + a_{m-1}) \\ &= a_m. \end{aligned}$$

This shows that $\mathscr{E}_{A_0} = \{a_i \mid i = 1, 2, \ldots, m\}$. Now let $w_j \in A$. Then there exists an $\ell \in \mathbb{N}$ such that $w_j = w_i \rho^{\ell k}$ for some $w_i \in A_0$ by the construction of A. So (recalling Equation (2)),

$$P(w_j, w_{j+1}) = P(w_i\rho^{\ell k}, w_{i+1}\rho^{\ell k}) = P(w_i, w_{i+1}) = a_i,$$

which proves the first property.

To prove the second, note if B is consecutive subset of A and $|B| \leq m+1$, then $\mathscr{E}_B$ has at most m elements. Each element of $\mathscr{E}_B$ must be a unique part of the partition of k, all of which are distinct.

(b) The subgroup of all rotations in H_A is isomorphic to $\mathbb{Z}_{n/k}$:

In order to show that the subgroup of all rotations in H_A is $\mathbb{Z}_{n/k}$, from Lemma 2.2 we need only show that k is the smallest positive integer such that $\iota_A\rho^k \in H_A$. Assume, rather, that $\iota_A\rho^j$ is an element of H_A for some j such that $0 < j < k$. Let $A_1 = A_0 \cup \{w_{m+1}\}$. Note that since $|A_1| = m+1$, every element of $\mathscr{E}_{A_1}$ is unique by Theorem 2.1a-ii. Also note that since $\iota_A\rho^j$ is a rotation, it preserves both adjacency and orientation of the vertices of A. With these two facts, we see that $w_1\rho^j$ and $w_2\rho^j$ cannot both lie in A_1. If they did, we would have $P(w_1, w_2) = P(w_1\rho^j, w_2\rho^j) \in \mathscr{E}_{A_1}$, contradicting Theorem 2.1a-ii. But recall that $P(w_1, w_1\rho^j) = j$, and by construction, $P(w_1, w_{m+1}) = k$. Since $j < k$, this implies that $w_1\rho^j = w_s$ for some $w_s \in A_1$. Since the vertices of A are indexed clockwise, we have that $s < m+1$. Now, since $w_1\rho^j = w_s$, and ρ^j is orientation preserving, we may write $w_2\rho^j = w_{s+1}$. From our argument above, since $w_1\rho^j \in A_1$, we know that $w_2\rho^j \notin A_1$. That is, $s+1 > m+1$. Since we have already established that $s < m+1$, we have $s < m+1 < s+1$. This is impossible since both m and s are positive integers. Thus, k is the smallest positive integer such that $\iota_A\rho^k \in H_A$. Therefore, the subgroup of all rotations in H_A is isomorphic to $\mathbb{Z}_{n/k}$.

(c) H_A contains no reflections: We again proceed by contradiction. Suppose that P_A is symmetric about some line L. Let $d \in D_n$ be the reflection of P about L. Then $\alpha = \iota_A d \in H_A$. There are two cases to consider.

(i) L passes through some vertex $w_i \in A$. Let $B = \{w_{i-1}, w_i, w_{i+1}\}$. Now, $w_i\alpha = w_i$, and $w_{i+1}\alpha = w_{i-1}$. Moreover, from Theorem 2.1a-i,

$$P(w_i, w_{i+1}) = P(w_i\alpha, w_{i+1}\alpha) = P(w_i, w_{i-1}) = a_j \text{ for some} j.$$

But $P(w_{i-1}, w_i), P(w_i, w_{i+1}) \in \mathscr{E}_B$, and B is a consecutive subset of A of order 3. Since $m \geq 3$, $|B| = 3 < m+1$. So $P(w_{i-1}, w_i)$ and $P(w_i, w_{i+1})$ must be distinct by Theorem 2.1a-i, which is a contradiction.

(ii) L passes through no vertices of A. Then there exists $w_i \in A$ such that $w_i\alpha = w_{i+1}$ and $w_{i-1}\alpha = w_{i+2}$. Thus, from Theorem 2.1a-i:

$$P(w_i, w_{i-1}) = P(w_i\alpha, w_{i-1}\alpha) = P(w_{i+1}, w_{i+2}) = a_j \quad \text{for some } j.$$

Let $B = \{w_{i-1}, w_i, w_{i+1}, w_{i+2}\}$. Then B is a consecutive subset of order 4. Since $m \geq 3$, $|B| \leq m + 1$. But clearly not every element of $\mathscr{E}_B$ is distinct. By Theorem 2.1a-ii, this is impossible.

Therefore, the subpolygon P_A has no line of symmetry. Thus H_A contains no reflections. The theorem now follows. □

Corollary 2.1. *If k divides n and $k \geq 6$, then S has a maximal subgroup isomorphic to $\mathbb{Z}_{n/k}$.*

Proof. Let k divide n and $k \geq 6$. Then $k = 1 + 2 + (k - 3)$, and $k - 3 > 2$, so k can be partitioned into at least 3 distinct parts. Thus by Theorem 2.1, S contains a maximal subgroup isomorphic to $\mathbb{Z}_{n/k}$. □

As an example of the construction of subpolygons with cyclic symmetry groups given in Theorem 2.1, let $n = 24$ and $k = 8$. Consider the partition $8 = 1+2+5$. Now let $A_0 \subseteq V(P)$ such that $A_0 = \{v_1, v_2, v_4\}$ (see Figure 1). Note that $P(v_1, v_2) = 1$ and $P(v_2, v_4) = 2$. If we then consider the set $A = \bigcup_{v \in A_0} [v]_8$, we get

$$\begin{aligned} A &= [v_1]_8 \cup [v_2]_8 \cup [v_4]_8 \\ &= \{v_1, v_9, v_{17}\} \cup \{v_2, v_{10}, v_{18}\} \cup \{v_4, v_{12}, v_{20}\} \\ &= \{v_1, v_2, v_4, v_9, v_{10}, v_{12}, v_{17}, v_{18}, v_{20}\}. \end{aligned}$$

So A is essentially made up of 3 copies of A_0 evenly spaced around the polygon, and one can see that the polygonal distances between adjacent vertices in P_A are all elements of the partition of k. The entire subpolygon in Figure 1 is P_A.

We have shown that, given a regular polygon with n sides and a positive integer $k \geq 6$ that divides n, we can construct a subpolygon with symmetry group $\mathbb{Z}_{n/k}$.

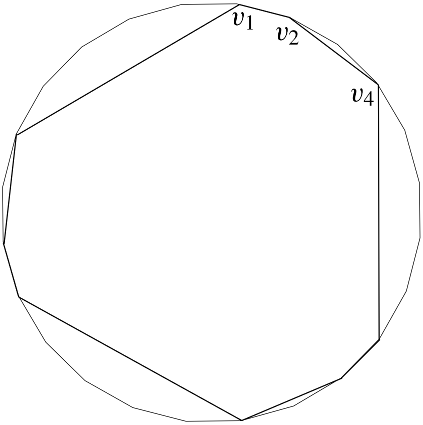


Figure 1. Subpolygon with symmetry group $\mathbb{Z}_3$ in a regular 24-gon.

With the following three results, we show that if $k < 6$ then no such subpolygon exists.

Lemma 2.5. *If A is the union of exactly two k classes, then H_A is isomorphic to a dihedral group.*

Proof. Let $A = [v]_k \cup [w]_k$, where $[v]_k \cap [w]_k = \varnothing$. If $[v]_k \cup [w]_k = [v]_j$ for some j, then P_A is regular and the result holds. So assume that is not the case. We know there exists some reflection $d \in D_n$ such that $vd = w$ and $wd = v$. Recall that $P_{[v]_k}$ and $P_{[w]_k}$ are the unique regular subpolygons with n/k vertices containing v and w respectively. Since d is an isometry, the subpolygon formed by the vertex set $d([v]_k)$ must also be a regular n/k-gon. Since $vd = w \in [w]_k$, we have $d([v]_k) = [w]_k$ by the uniqueness of $P_{[w]_k}$. Similarly $d([w]_k) = [v]_k$. So $d|_A(A) = A$, and $\iota_A d$ is a reflection in H_A. From Lemma 2.4, since A is the union of k classes and H_A contains a reflection, H_A is isomorphic to a dihedral group. □

Lemma 2.6. *For any $A \subseteq V(P)$, let $A^c = V(P) \setminus A$. If A is a subset of $V(P)$ such that $|A| > 2$ and $|A^c| > 2$, then $H_A \approx H_{A^c}$.*

Proof. Let A be a subset of $V(P)$ such that $|A| > 2$ and $|A^c| > 2$. Recall from Equation (1) that $H_A = \{\iota_A d \mid d \in D_n \text{ and } d|_A(A) = A\}$. Let

$$D_n(A) = \{d \in D_n \mid d|_A(A) = A\} \quad \text{and} \quad D_n(A^c) = \left\{d \in D_n \mid d|_{A^c}(A^c) = A^c\right\}.$$

Since all mappings in D_n are bijections from $V(P)$ to $V(P)$, we have $d|_A(A) = A$ if and only if $d|_{A^c}(A^c) = A^c$. So $D_n(A) = D_n(A^c)$. Now define the mapping $\phi : H_A \to H_{A^c}$ by $\phi(\iota_A d) = \iota_{A^c} d$. Such a mapping is guaranteed to exist since $D_n(A) = D_n(A^c)$, and furthermore, the same equality shows that ϕ maps onto H_{A^c}. Since $|A| > 2$, every element of H_A can be extended to a unique element of D_n. So $\iota_A d_1 = \iota_A d_2$ if and only if $d_1 = d_2$, and ϕ is thus well defined. To see that ϕ is one-to-one, suppose that $\phi(\iota_A d_1) = \phi(\iota_A d_2)$. Then $\iota_{A^c} d_1 = \iota_{A^c} d_2$. By assumption, $|A^c| > 2$, so uniqueness of extension gives $d_1 = d_2$. Thus $\iota_A d_1 = \iota_A d_2$. It follows from Lemma 2.1a that for any $d_1, d_2 \in D_n(A)$,

$$\phi((\iota_A d_1)(\iota_A d_2)) = \phi(\iota_A d_1 d_2) = \iota_{A^c} d_1 d_2 = (\iota_{A^c} d_1)(\iota_{A^c} d_2) = \phi(\iota_A d_1)\,\phi(\iota_A d_2).$$

So ϕ is a homomorphism, and $H_A \approx H_{A^c}$. □

Lemma 2.7. *Let k divide n, $k \leq 5$, and $k \neq n/2$. Then S has no maximal subgroup isomorphic to $\mathbb{Z}_{n/k}$.*

Proof. Let k be as described and $A \subseteq V(P)$ such that A is the union of m k classes. Since $k \leq 5$, we know $m \leq 5$. Then we can write A^c as the union of $(k - m)$ k classes. Since $k \neq n/2$, both A and A^c have more than two elements. Since $m + (k - m) = k \leq 5$, either $m \leq 2$ or $k - m \leq 2$. Thus from Lemma 2.5, H_A or H_{A^c} is isomorphic to a dihedral group. From Lemma 2.6, $H_A \approx H_{A^c}$, so both

H_A and H_{A^c} are isomorphic to a dihedral group. Thus the symmetry group of any subpolygon whose vertices are the union of k classes is dihedral. So $\mathbb{Z}_{n/k}$ does not occur as a maximal subgroup of S. □

We now have all of the information necessary to determine exactly for which regular polygons P, $S(P)$ contains cyclic maximal subgroups other than $\mathbb{Z}_1$ and $\mathbb{Z}_2$.

Theorem 2.2. *Let P be a regular polygon with n sides. Then S contains a maximal subgroup isomorphic to $\mathbb{Z}_m$, for some $m \geq 3$, if and only if $n = km$ for some $k \geq 6$. In particular*, 18 *is the smallest value of n for which this occurs.*

Proof. Suppose that S contains a cyclic maximal subgroup isomorphic to $\mathbb{Z}_m$ for some $m \geq 3$. Then, by Lemma 2.3, $m = n/k$ for some k. Since $m \geq 3$, we know that $k \neq n/2$. Then by the contrapositive of Lemma 2.7, since $\mathbb{Z}_m = \mathbb{Z}_{n/k}$ is a maximal subgroup of S, then $k \geq 6$. Thus $n = km$ where $k \geq 6$ and $m \geq 3$.

Conversely, suppose $n = km$ for some k and m such that $k \geq 6$ and $m \geq 3$. By Corollary 2.1, since $k \geq 6$ and k divides n, S contains a maximal subgroup isomorphic to $\mathbb{Z}_{n/k} = \mathbb{Z}_m$. The result follows since $|\mathbb{Z}_m| = m \geq 3$.

Thus $6 \times 3 = 18$ is the fewest number of vertices of a regular polygon P such that S has a cyclic maximal subgroup with more than two elements. □

For values of $n \leq 40$, the only n for which S contains a nontrivial (other than $\mathbb{Z}_1$ and $\mathbb{Z}_2$) cyclic maximal subgroup are 18, 21, 24, 27, 28, 30, 32, 33, 35, 36, 39, and 40. In contrast, there are 26 values of $n \leq 40$ for which S contains a nontrivial dihedral maximal subgroup.

The case where n is even and $k = n/2$ was ignored throughout the paper in order to ensure that each individual k class formed a regular polygon with at least 3 vertices. However, the consequences of allowing k to take the value of $n/2$ are nontrivial. If $k = n/2$, then any subpolygon formed by exactly one k class is simply a line segment connecting two antipodal vertices, so its symmetry group is $\mathbb{Z}_2$. But suppose that we take the union of two k classes of adjacent vertices, $[v_1]_k$ and $[v_2]_k$. Then the resulting subpolygon is a nonsquare rectangle (if $n > 4$). The symmetry group of a nonsquare rectangle contains 4 elements: the identity, one rotation of order 2, and two reflections. This symmetry group is known by more than one name, including the Klein four group, and $\mathbb{Z}_2 \times \mathbb{Z}_2$. But if we define the dihedral groups in terms of generators and relations as $D_n = \langle d_1, d_2 \mid d_1^2 = d_2^2 = (d_1 d_2)^n = \iota_P \rangle$, then the symmetry group of a nonsquare rectangle is seen to be isomorphic to D_2 [Gallian 2002, p. 442]. This maximal subgroup exists only for all even $n > 4$ (using the construction above).

With this final case considered, we have characterized all maximal subgroups of S for any regular polygon P.

n	Dihedral maximal subgroups	Cyclic maximal subgroups
4	D_4	$\mathbb{Z}_1$, $\mathbb{Z}_2$
8	D_2, D_4, D_8	$\mathbb{Z}_1$, $\mathbb{Z}_2$
13	D_{13}	$\mathbb{Z}_1$, $\mathbb{Z}_2$
18	D_2, D_3, D_6, D_9, D_{18}	$\mathbb{Z}_1$, $\mathbb{Z}_2$, $\mathbb{Z}_3$
24	D_2, D_3, D_4, D_6, D_8, D_{12}, D_{24}	$\mathbb{Z}_1$, $\mathbb{Z}_2$, $\mathbb{Z}_3$, $\mathbb{Z}_4$
30	D_2, D_3, D_5, D_6, D_{10}, D_{15}, D_{30}	$\mathbb{Z}_1$, $\mathbb{Z}_2$, $\mathbb{Z}_3$, $\mathbb{Z}_5$

Table 1. Maximal subgroups for various values of n.

Theorem 2.3. *Let P be a regular polygon with n sides, and let $n > 4$. Let S be the semigroup of partial symmetries of P. Then every maximal subgroup of S is isomorphic to one of the following groups, and S contains a maximal subgroup isomorphic to each of the following*:

(a) $\mathbb{Z}_1$;

(b) $\mathbb{Z}_2$;

(c) $D_{n/k}$ *for all k that divide n*;

(d) $\mathbb{Z}_{n/k}$ *for all k that divide n such that $k \geq 6$ and $n/k \geq 3$.*

Table 1 shows exactly which maximal subgroups occur for a variety of values of n.

Acknowledgement

This work was conducted by Tom Shelly as an undergraduate research project under the guidance of Dr. Janet Mills.

References

[Gallian 2002] J. A. Gallian, *Contemporary Abstract Algebra*, Houghton Mifflin Company, Boston, 2002.

[Mills 1990a] J. E. Mills, "The inverse semigroup of partial symmetries of a convex polygon", *Semigroup Forum* **41**:2 (1990), 127–143. MR 91j:20158 Zbl 0704.20051

[Mills 1990b] J. E. Mills, "Inverse semigroups through groups", *Internat. J. Math. Ed. Sci. Tech.* **21**:1 (1990), 93–98. MR 1036371 Zbl 0693.20059

[Mills 1993] J. E. Mills, "Factorizable semigroup of partial symmetries of a regular polygon", *Rocky Mountain J. Math.* **23**:3 (1993), 1081–1090. MR 94j:20077 Zbl 0815.20064

Received: 2007-06-11 Accepted: 2007-11-01

shellyt@seattleu.edu
Department of Mathematics, 901 12th Ave.,
P.O. Box 222000, Seattle, WA 98122-1090, United States

jemills@seattleu.edu
Department of Mathematics, 901 12th Ave.,
P.O. Box 222000, Seattle, WA 98122-1090, United States

Divisibility of class numbers of imaginary quadratic function fields

Adam Merberg

(Communicated by Ken Ono)

We consider applications to function fields of methods previously used to study divisibility of class numbers of quadratic number fields. Let K be a quadratic extension of $\mathbb{F}_q(x)$, where q is an odd prime power. We first present a function field analog to a Diophantine method of Soundararajan for finding quadratic imaginary function fields whose class groups have elements of a given order. We also show that this method does not miss many such fields. We then use a method similar to Hartung to show that there are infinitely many imaginary K whose class numbers are indivisible by any odd prime distinct from the characteristic.

1. Introduction and statement of results

The study of the structure of class groups of imaginary quadratic number fields dates back to Gauss, who posed the problem of finding all positive square-free d such that the class group of $\mathbb{Q}(\sqrt{-d})$, which we denote by $\mathrm{Cl}(-d)$, has some fixed order h. Heegner [1952], Baker [1967] and Stark [1967] solved Gauss's problem in the case $h = 1$, showing that there are only nine imaginary quadratic fields of class number 1. Baker [1971] and Stark [1975] later presented solutions to the case $h = 2$. A famous theorem of Siegel says that for $\epsilon > 0$, there exist positive constants $c_1(\epsilon)$ and $c_2(\epsilon)$ such that for each square-free d we have

$$c_1(\epsilon)d^{\frac{1}{2}-\epsilon} < h(-d) < c_2(\epsilon)d^{\frac{1}{2}+\epsilon}.$$

But this bound was ineffective. Goldfeld [1976] and Gross and Zagier [1983] showed that Gauss's problem is effectively computable for any h.

Of interest in the study of the structure of the class groups of the imaginary quadratic fields is the presence or absence of c-torsion for positive integers c. For $c = 2$, the answer to this question follows from Gauss's genus theory. For odd

MSC2000: primary 11R29; secondary 11R11.

Keywords: number theory, quadratic function fields, class numbers, class groups, divisibility.

primes c, a conjecture of Cohen and Lenstra [1984] states that the "probability" that $\mathrm{Cl}(-d)$ has an element of order c is

$$1-\prod_{1\le k<\infty}(1-c^{-k}).$$

With a few exceptions, little is known about divisibility of class numbers of imaginary quadratic number fields. A theorem of Davenport and Heilbronn [1971] shows that for $c=3$, the proportion of d for which the order of $\mathrm{Cl}(-d)$ is prime to 3 is at least $1/2$. Other results do not even give positive proportions. Soundararajan [2000] used a Diophantine construction to show that the number of $d<X$ such that $\mathrm{Cl}(-d)$ has an element of even order c is

$$\gg \begin{cases} X^{1/2+2/c-\epsilon}, & \text{if } c\equiv 0 \pmod 4,\\ X^{1/2+3/(c+2)-\epsilon}, & \text{if } c\equiv 2 \pmod 4.\end{cases} \tag{1}$$

This can also be used to give a bound for c odd since if $\mathrm{Cl}(-d)$ has an element of order $2c$, then it also has an element of order c. On the question of indivisibility of class numbers, Kohnen and Ono [1999] showed that for odd primes c, the number of $d<X$ such that $\mathrm{Cl}(-d)$ has no $c-$torsion is at least

$$\left(\frac{2(c-2)}{\sqrt{3}(c-1)}-\epsilon\right)\frac{\sqrt{X}}{\log X},$$

for any $\epsilon>0$.

In the setting of function fields, Friedman and Washington [1989] conjectured an analog of the Cohen–Lenstra heuristics. Achter [2006] used methods of algebraic geometry to prove this conjecture in a recent paper.

In this paper, we consider divisibility of class numbers of imaginary quadratic function fields. We use several styles of arguments applied to imaginary quadratic number fields prior to the work of Achter. We let q be a power of an odd prime, and define $k:=\mathbb{F}_q(x)$ and $A:=\mathbb{F}_q[x]$, the rational function field and polynomial ring over the finite field with q elements. Denote by $\mathrm{Cl}(f)$ the (divisor) class group of the function field $k(\sqrt{f})$ for f square-free and let $h(f)=\#\mathrm{Cl}(f)$. We look in particular at the case when $\deg f$ is odd. This is an analog of the case of an imaginary quadratic number field in which the prime at infinity ramifies and the unit group has rank 0. This case also has the property that the class number of the function field $k(\sqrt{f})$ is the same as the class number of its maximal order [Rosen 2002, Chapter 14]. Soundararajan [2000, Proposition 1] used solutions to the Diophantine equation $t^2d=m^c-n^2$ to find d such that $\mathrm{Cl}(-d)$ has an element of order c. The following is our analogous result for function fields.

Theorem 1.1. *Let $c\ge 3$ be a positive odd integer. Let $f\in A$ be a square-free polynomial of odd degree. If there exist nonzero $m,n,t\in A$ such that $m^c=n^2-t^2f$*

with $(m, n) = 1$ *and* $c \deg m < p \deg f$, *where* p *is the smallest prime dividing* c, *then* $\mathrm{Cl}(f)$ *has an element of order* c.

Remark 1. Cardon and Ram Murty [2001] used a similar Diophantine method to give a bound similar to Equation (1) in the function field case.

In the number field case, Soundararajan showed that any d such that $\mathrm{Cl}(-d)$ has an element of order c satisfies a Diophantine condition similar to that in his construction. The following is a function field analog of his result. In the theorem, the Diophantine condition from Theorem 1.1 corresponds to the case $l = 1$. Like Soundararajan's result, this is proven only in the case of c prime, but we expect a similar result to hold if c is composite.

Theorem 1.2. *Let* $c \geq 3$ *be prime and let* $f \in A$ *be a square-free polynomial of odd degree. Denote by* $h_c(f)$ *the number of elements of order* c *in* $\mathrm{Cl}(f)$. *Let* $C_+ = 2$ *and* $C_- = 1$. *If for each choice of* ϵ *in* $\{+, -\}$, D_ϵ *is the number of solutions in polynomials* l, m, n *and* t *with* l, m, n *monic to*

$$lm^c = n^2 l^2 - t^2 f, \quad \textit{where } l \mid f, \ (m, fn) = 1 \textit{ and } \deg lm < \frac{C_\epsilon}{2} \deg f, \tag{2}$$

then $D_- \leq h_c(f) \leq D_+$.

We also consider indivisibility of class numbers of quadratic function fields. Hartung [1974] used a famous class number relation to show that there are infinitely many imaginary quadratic number fields whose class numbers are not divisible by 3, and his method extends to any odd prime. We prove the following analog for function fields.

Theorem 1.3. *If* $c = 4$ *or* c *is an odd prime not dividing* q, *then there exist infinitely many quadratic imaginary function fields* K *over* k *with class number not divisible by* c.

Theorem 1.1 will be proven in Section 2, and Theorem 1.2 will be proven in Section 3. In Section 4, we prove Theorem 1.3. In Section 5, we conclude with some numerical examples.

2. Proof of Theorem 1.1

Some additional definitions and comments will be useful in proving these theorems. Given a quadratic extension K of the rational function field k, we can define a norm map $N : K \to k$ taking x to the product of its Galois conjugates (or $N(x) = x^2$ for $x \in k$). Furthermore, if B is the integral closure of A in K, then we can define the norm of an ideal $\mathfrak{I} \subset B$ as the ideal in A generated by elements of the set $\{N(b) : b \in \mathfrak{I}\}$. The ring B is a Dedekind domain and thus has unique factorization of ideals [Rosen 2002, Chapter 7]. In the special case that $\mathfrak{I} \subset B$ is principal, say

$\mathfrak{I} = (b)$, it is clear that $N(\mathfrak{I}) = (N(b))$. We also note that since A is a principal ideal domain, $N(\mathfrak{I})$ is principal even if $\mathfrak{I}$ is not. We also note that if $K = k(\sqrt{f})$, then $B = A[\sqrt{f}]$. This follows immediately from the formula for the roots of a quadratic equation. In the case of quadratic number fields, we have multiple cases depending on the parity of the discriminant, but in the function field case multiple cases do not arise since 2 is a unit of A so that an element of the form $(a+b\sqrt{f})/2$ (as would be given by the quadratic formula) can always be rewritten as $a'+b'\sqrt{f}$ with $a, b \in A$.

Proof of Theorem 1.1. Let $K = k(\sqrt{f})$ and consider the factorization of ideals $(m)^c = (n+t\sqrt{f})(n-t\sqrt{f})$ in the integral closure B of A in K. We claim that the ideals on the right side are relatively prime. If $\mathfrak{h}$ is a common prime divisor, then $(n+t\sqrt{f}) + (n-t\sqrt{f}) = 2n \in \mathfrak{h}$. However, we also have $m^c \in \mathfrak{h}$, which implies that $m \in \mathfrak{h}$ since $\mathfrak{h}$ is prime. Then $\mathfrak{h} | ((m, n))$, but this is a contradiction since $(m, n) = 1$.

Thus, the factorization of ideals shows that each of $(n \pm t\sqrt{f})$ is a cth power. Since $\{1, \sqrt{f}\}$ is a basis for B over A, let $\mathfrak{b}^c = (n+t\sqrt{f})$. We show that $\mathfrak{b}$ has order exactly c in $\mathrm{Cl}(f)$. Otherwise, $\mathfrak{b}$ has order $r < c$. Then $\mathfrak{b}^r = (u+v\sqrt{f})$ for some $u, v \in A$. We now consider the norms of each side of the equation $(n+t\sqrt{f}) = \mathfrak{b}^c$. On the left side, we have

$$N(n+t\sqrt{f}) \;=\; (n^2-t^2f) \;=\; (m)^c.$$

On the right side,

$$N(\mathfrak{b}^c) \;=\; N(\mathfrak{b}^r)^{c/r} \;=\; (u^2-v^2f)^{c/r}.$$

Comparing degrees now gives $c \deg m = c/r \cdot \deg(u^2-v^2f)$. Since the prime at infinity ramifies in $k(\sqrt{f})$, the unit group of the integral closure of A in $k(\sqrt{f})$ has rank 0, thus it follows that $n+t\sqrt{f} = [\alpha(u+v\sqrt{f})]^{c/r}$ for some $\alpha \in \mathbb{F}_q^\times$. Since $t \neq 0$, it is immediate that $v \neq 0$. Thus $v^2 f \neq 0$ has odd degree, and since u^2 has even degree, $\deg(u^2-v^2f) \geq \deg f$. Then $c \deg m \geq c/r \cdot \deg f$. But $c/r \geq p$, and $c \deg m < p \deg f$ by hypothesis, so it must be that $\mathfrak{b}$ has order exactly c in $\mathrm{Cl}(f)$. □

Remark 2. Soundararajan's Proposition 1 in [Soundararajan 2000] also holds if c is even. Indeed, in the function field case, the proof goes through without the explicit assumption that c is odd, but the conditions $m^c = n^2 - t^2 f$ and $c \deg m < p \deg f$ are never simultaneously satisfied for c even.

3. Proof of Theorem 1.2

Proof of Theorem 1.2. We first prove the lower bound. Suppose that we have a solution in l, m, n to t to Equation (2). We will show that the pair of solutions

(l, m, n, t) and $(l, m, n, -t)$ uniquely determines a pair of two ideals $\mathfrak{a}$ and $\bar{\mathfrak{a}}$ of order c in $\mathrm{Cl}(f)$ such that $N(\mathfrak{a}) = N(\bar{\mathfrak{a}})$ has degree less than $\frac{1}{2}\deg f$.

Let $K = k(\sqrt{f})$ and, as before, denote by B the integral closure of A in K. Consider the factorization of ideals in B:

$$(l)(m)^c \;=\; (lm^c) \;=\; (nl + t\sqrt{f})(nl - t\sqrt{f}). \tag{3}$$

Since $l|f$, l is a product of primes in K which are ramified over k, whence $(l) = \mathfrak{l}^2$ for some ideal $\mathfrak{l}$ of B. Letting $\mathfrak{h} = (nl + t\sqrt{f}, nl - t\sqrt{f})$, it follows from Equation (3) that $\mathfrak{l}^2|\mathfrak{h}^2$ so $\mathfrak{l}|\mathfrak{h}$. Furthermore, $\mathfrak{h}^2$ contains both lm^c and $(nl)^2 = n^2l^2$. Since $ln|fn$ and $(m, fn) = 1$, it follows that $(m^c, nl) = (m^c, n^2l) = 1$. Thus $(lm^c, n^2l) = l$, so $l \in \mathfrak{h}^2$. This shows that $\mathfrak{h}^2$ divides $(l) = \mathfrak{l}^2$, thus $\mathfrak{h} = \mathfrak{l}$.

We can now write $(nl + t\sqrt{f}) = \mathfrak{b}\mathfrak{l}$ and $(nl - t\sqrt{f}) = \bar{\mathfrak{b}}\mathfrak{l}$ where $\mathfrak{b}$ and $\bar{\mathfrak{b}}$ are relatively prime. Since $\mathfrak{l}^2 = (l)$, we have that $\bar{\mathfrak{b}}\mathfrak{b} = (m)^c$, so $\mathfrak{b}$ and $\bar{\mathfrak{b}}$ must both be cth powers, say $\mathfrak{b} = \beta^c$ and $\bar{\mathfrak{b}} = \bar{\beta}^c$ where β is an ideal of norm m. Define $\mathfrak{a} = \beta\mathfrak{l}$. Clearly $\mathfrak{a} \neq \bar{\mathfrak{a}}$ since otherwise we would have $\mathfrak{b} = \bar{\mathfrak{b}}$, from which it would follow that $t = 0$.

We now show that $\mathfrak{a}$ has order exactly c. Since

$$\mathfrak{a}^c \;=\; \beta^c\mathfrak{l}^c \;=\; \mathfrak{b}\mathfrak{l}(l)^{(c-1)/2} \;=\; (nl + t\sqrt{f})(l)^{(c-1)/2}$$

is principal, $\mathfrak{a}$ has order dividing c. Suppose $\mathfrak{a}$ is principal and write $\mathfrak{a} = (a + b\sqrt{f})$ with $a, b \in A$ and b nonzero (if b is 0 it follows from $\mathfrak{a}^c = (nl + t\sqrt{f})(l)^{(c-1)/2}$ that $t = 0$). Then $N(\mathfrak{a}) = N((a + b\sqrt{f})) = (a^2 - b^2 f)$. Since a^2 has even degree and $b^2 f$ has odd degree, the degree of this must be at least $\deg f$. However, this implies that $\frac{1}{2}\deg f > \deg lm = \deg N(\mathfrak{a}) \geq \deg f$, a contradiction. Thus $\mathfrak{a}$ is not principal and must have order c.

We now consider the degree of a generating element of $N(\mathfrak{a})$. We have

$$N(\mathfrak{a})^c \;=\; (nl + t\sqrt{f})(l)^{(c-1)/2} \cdot (nl - t\sqrt{f})(l)^{(c-1)/2} \;=\; (lm^c)(l)^{c-1} \;=\; ((lm)^c).$$

Thus $N(\mathfrak{a})^c$ is generated by $(lm)^c$ whence $N(\mathfrak{a}) = (lm)$ which has degree $\deg lm < \frac{1}{2}\deg f$.

We now show that different solutions to Equation (2) with $C_\epsilon = 1$ correspond to distinct pairs of ideals of order c in $\mathrm{Cl}(f)$. Consider two distinct solutions (l_1, m_1, n_1, t_1) and (l_2, m_2, n_2, t_2) with $\deg l_i m_i < \frac{1}{2}\deg f$. Let $\mathfrak{a}_i$ denote the corresponding ideals having order c in $\mathrm{Cl}(f)$. Suppose that $\mathfrak{a}_1$ and $\mathfrak{a}_2$ are in the same class in $\mathrm{Cl}(f)$. Then $\mathfrak{a}_1\bar{\mathfrak{a}}_2$ is principal, so let $\mathfrak{a}_1\bar{\mathfrak{a}}_2 = (a + b\sqrt{f})$. Then

$$(a^2 - b^2 f) \;=\; N(\mathfrak{a}_1\bar{\mathfrak{a}}_2) \;=\; (l_1m_1l_2m_2).$$

Considering degrees in this equality, we see that

$$\deg(a^2 - b^2 f) = \deg l_1m_1 + \deg l_2m_2 < \deg f,$$

so $b=0$ and $\mathfrak{a}_1\bar{\mathfrak{a}}_2=(a)$. Thus

$$(a)^c = \mathfrak{a}_1^c\bar{\mathfrak{a}}_2^c = (n_1l_1+t_1\sqrt{f})(n_2l_2-t_2\sqrt{f})(l_1l_2)^{(c-1)/2}.$$

Since the $\sqrt{f}$ term on the right side must be zero, we have $n_1l_1t_2=n_2l_2t_1$. From the equation in Equation (2), it is clear that $(n_il_i, t_i)^2|l_im_i^c$. Since l_i is square-free, this implies that $(n_il_i, t_i)|m_i^c$. Since also $(n_il_i, t_i)|fn_i$ and $(m_i^c, fn_i)=(m_i, fn_i)=1$, it follows that $(n_il_i, t_i)=1$. Using the fact that l_i and n_i are monic, we have that $t_1=t_2$ and $n_1l_1=n_2l_2$. Substituting into the Equation (2) gives $l_1m_1^c=l_2m_2^c$. Since l_i divides f and m_i is prime to f, it follows that $m_1=m_2$ and $l_1=l_2$, so the solutions are not distinct. A similar argument shows that $\mathfrak{a}_1$ and $\bar{\mathfrak{a}}_2$ are in different classes unless $(l_2, m_2, n_2, t_2)=(l_1, m_1, n_1, -t_1)$. Thus each pair of solutions of the form (l_1, m_1, n_1, t_1), $(l_1, m_1, n_1, -t_1)$ yields a unique pair of elements of order c. This completes the proof of the lower bound $D_- \le h_c(f)$.

It remains to show the upper bound. Define $s=h_c(f)/2$ for simplicity of notation. We note that s is an integer since if $\mathscr{C}\in \mathrm{Cl}(f)$ has order c, then so does $\mathscr{C}^{-1}$ (note also that since $c>2$ these elements must be distinct). Let $\mathscr{C}_1, \bar{\mathscr{C}}_1, \dots, \mathscr{C}_s, \bar{\mathscr{C}}_s$ be the classes of order c in $\mathrm{Cl}(f)$. By Theorem 4.4 in [Hayes 1999], we can choose integral ideals $\mathfrak{a}_i\in\mathscr{C}_i$ and $\bar{\mathfrak{a}}_i\in\bar{\mathscr{C}}_i$ of minimal degree less than $(\deg f-1)/2$. Furthermore, it is clear that ideals $\mathfrak{a}_i$ and $\bar{\mathfrak{a}}_i$ chosen to be minimal in this way are not divisible by any principal ideals.

Starting with a minimal pair of ideals $\mathfrak{a}_i$ and $\bar{\mathfrak{a}}_i$ we construct a solution to Equation (2) with $C_\epsilon=2$. Write $\mathfrak{a}_i=\mathfrak{b}_i\mathfrak{l}_i$ where $\mathfrak{l}_i$ is either the unit ideal or has order 2 in $\mathrm{Cl}(f)$ and $\mathfrak{b}_i$ is not divisible by any ideals of order 2. Similarly, write $\bar{\mathfrak{a}}_i=\bar{\mathfrak{b}}_i\bar{\mathfrak{l}}_i$ (in fact, $\mathfrak{l}_i=\bar{\mathfrak{l}}_i$). Then denote the unique monic generator of $N(\mathfrak{l}_i)$ by l_i. Note that each prime dividing l_i also divides f since any prime dividing $\mathfrak{l}_i$ is ramified over k. Since $\mathfrak{l}_i$ is not divisible by any principal ideal, l_i is not divisible by the square of any prime, so in particular $l_i|f$. Define m_i to be the monic generator for $N(\mathfrak{b}_i)$. Then

$$\deg l_im_i = 2\deg\mathfrak{b}_i\mathfrak{l}_i = 2\deg\mathfrak{a}_i \le \deg f-1 < \frac{C_+}{2}\deg f.$$

Since $\mathfrak{a}_i$ has order c, we can write $\mathfrak{a}_i^c=(a_i+b_i\sqrt{f})$ for some polynomials a_i and b_i, and we can assume that a_i is monic. Then $(l_i)^{(c-1)/2}=\mathfrak{l}_i^{c-1}$ divides $\mathfrak{a}_i^c$, so $l_i^{(c-1)/2}$ divides both a_i and b_i, so we may write $a_i=w_il_i^{(c-1)/2}$ and $b_i=t_il_i^{(c-1)/2}$. Since also $\bar{\mathfrak{a}}_i^c=(a_i-b_i\sqrt{f})$, we have that $(l_im_i)^c=l_i^{c-1}w_i^2-l_i^{c-1}t_i^2f$, and so $l_im_i^c=w_i^2-t_i^2f$. From the assumption that $l_i|f$ it follows that $l_i|w_i^2$, and since l_i is square-free this implies that $l_i|w_i$. Write $w_i=n_il_i$. Since $a_i=w_il_i^{(c-1)/2}$ and l_i are both monic, n_i is also monic. Thus, we have a solution to $l_im_i^c=n_i^2l_i^2-t_i^2f$ with $l_i|f$ and $\deg l_im_i<\frac{C_+}{2}\deg f$. Since t_i is not restricted to being monic, we note that substituting $-t$ for t gives another solution.

We now show that for the solutions constructed, $(m_i, n_i f) = 1$. Since $\mathfrak{b}$ and $\bar{\mathfrak{b}}$ are not divisible by any ideals of order 2, it follows that m_i, the monic generator for $N(\mathfrak{b}_i)$, is coprime to f. Since $(m_i, n_i)^2$ divides m_i^c and n_i^2 it also divides $t_i^2 f$, but since n_i is coprime to f, it follows that $(m_i, n_i)^2 | t_i^2$, so $(m_i, n_i) | t_i$. Since $n_i | a_i$, it follows that $(m_i, n_i) | (a_i + b_i\sqrt{f}) = \mathfrak{a}_i^c$. In particular, this means that each prime of A dividing (m_i, n_i) also divides $\mathfrak{a}_i$. However A is a principal ideal domain, but $\mathfrak{a}_i$ was taken not to be divisible by any principal ideal, so $(m_i, n_i) = 1$ and since also $(m_i, f) = 1$, we have $(m_i, n_i f) = 1$ as desired.

Finally, we must show that different pairs of ideals $\mathfrak{a}_i, \bar{\mathfrak{a}}_i$ and $\mathfrak{a}_j, \bar{\mathfrak{a}}_j$ give rise to distinct pairs of solutions as constructed above. If not, then it would follow that $a_i = a_j$ and $b_i = \pm b_j$. Then $\mathfrak{a}_i^c = \mathfrak{a}_j^c$, so $\mathfrak{a}_i = \mathfrak{a}_j$. Thus, we have shown that a pair of inverse elements of $\mathrm{Cl}(f)$ having order c gives a unique pair of solutions to Equation (2), concluding the proof of the upper bound $h_c(f) \le D_+$. □

4. Proof of Theorem 1.3

4.1. ***Background.*** We will use a class number relation over function fields proven by Yu. Before stating the proposition, we introduce some additional notation. If $m \in A$ is of odd degree but is not necessarily square-free, we define $h(m)$ to be the class number of the order $A[\sqrt{m}]$. This notation is consistent with our previous definition of $h(n)$ for n square-free because the class number of the maximal order $A[\sqrt{m}]$ is equal to the class number of the field $k(\sqrt{m})$ when m has odd degree and is square-free [Rosen 2002, Chapter 14]. We define $w(m) := \#A[\sqrt{m}]^\times/(q-1)$ and $h'(m) := h(m)/w(m)$. This allows us define the Hurwitz class number

$$H(m) := \sum_{n^2 | m} h'(m/n^2).$$

We now have defined all of the notation that we will need for the following class number relation, Proposition 7 of Yu [1995].

Proposition 4.1. *If $m \in A$ is monic, then*

$$\sum_{\substack{t \in A \\ \mu \in \mathbb{F}_q^\times/\mathbb{F}_q^{\times 2}}} H(t^2 - \mu m) = \sum_{d|m} \max(|d|, |m/d|) - \sum_{\substack{d|m \\ \deg d = 1/2 \deg m}} |m|^{-1/2} \frac{|m| - |m - d^2|}{q-1}, \quad (4)$$

where the sums on the right are over monic divisors and the sum on the left is over pairs (t, μ) such that $t^2 - \mu m$ is an imaginary discriminant. This is equivalent to the condition that either $t^2 - \mu m$ has odd degree or $t^2 - \mu m$ has a leading coefficient that is not a square in $\mathbb{F}_q$ [Rosen 2002, Chapter14].

We need one additional lemma regarding class numbers. Define the Kronecker symbol χ_f on the monic irreducible elements of A by

$$\chi_f(P) = \begin{cases} 1 & \text{if } P \text{ splits in } k(\sqrt{f}), \\ 0 & \text{if } P \text{ ramifies in } k(\sqrt{f}), \\ -1 & \text{otherwise,} \end{cases}$$

and extend χ_f to all monic polynomials in A by $\chi_f\left(\prod P_i^{e_i}\right) = \prod \chi_f(P_i)^{e_i}$. The following is Lemma 3 in [Yu 1995].

Lemma 1. For any square-free $f \in A$ and any $b \in A$,

$$\frac{h'(fb^2)}{h'(f)} = |f| \prod_{P|f} \left(1 - \frac{\chi_f(P)}{|P|}\right).$$

Corollary 1. Under the hypotheses of the lemma, $h'(fb^2)|h'(f)$.

We also prove a general proposition about polynomials.

Proposition 4.2. *Let $f_1, \dots, f_n \in A$ be monic polynomials of odd degree. There exists a monic irreducible polynomial $m \in A$ of odd degree such that each f_i for $1 \leq i \leq n$ is a quadratic nonresidue modulo m.*

Proof. Let $p_1, \dots, p_r$ be the monic irreducible polynomials of odd degree dividing any of the f_i, and let $l_1, \dots, l_s$ be the monic irreducibles of even degree dividing any of the f_i. By the multiplicativity of the Legendre symbol, it suffices to find a monic polynomial m such that $(p_u/m) = -1$ for each u and $(l_v/m) = 1$ for each v.

For each u with $1 \leq u \leq r$, let π_u be an irreducible polynomial such that $(\pi_u/p_u) = (-1)^{(q+1)/2}$. For $1 \leq v \leq s$, choose ν_v to be a monic irreducible such that $(\nu_v/l_v) = 1$. Such π_u and ν_v exist by Dirichlet's theorem on primes in arithmetic progressions. Applying this theorem again, choose m to be a monic irreducible of odd degree such that $m \equiv \pi_u \pmod{p_u}$ and $m \equiv \nu_v \pmod{l_v}$ for each choice of u and v.

We show that m satisfies the conclusion of the proposition. Applying quadratic reciprocity for function fields [Rosen 2002, Chapter 3], for each p_u we have

$$\left(\frac{p_u}{m}\right) = (-1)^{\frac{q-1}{2}\cdot\deg m\cdot\deg p_u}\cdot\left(\frac{m}{p_u}\right) = (-1)^{\frac{q-1}{2}}\cdot\left(\frac{\pi_u}{p_u}\right) = (-1)^{\frac{q-1}{2}}\cdot(-1)^{\frac{q+1}{2}} = -1.$$

Similarly, for the l_v, we have

$$\left(\frac{l_v}{m}\right) = (-1)^{\frac{q-1}{2}\cdot\deg m\cdot\deg l_v}\cdot\left(\frac{m}{l_v}\right) = 1\cdot\left(\frac{\nu_v}{l_v}\right) = 1.$$

Thus, m satisfies the conditions stated at the beginning of the proof and thus also the conclusion of the proposition. □

4.2. ***Proof of Theorem 1.3.*** Suppose that S is any finite set (possibly empty) of monic polynomials $f \in A$ of odd degree such that $c \nmid h(f)$. By Proposition 4.2, take m to be an irreducible monic polynomial of odd degree such that each $f \in S$ is a quadratic nonresidue modulo m (if $S = \varnothing$, take m to be any monic irreducible of odd degree). The class number relation Equation (4) gives us

$$\sum_{\substack{t\in A \\ \mu\in\mathbb{F}_q^\times/\mathbb{F}_q^{\times 2}}} H(t^2-\mu m) \;=\; \sum_{d|m} \max(|d|, |m/d|) \;=\; 2q^{\deg m}.$$

Since $c \nmid 2q^{\deg m}$, at least one of the terms on the left side of the equation is not divisible by c, we can take μ and t so that $H(t^2-\mu m)$ is not divisible by c. From the definition of the Hurwitz class number,

$$H(t^2-\mu m) \;=\; \sum_{n^2|m} h'\left(\frac{t^2-\mu m}{n^2}\right).$$

Since the left side of the equation is indivisible by c, we can choose n such that $h'\left(\frac{t^2-\mu m}{n^2}\right)$ is indivisible by c. We now write

$$\frac{t^2-\mu m}{n^2} \;=\; fb^2,$$

where f is square-free. From Corollary 1, we have that $h'(f)|h'(fb^2)$. In particular, $c \nmid h'(f)$. Furthermore, we have $h'(f) = h(f)/w(f)$. Since the prime at infinity is totally ramified in $k(\sqrt{f})$, the group of units of $A[\sqrt{f}]$ has rank 0 and thus is just $\mathbb{F}_q^\times$. This means that

$$w(f) \;=\; \frac{\#A[\sqrt{f}]^\times}{q-1} \;=\; 1.$$

So $h(f) = w(f)h'(f) = h'(f)$, whence $c \nmid h(f)$. This gives us an element $f \in A$ such that $h(f)$ is indivisible by c.

We show that $f \notin S$. We have that $f\cdot(bn)^2 = t^2-\mu m$. Reducing modulo m, we have $f \equiv (t/bn)^2 \pmod{m}$, so f is a quadratic residue modulo m. In particular, $f \notin S$. Thus, there are infinitely many quadratic imaginary discriminants f of odd degree such that c does not divide the class number of $K = k(\sqrt{f})$.

5. Examples

We consider first an example constructed by Theorem 1.1. Let $q = 3$ and $c = 17$, so that we aim to construct a quadratic imaginary discriminant $f \in \mathbb{F}_3[x]$ such that

$h(f)$ is divisible by 17. Take

$$\begin{aligned} f &= 2x^5 + 2x^4 + 1, \\ n &= x^7 + 2x^6 + x^5 + 2x^4 + x^3 + 2, \\ t &= x^6 + x^5 + 2x^3 + 1, \\ m &= x. \end{aligned}$$

Our choice of f is square-free (and, in fact, irreducible). The condition $c \deg m < p \deg f$ is clearly satisfied (note that since c is prime, we have $c = p$). We also have $(m, n) = 1$ and $m^{17} = n^2 - t^2 f$, so Theorem 1.1 says that $\mathrm{Cl}(f)$ has an element of order 17. Indeed, $h(f) = 17$. In fact, computation of a finite number of class numbers shows that there is no choice of f of smaller degree such that $17 | h(f)$.

We now provide an example of the method of the proof of Theorem 1.3. Let $q = 3$, and define $k = \mathbb{F}_q(x)$. Begin with the polynomials

$$f_1 = x + 2 \quad \text{and} \quad f_2 = x^3 + x^2 + 2x = x(x^2 + x + 2).$$

It can be computed that $h(f_1) = 1$ and $h(f_2) = 6$. We will use the method of the proof of Theorem 1.3 to find a third quadratic imaginary discriminant f_3 such that $h(f_3)$ is relatively prime to $c = 5$. Using the same notation as the proof of Proposition 4.2, we have $p_1 = x$ and $l_1 = x^2 + x + 2$. The method of the proof now calls for us to find irreducible polynomials π_1 and ν_1 such that

$$\left(\frac{\pi_1}{p_1}\right) = (-1)^{\frac{q+1}{2}} = 1 \quad \text{and} \quad \left(\frac{\nu_1}{l_1}\right) = 1.$$

It will thus suffice to take $\pi_1 \equiv 1 \pmod{p_1}$, and $\nu_1 \equiv 1 \pmod{l_1}$. Because the next step in the proof is to apply the Chinese Remainder Theorem, it is unnecessary (although a trivial exercise) to actually compute irreducible polynomials π_1 and ν_1. In the proof we use the existence of irreducible polynomials to apply quadratic reciprocity, but for the purpose of construction we need only find appropriate residue classes to apply the Chinese Remainder Theorem. We now need a monic irreducible polynomial m of odd degree such that

$$m \equiv \begin{cases} 1 & \pmod{p_1}, \\ 1 & \pmod{l_1}. \end{cases}$$

One such polynomial is $m = p_1 \cdot l_1 + 1 = x^3 + x^2 + 2x + 1$. By the class number relation Equation (4), we have

$$\sum_{\substack{t \in A \\ \mu \in \mathbb{F}_q^\times / \mathbb{F}_q^{\times 2}}} H(t^2 - \mu m) = \sum_{d | m} \max(|d|, |m/d|) = 54,$$

where the sum on the left is over all (μ, t) such that μ is either 1 or 2 and t has degree 0 or 1. Expanding the sum, we have

$$\begin{aligned} H(-m)+H(-2m)+2H(1-m)+2H(1-2m)+2H(x^2-m) \\ +2H(x^2-2m)+2H((x+1)^2-m)+2H((x+1)^2-2m) \\ +2H((x+2)^2-m)+2H((x+2)^2-2m)=54. \end{aligned}$$

Since $5 \nmid 54$, at least one of the Hurwitz class numbers on the left side of this equation is indivisible by 5. Although the first term, $H(-m)$ is 5, we find that the second term, $H(-2m)=H(m)$ is 3. Furthermore, since m is irreducible, we have by the definition of the Hurwitz class number that

$$H(-2m) \;=\; H(m) \;=\; \sum_{n^2|m} h'(m/n^2) \;=\; h'(m).$$

As discussed in the proof of Theorem 1.3, we have that $h(m)=h'(m)$, so $h(m)=3$. Thus choosing $f_3=m=x^3+x^2+2x+1$ gives a third polynomial f_3 with $5 \nmid h(f_3)$.

References

[Achter 2006] J. D. Achter, "The distribution of class groups of function fields", *J. Pure Appl. Algebra* **204**:2 (2006), 316–333. MR 2006h:11132

[Baker 1967] A. Baker, "Linear forms in the logarithms of algebraic numbers. I, II, III", *Mathematika 13* (1966)*, 204-216; ibid. 14* (1967)*, 102-107; ibid.* **14** (1967), 220–228. MR 36 #3732

[Baker 1971] A. Baker, "Imaginary quadratic fields with class number 2", *Ann. of Math.* (2) **94** (1971), 139–152. MR 45 #8631

[Cardon and Ram Murty 2001] D. A. Cardon and M. Ram Murty, "Exponents of class groups of quadratic function fields over finite fields", *Canad. Math. Bull.* **44**:4 (2001), 398–407. MR 2002g:11164

[Cohen and Lenstra 1984] H. Cohen and H. W. Lenstra, Jr., "Heuristics on class groups of number fields", pp. 33–62 in *Number theory, Noordwijkerhout 1983 (Noordwijkerhout, 1983)*, Lecture Notes in Math. **1068**, Springer, Berlin, 1984. MR 85j:11144

[Davenport and Heilbronn 1971] H. Davenport and H. Heilbronn, "On the density of discriminants of cubic fields. II", *Proc. Roy. Soc. London Ser. A* **322**:1551 (1971), 405–420. MR 58 #10816

[Friedman and Washington 1989] E. Friedman and L. C. Washington, "On the distribution of divisor class groups of curves over a finite field", pp. 227–239 in *Théorie des nombres (Quebec, PQ, 1987)*, de Gruyter, Berlin, 1989. MR 91e:11138

[Goldfeld 1976] D. M. Goldfeld, "The class number of quadratic fields and the conjectures of Birch and Swinnerton-Dyer", *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* (4) **3**:4 (1976), 624–663. MR 56 #8529

[Gross and Zagier 1983] B. Gross and D. Zagier, "Points de Heegner et dérivées de fonctions L", *C. R. Acad. Sci. Paris Sér. I Math.* **297**:2 (1983), 85–87. MR 85d:11062

[Hartung 1974] P. Hartung, "Proof of the existence of infinitely many imaginary quadratic fields whose class number is not divisible by 3", *J. Number Theory* **6** (1974), 276–278. MR 50 #4528

[Hayes 1999] D. R. Hayes, “Distribution of minimal ideals in imaginary quadratic function fields”, pp. 25–30 in *Applications of curves over finite fields (Seattle, WA, 1997)*, Contemp. Math. **245**, Amer. Math. Soc., Providence, RI, 1999. MR 2001a:11188

[Heegner 1952] K. Heegner, “Diophantische Analysis und Modulfunktionen”, *Math. Z.* **56** (1952), 227–253. MR 14,725j

[Kohnen and Ono 1999] W. Kohnen and K. Ono, “Indivisibility of class numbers of imaginary quadratic fields and orders of Tate-Shafarevich groups of elliptic curves with complex multiplication”, *Invent. Math.* **135**:2 (1999), 387–398. MR 2000c:11087

[Rosen 2002] M. Rosen, *Number theory in function fields*, Graduate Texts in Mathematics **210**, Springer, New York, 2002. MR 2003d:11171

[Soundararajan 2000] K. Soundararajan, “Divisibility of class numbers of imaginary quadratic fields”, *J. London Math. Soc.* (2) **61**:3 (2000), 681–690. MR 2001i:11128

[Stark 1967] H. M. Stark, “A complete determination of the complex quadratic fields of class-number one”, *Michigan Math. J.* **14** (1967), 1–27. MR 36 #5102

[Stark 1975] H. M. Stark, “On complex quadratic fields wth class-number two”, *Math. Comp.* **29** (1975), 289–302. Collection of articles dedicated to Derrick Henry Lehmer on the occasion of his seventieth birthday. MR 51 #5548

[Yu 1995] J.-K. Yu, “A class number relation over function fields”, *J. Number Theory* **54**:2 (1995), 318–340. MR 96i:11128

Received: 2007-08-03 Revised: 2007-10-28 Accepted: 2007-10-28

Adam_Merberg@brown.edu *Department of Mathematics, Brown University,*
151 Thayer Street, Providence, RI 02912, United States

Patch and crossover planar dyadic wavelet sets

A. J. Hergenroeder, Zachary Catlin, Brandon George and David R. Larson

(Communicated by Kenneth S. Berenhaut)

A single dyadic orthonormal wavelet on the plane $\mathbb{R}^2$ is a measurable square integrable function $\psi(x, y)$ whose images under translation along the coordinate axes followed by dilation by positive and negative integral powers of 2 generate an orthonormal basis for $\mathscr{L}^2(\mathbb{R}^2)$. A planar dyadic wavelet set E is a measurable subset of $\mathbb{R}^2$ with the property that the inverse Fourier transform of the normalized characteristic function $\frac{1}{2\pi}\chi(E)$ of E is a single dyadic orthonormal wavelet. While constructive characterizations are known, no algorithm is known for constructing all of them. The purpose of this paper is to construct two new distinct uncountably infinite families of dyadic orthonormal wavelet sets in $\mathbb{R}^2$. We call these the crossover and patch families. Concrete algorithms are given for both constructions.

Introduction

Wavelet theory is interesting to mathematicians both for its applications to signal analysis and image analysis and also because of the rich mathematical structure underlying the theory of wavelets. Wavelet sets are measurable sets whose normalized characteristic functions are the Fourier transforms of wavelets. Planar dyadic wavelet sets are interesting mathematically because they are fractal-like, and there are hands-on methods for working with them and constructing new examples. They are also interesting because while constructive characterizations are known, no algorithm is known for constructing all planar dyadic wavelets. There are open problems associated with their classification. Algorithms for constructing new examples or classes of examples can provide useful counterexamples to conjectures as well as be appreciated for their intrinsic mathematical beauty.

A single dyadic *orthonormal wavelet* on the plane $\mathbb{R}^2$ is a (Lebesgue) measurable square-integrable function $\psi(x, y)$ whose translations along the coordinate axes followed by dilations by positive and negative integral powers of 2 generate an orthonormal basis for $L^2(\mathbb{R}^2)$. A planar dyadic *wavelet* set E is a measurable

MSC2000: 47A13, 42C40, 42C15.

Keywords: wavelet, wavelet set, patch, crossover, congruence.

D. L. was partially supported by a grant from the National Science Foundation.

subset of $\mathbb{R}^2$ with the property that the inverse Fourier transform of the normalized characteristic function $1/2\pi\chi(E)$ of E is a single dyadic orthonormal wavelet. As usual, the Fourier transform on $\mathscr{L}^2(\mathbb{R}^2)$ is the tensor product of two copies of the Fourier transform on $\mathscr{L}^2(\mathbb{R})$. In this paper we discuss two algorithms which generate two distinct uncountably infinite classes of dyadic orthonormal wavelet sets in $\mathbb{R}^2$. We denote these classes the *crossover* and *patch* classes and denote the algorithms for these constructions the crossover and patch algorithms. A free parameter in both of the algorithms is a partition of the so-called inner square, $[-\pi/2, \pi/2) \times [-\pi/2, \pi/2)$, into four measurable subsets $X_\ominus$, $X_\oplus$, $Y_\ominus$, $Y_\oplus$, with the property that $X_\ominus$ is contained in the left half of the inner square, $X_\oplus$ is contained in the right half, $Y_\ominus$ is contained in the bottom half, and $Y_\oplus$ is contained in the top half. Notice that if the boundary of any two of these sets is the same, and if this boundary has Lebesgue measure 0, then these two sets are still essentially disjoint although their boundaries are the same.

Wavelets for dilations other than 2 (the dyadic case) on the line $\mathbb{R}$ and in $\mathbb{R}^n$ have been investigated by many authors. In higher dimensions both scalar dilations and matrix dilations have been studied. However, much of the interesting work in the literature has been for the dyadic case, which is the case we focus on.

For completeness, we give the form used for abstract matrix dilations: A dilation A wavelet is a function on $\mathbb{R}^n$ whose dilations by integral powers of A and translations along the coordinate axes (or, more generally, translations along a full-rank lattice) form an orthonormal basis for the space of all square-integrable measurable functions over $\mathbb{R}^n$ with respect to Lebesgue measure. In precise terms, a function f on $\mathbb{R}^n$ is a dilation A wavelet if and only if it is measurable with respect to product Lebesgue measure, and

$$\{|\det A|^{\frac{m}{2}} f\left(A^m t - l\right) : m \in \mathbb{Z}, l \in \mathbb{Z}^n\}$$

is an orthonormal basis of $L^2(\mathbb{R}^n)$. A dilation A wavelet set is a measurable set W for which the inverse Fourier transform of the normalized characteristic function is a dilation A orthonormal wavelet. The dyadic case is where $A := 2I_2$, where I_2 is the identity matrix in two dimensions.

Existence of wavelet sets in $\mathbb{R}^n$ was first demonstrated in 1994 [Dai et al. 1997]. The proof used an algorithmic approach which generated wavelet sets that were unbounded and had 0 as a limit point, rendering them difficult to visualize [Dai et al. 1997; Zhang and Larson $\geq$ 2008]. It showed that there are uncountably many such sets in $\mathbb{R}^2$ for many matrix dilations, including the dyadic case. Subsequently, several authors [Soardi and Weiland 1998; Dai et al. 1998; Benedetto and Leon 1999; 2001; Baggett et al. 1999; Gu and Han 2000] constructed wavelet sets in $\mathbb{R}^2$ which were more easily pictured due to their qualities of being bounded and bounded away from 0, and had other interesting structural properties. Two such sets were included in the final remarks section of [Dai and Larson 1998]. Recent papers that construct new planar wavelet sets with reasonable graphics and interesting properties can be found in [Zhang and Larson ≥ 2008] and [Merrill ≥ 2008]. A brief history of wavelet sets can be found in [Zhang and Larson ≥ 2008, Section 5].

In the summer of 2007, the first three authors were undergraduate student participants in the Texas A&M Mathematics REU on *matrix analysis and wavelets* (funded by the NSF), which was mentored by D. Larson. They set out to classify multiple categories of wavelet sets in $\mathbb{R}^2$ using an algorithmic approach. The present paper is the upshot of that project. Two algorithms were obtained. The wavelet sets in Figures 1 and 2 are called crossover wavelet sets because in their generation, regions are added to or subtracted from alternating sides of the inner square. Alternatively, patch wavelet sets are created by adding regions to the same side of the square for each translation; see, for example, the set in Figure 3. This category of sets is so named because in computer networking a patch cable is the opposite of a crossover cable.

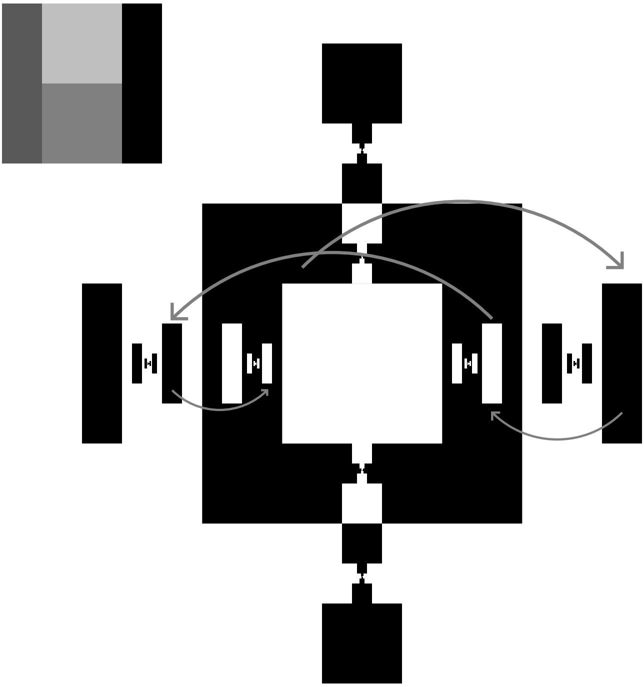


Figure 1. The two-dimensional wavelet set formed in Example 1.

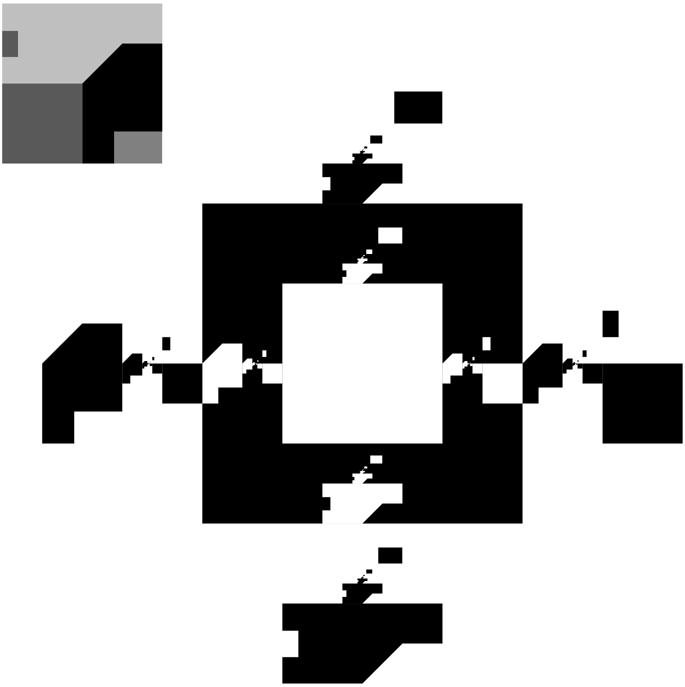


Figure 2. An arbitrary (conforming) partition, with wavelet set.

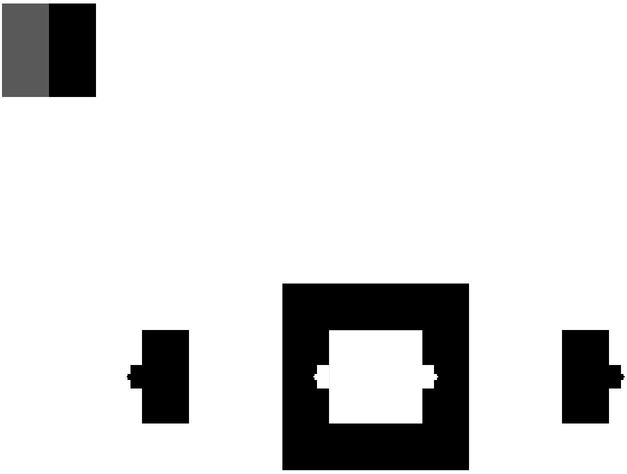


Figure 3. A patch wavelet set: the *wedding cake set*.

1. Preliminaries

We begin with some basic formal definitions.

A *single dyadic orthonormal wavelet* is a function $\psi \in \mathcal{L}^2(\mathbb{R})$ (Lebesgue measure) with the property that the set $\{ 2^{\frac{n}{2}} \psi(2^n \cdot -l) \mid n, l \in \mathbb{Z} \}$ forms an orthonormal basis for $\mathcal{L}^2(\mathbb{R})$ [Consortium 1998]; see also [Larson 2007a, pp. 6–7]. More generally, if A is any real invertible $n \times n$ matrix, then a single function $\psi \in L^2(\mathbb{R}^n)$ is a multivariate orthonormal wavelet for A if

$$\{| \det A|^{\frac{n}{2}} \psi(A^n \cdot -l) \mid n \in \mathbb{Z}, l \in \mathbb{Z}^{(n)}\}$$

is an orthonormal basis of $L^2(\mathbb{R}^n)$. It was shown in [Dai et al. 1997] that if A is *expansive* (equivalently, all eigenvalues of A are required to have absolute value strictly greater than 1) then orthonormal wavelets for A always exist. The dyadic case is the case where $A := 2I_2$ (two times the identity matrix on R^n). This is the simplest (and most investigated) case.

We let $\mathcal{F}$ denote the n-dimensional Fourier transform on $\mathcal{L}^2(\mathbb{R}^n)$ defined by

$$(\mathcal{F}f)(s) := \frac{1}{(2\pi)^{\frac{n}{2}}} \int_{\mathbb{R}^n} e^{-s \circ t} f(t) dm,$$

for all $f \in L^2(\mathbb{R}^n)$. Here, $s \circ t$ denotes the real inner product. A measurable set $E \subseteq \mathbb{R}^n$ is defined to be a wavelet set for a dilation matrix A if

$$\mathcal{F}^{-1}(\frac{1}{\sqrt{\mu(E)}} \chi_E)$$

is an orthonormal wavelet for A, where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform.

In this paper, we will not explicitly use properties of $\mathcal{F}$ and $\mathcal{F}^{-1}$; however we state the formal definition of Fourier transform because it is needed to give the proper definition of a wavelet set. The Fourier transform is a unitary transform from $\mathcal{L}(\mathbb{R}^n)$, where $\mathbb{R}^n$ is usually considered as a multivariate time domain, to another copy of $\mathcal{L}(\mathbb{R}^n)$, where $\mathbb{R}^n$ is considered a multivariate frequency domain. We will do our work with wavelet sets entirely in the frequency domain. We can do this because there is a set-theoretic characterization of wavelet sets, Proposition 1.1, which allows one to construct and otherwise work with wavelet sets without using the Fourier transform.

A sequence of measurable subsets $\{E_n\}$ of a measurable set E is called a *measurable partition* of E if the relative complement of $\bigcup_n E_n$ in E is a null set (that is, has measure zero) and $E_n \cap E_m$ is a null set whenever $n \neq m$.

Measurable subsets E and F of $\mathbb{R}$ are called *2π-translation congruent* to each other, denoted by $E \sim_{2\pi} F$, if there exists a measurable partition $\{E_n\}$ of E, such that $\{E_n + 2n\pi\}$ is a measurable partition of F. Similarly, E and F are called *2-dilation congruent* to each other, denoted by $E_2\!\sim F$, if there is a measurable partition $\{E_n\}$ of E, such that $\{2^n E_n\}$ is a measurable partition of F. A measurable set E is called a *2π-translation generator of a measurable partition* of $\mathbb{R}$ if $\{E+2n\pi\}_{n\in\mathbb{Z}}$ forms a measurable partition of $\mathbb{R}$. Similarly, a measurable set F is called a *2-dilation generator of a measurable partition of* $\mathbb{R}$ if $\{2^n F\}_{n\in\mathbb{Z}}$ forms a measurable partition of $\mathbb{R}$.

Lemma 4.3 in [Dai and Larson 1998] gives the following characterization of dyadic wavelet sets in $\mathbb{R}$, which was also obtained independently in [Fang and Wang 1996] using different techniques. *Let $E \subseteq \mathbb{R}$ be a measurable set. Then E is a dyadic wavelet set if and only if E is both a 2π-translation generator of a measurable partition of $\mathbb{R}$ and a 2-dilation generator of a measurable partition of $\mathbb{R}$.*

In [Dai et al. 1997], this criterion was generalized to the multivariate case. We will consider only the dyadic planar case in this paper because we will only use the criterion for that case, although the criterion actually applies for the arbitrary expansive case [Dai et al. 1997; 1998]. So from [Dai et al. 1997] we have that E is a dyadic wavelet set in $\mathbb{R}^n$ if and only if E is both a 2π-translation generator of a measurable partition of $\mathbb{R}^n$ and a 2-dilation generator of a measurable partition of $\mathbb{R}^n$. Here, to achieve a translation partition one translates by all integral multiples of 2π separately in each coordinate direction. To achieve a dilation partition, one dilates by all integral powers of 2 simultaneously in all coordinates. For example, it is clear that the set $[-\pi,\pi)\times[-\pi,\pi)$ is a 2π-translation generator of a measurable partition of $\mathbb{R}^2$, and

$$G_{TO}\backslash\left(\left[-\frac{\pi}{2},\frac{\pi}{2}\right)\times\left[-\frac{\pi}{2},\frac{\pi}{2}\right)\right)$$

is a 2-dilation generator of a measurable partition of $\mathbb{R}^2$.

Properly generalizing the one dimensional definition to the planar case, we say that two Lebesgue measurable sets $A, B \subset \mathbb{R}^2$ are 2π-translation congruent if there is a measurable partition $\{A_{k,l} : k,l \in \mathbb{Z}\}$ of A such that

$$\{A_{k,l}+\begin{bmatrix}2k\pi\\2l\pi\end{bmatrix} : k,l\in\mathbb{Z}\}$$

is a measurable partition of B, and they are 2-dilation congruent if there is a measurable partition $\{A_n : n\in\mathbb{Z}\}$ of A such that

$$\{2^n A_n : n\in\mathbb{Z}\}$$

is a measurable partition of B.

Translation congruence and dilation congruence are both equivalence relations on the class of all measurable subsets. If a set A is 2π-translation congruent to a 2π-translation generator of a measurable partition of $\mathbb{R}^2$, it is clear that A itself is a 2π-translation generator of a measurable partition of $\mathbb{R}^2$. Moreover, sets A

and B which are both 2π-translation generators of measurable partitions of $\mathbb{R}^2$ are necessarily translation congruent to each other. All this is in [Dai et al. 1997], and other expositions can be found in [Dai and Larson 1998; Larson 2007b; Zhang and Larson $\geq$ 2008]. This yields a useful criterion.

In the following proposition (and in the rest of the paper), let

$$G_{TO} := [-\pi,\pi)\times[-\pi,\pi)\,, \qquad \text{and} \qquad G_{SO} := G_{TO}\backslash\Big(\Big[-\frac{\pi}{2},\frac{\pi}{2}\Big)\times\Big[-\frac{\pi}{2},\frac{\pi}{2}\Big)\Big).$$

Proposition 1.1 (Working Principle Criterion). *A measurable set $W \subseteq \mathbb{R}^2$ is a dyadic wavelet set if and only if W is 2π-translation congruent to G_{TO} and 2-dilation congruent to G_{SO}.*

2. The crossover algorithm

We first consider a special case of a wavelet set to illustrate the *crossover algorithm*. We will then generalize this to obtain Theorem 2.1.

Example 2.1. Let

$$X_\ominus = \Big[-\frac{\pi}{2},-\frac{\pi}{4}\Big)\times\Big[-\frac{\pi}{2},\frac{\pi}{2}\Big), \qquad X_\oplus = \Big[\frac{\pi}{4},\frac{\pi}{2}\Big)\times\Big[-\frac{\pi}{2},\frac{\pi}{2}\Big),$$
$$Y_\ominus = \Big[-\frac{\pi}{4},\frac{\pi}{4}\Big)\times\Big[-\frac{\pi}{2},0\Big), \qquad Y_\oplus = \Big[-\frac{\pi}{4},\frac{\pi}{4}\Big)\times\Big[0,\frac{\pi}{2}\Big).$$

We can generate a wavelet set in the plane using the above partition of the inner square using an algorithm (the crossover algorithm) which we will illustrate with the following example.

Let $X_{\ominus 1} := X_\ominus$. Start by adding the vector $\begin{bmatrix}2\pi\\0\end{bmatrix}$ to the set $X_\ominus$, translating it (that is, crossing it over) to the right half-plane. The set formed is

$$\Big[\frac{3\pi}{2},\frac{7\pi}{4}\Big)\times\Big[-\frac{\pi}{2},\frac{\pi}{2}\Big).$$

Call this $X_{\ominus 2}$. Secondly, scale $X_{\ominus 2}$ by $\frac{1}{2}$ to obtain

$$\Big[\frac{3\pi}{4},\frac{7\pi}{8}\Big)\times\Big[-\frac{\pi}{4},\frac{\pi}{4}\Big).$$

Call this $X_{\ominus 3}$. Thirdly, translate $X_{\ominus 3}$ to the opposite half-plane by adding $\begin{bmatrix}-2\pi\\0\end{bmatrix}$ to obtain

$$\Big[-\frac{5\pi}{4},-\frac{9\pi}{8}\Big)\times\Big[-\frac{\pi}{4},\frac{\pi}{4}\Big),$$

and call this set $X_{\ominus 4}$. Notice that $X_{\ominus 4}$ is on the same side half-plane as $X_{\ominus}$. Finally, scale $X_{\ominus 4}$ by $\frac{1}{2}$ to form the set

$$\left[-\frac{5\pi}{8}, -\frac{9\pi}{16}\right) \times \left[-\frac{\pi}{8}, \frac{\pi}{8}\right)$$

and call it $X_{\ominus 5}$. Continue these four steps inductively for $X_{\ominus}$.

We perform four similar steps on the set $X_{\oplus}$; however, we translate by $\begin{bmatrix} -2\pi \\ 0 \end{bmatrix}$ for the first step (instead of $\begin{bmatrix} 2\pi \\ 0 \end{bmatrix}$) and translate by $\begin{bmatrix} 2\pi \\ 0 \end{bmatrix}$ for the third step (instead of $\begin{bmatrix} -2\pi \\ 0 \end{bmatrix}$). We obtain the following from the first four steps:

$$X_{\oplus 2} = \left[-\frac{7\pi}{4}, -\frac{3\pi}{2}\right) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right), \qquad X_{\oplus 3} = \left[-\frac{7\pi}{8}, -\frac{3\pi}{4}\right) \times \left[-\frac{\pi}{4}, \frac{\pi}{4}\right),$$
$$X_{\oplus 4} = \left[\frac{9\pi}{8}, \frac{5\pi}{4}\right) \times \left[-\frac{\pi}{4}, \frac{\pi}{4}\right), \qquad X_{\oplus 5} = \left[\frac{9\pi}{16}, \frac{5\pi}{8}\right) \times \left[-\frac{\pi}{8}, \frac{\pi}{8}\right).$$

Continue this process inductively for $X_{\oplus}$ as well. Perform similar steps for $Y_{\oplus}$ and $Y_{\ominus}$, translating by $\begin{bmatrix} 0 \\ \pm 2\pi \end{bmatrix}$ instead of $\begin{bmatrix} \pm 2\pi \\ 0 \end{bmatrix}$, beginning with a translation of $\begin{bmatrix} 0 \\ 2\pi \end{bmatrix}$ for $Y_{\ominus}$ and a translation of $\begin{bmatrix} 0 \\ -2\pi \end{bmatrix}$ for $Y_{\oplus}$.

Let W be the set

$$\begin{aligned}
&\Big(\bigcup_{i=1}^{\infty}[X_{\ominus 2i} \cup X_{\oplus 2i} \cup Y_{\ominus 2i} \cup Y_{\oplus 2i}]\Big) \\
&\qquad\qquad \cup\Big(G_{TO} \setminus \Big[\bigcup_{i=1}^{\infty}[X_{\ominus 2i-1} \cup X_{\oplus 2i-1} \cup Y_{\ominus 2i-1} \cup Y_{\oplus 2i-1}]\Big]\Big) \\
&= \Big(\bigcup_{i=1}^{\infty}[X_{\ominus 2i} \cup X_{\oplus 2i} \cup Y_{\ominus 2i} \cup Y_{\oplus 2i}]\Big) \\
&\qquad\qquad \cup\Big(G_{SO} \setminus \Big[\bigcup_{i=2}^{\infty}[X_{\ominus 2i-1} \cup X_{\oplus 2i-1} \cup Y_{\ominus 2i-1} \cup Y_{\oplus 2i-1}]\Big]\Big).
\end{aligned}$$

We can think of W as being the union of G_{TO} and the exterior *black pieces* of the form $X_{\oplus 2n}$, $X_{\ominus 2n}$, $Y_{\oplus 2n}$, $Y_{\ominus 2n}$, with *white spaces* of the form $X_{\oplus 2n+1}$, $X_{\ominus 2n+1}$, $Y_{\oplus 2n+1}$, $Y_{\ominus 2n+1}$ removed from G_{TO}.

This set W (see Figure 1) is a wavelet set. To see this, let

$$G(X_{\ominus \text{odd}}) := \bigcup_{i=1}^{\infty} X_{\ominus 2i-1}, \qquad \text{and} \qquad G(X_{\ominus \text{even}}) := \bigcup_{i=1}^{\infty} X_{\ominus 2i}.$$

Observe that

$$G(X_{\ominus\mathrm{odd}})\backslash X_{\ominus} \subset G_{SO}, \qquad \text{and} \qquad G(X_{\ominus\mathrm{even}}) \not\subseteq G_{SO}.$$

Similarly, define sets for $X_{\oplus}$, $Y_{\ominus}$, and $Y_{\oplus}$ with analogous characteristics. Observe that W is translation congruent to G_{TO} modulo 2π because

$$\Big[\Big(\bigcup_{i=0}^{\infty} X_{\ominus 4i+2}\Big) \cup \Big(\bigcup_{i=1}^{\infty} X_{\oplus 4i}\Big)\Big] - \begin{bmatrix} 2\pi \\ 0 \end{bmatrix} = \Big[\Big(\bigcup_{i=0}^{\infty} X_{\ominus 4i+1}\Big) \cup \Big(\bigcup_{i=0}^{\infty} X_{\oplus 4i+3}\Big)\Big],$$

$$\Big[\Big(\bigcup_{i=0}^{\infty} X_{\oplus 4i+2}\Big) \cup \Big(\bigcup_{i=1}^{\infty} X_{\ominus 4i}\Big)\Big] + \begin{bmatrix} 2\pi \\ 0 \end{bmatrix} = \Big[\Big(\bigcup_{i=0}^{\infty} X_{\oplus 4i+1}\Big) \cup \Big(\bigcup_{i=0}^{\infty} X_{\ominus 4i+3}\Big)\Big],$$

and

$$\Big[\Big(\bigcup_{i=0}^{\infty} X_{\oplus 4i+1}\Big) \cup \Big(\bigcup_{i=0}^{\infty} X_{\ominus 4i+3}\Big)\Big] \cup \Big[\Big(\bigcup_{i=0}^{\infty} X_{\ominus 4i+1}\Big) \cup \Big(\bigcup_{i=0}^{\infty} X_{\oplus 4i+3}\Big)\Big] = G(X_{\ominus\mathrm{odd}}) \cup G(X_{\oplus\mathrm{odd}}),$$

so that all gaps in the set G_{TO} due to the crossover algorithm applied to the sets $X_{\ominus 1}$ and $X_{\oplus 1}$ are filled when we translate the black sets formed by the crossover algorithm applied to the sets $X_{\ominus 1}$ and $X_{\oplus 1}$ by multiples of $\begin{bmatrix} 2\pi \\ 0 \end{bmatrix}$. Similar results will apply for $Y_{\ominus}$ and $Y_{\oplus}$.

Moreover, W is dilation congruent to G_{SO} because

$$\tfrac{1}{2}G(X_{\ominus\mathrm{even}}) = G(X_{\ominus\mathrm{odd}}) \in G_{SO}$$

(that is, the even pieces of the form $X_{\ominus n}$ scale into the odd pieces of the form $X_{\ominus n}$), with similar results for $G(X_{\oplus\mathrm{even}})$, $G(Y_{\ominus\mathrm{even}})$, and $G(Y_{\oplus\mathrm{even}})$.

The set of steps indicated above, applied to all four pieces of the partition of the inner square, is a special case of the crossover algorithm. An uncountably infinite family of wavelet sets in $\mathbb{R}^2$ can be similarly constructed using a natural generalization of this algorithm. The generalized crossover algorithm will be presented rigorously in the later sections of this paper in the context of the proof of Theorem 2.1 and the constructions involved in the proof.

A brief description of the general crossover algorithm is the following:

(i) Partition the inner square into a maximum of four subsets satisfying the conditions given in the statement of Theorem 2.1 below. (These conditions are necessary because not all partitions of the inner square will lead to a wavelet set in this way.)

(ii) Translate one piece of the partition by $\begin{bmatrix} \pm 2\pi \\ 0 \end{bmatrix}$ or $\begin{bmatrix} 0 \\ \pm 2\pi \end{bmatrix}$, moving it out of the inner square to the opposite side of the x- or y-axis.

(iii) Dilate the set formed in step 2 into G_{SO} by $\frac{1}{2}$.

(iv) Translate the set formed in step 3 in the opposite direction (compared to the first translation), that is, by $\begin{bmatrix} \mp 2\pi \\ 0 \end{bmatrix}$ or $\begin{bmatrix} 0 \\ \mp 2\pi \end{bmatrix}$.

(v) Dilate the set formed in step 4 into G_{SO} by $\frac{1}{2}$.

(vi) Repeat these steps inductively for this piece, and perform steps 2–5 on the other pieces of the partition inductively as well.

Theorem 2.1 (Crossover Algorithm). *Let $\{X_\ominus, X_\oplus, Y_\ominus, Y_\oplus\}$ be a partition of the set*

$$\left[-\frac{\pi}{2}, \frac{\pi}{2}\right) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right),$$

such that $X_\ominus$ is contained in the left half of the inner square (that is, the maximum possible x-coordinate of any point in the set $X_\ominus$ is 0), $X_\oplus$ is contained in the right half of the inner square (that is, the minimum possible x-coordinate of any point in the set $X_\oplus$ is 0), $Y_\ominus$ is contained in the bottom half of the inner square (that is, the maximum possible y-coordinate of the set $Y_\ominus$ is 0), and $Y_\oplus$ is contained in the top half of the inner square (that is, the minimum possible y-coordinate of the set $Y_\oplus$ is 0). Then the set W generated by this partition, under translation by

$$\begin{bmatrix} \pm 2\pi \\ 0 \end{bmatrix} \quad \textit{and} \quad \begin{bmatrix} 0 \\ \pm 2\pi \end{bmatrix}$$

and dilation by powers of 2 using steps (1)–(6) above, defined as

$$\left[\left(\bigcup_{i=1}^{\infty}\left[X_{\ominus 2i} \cup X_{\oplus 2i} \cup Y_{\ominus 2i} \cup Y_{\oplus 2i}\right]\right) \cup G_{TO}\right] \setminus \left[\bigcup_{i=1}^{\infty}[X_{\ominus 2i-1} \cup X_{\oplus 2i-1} \cup Y_{\ominus 2i-1} \cup Y_{\oplus 2i-1}]\right],$$

is a dyadic wavelet set in $\mathbb{R}^2$.

Remark 2.1. If either both $X_\oplus$ and $X_\ominus$ are defined, or both $Y_\oplus$ and $Y_\ominus$ are defined, then the other two sets are automatically determined due to our constraints.

3. Expressions for $X_{\ominus n}$, $X_{\oplus n}$, $Y_{\ominus n}$, and $Y_{\oplus n}$

Before proving our main result, Theorem 2.1, we first give rigorous expressions for the sets $X_{\ominus n}$, $X_{\oplus n}$, $Y_{\ominus n}$, and $Y_{\oplus n}$. We begin with the sets of the form $X_{\ominus n}$. Suppose first that n is odd and $n \geq 3$. We can derive the formula for $X_{\ominus n}$ in terms of n by looking at the first few terms. Let $X_{\ominus 1} := X_\ominus$, which has the above constraints listed according to the theorem. We can then find the next few odd terms using the crossover algorithm described in the example:

$$X_{\ominus 3} = \frac{1}{2}\left(X_{\ominus} + \begin{bmatrix} 2\pi \\ 0 \end{bmatrix}\right),$$

$$X_{\ominus 5} = \frac{1}{2}\left(X_{\ominus 3} - \begin{bmatrix} 2\pi \\ 0 \end{bmatrix}\right) = \left(\frac{X_{\ominus}}{4} + \frac{1}{4}\begin{bmatrix} 2\pi \\ 0 \end{bmatrix} - \frac{1}{2}\begin{bmatrix} 2\pi \\ 0 \end{bmatrix}\right),$$

$$X_{\ominus 7} = \frac{1}{2}\left(X_{\ominus 5} + \begin{bmatrix} 2\pi \\ 0 \end{bmatrix}\right) = \left(\frac{X_{\ominus}}{8} + \frac{1}{8}\begin{bmatrix} 2\pi \\ 0 \end{bmatrix} - \frac{1}{4}\begin{bmatrix} 2\pi \\ 0 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} 2\pi \\ 0 \end{bmatrix}\right),$$

and, in general, we find that $X_{\ominus n+2} = \frac{1}{2}\left(X_{\ominus n} + (-1)^{\frac{n-1}{2}}\begin{bmatrix} 2\pi \\ 0 \end{bmatrix}\right)$. Solving this recurrence relation, we find that

$$\begin{aligned} X_{\ominus n} &= \frac{X_{\ominus}}{2^{\frac{n-1}{2}}} + \begin{bmatrix} 2\pi \\ 0 \end{bmatrix}\Big(\frac{1}{2^{\frac{n-1}{2}}} - \ldots + \frac{(-1)^{\frac{n-3}{2}}}{2}\Big) \\ &= \frac{X_{\ominus}}{2^{\frac{n-1}{2}}} + \begin{bmatrix} 2\pi \\ 0 \end{bmatrix}(-1)^{\frac{n-3}{2}}\Big(\frac{(-1)^{\frac{n-3}{2}}}{2^{\frac{n-1}{2}}} + \frac{(-1)^{\frac{n-3}{2}-1}}{2^{\frac{n-1}{2}-1}} + \ldots - \frac{1}{4} + \frac{1}{2}\Big) \\ &= \frac{X_{\ominus}}{2^{\frac{n-1}{2}}} + \begin{bmatrix} 2\pi \\ 0 \end{bmatrix}(-1)^{\frac{n-3}{2}}\sum_{i=1}^{\frac{n-1}{2}}\frac{(-1)^{i-1}}{2^i} \\ &= \frac{X_{\ominus}}{2^{\frac{n-1}{2}}} + \begin{bmatrix} 2\pi \\ 0 \end{bmatrix}\frac{1}{3}(-1)^{\frac{n-3}{2}}\Big(1 - (-\frac{1}{2})^{\frac{n-1}{2}}\Big), \end{aligned} \tag{1}$$

using the formula for a geometric series summation. In order to formally verify that our formula for $X_{\ominus n}$ solves the recurrence relation, we merely plug in the expressions for $X_{\ominus n+2}$ and $X_{\ominus n}$ and carry out basic computations.

Now suppose n' is even and $n' > 2$. We can derive the formula for $X_{\ominus n'}$ in terms of n' similarly:

$$X_{\ominus 2} = X_{\ominus} + \begin{bmatrix} 2\pi \\ 0 \end{bmatrix},$$

$$X_{\ominus 4} = \frac{X_{\ominus 2}}{2} - \begin{bmatrix} 2\pi \\ 0 \end{bmatrix} = \frac{X_{\ominus}}{2} + \frac{1}{2}\begin{bmatrix} 2\pi \\ 0 \end{bmatrix} - \begin{bmatrix} 2\pi \\ 0 \end{bmatrix},$$

$$X_{\ominus 6} = \frac{X_{\ominus 4}}{2} + \begin{bmatrix} 2\pi \\ 0 \end{bmatrix} = \frac{X_{\ominus}}{4} + \frac{1}{4}\begin{bmatrix} 2\pi \\ 0 \end{bmatrix} - \frac{1}{2}\begin{bmatrix} 2\pi \\ 0 \end{bmatrix} + \begin{bmatrix} 2\pi \\ 0 \end{bmatrix},$$

$$X_{\ominus 8} = \frac{X_{\ominus 6}}{2} - \begin{bmatrix} 2\pi \\ 0 \end{bmatrix} = \frac{X_{\ominus}}{8} + \frac{1}{8}\begin{bmatrix} 2\pi \\ 0 \end{bmatrix} - \frac{1}{4}\begin{bmatrix} 2\pi \\ 0 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} 2\pi \\ 0 \end{bmatrix} - \begin{bmatrix} 2\pi \\ 0 \end{bmatrix},$$

and, in general, we find that

$$X_{\ominus n'+2} = \frac{1}{2} X_{\ominus n'} + (-1)^{\frac{n'}{2}} \begin{bmatrix} 2\pi \\ 0 \end{bmatrix}.$$

Observe that $\frac{X_{\ominus n'}}{2} = X_{\ominus n'+1}$ is, in general, true based on our construction using the crossover algorithm since n' is even. Since $n'+1$ is odd, we plug into our formula the values for $X_{\ominus n}$ (see the bottom of page 69), where n is odd, to find $X_{\ominus n'}$.

$$\begin{aligned} X_{\ominus n'} = 2X_{\ominus n'+1} &= 2\left[\frac{X_\ominus}{2^{\frac{n'}{2}}} + \frac{1}{3} \begin{bmatrix} 2\pi \\ 0 \end{bmatrix} (-1)^{\frac{n'-2}{2}} \left(1 - \left(-\frac{1}{2}\right)^{\frac{n'}{2}}\right)\right] \\ &= \left[\frac{X_\ominus}{2^{\frac{n'-2}{2}}} + \frac{2}{3} \begin{bmatrix} 2\pi \\ 0 \end{bmatrix} (-1)^{\frac{n'-2}{2}} \left(1 - \left(-\frac{1}{2}\right)^{\frac{n'}{2}}\right)\right]. \end{aligned}$$

Once again, the proof that the recurrence relation is satisfied involves plugging in the expressions for $X_{\ominus n'+2}$ and $X_{\ominus n'}$ and performing basic computations.

Notice that when n' is even,

$$X_{\ominus n'} = X_{\ominus n'-1} + (-1)^{\frac{n'-2}{2}} \begin{bmatrix} 2\pi \\ 0 \end{bmatrix},$$

consistent with the crossover algorithm, because of the following:

$$\begin{aligned} X_{\ominus n'} &= X_{\ominus n'-1} + (-1)^{\frac{n'-2}{2}} \begin{bmatrix} 2\pi \\ 0 \end{bmatrix} \\ &= \frac{X_\ominus}{2^{\frac{n'-2}{2}}} + \frac{1}{3} \begin{bmatrix} 2\pi \\ 0 \end{bmatrix} (-1)^{\frac{n'-4}{2}} \left(1 - \left(-\frac{1}{2}\right)^{\frac{n'-2}{2}}\right) + (-1)^{\frac{n'-2}{2}} \begin{bmatrix} 2\pi \\ 0 \end{bmatrix} \\ &= \frac{X_\ominus}{2^{\frac{n'-2}{2}}} + \frac{2}{3} \begin{bmatrix} 2\pi \\ 0 \end{bmatrix} (-1)^{\frac{n'-2}{2}} \left(-\frac{1}{2} - \left(-\frac{1}{2}\right)^{\frac{n'}{2}}\right) + (-1)^{\frac{n'-2}{2}} \begin{bmatrix} 2\pi \\ 0 \end{bmatrix} \\ &= \frac{X_\ominus}{2^{\frac{n'-2}{2}}} + \frac{2}{3} \begin{bmatrix} 2\pi \\ 0 \end{bmatrix} (-1)^{\frac{n'-2}{2}} \left(1 - \left(-\frac{1}{2}\right)^{\frac{n'}{2}}\right). \end{aligned}$$

Thus, we can say in general that

$$X_{\ominus n} = \begin{cases} \frac{X_\ominus}{2^{\frac{n-1}{2}}} + \frac{1}{3} \begin{bmatrix} 2\pi \\ 0 \end{bmatrix} (-1)^{\frac{n-3}{2}} \left(1 - (-\frac{1}{2})^{\frac{n-1}{2}}\right), & \text{for } n \text{ odd,} \\ \frac{X_\ominus}{2^{\frac{n'-2}{2}}} + \frac{2}{3} \begin{bmatrix} 2\pi \\ 0 \end{bmatrix} (-1)^{\frac{n'-2}{2}} \left(1 - (-\frac{1}{2})^{\frac{n'}{2}}\right), & \text{for } n \text{ even,} \end{cases}$$

but then clearly

$$X_{\oplus n} = \begin{cases} \frac{X_\oplus}{2^{\frac{n-1}{2}}} - \frac{1}{3}\begin{bmatrix} 2\pi \\ 0 \end{bmatrix}(-1)^{\frac{n-3}{2}}\left(1-(-\tfrac{1}{2})^{\frac{n-1}{2}}\right), & \text{for } n \text{ odd,} \\ \frac{X_\oplus}{2^{\frac{n'-2}{2}}} - \frac{2}{3}\begin{bmatrix} 2\pi \\ 0 \end{bmatrix}(-1)^{\frac{n'-2}{2}}\left(1-(-\tfrac{1}{2})^{\frac{n'}{2}}\right), & \text{for } n \text{ even.} \end{cases}$$

Analogously, we find

$$Y_{\ominus n} = \begin{cases} \frac{Y_\ominus}{2^{\frac{n-1}{2}}} + \frac{1}{3}\begin{bmatrix} 0 \\ 2\pi \end{bmatrix}(-1)^{\frac{n-3}{2}}\left(1-(-\tfrac{1}{2})^{\frac{n-1}{2}}\right), & \text{for } n \text{ odd,} \\ \frac{Y_\ominus}{2^{\frac{n'-2}{2}}} + \frac{2}{3}\begin{bmatrix} 0 \\ 2\pi \end{bmatrix}(-1)^{\frac{n'-2}{2}}\left(1-(-\tfrac{1}{2})^{\frac{n'}{2}}\right), & \text{for } n \text{ even,} \end{cases}$$

and

$$Y_{\oplus n} = \begin{cases} \frac{Y_\oplus}{2^{\frac{n-1}{2}}} - \frac{1}{3}\begin{bmatrix} 0 \\ 2\pi \end{bmatrix}(-1)^{\frac{n-3}{2}}\left(1-(-\tfrac{1}{2})^{\frac{n-1}{2}}\right), & \text{for } n \text{ odd,} \\ \frac{Y_\oplus}{2^{\frac{n'-2}{2}}} - \frac{2}{3}\begin{bmatrix} 0 \\ 2\pi \end{bmatrix}(-1)^{\frac{n'-2}{2}}\left(1-(-\tfrac{1}{2})^{\frac{n'}{2}}\right), & \text{for } n \text{ even.} \end{cases}$$

Comment: By our construction we have (analogous to the properties for $X_{\ominus n}$) that for n' even,

$$\frac{X_{\oplus n'}}{2} = X_{\oplus n'+1}, \quad \frac{Y_{\ominus n'}}{2} = Y_{\ominus n'+1}, \quad \frac{Y_{\oplus n'}}{2} = Y_{\oplus n'+1}.$$

Moreover,

$$X_{\oplus n'} = X_{\oplus n'-1} - (-1)^{\frac{n'-2}{2}}\begin{bmatrix} 2\pi \\ 0 \end{bmatrix}, \quad Y_{\ominus n'} = Y_{\ominus n'-1} + (-1)^{\frac{n'-2}{2}}\begin{bmatrix} 0 \\ 2\pi \end{bmatrix},$$
$$Y_{\oplus n'} = Y_{\oplus n'-1} - (-1)^{\frac{n'-2}{2}}\begin{bmatrix} 0 \\ 2\pi \end{bmatrix}.$$

4. Proof of Theorem 2.1

For the proof of Theorem 2.1 we will require three technical lemmas.

Lemma 4.1. *For all odd $n \geq 3$, $X_{\ominus n} \subseteq G_{SO}$.*

Proof. Since all such

$$X_\ominus \subseteq \left[-\frac{\pi}{2}, 0\right) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right),$$

by definition, for all functions f,

$$f(X_\ominus) \subseteq f\left(\left[-\frac{\pi}{2}, 0\right) \times \left[\frac{\pi}{2}, \frac{\pi}{2}\right)\right).$$

Therefore, we only need to prove that for all odd $n \geq 3$,

$$\left[\left[-\frac{\pi}{2},0\right)\times\left[-\frac{\pi}{2},\frac{\pi}{2}\right)\right]_{\ominus n}\subseteq G_{SO}.$$

Let

$$S_{\ominus n}:=\left[\left[-\frac{\pi}{2},0\right)\times\left[-\frac{\pi}{2},\frac{\pi}{2}\right)\right]_{\ominus n}$$

represent the result of the n^{th} step of the crossover algorithm applied to

$$\left[-\frac{\pi}{2},0\right)\times\left[-\frac{\pi}{2},\frac{\pi}{2}\right).$$

Notice that $S_{\ominus n}$ is of the form of the more general set $X_{\ominus n}$, and, therefore, we can use our derived bounds (see above on page 71) for $X_{\ominus n}$ in terms of n to determine the bounds for $S_{\ominus n}$.

We begin by showing that $S_{\ominus n}\subseteq[-\pi,\pi)\times[-\pi,\pi)$. That this is satisfied for the vertical bounds of $S_{\ominus n}$ is clear, so we will only consider the horizontal bounds. Note that when we use the phrase "vertical bound," we refer to both upper and lower bounds. By a vertical upper bound, we mean to say that such a number is greater than or equal to all y-coordinates of the points in that set. When we use the phrase "horizontal bound of a set," we refer to both left and right hand bounds of a set. By left hand bound, we mean to signify a number that is less than or equal to all of the horizontal coordinates of the points in that set.

Case 1. $n=4k+1$ for some $k\in\mathbb{Z}$. Then $S_{\ominus n}$ is on the left side of the origin. Thus, the horizontal left hand bound (LHB) for $S_{\ominus n}$ is

$$-\frac{1}{2^{\frac{n-1}{2}}}\begin{bmatrix}\pi\\0\end{bmatrix}+\frac{1}{3}\begin{bmatrix}2\pi\\0\end{bmatrix}(-1)^{\frac{n-3}{2}}\left(1-\left(-\frac{1}{2}\right)^{\frac{n-1}{2}}\right),$$

and we must show it is bounded below by $\begin{bmatrix}-\pi\\0\end{bmatrix}$. In the x-coordinate,

$$\begin{aligned}-\pi\leq-\frac{\pi}{2^{\frac{n-1}{2}}}+\frac{2\pi}{3}(-1)^{\frac{n-3}{2}}\left(1-\left(-\frac{1}{2}\right)^{\frac{n-1}{2}}\right)&\Longleftrightarrow 1\geq\frac{1}{2^{2k}}-\frac{2}{3}(-1)^{2k-1}\left(1-\left(-\frac{1}{2}\right)^{2k}\right)\\ &\Longleftrightarrow 1\geq\frac{1}{2^{2k}}+\frac{2}{3}\left(1-\frac{1}{2^{2k}}\right)\\ &\Longleftrightarrow\frac{1}{3}\geq\frac{1}{2^{2k}}\left(1-\frac{2}{3}\right)\\ &\Longleftrightarrow 1\geq\frac{1}{2^{2k}},\end{aligned}$$

which is clearly true because $3 \le n = 4k + 1 \Rightarrow \frac{1}{2} \le k$, and $k \in \mathbb{Z}$, so $1 \le k$. Moreover, since the LHB on the x-coordinates of $S_{\ominus n}$ is less than the horizontal right hand bound (RHB), we know that $-\pi <$ RHB.

Case 2. $n = 4k + 3$ for some $k \in \mathbb{Z}$. Then $S_{\ominus n}$ is on the right hand side of the origin. Therefore, we want to show that the horizontal RHB on $S_{\ominus n}$ is less than or equal to π, that is, that

$$\begin{aligned}\frac{2\pi}{3}(-1)^{\frac{n-3}{2}}\Big(1-\big(-\frac{1}{2}\big)^{\frac{n-1}{2}}\Big) \le \pi &\iff \frac{2}{3}(-1)^{2k}\Big(1-\big(-\frac{1}{2}\big)^{2k+1}\Big) \le 1\\ &\iff \frac{2}{3}\Big(1+\big(\frac{1}{2}\big)^{2k+1}\Big) \le 1\\ &\iff \Big(1+\big(\frac{1}{2}\big)^{2k+1}\Big) \le \frac{3}{2}\\ &\iff \Big(\frac{1}{2}\Big)^{2k} \le 1,\end{aligned}$$

which is trivially true since n is odd and $n \ge 3 \Rightarrow 3 \le 4k+3 \Rightarrow 0 \le k$. We know that in this case, LHB $\le$ RHB $\le \pi$, as needed. But then in all possible cases, it is true that $S_{\ominus n} \subseteq [-\pi, \pi) \times [-\pi, \pi)$, and therefore that $X_{\ominus n} \subseteq [-\pi, \pi) \times [-\pi, \pi)$.

Now we want to show that

$$X_{\ominus n} \nsubseteq \Big[-\frac{\pi}{2}, \frac{\pi}{2}\Big) \times \Big[-\frac{\pi}{2}, \frac{\pi}{2}\Big),$$

for all $n \ge 3$. Recall from our earlier discussion that this will follow from the fact that

$$S_{\ominus n} \nsubseteq \Big[-\frac{\pi}{2}, \frac{\pi}{2}\Big) \times \Big[-\frac{\pi}{2}, \frac{\pi}{2}\Big).$$

We can prove this fact by merely showing that the horizontal bounds on the set $S_{\ominus n}$ are not contained in the set $\big[-\frac{\pi}{2}, \frac{\pi}{2}\big]$. Thus, the vertical bounds on the set $S_{\ominus n}$ are irrelevant.

Case 1. $n = 4k + 1$ for some $k \in \mathbb{Z}$. Recall $S_{\ominus n}$ is on the left hand side of the origin, so we must show that the horizontal RHB on the x-coordinates of the set $S_{\ominus n}$ is less than or equal to $-\frac{\pi}{2}$.

$$\begin{aligned}-\frac{\pi}{2} \ge \frac{2\pi}{3}(-1)^{\frac{n-3}{2}}\left(1-\left(-\frac{1}{2}\right)^{\frac{n-1}{2}}\right) &\iff -\frac{1}{2} \ge \frac{2}{3}(-1)^{2k-1}\left(1-\left(-\frac{1}{2}\right)^{2k}\right)\\ &\iff -\frac{1}{2} \ge \frac{2}{3}\left(\frac{1}{2^{2k}}-1\right)\\ &\iff \frac{1}{4} \ge \frac{1}{2^{2k}},\end{aligned}$$

which is true since $3 \leq n = 4k+1 \Rightarrow \frac{1}{2} \leq k$ and $k \in \mathbb{Z}$ (so $\Rightarrow 1 \leq k$).

Case 2. $n = 4k+3$ for some $k \in \mathbb{Z}$. Then $S_{\ominus n}$ is on the right hand side of the origin, and therefore we want to show that the horizontal LHB on $S_{\ominus n}$ is greater than or equal to $\frac{\pi}{2}$.

$$\begin{aligned}\frac{\pi}{2} &\leq -\frac{\pi}{2^{\frac{n+1}{2}}} + \frac{2\pi}{3}(-1)^{\frac{n-3}{2}}\left(1-\left(-\frac{1}{2}\right)^{\frac{n-1}{2}}\right)\\ &\iff \frac{1}{2} \leq -\frac{1}{2^{2k+2}} + \frac{2}{3}(-1)^{2k}\left(1-\left(-\frac{1}{2}\right)^{2k+1}\right)\\ &\iff -\frac{1}{6} \leq \frac{1}{2^{k+1}}\left(\frac{2}{3}-\frac{1}{2}\right)\\ &\iff -1 \leq \frac{1}{2^{k+1}},\end{aligned}$$

which is trivially true since $0 < \frac{1}{2^{k+1}}\ \forall k$.

Thus, $\forall n$, the horizontal bounds of $S_{\ominus n}$ are not contained in the set $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ but are contained in the set $[-\pi, \pi]$, and the vertical bounds are contained in the set $[-\pi, \pi]$. Thus $S_{\ominus n} \subseteq G_{SO}$ for all odd $n \geq 3$. Recall from our earlier discussion that it therefore follows that $X_{\ominus n} \subseteq G_{SO}$ for all odd $n \geq 3$, as needed. □

Analogously, for all odd $n \geq 3$, $X_{\oplus n} \subseteq G_{SO}$, $Y_{\ominus n} \subseteq G_{SO}$, and $Y_{\oplus n} \subseteq G_{SO}$.

Lemma 4.2. *$X_{\ominus n+4}$ and $X_{\ominus n}$ are disjoint for all $n > 1 \in \mathbb{Z}$.*

Proof. Let

$$S_{\ominus n} := \left[\left[-\frac{\pi}{2}, 0\right) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right)\right]_{\ominus n}$$

represent the result of the n^{th} step of the crossover algorithm applied to

$$\left[-\frac{\pi}{2}, 0\right) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right).$$

Since all $X_{\ominus} \subseteq \left[-\frac{\pi}{2}, 0\right) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right)$, for all functions f,

$$f(X_{\ominus n}) \subseteq f\left(\left[-\frac{\pi}{2}, 0\right) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right)\right),$$

and therefore $X_{\ominus n} \subseteq S_{\ominus n}$. Therefore, we will prove that all $S_{\ominus n}$ and $S_{\ominus (n+4)}$ are disjoint, from which our lemma follows immediately.

Consider odd n ($n = 2k+1$ for some $k \in \mathbb{Z}$). The horizontal LHB of $S_{\ominus(2k+1)}$ is

$$-\frac{\pi}{2^{k+1}} + \frac{2\pi}{3}(-1)^{k-1}\left(1-\left(-\frac{1}{2}\right)^{k}\right),$$

as follows from the rigorous definition of the set $X_{\ominus n}$, and the RHB of $S_{\ominus(2k+5)}$ is

$$\frac{2\pi}{3}(-1)^{k+1}\left(1-\left(-\frac{1}{2}\right)^{k+2}\right).$$

Therefore, the RHB of $S_{\ominus(2k+5)}$ equals the LHB of $S_{\ominus(2k+1)}$ if and only if

$$\begin{aligned}
-\frac{\pi}{2^{k+1}}+\frac{2\pi}{3}(-1)^{k-1}\left(1-\left(-\frac{1}{2}\right)^{k}\right)&=\frac{2\pi}{3}(-1)^{k+1}\left(1-\left(-\frac{1}{2}\right)^{k+2}\right)\\
\iff -\frac{1}{2^{k+1}}+\frac{2}{3}(-1)^{k-1}+\frac{2}{3}\frac{1}{(2)^{k}}&=\frac{2}{3}(-1)^{k+1}+\frac{2}{3}\frac{1}{(2)^{k+2}}\\
\iff -\frac{1}{2}\frac{1}{2^{k}}&=\frac{2}{3}\frac{1}{2^{k}}\left(\frac{1}{4}-1\right),
\end{aligned}$$

which is clearly true. So for all odd n, the LHB of the set $S_{\ominus n}$ is equivalent to the RHB of the set $S_{\ominus(n+4)}$. Recall that for n' even,

$$\tfrac{1}{2}X_{\ominus n'}=X_{\ominus n'+1},\qquad \text{which implies}\qquad \tfrac{1}{2}S_{\ominus n'}=S_{\ominus(n'+1)},$$

that is,

$$\tfrac{1}{2}S_{\ominus(n-1)}=S_{\ominus n}\qquad\text{and}\qquad \tfrac{1}{2}S_{\ominus(n+3)}=S_{\ominus(n+4)}.$$

Thus, the LHB of $S_{\ominus(n-1)}$ is equivalent to the RHB of the set $S_{\ominus(n+3)}$. But since $n-1\in\mathbb{Z}^{+}$ is even, both even and odd cases are satisfied. Thus, we can say that for all $n''>1\in\mathbb{Z}$, the LHB of the set $S_{\ominus n''}$ is equivalent to the RHB of the set $S_{\ominus(n''+4)}$. Nonetheless, these two sets are still "essentially disjoint" because their intersection has measure 0 using Lebesgue Measure. Therefore, by our earlier argument, our lemma follows. □

Lemma 4.3. *All $X_{\oplus n}$, $X_{\ominus n'}$, $Y_{\oplus n''}$, $Y_{\ominus n'''}$ are disjoint for all natural numbers n, n', n'', and n'''. Moreover, $X_{\ominus i}$ and $X_{\ominus j}$ are disjoint when $i\neq j$, with analogous properties following for sets of the form $X_{\oplus n}$, $Y_{\ominus n}$, and $Y_{\oplus n}$.*

Proof. First we show that all $X_{\oplus n}$, $X_{\ominus n}$ are disjoint (a similar argument shows that all $Y_{\oplus n}$, $Y_{\ominus n}$ are disjoint). Consider the maximal case for $X_{\oplus 1}$ and $X_{\ominus 1}$, namely,

$$\begin{aligned}
X_{\oplus 1}&=\left[0,\frac{\pi}{2}\right)\times\left[-\frac{\pi}{2},\frac{\pi}{2}\right),\\
X_{\ominus 1}&=\left[-\frac{\pi}{2},0\right)\times\left[-\frac{\pi}{2},\frac{\pi}{2}\right).
\end{aligned}$$

Because all other $X_{\oplus n}$, $X_{\ominus n}$ are copies of $X_{\oplus 1}$ and $X_{\ominus 1}$ that have been translated along the x-axis and scaled, we will consider only their x-coordinates. We note that because sets of the form $X_{\oplus n}$, $X_{\ominus n}$ are never scaled by factors α, for $|\alpha|>1$, they are all contained in $[-\infty,\infty)\times\left[-\frac{\pi}{2},\frac{\pi}{2}\right)$, that is, their vertical bounds are contained in the set $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right)$. From the crossover algorithm, we have that the following hold for all m (except for $4m+1 = 1$) in the x-coordinate:

$$\begin{aligned}
X_{\oplus 4m+1} &= \tfrac{1}{2} X_{\oplus 4m}, & Xm4m+1 &= \tfrac{1}{2} X_{\ominus 4m},\\
X_{\oplus 4m+2} &= X_{\oplus 4m+1} - 2\pi, & X_{\ominus 4m+2} &= X_{\ominus 4m+1} + 2\pi,\\
X_{\oplus 4m+3} &= \tfrac{1}{2} X_{\oplus 4m+2}, & X_{\ominus 4m+3} &= \tfrac{1}{2} X_{\ominus 4m+2},\\
X_{\oplus 4m+4} &= X_{\oplus 4m+3} + 2\pi, & X_{\ominus 4m+4} &= X_{\ominus 4m+3} - 2\pi,\\
\Rightarrow X_{\oplus 4(m+1)+1} &= \tfrac{1}{4} X_{\oplus 4m+1} + \tfrac{\pi}{2}, & \Rightarrow X_{\ominus 4(m+1)+1} &= \tfrac{1}{4} X_{\ominus 4m+1} - \tfrac{\pi}{2}.
\end{aligned}$$

Solving the recurrence relation for $X_{\oplus 4m+1}$ and $X_{\ominus 4m+1}$, and using the solution to obtain the other cases, we obtain the following:

$$\begin{aligned}
X_{\oplus 4m+1} &= \frac{1}{4^m} X_{\oplus 1} + \frac{2}{3}\left(1 - \frac{1}{4^m}\right)\pi, & X_{\ominus 4m+1} &= \frac{1}{4^m} X_{\ominus 1} + \frac{2}{3}\left(\frac{1}{4^m} - 1\right)\pi,\\
X_{\oplus 4m+2} &= \frac{1}{4^m} X_{\oplus 1} - \frac{2}{3}\left(2 + \frac{1}{4^m}\right)\pi, & X_{\ominus 4m+2} &= \frac{1}{4^m} X_{\ominus 1} + \frac{2}{3}\left(\frac{1}{4^m} + 2\right)\pi,\\
X_{\oplus 4m+3} &= \frac{1}{2}\frac{1}{4^m} X_{\oplus 1} - \frac{1}{3}\left(2 + \frac{1}{4^m}\right)\pi, & X_{\ominus 4m+3} &= \frac{1}{2}\frac{1}{4^m} X_{\ominus 1} + \frac{1}{3}\left(\frac{1}{4^m} + 2\right)\pi,\\
X_{\oplus 4m+4} &= \frac{1}{2}\frac{1}{4^m} X_{\oplus 1} + \frac{1}{3}\left(4 - \frac{1}{4^m}\right)\pi, & X_{\ominus 4m+4} &= \frac{1}{2}\frac{1}{4^m} X_{\ominus 1} + \frac{1}{3}\left(\frac{1}{4^m} - 4\right)\pi.
\end{aligned}$$

Using our maximal $X_{\oplus 1}$ and $X_{\ominus 1}$, we find that

$$\begin{aligned}
X_{\oplus 4m+1} &= \left[\left(\tfrac{2}{3} - \tfrac{2}{3}\left(\tfrac{1}{4}\right)^m\right)\pi, \left(\tfrac{2}{3} - \tfrac{1}{6}\left(\tfrac{1}{4}\right)^m\right)\pi\right) \subset \left[0, \tfrac{2}{3}\pi\right),\\
X_{\oplus 4m+2} &= \left[\left(-\tfrac{4}{3} - \tfrac{2}{3}\left(\tfrac{1}{4}\right)^m\right)\pi, \left(-\tfrac{4}{3} - \tfrac{1}{6}\left(\tfrac{1}{4}\right)^m\right)\pi\right) \subset \left[-2\pi, -\tfrac{4}{3}\pi\right),\\
X_{\oplus 4m+3} &= \left[\left(-\tfrac{2}{3} - \tfrac{1}{3}\left(\tfrac{1}{4}\right)^m\right)\pi, \left(-\tfrac{2}{3} - \tfrac{1}{12}\left(\tfrac{1}{4}\right)^m\right)\pi\right) \subset \left[-\pi, -\tfrac{2}{3}\pi\right),\\
X_{\oplus 4m+4} &= \left[\left(\tfrac{4}{3} - \tfrac{1}{3}\left(\tfrac{1}{4}\right)^m\right)\pi, \left(\tfrac{4}{3} - \tfrac{1}{12}\left(\tfrac{1}{4}\right)^m\right)\pi\right) \subset \left[\pi, \tfrac{4}{3}\pi\right),\\
X_{\ominus 4m+1} &= \left[\left(\tfrac{1}{6}\left(\tfrac{1}{4}\right)^m - \tfrac{2}{3}\right)\pi, \left(\tfrac{2}{3}\left(\tfrac{1}{4}\right)^m - \tfrac{2}{3}\right)\pi\right) \subset \left[-\tfrac{2}{3}\pi, 0\right),\\
X_{\ominus 4m+2} &= \left[\left(\tfrac{1}{6}\left(\tfrac{1}{4}\right)^m + \tfrac{4}{3}\right)\pi, \left(\tfrac{2}{3}\left(\tfrac{1}{4}\right)^m + \tfrac{4}{3}\right)\pi\right) \subset \left[\tfrac{4}{3}\pi, 2\pi\right),\\
X_{\ominus 4m+3} &= \left[\left(\tfrac{1}{12}\left(\tfrac{1}{4}\right)^m + \tfrac{2}{3}\right)\pi, \left(\tfrac{1}{3}\left(\tfrac{1}{4}\right)^m + \tfrac{2}{3}\right)\pi\right) \subset \left[\tfrac{2}{3}\pi, \pi\right),\\
X_{\ominus 4m+4} &= \left[\left(\tfrac{1}{12}\left(\tfrac{1}{4}\right)^m - \tfrac{4}{3}\right)\pi, \left(\tfrac{1}{3}\left(\tfrac{1}{4}\right)^m - \tfrac{4}{3}\right)\pi\right) \subset \left[-\tfrac{4}{3}\pi, -\pi\right).
\end{aligned}$$

Trivially, we conclude that the eight different sets of intervals are disjoint. Within each set of intervals, note that both endpoints of the intervals either monotonically increase (for $X_{\oplus n}$) or monotonically decrease (for $X_{\ominus n}$) as n increases; we also find that the right endpoint of $X_{\oplus 4m+k}$ is equal to the left endpoint of $X_{\oplus 4(m+1)+k}$ for all possible values of k, and the left endpoint of $X_{\ominus 4m+k}$ is equal to the right endpoint of $X_{\ominus 4(m+1)+k}$ for all possible values of k (See the proof of Lemma 4.2.) Thus, all the $X_{\oplus 4m+k}$ and $X_{\ominus 4m+k}$ are disjoint for all k, meaning we have proved that all $X_{\oplus n}$, $X_{\ominus n}$ are disjoint. Moreover, all $X_{\ominus i}$ and $X_{\ominus j}$ and all $X_{\oplus i}$ and $X_{\ominus j}$ are disjoint when i does not equal j. Analogously, all $Y_{\ominus i}$ and $Y_{\ominus j}$ and all $Y_{\oplus i}$ and $Y_{\oplus j}$ are disjoint when i does not equal j.

To show that the sets of the form $X_{\oplus n}$, $X_{\ominus n}$, $Y_{\oplus n}$ and $Y_{\ominus n}$ are disjoint, consider the following: All the sets of the form $X_{\oplus n}$, $X_{\ominus n}$ are contained in the region

$$[-2\pi, 2\pi) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right).$$

Similarly, all the sets of the form $Y_{\oplus n}$, $Y_{\ominus n}$ are contained the region

$$\left[-\frac{\pi}{2}, \frac{\pi}{2}\right) \times [-2\pi, 2\pi).$$

The intersection between these two regions is

$$\left[-\frac{\pi}{2}, \frac{\pi}{2}\right) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right),$$

but the only sets in this region are $X_{\oplus 1}$, $X_{\ominus 1}$, $Y_{\oplus 1}$, and $Y_{\ominus 1}$, and by definition, these are disjoint, completing the proof. □

Remark 4.1. The proof of Lemma 4.3 shows that all sets of the form $X_{\oplus n}$, $X_{\ominus n}$ for odd $n \geq 3$ are in the area

$$[-\pi, \pi) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right) \setminus \left[-\frac{\pi}{2}, \frac{\pi}{2}\right) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right) \in G_{SO}.$$

Exclusion from the inner square is due to the fact that $X_{\oplus 1}$, $X_{\ominus 1}$, $Y_{\oplus 1}$, and $Y_{\ominus 1}$ occupy that space, and all $X_{\ominus i}$ and $X_{\ominus j}$ are disjoint if $i \neq j$ (with analogous results for $X_{\oplus}$, $Y_{\ominus}$, and $Y_{\oplus}$), implying no other sets of the form $X_{\ominus n}$, $X_{\oplus n}$, $Y_{\ominus n}$, and $Y_{\oplus n}$ can occupy that space, providing a short proof of Lemma 4.1.

We can now prove our main result.

Proof of Theorem 2.1. Let W be defined as in Example 1 to be

$$\Big(\bigcup_{i=1}^{\infty}[X_{\ominus 2i} \cup X_{\oplus 2i} \cup Y_{\ominus 2i} \cup Y_{\oplus 2i}]\Big)$$
$$\cup\Big(G_{TO}\setminus\Big[\bigcup_{i=1}^{\infty}[X_{\ominus 2i-1} \cup X_{\oplus 2i-1} \cup Y_{\ominus 2i-1} \cup Y_{\oplus 2i-1}]\Big]\Big)$$
$$= \Big(\bigcup_{i=1}^{\infty}[X_{\ominus 2i} \cup X_{\oplus 2i} \cup Y_{\ominus 2i} \cup Y_{\oplus 2i}]\Big)$$
$$\cup\Big(G_{SO}\setminus\Big[\bigcup_{i=2}^{\infty}[X_{\ominus 2i-1} \cup X_{\oplus 2i-1} \cup Y_{\ominus 2i-1} \cup Y_{\oplus 2i-1}]\Big]\Big).$$

Part I. We will first show that W is dilation congruent to G_{SO}. Let

$$\Phi_{X_\ominus} : \{X_{\ominus i} : i \text{ is even} \geq 2\} \to \{X_{\ominus j} : j \text{ is odd} \geq 3\},$$

be such that for all even n,

$$\Phi_{X_\ominus}(X_{\ominus n}) := \tfrac{1}{2}X_{\ominus n} = X_{\ominus n+1} \in G_{SO},$$

from Lemma 4.1 and the discussion on the top of Section 3 of our paper.

Claim 1. $\Phi_{X_\ominus}$ is surjective. Take an arbitrary $X_{\ominus j}$ such that $j \geq 3$ is odd. Then $\Phi_{X_\ominus}(X_{\ominus j-1}) = X_{\ominus j}$.

Claim 2. $\Phi_{X_\ominus}$ is injective. Suppose

$$\Phi_{X_\ominus}(X_{\ominus j}) = \Phi_{X_\ominus}(X_{\ominus i}),$$

for some even i and j. Then $X_{\ominus j+1} = X_{\ominus i+1}$, and therefore

$$X_{\ominus j} = 2X_{\ominus j+1} = 2X_{\ominus i+1} = X_{\ominus i},$$

as needed.

Therefore, $\Phi_{X_\ominus}$ is a bijection. Similarly, define $\Phi_{X_\oplus}$, $\Phi_{Y_\ominus}$ and $\Phi_{Y_\oplus}$, which are all bijections by analogous arguments. But then

$$\Phi_{X_\ominus}\Big(\bigcup_{i=1}^{\infty} X_{\ominus 2i}\Big) = \bigcup_{i=1}^{\infty} \Phi_{X_\ominus}(X_{\ominus 2i}) = \bigcup_{i=1}^{\infty} X_{\ominus 2i+1} \in G_{SO},$$

by Lemma 4.1, all of the white spaces ($X_{\ominus k}$ for $k \geq 3$ and odd) in G_{SO} are filled because $\Phi_{X_\ominus}$ is surjective, and all of the black pieces ($X_{\ominus n}$ for n even) have been mapped into G_{SO} injectively so that no two distinct black pieces map to the same white space. Analogous properties follow for $\Phi_{X_\oplus}$, $\Phi_{Y_\ominus}$, and $\Phi_{Y_\oplus}$.

Let

$$\Phi : \{X_{\ominus i} \cup X_{\oplus i} \cup Y_{\ominus i} \cup Y_{\oplus i} : i \text{ is even} \geq 2\} \to \{X_{\ominus j} \cup X_{\oplus j} \cup Y_{\ominus j} \cup Y_{\oplus j} : j \text{ is odd} \geq 3\}$$

be such that

$$\begin{aligned}\Phi\left(X_{\ominus i} \cup X_{\oplus i} \cup Y_{\ominus i} \cup Y_{\oplus i}\right) &= \Phi_{X_\ominus}\left(X_{\ominus i}\right) \cup \Phi_{X_\oplus}\left(X_{\oplus i}\right) \cup \Phi_{Y_\ominus}\left(Y_{\ominus i}\right) \cup \Phi_{Y_\oplus}\left(Y_{\oplus i}\right)\\ &= \left(X_{\ominus i+1} \cup X_{\oplus i+1} \cup Y_{\ominus i+1} \cup Y_{\oplus i+1}\right).\end{aligned}$$

Then

$$\begin{aligned}&\Phi\Big(\bigcup_{i=1}^{\infty}\left[X_{\ominus 2i} \cup X_{\oplus 2i} \cup Y_{\ominus 2i} \cup Y_{\oplus 2i}\right]\Big)\\ &\qquad= \bigcup_{i=1}^{\infty}\left[\Phi_{X_\ominus}\left(X_{\ominus 2i}\right) \cup \Phi_{X_\oplus}\left(X_{\oplus 2i}\right) \cup \Phi_{Y_\ominus}\left(Y_{\ominus 2i}\right) \cup \Phi_{Y_\oplus}\left(Y_{\oplus 2i}\right)\right]\\ &\qquad= \bigcup_{i=1}^{\infty}\left[X_{\ominus 2i+1} \cup X_{\oplus 2i+1} \cup Y_{\ominus 2i+1} \cup Y_{\oplus 2i+1}\right] \in G_{SO}.\end{aligned}$$

Φ is clearly a bijection, so that Φ maps all exterior black pieces into all interior white pieces such that no distinct black pieces map to the same white piece.

Thus,

$$\begin{aligned}&\Phi\Big(\bigcup_{i=1}^{\infty}\left[X_{\ominus 2i} \cup X_{\oplus 2i} \cup Y_{\ominus 2i} \cup Y_{\oplus 2i}\right]\Big)\\ &\qquad\bigcup\Big[G_{SO}\backslash\Big(\bigcup_{i=2}^{\infty}[X_{\ominus 2i-1} \cup X_{\oplus 2i-1} \cup Y_{\ominus 2i-1} \cup Y_{\oplus 2i-1}]\Big)\Big] = G_{SO},\end{aligned}$$

as needed. Therefore, W is dilation congruent to G_{SO}.

Part II. We will now prove that W is translation congruent to G_{TO}. Let

$$\Psi_{X_\ominus} : \{X_{\ominus i} : i \text{ is even} \geq 2\} \to \{X_{\ominus j} : j \text{ is odd} \geq 1\}$$

be such that for all even n,

$$\Psi_{X_\ominus}(X_{\ominus n}) := X_{\ominus n} - (-1)^{\frac{n-2}{2}}\begin{bmatrix}2\pi\\0\end{bmatrix} = \frac{1}{2}X_{\ominus(n-2)} = X_{\ominus n-1} \in G_{TO},$$

using the discussion on the top of page 70, Lemma 4.1, and the fact that

$$X_{\ominus 1} \in \Big[-\frac{\pi}{2}, \frac{\pi}{2}\Big) \times \Big[-\frac{\pi}{2}, \frac{\pi}{2}\Big).$$

Claim 1. $\Psi_{X_\ominus}$ is surjective. Take an arbitrary $X_{\ominus j}$ such that $j \geq 1$ is odd. Then $\Phi_{X_\ominus}(X_{\ominus j+1}) = X_{\ominus j}$.

Claim 2. $\Psi_{X_\ominus}$ is injective. Suppose $\Psi_{X_\ominus}(X_{\ominus j}) = \Psi_{X_\ominus}(X_{\ominus i})$, for some even i and j. Then $X_{\ominus j-1} = X_{\ominus i-1}$. Suppose that $(j-1) \neq (i-1)$. Then by Lemma 4.3, $X_{\ominus j-1} \cap X_{\ominus i-1} = \varnothing$, which contradicts $X_{\ominus j-1} = X_{\ominus i-1}$. Thus it must be true that $(j-1) = (i-1)$ and thus $X_{\ominus i} = X_{\ominus j}$, as needed.

Therefore, $\Psi_{X_\ominus}$ is a bijection. But then

$$\Psi_{X_\ominus}\Big(\bigcup_{i=1}^{\infty} X_{\ominus 2i}\Big) = \bigcup_{i=1}^{\infty} \Psi_{X_\ominus}(X_{\ominus 2i}) = \bigcup_{i=1}^{\infty} X_{\ominus 2i-1}.$$

Therefore, all blank spaces in G_{TO} of the form $X_{\ominus n}$ are filled when $\Psi_{X_\ominus}$ acts on $\bigcup_{i=1}^{\infty} X_{\ominus 2i}$ since $\Psi_{X_\ominus}$ is onto. Moreover, all black pieces of the form $X_{\ominus n}$ are contained in the set $\bigcup_{i=1}^{\infty} X_{\ominus 2i}$, and therefore have been mapped into G_{TO}. Since $\Psi_{X_\ominus}$ is injective, no two distinct black pieces will map to the same white piece.

Similarly, define $\Psi_{X_\oplus}$, $\Psi_{Y_\ominus}$, and $\Psi_{Y_\oplus}$, which are all bijections by analogous arguments.

Define

$$\Psi : \{X_{\ominus i} \cup X_{\oplus i} \cup Y_{\ominus i} \cup Y_{\oplus i} : i \text{ is even } \geq 2\} \to \{X_{\ominus j} \cup X_{\oplus j} \cup Y_{\ominus j} \cup Y_{\oplus j} : j \text{ is odd} \geq 1\},$$

to be such that

$$\begin{aligned}\Psi\,(X_{\ominus i} \cup X_{\oplus i} \cup Y_{\ominus i} \cup Y_{\oplus i}) &= \Psi_{X_\ominus}(X_{\ominus i}) \cup \Psi_{X_\oplus}(X_{\oplus i}) \cup \Psi_{Y_\ominus}(Y_{\ominus i}) \cup \Psi_{Y_\oplus}(Y_{\oplus i}) \\ &= \big[X_{\ominus i-1} \cup X_{\oplus i-1} \cup Y_{\ominus i-1} \cup Y_{\oplus i-1}\big] \in G_{TO}.\end{aligned}$$

Ψ is clearly a bijection, and therefore when Ψ acts on the entire domain, that is,

$$\begin{aligned}\Psi\Big(\bigcup_{i=1}^{\infty}\big[X_{\ominus 2i} \cup X_{\oplus 2i} \cup Y_{\ominus 2i} \cup Y_{\oplus 2i}\big]\Big) &= \bigcup_{i=1}^{\infty} \Psi\Big(X_{\ominus 2i} \cup X_{\oplus 2i} \cup Y_{\ominus 2i} \cup Y_{\oplus 2i}\Big) \\ &= \bigcup_{i=1}^{\infty}\big[X_{\ominus 2i-1} \cup X_{\oplus 2i-1} \cup Y_{\ominus 2i-1} \cup Y_{\oplus 2i-1}\big] \in G_{TO},\end{aligned}$$

every white space (of the form $X_{\ominus n}$ for odd n) in G_{TO} is filled by some black piece (of the form $X_{\ominus n'}$ for some even n') from the exterior of G_{TO} since Ψ is surjective. No two distinct black pieces map to the same white piece since Ψ is injective. Moreover, every black piece outside G_{TO} is contained in the domain of Ψ, and therefore every black piece outside G_{TO} is mapped into G_{TO}. Thus,

$$\begin{aligned}&\Psi\Big(\bigcup_{i=1}^{\infty}\big[X_{\ominus 2i} \cup X_{\oplus 2i} \cup Y_{\ominus 2i} \cup Y_{\oplus 2i}\big]\Big) \\ &\qquad \bigcup\Big[G_{TO} \setminus \Big(\bigcup_{i=1}^{\infty}[X_{\ominus 2i-1} \cup X_{\oplus 2i-1} \cup Y_{\ominus 2i-1} \cup Y_{\oplus 2i-1}]\Big)\Big] = G_{TO},\end{aligned}$$

and thus by definition, W is dilation congruent modulo 2π to G_{TO}. By definition W is a wavelet set. □

A different example of a partition of the inner square conforming to the requirements of Theorem 2.1 is shown in Figure 2 with the resulting wavelet set.

5. Patch wavelet sets

All of the wavelet sets we have considered thus far are crossover wavelet sets. In this class, regions are added to or subtracted from alternating sides of the inner square. Alternatively, we could add or subtract regions to the same side of the square for each translation. Such wavelet sets are called patch wavelet sets. To illustrate the patch algorithm, we give an example. The reader will note that this example is actually a well known wavelet set: the *wedding cake* wavelet set (Figure 3); see [Dai and Larson 1998, Example 6.6.1] and also [Dai et al. 1998].

Patch Example 1. Let

$$X_\ominus = \left[-\frac{\pi}{2}, 0\right) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right), \qquad X_\oplus = \left[0, \frac{\pi}{2}\right) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right),$$
$$Y_\ominus = \varnothing, \qquad Y_\oplus = \varnothing.$$

Consider the piece $X_\ominus$. Start by translating $X_\ominus$ by $\begin{bmatrix} -2\pi \\ 0 \end{bmatrix}$ (keeping it on the same side of the origin) to obtain $X_{\ominus 2}$. We find that

$$X_{\ominus 2} = \left[-\frac{5\pi}{2}, -2\pi\right) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right).$$

Secondly, scale $X_{\ominus 2}$ by $\frac{1}{4}$ to obtain

$$X_{\ominus 3} = \left[-\frac{5\pi}{8}, -\frac{\pi}{2}\right) \times \left[-\frac{\pi}{8}, \frac{\pi}{8}\right).$$

Thirdly, translate $X_{\ominus 3}$ in the same direction as that of the first translation (that is, by $\begin{bmatrix} -2\pi \\ 0 \end{bmatrix}$) to obtain

$$X_{\ominus 4} = \left[-\frac{21\pi}{8}, -\frac{5\pi}{2}\right) \times \left[-\frac{\pi}{8}, \frac{\pi}{8}\right).$$

Finally, scale $X_{\ominus 4}$ by $\frac{1}{4}$ to form the set

$$X_{\ominus 5} = \left[-\frac{21\pi}{32}, -\frac{5\pi}{32}\right) \times \left[-\frac{\pi}{32}, \frac{\pi}{32}\right).$$

Continue these two steps inductively for $X_\ominus$.

We perform two similar steps on the set $X_\oplus$ inductively as well; however, we translate by $\begin{bmatrix} 2\pi \\ 0 \end{bmatrix}$ (instead of $\begin{bmatrix} -2\pi \\ 0 \end{bmatrix}$). We obtain the following as a result from the first four steps of the patch algorithm:

$$X_{\oplus 2} = \left[2\pi, \frac{5\pi}{2}\right) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right), \qquad X_{\oplus 3} = \left[\frac{\pi}{2}, \frac{5\pi}{8}\right) \times \left[-\frac{\pi}{8}, \frac{\pi}{8}\right),$$

$$X_{\oplus 4} = \left[\frac{5\pi}{2}, \frac{21\pi}{8}\right) \times \left[-\frac{\pi}{8}, \frac{\pi}{8}\right), \qquad X_{\oplus 5} = \left[\frac{5\pi}{32}, \frac{21\pi}{32}\right) \times \left[-\frac{\pi}{32}, \frac{\pi}{32}\right).$$

Continue this process inductively for $X_\oplus$ as well. In theory, we would perform similar steps for $Y_\oplus$ and $Y_\ominus$, but in this example both are the null set, and thus we have no computations to carry out for the sets $Y_\ominus$ and $Y_\oplus$.

Let W' be the set

$$\left(\bigcup_{i=1}^{\infty}[X_{\ominus 2i} \cup X_{\oplus 2i}]\right) \cup \left(G_{TO} \backslash \left[\bigcup_{i=1}^{\infty}[X_{\ominus 2i-1} \cup X_{\oplus 2i-1}]\right]\right)$$
$$= \left(\bigcup_{i=1}^{\infty}[X_{\ominus 2i} \cup X_{\oplus 2i}]\right) \cup \left(G_{SO} \backslash \left[\bigcup_{i=2}^{\infty}[X_{\ominus 2i-1} \cup X_{\oplus 2i-1}]\right]\right),$$

see Figure 3. Similarly to the crossover case, we can think of the set W' as being the union of G_{TO} combined with the sets on the exterior of G_{TO} of the form $X_{\oplus n}$, $X_{\ominus n}$ where n is even and with subsets of G_{TO} of the form $X_{\oplus n}$, $X_{\ominus n}$ where n is odd erased from G_{TO}. The reader should check that this set W' is indeed a wavelet set.

This algorithm can be generalized as follows:

(i) Partition the inner square into a maximum of four pieces. The conditions on this partition are identical to those on the partition of the inner square using the crossover algorithm as given in Theorem 2.1, and the proof for the case of the patch algorithm is similar to the proof given for the crossover algorithm.

(ii) Translate one piece of the partition by $\begin{bmatrix} \pm 2\pi \\ 0 \end{bmatrix}$ or $\begin{bmatrix} 0 \\ \pm 2\pi \end{bmatrix}$ so that the piece is translated out of the inner square and onto the half of the plane in which the original piece of the partition previously lay.

(iii) Dilate the set formed in step 2 into G_{SO} by $\frac{1}{4}$.

(iv) Translate the set formed in step 3 out of G_{SO} in the same direction as the translation in step 2 (that is, by $\begin{bmatrix} \pm 2\pi \\ 0 \end{bmatrix}$ or $\begin{bmatrix} 0 \\ \pm 2\pi \end{bmatrix}$).

(v) Dilate the set formed in step 4 into G_{SO} by $\frac{1}{4}$.

(vi) Repeat steps 2 and 3 inductively for this piece of the partition, and perform the same steps inductively on the other three pieces of the partition of the inner square.

Theorem 5.1 (Patch Algorithm). *Let $\{X_\ominus, X_\oplus, Y_\ominus, Y_\oplus\}$ be a partition of the set*

$$\left[-\frac{\pi}{2},\frac{\pi}{2}\right)\times\left[-\frac{\pi}{2},\frac{\pi}{2}\right),$$

such that $X_\ominus$ is contained in the left half of the inner square, $X_\oplus$ is contained in the right half of the inner square, $Y_\ominus$ is contained in the bottom half of the inner square, and $Y_\oplus$ is contained in the top half of the inner square. Then the set W, defined as

$$\Big[\Big(\bigcup_{i=1}^{\infty}[X_{\ominus 2i}\cup X_{\oplus 2i}\cup Y_{\ominus 2i}\cup Y_{\oplus 2i}]\Big)\cup G_{TO}\Big] \setminus\Big[\bigcup_{i=1}^{\infty}[X_{\ominus 2i-1}\cup X_{\oplus 2i-1}\cup Y_{\ominus 2i-1}\cup Y_{\oplus 2i-1}]\Big],$$

generated by this partition under translation by

$$\begin{bmatrix}\pm 2\pi\\0\end{bmatrix}\quad \textit{and} \quad \begin{bmatrix}0\\ \pm 2\pi\end{bmatrix}$$

and dilation by powers of 2 *using steps* (*i*)–(*vi*) *above*, *is a dyadic wavelet set in* $\mathbb{R}^2$.

Proof. Begin by showing the following for natural numbers n odd and n' even:

$$X_{\ominus n+2}=\frac{1}{4}\left(X_{\ominus n}-\begin{bmatrix}2\pi\\0\end{bmatrix}\right),\qquad X_{\ominus n'+2}=\frac{1}{4}X_{\ominus n'}-\begin{bmatrix}2\pi\\0\end{bmatrix},$$

$$\frac{X_{\ominus n'}}{4}=X_{\ominus n'+1},\qquad\qquad X_{\ominus n'}=X_{\ominus n'-1}-\begin{bmatrix}2\pi\\0\end{bmatrix}.$$

First, we solve the recurrence relation for n odd, and use this and the fact that $\frac{X_{\ominus n'}}{4}=X_{\ominus n'+1}$ to obtain a form for n' odd. From this point forward let n be an arbitrary odd or even natural number. We find that

$$X_{\ominus n}=\begin{cases}\frac{X_\ominus}{4^{\frac{n-1}{2}}}-\frac{1}{3}\begin{bmatrix}2\pi\\0\end{bmatrix}\left(1-(\frac{1}{4})^{\frac{n-1}{2}}\right), & \text{for } n \text{ odd}\\[2ex] \frac{X_\ominus}{4^{\frac{n'+1}{2}}}-\frac{4}{3}\begin{bmatrix}2\pi\\0\end{bmatrix}\left(1-(\frac{1}{4})^{\frac{n'-1}{2}}\right), & \text{for } n \text{ even.}\end{cases}$$

We derive similar expressions for $X_{\oplus n}$, $Y_{\ominus n}$, and $Y_{\oplus n}$.

An analogous property to that of Lemma 4.1 can be seen for the patch algorithm. Once again, we use the maximal possible $X_{\ominus n}$, that is,

$$S_{\ominus n} := \left[\left[-\frac{\pi}{2}, 0\right) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right)\right]_{\ominus n},$$

the result of the n^{th} step of the patch algorithm applied to

$$\left[-\frac{\pi}{2}, 0\right) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right).$$

We use our derived bounds for $X_{\ominus n}$ in terms of n to determine the bounds for $S_{\ominus n}$. We begin by showing that $S_{\ominus n} \subseteq [-\pi, \pi) \times [-\pi, \pi)$. That this is satisfied for the vertical bounds of $S_{\ominus n}$ is clear, so we will only consider the horizontal bounds. There is only one case to consider for the patch algorithm, the case that $n = 2k+1$ for some nonnegative integer k. (The patch algorithm requires only one case because the algorithm always translates the odd pieces out to the same side of the inner square rather than to alternating sides, as in the crossover algorithm, leading to two cases for the crossover algorithm.) Second, show that

$$X_{\ominus n} \not\subseteq \left[-\frac{\pi}{2}, \frac{\pi}{2}\right) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right),$$

for all $n \geq 3$, by showing that

$$S_{\ominus n} \not\subseteq \left[-\frac{\pi}{2}, \frac{\pi}{2}\right) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right).$$

This follows from the fact that the horizontal bounds on the set $S_{\ominus n}$ are not contained in the set $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. Thus, the vertical bounds on the set $S_{\ominus n}$ are irrelevant. Once again, here we find that there is only one case to consider (the case that $n = 2k+1$ for some nonnegative integer k). We conclude that $S_{\ominus n} \subseteq G_{SO}$ for all odd $n \geq 3$, and therefore that $X_{\ominus n} \subseteq G_{SO}$ for all odd $n \geq 3$. Analogously, for all odd $n \geq 3$, $X_{\oplus n} \subseteq G_{SO}$, $Y_{\ominus n} \subseteq G_{SO}$, and $Y_{\oplus n} \subseteq G_{SO}$.

An analogous property is true for the patch case to Lemma 4.2 for the crossover algorithm, that $X_{\ominus n+2}$ and $X_{\ominus n}$ are disjoint for all $n > 0 \in \mathbb{Z}$. We modify the argument that was used for the crossover case by showing that the left hand bound of $S_{\ominus 2k+1}$ equals the right hand bound of $S_{\ominus 2k+3}$.

Next, an analogous property is true for the patch case to that of Lemma 4.3 for crossover sets, namely, that all $X_{\oplus n}$, $X_{\ominus n'}$, $Y_{\oplus n''}$, $Y_{\ominus n'''}$ are disjoint for all natural numbers n, n', n'', and n'''. Moreover, $X_{\ominus i}$ and $X_{\ominus j}$ are disjoint when $i \neq j$, with analogous properties following for sets of the form $X_{\oplus n}$, $Y_{\ominus n}$, and $Y_{\oplus n}$.

First we show that all $X_{\oplus n}$, $X_{\ominus n}$ are disjoint. Once again, consider the maximal case for $X_{\oplus 1}$ and $X_{\ominus 1}$. Because all other $X_{\oplus n}$, $X_{\ominus n}$ are copies of $X_{\oplus 1}$ and $X_{\ominus 1}$ that have been translated along the x-axis and scaled, consider only the x-coordinates.

Because sets of the form $X_{\oplus n}$, $X_{\ominus n}$ are never scaled by factors α, for $|\alpha| > 1$, they are all contained in

$$[-\infty, \infty) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right).$$

From the Patch Algorithm, observe that, where $2m+1>1$, the following hold for all m in the x-coordinate:

$$\begin{aligned}
& X_{\ominus 2m+1} = \tfrac{1}{4} X_{\ominus 2m}, \\
& X_{\ominus 2m+2} = X_{\ominus 2m+1} - 2\pi, \\
\Rightarrow\ & X_{\ominus 2(m+1)+1} = \tfrac{1}{16} X_{\ominus 2m} - \tfrac{\pi}{2}, \\
& X_{\oplus 2m+1} = \tfrac{1}{4} X_{\oplus 2m}, \\
& X_{\oplus 2m+2} = X_{\oplus 2m+1} + 2\pi, \\
\Rightarrow\ & X_{\oplus 2(m+1)+1} = \tfrac{1}{16} X_{\oplus 2m} + \tfrac{\pi}{2}.
\end{aligned}$$

Solving these recurrence relations, we find a collection of disjoint sets, each of which contains one of the following as a subset: $X_{\ominus 2m+1}$, $X_{\ominus 2m+2}$, $X_{\oplus 2m+1}$, and $X_{\oplus 2m+2}$. Trivially, we conclude that the four different sets of intervals are disjoint. Within each set of intervals, note that both endpoints of the intervals either monotonically increase (for $X_{\oplus n}$) or monotonically decrease (for $X_{\ominus n}$) as n increases. Recall from our argument for the property similar to Lemma 4.2 (but for the patch case) that the left hand bound of $S_{\ominus 2k+1}$ equals the right hand bound of $S_{\ominus 2k+3}$. We will also find that the right hand bound of $S_{\ominus 2k+1}$ equals the left hand bound of $S_{\ominus 2k+3}$. Thereby we conclude that all $X_{\oplus n}$, $X_{\ominus n}$ are disjoint along with all $X_{\ominus i}$, $X_{\ominus j}$ and all $X_{\oplus i}$, $X_{\oplus j}$ when $i \neq j$. Analogously, all $Y_{\oplus n}$ and $Y_{\ominus n}$ are disjoint along with all $Y_{\ominus i}$ and $Y_{\ominus j}$ and all $Y_{\oplus i}$ and $Y_{\oplus j}$ when $i \neq j$.

To show that the sets of the form $X_{\oplus n}$, $X_{\ominus n}$, $Y_{\oplus n}$ and $Y_{\ominus n}$ are disjoint, consider the following: All the sets of the form $X_{\oplus n}$, $X_{\ominus n}$ are contained in the region

$$[-\infty, \infty) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right).$$

Similarly, all the sets of the form $Y_{\oplus n}$, $Y_{\ominus n}$ are contained the region

$$\left[-\frac{\pi}{2}, \frac{\pi}{2}\right) \times [-\infty, \infty).$$

The intersection between these two regions is

$$\left[-\frac{\pi}{2}, \frac{\pi}{2}\right) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right),$$

but the only sets in this region are $X_{\oplus 1}$, $X_{\ominus 1}$, $Y_{\oplus 1}$, and $Y_{\ominus 1}$, and by definition, these are disjoint.

Define the set W in the same way it was defined in the proof of the theorem for the crossover case. To show that W is dilation congruent to G_{SO}, define the bijection

$$\Phi_{X_\ominus} : \{X_{\ominus i} : i \text{ is even} \geq 2\} \to \{X_{\ominus j} : j \text{ is odd} \geq 3\},$$

such that for all even n,

$$\Phi_{X_\ominus}(X_{\ominus n}) := \frac{1}{4} X_{\ominus n} = X_{\ominus n+1} \in G_{SO}.$$

Observe that

$$\Phi_{X_\ominus}\Big(\bigcup_{i=1}^{\infty} X_{\ominus 2i}\Big) = \bigcup_{i=1}^{\infty} \Phi_{X_\ominus}\Big(X_{\ominus 2i}\Big) = \bigcup_{i=1}^{\infty} X_{\ominus 2i+1} \in G_{SO},$$

using the property analogous to Lemma 4.1 Lemma 1 but applied to the patch case. All of the white spaces ($X_{\ominus k}$ for $k \geq 3$ and odd) in G_{SO} are filled, and all of the black pieces ($X_{\ominus n}$ for n even) have been mapped into G_{SO} injectively.

Similarly, define the bijections $\Phi_{X_\oplus}$, $\Phi_{Y_\ominus}$, and $\Phi_{Y_\oplus}$. Analogous properties follow for $\Phi_{X_\oplus}$, $\Phi_{Y_\ominus}$, and $\Phi_{Y_\oplus}$. Let

$$\Phi : \{X_{\ominus i} \cup X_{\oplus i} \cup Y_{\ominus i} \cup Y_{\oplus i} : i \text{ is even} \geq 2\} \\ \to \{X_{\ominus j} \cup X_{\oplus j} \cup Y_{\ominus j} \cup Y_{\oplus j} : j \text{ is odd} \geq 3\}$$

be such that

$$\Phi\left(X_{\ominus i} \cup X_{\oplus i} \cup Y_{\ominus i} \cup Y_{\oplus i}\right) = \Phi_{X_\ominus}\left(X_{\ominus i}\right) \cup \Phi_{X_\oplus}\left(X_{\oplus i}\right) \cup \Phi_{Y_\ominus}\left(Y_{\ominus i}\right) \cup \Phi_{Y_\oplus}\left(Y_{\oplus i}\right) \\ = \left(X_{\ominus i+1} \cup X_{\oplus i+1} \cup Y_{\ominus i+1} \cup Y_{\oplus i+1}\right).$$

Using Φ, we show that W is dilation congruent to G_{SO}.

To show that W is translation congruent to G_{TO}, let

$$\Psi_{X_\ominus} : \{X_{\ominus i} : i \text{ is even} \geq 2\} \to \{X_{\ominus j} : j \text{ is odd} \geq 1\}$$

be such that for all even n,

$$\Psi_{X_\ominus}(X_{\ominus n}) := X_{\ominus n} + \begin{bmatrix} 2\pi \\ 0 \end{bmatrix} = \frac{1}{2} X_{\ominus(n-2)} = X_{\ominus n-1} \in G_{TO}.$$

$\Psi_{X_\ominus}$ is a bijection. Observe that

$$\Psi_{X_\ominus}\Big(\bigcup_{i=1}^{\infty} X_{\ominus 2i}\Big) = \bigcup_{i=1}^{\infty} \Psi_{X_\ominus}\Big(X_{\ominus 2i}\Big) = \bigcup_{i=1}^{\infty} X_{\ominus 2i-1}.$$

Therefore, all blank spaces in G_{TO} of the form $X_{\ominus n}$ are filled when $\Psi_{X_\ominus}$ acts on

$$\bigcup_{i=1}^{\infty} X_{\ominus 2i},$$

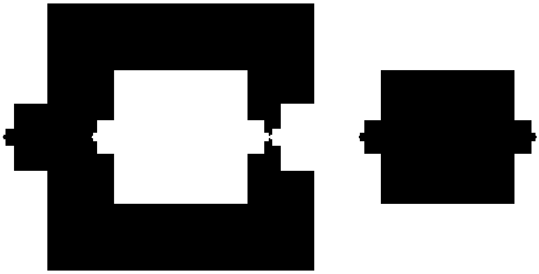


Figure 4. A wavelet set which has characteristics of both patch and crossover wavelet sets.

since $\Psi_{X_\ominus}$ is onto. Moreover, all black pieces of the form $X_{\ominus n}$ are contained in the set

$$\bigcup_{i=1}^{\infty} X_{\ominus 2i},$$

and therefore have been mapped into G_{TO}. Define similarly $\Psi_{X_\oplus}$, $\Psi_{Y_\ominus}$, and $\Psi_{Y_\oplus}$ which are all bijections by analogous arguments.

Define the bijection

$$\Psi : \{X_{\ominus i} \cup X_{\oplus i} \cup Y_{\ominus i} \cup Y_{\oplus i} : i \text{ is even } \geq 2\} \to \{X_{\ominus j} \cup X_{\oplus j} \cup Y_{\ominus j} \cup Y_{\oplus j} : j \text{ is odd} \geq 1\}$$

to be such that

$$\begin{aligned}\Psi\,(X_{\ominus i} \cup X_{\oplus i} \cup Y_{\ominus i} \cup Y_{\oplus i}) &= \Psi_{X_\ominus}\,(X_{\ominus i}) \cup \Psi_{X_\oplus}\,(X_{\oplus i}) \cup \Psi_{Y_\ominus}\,(Y_{\ominus i}) \cup \Psi_{Y_\oplus}\,(Y_{\oplus i}) \\ &= \left[X_{\ominus i-1} \cup X_{\oplus i-1} \cup Y_{\ominus i-1} \cup Y_{\oplus i-1}\right] \in G_{TO}.\end{aligned}$$

Using Φ, we show W is dilation congruent modulo 2π to G_{TO}. We conclude now that W is a wavelet set. □

6. Concluding remarks

In Figure 4, we partition the inner square in the following way:

$$X_\ominus = \left[-\frac{\pi}{2}, 0\right) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right), \qquad X_\oplus = \left[0, \frac{\pi}{2}\right) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right),$$
$$Y_\ominus = \varnothing, \qquad Y_\oplus = \varnothing.$$

To the piece $X_\ominus$ we apply the crossover algorithm. We obtain the following:

$$X_{\ominus 2} = \left[\frac{3\pi}{2}, 2\pi\right) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right), \qquad X_{\ominus 3} = \left[\frac{3\pi}{4}, \pi\right) \times \left[-\frac{\pi}{4}, \frac{\pi}{4}\right),$$

$$X_{\ominus 4} = \left[-\frac{5\pi}{4}, -2\pi\right) \times \left[-\frac{\pi}{4}, \frac{\pi}{4}\right), \qquad X_{\ominus 5} = \left[-\frac{5\pi}{8}, -\pi\right) \times \left[-\frac{\pi}{8}, \frac{\pi}{8}\right).$$

To the piece $X_\oplus$, we apply the patch algorithm and obtain the following as a result from the first four steps of the algorithm:

$$X_{\oplus 2} = \left[2\pi, \frac{5\pi}{2}\right) \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right), \qquad X_{\oplus 3} = \left[\frac{\pi}{2}, \frac{5\pi}{8}\right) \times \left[-\frac{\pi}{8}, \frac{\pi}{8}\right),$$
$$X_{\oplus 4} = \left[\frac{5\pi}{2}, \frac{21\pi}{8}\right) \times \left[-\frac{\pi}{8}, \frac{\pi}{8}\right), \qquad X_{\oplus 5} = \left[\frac{5\pi}{32}, \frac{21\pi}{32}\right) \times \left[-\frac{\pi}{32}, \frac{\pi}{32}\right).$$

We continue application of the patch algorithm to the piece $X_\oplus$ and application of the crossover algorithm to the piece $X_\ominus$ inductively. Once again we let W be the set

$$\left(\bigcup_{i=1}^{\infty}[X_{\ominus 2i} \cup X_{\oplus 2i}]\right) \cup \left(G_{TO} \setminus \left[\bigcup_{i=1}^{\infty}[X_{\ominus 2i-1} \cup X_{\oplus 2i-1}]\right]\right)$$
$$= \left(\bigcup_{i=1}^{\infty}[X_{\ominus 2i} \cup X_{\oplus 2i}]\right) \cup \left(G_{SO} \setminus \left[\bigcup_{i=2}^{\infty}[X_{\ominus 2i-1} \cup X_{\oplus 2i-1}]\right]\right),$$

where $X_{\ominus n}$ is defined according to our definition for a set of this form operated on by the crossover algorithm (see page 65), and $X_{\oplus n}$ is defined according to our definition given for a set of this form operated on by the Patch Algorithm.

This set W (see Figure 4) is a wavelet set. To see this, let

$$G(X_{\ominus \text{odd}}) := \bigcup_{i=1}^{\infty} X_{\ominus 2i-1}, \quad \text{and} \quad G(X_{\ominus \text{even}}) := \bigcup_{i=1}^{\infty} X_{\ominus 2i}.$$

Similarly, define sets for $X_\oplus$, $Y_\ominus$, $Y_\oplus$ with analogous characteristics. Observe that W is translation congruent to G_{TO} modulo 2π because

$$\bigcup_{i=1}^{\infty} X_{\ominus 4i} + \begin{bmatrix} 2\pi \\ 0 \end{bmatrix} = \bigcup_{i=0}^{\infty} X_{\ominus 4i+3},$$
$$\bigcup_{i=0}^{\infty} X_{\ominus 4i+2} - \begin{bmatrix} 2\pi \\ 0 \end{bmatrix} = \bigcup_{i=0}^{\infty} X_{\ominus 4i+1},$$
$$\bigcup_{i=1}^{\infty} X_{\oplus 2i} - \begin{bmatrix} 2\pi \\ 0 \end{bmatrix} = \bigcup_{i=0}^{\infty} X_{\oplus 2i+1}.$$

Notice

$$\bigcup_{i=0}^{\infty} X_{\oplus 2i+1} \cup \bigcup_{i=0}^{\infty} X_{\ominus 4i+1} \cup \bigcup_{i=0}^{\infty} X_{\ominus 4i+3} = G(X_{\ominus \text{odd}}) \cup G(X_{\oplus \text{odd}}),$$

and thus we observe that all of the white spaces in the set G_{TO} are filled when we translate the black sets on the exterior of G_{TO} by multiples of $\begin{bmatrix} 2\pi \\ 0 \end{bmatrix}$. Moreover, W is dilation congruent to G_{SO} because

$$\tfrac{1}{2}G(X_{\ominus\text{even}}) = G(X_{\ominus\text{odd}}) \in G_{SO},$$

that is, the even pieces of the form $X_{\ominus n}$ scale into the odd pieces of the form $X_{\ominus n}$, and

$$\tfrac{1}{4}G(X_{\oplus\text{even}}) = G(X_{\oplus\text{odd}}).$$

Thus, W is a wavelet set by definition.

Thanks to this example, we see that a wavelet set may demonstrate characteristics of both patch and crossover wavelet sets, and thereby not be classified as either type. The set contains both a patch region and a crossover region. Therefore, we have not made a complete classification of all two dimensional wavelet sets, but note that crossover wavelet sets seem to be maximally nonpatch. Finding a broader algorithm which encompasses both the patch and crossover algorithms would be an interesting problem to consider.

As a final comment, we remark that crossover and patch wavelet sets make perfect sense in one-dimension (that is, in $\mathbb{R}^1$). The reader can easily prove that all dyadic one-dimensional wavelet sets of two or three intervals are necessarily crossover wavelet sets. (Here crossover would mean through the origin.) On the other hand, the well known Journe wavelet set of 4 intervals (see [Dai and Larson 1998], Example 4.5 (i)), is easily seen to be a patch wavelet set. A characterization is not known at this time of all finite interval patch wavelet sets.

Acknowledgements

The first three authors thank our professors at Davidson College, Purdue University and The University of Tennessee — Knoxville, for their encouragement in this work. We thank the National Science Foundation for providing the funding for our research and the Texas A&M Department of Mathematics for wonderful hospitality during the Texas A&M Mathematics REU, June 10 — August 3, 2007.

References

[Baggett et al. 1999] L. W. Baggett, H. A. Medina, and K. D. Merrill, “Generalized multi-resolution analyses and a construction procedure for all wavelet sets in $\mathbb{R}^n$”, *J. Fourier Anal. Appl.* **5**:6 (1999), 563–573. MR 2001f:42055

[Benedetto and Leon 1999] J. J. Benedetto and M. T. Leon, “The construction of multiple dyadic minimally supported frequency wavelets on $\mathbb{R}^d$”, pp. 43–74 in *The functional and harmonic analysis of wavelets and frames (San Antonio, TX, 1999)*, Contemp. Math. **247**, Amer. Math. Soc., Providence, RI, 1999. MR 2001a:42034

[Benedetto and Leon 2001] J. J. Benedetto and M. Leon, "The construction of single wavelets in D-dimensions", *J. Geom. Anal.* **11**:1 (2001), 1–15. MR 2002g:42045

[Consortium 1998] T. W. Consortium, "Basic properties of wavelets", *J. Fourier Anal. Appl.* **4**:4-5 (1998), 575–594. MR 99i:42056

[Dai and Larson 1998] X. Dai and D. R. Larson, "Wandering vectors for unitary systems and orthogonal wavelets", *Mem. Amer. Math. Soc.* **134**:640 (1998), viii+68. MR 98m:47067

[Dai et al. 1997] X. Dai, D. R. Larson, and D. M. Speegle, "Wavelet sets in $\mathbb{R}^n$", *J. Fourier Anal. Appl.* **3**:4 (1997), 451–456. MR 98m:42048

[Dai et al. 1998] X. Dai, D. R. Larson, and D. M. Speegle, "Wavelet sets in $\mathbb{R}^n$. II", pp. 15–40 in *Wavelets, multiwavelets, and their applications (San Diego, CA, 1997)*, Contemp. Math. **216**, Amer. Math. Soc., Providence, RI, 1998. MR 99d:42054

[Fang and Wang 1996] X. Fang and X. Wang, "Construction of minimally supported frequency wavelets", *J. Fourier Anal. Appl.* **2**:4 (1996), 315–327. MR 97d:42030

[Gu and Han 2000] Q. Gu and D. Han, "On multiresolution analysis (MRA) wavelets in $\mathbb{R}^n$", *J. Fourier Anal. Appl.* **6**:4 (2000), 437–447. MR 2001d:42023

[Larson 2007a] D. R. Larson, "Unitary systems and wavelet sets", pp. 143–171 in *Wavelet analysis and applications*, Appl. Numer. Harmon. Anal., Birkhäuser, Basel, 2007. MR 2297918

[Larson 2007b] D. R. Larson, "Unitary systems and wavelet sets", pp. 143–171 in *Wavelet analysis and applications*, Appl. Numer. Harmon. Anal., Birkhäuser, Basel, 2007. MR 2297918

[Merrill $\geq$ 2008] K. D. Merrill, "Simple wavelet sets for scalar dilations in R^2", To appear.

[Soardi and Weiland 1998] P. M. Soardi and D. Weiland, "Single wavelets in n-dimensions", *J. Fourier Anal. Appl.* **4**:3 (1998), 299–315. MR 99k:42067

[Zhang and Larson $\geq$ 2008] X. Zhang and D. R. Larson, "Interpolation maps and congruence domains for wavelet sets", To appear.

Received: 2007-11-26 Accepted: 2007-12-12

ajhergenroeder@davidson.edu *Department of Mathematics, Davidson College, Box 5677,*
Davidson, NC 28035, United States

zcatlin@purdue.edu *Department of Mathematics, Purdue University,*
West Lafayette, IN 47907, United States

bgeorge@utk.edu *Department of Mathematics, University of Tennessee,*
Knoxville, TN 37996, United States

larson@math.tamu.edu *Department of Mathematics, Texas A&M University,*
College Station, TX 77843-3368, United States
http://www.math.tamu.edu/~larson

INVOLVE 1:1(2008)

Difference inequalities, comparison tests, and some consequences

Frank J. Palladino

(Communicated by Gerry Ladas)

We study the behavior of nonnegative sequences which satisfy certain difference inequalities. Several comparison tests involving difference inequalities are developed for nonnegative sequences. Using the aforementioned comparison tests, it is possible to determine the global stability and boundedness character for nonnegative solutions of particular rational difference equations in a range of their parameters.

1. Introduction

There has been a significant amount of work done at the University of Rhode Island pertaining to the boundedness character of rational difference equations. Recently a general boundedness result has appeared in the literature. This result proves the boundedness of solutions for many special cases of the k-th order rational difference equation [Camouzis et al. 2006, Theorem 6]. In this paper we intend to generalize this result.

Rather than working with solutions of difference equations, we intend to work with sequences which satisfy recursive inequalities, which we call difference inequalities. This approach bears relevance to the field of difference equations, as every solution to a difference equation satisfies several difference inequalities. The use of difference inequalities provides a general and efficient way to obtain bounds, attracting intervals, and convergence results for a variety of difference equations. These four theorems presented below provide the theoretical groundwork needed.

The first theorem demonstrates that the previously mentioned boundedness result extends to the framework of difference inequalities. In fact Theorem 1 demonstrates a much stronger result. Theorem 1 acts as a comparison test between difference inequalities, showing that any sequence of nonnegative real numbers which satisfies one of the assumed difference inequalities also satisfies a Riccati inequality.

MSC2000: 39A10.

Keywords: difference equation, boundedness, global stability, difference inequality.

A Riccati inequality is a difference inequality of the form

$$x_n \leq \frac{\alpha + \beta \max_{i=1,\ldots,k}(x_{n-i})}{A + B \max_{i=1,\ldots,k}(x_{n-i})}, \qquad n \geq J,$$

where J is a nonnegative integer. It is easy to see that with $A, B > 0$ and $\alpha, \beta \geq 0$ any nonnegative sequence which satisfies a Riccati inequality is bounded. Theorem 2 will demonstrate something stronger, however, namely a comparison between any nonnegative sequence which satisfies a Riccati inequality, and a solution to a particular associated Riccati equation.

Combining Theorem 1 and Theorem 2 a strong comparison is made between the solutions of certain rational difference equations and the solutions of associated Riccati equations. Using this comparison it is possible to prove global convergence results for certain rational difference equations in a range of their parameters. This global convergence result is given in Theorem 4.

2. Boundedness by iteration

Here the general theorem which proves boundedness through the method of iteration [Camouzis et al. 2006, Theorem 6] is extended to the framework of difference inequalities. Nonnegative sequences which satisfy certain difference inequalities are shown to satisfy a Riccati inequality. A direct result of this is that every solution of every rational difference equation which is bounded through the method of iteration satisfies a Ricatti inequality.

Theorem 1. *Suppose that we have a sequence of nonnegative real numbers $\{x_n\}_{n=1}^{\infty}$ which satisfies the inequality*

$$x_n \leq \frac{\alpha + \sum_{i=1}^{k} \beta_i x_{n-i}}{A + \sum_{i=1}^{k} B_i x_{n-i}}, \qquad n \geq J, \tag{1}$$

with nonnegative parameters.

Let us define the sets of indices

$$I_\beta = \{i \in \{1, 2, \ldots, k\} : \beta_i > 0\} \quad \text{and} \quad I_B = \{i \in \{1, 2, \ldots, k\} : B_i > 0\}.$$

Suppose that the following conditions hold true:

(1) $A > 0$.

(2) *There exists a positive integer η, such that for every sequence $\{c_m\}_{m=1}^{\infty}$ with $c_m \in I_\beta$, for $m = 1, 2, \ldots$, there exist positive integers, N_1, $N_2 \leq \eta$, such that $\sum_{m=N_1}^{N_2} c_m \in I_B$.*

Then $\{x_n\}_{n=1}^{\infty}$ satisfies a Riccati inequality for $n \geq J + k\eta$.

In particular, if $A \geq \sum_{i=1}^{k} \beta_i$, then, for $n \geq J + k\eta$,

$$x_n \leq \frac{\alpha\eta}{A} + \frac{\left(\sum_{i=1}^{k} \beta_i\right) \max_{i=1,\dots,k\eta} (x_{n-i})}{A + \min_{i \in I_B} (B_i) \max_{i=1,\dots,k\eta} (x_{n-i})}, \tag{2}$$

and if $A < \sum_{i=1}^{k} \beta_i$, then for $n \geq J + k\eta$,

$$x_n \leq \left(\frac{\alpha\eta}{A} + \frac{\left(\sum_{i=1}^{k} \beta_i\right) \max_{i=1,\dots,k\eta} (x_{n-i})}{A + \min_{i \in I_B} (B_i) \max_{i=1,\dots,k\eta} (x_{n-i})} \right) \left(\frac{\sum_{i=1}^{k} \beta_i}{A} \right)^{\eta-1}. \tag{3}$$

Proof. Let us consider a particular term x_N in $\{x_n\}_{n=1}^{\infty}$. Now for x_N, with $N \geq \max(J, k+1)$, let us define a finite sequence $\{c_m\}_{m=1}^{\tau}$ recursively based on x_N, $\{x_n\}_{n=1}^{\infty}$, and I_β. We will define this sequence by letting

$$c_1 = \min\left(i : x_{N-i} = \max_{\rho \in I_\beta} \left(x_{N-\rho} \right) \right), \tag{4}$$

and supposing that $c_1, \dots, c_{t-1}$ exist, and $N - \sum_{m=1}^{t-1} c_m \geq \max(J, k+1)$, and then letting

$$c_t = \min\left(i : x_{N-i-\sum_{m=1}^{t-1} c_m} = \max_{\rho \in I_\beta} \left(x_{N-\rho-\sum_{m=1}^{t-1} c_m} \right) \right).$$

Notice that this is a finite sequence, and that τ is the first integer such that $N - \sum_{m=1}^{\tau} c_m < \max(J, k+1)$. This finite sequence $\{c_m\}_{m=1}^{\tau}$ has two noteworthy properties. First it is a finite sequence $\{c_m\}_{m=1}^{\tau}$ with $c_m \in I_\beta$ for $m = 1, \dots, \tau$; second,

$$\max_{i \in I_\beta} \left(x_{N-i-\sum_{m=1}^{t-1} c_m} \right) = x_{N-\sum_{m=1}^{t} c_m}. \tag{5}$$

We will use these properties to establish bounds for the term x_N.

For the sake of notation let us define $c_0 = 0$. Now we will show by induction that when $N \geq \max(J, k+1)$, for all t such that $1 \leq t \leq \tau$, we have

$$x_N \leq \left(\frac{\alpha}{A}\right) \left(\sum_{D=0}^{t-1} \left(\frac{\sum_{i=1}^{k} \beta_i}{A} \right)^{D} \right) + \frac{\left(\sum_{i=1}^{k} \beta_i\right)^{t} \left(x_{N-\sum_{m=1}^{t} c_m} \right)}{\prod_{L=0}^{t-1} \left(A + \sum_{i=1}^{k} B_i x_{N-i-\sum_{m=0}^{L} c_m} \right)}. \tag{6}$$

First we will establish the base case

$$x_N \leq \frac{\alpha + \sum_{i=1}^{k} \beta_i x_{N-i}}{A + \sum_{i=1}^{k} B_i x_{N-i}} \leq \frac{\alpha}{A} + \frac{\left(\sum_{i=1}^{k} \beta_i\right) \max_{i \in I_\beta} (x_{N-i})}{A + \sum_{i=1}^{k} B_i x_{N-i}}.$$

Now using Equation (4) we get that

$$x_N \leq \frac{\alpha}{A} + \frac{\left(\sum_{i=1}^{k} \beta_i\right) x_{N-c_1}}{A + \sum_{i=1}^{k} B_i x_{N-i}}.$$

This is since $\max_{i \in I_\beta} (x_{N-i}) = x_{N-c_1}$, by (4). Thus (6) holds for $t = 1$. Now suppose (6) holds for $t < \tau$, we must show that it holds for $t + 1$.

$$x_N \leq \left(\frac{\alpha}{A}\right)\left(\sum_{D=0}^{t-1}\left(\frac{\sum_{i=1}^{k} \beta_i}{A}\right)^{D}\right) + \frac{\left(\sum_{i=1}^{k} \beta_i\right)^{t} \left(x_{N-\sum_{m=1}^{t} c_m}\right)}{\prod_{L=0}^{t-1}\left(A + \sum_{i=1}^{k} B_i x_{N-i-\sum_{m=0}^{L} c_m}\right)} \tag{7a}$$

$$\leq \left(\frac{\alpha}{A}\right)\left(\sum_{D=0}^{t-1}\left(\frac{\sum_{i=1}^{k} \beta_i}{A}\right)^{D}\right) + \frac{\left(\sum_{i=1}^{k} \beta_i\right)^{t} \left(\alpha + \sum_{i=1}^{k} \beta_i x_{N-i-\sum_{m=1}^{t} c_m}\right)}{\prod_{L=0}^{t}\left(A + \sum_{i=1}^{k} B_i x_{N-i-\sum_{m=0}^{L} c_m}\right)} \tag{7b}$$

$$\leq \left(\frac{\alpha}{A}\right)\left(\sum_{D=0}^{t}\left(\frac{\sum_{i=1}^{k} \beta_i}{A}\right)^{D}\right) + \frac{\left(\sum_{i=1}^{k} \beta_i\right)^{t+1} \left(\max_{i \in I_\beta}\left(x_{N-i-\sum_{m=1}^{t} c_m}\right)\right)}{\prod_{L=0}^{t}\left(A + \sum_{i=1}^{k} B_i x_{N-i-\sum_{m=0}^{L} c_m}\right)} \tag{7c}$$

$$\leq \left(\frac{\alpha}{A}\right)\left(\sum_{D=0}^{t}\left(\frac{\sum_{i=1}^{k} \beta_i}{A}\right)^{D}\right) + \frac{\left(\sum_{i=1}^{k} \beta_i\right)^{t+1} \left(x_{N-\sum_{m=1}^{t+1} c_m}\right)}{\prod_{L=0}^{t}\left(A + \sum_{i=1}^{k} B_i x_{N-i-\sum_{m=0}^{L} c_m}\right)}. \tag{7d}$$

Our induction assumption is (7a). We get (7b) from (7a) using our original inequality (1). We get (7c) from (7b), since $A > 0$ and our parameters are nonnegative. We get (7d) from (7c) since

$$\max_{i \in I_\beta}\left(x_{N-i-\sum_{m=1}^{t} c_m}\right) = x_{N-\sum_{m=1}^{t+1} c_m},$$

from (5). Thus we have shown that (6) holds for all t such that $1 \leq t \leq \tau$.

Since τ is the first integer such that $N - \sum_{m=1}^{\tau} c_m < \max(J, k+1)$, then

$$N - \max(J, k+1) < \sum_{m=1}^{\tau} c_m < k\tau.$$

Thus $\tau > (N - \max(J, k+1))/k$. So if we choose $N \geq J + k\eta$, where η is the integer η defined in Condition (2), then $\tau \geq \eta$. We know there exist positive integers $N_1, N_2 \leq \eta$, so that $\sum_{m=N_1}^{N_2} c_m \in I_B$; this is from Condition (2) in our original assumptions. Thus

$$\sum_{m=1}^{N_2} c_m = \sum_{m=0}^{N_1-1} c_m + i,$$

for some $i \in I_B$. Since $N_2 \leq \eta \leq \tau$, by Equation (6)

$$x_N \leq \left(\frac{\alpha}{A}\right)\left(\sum_{D=0}^{N_2-1}\left(\frac{\sum_{i=1}^{k}\beta_i}{A}\right)^{D}\right)+\frac{\left(\sum_{i=1}^{k}\beta_i\right)^{N_2}\left(x_{N-\sum_{m=1}^{N_2}c_m}\right)}{\prod_{L=0}^{N_2-1}\left(A+\sum_{i=1}^{k}B_i x_{N-i-\sum_{m=0}^{L}c_m}\right)} \tag{8a}$$

$$\leq \left(\frac{\alpha}{A}\right)\left(\sum_{D=0}^{N_2-1}\left(\frac{\sum_{i=1}^{k}\beta_i}{A}\right)^{D}\right)+\frac{\left(\sum_{i=1}^{k}\beta_i\right)^{N_2}\left(x_{N-\sum_{m=1}^{N_2}c_m}\right)}{A^{N_2-1}\left(A+\sum_{i=1}^{k}B_i x_{N-i-\sum_{m=0}^{N_1-1}c_m}\right)} \tag{8b}$$

$$\leq \left(\frac{\alpha}{A}\right)\left(\sum_{D=0}^{N_2-1}\left(\frac{\sum_{i=1}^{k}\beta_i}{A}\right)^{D}\right)+\frac{\left(\sum_{i=1}^{k}\beta_i\right)^{N_2}\left(x_{N-\sum_{m=1}^{N_2}c_m}\right)}{A^{N_2-1}\left(A+\left(\min_{i\in I_B}(B_i)\right)\left(x_{N-\sum_{m=1}^{N_2}c_m}\right)\right)}. \tag{8c}$$

We get (8a) directly from (6) with $t = N_2$. We get (8b) from (8a), since $A > 0$ and our parameters are nonnegative. This expression is obtained by reducing all of the terms of the product in the denominator of this fraction, except for the term where $L = N_1 - 1$, which is kept as it is needed to establish a bound. We get (8c) from (8b) since

$$\sum_{m=1}^{N_2} c_m = \sum_{m=0}^{N_1-1} c_m + i,$$

for some $i \in I_B$. Now we will consider two cases, namely

$$A \geq \sum_{i=1}^{k}\beta_i \quad \text{and} \quad A < \sum_{i=1}^{k}\beta_i.$$

Considering the former case, since $1 \leq N_2 \leq \eta$, we have that,

$$x_N \leq \left(\frac{\alpha}{A}\right)\left(\sum_{D=0}^{N_2-1}\left(\frac{\sum_{i=1}^{k}\beta_i}{A}\right)^{D}\right)+\frac{\left(\sum_{i=1}^{k}\beta_i\right)^{N_2}\left(x_{N-\sum_{m=1}^{N_2}c_m}\right)}{A^{N_2-1}\left(A+\left(\min_{i\in I_B}(B_i)\right)\left(x_{N-\sum_{m=1}^{N_2}c_m}\right)\right)} \tag{9a}$$

$$\leq \frac{\alpha\eta}{A}+\frac{\left(\sum_{i=1}^{k}\beta_i\right)x_{N-\sum_{m=1}^{N_2}c_m}}{A+\left(\min_{i\in I_B}(B_i)\right)x_{N-\sum_{m=1}^{N_2}c_m}}. \tag{9b}$$

Notice that our bound in (9b) is increasing with respect to $x_{N-\sum_{m=1}^{N_2}c_m}$, and that $1 \leq \sum_{m=1}^{N_2} c_m \leq k\eta$; thus, by (9b),

$$x_N \leq \frac{\alpha\eta}{A}+\frac{\left(\sum_{i=1}^{k}\beta_i\right)\max_{i=1,\ldots,k\eta}(x_{N-i})}{A+\left(\min_{i\in I_B}(B_i)\right)\max_{i=1,\ldots,k\eta}(x_{N-i})},$$

for all $N \geq J + k\eta$. Thus we have shown the inequality (2).

If $A < \sum_{i=1}^{k} \beta_i$, then, since $1 \leq N_2 \leq \eta$, we have,

$$x_N \leq \left(\frac{\alpha}{A}\right)\left(\sum_{D=0}^{N_2-1}\left(\frac{\sum_{i=1}^{k}\beta_i}{A}\right)^{D}\right)+\frac{\left(\sum_{i=1}^{k}\beta_i\right)^{N_2}\left(x_{N-\sum_{m=1}^{N_2}c_m}\right)}{A^{N_2-1}\left(A+\left(\min_{i\in I_B}(B_i)\right)\left(x_{N-\sum_{m=1}^{N_2}c_m}\right)\right)} \tag{10a}$$

$$\leq \left(\frac{\alpha\eta}{A}+\frac{\left(\sum_{i=1}^{k}\beta_i\right)x_{N-\sum_{m=1}^{N_2}c_m}}{A+\left(\min_{i\in I_B}(B_i)\right)x_{N-\sum_{m=1}^{N_2}c_m}}\right)\left(\frac{\sum_{i=1}^{k}\beta_i}{A}\right)^{\eta-1}. \tag{10b}$$

Notice that our bound in Equation (10b) is increasing with respect to $x_{N-\sum_{m=1}^{N_2}c_m}$, and that $1 \leq \sum_{m=1}^{N_2} c_m \leq k\eta$. Thus, by (10b),

$$x_n \leq \left(\frac{\alpha\eta}{A}+\frac{\left(\sum_{i=1}^{k}\beta_i\right)\max_{i=1,\dots,k\eta}(x_{n-i})}{A+\left(\min_{i\in I_B}(B_i)\right)\max_{i=1,\dots,k\eta}(x_{n-i})}\right)\left(\frac{\sum_{i=1}^{k}\beta_i}{A}\right)^{\eta-1},$$

for all $N \geq J + k\eta$. Thus we have shown the inequality (3) and the theorem is proved. □

Theorem 1 immediately establishes the boundedness character for a number of special cases of the k-th order rational difference equation. These boundedness results were completely established in [Camouzis et al. 2006]. For related works, see [Kocić and Ladas 1993; Kulenović and Ladas 2002; Camouzis et al. 2004a; 2004b; 2005a; 2005b; 2006; Ladas 2004; Camouzis and Ladas 2005; Grove and Ladas 2005; Camouzis 2006].

Since Theorem 1 only assumes that the inequality (1) eventually holds for our sequence $\{x_n\}_{n=1}^{\infty}$, it is also possible to quickly establish the boundedness character for several nonautonomous rational difference equations.

3. Comparison tests of the maximum and minimum

The following two theorems deal with comparison tests involving the maximum and minimum. One important consequence of these tests is that when combined with Theorem 1 they allow for the comparison between solutions of certain special cases of the k-th order rational difference equation and solutions of a Riccati type difference equation.

Theorem 2. *Let $g : [0, \infty) \to [0, \infty)$ be defined and increasing for all $x \in [0, \infty)$.*

Suppose that we have a sequence of nonnegative real numbers $\{x_n\}_{n=1}^{\infty}$ which satisfies the inequality, $x_n \leq g(\max(x_{n-1}, \dots, x_{n-k}))$, with $n \geq N$. Let $\{y_n\}_{n=0}^{\infty}$ be a solution of the difference equation $y_n = g(y_{n-1})$, given $n = 1, 2, \dots,$ and with $y_0 = \max(x_{N-1}, \dots, x_{N-k})$, then, for all $n \geq N$,

$$\max(x_{n-1}, \dots, x_{n-k}) \leq \max(y_{\lfloor \frac{n-N}{k} \rfloor}, \dots, y_{n-N}). \tag{11}$$

Proof. This result follows by strong induction. From our assumptions we have that $\max(x_{N-1}, \dots, x_{N-k}) = y_0$. This establishes the base case for $n = N$. Now suppose that

$$\max\left(x_{n-1}, \dots, x_{n-k}\right) \leq \max\left(y_{\lfloor \frac{n-N}{k} \rfloor}, \dots, y_{n-N}\right),$$

for all $N \leq n < J$. Then, for all $N \leq n < J$,

$$x_n \leq g\left(\max\left(x_{n-1}, \dots, x_{n-k}\right)\right) \leq g\left(\max\left(y_{\lfloor \frac{n-N}{k} \rfloor}, \dots, y_{n-N}\right)\right).$$

Since g is defined and increasing for all $x \in [0, \infty)$,

$$x_n \leq \max\left(g\left(y_{\lfloor \frac{n-N}{k} \rfloor}\right), \dots, g\left(y_{n-N}\right)\right) = \max\left(y_{1+\lfloor \frac{n-N}{k} \rfloor}, \dots, y_{n+1-N}\right).$$

From this it follows that,

$$\max\left(x_{J-1}, \dots, x_{J-k}\right) \leq \max_{n=J-1,\dots,J-k}\left(\max\left(y_{1+\lfloor \frac{n-N}{k} \rfloor}, \dots, y_{n+1-N}\right)\right).$$

Thus,

$$\max\left(x_{J-1}, \dots, x_{J-k}\right) \leq \max\left(y_{1+\lfloor \frac{J-k-N}{k} \rfloor}, \dots, y_{J-N}\right) = \max\left(y_{\lfloor \frac{J-N}{k} \rfloor}, \dots, y_{J-N}\right).$$

This proves that Equation (11) holds for J, and completes the proof by induction. □

Theorem 3. *Let $g : [0, \infty) \to [0, \infty)$ be defined and increasing for all $x \in [0, \infty)$.*

Suppose that we have a sequence of nonnegative real numbers $\{x_n\}_{n=1}^{\infty}$ which satisfies the inequality, $x_n \geq g(\min(x_{n-1}, \dots, x_{n-k}))$ with $n \geq N$. Let $\{y_n\}_{n=0}^{\infty}$ be a solution of the difference equation $y_n = g(y_{n-1})$, with $n = 1, 2, \dots,$ and with $y_0 = \min(x_{N-1}, \dots, x_{N-k})$, then for all $n \geq N$,

$$\min\left(x_{n-1}, \dots, x_{n-k}\right) \geq \min\left(y_{\lfloor \frac{n-N}{k} \rfloor}, \dots, y_{n-N}\right). \tag{12}$$

Proof. This result follows by strong induction. From our assumptions we have that $\min(x_{N-1}, \dots, x_{N-k}) = y_0$. This establishes the base case for $n = N$. Now suppose that

$$\min\left(x_{n-1}, \dots, x_{n-k}\right) \geq \min\left(y_{\lfloor \frac{n-N}{k} \rfloor}, \dots, y_{n-N}\right),$$

for all $N \leq n < J$. Then, for all $N \leq n < J$,

$$x_n \geq g\left(\min\left(x_{n-1}, \dots, x_{n-k}\right)\right) \geq g\left(\min\left(y_{\lfloor \frac{n-N}{k} \rfloor}, \dots, y_{n-N}\right)\right).$$

Since g is defined and increasing for all $x \in [0, \infty)$,

$$x_n \geq \min\left(g\left(y_{\lfloor \frac{n-N}{k} \rfloor}\right), \dots, g\left(y_{n-N}\right)\right) = \min\left(y_{1+\lfloor \frac{n-N}{k} \rfloor}, \dots, y_{n+1-N}\right).$$

From this it follows that

$$\min\left(x_{J-1},\ldots,x_{J-k}\right) \geq \min_{n=J-1,\ldots,J-k}\left(\min\left(y_{1+\lfloor\frac{n-N}{k}\rfloor},\ldots,y_{n+1-N}\right)\right).$$

Thus,

$$\min\left(x_{J-1},\ldots,x_{J-k}\right) \geq \min\left(y_{1+\lfloor\frac{J-k-N}{k}\rfloor},\ldots,y_{J-N}\right) = \min\left(y_{\lfloor\frac{J-N}{k}\rfloor},\ldots,y_{J-N}\right).$$

This proves that Equation (12) holds for J, and completes the proof by induction. □

Theorem 2 and its dual Theorem 3 provide a general and useful method for obtaining tighter bounds on both the solutions of difference equations and sequences which satisfy difference inequalities. Indeed using Theorem 2 it is sometimes possible to obtain upper bounds for the solutions of certain difference equations which are arbitrarily close to an equilibrium. The discovery of such bounds coupled with a thorough understanding of semicycle analysis may yield some interesting convergence results. We will leave this idea for future investigation.

4. A convergence result for difference inequalities

Here we will give one example which demonstrates convergence even in the framework of difference inequalities. The convergence result here also settles an open problem in rational difference equations in the case $A = \sum_{i=1}^{k}\beta_i$.

Theorem 4. *Suppose that we have a sequence of nonnegative real numbers* $\{x_n\}_{n=1}^{\infty}$ *which satisfies the inequality,*

$$x_n \leq \frac{\sum_{i=1}^{k}\beta_i x_{n-i}}{A+\sum_{i=1}^{k}B_i x_{n-i}}, \qquad n \geq J,$$

with nonnegative parameters.

Let us define the sets of indices

$$I_\beta = \{i \in \{1,2,\ldots,k\} : \beta_i > 0\} \quad \textit{and} \quad I_B = \{i \in \{1,2,\ldots,k\} : B_i > 0\}.$$

Suppose that the following conditions hold true:

(1) $A \geq \sum_{i=1}^{k}\beta_i$.

(2) *There exists a positive integer* η, *such that for every sequence* $\{c_m\}_{m=1}^{\infty}$ *with* $c_m \in I_\beta$ *for* $m = 1, 2, \ldots$ *there exists positive integers*, $N_1, N_2 \leq \eta$, *such that* $\sum_{m=N_1}^{N_2} c_m \in I_B$.

Then $\{x_n\}_{n=1}^{\infty}$ *converges to* 0.

Proof. By Theorem 1, for $n \geq J + k\eta$,

$$x_n \leq \frac{\left(\sum_{i=1}^{k} \beta_i\right) \max_{i=1,\ldots,k\eta} (x_{n-i})}{A + \left(\min_{i \in I_B} (B_i)\right) \max_{i=1,\ldots,k\eta} (x_{n-i})}.$$

Dividing the numerator and denominator by $\sum_{i=1}^{k} \beta_i$, we may rewrite the inequality in the form,

$$x_n \leq \frac{\max_{i=1,\ldots,k\eta} (x_{n-i})}{\rho + C \max_{i=1,\ldots,k\eta} (x_{n-i})},$$

where $\rho \geq 1$ and $C > 0$. Applying Theorem 2 we get that for $\{y_n\}_{n=0}^{\infty}$, a solution of the difference equation,

$$y_n = \frac{y_{n-1}}{\rho + C y_{n-1}}, \qquad n = 1, 2, \ldots, \tag{13}$$

with $y_0 = \max\left(x_{J+k\eta-1}, \ldots, x_{J+k\eta-k}\right)$, then for all $n \geq J + k\eta$,

$$\max(x_{n-1}, \ldots, x_{n-k}) \leq \max\left(y_{\lfloor \frac{n-J-k\eta}{k} \rfloor}, \ldots, y_{n-J-k\eta}\right).$$

Since $\{y_n\}_{n=0}^{\infty}$ is decreasing and bounded below by zero, $\{y_n\}_{n=0}^{\infty}$ converges. Since the only equilibrium of equation Equation (13) is zero, $\{y_n\}_{n=0}^{\infty}$ converges to zero.

Since $\{y_n\}_{n=0}^{\infty}$ converges to zero, given $\epsilon > 0$, there exists a natural number N sufficiently large so that $y_n < \epsilon$ for all $n \geq N$. Choose D to be a natural number so that $N = \lfloor (D - J - k\eta) / k \rfloor$. Then, for $n \geq D$,

$$x_{n-1} \leq \max(x_{n-1}, \ldots, x_{n-k}) \leq \max\left(y_{\lfloor \frac{n-J-k\eta}{k} \rfloor}, \ldots, y_{n-J-k\eta}\right) < \max(\epsilon, \ldots, \epsilon) = \epsilon.$$

Thus, given $\epsilon > 0$, there exists a natural number D sufficiently large so that $x_n < \epsilon$ for all $n \geq D$. Therefore $\{x_n\}_{n=1}^{\infty}$ converges to 0. □

References

[Camouzis 2006] E. Camouzis, "On the boundedness of some rational difference equations", *J. Difference Equ. Appl.* **12**:1 (2006), 69–94. MR 2006h:39005 Zbl 05016366

[Camouzis and Ladas 2005] E. Camouzis and G. Ladas, "On third-order rational difference equations. V", *J. Difference Equ. Appl.* **11**:6 (2005), 553–562. MR 2152557 Zbl 02198768

[Camouzis et al. 2004a] E. Camouzis, E. Chatterjee, G. Ladas, and E. P. Quinn, "On third-order rational difference equations. III", *J. Difference Equ. Appl.* **10**:12 (2004), 1119–1127. MR 2097116 Zbl 1055.39500

[Camouzis et al. 2004b] E. Camouzis, G. Ladas, and E. P. Quinn, "On third-order rational difference equations. II", *J. Difference Equ. Appl.* **10**:11 (2004), 1041–1047. MR 2082688

[Camouzis et al. 2005a] E. Camouzis, E. Chatterjee, G. Ladas, and E. P. Quinn, "Progress report on the boundedness character of third-order rational equations", *J. Difference Equ. Appl.* **11**:11 (2005), 1029–1035. MR 2174113 Zbl 02229066

[Camouzis et al. 2005b] E. Camouzis, G. Ladas, and E. P. Quinn, "On third-order rational difference equations. VI", *J. Difference Equ. Appl.* **11**:8 (2005), 759–777. MR 2156652 Zbl 1071.39502

[Camouzis et al. 2006] E. Camouzis, G. Ladas, F. Palladino, and E. P. Quinn, "On the boundedness character of rational equations. I", *J. Difference Equ. Appl.* **12**:5 (2006), 503–523. MR 2241391 Zbl 05040851

[Grove and Ladas 2005] E. A. Grove and G. Ladas, *Periodicities in nonlinear difference equations*, vol. 4, Advances in Discrete Mathematics and Applications, Chapman & Hall/CRC, Boca Raton, FL, 2005. MR 2006j:39002 Zbl 1078.39009

[Kocić and Ladas 1993] V. L. Kocić and G. Ladas, *Global behavior of nonlinear difference equations of higher order with applications*, vol. 256, Mathematics and its Applications, Kluwer Academic Publishers Group, Dordrecht, 1993. MR 94k:39005

[Kulenović and Ladas 2002] M. R. S. Kulenović and G. Ladas, *Dynamics of second order rational difference equations*, Chapman & Hall/CRC, Boca Raton, FL, 2002. With open problems and conjectures. MR 2004c:39001

[Ladas 2004] G. Ladas, "On third-order rational difference equations. I", *J. Difference Equ. Appl.* **10**:9 (2004), 869–879. MR 2074438

Received: 2007-10-28 Accepted: 2007-10-28

fpalladino@mail.uri.edu

Department of Mathematics, University of Rhode Island,
Kingston, RI 02881-0816, United States

INVOLVE 1:1(2008)

On the asymptotic behavior of unions of sets of lengths in atomic monoids

Paul Baginski, Scott Thomas Chapman, Natalie Hine and João Paixão

(Communicated by Kenneth S. Berenhaut)

Let M be a commutative cancellative atomic monoid. We use unions of sets of lengths in M to construct the $\mathcal{V}$-Delta set of M. We first derive some basic properties of $\mathcal{V}$-Delta sets and then show how they offer a method to investigate the asymptotic behavior of the sizes of unions of sets of lengths.

A central focus of number theory is the study of number theoretic functions and their asymptotic behavior. This has led to similar investigations concerning nonunique factorizations in integral domains and moniods. Suppose that M is a commutative cancellative monoid in which each nonunit can be factored into a product of irreducible elements (such a monoid is known as *atomic*). For a nonunit x in M, let $L(x)$ represent the maximum length of a factorization of x into irreducibles and $l(x)$ the minimum such length. The functions

$$\overline{L}(x) = \lim_{k\to\infty} \frac{L(x^n)}{n} \qquad \text{and} \qquad \overline{l}(x) = \lim_{k\to\infty} \frac{l(x^n)}{n}$$

have been studied in the literature by Anderson and Pruis [1991] and Geroldinger and Halter-Koch [1992]. Chapman and Smith [1998] defined the notion of a *generalized set of lengths*, and showed [Chapman and Smith 1993b] that the size of a generalized set of lengths (denoted $\Phi(n)$) satisfies

$$\overline{\Phi}(R) = \lim_{n\to\infty} \frac{\Phi(n)}{n} = \frac{D(G)^2 - 4}{2D(G)}, \tag{1}$$

for a ring of algebraic integers R where $D(G)$ represents Davenport's constant of the ideal class group G of R (the Davenport constant is defined in [Geroldinger and Halter-Koch 2006, Section 3.4]). Since a generalized set of lengths is actually

MSC2000: primary 20M14; secondary 20D60, 11B75.

Keywords: nonunique factorization, elasticity of factorization, unions of sets of lengths.

P. B. was supported by a Department of Homeland Security Graduate Fellowship. This work was completed during the 2006 Trinity University Research Experiences for Undergraduates in Mathematics Program. J. P. and N. H. were supported by the National Science Foundation, Grant no. DMS-0353488.

a union of certain length sets, we will refer to these sets with the more descriptive term *unions of sets of lengths*. The value $\overline{\Phi}(R)$ has also been explored for various semigroup rings over fields [Anderson et al. 1993, Theorem 3.3]. In this note, we examine the limit $\overline{\Phi}(R)$ in greater detail. By generalizing the well known notion of the Delta set of a monoid M [Geroldinger and Halter-Koch 2006, Section 1.4], we find new bounds for the value $\overline{\Phi}(M)$ which allows us to determine exact calculations in several instances recently addressed in the literature (see Examples 3 and 4). We will begin with a review of the necessary definitions and notations from the theory of nonunique factorizations. The reader is directed to the monograph [Geroldinger and Halter-Koch 2006] for a complete survey of recent results in this area.

Throughout our work, we assume that M is an atomic commutative cancellative monoid with sets $\mathcal{I}(M)$ of irreducible elements and $M^\bullet$ of nonunits. The set of lengths of $x \in M^\bullet$ is $\mathcal{L}(x) = \{n \mid x = x_1 \cdots x_n$ with each $x_i \in \mathcal{I}(M)\}$. Also, define $L(x) = \max \mathcal{L}(x)$ and $l(x) = \min \mathcal{L}(x)$. The quotient $L(x)/l(x)$ is called the *elasticity* of x and the constant $\rho(M) = \sup \left\{ \frac{L(x)}{l(x)} \mid x \in M^\bullet \right\}$ is known as the *elasticity* of M. A survey of the results in the literature concerning elasticity can be found in [Anderson 1997]. If $\mathcal{L}(x) = \{n_1, \ldots, n_t\}$ with the n_i's listed in increasing order, then the Delta set of x is $\Delta(x) = \{n_i - n_{i-1} \mid 2 \le i \le t\}$. The Delta set of M is then defined as $\Delta(M) = \cup_{x \in M^\bullet} \Delta(x)$. If $d = \gcd \Delta(M)$, Geroldinger [1988, Proposition 4] has shown that $d \in \Delta(M)$. Hence, it follows that

$$\{d, qd\} \subseteq \Delta(M) \subseteq \{d, 2d, \ldots, qd\}, \tag{2}$$

for some positive integer q. While the concept of the Delta set of a monoid M has been widely studied, there are few exact computations of specific Delta sets in the literature. If $\mathcal{B}(\mathbb{Z}_n)$ represents the block monoid ([Geroldinger and Halter-Koch 2006] or Example 2) on the cyclic group of order n, then

$$\Delta(\mathcal{B}(\mathbb{Z}_n)) = \{1, 2, \ldots, n-2\}$$

[Geroldinger and Halter-Koch 2006, Theorem 6.7.1]. The Delta sets of several numerical monoids [Bowles et al. 2006] and several congruence monoids [Baginski et al. 2008] have been computed under restricted conditions. In particular, an example is constructed in [Bowles et al. 2006, Proposition 4.9] where both containments in Equation (2) are strict.

The notion of a set of lengths was generalized in [Chapman and Smith 1998] as follows: With M as above, for each $n \in \mathbb{N}$ set $\mathcal{W}(n) = \{m \in M \mid n \in \mathcal{L}(m)\}$ and

$$\mathcal{V}(n) = \bigcup_{m \in \mathcal{W}(n)} \mathcal{L}(m).$$

We refer to the set $\mathcal{V}(n)$ as a *union of sets of lengths*. In [Chapman and Smith 1998], the basic properties of these sets are determined. Moreover, for block monoids $\mathcal{B}(G)$ where G is a finite abelian group, the authors argue that the sequence $\{\mathcal{V}(n)\}_{n=1}^{\infty}$ does not uniquely characterize G. We will often need to refer to the maximum and minimum values in $\mathcal{V}(n)$. Hence for each $n \in \mathbb{N}$ we set

$$\lambda_n(M) = \min \mathcal{V}(n) \quad \text{and} \quad \rho_n(M) = \sup \mathcal{V}(n).$$

When the monoid M is understood, we will merely use the notation λ_n and ρ_n. The sequence $\{\rho_n\}_{n=1}^{\infty}$ has been an object of study in its own right [Geroldinger and Halter-Koch 2006, Section 1.4] and [Geroldinger and Hassler $\geq$ 2008] and it is shown in [Geroldinger and Halter-Koch 2006, Proposition 1.4.2] that

$$\rho(M) = \lim_{n\to\infty} \frac{\rho_n(M)}{n}.$$

Finally, for each $n \in \mathbb{N}$, set $\Phi(n) = |\mathcal{V}(n)|$. Some basic properties of the Φ-function are explored in [Chapman and Smith 1990, Section 2] and several additional computations of the limit

$$\overline{\Phi}(M) = \lim_{n\to\infty} \frac{\Phi(n)}{n}$$

can be found in the literature [Chapman and Smith 1993a, Theorem 2.7 and Theorem 2.10].

For our purposes, we extend the notion of the Delta set to unions of sets of lengths as follows: For a fixed monoid M, suppose for each $n \in \mathbb{N}$ that

$$\mathcal{V}(n) = \{v_{1,n}, \ldots, v_{t,n}\},$$

where $v_{i,n} < v_{i+1,n}$ for $1 \leq i < t$. Define the $\mathcal{V}(n)$-Delta set of M to be

$$\Delta\mathcal{V}(n) = \{v_{i,n} - v_{i-1,n} \mid 2 \leq i \leq t\}$$

and the $\mathcal{V}$-Delta set of M to be

$$\Delta_{\mathcal{V}}(M) = \bigcup_{n\in\mathbb{N}} \Delta\big(\mathcal{V}(n)\big).$$

In addition, set $\mathcal{V}^*(M) = \sup \Delta_{\mathcal{V}}(M)$ and $\mathcal{V}_*(M) = \min \Delta_{\mathcal{V}}(M)$. Clearly,

$$\Delta\mathcal{V}(1) = \varnothing.$$

Example 1. Let $\mathbb{N}_0$ represent the nonnegative integers. Consider the additive submonoid $M = \{(x_1, x_2, x_3) \mid x_1 + 3x_2 = 4x_3 \text{ with each } x_i \in \mathbb{N}_0\}$ of $\mathbb{N}_0^3$. Such a monoid is known as a *Diophantine monoid* [Chapman et al. 2002]. A characterization of Diophantine monoids can be found in [Geroldinger and Halter-Koch 2006, Theorem 2.7.14]. It follows from [Chapman et al. 2000, Proposition 4.8], that $\Delta(M) = \{2\}$. Using elementary number theory, it follows that the irreducible

	λ_n	ρ_n
$n \equiv 0 \pmod 4$	$2\lfloor \frac{n}{4} \rfloor$	2n
$n \equiv 1 \pmod 4$	$2\lfloor \frac{n-1}{4} \rfloor$ +1	2n-1
$n \equiv 2 \pmod 4$	$2\lfloor \frac{n}{4} \rfloor$ + 2	2n
$n \equiv 3 \pmod 4$	$2\lfloor \frac{n-1}{4} \rfloor$ + 3	2n-1

Table 1. Example 1: values for λ_n and ρ_n for n = 0, 1, 2, 3.

elements of M are $v_1 = (4, 0, 1)$, $v_2 = (0, 4, 3)$ and $v_3 = (1, 1, 1)$. The following two facts will be key in determining $\Delta_{\mathcal{V}}(M)$:

- using the relation $v_1 + v_2 = 4v_3$, it is clear that an irreducible factorization in M which contains both v_1 and v_2 can be increased in length by 2;
- by [Chapman and Smith 1993a, Lemma 2.8], if a and b are in $\mathcal{V}(n)$, then $a \equiv b \pmod 2$.

By observing that λ_n is obtained by factoring nv_3 and ρ_n by factoring $2nv_3$, if n is even or $(2n-1)v_3$, if n is odd, we obtain the values given in Table 1. We list the first few values of $\mathcal{V}(n)$ below:

$$\begin{aligned} &\mathcal{V}(1) = \{1\}, && \mathcal{V}(5) = \{3, 5, 7, 9\}, \\ &\mathcal{V}(2) = \{2, 4\}, && \mathcal{V}(6) = \{4, 6, 8, 10, 12\}, \\ &\mathcal{V}(3) = \{3, 5\}, && \mathcal{V}(7) = \{5, 7, 9, 11, 13\}, \\ &\mathcal{V}(4) = \{2, 4, 6, 8\}, && \mathcal{V}(8) = \{4, 6, 8, 10, 12, 14, 16\}. \end{aligned}$$

We have that $\Delta\big(\mathcal{V}(n)\big) = \{2\}$ for all n and hence $\Delta_{\mathcal{V}}(M) = \{2\}$. Notice here that $\Delta_{\mathcal{V}}(M) = \Delta(M)$. □

Example 2. Let G be an abelian group and $\mathcal{F}(G)$ represent the free abelian monoid on G. Set

$$\mathcal{B}(G) = \left\{ \prod_{g_i \in G} g_i^{n_i} \mid \sum_{g_i \in G} n_i g_i = 0 \right\}.$$

$\mathcal{B}(G)$ is a submonoid of $\mathcal{F}(G)$ known as the *block monoid* on G. Its irreducible elements are known as *minimal zero-sequences*. Using the results of [Chapman and Smith 1998], we can write out the unions of sets of lengths, and in turn the $\mathcal{V}(n)$-Delta sets of block monoids on relatively simple groups. For instance, if

$G = \mathbb{Z}_5$, then [Chapman and Smith 1998, Example 5.4] yields:

$$\begin{aligned} &\rho_n = \lfloor \frac{5n}{2} \rfloor && \text{for } n \geq 2, \\ &\lambda_1 = 1,\ \lambda_k = 2 && \text{for } k = 2, 3, 4, 5, \\ &\lambda_k = \lambda_{(k-5)} + 2 && \text{for } k \geq 6, \end{aligned}$$

for all $n \geq 1$, $\mathcal{V}(n) = [\lambda_n, \rho_n] \cap \mathbb{Z}$ Hence, $\Delta\mathcal{V}(n) = \{1\}$ for each $n > 1$ in $\mathbb{N}$ and thus $\Delta_{\mathcal{V}}(\mathcal{B}(\mathbb{Z}_5)) = \{1\}$. Notice that our previous remark yields that $\Delta(\mathcal{B}(\mathbb{Z}_5)) = \{1, 2, 3\}$. □

We consider some basic properties of the $\mathcal{V}$-Delta set of M in the following lemma.

Lemma 1. *Let M be an atomic monoid with* $\min \Delta(M) = d$ *and* $\max \Delta(M) = qd$ *for $q \geq 1$.*

(1) $\mathcal{V}_*(M) = d$.

(2) $\mathcal{V}^*(M) \leq qd$.

(3) $\{d\} \subseteq \Delta_{\mathcal{V}}(M) \subseteq \{d, 2d, \ldots, qd\}$.

Proof. Choose $n \in \mathbb{N}$ and let $v_{i+1,n}$, $v_{i,n}$ be in $\mathcal{V}(n)$. We may choose x_1 and x_2 in $M^\bullet$ such that $\{n, v_{i+1,n}\} \subseteq \mathcal{L}(x_1)$ and $\{n, v_{i,n}\} \subseteq \mathcal{L}(x_2)$. By Equation (2), $\mathcal{L}(x_1)$ is a subset of $n + d\mathbb{Z}$ which contains n and whose consecutive elements are at most qd apart. The same statement holds for $\mathcal{L}(x_2)$, therefore the union, $\mathcal{L}(x_1) \cup \mathcal{L}(x_2)$, also possesses all these properties. Note that the union is a subset of $\mathcal{V}(n)$, so since $v_{i+1,n}$ and $v_{i,n}$ are consecutive elements of $\mathcal{V}(n)$, they in particular must be consecutive elements of $\mathcal{L}(x_1) \cup \mathcal{L}(x_2)$. Therefore $v_{i+1,n} - v_{i,n} = td$ for some $1 \leq t \leq q$. This shows that $\Delta\mathcal{V}(n) \subseteq \{d, 2d, \ldots, qd\}$, which in turn implies (2) and (3). It also determines that $\mathcal{V}_*(M) \geq d$, so we are left with just showing $d \in \Delta_{\mathcal{V}}(M)$.

Since $d \in \Delta(M)$, there is an $x \in M$ and $l_1, l_2 \in \mathcal{L}(x)$ with $l_2 - l_1 = d$. Consider $\mathcal{V}(l_1)$, to which both l_1 and l_2 belong. They must be consecutive elements of $\mathcal{V}(l_1)$ since we have just shown that consecutive elements are at least d apart. Hence $d \in \Delta(\mathcal{V}(l_1)) \subset \Delta_{\mathcal{V}}(M)$. □

Note that Example 2 indicates that the inequality in Lemma 1 regarding $\mathcal{V}^*(M)$ may be strict. The next corollary will later be useful and follows immediately from Lemma 1.

Corollary 1. *If $\Delta(M) = \{d\}$, then $\Delta_{\mathcal{V}}(M) = \{d\}$.*

We apply the $\mathcal{V}$-Delta set to limits of the form Equation (1). Unlike the $\overline{L}(x)$ and $\overline{l}(x)$ functions, there is no known argument that $\overline{\Phi}(M)$ exists for a general atomic monoid M. Hence, our analysis of Equation (1) will involve the use of lim inf and lim sup. Moreover, we must assume that $\Phi(n)$ is finite for all n, since this is necessary for $\limsup_{n\to\infty}$ to be finite. Indeed, if $\Phi(n)$ were infinite for some n, then so would be $\Phi(kn)$ for all k: if x has a factorization of length n and of length m, then x^k has factorizations of lengths kn and km. In [Chapman and Smith 1990], an atomic monoid which statisfies $\Phi(n) < \infty$ for all nonnegative n is called Φ*-finite*.

Our main theorem will use the stronger hypothesis that M has finite elasticity. The following proposition shows this is a necessary condition for

$$\limsup_{n\to\infty} \frac{\Phi(n)}{n}$$

to be finite, and the main theorem shows that it is sufficient as well.

Proposition 1. *Let M be an atomic Φ-finite monoid. If $\rho(M) = \infty$, then*

$$\limsup_{n\to\infty} \frac{\Phi(n)}{n} = \infty.$$

Proof.

Since $\rho(M) = \infty$, there are x_t such that $a_t = L(x_t)$ and $b_t = l(x_t)$ satisfying

$$\lim_{t\to\infty} \frac{a_t}{b_t} = \infty.$$

But all the $\mathcal{V}(n)$ are finite and $a_t \in \mathcal{V}(b_t)$, implying that for every $M > 0$ there is an $N > 0$ such that for all $t > N$, $b_t > M$. Therefore we may assume that the sequence is chosen such that the b_t are strictly increasing.

Since $\Phi(n)$ is finite for each n, $\mathcal{V}^*(b_t)$ exists and $\mathcal{V}^*(b_t) \geq a_t$. Pruning the sequence if necessary, we may assume that the b_t are chosen such that

$$\lim_{t\to\infty} \frac{\mathcal{V}^*(b_t)}{b_t} = \infty.$$

We may estimate

$$\Phi(b_t) \geq \frac{\mathcal{V}^*(b_t) - \mathcal{V}_*(b_t) + 1}{qd}.$$

Since $\mathcal{V}_*(b_t) \leq b_t$, we find that

$$\frac{\Phi(b_t)}{b_t} \geq \frac{\mathcal{V}^*(b_t)}{b_t qd} - \frac{1}{qd} + \frac{1}{b_t qd}.$$

Taking $\liminf$ of both sides, we see that

$$\liminf_{t\to\infty} \frac{\Phi(b_t)}{b_t} \geq \infty,$$

since the b_t are strictly increasing. Therefore

$$\limsup_{n\to\infty} \frac{\Phi(n)}{n} = \infty. \qquad \square$$

Now our main theorem:

Theorem 1. *Let M be an atomic monoid with $\rho(M) < \infty$. Then M is Φ-finite and moreover*

$$\frac{\rho(M)^2-1}{\rho(M)\mathcal{V}^*(M)} \le \liminf_{n\to\infty} \frac{\Phi(n)}{n} \le \limsup_{n\to\infty} \frac{\Phi(n)}{n} \le \frac{\rho(M)^2-1}{\rho(M)\mathcal{V}_*(M)}. \tag{3}$$

Proof. Let $n \in \mathbb{N}$ and suppose that $m \in \mathcal{V}(n)$. It follows that

$$\frac{1}{\rho(M)} \le \frac{m}{n} \le \rho(M)$$

and hence

$$\frac{n}{\rho(M)} \le m \le n\rho(M),$$

which shows that M is Φ-finite. We further obtain that

$$\frac{\left(\rho(M)-\frac{1}{\rho(M)}\right)n+1}{\mathcal{V}^*(M)} \le \Phi(n) \le \frac{\left(\rho(M)-\frac{1}{\rho(M)}\right)n+1}{\mathcal{V}_*(M)}.$$

Thus,

$$\left(\frac{\rho(M)^2-1}{\rho(M)\mathcal{V}^*(M)}\right)n + \frac{1}{\mathcal{V}^*(M)} \le \Phi(n) \le \left(\frac{\rho(M)^2-1}{\rho(M)\mathcal{V}_*(M)}\right)n + \frac{1}{\mathcal{V}_*(M)}.$$

After dividing by n and taking the respective lim inf and lim sup, we get that

$$\frac{\rho(M)^2-1}{\rho(M)\mathcal{V}^*(M)} \le \liminf_{n\to\infty} \frac{\Phi(n)}{n} \le \limsup_{n\to\infty} \frac{\Phi(n)}{n} \le \frac{\rho(M)^2-1}{\rho(M)\mathcal{V}_*(M)}. \qquad \square$$

If $\Delta(M) = \{d\}$, then Corollary 1 implies that $\mathcal{V}^*(M) = \mathcal{V}_*(M) = d$ and Theorem 1 reduces to the following.

Corollary 2. *Let M be an atomic monoid with $\rho(M) < \infty$. If $\Delta(M) = \{d\}$, then*

$$\overline{\Phi}(M) = \frac{\rho(M)^2-1}{\rho(M)d}. \tag{4}$$

Corollary 2 immediately has some nice applications.

Example 3. A numerical monoid is an additive submonoid of the nonnegative integers. Every numerical monoid S has a unique minimal set of generators, and we will use the notation $S = \langle a_1, a_2, \ldots, a_t \rangle$ to represent the minimal generating set (which we assume is written in linear order). S is *primitive* if

$$1 = \gcd\{s \mid s \in S\}.$$

Every numerical monoid S is isomorphic to a unique primitive numerical monoid, so when working with numerical monoids, we can always assume that S is a primitive numerical monoid. By [Bowles et al. 2006], there exists a method for calculating $\max \Delta(S)$ in finite time and

$$\min \Delta(S) = \gcd \{a_i - a_{i-1} \mid i \in \{2, 3, \ldots, t\}\} = d.$$

By [Chapman et al. 2006, Theorem 2.1], $\rho(S) = a_t/a_1$. Hence for a numerical monoid, Equation (3) reduces to

$$\frac{a_t^2 - a_1^2}{\mathcal{V}^* a_1 a_t} \leq \lim_{n\to\infty} \inf \frac{\Phi(n)}{n} \leq \lim_{n\to\infty} \sup \frac{\Phi(n)}{n} \leq \frac{a_t^2 - a_1^2}{\mathcal{V}_* a_1 a_t}.$$

If we know further that the generators of S form an arithmetic sequence (that is, $S = \langle a, a+d, a+2d, \ldots, a+kd\rangle$ for some positive integers d and k), then [Bowles et al. 2006, Theorem 3.9] indicates that $\Delta(S) = \{d\}$. In this case we obtain an exact calculation of $\overline{\Phi}(S)$ as

$$\overline{\Phi}(S) = \frac{k(2a+kd)}{a(a+kd)} = k\left(\frac{1}{a} + \frac{1}{a+kd}\right). \qquad \square$$

Example 4. Let a and b be positive integers with $a \leq b$ and $a^2 \equiv a \pmod b$. The set of numbers $M(a, b) = \{x \mid x \in \mathbb{N} \text{ and } x \equiv a \pmod b\} \cup \{1\}$ forms a multiplicative monoid known as an *arithmetical congruence monoid* (ACM). ACMs have been the focus of three recent papers in the literature [Banister et al. 2007a, 2007b, Baginski et al. 2008]. An ACM is called *local* if $\gcd(a, b) = p^\alpha$ for some prime number p and positive integer α. It follows from elementary number theory that a local ACM $M(a, b)$ has a minimal index, which we denote by β, for which $p^\beta \in M(a, b)$. There are two relevant known results for a local ACM $M(a, b)$:

- $\rho(M(a, b)) = \frac{\alpha+\beta-1}{\alpha}$ [Banister et al. 2007b, Theorem 2.4]
- if $\alpha = \beta > 1$, then $\Delta(M(a, b)) = \{1\}$ [Baginski et al. 2008, Theorem 3.1].

Hence, for an ACM as above where $\alpha = \beta > 1$ (for instance, $M(4, 12)$), Equation (4) reduces to

$$\overline{\Phi}(M(a, b)) = \frac{(2\alpha-1)^2 - \alpha^2}{\alpha(2\alpha-1)}. \qquad \square$$

We close with a few comments:

- The proof in [Chapman and Smith 1993b] of Equation (1) relies on a different technique than that used above. The proof relies on knowing the exact structure of the sets in an infinite subsequence of the sequence $\mathcal{V}(1), \mathcal{V}(2), \ldots$.
- By a recent result of [Freeze and Geroldinger ≥ 2008],

  $$\mathcal{V}^*(\mathcal{B}(G)) = \mathcal{V}_*(\mathcal{B}(G)) = 1$$

  for all abelian groups G. Combined with Theorem 1, this yields a simpler proof of Equation (1) than the original proof in [Chapman and Smith 1993b].

- Connected to the last remark is a question posed in [Chapman and Smith 1998, Section 5]: for $\mathscr{B}(\mathbb{Z}_n)$, does $\rho_3 = \max \mathscr{V}(3) = n+1$? This question has been answered in the affirmative by [Gao and Geroldinger $\geq$ 2008].

References

[Anderson 1997] D. F. Anderson, "Elasticity of factorizations in integral domains: a survey", pp. 1–29 in *Factorization in integral domains*, Lecture Notes in Pure and Appl. Math. **189**, Dekker, New York, 1997. MR 98j:13002 Zbl 0903.13008

[Anderson and Pruis 1991] D. F. Anderson and P. Pruis, "Length functions on integral domains", *Proc. Amer. Math. Soc.* **113**:4 (1991), 933–937. MR 92c:13015

[Anderson et al. 1993] D. F. Anderson, S. Chapman, F. Inman, and W. W. Smith, "Factorization in $K[X^2, X^3]$", *Arch. Math. (Basel)* **61**:6 (1993), 521–528. MR 94k:13027

[Baginski et al. 2008] P. Baginski, S. T. Chapman, and G. Schaeffer, "On the Delta set of a singular arithmetical congruence monoid", 2008, Available at http://www.trinity.edu/schapman/FinalRefereedVersionjT.pdf. To appear.

[Banister et al. 2007a] M. Banister, J. Chaika, S. T. Chapman, and W. Meyerson, "On a result of James and Niven concerning unique factorization in congruence semigroups", *Elem. Math.* **62**:2 (2007), 68–72. MR MR2314039

[Banister et al. 2007b] M. Banister, J. Chaika, S. T. Chapman, and W. Meyerson, "On the arithmetic of arithmetical congruence monoids", *Colloq. Math.* **108**:1 (2007), 105–118. MR 2007m:20096

[Bowles et al. 2006] C. Bowles, S. T. Chapman, N. Kaplan, and D. Reiser, "On delta sets of numerical monoids", *J. Algebra Appl.* **5**:5 (2006), 695–718. MR 2007j:20092

[Chapman and Smith 1990] S. T. Chapman and W. W. Smith, "Factorization in Dedekind domains with finite class group", *Israel J. Math.* **71**:1 (1990), 65–95. MR 91i:13022

[Chapman and Smith 1993a] S. T. Chapman and W. W. Smith, "An analysis using the Zaks–Skula constant of element factorizations in Dedekind domains", *J. Algebra* **159**:1 (1993), 176–190. MR 94e:13037

[Chapman and Smith 1993b] S. T. Chapman and W. W. Smith, "On the lengths of factorizations of elements in an algebraic number ring", *J. Number Theory* **43**:1 (1993), 24–30. MR 93k:11098

[Chapman and Smith 1998] S. T. Chapman and W. W. Smith, "Generalized sets of lengths", *J. Algebra* **200**:2 (1998), 449–471. MR 99i:13030

[Chapman et al. 2000] S. T. Chapman, M. Freeze, and W. W. Smith, "On generalized lengths of factorizations in Dedekind and Krull domains", pp. 117–137 in *Non-Noetherian commutative ring theory*, Math. Appl. **520**, Kluwer Acad. Publ., Dordrecht, 2000. MR 2002i:13025 Zbl 0987.13011

[Chapman et al. 2002] S. T. Chapman, U. Krause, and E. Oeljeklaus, "On Diophantine monoids and their class groups", *Pacific J. Math.* **207**:1 (2002), 125–147. MR 2004b:20089

[Chapman et al. 2006] S. T. Chapman, M. T. Holden, and T. A. Moore, "Full elasticity in atomic monoids and integral domains", *Rocky Mountain J. Math.* **36** (2006), 1437–1455. MR 2007j:20093

[Freeze and Geroldinger $\geq$ 2008] M. Freeze and A. Geroldinger, "Unions of sets of lengths", *Funct. Approximatio Comment. Math.* to appear.

[Gao and Geroldinger $\geq$ 2008] W. Gao and A. Geroldinger, "On products of k-atoms", *Monatsh. Math.*. to appear.

[Geroldinger 1988] A. Geroldinger, "Über nicht-eindeutige Zerlegungen in irreduzible Elemente", *Math. Z.* **197**:4 (1988), 505–529. MR 89d:11096

[Geroldinger and Halter-Koch 1992] A. Geroldinger and F. Halter-Koch, "On the asymptotic behaviour of lengths of factorizations", *J. Pure Appl. Algebra* **77**:3 (1992), 239–252. MR 93c:13001

[Geroldinger and Halter-Koch 2006] A. Geroldinger and F. Halter-Koch, *Non-unique factorizations*, vol. 278, Pure and Applied Mathematics (Boca Raton), Chapman & Hall/CRC, Boca Raton, FL, 2006. Algebraic, combinatorial and analytic theory. MR 2006k:20001

[Geroldinger and Hassler $\geq$ 2008] A. Geroldinger and W. Hassler, "Local tameness in v-noetherian monoids", *J. Pure Appl. Algebra*.. To appear.

Received: 2007-10-28 Revised: Accepted: 2007-10-29

baginski@gmail.com
University of California at Berkeley, Department of Mathematics, Berkeley CA 94720-3840, United States

schapman@trinity.edu
Trinity University, Department of Mathematics, One Trinity Place, San Antonio, TX 78212-7200, United States

cookiedoe68@gmail.com
The College of New Jersey, Mathematics and Statistics Department, P.O. Box 7718, Ewing, NJ 08628-0718, United States

paixao@vt.edu
Virginia Tech, Department of Mathematics, 460 McBryde, Blacksburg, VA 24061-0123, United States

An asymptotic for the representation of integers as sums of triangular numbers

Atanas Atanasov, Rebecca Bellovin,
Ivan Loughman-Pawelko, Laura Peskin and Eric Potash

(Communicated by Ken Ono)

Motivated by the result of Rankin for representations of integers as sums of squares, we use a decomposition of a modular form into a particular Eisenstein series and a cusp form to show that the number of ways of representing a positive integer n as the sum of k triangular numbers is asymptotically equivalent to the modified divisor function $\sigma^{\sharp}_{2k-1}(2n+k)$.

1. Introduction

1A. ***General problem.*** We wish to study $\delta_k(n)$, the number of ways to write n as the sum of k triangular numbers. This problem dates back to Gauss, who discovered that every nonnegative integer can be represented as a sum of three triangular numbers. The basic problem is similar to questions about representations of integers as sums of squares, and some of the basic techniques for attacking that problem carry over. We define the function

$$\Theta(q) := \sum_{n=-\infty}^{\infty} q^{n^2} = 1 + 2q + 2q^4 + 2q^9 + \cdots$$

so that

$$\Theta^k(q) = \sum_{n\geq 0} r_k(n) q^n$$

where $r_k(n)$ is the number of representations of n as the sum of k squares. It was exploited in [Rankin 1965] the fact that $\Theta(1)$ is a modular form of weight $\frac{1}{2}$ for $\Gamma_0(4)$ to study the functions $r_k(n)$. Ono, Robins, and Wahl [1995] defined an analogous modular form to study triangular numbers.

We begin by defining triangular numbers.

MSC2000: 11F11.

Keywords: modular form, triangular number, asymptotics.

Definition 1.1. The *n-th triangular number ($n \geq 0$)* is

$$T_n := \frac{n(n+1)}{2}.$$

These numbers may be geometrically interpreted as the number of dots in a grid with the shape of an equilateral triangle of side length n. We also introduce the generating functions

$$\Psi(q) := \sum_{n=0}^{\infty} q^{T_n} = 1 + q + q^3 + q^6 + \cdots$$

and

$$\Psi^k(q) = \sum_{n=0}^{\infty} \delta_k(n) q^n.$$

1B. ***Modular group and congruence subgroups.*** Before we formally define modular forms, we need to define the *modular group* and its subgroups.

Definition 1.2. Let $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. The *modular group* Γ is

$$\mathrm{SL}_2(\mathbb{Z}) \;=\; \{A \mid a, b, c, d \in \mathbb{Z} \text{ and } \det A = \pm 1\}.$$

It is well known that Γ is generated by $S = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ and $T = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$.

Definition 1.3. The *congruence subgroups of level $N \in \mathbb{N}$* are defined as follows:

(1) $\Gamma_0(N) := \{A \in \Gamma \mid c \equiv 0 \bmod N\}$;

(2) $\Gamma_1(N) := \{A \in \Gamma \mid c \equiv 0 \bmod N \text{ and } a \equiv d \equiv 1 \bmod N\}$;

(3) $\Gamma(N) := \{A \in \Gamma \mid c \equiv b \equiv 0 \bmod N \text{ and } a \equiv d \equiv 0 \bmod N\}$.

It is easy to check that they are, in fact, subgroups.

It is clear that for every level $N \in \mathbb{N}$, $\Gamma(N) \leq \Gamma_1(N) \leq \Gamma_0(N) \leq \Gamma$. More precisely, the following identities hold [Koblitz 1993, p. 231]:

$$[\Gamma : \Gamma_0(N)] = N \prod_{p \mid N} \Big(1 + \frac{1}{p}\Big),$$

$$[\Gamma : \Gamma_1(N)] = N^2 \prod_{p \mid N} \Big(1 - \frac{1}{p^2}\Big),$$

$$[\Gamma : \Gamma(N)] = N^3 \prod_{p \mid N} \Big(1 - \frac{1}{p^2}\Big).$$

We will make use of $\Gamma_0(4)$, which is generated by T and $ST^{-4}S$. In particular, $[\Gamma : \Gamma_0(4)] = 6$, with coset representatives

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad S^{-1}T^{-2}S = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}, \qquad S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$
$$ST = \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}, \qquad ST^2 = \begin{pmatrix} 0 & -1 \\ 1 & 2 \end{pmatrix}, \quad ST^3 = \begin{pmatrix} 0 & -1 \\ 1 & 3 \end{pmatrix}.$$

Any group of 2×2 matrices gives rise to an action on the complex plane, namely the linear fractional transformation

$$Az := \frac{az+b}{cz+d},$$

where $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. In particular, Γ and its subgroups act on the upper half plane

$$\mathcal{H} = \{z \in \mathbb{C} \mid \operatorname{Im}(z) > 0\}.$$

Considering the geometric meaning of orbits and equivalence classes under this action on $\mathcal{H}$ leads to the idea of a fundamental domain. This is a subset of $\mathcal{H}$ which possesses both convenient topological and geometric properties and is also algebraically related to some Γ or one of its subgroups.

Definition 1.4. A closed, simply connected region F in $\mathcal{H}$ is called a *fundamental domain* for a subgroup Γ' of Γ if every point in the plane is equivalent under Γ' to a point in F and no two points in the interior of F are equivalent under Γ'.

For example, a fundamental domain for Γ is the set

$$R_\Gamma = \{z \in \mathbb{C} \mid -\tfrac{1}{2} \le \operatorname{Re}(z) \le \tfrac{1}{2}, |z| \ge 1\}.$$

Figure 1 shows this fundamental domain, as well as the fundamental domain for $\Gamma_0(4)$

For the sake of consistency, we will use $R_{\Gamma'}$ to denote a fundamental domain for Γ'. The following lemma provides an algorithm to compute $R_{\Gamma'}$ using R_Γ, and coset representation of Γ' in Γ.

Lemma 1.5. *Let $\Gamma' \le \Gamma$ be of finite index n in Γ. If $\Gamma = \bigcup_{i=1}^n \gamma_i \Gamma'$ is its coset representation, then*

$$R_{\Gamma'} = \bigcup_{i=1}^{n} \gamma_i^{-1} R_\Gamma.$$

Proof. This is proved in [Koblitz 1993, p.105]. □

Definition 1.6. Let $\Gamma' \le \Gamma$, and fix a fundamental domain $R_{\Gamma'}$. The points where $R_{\Gamma'}$ intersects the boundary $\partial\mathcal{H} = \{i\infty\} \cup \mathbb{R}$ are called the *cusps of* Γ'.

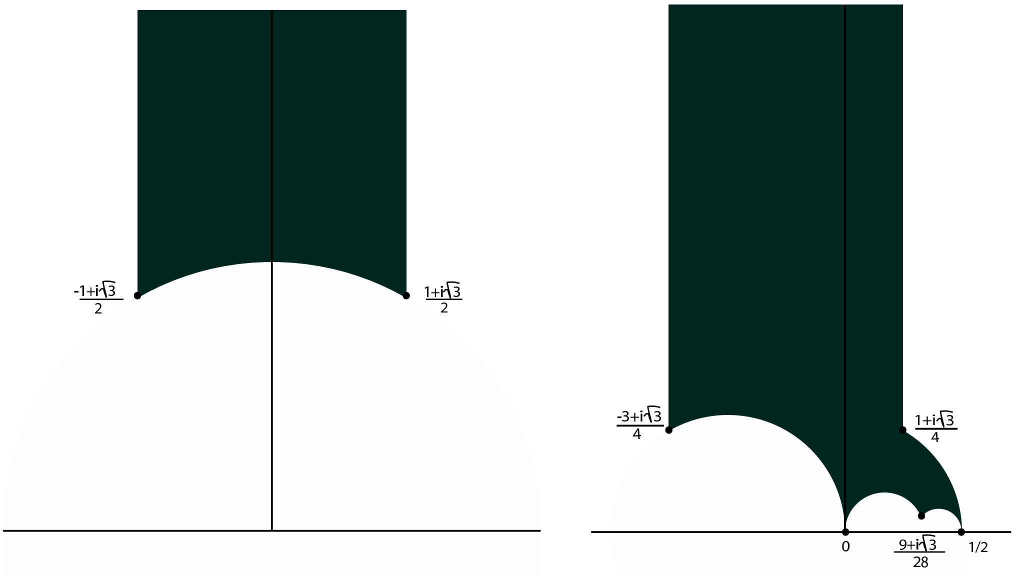


Figure 1. The fundamental domains for Γ (left) and for $\Gamma_0(4)$ (right).

The full modular group has a single cusp at $i\infty$. From the fundamental domain for $\Gamma_0(4)$ shown in Figure 1, we see that $\Gamma_0(4)$ has three cusps, namely 0, $\frac{1}{2}$ and $i\infty$.

1C. *Modular forms.* Modular forms are holomorphic functions on $\mathcal{H}$ which have nice symmetry properties under the action of Γ or one of its subgroups. Specifically, we say

Definition 1.7. $f : \mathcal{H} \to \mathbb{C}$ is a *modular form of weight* $k \in \mathbb{N}$ *over* $\Gamma_0(N)$ if

(i) f is holomorphic on $\mathcal{H}$;

(ii) f is holomorphic at the cusps of $\Gamma_0(N)$;

(iii) for all $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_0(N)$, the equation $f(Az) = (cz+d)^k f(z)$ holds for all $z \in \mathcal{H}$.

Definition 1.8. A modular form f over Γ' is called a *cusp form* if it vanishes at all cusps of Γ'.

If $T \in \Gamma'$, it follows that a modular form over Γ' always has period 1. In other words, $f(z) = f(z+1)$ for all $z \in \mathcal{H}$. Therefore f has a *Fourier expansion* (also called q*-expansion*) in $q = e^{2\pi i z}$:

$$f(z) = \sum_{n=0}^{\infty} c(n) q^n.$$

Modular forms of a given congruence subgroup and of fixed weight form a vector space. This structure can be of great help when trying to study a particular modular form. Using a suitable basis, we can decompose elements of a given space of modular forms in terms of basis vectors. This technique often produces expressions that are easy to work with.

Definition 1.9. The vector space of modular forms of weight k over the congruence subgroup Γ' of Γ is denoted $M_k(\Gamma')$. The subspace of cusp forms is denoted $S_k(\Gamma')$.

For all $\Gamma' \le \Gamma$ which contain $-I$, both $M_{2k+1}(\Gamma')$ and $S_{2k+1}(\Gamma')$ are trivial. This follows by applying the transformation $-I$ to a modular form $f(z)$, implying

$$f(z) = (-1)^{2k+1} f(z) = -f(z),$$

and hence $f(z) = 0$ for all $z \in \mathcal{H}$. Thus, we may consider only modular forms of even weight. Since $\Gamma_0(4)$ will be important in our work, we state without proof a characterization of its spaces of even-weight modular forms. Recall the definition of $\Theta(z) = \sum_{n=-\infty}^{\infty} q^{n^2}$.

Definition 1.10. Let $F(z)$ be the following modular form of weight 2 over $\Gamma_0(4)$:

$$F(z) = \sum_{n=1}^{\infty} \sigma_1(2n+1) q^{2n+1}.$$

Lemma 1.11. *$M_{2k}(\Gamma_0(4))$ is a $(k+1)$-dimensional vector space with basis*

$$\{F^k, F^{k-1}\Theta^4, \ldots, F\Theta^{4(k-1)}, \Theta^{2k}\}.$$

Furthermore, $S_{2k}(\Gamma_0(4))$ consists of all polynomials divisible by

$$\Theta^4 F(\Theta^4 - 16F) = \eta^{12}(2z).$$

Therefore, there exists an isomorphism between $S_{2k}(\Gamma_0(4))$ and $M_{2k-6}(\Gamma_0(4))$.

Proof. This is Exercise III.3.17 in [Koblitz 1993], proved on pp. 235–6. □

1D. ***Representations as sums of triangular numbers.*** In our study of $\delta_k(n)$, we focus on the expressions for $\delta_{4k}(n)$. The generating function $q^k \Psi^{4k}(q^2)$ is in $M_{2k}(\Gamma_0(4))$, which is a well-understood space of small dimension. By decomposing elements of $M_{2k}(\Gamma_0(4))$ for some k into basis vectors, it is possible to find identities between $q^k \Psi^{4k}(q^2)$ and other forms in the same space with accessible coefficients. This is the method used by Ono et al. [1995] for $\delta_k(n)$, $k = 2, 3, 4, 6, 8, 10, 12$ and 24. Their results for $\delta_{4k}(n)$ are summarized in the following lemma.

Lemma 1.12. *For $n \geq 0$,*

$$\begin{aligned} \delta_4(n) &= \sigma_1(2n+1), \\ \delta_8(n) &= -\tfrac{1}{8}\sigma_3^\sharp(n+1), \\ \delta_{12}(n) &= \tfrac{1}{256}(\sigma_5(2n+3) - a(2n+3)), \text{ and} \\ \delta_{24}(n) &= \frac{1}{17689}\Bigl(\sigma_{11}^\sharp(n+3) - \tau(n+3) - 2072\,\tau\Bigl(\frac{n+3}{2}\Bigr)\Bigr), \end{aligned}$$

where $a(n)$ is defined by $\eta^{12}(2z) = \sum_{n=1}^{\infty} a(n)q^n$ and $\tau(n)$ is the n-th Fourier coefficient of $\Delta(z) = (2\pi)^{12}\eta^{24}(z)$.

Looking at the case δ_4, this lemma states that

$$q\Psi^4(q^2) = \sum_{n=0}^{\infty} \delta_4(n)q^{2n+1} = F(z),$$

where F is as defined previously, so we have the following useful corollary:

Corollary 1.13.

$$q^k\Psi^{4k}(q^2) = F^k.$$

2. δ_{4k} as an Eisenstein series plus a cusp form

The generating function $\Theta^k(z)$ for $r_k(n)$ can be decomposed into a cusp form and a particular Eisenstein series. In the same vein as the work by Rankin [1965] on $\Theta^k(z)$, we would like to similarly decompose $q^k\Psi^{4k}(q^2)$.

Definition 2.1. Let $k \in \mathbb{N}$. Then let H_{2k} be the Eisenstein series of weight $2k$ on $\Gamma_0(4)$ defined by

$$H_{2k}(z) = \begin{cases} \displaystyle\sum_{\substack{n>0 \\ n \text{ odd}}}^{\infty} \sigma_{2k-1}^\sharp(n)q^n & \text{if } k \text{ is odd,} \\ \displaystyle\sum_{\substack{n>0 \\ n \text{ even}}}^{\infty} \sigma_{2k-1}^\sharp(n)q^n & \text{if } k \text{ is even.} \end{cases}$$

Definition 2.2. We define the *partial zeta function $\zeta^i(s)$ for i modulo N* to be

$$\zeta^i(s) := \sum_{n \equiv i \bmod N} \frac{1}{n^s}.$$

Proposition 2.3. *For a given congruence subgroup $\Gamma_0(N)$, let $G_k^{(a_1,a_2)}(z)$ be the Eisenstein series*

$$\sum_{\substack{m_1 \equiv a_1(N) \\ m_2 \equiv a_2(N)}} \frac{1}{(m_1 z + m_2)^k},$$

and let $B_{2k} \in \mathbb{Q}$ be the $2k$-th Bernoulli number. Then

(1) *as a modular form for $\Gamma_0(2)$,*

$$\sum_{n=1}^{\infty} \sigma^{\sharp}_{2k-1}(n) q^n = -\frac{B_{2k}}{8k\zeta(2k)} \big(G^{(1,0)}_{2k}(z) + G^{(1,1)}_{2k}(z)\big);$$

(2) *as a modular form for $\Gamma_0(4)$,*

$$\sum_{n=1}^{\infty} \sigma^{\sharp}_{2k-1}(n) q^n = -\frac{2^{2k} B_{2k}}{8k\zeta(2k)} \big(G^{(2,0)}_{2k}(z) + G^{(2,2)}_{2k}(z)\big).$$

Proof. Koblitz [1993, p. 133] shows that for $\Gamma_0(N)$,

$$G^{(a_1,a_2)}_k(z) = b_0^{(a_1,a_2)} + \frac{(-1)^{k-1} 2k\zeta(k)}{N^k B_k} \cdot \Bigg(\sum_{\substack{m_1 \equiv a_1 \bmod N \\ m_1 > 0}} \sum_{j=1}^{\infty} j^{k-1} \xi^{ja_2} q_N^{jm_1} + (-1)^k \sum_{\substack{m_1 \equiv -a_1 \bmod N \\ m_1 > 0}} \sum_{j=1}^{\infty} j^{k-1} \xi^{-ja_2} q_N^{jm_1} \Bigg),$$

where

$$\xi := e^{2\pi i/N}, \; q_N := e^{2\pi i z/N}, \; b_0^{(a_1,a_2)} = \begin{cases} 0 & \text{if } a_1 \neq 0, \\ \zeta^{a_1}(k) + (-1)^k \zeta^{-a_2}(k) & \text{if } a_1 = 0. \end{cases}$$

We can collect terms with $jm_1 = n$ to find explicit expansions of some particular $G^{(a_1,a_2)}_k(z)$. From the above expression, we have two assertions:

(i)
$$G^{(1,0)}_{2k}(z) = 2c_{2k} \sum_{n=1}^{\infty} \Bigg(\sum_{j \mid n, n/j \text{ odd}} j^{2k-1} \Bigg) q_2^n,$$
$$G^{(1,1)}_{2k}(z) = 2c_{2k} \sum_{n=1}^{\infty} \Bigg(\sum_{j \mid n, n/j \text{ odd}} j^{2k-1} (-1)^j \Bigg) q_2^n.$$

Adding these two series, we get

$$G^{(1,0)}_{2k}(z) + G^{(1,1)}_{2k}(z) = 2^{2k+1} c_{2k} \sum_{n=1}^{\infty} \sigma^{\sharp}_{2k-1}(n) q^n = -\frac{8k\zeta(2k)}{B_{2k}} \sum_{n=1}^{\infty} \sigma^{\sharp}_{2k-1}(n) q^n,$$

which is the first assertion.

(ii) The second assertion follows similarly, except that $c_{2k} = -\dfrac{4k\zeta(2k)}{2^{4k} B_{2k}}$ for the Eisenstein series of $\Gamma_0(4)$. □

If we take the first identity from Proposition 2.3 and substitute in $2z$, we obtain

$$\sum_{n=1}^{\infty} \sigma^{\sharp}_{2k-1}(n) q^{2n} = -\frac{B_{2k}}{8k\zeta(2k)} \big(G^{(1,0)}_{2k}(2z) + G^{(1,1)}_{2k}(2z)\big).$$

This is a modular form for $\Gamma_0(4)$. By the definition of $\sigma^\sharp_{2k-1}(n)$ and the definition of $G_k^{(a_1,a_2)}(z)$,

$$\sum_{n=1}^{\infty} \sigma^\sharp_{2k-1}(2n)q^{2n} = \\ -\frac{2^{2k}B_{2k}}{16\zeta(2k)}\big(G_{2k}^{(2,0)}(z)+G_{2k}^{(2,1)}(z)+G_{2k}^{(2,2)}(z)+G_{2k}^{(2,3)}(z)\big). \quad (2\text{-}1)$$

We have proven

Corollary 2.4. *$\sum_{\substack{n>0\\ even}} \sigma^\sharp_{2k-1}(n)q^n$ and $\sum_{\substack{n>0\\ odd}} \sigma^\sharp_{2k-1}(n)q^n$ are both modular forms for $\Gamma_0(4)$. The former is given by Equation (2-1), and the latter is equal to*

$$-\frac{2^{2k}B_{2k}}{16\zeta(2k)}\big(G_{2k}^{(2,0)}(z)-G_{2k}^{(2,1)}(z)+G_{2k}^{(2,2)}(z)-G_{2k}^{(2,3)}(z)\big).$$

We can now compute the desired values at the cusps. From [Koblitz 1993], we have

$$G_{2k}^{(2,i)}(z) = -\frac{4k\zeta(2k)}{4^{2k}B_{2k}}\sum_{n=1}^{\infty}\bigg(\sum_{\substack{j\,|\,n\\ n/j\equiv i\,(4)}} j^{2k-1} + \sum_{\substack{j\,|\,n\\ n/j\equiv -i\,(4)}} j^{2k-1}\bigg)q_4^n$$

so it follows that $\sum_{\substack{n>0\\ \text{even}}} \sigma^\sharp_{2k-1}(n)q^n$ and $\sum_{\substack{n>0\\ \text{odd}}} \sigma^\sharp_{2k-1}(n)q^n$ are both 0 at $i\infty$.

To find the values at the cusp 0, we use the transformation S. We have

$$G_{2k}^{(2,0)}(z)|[S]_{2k} = G^{(0,2)}(z); \quad G_{2k}^{(2,1)}(z)|[S]_{2k} = G^{(1,2)}(z);$$
$$G_{2k}^{(2,2)}(z)|[S]_{2k} = G^{(2,2)}(z); \quad G_{2k}^{(2,3)}(z)|[S]_{2k} = G^{(3,2)}(z).$$

Additionally, $G^{(1,2)}(i\infty) = G^{(2,2)}(i\infty) = G^{(3,2)}(i\infty) = 0$ (from [Koblitz 1993] again) and

$$G^{(0,2)}(i\infty) = 2\zeta^2(2k) = 2\sum_{\substack{n>0\\ n\equiv 2(4)}} \frac{1}{n^{2k}} = 2\Big(\frac{1}{2^{2k}}-\frac{1}{2^{4k}}\Big)\zeta(2k).$$

Hence, $\sum_{\substack{n>0\\ \text{even}}} \sigma^\sharp_{2k-1}(n)q^n$ and $\sum_{\substack{n>0\\ \text{odd}}} \sigma^\sharp_{2k-1}(n)q^n$ both equal

$$-\frac{B_{2k}}{8k}\Big(1-\frac{1}{4^k}\Big)$$

at 0.

To find the values at the cusp $\frac{1}{2}$, we use the transformation $ST^{-2}S$. We have

$$G_{2k}^{(2,0)}(z)|[ST^{-2}S]_{2k}=G^{(2,0)}(z);\quad G_{2k}^{(2,1)}(z)|[ST^{-2}S]_{2k}=G^{(0,1)}(z);$$
$$G_{2k}^{(2,2)}(z)|[ST^{-2}S]_{2k}=G^{(2,2)}(z);\quad G_{2k}^{(2,3)}(z)|[ST^{-2}S]_{2k}=G^{(0,3)}(z).$$

We know $G_{2k}^{(2,0)}(i\infty)=G_{2k}^{(2,2)}(i\infty)=0$, and

$$G_{2k}^{(0,1)}(i\infty)=G_{2k}^{(0,3)}(i\infty)=\zeta^1(2k)+\zeta^3(2k)=\sum_{\substack{n>0\\ n\text{ odd}}}\frac{1}{n^{2k}}=\Big(1-\frac{1}{2^{2k}}\Big)\zeta(2k).$$

Hence, at the cusp $\frac{1}{2}$,

$$\sum_{n>0,\text{even}}\sigma_{2k-1}^{\sharp}(n)q^n=-\frac{4^kB_{2k}}{8k}\Big(1-\frac{1}{4^k}\Big),$$
$$\sum_{n>0,\text{odd}}\sigma_{2k-1}^{\sharp}(n)q^n=\frac{4^kB_{2k}}{8k}\Big(1-\frac{1}{4^k}\Big).$$

Theorem 2.5. *Let $k\in\mathbb{N}$. Then*

$$q^k\Psi^{4k}(q^2)=\frac{1}{d_k}(H_{2k}(z)-T_{2k}(z)),\tag{2-2}$$

where

$$d_k=-\frac{(-16)^kB_{2k}(4^k-1)}{8k}\in\mathbb{Q}$$

and $T_{2k}(z)\in S_{2k}(\Gamma_0(4))$.

Proof. We know that $F^k(z)$ is 0 at $i\infty$, $\big(-\frac{1}{64}\big)^k$ at 0, and $\big(\frac{1}{16}\big)^k$ at $\frac{1}{2}$, as is $\frac{1}{d_k}H_{2k}(z)$. Hence,

$$q^k\Psi^{4k}(q^2)-\frac{1}{d_k}H_{2k}(z)$$

is a cusp form. □

Corollary 2.6.

$$\delta_{4k}(n)=\frac{1}{d_k}(\sigma_{2k-1}^{\sharp}(2n+k)-a(2n+k)),\tag{2-3}$$

where

$$T_{2k}(z)=\sum_n a(n)q^n\in S_{2k}(\Gamma_0(4)).$$

Proof. This follows from equating the coefficients of the Fourier series in (2-2). □

k	$\sum a(n)q^n$
1	0
2	0
3	$\Theta^8 F - 16\Theta^4 F^2$
4	$2^7(\Theta^8 F^2 - 16\Theta^4 F^3)$
5	$\Theta^{16} F - 32\Theta^{12} F^2 + 19968\Theta^8 F^3 - 3159\Theta^4 F^4$
6	$2^{11}(\Theta^{16} F^2 - 32\Theta^{12} F^3 + 2328\Theta^8 F^4 - 33152\Theta^4 F^5)$
7	$\Theta^{24} F - 48\Theta^{20} F^2 + 1595136\Theta^{16} F^3$
	$-51023872\Theta^{12} F^4 + 1660747776\Theta^8 F^5 - 20041433088\Theta^4 F^6$
8	$2^{15}\big(\Theta^{24} F^2 - 48\Theta^{20} F^3 + 33576\Theta^{16} F^4$
	$-1053952\Theta^{12} F^5 + 23271936\Theta^8 F^6 - 237969408\Theta^4 F^7\big)$

Table 1

Corollary 2.7. $\delta_{4k}(n) \sim \sigma^{\sharp}_{2k-1}(2n+k)$.

Proof. The cusp form coefficients in (2-3) $a(2n+k) \in O(n^k)$ [Apostol 1990]. The $\sigma^{\sharp}_{2k-1}(2n+k)$ term has lower bound n^{2k-1}, and thus this term is asymptotically dominant. Therefore

$$\lim_{n\to\infty} \frac{\delta_{4k}(n)}{\sigma^{\sharp}_{2k-1}(2n+k)} = 1\,. \qquad \square$$

For particular k, we can compute the value of c_{2k}, and then, by equating finitely many coefficients, compute the remaining cusp form $\sum a(n)q^n$ as a homogeneous polynomial in F and Θ^4. We list the result of this computation for several values of k in Table 1.

We can rewrite (2-3) using

$$\sigma^{\sharp}_k(n) = \begin{cases} \sigma_k(n) & \text{if } n \text{ is odd,} \\ 2^k \sigma^{\sharp}_k(\frac{n}{2}) & \text{if } n \text{ is even,} \end{cases}$$

and the values in Table 1. The resulting formulae for $k = 1$, 2, 3, and 6 agree with those in Lemma 1.12.

Acknowledgments

We would like to thank Kimball Martin, Joe Ross, and Mirela Cipriani for their help as part of the Columbia Mathematics Department's 2007 Summer Undergraduate Research Program. We would also like to thank the department for her support, which made our summer research possible.

References

[Apostol 1990] T. M. Apostol, *Modular functions and Dirichlet series in number theory*, 2nd ed., Graduate Texts in Mathematics **41**, Springer, New York, 1990. MR 90j:11001 Zbl 0697.10023

[Koblitz 1993] N. Koblitz, *Introduction to elliptic curves and modular forms*, 2nd ed., Graduate Texts in Mathematics **97**, Springer, New York, 1993. MR 94a:11078 Zbl 0804.11039

[Ono et al. 1995] K. Ono, S. Robins, and P. T. Wahl, "On the representation of integers as sums of triangular numbers", *Aequationes Math.* **50**:1-2 (1995), 73–94. MR 96i:11044 Zbl 0828.11057

[Rankin 1965] R. A. Rankin, "Sums of squares and cusp forms", *Amer. J. Math.* **87** (1965), 857–860. MR 32 #5605 Zbl 0132.30802

Received: 2007-10-30 Accepted: 2008-01-19

ava2102@columbia.edu
7653 Lerner Hall, Columbia University,
New York, NY 10027, United States

rmb2113@columbia.edu
1603 Lerner Hall, Columbia University,
New York, NY 10027, United States

il2139@columbia.edu
4067 Lerner Hall, Columbia University,
New York, NY 10027, United States

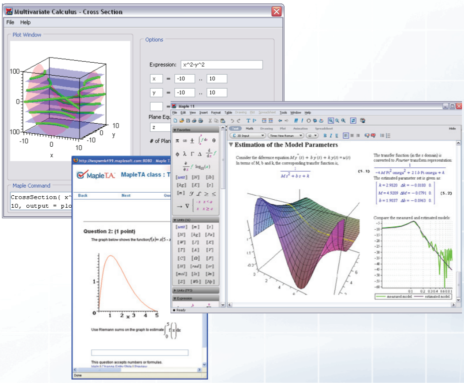
lp2153@columbia.edu
5991 Lerner Hall, Columbia University,
New York, NY 10027, United States

elp2109@columbia.edu
4926 Lerner Hall, Columbia University,
New York, NY 10027, United States

Accelerate Your Research AND Enhance Your Classroom

You probably already know that Maple™ 11 has thousands of mathematical functions, hundreds of plotting routines, a sophisticated programming language designed for mathematics, and professional publishing tools, making it the ideal tool for your mathematical research.

Did you also know…?


Multivariate Calculus - Cross Section

Expression: x^2-y^2

x = -10 .. 10

y = -10 .. 10

```
CrossSection( x^
10, output = plo
```

Estimation of the Model Parameters

Consider the difference equation $M y''(t) + b y'(t) + k y(t) = u(t)$
In terms of M, b and k, the corresponding transfer function is,

$$\frac{1}{M s^2 + b s + k} \quad (5.1)$$

The transfer function (in the s domain) is converted to *Fourier* transform representation:

$$\frac{1}{-4 M \text{Pi}^2 \text{omega}^2 + 2 I b \text{Pi omega} + k}$$

The estimated parameter set is given as:

$$\begin{bmatrix} k = 2.9820 & \Delta k = -0.0180 \\ M = 4.9209 & \Delta M = -0.0791 \\ b = 1.9037 & \Delta b = -0.0963 \end{bmatrix} \quad (5.2)$$

Compare the measured and estimated models:

Question 2: (1 point)

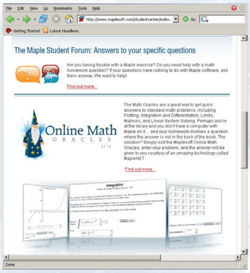
The graph below shows the function $f(x) = x(5-x$

Use Riemann sums on the graph to estimate $\int_0^5 f(x)\,dx$

This question accepts numbers or formulas.

· With Clickable Calculus™ techniques, Maple's point-and-click interface makes math easy to do, easy to learn, and easy to teach.

· With Maple T.A.™ and the Maplesoft™-MAA® Placement Test Suite, students can take online placement tests, do assignments, and receive instant feedback all online. The system also makes it easy to create meaningful mathematical questions, so you can ask just what you want to ask. And the grading? All done at the click of a button!


http://www.maplesoft.com/studentcenter/index...

The Maple Student Forum: Answers to your specific questions

Are you having trouble with a Maple exercise? Do you need help with a math homework question? If your questions have nothing to do with Maple software, ask them anyway. We want to help!

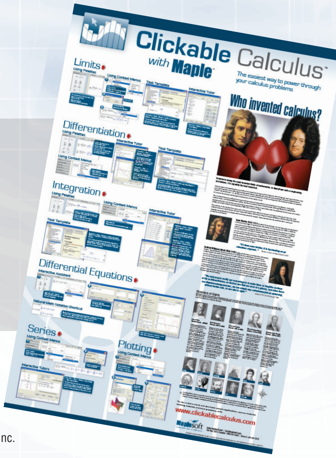
Find out more...

The Math Oracles are a great way to get quick answers to standard math problems, including Plotting, Integration and Differentiation, Limits, Matrices, and Linear System Solving. Perhaps you're at the library and you don't have a computer with Maple on it ... and your homework involves a question where the answer is not in the back of the book. The solution? Simply visit the Maplesoft Online Math Oracles, enter your problem, and the answer will be given to you courtesy of an amazing technology called MapleNET.

Find out more...

· Maplesoft provides Clickable Calculus applications on the Teacher Resource Center, and video illustrations on the Student Help Center – a web site full of applications, training materials, question content, user forums, and more. Help is just a click away!

With Maplesoft, math just clicks!



Want to learn more?

Contact Maplesoft by calling **1.800.267.6583** (US & Canada) or **info@maplesoft.com**.

Order your **FREE** copy of the Clickable Calculus poster at **http://www.maplesoft.com/freeposter**

Maplesoft
command the brilliance

www.maplesoft.com | info@maplesoft.com

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

2008 Volume 1 No. 1