involve

a journal of mathematics

mathematical sciences publishers

2009 vol. 2, no. 4

involve

pjm.math.berkeley.edu/involve

See inside back cover or http://pjm.math.berkeley.edu/involve for submission instructions and subscription prices.
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 2:4(2009)

Automatic growth series for right-angled Coxeter groups

Rebecca Glover and Richard Scott

(Communicated by Vadim Ponomarenko)

Right-angled Coxeter groups have a natural automatic structure induced by their action on a CAT(0) cube complex. The normal form for this structure is defined with respect to the generating set consisting of all cliques in the defining graph for the group. In this paper we study the growth series for right-angled Coxeter groups with respect to this *automatic* generating set. In particular, we show that there exist nonisomorphic Coxeter groups with the same automatic growth series, and give a comparison with the usual growth series defined with respect to the *standard* generating set.

1. Introduction

Given a group and a generating set the corresponding *growth series* is the power series whose coefficients are the numbers of group elements of a given length (measured with respect to the generating set). If elements of the group have a suitable normal form with respect to the generating set, then the growth series can be computed as a rational function. A classical example is the case of Coxeter groups. A Coxeter group W is by definition a group given by a certain type of presentation, hence comes equipped with a set of generators, usually denoted by S. Growth series for Coxeter groups with respect to this standard generating set are known to be rational and are characterized by a simple inductive formula due to Steinberg [1968].

A Coxeter group is *right-angled* if any two generators either commute or generate an infinite dihedral group. Right-angled Coxeter groups are particularly nice to work with because they are completely characterized by the graph Γ consisting of a vertex for each generator and an edge for each pair of commuting generators.

MSC2000: 05A15, 20F10, 20F55.

Keywords: Coxeter groups, growth series.

Glover was supported by the Pennello Fund of the Department of Mathematics and Computer Science at Santa Clara University.

In this case, Steinberg's formula reduces to a simple expression involving only the numbers of cliques (complete subgraphs) in Γ of each size.

Any right-angled Coxeter group admits a natural action on a CAT(0) cube complex [Davis 2002; 2008], hence by a result of Niblo and Reeves [1998] it acquires an induced automatic structure. In particular, this structure includes a normal form that is recognized by a finite state automaton (see [Epstein et al. 1992] for details). Although Coxeter groups were already known to be automatic, the automatic structure coming from the action on the Davis complex is with respect to a different generating set than S. The relevant generating set, which we call the *automatic generating set*, consists of a single generator for each clique in Γ. More precisely, the generator corresponding to a clique σ is the product of the vertices of σ.

The purpose of this paper is to study the growth series of right-angled Coxeter groups with respect to this automatic generating set. In Section 2, we review relevant properties of right-angled Coxeter groups. In Section 3, we introduce a multivariate growth series that specializes (with suitable substitutions of variables) to either the standard growth series or the automatic growth series. In Section 4, we describe a procedure for computing this multivariate growth series, proving as a corollary that it is also a rational function. Although the automatic growth series is an invariant of the graph Γ, there does not seem to be any simple formula analogous to the formula of Steinberg. In particular, the growth series definitely depends on more than the numbers of cliques of each size. In Section 5, however, we impose a strong form of regularity on the graph Γ, and show that under this restriction the growth series does indeed depend only on the clique numbers. We use this fact to construct examples of nonisomorphic groups with the same automatic (and standard) growth series. Finally, in Section 6, we compare the standard and automatic growth rates of a right-angled Coxeter group. In general, one expects that enlarging the generating set will increase the growth rate. We prove this directly by proving the stronger result that if W is infinite, then the sphere of radius r is always smaller with respect to the standard generators than with respect to the automatic generators.

2. Right-angled Coxeter groups

Let Γ be a simplicial graph with vertex set S and edge set E. The *right-angled Coxeter group* (RACG) determined by Γ is the group defined by the presentation

$$W = \langle s \in S \mid s^2 = 1 \text{ for all } s \in S, \text{ and } (ss')^2 = 1 \text{ for all } \{s, s'\} \in E \rangle.$$

In other words, W has an (involutive) generator for each vertex of Γ, and two generators commute whenever the corresponding vertices are joined by an edge in Γ. The *flag completion of* Γ (also known as the *nerve of* W) is the largest

simplicial complex K on the vertex set S such that Γ coincides with the 1-skeleton of K. Thus, K consists of all subsets $\{s_1, \dots, s_k\} \subset S$ such that $\{s_i, s_j\} \in E$ for all $1 \le i < j \le k$. The following proposition will allow us to work with any one of Γ, W, or K interchangeably.

Proposition 1. *Let W and W' be the RACGs associated with Γ and Γ', respectively, and let K, K' be the corresponding nerves. Then*

$$W \cong W' \iff \Gamma \cong \Gamma' \iff K \cong K'.$$

Proof. The only implication that is not immediate is that isomorphic RACGs must come from isomorphic graphs. This is a theorem of Radcliffe [2001]. □

The *length* of an element $w \in W$ is the minimal k such that $w = s_1 \cdots s_k$ for $s_i \in S$. More generally, for any generating set $R \subset W$, we define the *R-length* of $w \in W$ to be the minimal k such that $w = r_1 \cdots r_k$ for $r_i \in R$. We let $l_R(w)$ denote the R-length of $w \in W$. Our primary interest in this paper is the generating set defined as follows. For any simplex $\sigma \in K$, let $w_\sigma = \prod_{s \in \sigma} s$. The element w_σ is well-defined since any pair of vertices in σ are joined by an edge, so the corresponding generators in W commute. By convention, we let $w_\varnothing = 1$. We let $K_{>\varnothing}$ denote the set of nonempty simplices in K, and we let A denote the set $A = \{w_\sigma \in W \mid \sigma \in K_{>\varnothing}\}$. Note that since $w_{\{s\}} = s$ for all $s \in S$, A is also a generating set for W. To distinguish between the two generating sets S and A, we shall call the first the *standard* generating set and the second the *automatic* generating set.

Definition 2. The *standard growth series* and the *automatic growth series* for W are the power series $W_S(t)$ and $W_A(t)$ defined by

$$W_S(t) = \sum_{w \in W} t^{l_S(w)}, \qquad W_A(t) = \sum_{w \in W} t^{l_A(w)}.$$

The standard growth series for arbitrary Coxeter groups is known to be a rational function [Steinberg 1968]. In the right-angled case, this rational function has a particularly simple form in terms of the combinatorics of the simplicial complex K. Recall that the *f-polynomial* of K is the generating function for the numbers of simplices; that is, $f(t) = 1 + f_0 t + f_1 t^2 + \cdots$ where f_i is the number of i-dimensional simplices in K. (Note that if K is the flag completion of a graph Γ, then f_i is the number of $(i+1)$-cliques in Γ.) The following formula can be found, for example, in [Davis 2008, Proposition 17.4.2].

Proposition 3. *Let W be a RACG, and let $f(t)$ be the f-polynomial of the nerve. Then the standard growth series is given by the formula*

$$\frac{1}{W_S(t)} = f\left(\frac{-t}{1+t}\right).$$

This formula shows that, as an invariant for RACGs, the standard growth series is fairly coarse: it is easy to construct examples of nonisomorphic groups with the same standard growth.

Example 4. Let K $(= \Gamma)$ be a tree with n vertices. The f-polynomial of K is $f(t) = 1 + nt + (n-1)t^2$, so using the formula above, one obtains the standard growth series

$$W_S(t) = \frac{(1+t)^2}{1+(2-n)t}$$

for the corresponding Coxeter group. In particular, any two trees with the same number of vertices yield RACGs with the same standard growth series.

The purpose of this paper is to study the automatic growth series $W_A(t)$, which appears to be a much more subtle invariant of the group than $W_S(t)$.

3. The total growth series

Both the automatic and standard growth series for a RACG W can be regarded as specializations of a certain multivariable growth series. This "total" growth series is defined in terms of a certain regular language associated to any simplicial complex K. In the case where K is a flag complex (that is, K is the flag completion of its 1-skeleton), this regular language determines a normal form for the corresponding RACG.

Let K be a simplicial complex and let $\mathcal{A} = K_{>\varnothing}$. For any $\sigma \in K$, we let $\mathrm{St}(\sigma)$ denote the (vertices of the) closed star of σ. That is,

$$\mathrm{St}(\sigma) = \{s \in K \mid \{s\} \cup \sigma \in K\}.$$

The regular language we have in mind for K consists of certain words over the alphabet $\mathcal{A}$. Let $\mathcal{A}^*$ denote the free monoid on $\mathcal{A}$. We shall say that a word $\omega = \sigma_1 \cdots \sigma_n \in \mathcal{A}^*$ is *right-greedy* if $\mathrm{St}(\sigma_{i+1}) \cap \sigma_i = \varnothing$ for all $1 \le i < n$. We then let $\mathcal{L}_K \subset \mathcal{A}^*$ denote the language consisting of all right-greedy words.

Remark 5. It is not hard to see that $\mathcal{L}_K$ is in fact a *regular* language (see, for example, [Epstein et al. 1992], for a definition). We take as our states the set $\mathcal{S} = \mathcal{A} \cup \{\varnothing, \rho\}$ where $\varnothing$ is the start state, and ρ is a single fail state (thus, the set of accept states is $\mathcal{Y} = \mathcal{A} \cup \{\varnothing\}$). We define the transition function $\mu : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ by

- $\mu(\tau, \sigma) = \sigma$ whenever $\tau \in \mathcal{Y}$ and $\mathrm{St}(\sigma) \cap \tau = \varnothing$, and
- $\mu(\tau, \sigma) = \rho$ otherwise.

It follows easily that $\mathcal{L}_K$ is the accepted language of the automaton $(\mathcal{S}, \mathcal{A}, \mu, \mathcal{Y}, \varnothing)$.

For any word $\omega \in \mathcal{A}^*$ and $\sigma \in \mathcal{A}$ we let $n_\sigma(\omega)$ denote the number of occurrences of σ in ω. Let $\boldsymbol{t}$ be an $\mathcal{A}$-tuple $(t_\sigma)_{\sigma\in\mathcal{A}}$ of indeterminates, and for each word ω, we let t^ω be the monomial

$$t^\omega = \prod_{\sigma\in\mathcal{A}} t_\sigma^{n_\sigma(\omega)}.$$

In particular, $t^\varnothing = 1$.

Definition 6. Let K be a simplicial complex and let $\mathcal{L}_K$ be the corresponding right-greedy language. Then the *total growth series* of $\mathcal{L}_K$ is the generating function

$$\mathcal{L}_K(\boldsymbol{t}) = \sum_{\omega\in\mathcal{L}_K} t^\omega.$$

In the event that K is the nerve of a right-angled Coxeter group W (that is, whenever K is a flag complex), we let $\mathcal{L}_W$ denote the language $\mathcal{L}_K$. Note that in this case $\sigma \mapsto w_\sigma$ defines a bijection from $\mathcal{A}$ to the automatic generating set $A \subset W$, and hence induces a surjection $\mathcal{A}^* \to W$. We let $\pi : \mathcal{L}_W \to W$ denote the restriction of this map to the set of right-greedy words. More precisely, if $\omega = \sigma_1 \cdots \sigma_n \in \mathcal{L}_W$, then $\pi(\omega) = w_{\sigma_1} \cdots w_{\sigma_n} \in W$. The following proposition says that $\mathcal{L}_W$ gives a normal form for elements of W. We omit the proof, which follows directly from Tits' solution to the word problem for Coxeter groups [Tits 1969].

Proposition 7. *Let W be a RACG. Then*

(i) *The map $\pi : \mathcal{L}_W \to W$ is a bijection.*

(ii) *For $\omega = \sigma_1 \cdots \sigma_n \in \mathcal{L}_W$, the A-length of $\pi(\omega)$ is n.*

(iii) *For $\omega = \sigma_1 \cdots \sigma_n \in \mathcal{L}_W$, the S-length of $\pi(\omega)$ is $l_S(w_{\sigma_1}) + \cdots + l_S(w_{\sigma_n})$.*

It follows from Proposition 7 that both the standard and automatic growth series for W are specializations of the total growth series for $\mathcal{L}_W$.

Corollary 8. *Let W be a RACG. Then*

(i) *$W_S(t) = \mathcal{L}_W(\boldsymbol{t})$ after the substitutions $t_\sigma = t^{|\sigma|}$, $\sigma \in \mathcal{A}$, and*

(ii) *$W_A(t) = \mathcal{L}_W(\boldsymbol{t})$ after the substitutions $t_\sigma = t$, $\sigma \in \mathcal{A}$.*

In light of this corollary, it makes sense to call the total growth series of the language $\mathcal{L}_W$ the *total growth series of W*.

4. Calculating the total growth series

Let K be an arbitrary finite simplicial complex. In this section we describe a method for computing the total growth series of the right-greedy language $\mathcal{L}_K$. As a corollary we obtain that the total growth series is a rational function.

Let $\mathcal{A} = K_{>\varnothing}$. We let $\mathbb{C}[\mathcal{A}]$ denote the polynomial ring in the indeterminates t_σ, $\sigma \in \mathcal{A}$, and let $\mathbb{C}(\mathcal{A})$ denote the quotient field of rational functions. Similarly, we

let $\mathbb{C}[\![\mathscr{A}]\!]$ denote the ring of formal power series and $\mathbb{C}((\mathscr{A}))$ denote the quotient field. Note that $\mathbb{C}(\mathscr{A})$ is a subfield of $\mathbb{C}((\mathscr{A}))$.

We define a *transition matrix* $M : K \times K \to \mathbb{C}[\mathscr{A}]$ for $\mathscr{L}_K$ as follows:

(1) If $\sigma = \tau = \varnothing$, then $M(\sigma, \tau) = 1$.

(2) If $\mathrm{St}(\sigma) \cap \tau = \varnothing$, then $M(\sigma, \tau) = t_\sigma$.

(3) Otherwise, $M(\sigma, \tau) = 0$.

We let $\mathbb{C}(\mathscr{A})^K$ (respectively $\mathbb{C}((\mathscr{A}))^K$) denote the vector space over $\mathbb{C}(\mathscr{A})$ (resp., $\mathbb{C}((\mathscr{A}))$) with standard basis $\{\boldsymbol{e}_\sigma \mid \sigma \in K\}$, and we regard M as a $(K \times K)$-matrix over $\mathbb{C}(\mathscr{A})$ (resp., over $\mathbb{C}((\mathscr{A}))$). Our goal in this section is to prove the following.

Theorem 9. *Let $\mathscr{L} = \mathscr{L}_K$ be the right-greedy language over the simplicial complex K, and let M be the transition matrix.*

(i) *The $\lambda = 1$ eigenspace of M is 1-dimensional over $\mathbb{C}(\mathscr{A})$, and therefore also over $\mathbb{C}((\mathscr{A}))$.*

(ii) *If $\boldsymbol{v}(\boldsymbol{t}) = (v_\sigma(\boldsymbol{t}))_{\sigma \in K}$ is an eigenvector in $\mathbb{C}((\mathscr{A}))^K$ corresponding to the eigenvalue 1, then $v_\varnothing(\boldsymbol{t}) \neq 0$ and the total growth series for $\mathscr{L}_K$ is given by*

$$\mathscr{L}_K(\boldsymbol{t}) = \sum_{\sigma \in K} \frac{v_\sigma(\boldsymbol{t})}{v_\varnothing(\boldsymbol{t})}.$$

In particular, the total growth series is a rational function which can be computed explicitly via Gaussian elimination on M.

Proof. Since $M(\varnothing, \varnothing) = 1$ and $M(\varnothing, r) = 0$ for $r \neq \varnothing$, the $\varnothing$-row of the matrix $M - 1$ is all zeros. Thus, to prove the first part of the theorem, it suffices to show that the $(K - \varnothing) \times (K - \varnothing)$ minor of $M - 1$ is nonsingular (over $\mathbb{C}(\mathscr{A})$). This submatrix has polynomial entries whose constant terms are all zero off the diagonal and are all -1 on the diagonal. It follows that the determinant of this submatrix must be a polynomial which evaluates to ± 1 when all of the indeterminates are zero, hence the determinant is nonzero in $\mathbb{C}(\mathscr{A})$. This proves that the $\lambda = 1$ eigenspace of M is 1-dimensional *and* that any eigenvector $\boldsymbol{v}(\boldsymbol{t})$ must have $v_\varnothing(\boldsymbol{t}) \neq 0$.

For the second part of the theorem, suppose $\boldsymbol{v}(\boldsymbol{t}) \in \mathbb{C}((\mathscr{A}))^K$ is any $\lambda = 1$ eigenvector. For each $\sigma \in K$, let $\mathscr{L}_\sigma$ denote the set of words in $\mathscr{L}$ ending in σ, and let $u_\sigma(\boldsymbol{t})$ be the total growth series for $\mathscr{L}_\sigma$, that is,

$$u_\sigma(\boldsymbol{t}) = \sum_{\omega \in \mathscr{L}_\sigma} t^\omega.$$

Then

$$\mathscr{L}_K(\boldsymbol{t}) = \sum_{\sigma \in K} u_\sigma(\boldsymbol{t}),$$

and since only the trivial word ends in $\varnothing$, we have $u_\varnothing(\boldsymbol{t}) = 1$. In light of the first part of the theorem, it suffices to show that the vector $\boldsymbol{u}(\boldsymbol{t}) = (u_\sigma(\boldsymbol{t}))_{\sigma \in K} \in \mathbb{C}((\mathcal{A}))^K$ is also an eigenvector for M with eigenvalue 1 (since this will then imply that $u_\sigma(\boldsymbol{t}) = v_\sigma(\boldsymbol{t})/v_\varnothing(\boldsymbol{t})$.)

For any integer $n \geq 0$ we let $\mathcal{L}_\sigma^{(n)}$ be the subset of $\mathcal{L}_\sigma$ consisting of words of length $\leq n$, and let $\boldsymbol{u}^{(n)}(\boldsymbol{t}) = (u_\sigma^{(n)}(\boldsymbol{t}))_{\sigma \in K}$ where $u_\sigma^{(n)}(\boldsymbol{t})$ is the growth series for $\mathcal{L}_\sigma^{(n)}$ (a polynomial of degree n). In particular, if we regard $\boldsymbol{u}(\boldsymbol{t})$ as a power series with vector coefficients, then $\boldsymbol{u}^{(n)}(\boldsymbol{t})$ is the nth partial sum. Now any word $\omega \in \mathcal{L}_\tau^{(n)}$ can be extended to a word $\omega\sigma \in \mathcal{L}_\sigma^{(n+1)}$ precisely when $\mathrm{St}(\sigma) \cap \tau = \varnothing$, and in this case $t^{\omega\sigma} = t^\omega \cdot t_\sigma$. It then follows from the definition of M that

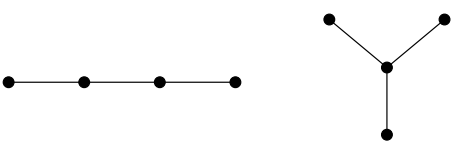
$$M\boldsymbol{u}^{(n)}(\boldsymbol{t}) = \boldsymbol{u}^{(n+1)}(\boldsymbol{t})$$

for all $n \geq 0$. But since $\boldsymbol{u}^{(n)}(\boldsymbol{t})$ is the nth partial sum of $\boldsymbol{u}(\boldsymbol{t})$, this means that $\boldsymbol{u}(\boldsymbol{t})$ must be a $\lambda = 1$ eigenvector of M. □

Corollary 10. *Let W be a RACG. Then the total growth series $\mathcal{L}_W(\boldsymbol{t})$ is a rational function in the indeterminates t_σ, $\sigma \in K_{>\varnothing}$; thus* (by Corollary 8) *the standard and automatic growth series for W are rational functions of the single indeterminate t.*

More importantly, Theorem 9 describes exactly how to obtain these power series.

Example 11. Consider the Coxeter groups W and W' corresponding to these trees:



The matrix M for W (with respect to the ordering $\varnothing$, 1, 2, 3, 4, 12, 23, 34) is

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ t_1 & 0 & 0 & t_1 & t_1 & 0 & 0 & t_1 \\ t_2 & 0 & 0 & 0 & t_2 & 0 & 0 & 0 \\ t_3 & t_3 & 0 & 0 & 0 & 0 & 0 & 0 \\ t_4 & t_4 & t_4 & 0 & 0 & t_4 & 0 & 0 \\ t_{12} & 0 & 0 & t_{12} & t_{12} & 0 & 0 & t_{12} \\ t_{23} & t_{23} & 0 & 0 & t_{23} & 0 & 0 & 0 \\ t_{34} & t_{34} & t_{34} & 0 & 0 & t_{34} & 0 & 0 \end{bmatrix}.$$

We find the total growth series by finding any vector in the nullspace of $M - 1$, dividing by the first entry of that vector, and then adding up the entries of the

vector:

$$\mathscr{L}(\boldsymbol{t}) = \frac{\begin{pmatrix} 1 + t_{34}t_{12} + t_2 + t_3 + t_4 + t_1 + t_{23} + t_{12} + t_{34} \\ -t_3t_4t_2 + t_{12}t_3 + t_4t_2 + t_{23}t_4 + t_{34}t_1 + t_{34}t_2 + t_4t_{12} + t_1t_3 + t_4t_1 \\ -t_1t_3t_2 - t_1t_3t_4t_2 - t_{23}t_{34}t_{12} + t_{23}t_2t_{34}t_1 + t_{23}t_{12}t_3t_4 \\ +t_{23}t_1 + t_{23}t_4t_1 - t_{23}t_1t_3t_4t_2 \end{pmatrix}}{1 - t_4t_{12} - t_{34}t_1 - t_1t_3 - t_{34}t_{12} + t_1t_3t_4t_2 - t_4t_2 - t_4t_1}.$$

By making the substitutions in Corollary 8, we get the automatic growth series

$$W_A(t) = \frac{1 + 5t + t^2 + t^3}{1 - 2t - t^2}.$$

By performing the same steps for W' we find the automatic growth series

$$W'_{A'}(t) = \frac{1 + 5t - 2t^2}{1 - 2t}.$$

Thus, the automatic growth series can tell these groups apart, while the standard growth series cannot (see Example 4).

The example above shows that the automatic growth series is not determined by the f-polynomial (as is the standard growth series). In general, the properties of the complex K that determine the automatic growth series seem to be fairly subtle. However, in the next section we describe a class of simplicial complexes for which the automatic growth series is still determined by the f-polynomial (but even in this case, the formula for the automatic growth series in terms of the f-polynomial is very complicated).

5. Link-regular simplicial complexes

In this section we consider certain simplicial complexes K whose structure allows for a substantial reduction in the recursion defining the language $\mathscr{L}_K$. For this class of simplicial complexes, both the automatic growth series *and* the standard growth series are determined by the f-polynomial of K. We use this fact to obtain nonisomorphic Coxeter groups that have the same automatic growth series. To define our regularity condition, we first recall the that the *link* of a simplex σ in K is, by definition, the subcomplex of K consisting of all $\tau \in K$ such that $\sigma \cup \tau \in K$ and $\sigma \cap \tau = \varnothing$.

Definition 12. Let K be a finite flag simplicial complex of dimension d. We say that K is *link-regular* if for every $0 \leq j \leq d$, the link of every j-simplex $\sigma \in K$ has the same f-polynomial. In this case, we let $F_j(t)$ denote the f-polynomial of the link of a j-simplex in K. (Since we regard $\dim \varnothing = -1$, the f-polynomial for K itself is $F_{-1}(t)$.)

Proposition 13. *Let K be a link-regular simplicial complex of dimension d. Then*

$$F_j(t) = \frac{f^{(j+1)}(t)}{(j+1)!\, f_j}$$

where $f(t) = 1 + f_0 t + \cdots + f_d t^{d+1}$ is the f-polynomial for K and $f^{(j+1)}(t)$ denotes its $(j+1)$st derivative.

Proof. Let f_k^j denote the number of k-simplices in the link of a j-simplex, so

$$F_j(t) = 1 + f_0^j t + f_1^j t^2 + \cdots + f_{d-j}^j t^{d-j+1}.$$

Any k-simplex τ in the link of a j-simplex σ determines a $(k+j+1)$-simplex $\sigma \cup \tau$ in K, and there are $\binom{k+j+1}{j+1}$ such ways to obtain this simplex. This gives the relation

$$f_j f_j^k = \binom{k+j+1}{j+1} f_{k+j+1}.$$

Solving for f_j^k and substituting for the coefficients in the polynomial $F_j(t)$ gives the desired identity. □

By Proposition 3, we know that two Coxeter groups will have the same standard growth series if their nerves have the same f-polynomial. If in addition, we assume their nerves are link-regular, we obtain the following theorem.

Theorem 14. *Let W and W' be two right-angled Coxeter groups with corresponding nerves K and K'. Assume further that K and K' are both link-regular and have the same f-polynomial. Then W and W' have the same automatic growth series.*

Proof. Let K be a link-regular simplicial complex of dimension d. By Proposition 13, it suffices to show that the automatic growth series depends only on the polynomials $F_j(t)$, $-1 \le j \le d$.

Let $\mathcal{L} = \mathcal{L}_K$, and for each $i \in \mathbb{Z}_{\ge 0}$, let $B_i(m)$ denote the number of words of length m in $\mathcal{L}$ that end in an i-simplex. Then the automatic growth series can be written as the double sum

$$W_A(t) = 1 + \sum_{m=1}^{\infty} \sum_{i=0}^{d} B_i(m) t^m.$$

We form a recursion relation for $B_i(m)$, $m \ge 2$, as follows. Any word ω of length m that ends in an i-simplex is obtained by multiplying a word ω' of length $m-1$ by some i-simplex σ. Given $\omega' \in W$ of length $m-1$ ending in τ, let $\beta_i(\tau)$ be the number of i-simplices σ such that $\tau\sigma \in \mathcal{L}$. That is, $\beta_i(\tau)$ is the number of i-simplices σ such that $St(\sigma) \cap \tau = \varnothing$. Since K is link-regular, $\beta_i(\tau)$ depends only on the dimension of τ and not on τ itself. We denote this number by β_{ij} where

$j = \dim(\tau)$. The number of words of length m ending in an i-simplex is then given by the recurrence

$$B_i(m) = \beta_{i0} B_0(m-1) + \beta_{i1} B_1(m-1) + \cdots + \beta_{id} B_d(m-1)$$

for $m \geq 2$. A straightforward inclusion-exclusion argument gives an explicit formula for β_{ij} in terms of the f-polynomials of the various links; more explicitly, if one writes the polynomials of the links as

$$F_j(t) = 1 + f_0^j t + f_1^j t^2 + \cdots + f_{d-j}^j t^{d-j+1},$$

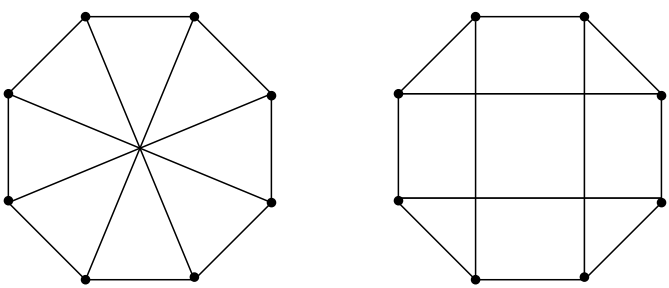
then the β_{ij} are given by

$$\beta_{ij} = f_i + \sum_{q=1}^{j+1} (-1)^q \binom{j+1}{q} \sum_{p=0}^{q} \binom{q}{p} f_{i-p}^{q-1}.$$

Since the coefficients of the recurrence relations and the initial values $B_i(1)$ (which equals $f_i^{-1} = f_i$) can all be expressed explicitly in terms of the f-polynomials of the links, so can all of the $B_i(m)$s. Thus $W_A(t)$ depends only on the f-polynomial of K. □

We now give several pairs of examples of nonisomorphic RACGs that have the same automatic (and standard) growth series.

Example 15. The Coxeter groups corresponding to these two graphs have the same automatic growth series:



To see this, note that any vertex-regular graph with no 3-cycles is a 1-dimensional link-regular simplicial complex; hence these graphs are both link-regular and have the same f-polynomial $f(t) = 1 + 8t + 12t^2$. By Theorem 14, they have the same automatic growth series, which we can compute explicitly as

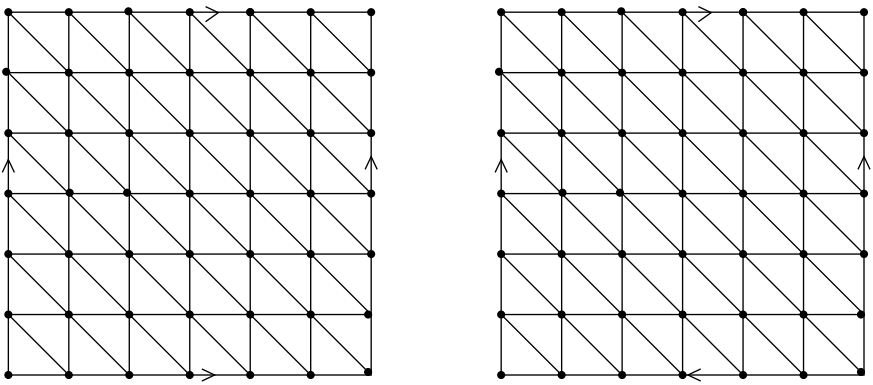
$$W_A(t) = \frac{1 + 9t + 2t^2}{1 - 11t + 10t^2}.$$

The graphs are clearly not isomorphic (the one on the right is bipartite, the other is not).

It is easy to generalize the example above to other pairs of Coxeter groups with the same automatic growth series by adding diagonals to a $2n$-gon to get a regular graph. As long as there are no 3-cycles, these graphs will have simplicial complexes that are link-regular. (In fact, one can construct such a pair Γ_{hyp} and Γ_{euc} so that the first has no 4-cycles while the second has 4-cycles. The corresponding Coxeter groups will have the same standard and automatic growth series even though one is a Gromov hyperbolic group and the other is not.)

Constructing examples with $\dim K \geq 2$ is a little more subtle:

Example 16. Let Γ_1 and Γ_2 be the following two graphs, embedded on the torus and the Klein bottle, respectively:



They are not isomorphic (Γ_1 is three-colorable and Γ_2 is not), so they correspond to nonisomorphic Coxeter groups. On the other hand, the corresponding nerves are precisely the triangulations shown in the figure. These simplicial complexes are clearly link-regular and have the same f-polynomials, hence have the same automatic growth series.

We have seen examples of two nonisomorphic right-angled Coxeter groups where the standard growth series are the same, but the automatic growth series are different and where both the standard and the automatic growth series are the same. This suggests the following question, for which we do not yet have an answer.

Question 17. Are there two right-angled Coxeter groups with the same automatic growth series but different standard growth series?

6. Comparing automatic and standard growth rates

Given a power series $\sum a_n t^n$ (over $\mathbb{C}$), we define its *growth rate* γ to be

$$\gamma = \limsup_n \sqrt[n]{|a_n|}.$$

Equivalently, by the root test, $\gamma = 1/\rho$ where ρ is the radius of convergence. If a power series is a rational function say $p(t)/q(t)$, its growth rate can therefore be calculated from

$$\frac{1}{\gamma} = \rho = \min(|r_1|, |r_2|, \ldots, |r_n|),$$

where $r_1, \ldots, r_n$ are the roots of the denominator $q(t)$.

In this section, we consider the growth rates for the standard and automatic growth series of a RACG. We denote them by γ_S and γ_A, respectively. Our first observation is following.

Proposition 18. *If (W, S) is a right-angled Coxeter group, then $\gamma_S \leq \gamma_A$. More precisely*:

(i) *If W is finite, $\gamma_S = \gamma_A = 0$.*

(ii) *If W is infinite and virtually abelian, $\gamma_S = \gamma_A = 1$.*

(iii) *If W is not virtually abelian, $1 < \gamma_S \leq \gamma_A$.*

Proof. Statement (i) is clear. Statement (ii) follows from the fact that for an arbitrary infinite Coxeter group W and any generating set, the radius of convergence for the growth series is 1 if and only if W is virtually abelian [Davis 2008, Proposition 17.2.1]. This also ensures that if W is not virtually abelian, then the two growth rates γ_S and γ_A are both > 1. For the final inequality in (iii), suppose W is not virtually abelian. Let $BW_S(t)$ (respectively, $BW_A(t)$) denote the growth series for *balls* in W with respect to the S-length (resp., A-length). In other words, $BW_S(t) = \sum_n b_n t^n$ where b_n is the number of elements in W of S-length $\leq n$. A geometric series calculation shows that

$$BW_S(t) = \frac{W_S(t)}{1-t}, \qquad BW_A(t) = \frac{W_A(t)}{1-t}.$$

Since the growth rate of $W_S(t)$ (respectively, $W_A(t)$) is > 1, $BW_S(t)$ (resp., $BW_A(t)$) will have the same growth rate as $W_S(t)$ (resp., $W_A(t)$). On the other hand since $S \subset A$, the A-length of an element is always $\leq$ the S-length, hence the terms of the series $BW_A(t)$ are all $\geq$ the terms of the series $BW_S(t)$. It follows that the growth rate for $BW_S(t)$ is $\leq$ the growth rate for $BW_A(t)$; in other words, $\gamma_S \leq \gamma_A$. □

The following examples illustrate these three cases.

Example 19. Let W be the Coxeter group whose graph Γ is the complete graph K_n. Then W is the (finite) group $(\mathbb{Z}_2)^n$ and the standard and automatic growth series are

$$W_S(t) = (1+t)^n = 1 + nt + \cdots + nt^{n-1} + t^n,$$
$$W_A(t) = 1 + (2^n - 1)t.$$

Since these are both polynomials, $\gamma_S = \gamma_A = 0$.

Example 20. Let (W, S) be the Coxeter group corresponding to the graph $K_4 - e$, the complete graph on 4 vertices with one edge removed. The standard growth series and automatic growth series are

$$W_S(t) = \frac{(1+t)^3}{1-t} = 1 + 4t + 7t^2 + 8t^3 + 8t^4 \ldots,$$
$$W_A(t) = \frac{1 + 10t - 3t^2}{1-t} = 1 + 11t + 8t^2 + 8t^3 + 8t^4 \ldots.$$

Hence both have growth rates equal to 1. In this case the group W is a product of $(\mathbb{Z}_2)^2$ with the infinite dihedral group (hence is virtually $\mathbb{Z}$).

Example 21. Recall the Coxeter group W from Example 11. The growth series are given by

$$W_S(t) = \frac{(1+t)^2}{1-2t},$$
$$W_A(t) = \frac{1 + 5t + t^2 + t^3}{1 - 2t - t^2},$$

so we obtain

$$\gamma_S = 2, \qquad \gamma_A = \frac{1}{-1+\sqrt{2}} \approx 2.41.$$

It turns out that the inequality $\gamma_S \leq \gamma_A$ can also be deduced from a stronger statement about the relative growth rates of the coefficient sequences in $W_S(t)$ and $W_A(t)$. Namely, let $\mathcal{L} = \mathcal{L}_W$ be the language defining the right-greedy normal form for W, and let S_n (respectively, A_n) denote the set of words in $\mathcal{L}$ with S-length n (resp., A-length n). Then the relevant growth series are given by

$$W_S(t) = 1 + |S_1|t + |S_2|t^2 + |S_3|t^3 + \cdots,$$
$$W_A(t) = 1 + |A_1|t + |A_2|t^2 + |A_3|t^3 + \cdots.$$

It is clear that $\gamma_S \leq \gamma_A$ would be implied by the stronger statement $|S_n| \leq |A_n|$ for all n. Of course, this statement is false if K is a simplex (that is, if W is finite) as in Example 19, but otherwise, we have the following.

Theorem 22. *Let (W, S) be an infinite right-angled Coxeter group. Then the coefficients of the standard and automatic growth series satisfy $|S_n| \leq |A_n|$ for all $n \geq 0$.*

Before proving the theorem, we recall some terminology for simplicial complexes. The *m-skeleton* $K^{(m)}$ of a simplicial complex K is the subcomplex consisting of all $\sigma \in K$ such that $\dim(\sigma) \leq m$. Thus, the vertex set of K is denoted $K^{(0)}$ and when K is the flag completion of a graph Γ, the 1-skeleton $K^{(1)}$ is precisely the graph Γ.

Given a simplex $\sigma \in K$, we let $\hat{\sigma}$ denote the subcomplex consisting of σ and its faces. And given two simplicial complexes K_1 and K_2, we define their *join* $K_1 * K_2$ to be the simplicial complex with vertex set $K_1^{(0)} \cup K_2^{(0)}$ and simplices given by

$$K_1 * K_2 = \{\sigma_1 \cup \sigma_2 \mid \sigma_1 \in K_1 \text{ and } \sigma_2 \in K_2\}.$$

Lemma 23. *Let K be a flag simplicial complex. Then there exists a $\sigma \in K$ and a subcomplex $L \subseteq K$ such that $K = \hat{\sigma} * L$ and for any $\tau \in L$, $\mathrm{St}(\tau) \neq L^{(0)}$.*

Proof. Let $I = \{\sigma_1, \sigma_2, \ldots, \sigma_q\}$ be the set of all σ_i such that $\mathrm{St}(\sigma_i) = K^{(0)}$. Then all of the vertices of σ_i are adjacent to all of the vertices of σ_j for $1 \leq i \leq j \leq q$. Since K is a flag complex this means that $\sigma = \sigma_1 \cup \sigma_2 \cup \ldots \cup \sigma_q$ is a simplex in K. Let L be the induced subcomplex on the vertex set $K^{(0)} - \sigma$. Since K is a flag complex and all of the vertices in L are adjacent to all of the vertices in σ, we know that $\hat{\sigma} * L = K$. If there exists a $\tau \in L$ such that $\mathrm{St}(\tau) = L^{(0)}$, then in fact $\mathrm{St}(\tau) = K^{(0)}$ by the definition of the join, contradicting our definition of I. Hence, L has the desired property. □

Proof of Theorem 22. To show that $|S_n| \leq |A_n|$, it suffices to construct an injective map $S_n \to A_n$. By Lemma 23, we can write K in the form $K = \hat{\sigma} * L$ where $L \neq \{\varnothing\}$ and for any $\tau \in L$, $\mathrm{St}(\tau) \neq L^{(0)}$. Let $\omega = \sigma_1\sigma_2 \cdots \sigma_p$ be an element in S_n. Then for each i, $\sigma_i = \tau_i \cup \sigma_i'$ for some $\tau_i \in \hat{\sigma}$ and $\sigma_i' \in L$. If $\sigma_i' = \varnothing$ for some $1 < i \leq p$, $\mathrm{St}(\sigma_{i+1}) \cap \sigma_i \neq \varnothing$, contradicting the fact that $\omega \in \mathcal{L}$. Therefore, $\sigma_i' \neq \varnothing$ for all $1 < i \leq p$.

In order to map ω to an element of A_n, we will append $n - p$ 0-simplices to the front of ω while keeping it in $\mathcal{L}$. There are two cases.

Case 1. $\sigma_1' \neq \varnothing$. We know that $\mathrm{St}(\sigma_1') \neq L^{(0)}$ so $\mathrm{St}(\sigma_1) \neq K^{(0)}$. Therefore, there exists a $v \in L^{(0)}$ such that $\mathrm{St}(\sigma_1) \cap \{v\} = \varnothing$, this means that $\{v\}\omega$ is still in $\mathcal{L}$. Since $v \in L^{(0)}$, there exists a $u \in L^{(0)}$ such that $\mathrm{St}(\{v\}) \cap \{u\} = \varnothing$. It follows that the words, $\{v\}\omega$, $\{u\}\{v\}\omega$, $\{v\}\{u\}\{v\}\omega$, ... are all in $\mathcal{L}$, so by adjoining the alternating word $\nu = \{u\}\{v\}\cdots\{v\}$ (or $\nu = \{v\}\{u\}\{v\}\cdots\{v\}$ depending on the parity of $n - p$) of length $(n - p)$ to the beginning of ω, we obtain an element $\nu\omega \in A_n$.

Case 2. $\sigma_1' = \varnothing$. In this case $\omega = \tau_1\sigma_2 \cdots \sigma_p$. Since $\tau_1 \in \hat{\sigma}$, this means that $\tau_1 \subset \mathrm{St}(\sigma_2)$ contradicting the fact that $\omega \in \mathcal{L}$. It follows that in this case, we must have $p = 1$ and $\omega = \tau_1$. Put $\tau = \tau_1$. Since $\omega \in S_n$ we know $|\tau| = n$. Every element in τ is adjacent to every vertex in L so pick a $v \in L^{(0)}$. Then $\{v\} \cup \tau \in \mathcal{L}$. Moreover, $\mathrm{St}(\{v\} \cup \tau) \neq K^{(0)}$, so there exists a $u \in L^{(0)}$ such that $\mathrm{St}(\{v\} \cup \tau) \cap \{u\} = \varnothing$. In particular, u and v are not adjacent, so if we let ν be the alternating word $\{u\}\{v\}\cdots\{u\}$ (or $\{v\}\{u\}\{v\}\cdots\{u\}$ depending on the parity of n) of length $n - 1$ to the beginning of $\{v\} \cup \tau$, we obtain an element $\nu(\{v\} \cup \tau) \in A_n$.

Since two words in $\mathcal{L}$ with different endings must be different, the map $S_n \to A_n$ given by $\omega \mapsto v\omega$ (Case 1) or $\omega \mapsto v(\{v\} \cup \tau)$ (Case 2) is injective. □

References

[Davis 2002] M. W. Davis, "Nonpositive curvature and reflection groups", pp. 373–422 in *Handbook of geometric topology*, edited by R. J. Daverman and R. B. Sher, North-Holland, Amsterdam, 2002. MR 2002m:53061 Zbl 0998.57002

[Davis 2008] M. W. Davis, *The geometry and topology of Coxeter groups*, London Mathematical Society Monographs Series **32**, Princeton University Press, Princeton, NJ, 2008. MR 2008k:20091 Zbl 1142.20020

[Epstein et al. 1992] D. B. A. Epstein, J. W. Cannon, D. F. Holt, S. V. F. Levy, M. S. Paterson, and W. P. Thurston, *Word processing in groups*, Jones and Bartlett, Boston, MA, 1992. MR 93i:20036 Zbl 0764.20017

[Niblo and Reeves 1998] G. A. Niblo and L. D. Reeves, "The geometry of cube complexes and the complexity of their fundamental groups", *Topology* **37**:3 (1998), 621–633. MR 99a:20037 Zbl 0911.57002

[Radcliffe 2001] D. Radcliffe, *Unique presentation of Coxeter groups and related groups*, Ph.D. thesis, University of Wisconsin, Milwaukee, 2001. Ph.D. thesis.

[Steinberg 1968] R. Steinberg, *Endomorphisms of linear algebraic groups*, Memoirs of the American Mathematical Society **80**, American Mathematical Society, Providence, 1968. MR 37 #6288 Zbl 0164.02902

[Tits 1969] J. Tits, "Le problème des mots dans les groupes de Coxeter", pp. 175–185 in *Symposia Mathematica* (INDAM, Rome, 1967/68), vol. 1, Academic Press, London, 1969. MR 40 #7339 Zbl 0206.03002

Received: 2008-03-11 Revised: 2009-09-08 Accepted: 2009-09-26

reglover@email.unc.edu *Department of Mathematics, The University of North Carolina, CB No. 3250, Phillips Hall, Chapel Hill, NC 27599-3250, United States*

rscott@schubert.scu.edu *Department of Mathematics and Computer Science, Santa Clara University, 500 El Camino Real, Santa Clara, CA 95053-0290, United States*
http://schubert.scu.edu/rscott

INVOLVE 2:4(2009)

Contributions to Seymour's second neighborhood conjecture

James Brantner, Greg Brockman, Bill Kay and Emma Snively

(Communicated by Vadim Ponomarenko)

Let D be a simple digraph without loops or digons. For any $v \in V(D)$ let $N_1(v)$ be the set of all nodes at out-distance 1 from v and let $N_2(v)$ be the set of all nodes at out-distance 2. We show that if the underlying graph is triangle-free, there must exist some $v \in V(D)$ such that $|N_1(v)| \leq |N_2(v)|$. We provide several properties a "minimal" graph which does not contain such a node must have. Moreover, we show that if one such graph exists, then there exist infinitely many.

1. Introduction

In this article, we consider only simple nonempty finite digraphs (those containing no loops or multiple edges and having a nonempty vertex set), unless stated otherwise. We also require that our digraphs contain no digons, that is, if D is a digraph then $(u, v) \in E(D) \Rightarrow (v, u) \notin E(D)$. If i is a positive integer, we denote the ith neighborhood of a vertex u in D by $N_{i,D}(u) = \{v \in V(D) | \operatorname{dist}_D(u, v) = i\}$, where $\operatorname{dist}_D(u, v)$ is the length of the shortest directed path from u to v in D (if there is no directed path from u to v, we set $\operatorname{dist}_D(u, v) = \infty$). If D is clear from context, we simply write $N_i(u)$ and $\operatorname{dist}(u, v)$. We will also consider the ith in-neighborhood of a node $N_{-i}(u) = \{v \in V(D) | \operatorname{dist}(v, u) = i\}$. In addition, if $V' \subseteq V(D)$, we let $D[V']$ be the subgraph of D induced by V'.

Graph theorists will be familiar with the following conjecture by Seymour.

Conjecture 1.1 (Seymour's second neighborhood conjecture). *Let D be a directed graph. Then there exists a vertex $v_0 \in V(D)$ such that $|N_1(v_0)| \leq |N_2(v_0)|$.*

Dean and Latka [1995] conjectured this to be true when D is a tournament. Dean's conjecture was subsequently proven by Fisher [1996]. Further, Kaneko and Locke [2001] showed Conjecture 1.1 to be true if the minimum out-degree of vertices in D is less than 7, while Cohn, Wright and Godbole [Cohn et al. 2009]

MSC2000: 05C20.

Keywords: graph theory, second neigbhorhood conjecture, graph properties, open problems in graph theory.

showed that it holds for random graphs almost always. Finally, Fidler and Yuster [2007] proved that Conjecture 1.1 holds for graphs with minimum out-degree $|V(D)| - 2$, tournaments minus a star, and tournaments minus a subtournament. While over the years there have been several attempts at a proof of Conjecture 1.1, none of these has yet been successful.

For completeness, we introduce the related Caccetta–Häggkvist conjecture.

Conjecture 1.2 (Caccetta–Häggkvist). *If D is a directed graph with minimum out-degree at least $|V(D)|/k$, then D has a directed cycle of length at most k.*

Conjecture 1.1 would imply the $k = 3$ case of Conjecture 1.2. Much work has been done on Conjecture 1.2, including an entire workshop sponsored by the American Institute of Mathematics and the National Science Foundation, but still both Conjectures 1.1 and 1.2 remain wide open.

In this paper, we will show that Conjecture 1.1 holds for digraphs where the underlying graph is triangle-free. We then take a different tack and provide conditions that must be satisfied by any appropriately-defined minimal counterexample to Seymour's second neighborhood conjecture.

2. Definitions

Definition 2.1. Suppose that D is digraph and $u \in V(D)$. We say that u is *satisfactory* if $|N_1(u)| \leq |N_2(u)|$. Also, u is a *sink* if $|N_1(u)| = 0$. Note that a sink is trivially satisfactory.

Definition 2.2. Let $\mathscr{A}$ be the set of *Seymour counterexamples*, i.e., simple directed graphs D with no satisfactory vertices (in other words, counterexamples to Seymour's second neighborhood conjecture). Let

$$\mathscr{A}' = \{D| \ |E(D)| = \min_{H \in \mathscr{A}} |E(H)|\}$$

be the set of graphs in $\mathscr{A}$ with the fewest number of edges. Finally, let $\mathscr{A}'' = \{D| \ |V(D)| = \min_{H \in \mathscr{A}'} |V(H)|\}$ be the set of graphs in $\mathscr{A}'$ with the fewest number of vertices. We will refer to any element of $\mathscr{A}''$ as a *minimal counterexample*. Note that $\mathscr{A}''$ is empty if and only if Conjecture 1.1 is true.

Definition 2.3. Define $A_{s,G}(u) = |N_1(u)| - |N_2(u)|$ to be the *antisatisfaction* of u. As usual, if G is clear from context, we simply write $A_s(v)$. Notice that u is satisfactory if and only if $A_s(u) \leq 0$.

Definition 2.4. Again let D be a digraph. If $(u, v) \in E(D)$, we say that edge (u, v) is the *base* of a transitive triangle if u and v share a common first neighbor; that is, $|N_1(u) \cap N_1(v)| \geq 1$ (see Figure 1).

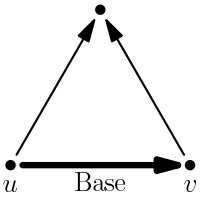


Figure 1. Demonstration of an edge that is the base of a transitive triangle.

If, for distinct $t, u, v, w \in V(D)$, we have $(t, u), (u, w), (t, v), (v, w) \in E$ then we call $\{(t, u), (u, w), (t, v), (v, w)\}$ a *Seymour diamond*. We say the edges (t, u), (t, v) are the *bases* of the Seymour diamond (see Figure 2).

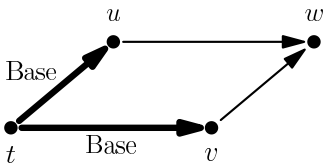


Figure 2. Demonstration of the bases of a Seymour diamond.

3. Directed cycles and underlying girth

In this section we show that certain classes of graphs satisfy Seymour's second neighborhood conjecture. The following theorem shows that directed cycles are necessary for a graph to be a Seymour counterexample.

Observation 3.1. *If a digraph contains no directed cycles, then it must have a satisfactory vertex.*

Proof. Let D be a directed graph. Suppose that D contains no satisfactory vertices. Then D has no sink, as noted in Definition 2.1. Thus if $u \in V(D)$, we have $|N_1(u)| \geq 1$. Now pick an arbitrary vertex $v_0 \in V(D)$, and consider the infinite sequence $\{v_i\}_{i=0}^{\infty}$ defined recursively by $v_{i+1} \in N_1(v_i)$ for $i \geq 0$. Since V is finite, we then have that there exist some $r \neq s$ such that $v_r = v_s$. Then we note that the sequence of edges $(v_r, v_{r+1}), (v_{r+1}, v_{r+2}), \ldots, (v_{s-1}, v_s = v_r)$ defines a dicycle in D, thus completing our proof. □

Recall that the girth of an undirected graph is the length of its shortest cycle. We show that any Seymour counterexample must have underlying girth of exactly 3:

Theorem 3.2. *Let G be a simple graph with girth strictly larger than 3. Then any orientation of G will result in a directed graph with a satisfactory vertex.*

Proof. Let D be any orientation of G. Clearly there must exist some vertex v_0 with minimal out-degree. If $|N_1(v_0)| = 0$, then v_0 is a sink and hence a satisfactory vertex. Otherwise, let $v_1 \in N_1(v_0)$. By construction, we have that $|N_1(v_1)| \geq$

$|N_1(v_0)|$. Furthermore, the underlying graph has girth at least 4, so $|N_1(v_0) \cap N_1(v_1)| = 0$. Thus, $|N_2(v_0)| \geq |N_1(v_1)| \geq |N_1(v_0)|$, and by definition v_0 is a satisfactory vertex. □

Remark. A similar argument will show that any digraph D which has no transitive triangle as a subgraph must have a satisfactory vertex. We will prove a stronger version of this result in the following section.

4. Properties of counterexamples to Seymour's second neighborhood conjecture

To this point, we have been showing that classes of graphs satisfy Conjecture 1.1. In this section we reverse course and explore necessary properties of the minimal counterexample graphs of $\mathscr{A}''$ from Definition 2.2.

Theorem 4.1. *Suppose $\mathcal{M} \in \mathscr{A}''$.*

(i) *$\mathcal{M}$ is strongly connected.*

(ii) *For each $u \in V(\mathcal{M})$, $A_s(u) \in \{1, 2\}$.*

(iii) *For every edge $e = (u, v) \in E(\mathcal{M})$, there exists a path of length 1 or 2 avoiding e from u to all but at most 1 element of $\{v\} \cup N_1(v)$.*

(iv) *Every edge of $\mathcal{M}$ is the base of either a transitive triangle or a Seymour diamond.*

(v) *For any node $u \in V(\mathcal{M})$, there exists a node $v \in N_{-1}(u)$ such that $A_s(v) = 1$.*

(vi) *There exists a cycle $C = (v_1, v_2), (v_2, v_3), \ldots, (v_k, v_1)$ in $\mathcal{M}$ such that for $1 \leq i \leq k$, we have that $A_s(v_i) = 1$.*

Proof.
(i) Let D be a digraph with $u \in V(D)$. We define

$$W_D(u) = \{v \mid \operatorname{dist}(u, v) \neq \infty\}$$

to be the *reachable vertices from u* with respect to D. If D is clear from context, we simply write $W(u)$. Pick an arbitrary node u from the vertex set of $\mathcal{M}$. Now consider $\mathcal{M}' = \mathcal{M}[W(u)]$. We now pick an arbitrary node $v \in W(u)$. Clearly, $N_{1,\mathcal{M}}(v) \subseteq W(u)$ and $N_{2,\mathcal{M}}(v) \subseteq W(u)$. But this implies that

$$A_{s,\mathcal{M}'} = |N_{1,\mathcal{M}'}(v)| - |N_{2,\mathcal{M}'}(v)| = |N_{1,\mathcal{M}}(v)| - |N_{2,\mathcal{M}}(v)| = A_{s,\mathcal{M}},$$

and hence v is satisfactory in $\mathcal{M}'$ if and only if v is satisfactory in $\mathcal{M}$. Since by construction $\mathcal{M}$ contains no satisfactory vertices, v cannot be satisfactory in $\mathcal{M}'$. Thus $\mathcal{M}'$ contains no satisfactory vertices. But $\mathcal{M}'$ is a subgraph of $\mathcal{M}$, and so by minimality of $\mathcal{M}$ we have that $\mathcal{M} = \mathcal{M}'$. Since u was arbitrary, we are done.

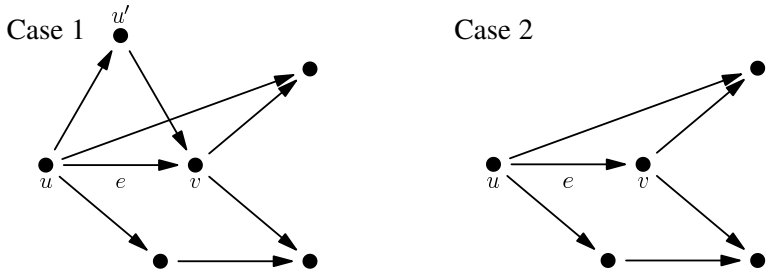


Figure 3. Two possible cases resulting from deleting an edge from $\mathcal{M}$. In case 1, there is a length 2 path from u to v, while in case 2 no such path exists. Note that it is possible that deleting e will increase the size of u's second neighborhood, as shown in case 1.

(ii) Pick an arbitrary edge $e = (u, v) \in E(\mathcal{M})$. Consider the digraph M obtained by deleting e from $\mathcal{M}$. Since M has fewer edges than $\mathcal{M}$, we have that M contains a satisfactory vertex. For each vertex $w \in V(\mathcal{M})$, we note that $|N_{1,M}(w)| = |N_{1,\mathcal{M}}(w)|$ unless $w = u$, in which case $|N_{1,M}(u)| = |N_{1,\mathcal{M}}(u)| - 1$. Furthermore, we have that $|N_{2,M}(w)| \leq |N_{2,\mathcal{M}}(w)|$, except if $w = u$, in which case we have that $|N_{2,M}(u)| \leq |N_{2,\mathcal{M}}(u)| + 1$. (See Figure 3.)

Thus, we obtain that in M for $w \neq u \in V(M)$, $A_{s,M}(w) \geq A_{s,\mathcal{M}}(w)$, and hence all vertices in M besides u are not satisfactory. Thus by process of elimination we have that u is satisfactory in M. Thus

$$0 \geq A_{s,M}(u) = |N_{1,M}(u)| - |N_{2,M}(u)| \geq (|N_{1,\mathcal{M}}(u)| - 1) - (|N_{2,\mathcal{M}}(u)| + 1),$$

and hence we have that $0 < A_{s,\mathcal{M}}(u) = |N_{1,\mathcal{M}}(u)| - |N_{2,\mathcal{M}}(u)| \leq 2$. Result (ii) follows immediately.

(iii) Pick an arbitrary edge $e = (u, v) \in E(\mathcal{M})$. Consider the graph M obtained by deleting e from $\mathcal{M}$. We see that $|N_{2,M}(u)| \geq |N_{2,\mathcal{M}}(u)|$, since otherwise $A_{s,M}(u) \leq 0$ and u is not satisfactory in M, a contradiction. Consider now $X = N_{2,M}(u) \setminus N_{2,\mathcal{M}}(u)$. We note that $X \subseteq \{v\}$, since v is the only vertex that could have been added to u's second neighborhood in M (case 1 in Figure 3). Thus we see that

$$|N_{2,\mathcal{M}}(u) \setminus N_{2,M}(u)| \leq 1,$$

with equality only if $v \in N_{2,M}(u)$.

Note that $N_{1,\mathcal{M}}(v) \subseteq N_{1,\mathcal{M}}(u) \cup N_{2,\mathcal{M}}(u)$. Let $Y = N_{1,\mathcal{M}}(u) \cap N_{1,\mathcal{M}}(v)$ and $Z = N_{2,\mathcal{M}}(u) \cap N_{1,\mathcal{M}}(v)$. For $y \in Y$, we clearly have a path of length 1 from u to y avoiding e (namely the edge (u, y)). If $|N_{2,\mathcal{M}}(u) \setminus N_{2,M}(u)| = 0$, then for $z \in Z$, we have a path of length 2 from u to z in M, and considering this path in $\mathcal{M}$ yields

a path from u to z avoiding e. And finally, if $|N_{2,\mathcal{M}}(u) \setminus N_{2,M}(u)| = 1$, then we have a path of length 2 from u to z in M for all but 1 vertex in Z, and as before we have a corresponding path from u to z avoiding e. But in this case, there is a path of length 2 from u to v avoiding e, and hence we have obtained the desired result.

(iv) Paths of length 1 from u to $v' \in N_1(v)$ yield transitive triangles with e as the base, and paths of length 2 from u to $v' \in \{v\} \cup N_1(v)$ yield Seymour diamonds with e as one of the bases. By part 3, at least one of these structures exists, and hence we are done.

(v) In $\mathcal{M}$, pick an arbitrary vertex u. Delete this vertex and label the resulting graph M. Then in a similar manner to before, one of the nodes in $N_{-1,\mathcal{M}}(u)$ must be satisfactory in M by vertex minimality of $\mathcal{M}$. Label this node t. Since $|N_{1,M}(t)| = |N_{1,\mathcal{M}}(t)| - 1$, t is satisfactory, and $|N_{2,M}(t)| \subseteq |N_{2,\mathcal{M}}(t)|$ (note that in contrast to deleting an edge, deleting a vertex does not allow any vertices to add nodes to their second neighborhoods), we see that we must have $|N_{2,M}(t)| = |N_{2,\mathcal{M}}(t)|$. It is then necessary that $A_{s,\mathcal{M}}(t) = 1$. Since u was arbitrary, we have obtained the desired result.

(vi) We apply the same technique as we used Observation 3.1. We present a brief sketch of our proof: by part (v), each node in $\mathcal{M}$ has an in-neighbor having antisatisfaction of exactly 1. If we begin at an arbitrary vertex and choose one of its in-neighbors having antisatisfaction of exactly 1, do the same for the resulting vertex, and iterate this process, at some point we must arrive back at a vertex we have already visited. Thus we have constructed a dicycle of nodes having antisatisfaction exactly 1. □

We now extend some of our results from the previous theorem. In particular, we turn to a count of the number of transitive triangles and Seymour diamonds that certain edges must belong to.

Theorem 4.2. *If $\mathcal{M} \in A''$, suppose that $e = (u, v) \in E(\mathcal{M})$ and $|N_1(u)| \leq |N_1(v)|$. Then e must be the base of at least $|N_1(v)| - |N_1(u)| + 1$ transitive triangles and the base of at least $|N_1(v)| - |N_1(u)| + 1$ Seymour diamonds.*

Proof. Since $N_1(v) \setminus (N_1(u) \cap N_1(v)) \subseteq N_2(u)$, we have

$$|N_2(u)| \geq |N_1(v)| - |N_1(u) \cap N_1(v)|.$$

But since $\mathcal{M}$ contains no satisfactory vertices, we have that $|N_2(u)| < |N_1(u)|$. By transitivity, we obtain $|N_1(v)| - |N_1(v) \cap N_1(u)| < |N_1(u)|$. It then follows that $|N_1(v)| - |N_1(u)| < |N_1(v) \cap N_1(u)|$, but $|N_1(v) \cap N_1(u)|$ is the number of transitive triangles having base e, so we have proved the first half of the theorem.

To prove the second half of the theorem, we consider the following cases.

Case 1. Suppose there exists a vertex u' such that $(u, u'), (u', v) \in E(\mathcal{M})$. By part (iii) of Theorem 4.1, we know that u must be connected to at least $|N_1(v)| - 1$ elements of $N_1(v)$ via a path of length 1 or 2 avoiding e. But we see that u is adjacent to at most $|N_1(u) - 2|$ nodes in $N_1(v)$. Subtracting, we see that u is connected via a path of length 2 avoiding e to at least $|N_1(v)| - 1 - (|N_1(u)| - 2) = (|N_1(v)| - |N_1(u)|) + 1$ nodes in $N_1(v)$; each of which yields a Seymour diamond of which e is the base, which is the desired result.

Case 2. Suppose there is no such u'. Then again applying part (iii) of Theorem 4.1, it must be that there exists a path of length 1 or 2 avoiding e to each node in $N_1(v)$. But u is adjacent to at most $|N_1(u)| - 1$ of these nodes, and as before we count that there is a path of length 2 avoiding e from u to at least $|N_1(v)| - (|N_1(u)| - 1) = |N_1(v)| - |N_1(u)| + 1$ nodes in $|N_1(v)|$. Since each of these paths yields a Seymour diamond with e as the base, we are done. □

Finally, we show that there is not some finite nonzero number of counterexamples to the conjecture. That is, either the conjecture is true, or there are an infinite number of (non-isomorphic) graphs that violate Conjecture 1.1. We provide a constructive proof below.

Theorem 4.3. *If Seymour's second neighborhood conjecture is false, there are infinitely many non-isomorphic strongly-connected counterexamples to Seymour's second neighborhood conjecture.*

Proof. Suppose that Seymour's second neighborhood conjecture is false, and suppose that digraph D is any strongly-connected counterexample to Seymour's second neighborhood conjecture. (By Theorem 4.1(i) , such a D must exist.) Let H be any digraph satisfying the condition $A_s(v) \geq 0$ for all $v \in V(H)$; that is, all of H's vertices have nonnegative antisatisfaction. Note that any dicycle satisfies the relevant condition, and hence there exists a choice of H on any number n of vertices, $n \geq 3$.

We now construct a graph D' on $|V(D)| \cdot |V(H)|$ vertices such that D' is a counterexample to Seymour's second neighborhood conjecture, thus proving our theorem. We define our graph D' as follows:

(i) $V(D') = V(D) \times V(H)$.

(ii) If $u = (d_1, h_1), v = (d_2, h_2) \in V(D')$, then $(u, v) \in E(D')$ if and only if either
   (a) $d_1 = d_2$ and $(h_1, h_2) \in E(H)$, or
   (b) $d_1 \neq d_2$ and $(d_1, d_2) \in E(D)$.

For any vertex $v = (d, h) \in V(D')$, we calculate that

$$|N_{1,D'}(v)| = |N_{1,H}(h)| + |V(H)| \cdot |N_{1,D}(d)|,$$

since the neighborhood is equivalent to the set of vertices reachable by stepping in H, holding d constant, or stepping in D and allowing h to be arbitrary.

Similarly, we have

$$|N_{2,D'}(v)| = |N_{2,H}(h)| + |V(H)| \cdot |N_{2,D}(d)|,$$

since we may consider walking two steps in H or two steps in D. Note that taking one step in H and one step in D or one step in D and then one in H will result in reaching a vertex that is in $N_{1,D'}(v)$, and hence this is not an overcount.

We then calculate that

$$\begin{aligned} A_{s,D'}(v) &= |N_{1,D'}(v)| - |N_{2,D'}(v)| \\ &= (|N_{1,H}(h)| - |N_{2,H}(h)|) + |V(H)|(|N_{1,D}(d)| - |N_{2,D}(d)|). \end{aligned}$$

But by our choice of H, we have $|N_{1,H}(h)| - |N_{2,H}(h)| \geq 0$, and by our choice of D we have $|N_{1,D}(d)| - |N_{2,D}(d)| > 0$. Hence we obtain $A_{s,D'}(v) > 0$, thus implying that every vertex in D' has positive antisatisfaction.

Furthermore, D' is strongly connected: fix $(d_1, h_1), (d_2, h_2) \in V(D')$. If $d_1 \neq d_2$, let $d_1, \delta_1, \ldots, \delta_i, d_2$ define a directed path in D from d_1 to d_2. Then

$$(d_1, h_1), (\delta_1, h_2), \ldots, (\delta_i, h_2), (d_2, h_2)$$

defines a directed path in D' from (d_1, h_1) to (d_2, h_2). If $d_1 = d_2$, let $d_3 \in N_{1,D}(d_1)$; we know that $(d_1, h_1), (d_3, h_2)$ are adjacent in D', and since $d_2 \neq d_3$ there is a path from (d_3, h_2) to (d_2, h_2) in D', the existence of a path from (d_1, h_1) to (d_2, h_2) follows.

By definition, we then have that D' is a strongly-connected counterexample to Seymour's second neighborhood conjecture. □

5. Conclusions and future directions

In total, this paper has been an exploration of Seymour's second neighborhood conjecture. We have neither proven nor disproved the conjecture, but instead determined some classes of graphs that do satisfy the conjecture; we have also described some properties of a family of minimal counterexamples. Moreover, we have shown that the existence of one counterexample graph implies the existence of infinitely many such graphs. Our work is intended as a stepping stone for further analysis of Conjecture 1.1, which we hope will ultimately lead to its resolution.

Acknowledgment

This work was done at the East Tennessee State University REU, supported by NSF grant 0552730, under the supervision of Dr. Anant Godbole.

References

[Cohn et al. 2009] Z. Cohn, E. Wright, and A. Godbole, "Probabilistic versions of Seymour's distance two conjecture", preprint, 2009, Available at www.etsu.edu/math/godbole/seymour.pdf.

[Dean and Latka 1995] N. Dean and B. J. Latka, "Squaring the tournament—an open problem", pp. 73–80 in *Proceedings of the Twenty-sixth Southeastern International Conference on Combinatorics, Graph Theory and Computing (Boca Raton, FL, 1995)*, vol. 109, 1995. MR 96h:05085 Zbl 0904.05034

[Fidler and Yuster 2007] D. Fidler and R. Yuster, "Remarks on the second neighborhood problem", *J. Graph Theory* **55**:3 (2007), 208–220. MR 2009b:05122 Zbl 1122.05040

[Fisher 1996] D. C. Fisher, "Squaring a tournament: a proof of Dean's conjecture", *J. Graph Theory* **23**:1 (1996), 43–48. MR 97i:05051 Zbl 0857.05042

[Kaneko and Locke 2001] Y. Kaneko and S. C. Locke, "The minimum degree approach for Paul Seymour's distance 2 conjecture", pp. 201–206 in *Proceedings of the Thirty-second Southeastern International Conference on Combinatorics, Graph Theory and Computing (Baton Rouge, LA, 2001)*, vol. 148, 2001. MR 2002k:05078 Zbl 0996.05042

Received: 2008-08-18 Revised: 2009-08-03 Accepted: 2009-08-13

jbrantne@brandeis.edu *Department of Philosophy, Brandeis University, 415 South Street, Waltham, MA 02453, United States*

brockman@hcs.harvard.edu *Department of Mathematics, Harvard University, 1 Oxford Street, Cambridge, MA 02138, United States*

kayw@mailbox.sc.edu *Department of Mathematics, University of South Carolina, Leconte 411, Columbia, SC 29208, United States*

snivelee@rose-hulman.edu *Department of Mathematics, Rose–Hulman Institute of Technology, 5500 Wabash Avenue, Terre Haute, IN 47803, United States*

Yet another generalization of frames and Riesz bases

Reza Joveini and Massoud Amini

(Communicated by David Larson)

A frame is a sequence of vectors in a Hilbert space satisfying certain inequalities that make it valuable for signal processing and other purposes. There is a formula giving the reconstruction of a signal (a vector in the space) from its sequence of inner products (the Fourier coefficients) with the elements of the frame sequence. A g-frame, or operator-valued frame, is a sequence of operators defined on a countable ordered index set that has properties analogous to those of a frame sequence.

We present a new approach to the matter of defining a Hilbert space frame, indexed by an ordered set, when the set is a measure space which is not necessarily purely atomic. Continuous frames have been widely studied in the literature, but the measure spaces they are associated with are not necessarily ordered in any way. Our approach is to make the measure space a directed set, and then replace the sequence of vectors (or operators) with a net indexed by the directed set, obtaining a natural generalization of the usual notion of generalized frame. We show that this definition makes sense mathematically, and proceed to obtain generalizations of several of the standard results for frame and Bessel sequences, and also Riesz bases, g-frames and operator-valued frames.

1. Introduction

Frames are generalizations of bases in a Hilbert space. They were introduced and studied in [Duffin and Schaeffer 1952] and [Daubechies et al. 1986]. They have been recently of special interest because of their applications in signal processing. The interested reader is referred to [Christensen 2003; Daubechies 1992; Feichtinger and Strohmer 1998; Gröchenig 2001; Han and Larson 2000; Heil and Walnut 1989; Yong 1980] for theory and applications of frames. Throughout this paper, U and V are two Hilbert spaces and $\{V_m : m \in M\}$ is a net of subspaces of

MSC2000: primary 42C15, 42C99; secondary 42C40.
Keywords: g-frames, g-Riesz bases.
This is part of the first author's M. Sc. Thesis at Tarbiat Modares University.

V, $\mathcal{B}(U, V_m)$ is the collection of all bounded linear operators from U into V_m and (M, μ) is a measure space.

Definition 1.1. We call a net $\{\Lambda_m \in \mathcal{B}(U, V_m) : m \in M\}$ a generalized frame or simply a g-frame for U with respect to $\{V_m : m \in M\}$ if

(a) for each $f \in U$, there is a measurable function $\tilde{f} : M \to V_m$ such that $\tilde{f}(m) = \Lambda_m f$, and

(b) there are positive constants A and B such that

$$A\|f\|_U^2 \leq \|\tilde{f}\|_{L^2_{(M,V_m)}}^2 \leq B\|f\|_U^2 \quad (f \in U). \tag{1-1}$$

We call A and B the lower and upper frame bounds. We call $\{\Lambda_m : m \in M\}$

(i) a tight g-frame if $A = B$;

(ii) an exact g-frame if it ceases to be a g-frame whenever any of its elements is removed;

(iii) a g-frame for U whenever the net $\{V_m : m \in M\}$ is clear from the context;

(iv) a g-frame for U with respect to V whenever $V_m = V$, for each $m \in M$.

Various generalizations of frames have been proposed [Aldroubi et al. 2004; Asgari and Khosravi 2005; Casazza and Kutyniok 2004; Christensen and Eldar 2004; Feichtinger and Strohmer 1998; Fornasier 2003; Gröchenig 2001]. We take as our starting point the generalization presented in [Sun 2006]. Our definition above is just the definition of g-frames in [Sun 2006] when the measure space M is countable, μ is the counting measure and $V_m = \mathbb{C}$ for $m \in M$. The case when M is not countable could be of interest when one deals with nonseparable Hilbert spaces (such as Hilbert space completions of the space of almost periodic functions). Thus our present work can be regarded as a nonseparable version of [Sun 2006]. For instance Examples 3.4 and 3.5 in that reference are outside the scope of Sun's definition when the Hilbert space has no countable Riesz basis, but they fit in our framework.

2. g-frame operators and dual g-frames

Let $\{\Lambda_m : m \in M\}$ be a g-frame for U with respect to $\{V_m : m \in M\}$. Define the g-frame operator S by

$$Sf = \int_M \Lambda_m^* \Lambda_m f \, d\mu(m), \tag{2-1}$$

where Λ^* is the adjoint operator of Λ.

Lemma 2.1. *Let (Ω, μ) be a measure space, X and Y are two Banach spaces, $\lambda : X \to Y$ be a bounded linear operator and $f : \Omega \to X$ be a measurable function. Then*

$$\lambda\Big(\int_\Omega f\, d\mu\Big) = \int_\Omega (\lambda f)\, d\mu.$$

Next we want to show S is self-adjoint operator. For any $f_1, f_2 \in U$ we have

$$\begin{aligned}\langle Sf_1, f_2\rangle &= \Big\langle \int_M \Lambda_m^*\Lambda_m f_1\, d\mu(m), f_2\Big\rangle = \int_M \langle \Lambda_m^*\Lambda_m f_1, f_2\rangle\, d\mu(m)\\ &= \int_M \langle f_1, \Lambda_m^*\Lambda_m f_2\rangle\, d\mu(m) = \Big\langle f_1, \int_M \Lambda_m^*\Lambda_m f_2\, d\mu(m)\Big\rangle = \langle f_1, Sf_2\rangle.\end{aligned}$$

Moreover, S is a bounded operator because

$$\begin{aligned}\|S\| &= \sup_{\|f\|=1}\|Sf\| = \sup_{\|f\|=1}\Big\|\int_M \Lambda_m^*\Lambda_m f\, d\mu(m)\Big\| \le \sup_{\|f\|=1}\Big(\int_M \|\Lambda_m^*\Lambda_m f\|\, d\mu(m)\Big)\\ &= \sup_{\|f\|=1}\int_M \big(\langle \Lambda_m f, \Lambda_m f\rangle\big)^{1/2}\, d\mu(m) = \sup_{\|f\|=1}\int_M \|\tilde{f}(m)\|\, d\mu(m) \le B.\end{aligned}$$

Since $A\|f\|^2 \le \langle Sf, f\rangle \le \|Sf\|\,\|f\|$, we have

$$\|Sf\| \ge A\|f\|,$$

which implies that S is injective and SU is closed in U. Let $f_2 \in U$ be such that $\langle f_1, Sf_2\rangle = 0$, for each $f_1 \in U$. This implies that $Sf_2 = 0$ and therefore $f_2 = 0$. Hence $SU = U$. Consequently S is invertible and $\|S^{-1}\| \le \frac{1}{A}$. For any $f \in U$ we have

$$f = SS^{-1}f = S^{-1}Sf = \int_M \Lambda_m^*\Lambda_m S^{-1} f\, d\mu(m) = \int_M S^{-1}\Lambda_m^*\Lambda_m f\, d\mu(m).$$

Let $\tilde{\Lambda}_m = \Lambda_m S^{-1}$. Then the above equalities become

$$f = \int_M \Lambda_m^*\tilde{\Lambda}_m f\, d\mu(m) = \int_M \tilde{\Lambda}_m^*\Lambda_m f\, d\mu(m). \tag{2-2}$$

We now prove that $\{\tilde{\Lambda}_m : m \in M\}$ is also a g-frame for U with respect to the set $\{V_m : m \in M\}$; in fact for any $f \in U$ we have

$$\begin{aligned}\int_M \|\tilde{\Lambda}_m f\|^2\, d\mu &= \int_M \|\Lambda_m S^{-1} f\|^2\, d\mu = \int_m \langle \Lambda_m S^{-1} f, \Lambda_m S^{-1} f\rangle\, d\mu\\ &= \int_M \langle \Lambda_m^*\Lambda_m S^{-1} f, S^{-1} f\rangle\, d\mu\\ &= \langle SS^{-1} f, S^{-1} f\rangle = \langle f, S^{-1} f\rangle \le \frac{1}{A}\|f\|^2.\end{aligned}$$

On the other hand, since

$$\|f\|^2 = \int_M \langle \tilde{\Lambda}_m^* \Lambda_m f, f\rangle \, d\mu = \int_M \langle \Lambda_m f, \tilde{\Lambda}_m f\rangle \, d\mu$$
$$\leq \Big(\int_M \|\Lambda_m f\|^2 \, d\mu\Big)^{1/2} \Big(\int_M \|\tilde{\Lambda}_m f\|^2 d\,\mu\Big)^{1/2} \leq B^{1/2}\|f\| \Big(\int_M \|\tilde{\Lambda}_m f\|^2 \, d\mu\Big)^{1/2},$$

we have

$$\int_M \|\tilde{\Lambda}_m f\|^2 d\mu \geq \frac{1}{B}\|f\|^2.$$

Hence, $\{\tilde{\Lambda}_m : m \in M\}$ is a g-frame for U with frame bounds A^{-1} and B^{-1}. We call it the (canonical) dual g-frame of $\{V_m : m \in M\}$.

Let $\tilde{S}$ be the g-frame operator associated with $\{\tilde{\Lambda}_m : m \in M\}$. Then, for $f \in U$,

$$S\tilde{S}f = \int_M S\tilde{\Lambda}_m^* \tilde{\Lambda}_m f \, d\mu = \int_M SS^{-1}\Lambda_m^* \Lambda_m S^{-1} f \, d\mu$$
$$= \int_M \Lambda_m^* \Lambda_m S^{-1} f \, d\mu = SS^{-1} f = f.$$

Hence $\tilde{S} = S^{-1}$ and $\tilde{\Lambda}_m \tilde{S}^{-1} = \Lambda_m S^{-1} S = \Lambda_m$. In other words, $\{\Lambda_m : m \in M\}$ and $\{\tilde{\Lambda}_m : m \in M\}$ are dual g-frames with respect to each other.

Remark 2.2. We can always get a tight g-frame from any g-frame $\{\Lambda_m : m \in M\}$. In fact if we put $\{Q_m = \Lambda_m S^{1/2} : m \in M\}$, it is easy to check that $\{Q_m : m \in M\}$ is a tight g-frame with the frame bound 1.

Lemma 2.3. *Let $\{\Lambda_m : m \in M\}$ be a g-frame for U with respect to $\{V_m : m \in M\}$ and $\tilde{\Lambda}_m = \Lambda_m S^{-1}$. For any $g_m \in V_m$ satisfying $f = \int_M \Lambda_m^* g_m d\mu$, we have*

$$\int_M \|g_m\|^2 d\mu = \int_M \|\tilde{\Lambda}_m f\|^2 d\mu + \int_M \|g_m - \tilde{\Lambda}_m f\|^2 d\mu.$$

Proof. It is easy to check that for every $f \in U$,

$$\int_M \|\tilde{\Lambda}_m f\|^2 \, d\mu = \int_M \langle \tilde{\Lambda}_m f, \Lambda_m S^{-1} f\rangle \, d\mu = \int_M \langle \Lambda_m^* \tilde{\Lambda}_m f, S^{-1} f\rangle \, d\mu$$
$$= \int_M \langle \Lambda_m^* g_m, S^{-1} f\rangle \, d\mu = \int_M \langle g_m, \Lambda_m S^{-1} f\rangle \, d\mu$$
$$= \int_M \langle g_m, \tilde{\Lambda}_m f\rangle \, d\mu,$$

and the conclusion follows. □

3. Generalized Bessel nets, Riesz bases and orthonormal bases

Similar to generalized frames, we can define generalized Bessel nets, Riesz bases, and orthonormal bases.

Definition 3.1. Let $\Lambda_m \in \mathcal{B}(U, V_m)$, for $m \in M$.

(i) If the right hand inequality of (1-1) holds, then we say that $\{\Lambda_m : m \in M\}$ is a g-Bessel net for U with respect to $\{V_m : m \in M\}$.

(ii) If $\{f : \Lambda_m f = 0\ (m \in M)\} = \{0\}$ then we say that $\{\Lambda_m : m \in M\}$ is g-complete.

(iii) If $\{\Lambda_m : m \in M\}$ is g-complete and there are two positive constant A and B such that for any measurable subset $M_1 \subset M$ of finite measure, and $g_m \in V_m, m \in M_1$,

$$A \int_{M_1} \|g_m\|^2 \, d\mu \leq \left\| \int_{M_1} \Lambda_m^* \, d\mu \right\|^2 \leq B \int_{M_1} \|g_m\|^2 \, d\mu,$$

then we say $\{\Lambda_m : m \in M\}$ is a g-Riesz bases for U with respect to the set $\{V_m : m \in M\}$.

(iv) We say $\{\Lambda_m : m \in M\}$ is a g-orthonormal basis for U with respect to the set $\{V_m : m \in M\}$ if it satisfies the following equalities:

$$\langle \Lambda_{m_1}^* g_{m_1}, \Lambda_{m_2}^* g_{m_2} \rangle = \delta_{m_1 m_2} \langle g_{m_1}, g_{m_2} \rangle \ (m_1, m_2 \in M, g_{m_1} \in V_{m_1}, g_{m_2} \in V_{m_2}) \quad (3\text{-}1)$$

$$\int_M \|\Lambda_m f\|^2 \, d\mu(m) = \|f^2\|, \quad (f \in U). \quad (3\text{-}2)$$

Characterization of g-frames, g-Riesz bases and g-orthonormal bases. Consider $\Lambda_m \in \mathcal{B}(U, V_m)$; we do not have other assumptions on Λ_m at the moment. Suppose that $\{e_{m,n}, n \in N_m\}$ is an orthonormal basis for V_m, where N_m is an index set of arbitrary cardinality. Then

$$f \mapsto \langle \Lambda_m f, e_{m,n} \rangle$$

defines a bounded linear functional on U, so we can find $u_{m,n} \in U$ such that

$$\langle f, u_{m,n} \rangle = \langle \Lambda_m f, e_{m,n} \rangle; \quad (3\text{-}3)$$

hence

$$\Lambda_m f = \sum_{n \in N_m} \langle f, u_{m,n} \rangle e_{m,n}. \quad (3\text{-}4)$$

Since $\sum_{n \in N_m} |\langle f, u_{m,n} \rangle|^2 = \|\Lambda_m f\|^2 \leq \|\Lambda_m\|^2 \|f\|^2$, the family $\{u_{m,n} : n \in N_m\}$ is a Bessel net for U, and it follows that for any $f \in U$ and $g \in V_m$,

$$\langle f, \Lambda_m^* g \rangle = \langle \Lambda_m f, g \rangle = \sum_{n \in N_m} \langle f, u_{m,n} \rangle \langle e_{m,n}, g \rangle = \left\langle f, \sum_{n \in N_m} \langle g, e_{m,n} \rangle u_{m,n} \right\rangle.$$

Hence

$$\Lambda_m^* g = \sum_{n \in N_m} \langle g, e_{m,n} \rangle u_{m,n} \quad (g \in V_m). \quad (3\text{-}5)$$

In particular,

$$u_{m,n} = \Lambda_m^* e_{m,n} \quad (m \in M, n \in N_m). \quad (3\text{-}6)$$

We call $\{u_{m,n} : m \in M, n \in N_m\}$ the net induced by $\{\Lambda_m : m \in M\}$ with respect to $\{e_{m,n} : n \in N_m, m \in M\}$. With these representations for Λ_m^* and Λ_m, we get characterizations of generalized frames, Riesz bases and orthonormal bases.

Theorem 3.2. *Let $\Lambda_m \in \mathcal{B}(U, V_m)$ and $u_{m,n}$ be defined as in (3-3).*

(i) *$\{\Lambda_m : m \in M\}$ is a g-frame (alternatively a g-Bessel net, tight g-frame, g-Riesz basis, or g-orthonormal basis) for U if and only if $\{u_{m,n} : m \in M, n \in N_m\}$ is a frame (Bessel net, tight frame, Riesz basis, orthonormal basis) for U.*

(ii) *If $\{\Lambda_m : m \in M\}$ is a g-frame, then*

$$\sum_{m \in M} \dim V_m \geq \dim U$$

and the equality holds whenever $\{\Lambda_m : m \in M\}$ is a g-Riesz basis.

(iii) *The g-frame operator for $\{\Lambda_m : m \in M\}$ coincides with the frame operator for $\{u_{m,n} : m \in M, n \in N_m\}$.*

(iv) *$\{\Lambda_m : m \in M\}$ and $\{\tilde{\Lambda}_m : m \in M\}$ are a pair of (canonical) dual g-frames if and only if the induced net are a pair of (canonical) dual frames.*

Proof. (i) We see from (3-4) that

$$\int_M \|\Lambda_m f\|^2 \, d\mu(m) = \int_M \sum_{n \in N_m} |\langle f, u_{m,n} \rangle|^2 \, d\mu(m), \quad (f \in U).$$

Hence $\{\Lambda_m : m \in M\}$ is a g-frame (respectively g-Bessel net, tight-frame) for U if and only if $\{u_{m,n} : m \in M, n \in N_m\}$ is a frame (respectively Bessel net, tight frame) for U.

Next assume that $\{\Lambda_m : m \in M\}$ is a g-Riesz bases for U. Since $\{e_{m,n} : n \in N_m\}$ is an orthonormal basis for V_m, every $g_m \in V_m$ has an expansion of the form $g_m = \sum_{n \in N_m} c_{m,n} e_{m,n}$, where $\{c_{m,n} : n \in N_m\} \in l^2(N_m)$. It follows that

$$A \int_{M_1} \|g_m\|^2 \, d\mu \leq \| \int_{M_1} \Lambda_m^* g_m \, d\mu \|^2 \leq B \int_{M_1} \|g_m\|^2 \, d\mu$$

is equivalent to

$$A \int_{M_1} \Big(\sum_{n \in N_m} |c_{m,n}|^2 \Big) \, d\mu \leq \Big\| \int_{M_1} \Big(\sum_{n \in N_m} c_{m,n} u_{m,n} \Big) \, d\mu \Big\|^2 \leq B \int_{M_1} \Big(\sum_{n \in N_m} |c_{m,n}|^2 \Big) \, d\mu.$$

On the other hand, we see from $\Lambda_m f = \sum_{n \in N_m} \langle f, u_{m,n} \rangle e_{m,n}$ that

$$\big\{ f : \Lambda_m f = 0 \, (m \in M) \big\} = \big\{ f : \langle f, u_{m,n} \rangle = 0 \, (m \in M, n \in N_m) \big\}.$$

Hence $\{\Lambda_m : m \in M\}$ is g-complete if and only if $\{u_{m,n} : m \in M, n \in N_m\}$ is complete. Therefor $\{\Lambda_m : m \in M\}$ is a g-Riesz basis if and only if $\{u_{m,n} : m \in M, n \in N_m\}$ is a Riesz basis.

Now assume that $\{\Lambda_m : m \in M\}$ is a g-orthonormal basis. It follows from (3-1) and (3-3) that

$$\begin{aligned}\langle u_{m_1,n_1}, u_{m_2,n_2}\rangle &= \langle \Lambda_{m_2} u_{m_1,n_1}, e_{m_2,n_2}\rangle = \overline{\langle \Lambda_{m_2}^* e_{m_2,n_2}, u_{m_1,n_1}\rangle} \\ &= \overline{\langle \Lambda_{m_1}\Lambda_{m_2}^* e_{m_2,n_2}, e_{m_1,n_1}\rangle} = \langle \Lambda_{m_1}^* e_{m_1,n_1}, \Lambda_{m_2}^* e_{m_2,n_2}\rangle = \delta_{m_1,m_2}\delta_{n_1,n_2}\end{aligned}$$

for $m_1, m_2 \in M$, $n_1 \in N_{m_1}$, $n_2 \in N_{m_2}$. Hence $\{u_{m,n}, m \in M, n \in N_m\}$ is an orthonormal net. Moreover, observe that

$$\|f\|^2 = \int_M \|\Lambda_m f\|^2 \, d\mu = \int_M \Big(\sum_{n \in N_m} |\langle f, u_{m,n}\rangle| \Big) \, d\mu \quad (f \in U).$$

Therefore $\{u_{m,n} : m \in M, n \in N_m\}$ is an orthonormal basis.

For the converse, we need only to show that (3-2) holds. In fact, we see from (3-5) that for any $m_1 \neq m_2 \in M$, $g_{m_1} \in V_{m_1}$, $g_{m_2} \in V_{m_2}$,

$$\langle \Lambda_{m_1}^* g_{m_1}, \Lambda_{m_2}^* g_{m_2}\rangle = \Big\langle \sum_{n_1 \in N_{m_1}} \langle g_{m_1}, e_{m_1,n_1}\rangle u_{m_1,n_1}, \sum_{n_2 \in N_{m_2}} \langle g_{m_2}, e_{m_2,n_2}\rangle u_{m_2,n_2} \Big\rangle = 0,$$

and for $m \in M$, $g_1, g_2 \in V_m$,

$$\langle \Lambda_m^* g_1, \Lambda_m^* g_2\rangle = \Big\langle \sum_{n_1 \in N_m} \langle g_1, e_{m,n_1}\rangle u_{m,n_1}, \sum_{n_2 \in N_m} \langle g_2, e_{m,n_2}\rangle u_{m,n_2} \Big\rangle = \langle g_1, g_2\rangle.$$

Now the conclusion follows.

(ii) Since the cardinality of a frame is no less than that of the basis, we have $\#\{u_{m,n} : m \in M, n \in N_m\} \geq \dim U$. Moreover, we see from (i) that the equality holds whenever $\{\Lambda_m : m \in M\}$ is a g-Riesz basis.

(iii) We see from (3-4) and (3-5) that, for $f \in U$,

$$\begin{aligned}\int_M \Lambda_m^* \Lambda_m f \, d\mu &= \int_M \Big(\sum_{n \in N_m} \langle \Lambda_m f, e_{m,n}\rangle u_{m,n} \Big) \, d\mu \\ &= \int_M \Big(\sum_{n \in N_m} \Big\langle \sum_{n' \in N_m} \langle f, u_{m,n'}\rangle e_{m,n'}, e_{m,n} \Big\rangle u_{m,n} \Big) \, d\mu \\ &= \int_M \Big(\sum_{n \in N_m} \langle f, u_{m,n}\rangle u_{m,n} \Big) \, d\mu\end{aligned}$$

Hence the g-frame operator for $\{\Lambda_m : m \in M\}$ coincides with the frame operator for $\{u_{m,n} : m \in M, n \in N_m\}$.

(iv) This is the content of (i) and (iii). □

Corollary 3.3. *$\{\Lambda_m : m \in M\}$ is a g-Bessel net with an upper bound B if and only if for any measurable subset $M_1 \subset M$ of finite measure,*

$$\left\| \int_{M_1} \Lambda_m^* \, d\mu \right\|^2 \leq B \int_{M_1} \|g_m\|^2 \, d\mu.$$

Corollary 3.4. *A g-Riesz basis $\{\Lambda_m : m \in M\}$ is an exact g-frame. Moreover, it is g-biorthonormal with respect to its dual $\{\tilde{\Lambda}_m : m \in M\}$ in the sense that*

$$\langle \Lambda_{m_1}^* g_{m_1}, \Lambda_{m_2}^* g_{m_2} \rangle = \delta_{m_1,m_2} \langle g_{m_1}, g_{m_2} \rangle \quad (m_1, m_2 \in M,\ g_{m_1} \in V_{m_1},\ g_{m_2} \in V_{m_2}).$$

Corollary 3.5. *A net $\{\Lambda_m : m \in M\}$ is a g-Riesz basis for U with respect to the set $\{V_m : m \in M\}$ if and only if there is a g-orthonormal basis $\{Q_m : m \in M\}$ for U and a bounded invertible linear operator T on U such that $\Lambda_m = Q_m T$, $m \in M$.*

Proof. Let $\{e_{m,n} : n \in N_m\}$ be an orthonormal basis for V_m, $m \in M$. First, we assume that $\{\Lambda_m : m \in M\}$ is a g-Riesz basis for U. By Theorem 3.2, we can find some Riesz basis $\{u_{m,n} : m \in M, n \in N_m\}$ for U such that

$$\Lambda_m f = \sum_{n \in N_m} \langle f, u_{m,n} \rangle e_{m,n}.$$

Take an orthonormal basis $\{u_{m,n}^0\}$ for U and define the operator T on U by

$$T^* u_{m,n}^0 = u_{m,n}.$$

Obviously, T is a bounded invertible operator. Let $Q_m \in \mathcal{B}(U, V_m)$ be such that $Q_m f = \sum_{n \in N_m} \langle f, u_{m,n}^0 \rangle e_{m,n}$. Again by Theorem 3.2, $\{Q_m : m \in M\}$ is a g-orthonormal basis for any $f \in U$, and

$$\begin{aligned} Q_m T f &= \sum_{n \in N_m} \langle Tf, u_{m,n}^0 \rangle e_{m,n} = \sum_{n \in N_m} \langle f, T^* u_{m,n}^0 \rangle e_{m,n} \\ &= \sum_{n \in N_m} \langle Tf, u_{m,n} \rangle e_{m,n} = \Lambda_m f. \end{aligned}$$

Hence $\Lambda_m = Q_m T$, for each $m \in M$.

Next we assume that $\{Q_m : m \in M\}$ is a g-orthonormal basis and $\Lambda_m = Q_m T$ for some bounded invertible operator T. Then $\{\Lambda_m : m \in M\}$ is g-complete in U and we can find orthonormal basis $\{u_{m,n}^0 : m \in M, n \in N_m\}$ for U such that $Q_m f = \sum_{n \in N_m} \langle f, u_{m,n}^0 \rangle e_{m,n}$. Hence

$$\Lambda_m f = \sum_{n \in N_m} \langle Tf, u_{m,n}^0 \rangle e_{m,n} = \sum_{n \in N_m} \langle f, T^* u_{m,n}^0 \rangle e_{m,n},$$

and we see from Theorem 3.2 that $\{\Lambda_m : m \in M\}$ is a g-Riesz basis. □

Excess of g-frames. By Theorem 3.2, g-frames, g-Riesz basis and g-orthonormal bases have properties similar to those of frame, Riesz bases and orthonormal bases, respectively. However, not all the properties are similar. For example, Riesz bases are equivalent to exact frames, but this is not the case for g-Riesz bases and exact g-frames.

In fact, we see from Theorem 3.2 that a g-Riesz basis is also a g-frame, while the converse is not true. This is not surprising, since one element of a g-frame might correspond to several elements of the induced frame.

A natural problem arises: given a measurable subset $M_1 \subset M$ with $\mu(M_1) > 0$, when is $\{\Lambda_m : m \in M - M_1\}$ a g-frame?

Theorem 3.6. *Let $\{\Lambda_m : m \in M\}$ be a g-frame for U with respect to $\{V_m : m \in M\}$ and $\{\tilde{\Lambda}_m : m \in M\}$ be the canonical dual g-frame. Suppose that $M_1 \subset M$ and $\mu(M_1) > 0$.*

(i) *If there is some $g_0 \in V_{m'} - \{0\}$ such that $\tilde{\Lambda}_{m'}\Lambda^*_{m'}g_0 = g_0$, for any $m' \in M_1$, then $\{\Lambda_m : m \in M - M_1\}$ is not g-complete in U.*

(ii) *If there is some $f_0 \in U - \{0\}$ such that $\Lambda^*_{m'}\tilde{\Lambda}_{m'}f_0 = f_0$, for any $m' \in M_1$, then $\{\Lambda_m : m \in M - M_1\}$ is not g-complete in U.*

(iii) *If $I - \Lambda^*_{m'}\tilde{\Lambda}^*_{m'}$ or $I - \tilde{\Lambda}_{m'}\Lambda^*_{m'}$ is bounded invertible on $V_{m'}$, for any $m' \in M_1$, then $\{\Lambda_m : m \in M - M\}$ is a g-frame for U.*

Proof. (i) Since $\Lambda^*_{m'}g_0 \in U$, we have from (2-2)

$$\Lambda^*_{m'}g_0 = \int_M \Lambda^*_m \tilde{\Lambda}_m \Lambda^*_{m'}g_0 \, d\mu.$$

Put $v_{m',m} = \delta_{m',m}g_0$. We have

$$\Lambda^*_{m'}g_0 = \int_M \Lambda^*_m v_{m',m} \, d\mu.$$

It follows from Lemma 2.3 that

$$\int_M \|v_{m_0,m}\|^2 \, d\mu = \int_M \|\tilde{\Lambda}_m \Lambda^*_{m_0}g_0\|^2 \, d\mu + \int_M \|\tilde{\Lambda}_m \Lambda^*_{m_0}g_0 - v_{m_0,m}\|^2,$$

and so

$$\|g_0\|^2 = \|g_0\|^2 + 2\int_{M-M_1} \|\tilde{\Lambda}_m \Lambda^*_{m'}g_0\|^2 \, d\mu.$$

Hence $\tilde{\Lambda}_m \Lambda^*_{m'}g_0 = 0$, for each $m' \in M_1$ and almost all $m \in M - M_1$. This means that $\Lambda_m \tilde{\Lambda}^*_{m'}g_0 = \Lambda_m S^{-1}\Lambda^*_{m'}g_0 = \tilde{\Lambda}_m \Lambda^*_{m'}g_0 = 0$, for almost all $m \neq m'$. But $\langle \Lambda^*_{m'}g_0, \tilde{\Lambda}^*_{m'}g_0\rangle = \langle \tilde{\Lambda}_{m'}\Lambda^*_{m'}g_0, g_0\rangle = \|g_0\|^2 > 0$, which implies that $\tilde{\Lambda}^*_{m'}g_0 \neq 0$. Hence $\{\Lambda_m : m \in M - M_1\}$ is not g-complete in U.

(ii) Since $\Lambda_{m'}^*\tilde{\Lambda}_{m'} f_0 = f_0 \neq 0$, we have $\tilde{\Lambda}_{m'} f_0 \neq 0$ and $\tilde{\Lambda}_{m'}\Lambda_{m'}^* f_0 = \tilde{\Lambda}_{m'} f_0$. Now the conclusion follows from (i).

(iii) Since $\tilde{\Lambda}_m = \Lambda_m S^{-1}$ where S is the g-frame operator for $\{\Lambda_m : m \in M\}$, we have

$$I - \Lambda_{m'}\tilde{\Lambda}_{m'} = I - \Lambda_{m'} S^{-1}\tilde{\Lambda}_{m'}^* = I - \tilde{\Lambda}_{m'}\Lambda_{m'}^*.$$

Let A and B be the lower and upper frame bounds for $\{\Lambda_m : m \in M\}$. For any $f \in U$, we have

$$f = \int_M \tilde{\Lambda}_m^* \Lambda_m f \, d\mu.$$

Hence, for each $m' \in M_1$,

$$\Lambda_{m'} f = \int_M \Lambda_{m'}\tilde{\Lambda}_m^* \Lambda_m f \, d\mu.$$

Therefore

$$(I - \Lambda_{m'}\tilde{\Lambda}_{m'}^*)\Lambda_{m'} f = \int_{M-M_1} \Lambda_{m'}\tilde{\Lambda}_m^* \Lambda_m f \, d\mu. \tag{3-7}$$

Note that

$$\begin{aligned}
\left\| \int_{M-M_1} \Lambda_{m'}\tilde{\Lambda}_m^* \Lambda_m f \, d\mu \right\|^2 &= \sup_{g \in V_{m',\|g\|=1}} \left| \left\langle \int_{M-M_1} \Lambda_{m'}\tilde{\Lambda}_m^* \Lambda_m f \, d\mu, g \right\rangle \right|^2 \\
&= \sup_{\|g\|=1} \left| \int_{M-M_1} \langle \Lambda_m f, \tilde{\Lambda}_m \Lambda_{m'}^* g \rangle \, d\mu \right|^2 \\
&\leq \int_{M-M_1} \|\Lambda_m f\|^2 \, d\mu \sup_{\|g\|=1} \int_M \|\tilde{\Lambda}_m \Lambda_{m'}^* g\|^2 \, d\mu \\
&\leq \frac{1}{A}\|\Lambda_{m'}\|^2 \int_{M-M_1} \|\Lambda_m f\|^2 \, d\mu.
\end{aligned}$$

We see from (3-7) that

$$\|\Lambda_{m'} f\|^2 \leq \|(I - \Lambda_{m'}\tilde{\Lambda}_{m'}^*)^{-1}\| \, \frac{1}{A}\|\Lambda_{m'}\|^2 \int_{M-M_1} \|\Lambda_m f\|^2 \, d\mu \quad (m' \in M_1).$$

Hence

$$\int_M \|\Lambda_m f\|^2 \, d\mu \leq C \int_{M-M_1} \|\Lambda_m f\|^2 \, d\mu.$$

Therefore, for $f \in U$,

$$\frac{A}{C}\|f\|^2 \leq \int_{M-M_1} \|\Lambda_m f\|^2 \, d\mu \leq B\|f\|^2.$$

This completes the proof. □

4. Applications of g-frames

Atomic resolution of bounded linear operators. Here we give an application of g-frames. Let $\{\Lambda_m : m \in M\}$ be a g-frame for U with respect to $\{V_m : m \in M\}$. Suppose that $\{\tilde{\Lambda}_m : m \in M\}$ is the canonical dual g-frame. Then for any $f \in U$, we have $f = \int_M \Lambda_m^* \tilde{\Lambda}_m f \, d\mu = \int_M \tilde{\Lambda}_m^* \Lambda_m f \, d\mu$. It follows that

$$I_U = \int_M \Lambda_m^* \tilde{\Lambda}_m \, d\mu = \int_M \tilde{\Lambda}_m^* \Lambda_m \, d\mu. \tag{4-1}$$

Let T be a bounded linear operator on U. We see from (4-1) that

$$T = \int_M T \Lambda_m^* \tilde{\Lambda}_m \, d\mu = \int_M T \tilde{\Lambda}_m^* \Lambda_m \, d\mu = \int_M \Lambda_m^* \tilde{\Lambda}_m T \, d\mu = \int_M \tilde{\Lambda}_m^* \Lambda_m T \, d\mu. \tag{4-2}$$

Construction of frames via g-frames. Let $\{\Lambda_m : m \in M\}$ be a g-frame for U with respect to $\{V_m : m \in M\}$. We see from Theorem 3.2 that

$$\{u_{m,n} : m \in M, n \in N_m\} = \{\Lambda_m^* e_{m,n} : m \in M, n \in N_m\}$$

is a frame for U, where $\{e_{m,n} : n \in N_m\}$ is an orthonormal basis for V_m. However, it might be difficult to find an orthonormal basis for V_m in practice. Fortunately, orthonormality is not necessary to get a frame. In fact:

Theorem 4.1. *Let $\{\Lambda_m : m \in M\}$ and $\{\tilde{\Lambda}_m : m \in M\}$ be a pair of dual g-frames for U with respect to $\{V_m : m \in M\}$, and $\{g_{m,n} : n \in N_m\}$ and $\{\tilde{g}_{m,n} : n \in N_m\}$ be the corresponding pair of dual frames for V_m, respectively. Then*

$$\{\Lambda_m^* g_{m,n} : m \in M, n \in N_m\} \quad \text{and} \quad \{\tilde{\Lambda}_m \tilde{g}_{m,n} : m \in M, n \in N_m\}$$

are a pair of dual frames for U, provided that the frame bounds for $\{g_{m,n} : n \in N_m\}$ satisfy $C_1 \le A_m \le B_m \le C_2$, for some constants $C_1, C_2 > 0$.

Moreover, suppose that $\{\Lambda_m : m \in M\}$ and $\{\tilde{\Lambda}_m : m \in M\}$ are canonical dual g-frames for U, $\{g_{m,n} : n \in N_m\}$ and $\{\tilde{g}_{m,n} : n \in N_m\}$ are canonical dual frames for V_m, and $\{g_{m,n} : n \in N_m\}$ is a tight g-frame with frame bounds $A_m = B_m = A$, $m \in M$. Then $\{\Lambda_m^* g_{m,n} : m \in M, n \in N_m\}$ and $\{\tilde{\Lambda}_m^* \tilde{g}_{m,n} : m \in M, n \in N_m\}$ are canonical dual frames for U.

Proof. Note that $\langle f, \Lambda_m^* g_{m,n} \rangle = \langle \Lambda_m f, g_{m,n} \rangle$. It is easy to see that $\{\Lambda_m^* g_{m,n} : m \in M, \ n \in N_m\}$ and $\{\tilde{\Lambda}_m^* \tilde{g}_{m,n} : m \in M, n \in N_m\}$ are frames for U. On the other hand, for any $f \in U$, we have

$$\begin{aligned} \int_M \Big(\sum_{n \in N_m} \langle f, \Lambda_m^* g_{m,n} \rangle \tilde{\Lambda}_m^* \tilde{g}_{m,n} \Big) \, d\mu &= \int_M \tilde{\Lambda}_m^* \sum_{n \in N_m} \langle \Lambda_m f, g_{m,n} \rangle \tilde{g}_{m,n} \, d\mu \\ &= \int_M \tilde{\Lambda}_m^* \Lambda_m f = f. \end{aligned}$$

Similarly we can get

$$\int_M \Big(\sum_{n\in N_m} \langle f, \tilde{\Lambda}_m^* \tilde{g}_{m,n} \rangle \Lambda_m^* g_{m,n} \Big)\, d\mu = f.$$

Hence $\{\Lambda_m^* g_{m,n} : m \in M, n \in N_m\}$ and $\{\tilde{\Lambda}_m^* \tilde{g}_{m,n} : m \in M, n \in N_m\}$ are dual frames for U.

Next we assume that $\{\Lambda_m : m \in M\}$ and $\{\tilde{\Lambda}_m : m \in M\}$ are canonical dual g-frames and $\{g_{m,n} : n \in N_m\}$ is a tight frame with frames bounds $A_m = B_m = A$. Then $\tilde{g}_{m,n} = A^{-1} g_{m,n}$. Let S_Λ and $S_{\Lambda,g}$ be the frame operators associated with $\{\Lambda_m : m \in M\}$ and $\{\Lambda_m^* g_{m,n} : m \in M, n \in N_m\}$, respectively. Then, for $f \in U$,

$$\begin{aligned} S_{\Lambda,g} f &= \int_M \Big(\sum_{n\in N_m} \langle f, \Lambda_m^* g_{m,n} \rangle \Lambda_m^* g_{m,n} \Big)\, d\mu \\ &= \int_M \Big(\Lambda_m^* \sum_{n\in N_m} \langle \Lambda_m f, g_{m,n} \rangle g_{m,n} \Big)\, d\mu = A \int_M \Lambda_m^* \Lambda_m f \, d\mu = A S_\Lambda f. \end{aligned}$$

Hence

$$S_{\Lambda,g}^{-1} \Lambda_m^* g_{m,n} = \frac{1}{A} S_\Lambda^{-1} \Lambda_m^* g_{m,n} = \Lambda_m^* \tilde{g}_{m,n} \quad (m \in M, n \in N_m).$$

This completes the proof. □

References

[Aldroubi et al. 2004] A. Aldroubi, C. Cabrelli, and U. M. Molter, "Wavelets on irregular grids with arbitrary dilation matrices and frame atoms for $L^2(\mathbb{R}^d)$", *Appl. Comput. Harmon. Anal.* **17**:2 (2004), 119–140. MR 2006f:42035 Zbl 1060.42025

[Asgari and Khosravi 2005] M. S. Asgari and A. Khosravi, "Frames and bases of subspaces in Hilbert spaces", *J. Math. Anal. Appl.* **308**:2 (2005), 541–553. MR 2006b:42042 Zbl 1091.46006

[Casazza and Kutyniok 2004] P. G. Casazza and G. Kutyniok, "Frames of subspaces", pp. 87–113 in *Wavelets, frames and operator theory*, edited by C. Heil et al., Contemp. Math. **345**, Amer. Math. Soc., Providence, RI, 2004. MR 2005e:42090 Zbl 1058.42019

[Christensen 2003] O. Christensen, *An introduction to frames and Riesz bases*, Birkhäuser, Boston, 2003. MR 2003k:42001 Zbl 1017.42022

[Christensen and Eldar 2004] O. Christensen and Y. C. Eldar, "Oblique dual frames and shift-invariant spaces", *Appl. Comput. Harmon. Anal.* **17**:1 (2004), 48–68. MR 2005f:42065 Zbl 1043.42027

[Daubechies 1992] I. Daubechies, *Ten lectures on wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics **61**, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992. MR 93e:42045 Zbl 0776.42018

[Daubechies et al. 1986] I. Daubechies, A. Grossmann, and Y. Meyer, "Painless nonorthogonal expansions", *J. Math. Phys.* **27**:5 (1986), 1271–1283. MR 87e:81089 Zbl 0608.46014

[Duffin and Schaeffer 1952] R. J. Duffin and A. C. Schaeffer, "A class of nonharmonic Fourier series", *Trans. Amer. Math. Soc.* **72** (1952), 341–366. MR 13,839a Zbl 0049.32401

[Feichtinger and Strohmer 1998] H. G. Feichtinger and T. Strohmer (editors), *Gabor analysis and algorithms: Theory and applications*, Birkhäuser, Boston, 1998. MR 98h:42001

[Fornasier 2003] M. Fornasier, "Decompositions of Hilbert spaces: local construction of global frames", pp. 275–281 in *Constructive theory of functions*, edited by B. Bojanov, DARBA, Sofia, 2003. MR 2005f:42085 Zbl 1031.42035

[Gröchenig 2001] K. Gröchenig, *Foundations of time-frequency analysis*, Birkhäuser, Boston, 2001. MR 2002h:42001 Zbl 0966.42020

[Han and Larson 2000] D. Han and D. R. Larson, "Frames, bases and group representations", *Mem. Amer. Math. Soc.* **147**:697 (2000), x+94. MR 2001a:47013

[Heil and Walnut 1989] C. E. Heil and D. F. Walnut, "Continuous and discrete wavelet transforms", *SIAM Rev.* **31**:4 (1989), 628–666. MR 91c:42032 Zbl 0683.42031

[Sun 2006] W. Sun, "G-frames and g-Riesz bases", *J. Math. Anal. Appl.* **322**:1 (2006), 437–452. MR 2007b:42047 Zbl 1129.42017

[Yong 1980] R. Yong, *An Introduction to non-harmonic Fourier series*, Academic Press, New York, 1980.

Received: 2008-09-29 Accepted: 2009-03-19

reza_joveini@bojnourdiau.ac.ir *Faculty of Science, Islamic Azad University of Bojnourd, Bojnourd 94176-94686, Iran*

mamini@modraes.ac.ir *Faculty of Mathematical Sciences, Tarbiat Modares University, Tehran 14115-175, Iran*

Current address: *Institut Penyelidikan Matematik, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia*

INVOLVE 2:4(2009)

A complete classification of $\mathbb{Z}_p$-sequences corresponding to a polynomial

Leonard Huang

(Communicated by Andrew Granville)

Let p be a prime number and set $\mathbb{Z}_p = \mathbb{Z}/p\mathbb{Z}$. A $\mathbb{Z}_p$-sequence is a function $S : \mathbb{Z} \to \mathbb{Z}_p$. Let $\mathcal{R}$ be the set $\{P \in \mathbb{R}[X] \mid P(\mathbb{Z}) \subseteq \mathbb{Z}\}$. We prove that the set of sequences of the form $(P(n) \pmod p)_{n\in\mathbb{Z}}$, where $P \in \mathcal{R}$, is precisely the set of periodic $\mathbb{Z}_p$-sequences with period equal to a p-power. Given a $\mathbb{Z}_p$-sequence, we will also determine all $P \in \mathcal{R}$ that correspond to the sequence according to the manner above.

1. Preliminaries

Let $\mathbb{N} = \{1, 2, 3, \ldots\}$ and $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$.

Definition 1. Define the sequence $(P_i)_{i\in\mathbb{N}_0}$ of polynomials in $\mathbb{R}[X]$ as follows:

$$P_0 = 1 \quad \text{and} \quad \text{for all } i \in \mathbb{N} : P_i = \binom{X}{i} = \frac{\prod_{j=0}^{i-1}(X-j)}{i!}.$$

Lemma 2 [Niven et al. 1991, pp. 42–43, Problems 11, 14, 15]. *We have*

$$\mathcal{R} = \left\{ \sum_{i=0}^{m} c_i P_i \mid m \in \mathbb{N}_0, c_0, \ldots, c_m \in \mathbb{Z} \right\}.$$

Proof. Clearly, $\mathcal{R} \supseteq \left\{\sum_{i=0}^{m} c_i P_i \mid m \in \mathbb{N}_0, c_0, \ldots, c_m \in \mathbb{Z}\right\}$, so we only need to prove the reverse inclusion.

Let $P \in \mathcal{R}$ have degree m. If the system of equations

$$P(j) = \sum_{i=0}^{m} c_i P_i(j), \quad j = 0, \ldots, m, \tag{1}$$

MSC2000: 11B83.
Keywords: $\mathbb{Z}_p$-sequences, polynomials, free abelian group.

in the unknowns $c_0, \ldots, c_m$ has a solution $(c_0, \ldots, c_m) \in \mathbb{Z}^{m+1}$, then $P = \sum_{i=0}^{m} c_i P_i$ because P and $\sum_{i=0}^{m} c_i P_i$ are polynomials of degree at most m that agree at the $m+1$ points $0, \ldots, m$. However, (1) is equivalent to the system

$$c_j = P(j) - \sum_{i=0}^{j-1} \binom{j}{i} c_i, \quad j = 0, \ldots, m,$$

which clearly has a unique solution $(c_0, \ldots, c_m) \in \mathbb{Z}^{m+1}$. □

Lemma 3. *Let p be a prime number. For every $k \in \mathbb{N}$,*

$$\binom{p^k}{0} \equiv 1 \pmod{p} \quad \textit{and} \quad \textit{for all } i \in \{1, \ldots, p^k - 1\} : \binom{p^k}{i} \equiv 0 \pmod{p}.$$

Proof. The first identity is clearly true. When $i \in \{1, \ldots, p^k - 1\}$, we have

$$\frac{p^k}{i} = \frac{\binom{p^k}{i}}{\binom{p^k - 1}{i - 1}}.$$

Write i as $p^l m$, where $l \in \mathbb{N}_0$ and m is a positive integer not divisible by p. From the equation

$$\frac{p^{k-l}}{m} = \frac{p^k}{i} = \frac{\binom{p^k}{i}}{\binom{p^k - 1}{i - 1}},$$

we immediately obtain

$$p^{k-l} \binom{p^k - 1}{i - 1} = m \binom{p^k}{i}.$$

Since $i < p^k$, we have $k - l \geq 1$. Thus p divides $m\binom{p^k}{i}$, and since it does not divide m, it must divide $\binom{p^k}{i}$. This proves that

$$\binom{p^k}{i} \equiv 0 \pmod{p}.$$

As $i \in \{1, \ldots, p^k - 1\}$ was arbitrary, Lemma 3 is true. □

Lemma 4. *Let p be a prime number. Then, for every $n \in \mathbb{Z}$, $k \in \mathbb{N}_0$ and $i \in \{0, \ldots, p^k - 1\}$,*

$$\binom{n + p^k}{i} \equiv \binom{n}{i} \pmod{p}. \tag{2}$$

Proof. Let $k \in \mathbb{N}_0$. Define a well-ordering $\prec$ on $\{0, \ldots, p^k - 1\} \times \mathbb{N}_0$ by setting $(i, n) \prec (i', n')$ if either (i) $i < i'$, or (ii) $i = i'$ and $n < n'$. By the principle of induction, it suffices to prove the following statements:

(A) For every $i \in \{0, \ldots, p^k - 1\}$,

$$\binom{0+p^k}{i} \equiv \binom{0}{i} \pmod{p}.$$

(B) Given $(i^*, n^*) \in \{0, \ldots, p^k - 1\} \times \mathbb{N}_0$, if

$$\text{for every } (i,n) \preceq (i^*, n^*): \quad \binom{n+p^k}{i} \equiv \binom{n}{i} \pmod{p}, \tag{3}$$

and

$$\text{for every } (i,n) \preceq (i^*, n^*): \quad \binom{-n+p^k}{i} \equiv \binom{-n}{i} \pmod{p}, \tag{4}$$

then, respectively,

$$\binom{n^*+1+p^k}{i^*} \equiv \binom{n^*+1}{i^*} \pmod{p} \tag{5}$$

and

$$\binom{-(n^*+1)+p^k}{i^*} \equiv \binom{-(n^*+1)}{i^*} \pmod{p}. \tag{6}$$

Statement (A) holds by Lemma 3. For Statement (B), we consider two cases: (i) $i^* = 0$ and (ii) $i^* > 0$. In Case (i), (B) is vacuously true. In Case (ii), we deduce (5) from (3) by applying Pascal's Rule:

$$\begin{aligned}\binom{n^*+1+p^k}{i^*} &= \binom{n^*+p^k}{i^*-1} + \binom{n^*+p^k}{i^*} && \text{(by Pascal's Rule)}\\ &\equiv \binom{n^*}{i^*-1} + \binom{n^*}{i^*} && \text{(from (3))}\\ &\equiv \binom{n^*+1}{i^*} \pmod{p} && \text{(by Pascal's Rule again).}\end{aligned}$$

In a similar fashion, we deduce (6) from (4):

$$\begin{aligned}\binom{-(n^*+1)+p^k}{i^*} &= \binom{-n^*+p^k}{i^*} - \binom{-(n^*+1)+p^k}{i^*-1}\\ &\equiv \binom{-n^*}{i^*} - \binom{-(n^*+1)}{i^*-1} \equiv \binom{-(n^*+1)}{i^*} \pmod{p}.\end{aligned}$$

Therefore (B) is true in Case (ii). Since $k \in \mathbb{N}_0$ was arbitrary, Lemma 4 is true. □

Corollary 5. *For every $i \in \mathbb{N}_0$, the sequence $\left(\binom{n}{i} \pmod{p}\right)_{n\in\mathbb{Z}}$ is periodic with period equal to a p-power.*

Proof. Choose $k \in \mathbb{N}_0$ such that $i < p^k$. By Lemma 4, $\binom{n+p^k}{i} \equiv \binom{n}{i} \pmod{p}$ for every $n \in \mathbb{Z}$. This clearly implies the claim. □

Corollary 6. *For every $P \in \mathcal{R}$, the sequence $(P(n) \pmod{p})_{n\in\mathbb{Z}}$ is periodic with period equal to a p-power.*

Proof. Let $P \in \mathcal{R}$. By Lemma 2, there exist $m \in \mathbb{N}_0$ and $c_0, \ldots, c_m \in \mathbb{Z}$ such that $P = \sum_{i=0}^{m} c_i P_i$. Then,

$$(P(n) \pmod p)_{n\in\mathbb{Z}} = \left(\sum_{i=0}^{m} c_i P_i(n) \pmod p \right)_{n\in\mathbb{Z}}.$$

By Corollary 5, each $(P_i(n) \pmod p)_{n\in\mathbb{Z}}$ is periodic with period equal to a p-power. We conclude that $(P(n) \pmod p)_{n\in\mathbb{Z}}$ is also periodic with period equal to a p-power. □

2. Main results

Theorem 7. *Let $p\mathcal{R}$ be the subset of $\mathcal{R}$ obtained by multiplying every $P \in \mathcal{R}$ by p. A polynomial $P \in \mathcal{R}$ lies in $p\mathcal{R}$ if and only if p divides $P(n)$ for all $n \in \mathbb{Z}$; in symbols,*

$$p\mathcal{R} = \left\{ P \in \mathcal{R} \mid (P(n) \pmod p)_{n\in\mathbb{Z}} = (0)_{n\in\mathbb{Z}} \right\}.$$

Proof. It is clear that every polynomial in $p\mathcal{R}$ corresponds to $(0)_{n\in\mathbb{Z}}$, so let us suppose that $P \in \mathcal{R}$ satisfies

$$(P(n) \pmod p)_{n\in\mathbb{Z}} = (0)_{n\in\mathbb{Z}}.$$

Then, by Lemma 2, there exist $m \in \mathbb{N}_0$ and $c_0, \ldots, c_m \in \mathbb{Z}$ such that $P = \sum_{i=0}^{m} c_i P_i$. We claim that $c_0, \ldots, c_m \equiv 0 \pmod p$.

To prove the claim, we use mathematical induction. By our hypothesis,

$$P(0) = \sum_{i=0}^{m} c_i P_i(0) = c_0 \equiv 0 \pmod p.$$

Hence, the claim is true for c_0. Next, suppose that $k \in \mathbb{N}_0$ and that the claim is true for c_j for every $j \le k$. If $j = m$, we are done. If $j < m$, then

$$P(j+1) = \sum_{i=0}^{m} c_i P_i(j+1) \equiv c_{j+1} \equiv 0 \pmod p.$$

Hence, the claim is true for c_{j+1} as well. By induction, the claim is true for all $c_0, \ldots, c_m$. This shows that $P \in p\mathcal{R}$. □

Theorem 8. *The set of sequences of the form $(P(n) \pmod p)_{n\in\mathbb{Z}}$, where $P \in \mathcal{R}$, is precisely the set of periodic $\mathbb{Z}_p$-sequences with period equal to a p-power.*

Proof. By virtue of Corollary 6, we only have to prove that every periodic $\mathbb{Z}_p$-sequence with period equal to a p-power corresponds to some $P \in \mathcal{R}$.

Let $k \in \mathbb{N}_0$. Define A to be the set

$$\left\{ \sum_{i=0}^{p^k-1} c_i P_i \;\middle|\; c_1, \ldots, c_{p^k-1} \in \{0, \ldots, p-1\} \right\},$$

and B to be set of all periodic $\mathbb{Z}_p$-sequences with period equal to p^l, where $0 \le l \le k$. By Lemma 4 and Theorem 7, every polynomial in A corresponds to a unique sequence in B. Since $|A| = |B| = p^{p^k}$, the correspondence is actually one-to-one. Therefore, every periodic $\mathbb{Z}_p$-sequence with period p^k corresponds to a unique polynomial of the form

$$\sum_{i=0}^{p^k-1} c_i P_i,$$

where $c_1, \ldots, c_{p^k-1} \in \{0, \ldots, p-1\}$. Since k was arbitrary, Theorem 8 is proven.

The theorem, however, would not be of much use unless the coefficients c_i can be determined. Hence, let S be a periodic $\mathbb{Z}_p$-sequence with period p^k, where $k \in \mathbb{N}_0$. By the first part, there exist $c_1, \ldots, c_{p^k-1} \in \{0, \ldots, p-1\}$ such that

$$S = \left(\sum_{i=0}^{p^k-1} c_i P_i(n) \pmod p \right)_{n \in \mathbb{Z}}.$$

From this identity, we obtain the equations

$$S(j) = \sum_{i=0}^{p^k-1} c_i \binom{j}{i}, \quad j = 0, \ldots, p^k - 1.$$

Some algebraic manipulation shows that the c_i's satisfy

$$c_i \equiv \sum_{j=0}^{i} (-1)^j \binom{i}{j} S(i-j) \pmod p, \quad i = 0, \ldots, p^k - 1. \qquad \square$$

Corollary 9. *Let S be a periodic $\mathbb{Z}_p$-sequence with period p^k, where $k \in \mathbb{N}_0$. Then, the set of all $P \in \mathcal{R}$ which correspond to S is*

$$\left(\sum_{i=0}^{p^k-1} c_i P_i \right) + p\mathcal{R},$$

where c_i is the least positive residue of $\sum_{j=0}^{i} (-1)^j \binom{i}{j} S(i-j) \pmod p$ for every $i = 0, \ldots, p^k - 1$.

Proof. Let c_i satisfy the hypothesis given in the corollary. By Theorem 8,

$$S - \left(\sum_{i=0}^{p^k-1} c_i P_i(n) \pmod{p} \right)_{n\in\mathbb{Z}} = (0)_{n\in\mathbb{Z}}.$$

The corollary now follows directly from Theorem 7. □

With both theorems and Corollary 9, we have a complete classification of $\mathbb{Z}_p$-sequences corresponding to a polynomial. Let us now look at some examples.

3. Examples

Example 10. To determine whether or not a $\mathbb{Z}_p$-sequence corresponds to a polynomial, simply investigate its periodicity. For example, the $\mathbb{Z}_7$-sequence

$$\ldots, \widehat{4}, 0, 6, 4, 0, 6, \ldots$$

(the $\widehat{\ }$ marks the zeroth element of the sequence) does not correspond to any polynomial because, although periodic, it has period 3, which is not a power of 7.

Example 11. The $\mathbb{Z}_3$-sequence

$$\ldots, \widehat{1}, 0, 1, 2, 0, 1, 1, 0, 2, \ldots$$

is periodic with period $9 = 3^2$, so it corresponds to a polynomial. The proof of Theorem 8 says that the sequence corresponds to

$$\binom{X}{0} + 2\binom{X}{1} + 2\binom{X}{2} + \binom{X}{3} + 2\binom{X}{4} + \binom{X}{5} + \binom{X}{6} + 2\binom{X}{7} + 2\binom{X}{8}.$$

4. Conclusion

We can add some algebraic flavor to the classification problem as follows. Let $\mathcal{S}$ denote the set of periodic $\mathbb{Z}_p$-sequences with period equal to a p-power. It is not difficult to see that $\mathcal{S}$ forms an abelian group under component-wise addition. Notice also that $\mathcal{R}$ is a free abelian group generated by the set $\{P_i \mid i \in \mathbb{N}_0\}$ and that the mapping

$$\begin{aligned} &\phi : \mathcal{R} \to \mathcal{S}, \\ &\phi : P \mapsto (P(n) \pmod{p})_{n\in\mathbb{Z}} \end{aligned}$$

is a surjective group homomorphism. By the first isomorphism theorem for groups, $\mathcal{R}/\ker(\phi) \cong \mathcal{S}$. However, Theorem 7 says that $\ker(\phi) = p\mathcal{R}$, so we obtain $\mathcal{R}/p\mathcal{R} \cong \mathcal{S}$. This elegant algebraic identity summarizes much of the effort invested in this paper.

Theorem 8 may be generalized so as to obtain a complete classification of all $\mathbb{Z}_m$-sequences corresponding to a polynomial for an arbitrary integer $m \geq 2$. The

first step to doing this is to consider the case when m is a prime-power. It would be too good to be true for Lemma 4 to hold if we replace p in (2) by p^a for arbitrary $a \in \mathbb{N}$, and indeed it is.[1] We have the following counterexample:

$$\binom{4+3^2}{3} \equiv 7 \pmod 9 \quad \text{but} \quad \binom{4}{3} \equiv 4 \pmod 9.$$

However, an analogous equation for prime-powers may be obtained from the following proposition (see Theorem 1 of [Granville 1997]):

Proposition 12. *Let p be a prime number. For any positive integer a, define $(a!)_p$ to be the product of those positive integers $\le a$ which are not divisible by p. Let q, m, n and r be positive integers such that $n = m + r$. Write n in base p as $\sum_{i=0}^{d} n_i p^i$ and let N_j be the least positive residue of $[n/p^j] \pmod{p^q}$ for each $j \ge 0$ (so that $N_j = \sum_{i=0}^{q-1} n_{j+i} p^i$). Also, make the corresponding definitions for m_j, M_j, r_j and R_j. Let e_j denote the number of 'carries', when adding m and r in base p, on and beyond the j-th digit. Then,*

$$\frac{(\pm 1)^{e_{q-1}}}{p^{e_0}} \binom{n}{m} \equiv \left(\prod_{j=0}^{d} \frac{(N_j!)_p}{(M_j!)_p (R_j!)_p} \right) \pmod{p^q},$$

where $(\pm 1) = -1$ except if $p = 2$ and $q \ge 3$.

We will not attempt to generalize Theorem 8 in this paper because it would take us too far afield.

Acknowledgments

I thank Dr. Fedor Duzhin from the School of Physical and Mathematical Sciences (SPMS), Nanyang Technological University, for first sparking my interest in this problem. Many thanks also go to Dr. Sinai Robins (SPMS) for the initial advice he gave on preparing this paper, and to Magdalene Lee for her unwavering support of my mathematical endeavors. I express my warmest gratitude, however, to the anonymous referee who painstakingly reviewed this paper and, at the same time, corrected many of the flaws in my practice of mathematical exposition.

References

[Granville 1997] A. Granville, "Arithmetic properties of binomial coefficients, I: Binomial coefficients modulo prime powers", pp. 253–276 in *Organic mathematics* (Burnaby, BC, 1995), edited by J. Borwein et al., CMS Conf. Proc. **20**, Amer. Math. Soc., Providence, RI, 1997. MR 99h:11016 Zbl 0903.11005

[1] Note that Theorem 7 still holds if we replace p by p^a, or even by any integer ≥ 2. It is only for Theorem 8 that this method fails, which, in turn, happens because it fails for Lemmas 3 and 4.

[Niven et al. 1991] I. Niven, H. S. Zuckerman, and H. L. Montgomery, *An introduction to the theory of numbers*, 5th ed., Wiley, New York, 1991. MR 91i:11001 Zbl 0742.11001

Received: 2008-10-25 Revised: 2009-09-15 Accepted: 2009-09-15

huan0074@ntu.edu.sg *School of Physical and Mathematical Sciences, Nanyang Technological University, SPMS-04-01, 21 Nanyang Link, 637371, Singapore*

INVOLVE 2:4(2009)

Newton's law of heating and the heat equation

Mark Gockenbach and Kristin Schmidtke

(Communicated by Suzanne Lenhart)

Newton's law of heating models the average temperature in an object by a simple ordinary differential equation, while the heat equation is a partial differential equation that models the temperature as a function of both space and time. A series solution of the heat equation, in the case of a spherical body, shows that Newton's law gives an accurate approximation to the average temperature if the body is not too large and it conducts heat much faster than it gains heat from the surroundings. Finite element simulation confirms and extends the analysis.

1. Introduction

A popular application involving an elementary differential equation is Newton's law of heating, which describes the change in temperature in an object whose surroundings are hotter than it is.[1] If the temperature at time t is $T(t)$, then Newton's law of heating is

$$T' = \overline{\alpha}(T_s - T), \quad T(0) = T_0, \tag{1}$$

where T_s and T_0 are constants representing the temperature of the surroundings and the initial temperature of the object, respectively.[2] The differential equation in (1) simply states that the rate of change of the temperature is proportional to the difference between the temperatures of the surroundings and the object. The solution of (1) is

$$T(t) = T_s - (T_s - T_0)e^{-\overline{\alpha}t}. \tag{2}$$

MSC2000: 35K05.

Keywords: heat equation, Newton's law of heating, finite elements, Bessel functions.

[1] If the surroundings are colder, then the differential equation is called Newton's law of cooling. For definiteness of language, we will usually assume that heating is occurring.

[2] We take T_0 and T_s to be constants, as this agrees with the usual textbook presentation of Newton's law of heating, which we wish to analyze in this paper. There is nothing that would prevent us from allowing T_s to depend on time, or T_0 to depend on space (so that $T_0(x, y, z)$ is the initial temperature at the point (x, y, z) in the object). If T_0 were variable, then we would use the average value $\overline{T}_0$ of T_0 in the ordinary differential equation model (1) and the variable T_0 itself in the partial differential equation model presented below. Allowing nonconstant T_0 and/or T_s would considerably complicate the analysis in this paper.

Newton's law of heating is included in virtually every introductory textbook on differential equations, such as [Boyce and DiPrima 1992; Zill 2005].

Newton's law of heating assumes that the temperature of the object is represented by a single number. A more sophisticated model represents the object as occupying a domain Ω in R^3 and its temperature as a function $u(x, y, z, t)$, where $(x, y, z) \in \Omega$. The governing partial differential equation is the heat equation

$$\rho c \frac{\partial u}{\partial t} = \kappa \Delta u \text{ in } \Omega, \quad t > 0. \tag{3}$$

In this equation, Δu is the Laplacian of u:

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}.$$

The quantities ρ, c, and κ describe the material properties of the object; ρ is the density, c is the specific heat, and κ is the thermal conductivity. Typically, κ is given in J/s cm K, ρ in g/cm^3, and c in J/g K. For a derivation of the heat equation, the reader can consult introductory books on partial differential equations, such as [Gockenbach 2002; Haberman 2004].

To obtain a complete description of u, we must model how Ω exchanges heat energy with its surroundings. We adopt a model much like Newton's law of heating:

$$\kappa \frac{\partial u}{\partial n} = \alpha(T_s - u) \quad \text{on } \partial\Omega. \tag{4}$$

This is called a *Robin boundary condition*; it states that the heat flux across the boundary is proportional to the difference between T_s and the temperature on $\partial\Omega$. Although the form of the boundary condition is analogous to Newton's law of heating, there is no reason to expect that the constants $\overline{\alpha}$ and α are the same; indeed, since they have different units, it would be surprising if their numerical values were the same.

The PDE (3) and the boundary condition (4), together with an initial condition, form a well-posed problem that determines u uniquely. We assume that the initial temperature in Ω is constant and obtain the following initial boundary value problem (IBVP):

$$\begin{aligned} \rho c \frac{\partial u}{\partial t} - \kappa \Delta u &= 0 \qquad \text{in } \Omega,\ t > 0, \\ u(x, y, z, 0) &= T_0 \qquad \text{in } \Omega, \\ \kappa \frac{\partial u}{\partial n} + \alpha u &= \alpha T_s \quad \text{on } \partial\Omega,\ t > 0. \end{aligned} \tag{5}$$

We now have two models to describe the temperature of the given object, namely, Newton's law of heating and the heat equation together with a Robin boundary

condition. The simpler model (1) would be an adequate substitute for the more complicated model (5) if T is close to the average temperature in Ω as predicted by (5):

$$\overline{u} = \frac{1}{|\Omega|}\int_{\Omega} u. \tag{6}$$

Before we can compare the solutions T and $\overline{u}$, we must determine the relative values of the constants $\overline{\alpha}$ and α appearing in (1) and (5), respectively. Fortunately, given α, the value of $\overline{\alpha}$ is suggested by the following calculation:

$$\begin{aligned}\frac{d\overline{u}}{dt} &= \frac{1}{|\Omega|}\int_{\Omega}\frac{\partial u}{\partial t} = \frac{1}{\rho c|\Omega|}\int_{\Omega}\rho c\frac{\partial u}{\partial t} = \frac{1}{\rho c|\Omega|}\int_{\Omega}\kappa\,\Delta u \\ &= \frac{1}{\rho c|\Omega|}\int_{\partial\Omega}\kappa\frac{\partial u}{\partial n} = \frac{1}{\rho c|\Omega|}\int_{\partial\Omega}\alpha(T_s - u).\end{aligned}$$

If we define $\overline{u}_b = \overline{u}_b(t)$ to be the average value of u on $\partial\Omega$,

$$\overline{u}_b = \frac{1}{|\partial\Omega|}\int_{\partial\Omega} u,$$

then

$$\int_{\partial\Omega}\alpha(T_s - u) = \alpha|\partial\Omega|(T_s - \overline{u}_b),$$

and we obtain

$$\overline{u}' = \frac{\alpha|\partial\Omega|}{\rho c|\Omega|}(T_s - \overline{u}_b). \tag{7}$$

We then see that $\overline{u}$ satisfies a differential equation similar to Newton's law of heating, but with α replaced with

$$\overline{\alpha} = \frac{|\partial\Omega|}{\rho c|\Omega|}\alpha. \tag{8}$$

Of course, even with this value of $\overline{\alpha}$, (1) and (7) are not the same, since (7) shows that the rate of change of $\overline{u}$ is determined not by $\overline{u}$ itself, but by $\overline{u}_b$. However, the equations are similar enough that we might expect T to be a good approximation to $\overline{u}$.

The primary purpose of this paper is to compare the solutions of the heat equation (5) and Newton's law of heating (1), where $\overline{\alpha}$ is given by (8). We will use both analytical and numerical methods; to make the analysis tractable and the numerics simpler, we will assume that the object is spherical.

2. Solution of the heat equation on a spherical domain

We will henceforth assume that Ω is the ball of radius R centered at the origin. Since the initial value of u is a constant and the boundary conditions are constant on $\partial\Omega$, the solution to the IBVP (5) depends only on the radial variable $r = \sqrt{x^2+y^2+z^2}$. This follows from the fact that the Laplacian Δ is invariant under any rotation of the coordinate system (for a nice discussion of this, the reader can consult [Folland 1995, Section 2.A]). We can therefore write $u = u(r,t)$, and we recall that

$$\Delta u = \frac{\partial^2 u}{\partial r^2} + \frac{2}{r}\frac{\partial u}{\partial r}$$

(see, for example, [Haberman 2004]). The IBVP (5) can thus be rewritten as

$$\begin{aligned} \rho c \frac{\partial u}{\partial t} - \kappa\left(\frac{\partial^2 u}{\partial r^2} + \frac{2}{r}\frac{\partial u}{\partial r}\right) &= 0, && 0 < r < R,\ t > 0, \\ u(r,0) &= T_0, && 0 < r < R, \\ \kappa \frac{\partial u}{\partial r}(R,t) + \alpha u(R,t) &= \alpha T_s, && t > 0. \end{aligned} \tag{9}$$

In addition to the equations explicitly listed above, there is the implicit requirement that u be finite at $r = 0$. We use the technique of *shifting the data* to transform (9) to a problem with homogeneous boundary conditions. We define $Y(r) = ar^2$, where a is chosen so that Y satisfies the boundary condition $\kappa Y'(R) + \alpha Y(R) = \alpha T_s$,

$$a = \frac{\alpha T_s}{2\kappa R + \alpha R^2}, \tag{10}$$

and write $u(r,t) = U(r,t) + Y(r)$. Then U satisfies

$$\begin{aligned} \rho c \frac{\partial U}{\partial t} - \kappa\left(\frac{\partial^2 U}{\partial r^2} + \frac{2}{r}\frac{\partial U}{\partial r}\right) &= 6a\kappa, && 0 < r < R,\ t > 0, \\ U(r,0) &= T_0 - ar^2, && 0 < r < R, \\ \kappa \frac{\partial U}{\partial r}(R,t) + \alpha U(R,t) &= 0, && t > 0. \end{aligned} \tag{11}$$

We can derive a solution to (11) by expanding U in terms of the eigenfunctions of the spatial operator

$$L = -\kappa\left(\frac{d^2}{dr^2} + \frac{2}{r}\frac{d}{dr}\right),$$

where homogeneous Robin conditions are imposed on the eigenfunctions:

$$\kappa v'(R) + \alpha v(R) = 0. \tag{12}$$

We will now briefly derive these eigenfunctions and the corresponding eigenvalues. First, using integration by parts, we can show that if v_1, v_2 satisfy (12), then

$$\int_0^R -\kappa\left(\frac{d^2v_1}{dr^2}+\frac{2}{r}\frac{dv_1}{dr}\right)v_2(r)r^2\,dr = \alpha v_1(R)v_2(R)R^2 + \int_0^R \kappa\frac{dv_1}{dr}\frac{dv_2}{dr}r^2\,dr.$$

This shows that L is symmetric with respect to the inner product

$$\langle v_1, v_2\rangle = \int_0^r v_1(r)v_2(r)r^2\,dr,$$

that is, that $\langle L(v_1), v_2\rangle = \langle v_1, L(v_2)\rangle$ for all v_1, v_2 satisfying the given boundary conditions. The standard argument then shows that the eigenvalues of L are all real, and that eigenfunctions of L corresponding to distinct eigenvalues are orthogonal with respect to the given inner product [Gockenbach 2002, Section 5.1]. We also see that

$$\langle L(v), v\rangle = \alpha v(R)^2 + \int_0^R \kappa\left(\frac{dv}{dr}(r)\right)^2 r^2\,dr,$$

which is positive for every nonzero function v. This implies that all the eigenvalues of L are positive.

We now wish to solve

$$\begin{aligned} -\kappa\left(\frac{d^2v}{dr^2}+\frac{2}{r}\frac{dv}{dr}\right) &= \lambda v, \quad 0 < r < R, \\ \kappa v'(R) + \alpha v(R) &= 0, \end{aligned} \tag{13}$$

for $\lambda > 0$ and $v = v(r)$. It is well known [Arfken and Weber 2005] that the only solutions to $(13)_1$ that are bounded at the origin are multiples of

$$j_0\left(\sqrt{\lambda/\kappa}\,r\right),$$

where j_0 is the first spherical Bessel function:

$$j_0(s) = \frac{\sin(s)}{s}.$$

(For more information about Bessel functions, including the properties cited below, the reader can consult [Trantor 1968] or the comprehensive reference [**?**].) The problem of finding the eigenvalues and eigenfunctions then reduces to finding the values of $\lambda > 0$ such that $v(r) = j_0\left(\sqrt{\lambda/\kappa}\,r\right)$ satisfies the boundary condition $(13)_2$.

Substituting v into $(13)_2$ and simplifying yields

$$\tan(s) = ms, \quad m = \frac{\kappa}{\kappa - \alpha R}, \quad s = R\sqrt{\lambda/\kappa}. \tag{14}$$

We will henceforth make the important assumption that $\kappa - \alpha R > 0$ or, equivalently, that

$$\beta = \frac{\alpha R}{\kappa} < 1. \tag{15}$$

Recalling that κ is the thermal conductivity within the object and α describes how well thermal energy is transmitted to the environment, this assumption means that the object conducts energy more quickly than it transmits energy to the surroundings and also that the radius of the object is not too large. Our intuition ought to tell us that these are precisely the conditions under which Newton's law of heating should give accurate results. In fact, below we will expand $\overline{u}(t) - T(t)$ in powers of β and show that $\overline{u}(t) - T(t) = O(\beta)$ uniformly for $t \geq 0$.

Assumption (15) implies that $m = 1/(1-\beta) > 1$ in (14), and a simple graph then shows that (14) has positive solutions $s_1, s_2, s_3, \ldots$, with

$$(k-1)\pi < s_k < \left(k - \frac{1}{2}\right)\pi, \quad k = 1, 2, 3, \ldots$$

and $s_k \approx \left(k - \frac{1}{2}\right)\pi$ for $k \geq 2$. For our analysis below, we need accurate estimates of the s_k's. To estimate s_1, we can expand $\tan(s)$ in powers of s, truncate the series, and obtain

$$\begin{aligned} s_1 &= \sqrt{3}\beta^{1/2} - \frac{\sqrt{3}}{10}\beta^{3/2} + O(\beta^{5/2}), \\ \lambda_1 &= \frac{3\kappa}{R^2}\left(\beta - \tfrac{1}{5}\beta^2 + O(\beta^3)\right). \end{aligned} \tag{16}$$

For $k \geq 2$, we see that each s_k is greater than the corresponding solution $\overline{s}_k$ to $\tan(s) = s$. We write $\overline{s}_k = \left(k - \frac{1}{2}\right)\pi - \epsilon_k$, expand $\tan\left(\left(k - \frac{1}{2}\right)\pi - \epsilon_k\right)$ in powers of ϵ_k, and solve to get

$$s_k \geq \overline{s}_k = \left(k - \frac{1}{2}\right)\pi\left(1 - \frac{1}{\left(k - \frac{1}{2}\right)^2\pi^2} - \frac{2}{3\left(k - \frac{1}{2}\right)^4\pi^4} + \ldots\right). \tag{17}$$

In particular, we find that

$$0.95\left(k - \tfrac{1}{2}\right)\pi < s_k < \left(k - \tfrac{1}{2}\right)\pi, \quad k = 2, 3, \ldots. \tag{18}$$

This estimate will suffice for our purposes below.

We now write

$$v_k(r) = j_0\left(\sqrt{\lambda_k/\kappa}\, r\right) = \frac{\sin\left(\sqrt{\lambda_k/\kappa}\, r\right)}{\sqrt{\lambda_k/\kappa}\, r}, \quad k = 1, 2, 3, \ldots, \tag{19}$$

for the eigenfunctions of L. The standard theory of (spherical) Bessel functions guarantees that any function $u = u(r)$ (finite at the origin) can be expanded in terms

of the orthogonal functions $v_1, v_2, v_3, \ldots$. We can therefore expand the solution U of (11) as

$$U(r,t) = \sum_{k=1}^{\infty} C_k(t) v_k(r).$$

Substituting this expression into the PDE $(11)_1$ yields

$$\sum_{k=1}^{\infty} \left\{ \rho c C_k'(t) + \lambda_k C_k(t) \right\} v_k(r) = \sum_{k=1}^{\infty} b_k v_k(r),$$

where $\sum_{k=1}^{\infty} b_k v_k(r)$ is the expansion of the constant function $6a\kappa$:

$$b_k = \frac{6a\kappa \int_0^R v_k(r) r^2 \, dr}{\int_0^R v_k(r)^2 r^2 \, dr}, \quad k = 1, 2, 3, \ldots. \tag{20}$$

From the initial condition $(11)_2$, we have

$$\sum_{k=1}^{\infty} C_k(0) v_k(r) = \sum_{k=1}^{\infty} d_k v_k(r),$$

where $\sum_{k=1}^{\infty} d_k v_k(r)$ is the expansion of $T_0 - ar^2$:

$$d_k = \frac{\int_0^R (T_0 - ar^2) v_k(r) r^2 \, dr}{\int_0^R v_k(r)^2 r^2 \, dr}, \ k = 1, 2, 3, \ldots. \tag{21}$$

We then obtain the following initial value problem for C_k:

$$\rho c C_k' + \lambda_k C_k = b_k, \ C_k(0) = d_k, \quad k = 1, 2, 3, \ldots.$$

The solution is

$$C_k(t) = \frac{b_k}{\lambda_k} + \left(d_k - \frac{b_k}{\lambda_k} \right) e^{-\lambda_k t/(\rho c)}, \quad k = 1, 2, 3, \ldots. \tag{22}$$

We now have the following solution to the IBVP (9):

$$u(r,t) = ar^2 + \sum_{k=1}^{\infty} C_k(t) v_k(r). \tag{23}$$

The reader will recall that

$$\overline{u} = \frac{1}{|\Omega|} \int_{\Omega} u.$$

With Ω equal to the ball of radius R, this reduces to

$$\overline{u}(t) = \frac{4\pi}{(4/3)\pi R^3} \int_0^R u(r,t) r^2 \, dr = \frac{3}{R^3} \int_0^R u(r,t) r^2 \, dr,$$

and so

$$\overline{u}(t) = \frac{3}{R^3}\int_0^R ar^4\,dr + \sum_{k=1}^{\infty}\left\{\frac{3}{R^3}\int_0^R v_k(r)r^2\,dr\right\}C_k(t) \tag{24}$$

(since the solution u to the heat equation is known to be smooth, the series for u can be integrated term-by-term to produce (24)).

3. Comparing the solutions

We now have formulas for both $T(t)$ and $\overline{u}(t)$, and we wish to bound $|T(t)-\overline{u}(t)|$ for $t \geq 0$. Since v_k is oscillatory for $k \geq 2$ (see Figure 1 and also the definition of v_k in (19)), we expect

$$\frac{3}{R^3}\int_0^R v_k(r)r^2\,dr$$

to be small for $k \geq 2$. We hypothesize, then, that $\overline{u}(t)$ will be well approximated by

$$\overline{u}_1(t) = \frac{3}{R^3}\int_0^R ar^4\,dr + \left(\frac{3}{R^3}\int_0^R v_1(r)r^2\,dr\right)C_1(t).$$

We therefore wish to show that $|T(t)-\overline{u}_1(t)|$ and $|\overline{u}_1(t)-\overline{u}(t)|$ are both small for $t \geq 0$.

A straightforward calculation shows that

$$\frac{3}{R^3}\int_0^R v_k(r)r^2\,dr = 3\frac{\sin\left(R\sqrt{\lambda_k/\kappa}\right) - R\sqrt{\lambda_k/\kappa}\,\cos\left(R\sqrt{\lambda_k/\kappa}\right)}{\left(R\sqrt{\lambda_k/\kappa}\right)^3},$$

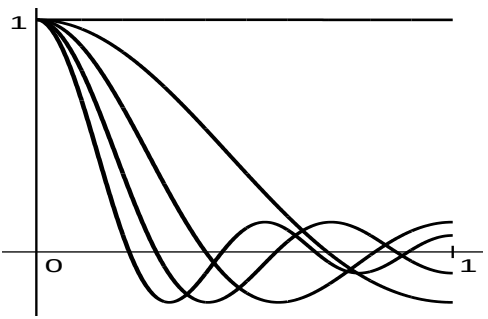


Figure 1. The first five eigenfunctions v_1, v_2, v_3, v_4, v_5 on the interval $[0, R]$, $R = 1.0$. To construct this graph, we have taken $\kappa = 1.0$ and $\alpha = 0.001$. The first eigenfunction is nearly constant on the interval $[0, R]$, while, for $k \geq 2$, v_k is increasingly oscillatory as k increases.

or

$$\frac{3}{R^3}\int_0^R v_k(r)r^2\,dr = 3\frac{\sin(s_k) - s_k\cos(s_k)}{s_k^3}.$$

Since $\tan(s_k) = ms_k$, or $s_k\cos s_k = m^{-1}\sin s_k$, we obtain

$$\frac{3}{R^3}\int_0^R v_k(r)r^2\,dr = 3\frac{m-1}{m}\frac{\sin s_k}{s_k^3} = 3\beta\frac{\sin s_k}{s_k^3}. \tag{25}$$

For $k \geq 2$, we can apply (18) to obtain

$$\left|\frac{3}{R^3}\int_0^R v_k(r)r^2\,dr\right| \leq \frac{3\beta}{s_k^3} \leq \frac{3\beta}{(0.95\,(k-1/2)\,\pi)^3},$$

which yields

$$\left|\frac{3}{R^3}\int_0^R v_k(r)r^2\,dr\right| \leq \frac{1}{8\,(k-1/2)^3}\beta, \quad k = 2, 3, \ldots. \tag{26}$$

For $k = 1$, we expand the integral in powers of β, which is a straightforward calculation:[3]

$$\frac{3}{R^3}\int_0^R v_1(r)r^2\,dr = 1 - \frac{3}{10}\beta + O\left(\beta^2\right). \tag{27}$$

The results given in (25) and (27) support our hypothesis that $\overline{u}(t)$ should be well approximated by $\overline{u}_1(t)$.

We can now complete the bound $|T(t) - \overline{u}_1(t)|$ in short order. From $(16)_2$, (20), and (21), we find that

$$\frac{b_1}{\lambda_1} = T_s + O\left(\beta^2\right)$$

and

$$d_1 - \frac{b_1}{\lambda_1} = T_0 - T_s + \frac{3}{10}(T_0 - T_s)\beta + O\left(\beta^2\right). \tag{28}$$

We can also expand the constant term in the series for $\overline{u}(t)$ in powers of β:

$$\frac{3}{R^3}\int_0^R ar^4\,dr = \frac{3aR^2}{5} = \frac{3T_s}{10}\beta + O\left(\beta^2\right).$$

Putting these results together, we obtain

$$\overline{u}_1(t) = T_s - (T_s - T_0)e^{-\lambda_1 t/(\rho c)} + O\left(\beta^2\right). \tag{29}$$

The reader should notice how the $O(\beta)$ term has canceled (compare the product of (27) and (28)). The only dependence of the $O\left(\beta^2\right)$ term on t is through the exponential (this dependence is not shown explicitly here), which is bounded by

[3] We used Mathematica to generate this and other series expansions.

one for $t \geq 0$. Thus the $O\left(\beta^2\right)$ term is uniformly small for $t \geq 0$ (assuming β is small).

The similarity between $T(t)$ and $\overline{u}_1(t)$ is now obvious (compare (2) and (29)). It remains only to compare the exponentials $e^{-\lambda_1 t/(\rho c)}$ and $e^{-\overline{\alpha} t} = e^{-3\alpha t/(\rho c R)}$. We have

$$\lambda_1 = \frac{3\kappa}{R^2}\beta\left(1 - \frac{1}{5}\beta + O(\beta^2)\right) = \frac{3\alpha}{R}\left(1 - \frac{1}{5}\beta + O(\beta^2)\right) \tag{30}$$

(using $\beta = \alpha R/\kappa$), so we see that $\lambda_1/(\rho c)$ and $\overline{\alpha} = 3\alpha/(\rho c R)$ are quite similar, with

$$e^{-\lambda_1 t/(\rho c)} \geq e^{-\overline{\alpha} t}, \quad t \geq 0.$$

To obtain a useful bound, we maximize the function

$$f(t) = e^{-\lambda_1 t/(\rho c)} - e^{-\overline{\alpha} t}, \quad t \geq 0.$$

The function f has a unique stationary point, and we easily obtain

$$0 \leq e^{-\lambda_1 t/(\rho c)} - e^{-\overline{\alpha} t} \leq \frac{1}{5e}\beta + O\left(\beta^2\right), \quad t \geq 0. \tag{31}$$

We can now bound the difference between $T(t)$ and $\overline{u}_1(t)$:

$$\begin{aligned} |T(t) - \overline{u}_1(t)| &= \left|T_s - (T_s - T_0)e^{-\overline{\alpha} t} - T_s + (T_s - T_0)e^{-\lambda_1 t/(\rho c)} + O(\beta^2)\right| \\ &= \left|(T_s - T_0)(e^{-\lambda_1 t/(\rho c)} - e^{-\overline{\alpha} t})\right| + O(\beta^2) \\ &\leq \frac{|T_s - T_0|}{5e}\beta + O(\beta^2). \end{aligned}$$

As noted above, this bound is uniform over the interval $0 \leq t < \infty$.

Finally, we bound $|\overline{u}_1(t) - \overline{u}(t)|$ for $t \geq 0$. We will merely sketch the results, which the interested reader can verify. We already have the upper bound (26) for

$$\left|\frac{3}{R^3}\int_0^R v_k(r) r^2\, dr\right|.$$

We will need upper bounds for d_k and b_k/λ_k, which will require a lower bound for

$$\int_0^R v_k(r)^2 r^2\, dr.$$

A straightforward calculation gives

$$\int_0^R v_k(r)^2 r^2\, dr = \frac{\kappa}{\lambda_k}\left(\frac{R}{2} - \frac{\sin\left(2\sqrt{\lambda_k/\kappa}\, R\right)}{4\sqrt{\lambda_k/\kappa}}\right) \geq \frac{R^3}{4\,(k - 1/2)^2\,\pi^2} \tag{32}$$

(applying the upper bound for λ_k implied by (18)). We then have

$$b_k = \frac{6a\kappa \int_0^R v_k(r) r^2\, dr}{\int_0^R v_k(r)^2 r^2\, dr},$$

and (26) gives an upper bound for the numerator. Applying this upper bound together with (32) and simplifying yields

$$b_k \le \frac{3\kappa T_s \pi^2}{2R^5\left(k-\frac{1}{2}\right)}\beta^2.$$

Since (18) implies

$$\lambda_k \ge 0.95^2\left(k-\tfrac{1}{2}\right)^2\pi^2\kappa, \quad k=2,3,\ldots, \tag{33}$$

we obtain (after a little manipulation)

$$\frac{b_k}{\lambda_k} \le \frac{2T_s}{R^5\left(k-\frac{1}{2}\right)^3}\beta^2. \tag{34}$$

Obtaining a bound for d_k is more work. We have

$$\begin{aligned}
&\int_0^R (T_0 - ar^2)v_k(r)r^2\,dr \\
&\quad= \frac{\kappa}{(2+\beta)\lambda_k^{5/2}R^2}\cdot\Big\{-\sqrt{\lambda_k}R\big(6\beta\kappa T_s + \lambda_k R^2((2+\beta)T_0 - \beta T_s)\big)\cos\big(\sqrt{\lambda_k/\kappa}\,R\big) \\
&\qquad\qquad + \sqrt{\kappa}\big(6\beta\kappa T_s + \lambda_k R^2((2+\beta)T_0 - 3\beta T_s)\big)\sin\big(\sqrt{\lambda_k/\kappa}\,R\big)\Big\} \\
&\quad= \frac{\kappa^{3/2}}{(2+\beta)\lambda_k^{5/2}R^2}\cdot\Big\{(\sin s_k - s_k\cos s_k)\big(6\beta\kappa T_s + \lambda_k R^2((2+\beta)T_0 - \beta T_s)\big) \\
&\qquad\qquad - 2\beta T_2 \sin s_k\Big\}.
\end{aligned}$$

Since $\sin s_k = m s_k \cos s_k$, we have

$$\sin s_k - s_k\cos s_k = (m-1)s_k\cos s_k = \frac{\beta}{1-\beta}s_k\cos s_k.$$

Also, s_k is an approximate root of cosine; using (17) and the Taylor expansion of cosine around $s=(k-1/2)\,\pi$, we obtain $s_k\cos s_k = O(1)$. Using this and some more manipulation, we find positive constants γ_1 and γ_2 such that

$$\left|\int_0^R (T_0 - ar^2)v_k(r)r^2\,dr\right| \le \frac{\gamma_1\beta + \gamma_2\beta^2}{\lambda_k^{3/2}}, \quad k=2,3,\ldots.$$

This, together with the lower bound (32), yields

$$d_k = \frac{\int_0^R (T_0 - ar^2)v_k(r)r^2\,dr}{\int_0^R v_k(r)^2 r^2\,dr} \le \bar{\gamma}_1\beta + \bar{\gamma}_2\beta^2, \tag{35}$$

where $\bar{\gamma}_1$ and $\bar{\gamma}_2$ are positive constants.

We can finally use (26), (34), and (35) to bound $|\overline{u}_1(t) - \overline{u}(t)|$. We have

$$|\overline{u}_1(t) - \overline{u}(t)| \leq \sum_{k=2}^{\infty} \left(\frac{3}{R^3} \int_0^R v_k(r) r^2 \, dr \right) \left| \frac{b_k}{\lambda_k} + \left(d_k - \frac{b_k}{\lambda_k} \right) e^{-\lambda_k t/(\rho c)} \right|$$
$$\leq \sum_{k=2}^{\infty} \frac{\beta}{8(k-1/2)^3} \left(2 \left| \frac{b_k}{\lambda_k} \right| + |d_k| \right).$$

Since

$$\sum_{k=2}^{\infty} \frac{1}{\left(k - \frac{1}{2}\right)^3}$$

is finite, (34) and (35) yield positive constants $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ such that

$$|\overline{u}_1(t) - \overline{u}(t)| \leq \tilde{\gamma}_1 \beta^2 + \tilde{\gamma}_2 \beta^3, \ t \geq 0.$$

This, together with our earlier bound on $|T(t) - \overline{u}_1(t)|$, yields our final result:

$$|T(t) - \overline{u}(t)| \leq \frac{|T_s - T_0|}{5e} \beta + O(\beta^2), \quad t \geq 0. \tag{36}$$

The reader will recall that

$$\beta = \frac{\alpha R}{\kappa},$$

where κ is the thermal conductivity with the object, α describes how well the object transmits heat energy to its surroundings (or vice versa), and R is the radius of Ω. As long as $\alpha \ll \kappa$ and Ω is not too large, (36) shows that the average temperature in Ω will be well approximated by Newton's law of heating.

4. The finite element method

We wish to give some numerical examples to illustrate the above analysis. This requires that we be able to compute accurate solutions to the initial-boundary value problem (9). We will use the standard Galerkin–Crank–Nicolson finite element method, which we now briefly describe.

To compute the solution of (9), we first rewrite the problem in its variational form:

$$\int_0^R \rho c \frac{\partial u}{\partial t}(r,t) v(r) r^2 \, dr + \int_0^R \kappa \frac{\partial u}{\partial r}(r,t) v'(r) r^2 \, dr + \alpha R^2 u(R,t) v(R)$$
$$= \alpha R^2 T_s v(R), \quad \text{for all } v \in V. \tag{37}$$

Here V is the space of *test functions*,

$$V = \left\{ v \in H^1(0,R) \, : \, rv \in L^2(0,R), \ rv' \in L^2(0,R) \right\},$$

based on the Sobolev space $H^1(0, R)$ (the space of functions with one square-integrable (weak) derivative). The variational form (37) results from multiplying the PDE (9) by a test function, integrating over Ω, integrating the $\partial^2 u/\partial r^2$ term by parts, and applying the boundary condition. It is well known that the variational form is equivalent to the original initial-boundary value problem (at least when, as in this case, the original problem is known to have a smooth solution).

We obtain the semidiscrete form of (37) by discretizing in space using piecewise linear functions on a mesh defined by $r_i = ih$, $h = R/n$, and applying Galerkin's method. We will write V_h for the space of continuous piecewise linear functions on the given mesh, and $\{\phi_0, \phi_1, \ldots, \phi_n\}$ for the usual nodal basis defined by

$$\phi_i(r_j) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

The semidiscrete solution is

$$u_h(r, t) = \sum_{j=0}^{n} \alpha_j(t)\phi_j(r),$$

satisfying

$$\int_0^R \rho c \frac{\partial u_h}{\partial t}(r, t)v(r)r^2\,dr + \int_0^R \kappa \frac{\partial u_h}{\partial r}(r, t)v'(r)r^2\,dr + \alpha R^2 u_h(R, t)v(R) = \alpha R^2 T_s v(R), \quad \text{for all } v \in V_h. \tag{38}$$

Choosing $v = \phi_i$, $i = 0, 1, \ldots, n$, (38) is equivalent to

$$Ma' + (K + G)a = F, \tag{39}$$

where M and K are the mass and stiffness[4] matrices,

$$M_{ij} = \int_0^R \rho c\phi_j(r)\phi_i(r)r^2\,dr, \quad K_{ij} = \int_0^R \kappa\phi_j'(r)\phi_i'(r)r^2\,dr, \quad i, j = 0, 1, \ldots, n.$$

Every entry in the matrix G is zero except the n, n entry, and similarly only the n-th component of the vector F is nonzero:

$$G_{nn} = \alpha R^2, \quad F_n = \alpha R^2 T_s.$$

This scheme is $O(h^2)$ in the sense that there is a constant $C > 0$ (depending on the true solution u) such that

$$\|u(\cdot, t) - u_h(\cdot, t)\| \leq Ch^2 \quad \text{for all } t \geq 0,$$

[4]The terminology comes from mechanics, the discipline that popularized finite element methods.

where

$$\|u(\cdot,t)-u_h(\cdot,t)\| = \left[\int_0^R (u(r,t)-u_h(r,t))^2 r^2\,dr\right]^{1/2}.$$

To obtain a fully discrete scheme, (39) is discretized in time by the Crank–Nicolson method,

$$M\left(\frac{a^{(k+1)}-a^{(k)}}{\Delta t}\right)+(K+G)\left(\frac{a^{(k+1)}+a^{(k)}}{2}\right)=F,$$

to obtain

$$\left(M+\frac{\Delta t}{2}B\right)a^{(k+1)}=\left(M-\frac{\Delta t}{2}B\right)a^{(k)}+\Delta t F, \tag{40}$$

where $a^{(k)}$ is the approximation to $a(t_k)$, $t_k = k\Delta t$, $k = 0, 1, 2, \ldots$.

We write $u^{(k)}(r)$ for the approximation to $u_h(r, t_k)$ obtained by estimating $a(t_k)$ by $a^{(k)}$. Then there exists constants $C_1, C_2 > 0$ such that, for all k,

$$\|u(\cdot,t_k)-u^{(k)}\| \le C_1 h^2 + C_2 \Delta t^2.$$

This error bound (and the earlier bound on the error in the semidiscrete solution) can be obtain by a straightforward generalization of the standard error analysis found in Thomée [2006].

5. Examples

We will now show several examples, demonstrating the effectiveness of the above analysis. In these examples, we compare the solution (2) of Newton's law of heating with an accurate solution of the heat equation computed by the finite element method described above.

Example 1 (A small iron ball). We first consider an iron ball approximately the size of a baseball: $R = 3.7\,\text{cm}$. The physical constants describing iron are $c = 0.437\,\text{J/g K}$, $\rho = 7.88\,\text{g/cm}^3$, and $\kappa = 0.802\,\text{J/s cm}$. Various references suggest values of α (the *convection heat transfer coefficient* in air) from 10^{-2} to 10^{-3} $\text{W/cm}^2\text{K}$; we will use a value of $\alpha = 0.0045$. The corresponding value of $\overline{\alpha}$ is

$$\overline{\alpha} = \frac{3\alpha}{\rho c R} \approx 0.0010596.$$

We assume that the initial temperature of the ball is $T_0 = 0^\circ\,\text{C}$ and that the temperature of the surrounding air is $T_s = 25^\circ\,\text{C}$, and simulate the temperature in the ball for one hour. The average temperature $\overline{u}$ computed by solving the heat equation and the temperature T predicted by Newton's law of heating are indistinguishable on a graph (see Figure 2); the maximum difference between the two is about 0.038187.

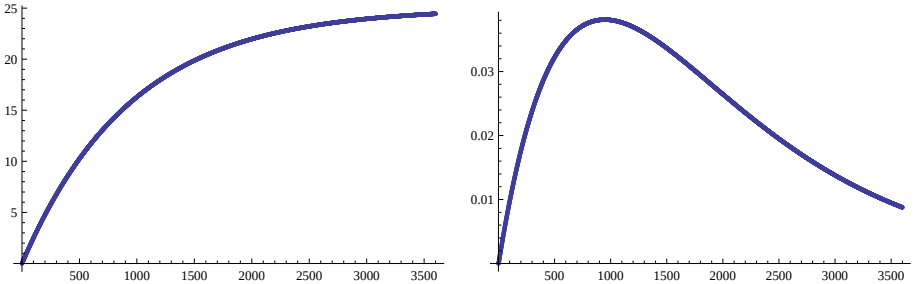


Figure 2. Left: The average temperature of the iron ball in Example 1. Right: The difference $\overline{u}(t) - T(t)$ between the temperatures calculated from the heat equation and Newton's law of heating. In both graphs, the horizontal axis is time in seconds, and the vertical is degrees Celsius.

In this example, we have $\beta = \dfrac{\alpha R}{\kappa} \approx 0.020761$, and the first-order bound on the error is

$$\frac{|T_s - T_0|}{5e}\beta \approx 0.038107.$$

With a small value of β, the analysis suggests that Newton's law is an accurate substitute for the heat equation, and that conclusion is confirmed by the numerical results. Moreover, the first-order bound on the difference between the two solutions is an excellent estimate of the actual difference.

Example 2 (A large iron ball). The second example is the same as the first, except now the radius of the ball is $R = 100\,\text{cm}$. The value of β is now approximately 0.56110, so we do not expect that Newton's law will yield a particularly accurate estimate of the true average temperature. We simulate the temperature for 20 hours (since it takes a long time to appreciably change the temperature in such a large ball).

As Figure 3 shows, the maximum difference between the two solutions is about 0.97333. The first-order bound on the error is

$$\frac{|T_s - T_0|}{5e}\beta \approx 1.0321.$$

Once again, the analysis proves to be quite accurate.

Example 3 (A small styrofoam ball). In the last example, Newton's law of heating, while not a bad approximation, did not accurately model the average temperature in the ball because the ball was so large. In this example, we consider the other reason why Newton's law might not work particularly well, namely, that heat flows slowly through the object compared to how quickly it flows from the surroundings to the object. We consider a styrofoam ball of radius $R = 3.7\,\text{cm}$. The physical

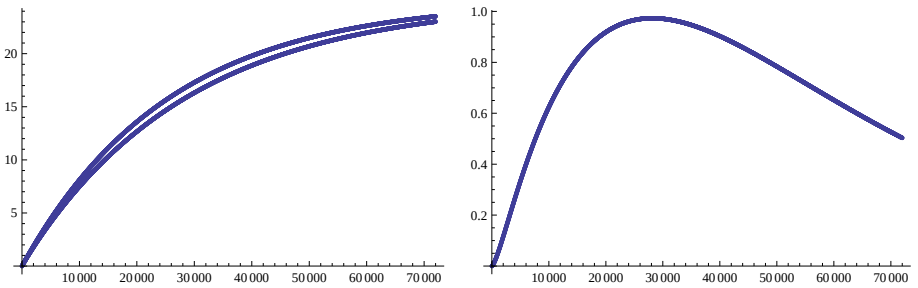


Figure 3. Left: The temperatures of the iron ball in Example 2, as predicted by the heat equation and Newton's law of heating. (The larger temperature is predicted by Newton's law.) Right: The difference $\overline{u}(t) - T(t)$ between the temperatures calculated from the heat equation and Newton's law.

parameters describing styrofoam are $c = 0.209\,\mathrm{J/g\,K}$, $\rho = 0.1\,\mathrm{g/cm^3}$, and $\kappa = 3.3 \cdot 10^{-4}\,\mathrm{J/s\,cm}$. We continue to use $\alpha = 0.0045$, so now $\beta \approx 50.455$. Since $\beta \gg 1$, we expect Newton's law to yield a poor approximation to the true average temperature. This is confirmed in Figure 4, which shows a maximum error of about 14.363. (Since β is so large, we should not expect the first-order error bound to be a good approximation to the actual error, and indeed it is not; the bound is about 92.806.) The value of $\overline{\alpha}$ is

$$\overline{\alpha} \approx 0.17458.$$

A natural question arises in regard to this example: The results show that Newton's law does not produce good results with $\overline{\alpha} = 3\alpha/(\rho c R)$, but what if we use a

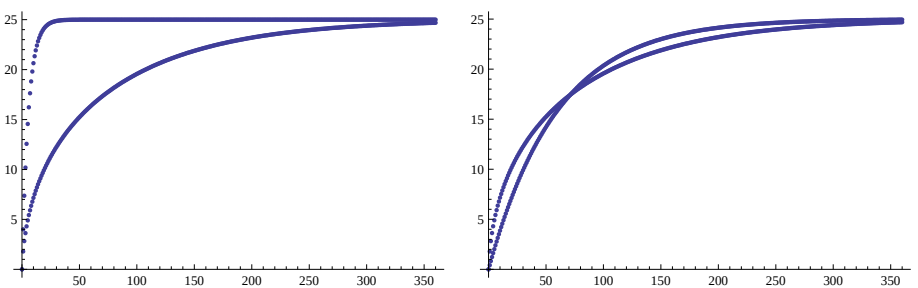


Figure 4. Left: The temperatures of the styrofoam ball in Example 3, as predicted by the heat equation and Newton's law of heating. (The larger temperature is predicted by Newton's law.) Right: The temperatures of the styrofoam ball in Example 3, as predicted by the heat equation and Newton's law with a better value of $\overline{\alpha}$. (The temperature predicted by Newton's law is eventually larger.)

different value of $\overline{\alpha}$? To answer this question, we found the value of $\overline{\alpha}$ that produces a solution (2) as close as possible to $\overline{u}$ in the least-squares sense; that is, we found $\overline{\alpha}$ to minimize

$$J(\overline{\alpha}) = \sum_{k=1}^{N} (\overline{u}(t_k) - T_s + (T_s - T_0)e^{-\overline{\alpha} t_k})^2$$

(where N is the number of time steps in the finite element simulation). We denote the optimal value of $\overline{\alpha}$ by $\tilde{\alpha}$; the result in this example is

$$\tilde{\alpha} \approx 0.016805.$$

With this value of $\overline{\alpha}$, Newton's law yields a much improved estimate of the average temperature of the ball. Nevertheless, the result is still not very good (see Figure 4), which shows that, for this example, the true average temperature in the styrofoam ball is simply not well modeled by Newton's law of heating.

6. Concluding remarks

Our results show that for a small spherical object with the property that heat flows through the object more quickly than it flows to the surroundings ($\beta = \alpha R/\kappa \ll 1$), Newton's law of heating provides a satisfactory model of the average temperature of the object. Moreover, the value

$$\overline{\alpha} = \frac{3\alpha}{\rho c R}$$

is a satisfactory constant of proportionality in Newton's law. To carry out this analysis, we have assumed that the initial temperature of the object is constant throughout, and also that the temperature of the surroundings is held constant.

The alert reader may have noticed that the analysis suggests an even better value of $\overline{\alpha}$. The estimate of λ_1 in (30) suggests that

$$\overline{\alpha} = \frac{3\alpha}{\rho c R}\left(1 - \frac{1}{5}\beta\right)$$

would be an improved estimate of $\lambda_1/(\rho c)$ and hence lead to a better estimate of $\overline{u}$ by T. Indeed, the reader can easily check that this value of $\overline{\alpha}$ leads to an $O(\beta^2)$ bound for $|\overline{u}(t) - T(t)|$, $t \geq 0$.

Many textbook problems on Newton's law of cooling refer to inhomogeneous objects; perhaps the classic example is the cooling of a cup of coffee; see, for example, [Boyce and DiPrima 1992, Section 2.5, problem 14]. (Another popular example is the cooling of a corpse!) Nothing in our analysis allows us to address either inhomogeneities in the object or complex geometries. However, we can apply finite element simulation to an inhomogeneous sphere. For example, we can consider a hollow styrofoam ball filled with water, the closest we can get to

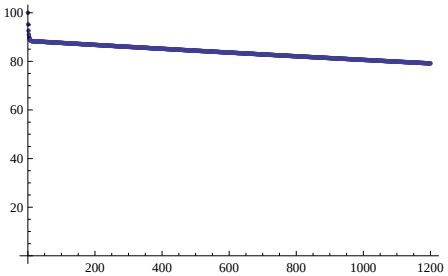


Figure 5. The average temperature in a hollow styrofoam ball filled with water, as computed by finite element simulation.

a coffee cup with our current work. We set the outer radius of the ball to 4.4 cm and the inner radius to 3.9 cm (so that it holds about 250 ml of water), and assume that the initial temperature in both the water and the 'cup' is $T_0 = 100°$ C. Finite element simulation (for 20 minutes) produces the average temperature shown in Figure 5. The results show that the average temperature initially drops quite rapidly, after which it decreases at a more moderate rate. The initial decrease (see the first few seconds in Figure 5) is due to the styrofoam cup's initial loss of heat to the surroundings; since styrofoam has a very small volumetric heat capacity (that is, c measured in J/cm^3 K), a small loss of heat energy translates to a relatively large decrease in temperature in the styrofoam. Once this decrease of average temperature in the styrofoam is complete, the average temperature in the entire ball decreases in a rate well modeled by a function of the form (2) (as the authors have verified), and so Newton's law is a good model after the first few seconds.

A more realistic initial condition would have the temperature of the water at, say, 100° C and the temperature of the styrofoam at room temperature. In this case, the average temperature in the entire ball is less than 100° C and initially increases as the hot water heats the styrofoam. Thereafter, again, Newton's law provides an adequate model.

Acknowledgments

The authors thank the anonymous referees for their careful reading of the paper. Their suggestions improved the final version noticeably.

References

[Arfken and Weber 2005] G. B. Arfken and H. J. Weber, *Mathematical methods for physicists*, 6th ed., Elsevier, New York, 2005. Zbl 1066.00001

[Boyce and DiPrima 1992] W. E. Boyce and R. C. DiPrima, *Elementary differential equations and boundary value problems*, 5th ed., Wiley, New York, 1992. Zbl 0807.34002

[Folland 1995] G. B. Folland, *Introduction to partial differential equations*, 2nd ed., Princeton University Press, Princeton, NJ, 1995. MR 96h:35001 Zbl 0841.35001

[Gockenbach 2002] M. S. Gockenbach, *Partial differential equations: analytical and numerical methods*, Soc. Ind. Appl. Math., Philadelphia, 2002. MR 2003m:35001

[Haberman 2004] R. Haberman, *Applied partial differential equations with fourier series and boundary value problems*, 4th ed., Prentice Hall, Upper Saddle River, NJ, 2004.

[Thomée 2006] V. Thomée, *Galerkin finite element methods for parabolic problems*, 2nd ed., Series in Computational Math. **25**, Springer, Berlin, 2006. MR 2007b:65003 Zbl 1105.65102

[Trantor 1968] C. J. Trantor, *Bessel functions with some physical applications*, Hart, New York, 1968.

[Watson 1944] G. N. Watson, *A treatise on the theory of Bessel functions*, 2nd ed., Cambridge University Press, Cambridge, 1944. Reprinted 1995. MR 96i:33010 Zbl 0849.33001

[Zill 2005] D. G. Zill, *A first course in differential equations*, 8th ed., Brooks/Cole, Belmont, 2005. Zbl 0785.34002

Received: 2008-11-25 Revised: 2009-06-15 Accepted: 2009-07-13

msgocken@mtu.edu *Department of Mathematical Sciences, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931-1295, United States*

kkschmid@mtu.edu *Department of Mathematical Sciences, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931-1295, United States*

Minimum spanning trees

Pallavi Jayawant and Kerry Glavin

(Communicated by Arthur T. Benjamin)

The minimum spanning tree problem originated in the 1920s when O. Borůvka identified and solved the problem during the electrification of Moravia. This graph theory problem and its numerous applications have inspired many others to look for alternate ways of finding a spanning tree of minimum weight in a weighted, connected graph since Borůvka's time. This note presents a variant of Borůvka's algorithm that developed during the graph theory course work of undergraduate students. We discuss the proof of the algorithm, compare it to existing algorithms, and present an implementation of the procedure in Maple.

1. Introduction

Minimum spanning trees (MSTs) have long been of interest to mathematicians because of their many applications. Most commonly, cable and communications companies can represent the task of connecting every house in a network in the least expensive way possible as an MST problem. In this case, the cost of laying cables between houses corresponds to the weights of the edges. There are analogous applications to transportation networks, such as determining the least expensive method of connecting a number of islands or bodies of land. For more applications, see [Wu and Chao 2004; Graham and Hell 1985].

Many algorithms have been developed over the years to find MSTs efficiently. The problem originated in the 1920s when O. Borůvka identified and solved the problem during the electrification of Moravia. However, the language of graph theory is not used to describe the algorithm in his papers from 1926 [Borůvka 1926a; 1926b] which have been translated recently into English [Nešetřil et al. 2001]. In the 1950s, many people contributed to the MST problem. Among them were R. C. Prim and J. B. Kruskal, whose algorithms are very widely used today. The algorithm known as Prim's algorithm was in fact discovered earlier by V. Jarník in 1930. A history of the MST problem appears in [Graham and Hell 1985; Nešetřil et al. 2001; Milková 2007; Nešetřil 1997]. In this note we present a variant

MSC2000: 05C85, 68R10.

Keywords: minimum spanning tree, graph algorithm, graph theory, Maple.

of Borůvka's algorithm and compare it to the algorithms given by Borůvka, Prim and Kruskal which have been central to the history of the problem.

In Section 2, we introduce the graph theory terminology used in this note. We outline the steps of our algorithm in Section 3, provide an illustrative example in Section 4, and prove the algorithm works as intended in Section 5. Section 6 highlights the differences between our algorithm and the work of Borůvka, Prim, and Kruskal. Finally, Section 7 discusses the Maple™ implementation of our algorithm.

2. Terminology

We use the following terminology throughout this note. An *undirected graph* G consists of a set of vertices, denoted by $V(G)$, and a set of unordered pairs of vertices called edges, denoted by $E(G)$. Since we focus on undirected graphs, henceforth we use the word graph to mean an undirected graph. If there exists a path from each vertex u to every other vertex v in G, we call G a *connected graph*. In a *weighted graph*, a real number (usually positive) is assigned to each edge and is called the *weight of the edge*. An example of a connected, weighted graph is provided in Figure 1. The sequence of edges $\{1, 2\}$, $\{2, 4\}$, $\{4, 5\}$ and $\{5, 1\}$ is called a *cycle*. There are many cycles in the graph in Figure 1. The *ends* of the edge $\{1, 2\}$ are the vertices 1 and 2 and its weight is 1, that is, $w(\{1, 2\}) = 1$.

A *minimum spanning tree* (MST) T in a connected, weighted graph G is a connected, acyclic subgraph of G with minimum total weight. To further clarify this definition, we use Figure 2 to explain the various graph theory terms embedded in an MST. The top left diagram shows a tree in G. A *tree* is a connected graph which is acyclic, that is, it has no cycles. A graph with multiple connected components such that each component is a tree is called a *forest*.

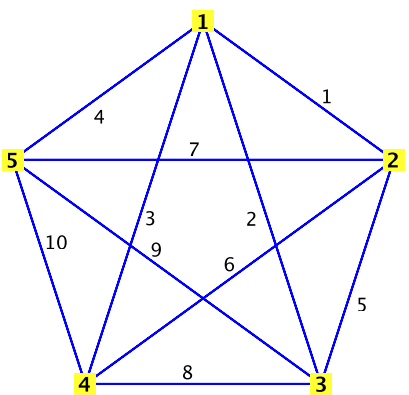


Figure 1. A connected, weighted graph G.

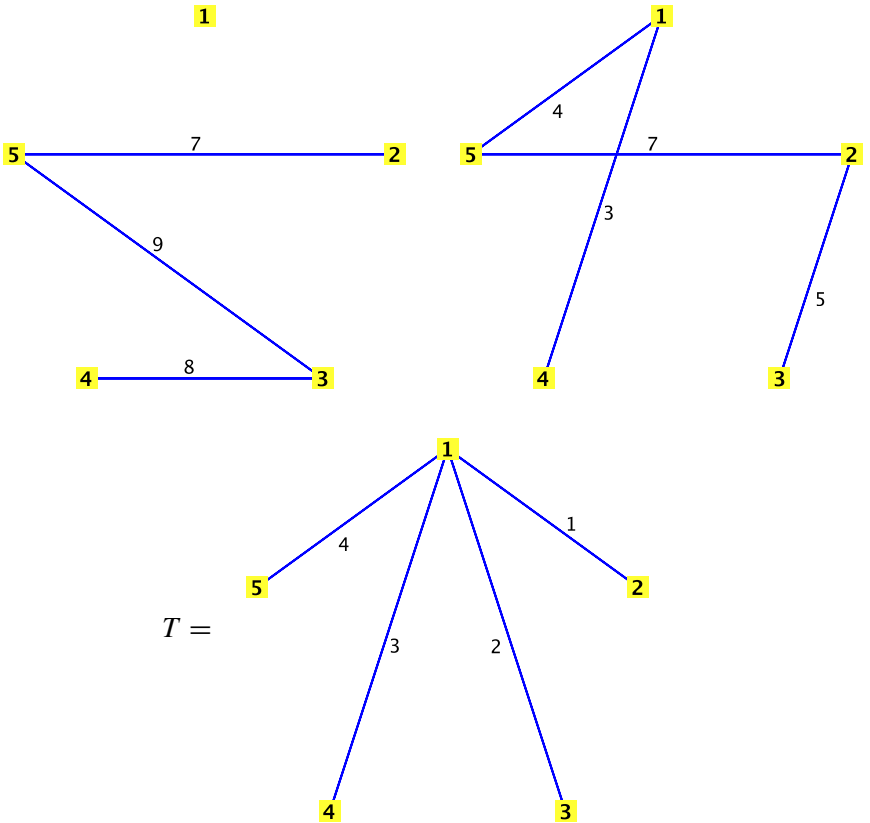


Figure 2. Clockwise from top left: a tree in G; a spanning tree in G; a minimum spanning tree T in G.

The top right diagram represents a spanning tree in G, that is, a tree with vertex set equal to $V(G)$. Finally, the bottom diagram shows a minimum spanning tree in G, which is a spanning tree with minimum total weight. The total weight of a tree is the sum of the weights of all the edges in the tree. The weight of the minimum spanning tree T in Figure 2 is $w(T) = 10$.

3. Steps of the algorithm

First, we establish the input and output for the algorithm, in addition to any necessary notation. The input is a weighted, connected graph G and the output is a spanning tree in G with minimum total weight, which we will call H. We start with H having no vertices and edges. We then construct H by adding vertices and edges as we go through the steps of the algorithm. In this procedure, we assume that there

is an ordering of the vertices in $V(G)$. This is quite standard in a computer algebra system such as Maple. We designate the number of vertices in G as n.

3.1. ***Identify incident edge with smallest weight.*** For each vertex v_i for i from 1 to n, identify the edge incident to v_i with the smallest weight. In the case of multiple edges with the same weight, identify only one of these edges. If the ends of this edge are not already in H, then add the edge to H. Otherwise, do not make any changes to H. This ensures that H does not contain any cycles.

At the end of this step, H may contain one or more connected components. If there is only one connected component, then the procedure is finished. If the number of connected components of H is greater than one, then the procedure continues in the next step.

3.2. ***If necessary, create one connected component.*** If H consists of more than one connected component, then evaluate the weights of all edges connecting the distinct components of H. Add the edge with the smallest weight to H. In the case of multiple edges with the same weight, add only one of these edges.

Repeat this step until H has just one connected component.

4. An illustrative example

To demonstrate the two separate steps involved in this algorithm, we will build a minimum spanning tree H in the weighted graph shown in Figure 3. This is a complete bipartite graph, denoted by $K_{3,3}$.

The first step of the algorithm calls for an evaluation of the weights of the edges incident to each vertex. Starting with vertex 1, we evaluate the weights of the edges $\{1, 4\}$, $\{1, 5\}$, and $\{1, 6\}$. We note that $w(\{1, 6\}) = 6$ and this is the edge of smallest weight at vertex 1. Since neither vertex 1 nor vertex 6 is already part of H, the

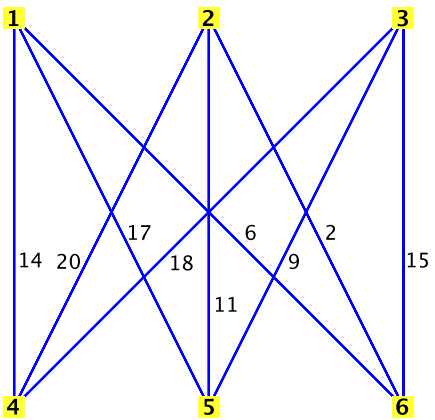


Figure 3. A weighted complete bipartite graph $K_{3,3}$.

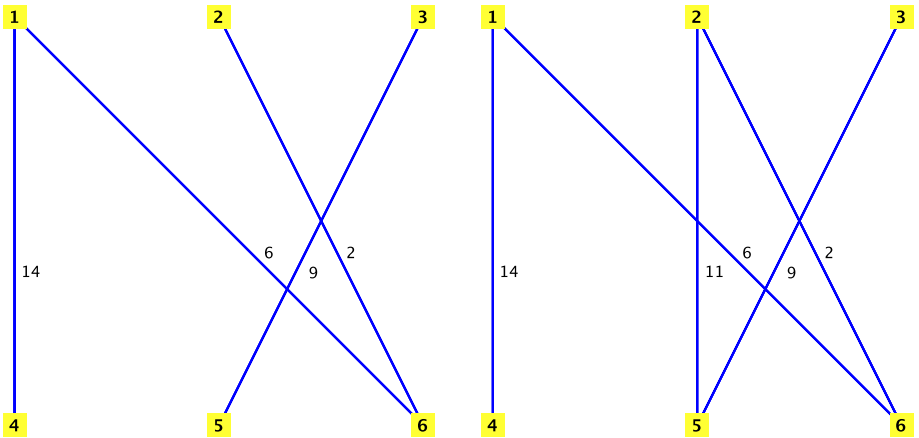


Figure 4. The generation of a minimum spanning tree H in a weighted $K_{3,3}$. We begin with the smallest edge incident to vertex 1, namely edge $\{1, 6\}$ in Figure 3. At the end of the first step of the algorithm, H is the graph shown on the left. At completion, H is the minimum spanning tree shown on the right.

edge $\{1, 6\}$ is the first addition to H. Even though it is obvious that the ends of the first edge added are not already part of the MST, this check is extremely important in later iterations to guarantee that H does not contain any cycles.

Once the algorithm runs through the entire first step, we have two distinct components to H. Figure 4, left, shows these components. The second step of the algorithm is necessary in this case to achieve a connected graph.

The second step begins with an evaluation of the weights of all edges that connect these two distinct components. The algorithm evaluates the weights of edges $\{3, 4\}$, $\{3, 6\}$, $\{2, 5\}$, and $\{1, 5\}$ and finds that edge $\{2, 5\}$ has the smallest weight. Thus this edge is added to H. We know that the addition of this edge will not create any cycles by the nature of distinct components. Now that H contains just one connected component with no cycles, we have our final desired result, as shown in Figure 4, right. The weight of the MST is $w(H) = 42$.

5. Proof of the algorithm

The proof of our algorithm uses similar techniques as the existing proofs for other MST algorithms, including the work of both Prim and Kruskal [Wilson and Watkins 1990; Rosen 2007]. In order to prove that the final output of the algorithm, H, is a minimum spanning tree, it is necessary to prove two separate properties: H is a spanning tree of the weighted, connected graph G; and H is of minimum weight.

5.1. ***H is a spanning tree of G.*** First, we show that H is a tree, that is, H is both connected and acyclic. The second step of the algorithm (3.2) guarantees that edges will be added to H until it has only one connected component. There are two different parts of the algorithm to consider when determining if H contains a cycle. In the first step (3.1), we do not add any edges whose ends are already in H. This prevents the creation of any cycles. In the second step (3.2), we only add edges connecting distinct components. There is no way to create a cycle in H by adding an edge that connects distinct components.

Second, we show that H spans G, that is, we show $V(H) = V(G)$. The first step of the algorithm, the loop for each vertex v_i for i from 1 to n, ensures that H contains all vertices of G.

5.2. ***H is of minimum weight.*** Suppose M is a minimum spanning tree in G. We know $w(M) \leq w(H)$ and M and H are both subgraphs of G. Our goal is to transform M into H in a way that shows $w(H) \leq w(M)$. This implies $w(H) = w(M)$ and proves that H is of minimum weight.

A tree on n vertices has $n-1$ edges. We name the edges in H according to the order in which they were added by the algorithm: $e_1, e_2, \ldots, e_{n-1}$. Assume $e_1, e_2, \ldots, e_{k-1}$ are all in M as well. So $e_k = \{u, v\}$ is the first edge in H that we find is not also in M. We want to add e_k to M and delete an edge from it such that the resulting subgraph L is a spanning tree of G and $w(L) \leq w(M)$. We know that there must be a path between u and v in M since the MST is both connected and spanning. Thus the addition of $e_k = \{u, v\}$ to M creates a cycle C in M. Let e be the other edge incident to u in C.

To select the edge to delete so that we obtain L, we now consider two cases: either e_k was added to H during the first step (3.1) or e_k was added during the second step (3.2).

If e_k was added to H during the first step, it must be true that e_k is the edge of smallest weight at one of its endpoints. Without loss of generality, assume that e_k is the edge of smallest weight incident to u. We then delete the edge e from M to obtain L. We know that $w(e) \geq w(e_k)$ because e_k is an edge incident to u with the smallest weight and hence $w(L) \leq w(M)$.

If e_k was added to H during the second step, let K be the subgraph of H to which e_k was added. Then we know that u and v must have been in distinct components of K. Note that K is also a subgraph of M because all the edges added to H before e_k are in M as well. If we start traversing the cycle C along the edge e, then we must reach an edge f such that its ends are in distinct components of K and we delete f from M to obtain L. Again, $w(f) \geq w(e_k)$ because the algorithm adds an edge of smallest weight that connects distinct components of K, and hence $w(L) \leq w(M)$.

Since M is of minimum weight, $w(L) = w(M)$ and thus L is a minimum spanning tree of G. We repeat the process of adding and deleting an edge with M replaced by L. We continue in this way until we get H.

6. Comparison with other algorithms

We now compare our algorithm to the three algorithms by Prim, Kruskal and Borůvka "that have played a central role in the history of the MST problem" as stated in [Milková 2007].

Prim's algorithm [1957] generates a minimum spanning tree by identifying an edge with minimum weight incident to the initial vertex and spreading the tree from this edge. At every iterative step, the algorithm finds and adds the edge of smallest weight such that one vertex is already part of the MST and the other vertex is not. The procedure is complete once the vertex set of the tree is equal to the vertex set of the original graph, that is, the tree spans all of the vertices. Prim's algorithm allows for just one tree at any given step. By comparison, there may be multiple trees, or a forest, that are ultimately connected in our algorithm.

We use graphs created in Maple to highlight the differences between our algorithm and Prim's algorithm. Figure 5 shows MSTs in a complete graph on six vertices, denoted by K_6, with each edge weighted 5. The left diagram shows the MST generated by the `spantree` procedure in Maple, which uses Prim's algorithm. The Maple implementation of our algorithm created the MST shown in the diagram on the right.

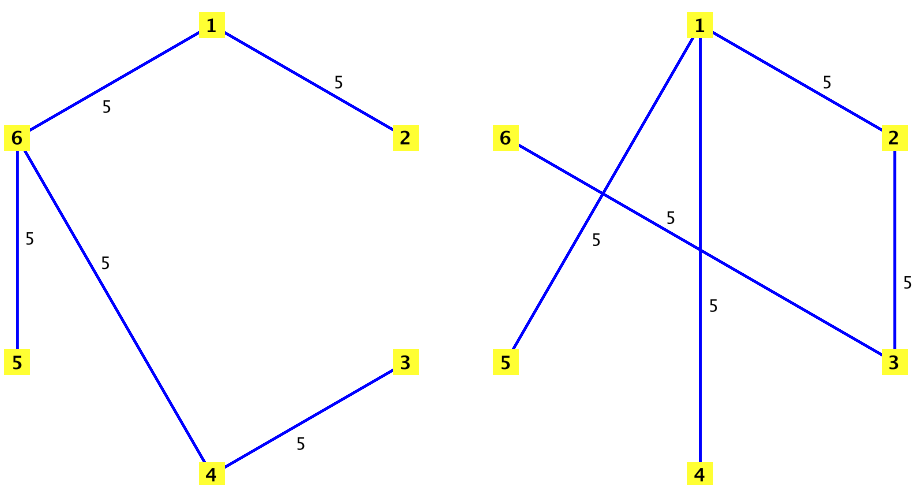


Figure 5. MSTs in K_6 with each edge weighted 5. The one on the left was created by Prim's algorithm, the one on the right by our algorithm.

Kruskal's algorithm [Kruskal 1956] is a "greedy" algorithm that constructs a minimum spanning tree by adding edges with minimum weight as long as doing so does not form a cycle. For detailed steps of the algorithm, see [Wilson and Watkins 1990]. Thus, Kruskal's algorithm chooses edges of smallest weight from the entire graph at every iteration whereas our algorithm identifies an edge with minimum weight at each vertex and later between two connected components.

Borůvka's algorithm [Borůvka 1926a; 1926b; Nešetřil et al. 2001] is a recursive algorithm which at every recursive step repeats the first step of the algorithm on a new graph formed by contraction. The first step of Borůvka's algorithm and the first step of our algorithm are identical. Thus, at the end of the first step, there is a set of chosen edges that may not form a single connected component. Borůvka's algorithm then forms a new graph by contraction as follows. Each connected component is replaced by a single vertex. All edges connecting vertices in the same connected component are eliminated. All edges between two distinct components are eliminated except for the edge with the smallest weight. If edges do not have distinct weights, a tie-breaking procedure is used to retain only one edge between two distinct components. Borůvka's algorithm then repeats the first step on this newly formed graph. The recursion continues until only one vertex remains in the contracted graph; that is, until the chosen edges form only one connected component. The set of all the edges chosen each time the first step is executed constitutes an MST. The contraction to form the new graph and the subsequent recursion of Borůvka's algorithm are replaced in our algorithm with the iterative process of joining the connected components obtained at the end of the first step with edges of minimum weight.

Thus of the three algorithms mentioned here, our algorithm is most similar to Borůvka's algorithm. The output of our algorithm may differ from the output of Borůvka's algorithm because both depend on the particular tie-breaking procedures used when all edges do not have distinct weights.

7. Maple implementation of algorithm

Our implementation of the algorithm takes advantage of the `networks` package in Maple. This package contains many commands useful in graph theory, a number of which we will discuss later in this section. The Maple implementation of the first step of the algorithm (3.1) is shown on the next page

The implementation employs a number of useful commands. In general, the `networks` package makes it easy to work with both vertices and edges in a graph. The `incident` command returns a set of the edges incident to a vertex v_i. The `eweight` command gives the weights of the edges in a graph. Both of these commands are important when we create the list of the weights of all edges incident to

vertex v_i. The `nops` command counts the number of elements in its argument, and the `member` command tests if an element belongs to a set or list. While `nops` and `member` are not specific to the `networks` package, both of these commands play a large role in the procedure as well.

```
pathset:={}; # This is where we will keep track of edges in our MST.
for i from 1 to nops(vertices(G)) do
  listofweights:=[];
  currentedge:={};
  for j in ends(incident(i,G),G) do
    listofweights:=[op(listofweights),eweight(edges({j[1],j[2]},G)[1],G)];
  end do;
  # In this loop, we create a list of the weights (called "listofweights")
  # of all edges incident to vertex v_i.
  smallest:=listofweights[1];
     for k from 2 to nops(listofweights) do
         if (listofweights[k]<smallest) then smallest:=listofweights[k]
         end if;
     end do;
     # Here, we identify the smallest value in "listofweights".
  for x in ends(incident(i,G),G) do
    if eweight(edges({x[1],x[2]},G)[1],G)=smallest then
       currentedge:=currentedge union {x};
    end if;
  end do;
  # We match the smallest value to the edge(s) with this weight and add it
  # to a set called "currentedge".
  found:=0;
  foundvertex:=0;
  for j in pathset do
      if (member(currentedge[1][1],j) and currentedge[1][1]<>foundvertex)
       then found:=found+1:foundvertex:=currentedge[1][1]
      end if;
      if found=2 then break end if;
      if (member(currentedge[1][2],j) and currentedge[1][2]<>foundvertex)
       then found:=found+1:foundvertex:=currentedge[1][2]
      end if;
      if found=2 then break end if;
  end do;
  if found=2 then pathset:=pathset
      else pathset:=pathset union {currentedge[1]}
  end if;
  # If the ends of the edge incident to v_i with the smallest weight
  # are already part of pathset, then pathset remains the same.
  # Otherwise, we add just one edge to pathset.
end do;
```

Maple implementation of the first step (3.1) of our algorithm (# introduces a comment line). The input is a weighted, connected graph G.

Following the first step of the procedure, we begin to build H by inserting vertices 1 to n and the edges from `pathset` into H. If H has more than one component, the second part of the procedure starts by identifying all edges in G that connect distinct components in H, as shown below. These edges are then added to a set called `connectingedges`. The `components` command, which identifies the components of a graph as a set of sets, is especially valuable in this part of the procedure.

```
possibleedges:=ends(G) minus pathset;
connectingedges:={};
for r in possibleedges do
  d:=nops(components(H));
  addedge(r,H);
    if nops(components(H))<d
      then connectingedges:=connectingedges union {r}
    end if;
  delete(edges({r[1],r[2]},H),H);
end do;
```

The implementation of the rest of the second step (3.2) is similar to the procedure for the first step. After establishing the set `connectingedges`, we evaluate the weights of the edges, identify the smallest value, and add the associated edge to H. This step is repeated until H contains just one connected component.

Now we take a look at the complexity of our implementation. Let n be the number of vertices in G and m the number of edges in G. If we assume that each of the Maple command runs in unit time, then our implementation runs in time $O(mn)$. This could be improved with a more efficient sorting procedure in each step. For a discussion of the complexity of MST algorithms and recent work on MSTs see [Wu and Chao 2004; Graham and Hell 1985; Cheriton and Tarjan 1976].

8. Conclusion

Due to the numerous applications of minimum spanning trees to communications and transportation networks, it is important to have efficient algorithms to find minimum spanning trees in weighted, connected graphs. Borůvka, Jarník, Prim, and Kruskal, among others, have made important contributions to this area of graph theory. We have presented an algorithm that is a variant of the original solution by Borůvka and unlike the proof by Borůvka, we have provided a proof of the algorithm using the language of modern graph theory. The running time of the implementation could be improved and we hope the reader will try to do so.

Acknowledgments

The Graph Algorithms course is an elective course offered by the Department of Mathematics at Bates College and it aims to teach students the basics of graph

theory (with an emphasis on algorithms) and computer programming skills. Two thirds of class time is devoted to learning graph theory and one third is spent learning basic programming skills through the use of the `networks` package in the computer algebra system Maple. In the class of Fall 2007, after learning the definitions and basic properties of trees and MSTs, we discussed briefly the algorithms by Prim and Kruskal in class. The assignment for the next class was to write the pseudocode for an algorithm (not necessarily one of the algorithms we had discussed in class) to find an MST in a weighted, connected graph. The next class began with one of the students (the second author) presenting the first step of an algorithm to the whole class. As a class, we then completed the algorithm to produce the desired result. The second author and another student implemented the algorithm in Maple as their final exam project, with help from the instructor (the first author).

The second author was supported by a Hoffman Research Support Grant, a Bates College program funded by an endowment established by the Hoffman Foundation, during our work on this note in Summer 2008 when we wrote the proof of the algorithm. We thank Emmanuel Drabo, Philip Greengard, Alex Jorge, Binit Malla, Dylan Mogk, Razin Mustafiz, Duane Pelz and Dan Perry — the Graph Algorithms class of Fall 2007 for the class discussion to complete the second step of the algorithm. We are particularly grateful to Dan Perry for teaming up with the second author to implement the algorithm in Maple. We would also like to thank Eric Towne of the Department of Mathematics at Bates College for conducting the labs through which the students learned to program in Maple and for helpful comments on this note.

References

[Borůvka 1926a] O. Borůvka, "O jistém problému minimálním", *Práce Mor. Přírodověd., Spol. v Brně* **3** (1926), 37–58.

[Borůvka 1926b] O. Borůvka, "Příspěvek k řešení otázky ekonomické stavby elektrovodních sítí", *Elektrotechnický obzor* **15** (1926), 153–154.

[Cheriton and Tarjan 1976] D. Cheriton and R. E. Tarjan, "Finding minimum spanning trees", *SIAM J. Comput.* **5**:4 (1976), 724–742. MR 56 #4783 Zbl 0358.90069

[Graham and Hell 1985] R. L. Graham and P. Hell, "On the history of the minimum spanning tree problem", *Ann. Hist. Comput.* **7**:1 (1985), 43–57. MR 86g:68005 Zbl 0998.68003

[Kruskal 1956] J. B. Kruskal, Jr., "On the shortest spanning subtree of a graph and the traveling salesman problem", *Proc. Amer. Math. Soc.* **7** (1956), 48–50. MR 17,1231d Zbl 0070.18404

[Milková 2007] E. Milková, "The minimum spanning tree problem: Jarník's solution in historical and present context", pp. 309–316 in *6th Czech-Slovak International Symposium on Combinatorics, Graph Theory, Algorithms and Applications*, edited by P. Hliněný et al., Electron. Notes Discrete Math. **28**, Elsevier, Amsterdam, 2007. MR 2324010

[Nešetřil 1997] J. Nešetřil, "A few remarks on the history of MST-problem", *Arch. Math.* (*Brno*) **33**:1-2 (1997), 15–22. MR 98g:01065 Zbl 0909.05022

[Nešetřil et al. 2001] J. Nešetřil, E. Milková, and H. Nešetřilová, "Otakar Borůvka on minimum spanning tree problem: translation of both the 1926 papers, comments, history", *Discrete Math.* **233**:1-3 (2001), 3–36. MR 2002f:05053

[Prim 1957] R. C. Prim, "The shortest connecting network and some generalizations", *Bell System Tech. J.* **36** (1957), 1389–1401.

[Rosen 2007] K. H. Rosen, *Discrete mathematics and its applications*, McGraw-Hill, New York, 2007. Zbl 0691.05001

[Wilson and Watkins 1990] R. J. Wilson and J. J. Watkins, *Graphs: An introductory approach*, Wiley, New York, 1990. MR 91b:05001 Zbl 0712.05001

[Wu and Chao 2004] B. Y. Wu and K.-M. Chao, *Spanning trees and optimization problems*, Discrete Mathematics and its Applications, Chapman & Hall/CRC, Boca Raton, FL, 2004. MR 2004i:90008 Zbl 1072.90047

Received: 2009-03-01 Accepted: 2009-05-02

pjayawan@bates.edu *Department of Mathematics, Bates College, Lewiston, ME 04240, United States*

kerry.glavin@gmail.com *Department of Mathematics, Bates College, Lewiston, ME 04240, United States*

INVOLVE 2:4(2009)

Geometric properties of Shapiro–Rudin polynomials

John J. Benedetto and Jesse D. Sugar Moore

(Communicated by David Larson)

The Shapiro–Rudin polynomials are well traveled, and their relation to Golay complementary pairs is well known. Because of the importance of Golay pairs in recent applications, we spell out, in some detail, properties of Shapiro–Rudin polynomials and Golay complementary pairs. However, the theme of this paper is an apparently new elementary geometric observation concerning cusp-like behavior of certain Shapiro–Rudin polynomials.

1. Introduction

We begin by defining Shapiro–Rudin polynomials [Shapiro 1951; Rudin 1959] (see also [Tseng and Liu 1972]). $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}$, and $\mathbb{C}$ are the sets of natural numbers, integers, real numbers, and complex numbers, respectively.

Definition 1.1. The *Shapiro–Rudin polynomials*, P_n, Q_n, $n = 0, 1, 2, \ldots$, are defined recursively as follows. For $t \in \mathbb{R}/\mathbb{Z}$, we set $P_0(t) = Q_0(t) = 1$ and

$$P_{n+1}(t) = P_n(t) + e^{2\pi i 2^n t} Q_n(t), \quad Q_{n+1}(t) = P_n(t) - e^{2\pi i 2^n t} Q_n(t). \tag{1-1}$$

The number of terms in the n-th polynomial, P_n or Q_n, is 2^n. Thus, the sequence of coefficients of each polynomial, P_n or Q_n, is a sequence of length 2^n consisting of ± 1s.

Definition 1.2. For any sequence $z = \{z_k\}_{k=0}^{n-1} \subseteq \mathbb{C}$ and for any $m \in \{0, 1, \ldots, n-1\}$, the m-th *aperiodic autocorrelation coefficient*, $A_z(m)$, is defined as

$$A_z(m) = \sum_{j=0}^{n-1-m} z_j \overline{z_{m+j}}. \tag{1-2}$$

We now define a *Golay complementary pair* of sequences. The concept was introduced by Golay [1951; 1961; 1962], but a significant precursor is found in [Golay 1949].

MSC2000: 42A05.

Keywords: Shapiro–Rudin polynomials, Golay pairs, cusp properties.

Definition 1.3. Two sequences, $p = \{p_k\}_{k=0}^{n-1} \subseteq \mathbb{C}$ and $q = \{q_k\}_{k=0}^{n-1} \subseteq \mathbb{C}$, are a *Golay complementary pair* if $A_p(0) + A_q(0) \neq 0$ and

$$A_p(m) + A_q(m) = 0 \quad \text{for all } m = 1, 2, \ldots, n-1. \tag{1-3}$$

It is well known that the Shapiro–Rudin coefficients are Golay pairs; see Proposition 2.1. Further, Welti codes [1960] are intimately related to Golay pairs and Shapiro–Rudin polynomials. In Section 3, we begin with a useful formula for the Shapiro–Rudin polynomials, then record MATLAB code for their evaluation. Page 459 is devoted to graphs of Shapiro–Rudin polynomials; these graphs served as the basis for our geometrical observations about cusps, quantified in Section 4. In fact, in Theorem 4.8, we shall prove that the graph or trajectory of P_{2n} in $\mathbb{C}$, as a function of $t \in \mathbb{R}$, has a quadratic cusp at $t = 2\pi j$, $j \in \mathbb{Z}$. Clearly, P_{2n} is 1-periodic and infinitely differentiable as a function of $t \in \mathbb{R}$.

Remark 1.4. (a) Shapiro–Rudin polynomials have the Pythagorean and quadrature mirror filter (QMF or CMF) property:

$$|P_n(t)|^2 + |Q_n(t)|^2 = 2^{n+1} \quad \text{for all } n \geq 0 \text{ and } t \in \mathbb{R}$$

(see [Vaidyanathan 1993; Daubechies 1992; Mallat 1998]), as well as the sup-norm bound or "flatness" property,

$$\|P_n\|_{C(\mathbb{R}/\mathbb{Z})} \leq 2^{(n+1)/2} \quad \text{and} \quad \|Q_n\|_{C(\mathbb{R}/\mathbb{Z})} \leq 2^{(n+1)/2}, \tag{1-4}$$

where $\|f\|_{C(\mathbb{R}/\mathbb{Z})} = \sup_{t\in\mathbb{R}} |f(t)|$, for continuous and 1-periodic functions $f : \mathbb{R} \to \mathbb{C}$. Note that the $L^2(\mathbb{R}/\mathbb{Z})$ norms of the Shapiro–Rudin polynomials are

$$\|P_n\|_{L^2(\mathbb{R}/\mathbb{Z})} = \left(\int_0^1 |P_n(t)|^2 dt\right)^{1/2} = 2^{n/2} \quad \text{and} \quad \|Q_n\|_{L^2(\mathbb{R}/\mathbb{Z})} = 2^{n/2}.$$

The sup-norm estimates have deep analytic implications in bounding the pseudomeasure norms of important measures arising in the study of restriction algebras of the Fourier algebra of absolutely convergent Fourier series (see, for example, [Kahane 1970]). Benke's analysis and generalization of Shapiro–Rudin polynomials [Benke 1994] provide an understanding of the importance of unitarity in obtaining the low sup-norm bound in (1-4) vis a vis the exponential growth, 2^n, of P_n and Q_n. This issue is central in the Littlewood flatness problem and associated applications dealing with crest factors, $\|f\|_{C(\mathbb{R}/\mathbb{Z})}/\|f\|_{L^2(\mathbb{R}/\mathbb{Z})}$ (see, for example, [Benedetto 1997, page 238]).

(b) In classical Fourier series, Shapiro–Rudin polynomials can be used to construct continuous and 1-periodic functions $f : \mathbb{R} \to \mathbb{C}$ which are of Lipschitz order 1/2, but which do not have an absolutely convergent Fourier series [Katznelson 1976, pages 33-34].

(c) There is a large literature, several research areas, and a plethora of fiendish unresolved problems associated with Shapiro–Rudin polynomials, Golay complementary pairs, and Welti codes. For a sampling of the literature, besides [Benke 1994], we mention [Brillhart and Carlitz 1970; Brillhart 1973; Saffari 1986; 1987; Eliahou et al. 1990; 1991; Brillhart and Morton 1996; Saffari 2001; Jedwab 2005; Jedwab and Yoshida 2006]. This is truly the tip of the iceberg, even for the one-dimensional case, and the references in these articles give a hint of the breadth of the area.

(d) Besides applications to coding theory and to antenna theory, reflected by the analysis of crest factors mentioned above, Golay complementary pairs are now being used in radar waveform design [Levanon and Mozeson 2004; Howard et al. 2006; Searle and Howard 2007; Pezeshki et al. 2008], perhaps inspired by [Lüke 1985; Budišin 1990], and certainly going back to [Welti 1960].

2. Shapiro–Rudin polynomials and Golay complementary pairs

Let $\hat{P}_n = \{\hat{P}_n(k)\}_{k=0}^{2^n-1}$ denote the sequence of ± 1 coefficients of P_n, and let $\hat{Q}_n = \{\hat{Q}_n(k)\}_{k=0}^{2^n-1}$ denote the sequence of ± 1 coefficients of Q_n. Note that $k = 0$ corresponds to the first coefficient, $k = 1$ to the second, and so on.

As a result of the recursive construction of the Shapiro–Rudin polynomials, the coefficients of the $(n+1)$-st polynomials can be given in terms of the coefficients of the n-th polynomials:

$$\begin{aligned} \{\hat{P}_{n+1}(k)\}_{k=0}^{2^{n+1}-1} &= \big\{\{\hat{P}_n(k)\}_{k=0}^{2^n-1}, \{\hat{Q}_n(k)\}_{k=0}^{2^n-1}\big\}, \\ \{\hat{Q}_{n+1}(k)\}_{k=0}^{2^{n+1}-1} &= \big\{\{\hat{P}_n(k)\}_{k=0}^{2^n-1}, -\{\hat{Q}_n(k)\}_{k=0}^{2^n-1}\big\}. \end{aligned} \tag{2-1}$$

For example, we have

$$\begin{aligned} \{\hat{P}_1(k)\}_{k=0}^{1} &= \big\{\{\hat{P}_0\}, \{\hat{Q}_0\}\big\} = \{1, -1\}, \\ \{\hat{Q}_1(k)\}_{k=0}^{1} &= \big\{\{\hat{P}_0\}, -\{\hat{Q}_0\}\big\} = \{1, -1\}, \\ \{\hat{P}_2(k)\}_{k=0}^{3} &= \big\{\{\hat{P}_1(k)\}_{k=0}^{1}, \{\hat{Q}_1(k)\}_{k=0}^{1}\big\} = \{1, 1, 1, -1\}, \\ \{\hat{Q}_2(k)\}_{k=0}^{3} &= \big\{\{\hat{P}_1(k)\}_{k=0}^{1}, -\{\hat{Q}_1(k)\}_{k=0}^{1}\big\} = \{1, 1, -1, 1\}, \\ \{\hat{P}_3(k)\}_{k=0}^{7} &= \big\{\{\hat{P}_2(k)\}_{k=0}^{3}, \{\hat{Q}_2(k)\}_{k=0}^{3}\big\} = \{1, 1, 1, -1, 1, 1, -1, 1\}, \\ \{\hat{Q}_3(k)\}_{k=0}^{7} &= \big\{\{\hat{P}_2(k)\}_{k=0}^{3}, -\{\hat{Q}_2(k)\}_{k=0}^{3}\big\} = \{1, 1, 1, -1, -1, -1, 1, -1\}. \end{aligned}$$

This recursive method of constructing sequences is the *append rule* [Benke 1994]. The following result is well known.

Proposition 2.1. *For each $n \in \mathbb{N}$, the sequences $\hat{P}_n = \{\hat{P}_n(k)\}_{k=0}^{2^n-1}$ and $\hat{Q}_n = \{\hat{Q}_n(k)\}_{k=0}^{2^n-1}$ are a Golay complementary pair, i.e., $A_{\hat{P}_n}(0) + A_{\hat{Q}_n}(0) = 2^{n+1}$ and*

$$A_{\hat{P}_n}(m) + A_{\hat{Q}_n}(m) = 0 \quad \text{for all } m = 1, 2, \cdots, 2^n - 1. \tag{2-2}$$

Proof. Since $\{\hat{P}_n(k)\}_{k=0}^{2^n-1}, \{\hat{Q}_n(k)\}_{k=0}^{2^n-1} \subseteq \mathbb{R}$, complex conjugation is ignored in the summands $A_{\hat{P}_n}(m)$ and $A_{\hat{Q}_n}(m)$.

Let $n \in \mathbb{N}$. If $m = 0$, then

$$\begin{aligned} A_{\hat{P}_n}(0) + A_{\hat{Q}_n}(0) &= \sum_{j=0}^{2^n-1} \hat{P}_n(j)\hat{P}_n(j) + \sum_{j=0}^{2^n-1} \hat{Q}_n(j)\hat{Q}_n(j) \\ &= \sum_{j=0}^{2^n-1} \big(\hat{P}_n(j)^2 + \hat{Q}_n(j)^2\big) = \sum_{j=0}^{2^n-1} 2 = 2^{n+1}. \end{aligned}$$

For $m \neq 0$, we shall use induction. Two separate cases arise when proving the inductive step. In the first case, we consider m such that $1 \leq m \leq 2^n - 1$, and, in the second case, we consider m such that $2^n \leq m \leq 2^{n+1} - 1$. In both cases, we shall use the fact that, for any $n \in \mathbb{N}$, $\hat{P}_n(j) = \hat{Q}_n(j)$ for $j = 0, 1, \ldots, 2^{n-1} - 1$ and $\hat{P}_n(j) = -\hat{Q}_n(j)$ for $j = 2^{n-1}, \ldots, 2^n - 1$.

For $n = 1$, the only nonzero value m takes is $m = 1$. Consequently,

$$A_{\hat{P}_1}(1) + A_{\hat{Q}_1}(1) = \sum_{j=0}^{0} \hat{P}_1(0)\hat{P}_1(1) + \sum_{j=0}^{0} \hat{Q}_1(0)\hat{Q}_1(1) = 1 + (-1) = 0.$$

We now assume that (2-2) is true for some $n \in \mathbb{N}$ and for each m such that $1 \leq m \leq 2^n - 1$, and we consider the $n+1$ case.

Case 1. If $1 \leq m \leq 2^n - 1$, then

$$\begin{aligned} A_{\hat{P}_{n+1}}(m) + A_{\hat{Q}_{n+1}}(m) &= \sum_{j=0}^{2^{n+1}-1-m} \big(\hat{P}_{n+1}(j)\hat{P}_{n+1}(m+j) + \hat{Q}_{n+1}(j)\hat{Q}_{n+1}(m+j)\big) \\ &= \sum_{j=0}^{2^n-1-m} \big(\hat{P}_{n+1}(j)\hat{P}_{n+1}(m+j) + \hat{Q}_{n+1}(j)\hat{Q}_{n+1}(m+j)\big) \\ &\quad + \sum_{j=2^n-m}^{2^n-1} \big(\hat{P}_{n+1}(j)\hat{P}_{n+1}(m+j) + \hat{Q}_{n+1}(j)\hat{Q}_{n+1}(m+j)\big) \\ &\quad + \sum_{j=2^n}^{2^{n+1}-1-m} \big(\hat{P}_{n+1}(j)\hat{P}_{n+1}(m+j) + \hat{Q}_{n+1}(j)\hat{Q}_{n+1}(m+j)\big) \\ &= \sum_{j=0}^{2^n-1-m} \big(\hat{P}_n(j)\hat{P}_n(m+j) + \hat{P}_n(j)\hat{P}_n(m+j)\big) \\ &\quad + \sum_{j=2^n-m}^{2^n-1} \big(\hat{Q}_{n+1}(j)\hat{P}_{n+1}(m+j) + \hat{Q}_{n+1}(j)\big(-\hat{P}_{n+1}(m+j)\big)\big) \\ &\quad + \sum_{j=2^n}^{2^{n+1}-1-m} \big(\hat{P}_{n+1}(j)\hat{P}_{n+1}(m+j) + \big(-\hat{P}_{n+1}(j)\big)\big(-\hat{P}_{n+1}(m+j)\big)\big) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=0}^{2^n-1-m} 2\big(\hat{P}_n(j)\hat{P}_n(m+j)\big) + 0 + \sum_{j=0}^{2^n-1-m} 2\big(\hat{Q}_n(j)\hat{Q}_n(m+j)\big) \\
&= 2\sum_{j=0}^{2^n-1-m} \big(\hat{P}_n(j)\hat{P}_n(m+j) + \hat{Q}_n(j)\hat{Q}_n(m+j)\big) \\
&= 2\big(A_{\hat{P}_n}(m) + A_{\hat{Q}_n}(m)\big).
\end{aligned}
$$

Since $2\big(A_{\hat{P}_n}(m) + A_{\hat{Q}_n}(m)\big) = 0$ for all m such that $1 \le m \le 2^n - 1$ by the inductive hypothesis, we have that $A_{\hat{P}_{n+1}}(m) + A_{\hat{Q}_{n+1}}(m) = 0$ for all m such that $1 \le m \le 2^n - 1$.

Case 2. If $2^n \le m \le 2^{n+1} - 1$, then

$$
\begin{aligned}
&A_{\hat{P}_{n+1}}(m) + A_{\hat{Q}_{n+1}}(m) \\
&\qquad = \sum_{j=0}^{2^{n+1}-1-m} \big(\hat{P}_{n+1}(j)\hat{P}_{n+1}(m+j) + \hat{Q}_{n+1}(j)\hat{Q}_{n+1}(m+j)\big) \\
&\qquad = \sum_{j=0}^{2^{n+1}-1-m} \big(\hat{P}_{n+1}(j)\hat{P}_{n+1}(m+j) + \big(\hat{P}_{n+1}(j)\big)\big(-\hat{P}_{n+1}(m+j)\big)\big) = 0.
\end{aligned}
$$

This gives $A_{\hat{P}_{n+1}}(m) + A_{\hat{Q}_{n+1}}(m) = 0$ for all m such that $2^n \le m \le 2^{n+1} - 1$, which completes the inductive step, as well as the proof of the proposition. □

Remark 2.2. This proof remains valid if we begin with any complementary pair of sequences, $\{a_0(j)\}_{j=0}^{k-1}$ and $\{b_0(j)\}_{j=0}^{k-1}$, of length k, and we use the append rule to construct a family, $\mathcal{F}$, of pairs of sequences of length $k2^n$, viz., $\mathcal{F} = \big\langle \{a_n(j)\}_{j=0}^{k2^n-1}, \{b_n(j)\}_{j=0}^{k2^n-1} \big\rangle$ for each $n \in \mathbb{N}$. By changing 2^n to $k2^n$ and 2^{n+1} to $k2^{n+1}$ in the proof of Proposition 2.1 we find that each equilength pair of sequences in $\mathcal{F}$ is a Golay complementary pair. Thus, to show the existence of a Golay pair of sequences each of length k is to show the existence of Golay pairs of sequences of length $k2^n$ for each $n \in \mathbb{N}$.

We have proved that the coefficients of Shapiro–Rudin polynomials form Golay complementary pairs. There are many examples of pairs of sequences that are Golay complementary pairs and are not necessarily the coefficients of Shapiro–Rudin polynomials.

Example 2.3. Let $p = \{2, 3\}$ and $q = \{1, -6\}$. Then

$$A_p(0) + A_q(0) = 2^2 + 3^2 + 1^2 + (-6)^2 = 50 \neq 0$$

and $A_p(1) + A_q(1) = 2 \cdot 3 + 1 \cdot (-6) = 0$. Therefore, p and q form a Golay complementary pair, but the corresponding polynomials P and Q are not Shapiro–Rudin polynomials.

Example 2.4. Let $a, b, c, d \in \mathbb{R}$, and let at least one of a, b, c, d be nonzero. Let $ab + cd = 0$, and let $p = \{a, b, c, d\}$ and $q = \{a, b, -c, -d\}$. Then

$$A_p(0) + A_q(0) = 2(a^2 + b^2 + c^2 + d^2) \neq 0 \quad \text{since one of } a, b, c, d \text{ is nonzero,}$$
$$A_p(1) + A_q(1) = (ab + bc + cd) + (ab - bc + cd) = 2(ab + cd) = 0,$$
$$A_p(2) + A_q(2) = (ac + bd) + (-ac - bd) = 0,$$
$$A_p(3) + A_q(3) = (ad + (-ad)) = 0.$$

Thus, p and q form a Golay complementary pair. By letting $a = b = c = 1$ and $d = -1$, we obtain the special case where $p = \{\hat{P}_2\}$ and $q = \{\hat{Q}_2\}$. Letting a be any nonzero real number and $b = c = -d = a$, we can generate Golay pairs that are not the coefficients of P_2 or Q_2.

Example 2.5. Using the append rule (2-1) and Remark 2.2, we can readily construct a nonbinary Golay complementary pair of sequences of length 2^n for any $n \in \mathbb{N}$. Starting with $p = \{2, 3\}$ and $q = \{1, -6\}$ from Example 2.3, we obtain $\tilde{p} = \{2, 3, 1, -6\}$ and $\tilde{q} = \{2, 3, -1, 6\}$ after one application of the append rule. By Example 2.4, $\tilde{p}$ and $\tilde{q}$ are a Golay complementary pair. After two applications of the append rule, we obtain

$$\tilde{\tilde{p}} = \{2, 3, 1, -6, 2, 3, -1, 6\} \quad \text{and} \quad \tilde{\tilde{q}} = \{2, 3, 1, -6, -2, -3, 1, -6\}.$$

By Remark 2.2, $\tilde{\tilde{p}}$ and $\tilde{\tilde{q}}$ are a Golay complementary pair. Repeated application of the append rule will continue to produce nonbinary Golay complementary pairs of length 2^n for any $n \in \mathbb{N}$.

Example 2.6. It is known that binary Golay complementary pairs of sequences of length $2^a 10^b 26^c$ exist for any nonnegative integers a, b, and c [Turyn 1974]. Earlier, Golay gave examples of Golay complementary sequences of length 10 and 26 [Golay 1961; 1962]. The operation used when calculating the aperiodic autocorrelation coefficients is parity of elements of the sequences (+1 if two elements match, and −1 if they do not). Golay's examples are $p = \{1, 0, 0, 1, 0, 1, 0, 0, 0, 1\}$, $q = \{1, 0, 0, 0, 0, 0, 0, 1, 1, 0\}$ for length 10 sequences, and

$$p = \{1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0\},$$
$$q = \{0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0\}$$

for length 26 sequences. Using the parity operation on these sequences, as Golay did, is equivalent to replacing the zeros in each sequence with (-1)s and using multiplication in the definition of the aperiodic autocorrelation coefficients, as in Definition 1.2.

3. A formula for Shapiro–Rudin coefficients, and some useful MATLAB code

Coefficient formula. Given an $n \in \mathbb{N}$ and k such that $0 \leq k \leq 2^n - 1$, the k-th coefficient of P_n is given in [Brillhart and Carlitz 1970] and [Benke 1994] by the formula $\hat{P}_n(k) = (-1)^{\langle B\omega,\omega\rangle}$, where ω is the $j \times 1$ column vector containing coefficients of the binary expansion of k, and B is the $j \times j$ shift operator matrix given by $B_{m,n} = \delta_{m,n+1}$. The expression $\langle B\omega, \omega\rangle$ is interpreted as the number of occurrences of two consecutive 1s in ω. Note that $k = 0$ corresponds to the first coefficient, $k = 1$ corresponds to the second coefficient, and so on.

MATLAB codes for Shapiro–Rudin coefficients. The following programs were coded using MATLAB v.7.0. The first program, `shapcoef.m`, is a function used in the second program, `shapvector.m`.

`shapcoef.m`

```
function matches=shapcoef(n);
    binary=dec2bin(n);
    binaryShifted=binary;
    binaryShifted(1)='0';
    for c=2:length(binary);
       binaryShifted(c)=binary(c-1);
    end;
    binary;
    binaryShifted;
    matches=0;
    for c=1:length(binary);
       if binary(c)==binaryShifted(c) && binary(c)=='1';
          matches=matches+1;
       end;
    end;
```

`shapvector.m`

```
function shapvector(a,b);
for t=a:b;
    coeff(t+1)=(-1)^shapcoef(t);
end;
B = nonzeros(coeff);
transpose(B)
```

One should use the program `shapvector` by choosing two integers a and b such that $0 \leq a \leq b$, and typing `shapvector(a,b)` into the MATLAB editor window. The program will return the a-th through b-th coefficients of P_n for sufficiently large values of n.

Example 3.1. To compute the coefficients of some P_n, one should use the input `shapvector(0,(2^n)-1)`. For example, the output for $n = 3$ is

```
1     1     1    -1     1     1    -1     1
```

Example 3.2. Suppose we want the coefficients of Q_3. By the append rule (2-1), they coincide with coefficients 8 through 15 of P_4, so we type `shapvector(8,15)`. The output is

```
1     1     1    -1    -1    -1     1    -1
```

Example 3.3. To find the hundredth coefficient of P_n, where $2^n \geq 100$, we type `shapvector(100,100)`. The output is -1.

The program above can be used to construct symbolic Shapiro–Rudin polynomials in MATLAB. One would simply use a for-loop with $k = 0, 1, 2, \ldots, 2^n - 1$ to construct a symbolic vector V whose k-th entry is $e^{2\pi ikt}$, then use the program to compute the vectors C_P of coefficients of P_n, and C_Q of coefficients of Q_n. The dot products $\langle C_P, V\rangle$ and $\langle C_Q, V\rangle$ are P_n and Q_n, respectively.

Parametric images. The parametric image of both P_1 and Q_1 is a circle of unit radius centered at $(1, 0)$. For the next three values of n, we illustrate on the next page the parametric images of P_n and Q_n, with the usual convention: a complex number z is represented by $(\operatorname{Re} z, \operatorname{Im} z)$. Note the complexity of some of these graphs.

4. Geometric descriptions of the curves $(\operatorname{Re} P_n, \operatorname{Im} P_n)$ and $(\operatorname{Re} Q_n, \operatorname{Im} Q_n)$

In Theorem 4.8, we shall show that, for any $n \in \mathbb{N}$, P_{2n} gives rise to a cusp at $t = 0$ while P_{2n+1} and Q_n do not give rise to cusps at $t = 0$. In fact, we shall prove that the cusp of $P_{2n} : \mathbb{R}/\mathbb{Z} \to \mathbb{C}$ occurs at the point $(2^n, 0) \in \mathbb{C}$, and that it is a so-called *quadratic cusp*.

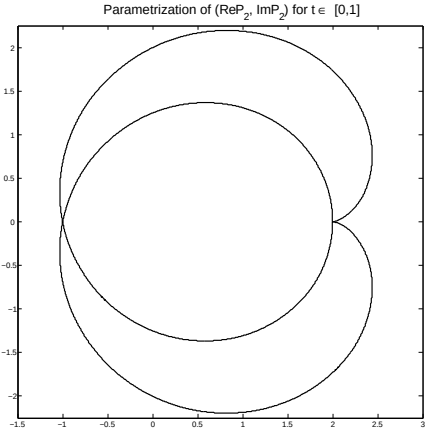
We begin by reinforcing our intuitive notion of a cusp with the following definition [Rutter 2000].

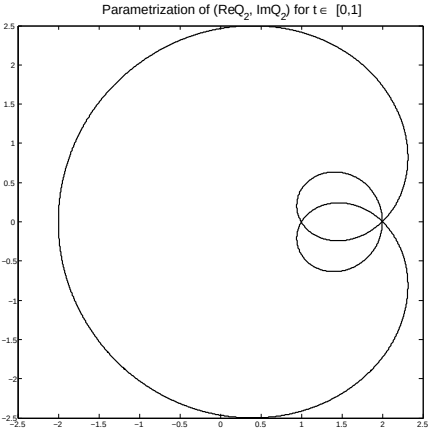
Definition 4.1. A parametrized curve $\gamma : \mathbb{R} \to \mathbb{R}^2$, defined by $\gamma(t) = (u(t), v(t))$, has a *nonregular point* at $t = t_0$ if

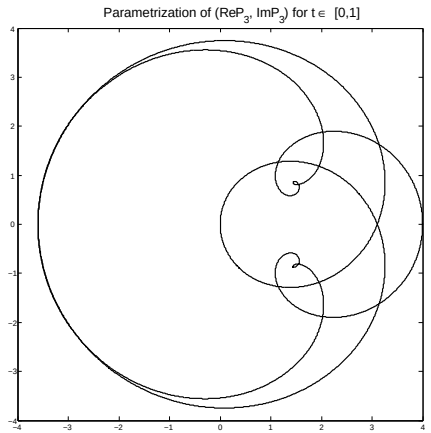
$$\left.\frac{du}{dt}\right|_{t=t_0} = \left.\frac{dv}{dt}\right|_{t=t_0} = 0.$$

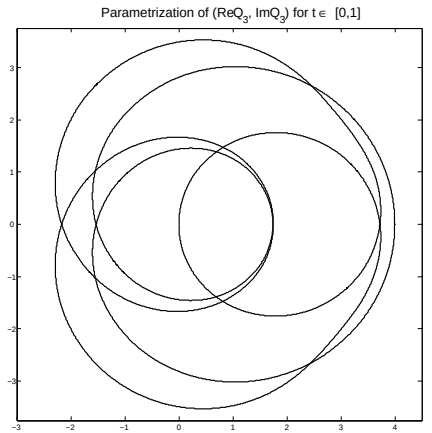
Otherwise, t_0 is a *regular point*. A nonregular point t_0 gives rise to a *quadratic cusp* for γ if

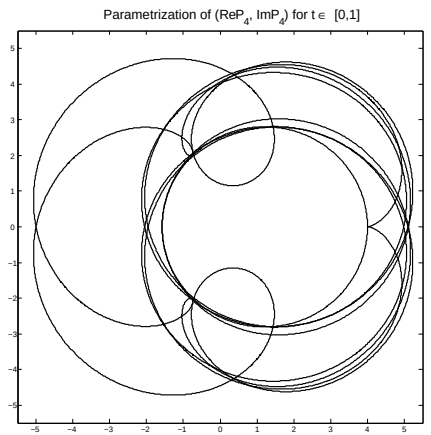
$$\left(\left.\frac{d^2u}{dt^2}\right|_{t=t_0}, \left.\frac{d^2v}{dt^2}\right|_{t=t_0}\right) \neq (0, 0).$$

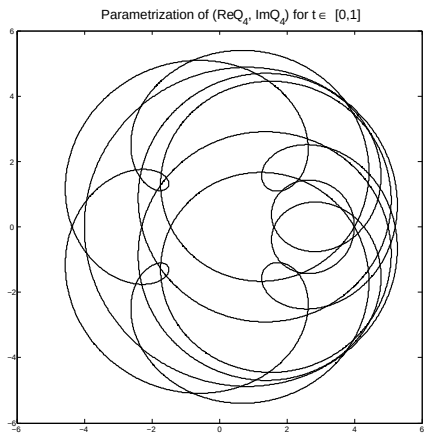

Parametrization of (ReP2, ImP2) for t ∈ [0,1]


Parametrization of (ReQ2, ImQ2) for t ∈ [0,1]


Parametrization of (ReP3, ImP3) for t ∈ [0,1]


Parametrization of (ReQ3, ImQ3) for t ∈ [0,1]


Parametrization of (ReP4, ImP4) for t ∈ [0,1]


Parametrization of (ReQ4, ImQ4) for t ∈ [0,1]

A nonregular point t_0 gives rise to an *ordinary cusp* if it gives rise to a quadratic cusp, and

$$\left(\frac{d^2u}{dt^2}\Big|_{t=t_0}, \frac{d^2v}{dt^2}\Big|_{t=t_0}\right) \quad \text{and} \quad \left(\frac{d^3u}{dt^3}\Big|_{t=t_0}, \frac{d^3v}{dt^3}\Big|_{t=t_0}\right)$$

are linearly independent points of the real vector space $\mathbb{R}^2$, that is, they are not parallel vectors in $\mathbb{R}^2$.

Example 4.2. Let $P(z) = z^2 - 2z$ on $\mathbb{C}$. Then, P' has a zero of multiplicity 1 at $z_0 = 1$. In the notation of Definition 4.1, we consider $\gamma : \mathbb{R} \to \mathbb{R}^2$, where $\gamma(t) = P(e^{2\pi it})$, $t \in \mathbb{R}$, and so

$$u(t) = \cos(4\pi t) - 2\cos(2\pi t) \quad \text{and} \quad v(t) = \sin(4\pi t) - 2\sin(2\pi t).$$

We compute that γ has a nonregular point at $t_0 = 0$, and, in fact, $t_0 = 0$ gives rise to a quadratic cusp.

Further, if $Q : \mathbb{C} \to \mathbb{C}$ is any polynomial with complex coefficients, then $t = t_0$ gives rise to a quadratic cusp for γ, where $\gamma(t) = Q(e^{2\pi it})$, if and only if Q' vanishes at $e^{2\pi it_0}$ with odd multiplicity. The angle at the cusp point $z_0 = e^{2\pi it_0}$ naturally depends on the order of the multiplicity. This assertion of odd order of multiplicity to characterize a cusp is not restricted to polynomials, but is valid for any complex valued analytic function.

Remark 4.3. To show that P_{2n} gives rise to a quadratic cusp at $t = 0$, we must first show the existence of a nonregular point at $t = 0$, and to show that P_{2n} has a nonregular point at $t = 0$, we must show

$$\frac{d}{dt}\operatorname{Re} P_{2n}\Big|_{t=0} = \frac{d}{dt}\operatorname{Im} P_{2n}\Big|_{t=0} = 0. \tag{4-1}$$

To show that P_{2n+1} and Q_n have regular points at $t = 0$, we shall verify that

$$\frac{d}{dt}\operatorname{Re} P_{2n+1}\Big|_{t=0} \neq 0 \quad \text{or} \quad \frac{d}{dt}\operatorname{Im} P_{2n+1}\Big|_{t=0} \neq 0 \tag{4-2}$$

and

$$\frac{d}{dt}\operatorname{Re} Q_n\Big|_{t=0} \neq 0 \quad \text{or} \quad \frac{d}{dt}\operatorname{Im} Q_n\Big|_{t=0} \neq 0, \tag{4-3}$$

respectively. Clearly, (4-1) is equivalent to showing $(dP_{2n}/dt)|_{t=0} = 0$, while (4-2) is equivalent to showing $(dP_{2n+1}/dt)|_{t=0} \neq 0$ and (4-3) is equivalent to showing $(dQ_n/dt)|_{t=0} \neq 0$. These calculations are contained in the proof of Theorem 4.8.

Example 4.4. We calculate the derivatives of P_n and Q_n. By writing the coefficients of P_n and Q_n as $\{\hat{P}_n(k)\}_{k=0}^{2^n-1}$ and $\{\hat{Q}_n(k)\}_{k=0}^{2^n-1}$, we have

$$P_n(t) = \sum_{k=0}^{2^n-1} \hat{P}_n(k) e^{2\pi ikt} \quad \text{and} \quad Q_n(t) = \sum_{k=0}^{2^n-1} \hat{Q}_n(k) e^{2\pi ikt}.$$

Consequently,

$$\frac{dP_n(t)}{dt} = \frac{d}{dt}\sum_{k=0}^{2^n-1} \hat{P}_n(k)e^{2\pi ikt} = 2\pi i \sum_{k=0}^{2^n-1} k\hat{P}_n(k)e^{2\pi ikt},$$
$$\frac{dQ_n(t)}{dt} = \frac{d}{dt}\sum_{k=0}^{2^n-1} \hat{Q}_n(k)e^{2\pi ikt} = 2\pi i \sum_{k=0}^{2^n-1} k\hat{Q}_n(k)e^{2\pi ikt}.$$

The following well-known formulas for the sums of coefficients of Shapiro–Rudin polynomials are used in the verification of Proposition 4.6.

Proposition 4.5. *For each $n \in \mathbb{N}$,*

$$\sum_{k=0}^{2^n-1} \hat{P}_n(k) = \begin{cases} 2^{(n+1)/2} & \text{if } n \text{ is odd,} \\ 2^{n/2} & \text{if } n \text{ is even;} \end{cases} \qquad \sum_{k=0}^{2^n-1} \hat{Q}_n(k) = \begin{cases} 0 & \text{if } n \text{ is odd,} \\ 2^{n/2} & \text{if } n \text{ is even.} \end{cases} \tag{4-4}$$

Proof. From the append rule (2-1), we have

$$\sum_{k=0}^{2^{n+1}-1} \hat{P}_{n+1}(k) = \sum_{k=0}^{2^n-1} \hat{P}_n(k) + \sum_{k=0}^{2^n-1} \hat{Q}_n(k), \tag{4-5}$$
$$\sum_{k=0}^{2^{n+1}-1} \hat{Q}_{n+1}(k) = \sum_{k=0}^{2^n-1} \hat{P}_n(k) - \sum_{k=0}^{2^n-1} \hat{Q}_n(k). \tag{4-6}$$

We complete the proof using induction. To verify the basic cases, we observe: for $n = 1$, $\sum_{k=0}^{1} \hat{P}_1(k) = 1 + 1 = 2^1$ and $\sum_{k=0}^{1} \hat{Q}_1(k) = 1 - 1 = 0$, and for $n = 2$, $\sum_{k=0}^{3} \hat{P}_2(k) = 1+1+1-1 = 2^{(3-1)/2}$ and $\sum_{k=0}^{3} \hat{Q}_2(k) = 1+1-1+1 = 2^{(3-1)/2}$. For the inductive step, suppose (4-4) holds for some $n \in \mathbb{N}$. Then, if n is even, $\sum_{k=0}^{2^n-1} \hat{P}_n(k) = 2^{n/2}$ and $\sum_{k=0}^{2^n-1} \hat{Q}_n(k) = 2^{n/2}$. Hence,

$$\sum_{k=0}^{2^{n+1}-1} \hat{P}_{n+1}(k) = \sum_{k=0}^{2^n-1} \hat{P}_n(k) + \sum_{k=0}^{2^n-1} \hat{Q}_n(k) = 2^{n/2} + 2^{n/2} = 2^{(n/2)+1} = 2^{((n+1)+1)/2},$$
$$\sum_{k=0}^{2^{n+1}-1} \hat{Q}_{n+1}(k) = \sum_{k=0}^{2^n-1} \hat{P}_n(k) - \sum_{k=0}^{2^n-1} \hat{Q}_n(k) = 2^{n/2} - 2^{n/2} = 0,$$

completing the induction step. The verification in the case of n odd is entirely analogous. □

We define the finite sums

$$S_P(n) = \frac{1}{2\pi i}\frac{dP_n}{dt}\Big|_{t=0} = \sum_{k=0}^{2^n-1} k\hat{P}_n(k), \qquad S_Q(n) = \frac{1}{2\pi i}\frac{dQ_n}{dt}\Big|_{t=0} = \sum_{k=0}^{2^n-1} k\hat{Q}_n(k).$$

Using this notation, relations (4-1)–(4-3) become, respectively,

$$S_P(2n) = 0, \tag{4-7}$$

$$S_P(2n+1) \neq 0, \tag{4-8}$$

$$S_Q(n) \neq 0. \tag{4-9}$$

The following result is used in the proof of Theorem 4.8.

Proposition 4.6. *For all* $n \in \mathbb{N}$,

$$\begin{aligned} S_P(n+1) &= \begin{cases} S_P(n) + S_Q(n) & \text{if } n \text{ is odd}, \\ S_P(n) + (S_Q(n) + 2^{3n/2}) & \text{if } n \text{ is even}; \end{cases} \\ S_Q(n+1) &= \begin{cases} S_P(n) + S_Q(n) & \text{if } n \text{ is odd}, \\ S_P(n) - (S_Q(n) + 2^{3n/2}) & \text{if } n \text{ is even}. \end{cases} \end{aligned} \tag{4-10}$$

Proof. Using (4-5), we have, for every $n \in \mathbb{N}$,

$$\begin{aligned} S_P(n+1) &= S_P(n) + \Big(S_Q(n) + 2^n \sum_{k=0}^{2^n-1} \hat{Q}_n(k)\Big) \\ &= \sum_{k=0}^{2^n-1} k \hat{P}_{n+1}(k) + \sum_{k=2^n}^{2^{n+1}-1} k \hat{P}_{n+1}(k) = \sum_{k=0}^{2^{n+1}-1} k \hat{P}_{n+1}(k) \\ &= \sum_{k=0}^{2^n-1} k \hat{P}_n(k) + \sum_{k=2^n}^{2^{n+1}-1} \big((k-2^n) + 2^n\big) \hat{P}_{n+1}(k) \\ &= \sum_{k=0}^{2^n-1} k \hat{P}_n(k) + \sum_{k=2^n}^{2^{n+1}-1} (k-2^n) \hat{P}_{n+1}(k) + \sum_{k=2^n}^{2^{n+1}-1} 2^n \hat{P}_{n+1}(k) \\ &= \sum_{k=0}^{2^n-1} k \hat{P}_n(k) + \sum_{k=0}^{2^n-1} k \hat{Q}_n(k) + 2^n \sum_{k=0}^{2^n-1} \hat{Q}_n(k) \\ &= \begin{cases} S_P(n) + S_Q(n) & \text{if } n \text{ is odd}, \\ S_P(n) + \big(S_Q(n) + 2^{3n/2}\big) & \text{if } n \text{ is even}. \end{cases} \end{aligned}$$

The expression for $S_Q(n+1)$ is proved analogously, starting from (4-6). □

Example 4.7. Define the finite sums

$$S_{P,2}(n) = -\frac{1}{4\pi^2} \frac{d^2 P_n}{dt^2}\Big|_{t=0} = \sum_{k=0}^{2^n-1} k^2 \hat{P}_n(k),$$

$$S_{Q,2}(n) = -\frac{1}{4\pi^2} \frac{d^2 Q_n}{dt^2}\Big|_{t=0} = \sum_{k=0}^{2^n-1} k^2 \hat{Q}_n(k).$$

In [Brillhart 1973], the following formulas relating to the second derivatives of Shapiro–Rudin polynomials are proved. These formulas will be used in Theorem 4.8 to classify the cusps of P_n and Q_n.

$$S_{P,2}(2n) = \frac{-2^{n+1}(2^n-1)(2^{2n+2}-1)}{45}, \tag{4-11}$$

$$S_{P,2}(2n+1) = \frac{2^{n+2}(2^{2n}-1)(2^{2n+2}-1)}{9}, \tag{4-12}$$

$$S_{Q,2}(2n) = \frac{2^{n+1}(2^{2n}-1)(13\cdot 2^{2n-1}-11)}{45}, \tag{4-13}$$

$$S_{Q,2}(2n+1) = \frac{-2^{n+3}(2^{2n}-1)(2^{2n+2}-1)}{15}. \tag{4-14}$$

We shall now prove that P_{2n} gives rise to a quadratic cusp at $t=0$. We shall also prove that this cusp occurs at the point $(2^n, 0)$. Lastly, we shall prove that P_{2n+1} and Q_n do not give rise to cusps at $t=0$ as a result of the fact that $t=0$ is a regular point of each of these curves.

Theorem 4.8. *For each $n \in \mathbb{N}$, the parametrization* $(\operatorname{Re} P_{2n}, \operatorname{Im} P_{2n})$ *gives rise to a quadratic cusp at* $(2^n, 0)$, *that is, when* $t=0$, *and neither* $(\operatorname{Re} P_{2n+1}, \operatorname{Im} P_{2n+1})$ *nor* $(\operatorname{Re} Q_n, \operatorname{Im} Q_n)$ *gives rise to a cusp when* $t=0$.

Proof. (i) We notice that $P_{2n}(0) = \sum_{k=0}^{2^{2n}-1} \hat{P}_{2n}(k) = 2^{2n/2} = 2^n$ by (4-4). This implies that $\operatorname{Re} P_{2n}(0) = 2^n$ and $\operatorname{Im} P_n(0) = 0$. Thus, at $t=0$, $(\operatorname{Re} P_{2n}, \operatorname{Im} P_{2n}) = (2^n, 0)$. It is clear that none of (4-11), (4-12), (4-13), or (4-14) can ever equal zero, and, hence, none of the second derivatives can equal zero. This proves that $t=0$ is at least a quadratic cusp of the parametrization $(\operatorname{Re} P_{2n}, \operatorname{Im} P_{2n})$, provided $t=0$ is, in fact, a nonregular point of the curve.

To prove that $t=0$ is a nonregular point of P_{2n}, it suffices to prove (4-7). We shall also prove (4-8) and (4-9), which will, in turn, prove that $t=0$ is a regular point of P_{2n+1} and Q_n.

(ii) Using induction, we shall prove (4-7), (4-8), and (4-9) by showing that, for each $n \in \mathbb{N}$,

$$S_P(n) = \begin{cases} 0 & \text{if } n \text{ is even,} \\ \frac{4}{3}(2^{3(n-1)/2} - 2^{(n-1)/2}) + 2^{(n-1)/2} & \text{if } n \text{ is odd} \end{cases} \tag{4-15}$$

and

$$S_Q(n) = \begin{cases} \frac{1}{3}(2^{3n/2} - 2^{n/2}) & \text{if } n \text{ is even,} \\ -S_P(n) = -\frac{4}{3}(2^{3(n-1)/2} - 2^{(n-1)/2}) - 2^{(n-1)/2} & \text{if } n \text{ is odd.} \end{cases} \tag{4-16}$$

We start with $n=1$, where $S_P(1) = 0+1 = 1 = \frac{4}{3}(2^0-1)(2^0)+2^0$ and $S_Q(1) = 0-1 = -1 = -\frac{4}{3}(2^0-1)(2^0)-2^0$, and with $n=2$, we have $S_P(2) = 0+1+2-3 = 0$ and $S_Q(2) = \frac{1}{3}(2^2-1)(2^{2/2}) = 2$.

To prove the inductive step, assume (4-15) and (4-16) hold for some $n \in \mathbb{N}$. Assume first that the case where n is even, n is even, so $n+1$ is odd. By (4-10) we have

$$\begin{aligned} S_P(n+1) &= S_P(n) + S_Q(n) + 2^{3n/2} = 0 + \tfrac{1}{3}(2^{3n/2}) - \tfrac{1}{3}(2^{n/2}) + 2^{3n/2} \\ &= \tfrac{4}{3}(2^{3n/2}) - \tfrac{1}{3}(2^{n/2}) = \tfrac{4}{3}(2^{3n/2}) - \tfrac{4}{3}(2^{n/2}) + 2^{n/2} = \tfrac{4}{3}(2^{3n/2} - 2^{n/2}) + 2^{n/2}, \\ S_Q(n+1) &= S_P(n) - S_Q(n) - 2^{3n/2} = -\tfrac{1}{3}(2^{3n/2}) + \tfrac{1}{3}(2^{n/2}) - 2^{3n/2}, \\ &= -\tfrac{4}{3}(2^{3n/2}) + \tfrac{4}{3}(2^{n/2}) - 2^{n/2} = -\tfrac{4}{3}(2^{3n/2} - 2^{n/2}) - 2^{n/2}, \end{aligned}$$

completing the induction step in this case. The complementary case is proved similarly. □

Appendix

The cusps arising in P_{2n} can be explicitly studied using only elementary calculations. Although such calculations are not very illuminating, they illustrate the difficulty of discovering and verifying the assertion of Theorem 4.8 by a direct approach, as opposed to the way we have proceeded. In this appendix we spell out the details of the special case $P_2(t)$.

We have

$$\begin{aligned} P_2(t) = P_{1+1}(t) &= P_1(t) + e^{2\pi i 2t} Q_1(t) = P_{0+1} + e^{2\pi i 2t} Q_{0+1} \\ &= P_0(t) + e^{2\pi i t} Q_0(t) + e^{2\pi i 2t} \left(P_0(t) - e^{2\pi i t} Q_0(t)\right) \\ &= 1 + e^{2\pi i t} + e^{2\pi i 2t} - e^{2\pi i 3t}. \end{aligned}$$

Define

$$\begin{aligned} P_r(t) &= \operatorname{Re} P_2(t) = 1 + \cos(2\pi t) + \cos(2\pi 2t) - \cos(2\pi 3t), \\ P_i(t) &= \operatorname{Im} P_2(t) = \sin(2\pi t) + \sin(2\pi 2t) - \sin(2\pi 3t). \end{aligned}$$

We know that $P_2(t) = \operatorname{Re} P_2(t) + i \operatorname{Im} P_2(t)$ for $t \in [0, 1]$, and so $P_2(t) = 2 + i0 = (2, 0) \in \mathbb{C}$ at $t = 0$.

Let $\alpha = 1/\pi^5$. We must show several facts:

(a) $P_i(t) > 0$ for $t \in (0, \alpha]$.

(b) $P_i(t) < 0$ for $t \in [-\alpha, 0)$.

(c) $P_r(t) > 0$ for $t \in [-\alpha, \alpha] \setminus \{0\}$.

(d) $P_r(t)$ is strictly increasing on $(0, \alpha]$.

(e) $P_r(t)$ is strictly decreasing on $[-\alpha, 0)$.

(f) $P_i(t)$ is strictly increasing on $[-\alpha, \alpha]\setminus\{0\}$.

(g) $\lim_{t\to 0^+} P_i'(t)/P_r'(t)$ and $\lim_{t\to 0^-} P_i'(t)/P_r'(t)$ both exist as finite real numbers.

These seven facts imply that P_2 gives rise to a cusp at $(2,0) \in \mathbb{C}$, as follows. Conditions (a), (b), and (f) together show that P_2 is traced out in the complex plane from below the real axis to above it, crossing only when $t = 0$. Conditions (c), (d), and (e) together show that P_2 crosses the real axis on the right side of the line $\{2 + xi \ : \ x \in \mathbb{R}\}$, only touching the line when $t = 0$. Finally, (g) shows that the curve is not smooth at $(2, 0)$; in conjunction with (a)–(f), the limits would need to be $\pm\infty$ for no cusp to arise.

We shall use the following Taylor series estimates. For all $x \in \mathbb{R}$,

$$x - \frac{x^3}{3!} \le \sin x \le x - \frac{x^3}{3!} + \frac{x^5}{5!} \tag{A.1}$$

and

$$1 - \frac{x^2}{2!} \le \cos x \le 1 - \frac{x^2}{2!} + \frac{x^4}{4!}. \tag{A.2}$$

Verification of (a), viz., $P_i(t) = \sin(2\pi t) + \sin(4\pi t) - \sin(6\pi t) > 0$ for all $t \in (0, \alpha]$. Using (A.1), we make the estimates

$$\sin(2\pi t) + \sin(4\pi t) \ge 2\pi t - \frac{(2\pi t)^3}{3!} + 4\pi t - \frac{(4\pi t)^3}{3!} = 6\pi t - \frac{1}{3!}\left((2\pi t)^3 + (4\pi t)^3\right),$$

$$\sin(6\pi t) \le 6\pi t - \frac{(6\pi t)^3}{3!} + \frac{(6\pi t)^5}{5!}.$$

Hence, it suffices to show that for all $t \in (0, \alpha]$,

$$6\pi t - \frac{(6\pi t)^3}{3!} + \frac{(6\pi t)^5}{5!} < 6\pi t - \frac{1}{3!}\left((2\pi t)^3 + (4\pi t)^3\right),$$

that is,

$$\frac{(6\pi t)^5}{5!} < \frac{1}{3!}(2\pi)^3\left(-t^3 - (2t)^3 + (3t)^3\right) = \frac{18}{3!}(2\pi)^3 t^3.$$

Since $t > 0$, this simplifies to

$$t^2 < \frac{20}{(2\pi)^2}\frac{18}{3^5},$$

which in turn is solved by $0 < t < \frac{\sqrt{5}}{\pi}\frac{3\sqrt{2}}{3^{5/2}} = \frac{\sqrt{10}}{3^{3/2}\pi}$. Since $\alpha = \frac{1}{\pi^5} < \frac{\sqrt{10}}{3^{3/2}\pi}$, we have proved (a).

Verification of (b), viz., $P_i(t) = \sin(2\pi t) + \sin(4\pi t) - \sin(6\pi t) < 0$ for all $t \in [-\alpha, 0)$. The proof of (b) relies on the fact that the sine function is odd. Let $t = -s$, $s \in (0, \alpha]$. Then

$$\sin(2\pi t) + \sin(4\pi t) = -\sin(2\pi s) - \sin(4\pi s) = -\left(\sin(2\pi s) + \sin(4\pi s)\right).$$

We know from (a) that $\sin(2\pi s) + \sin(4\pi s) > \sin(6\pi s)$ for $s \in (0, \alpha]$. Hence $-\sin(6\pi s) > -(\sin(2\pi s) + \sin(4\pi s))$ for $s \in (0, \alpha]$, and therefore, for $t \in [-\alpha, 0)$,

$$\sin(6\pi t) > \sin(2\pi t) + \sin(4\pi t).$$

Hence, (b) is proved.

Verification of (c), viz., $P_r(t) = 1 + \cos(2\pi t) + \cos(4\pi t) - \cos(6\pi t) > 0$ for all $t \in [-\alpha, \alpha]\backslash\{0\}$. It suffices to verify the inequality for $t \in (0, \alpha]$ since the cosine function is even.

Using (A.2), we make the estimates

$$1 + \cos(6\pi t) \le 2 - \frac{(6\pi t)^2}{2!} + \frac{(6\pi t)^4}{4!}$$

$$\cos(2\pi t) + \cos(4\pi t) \ge 1 - \frac{(2\pi t)^2}{2!} + 1 - \frac{(4\pi t)^2}{2!}.$$

Hence, to prove (c), it suffices to show that, for all $t \in (0, \alpha]$,

$$2 - \frac{(6\pi t)^2}{2!} + \frac{(6\pi t)^4}{4!} < 2 - \Big(\frac{(2\pi t)^2 + (4\pi t)^2}{2!}\Big).$$

Simplifying, we obtain $\frac{(6\pi t)^4}{4!} < -6\pi^2 t^2 + \frac{36\pi^2 t^2}{2}$, which turns into

$$54\pi^4 t^4 < 12\pi^2 t^2.$$

Since $t > 0$, we divide by $6\pi^2 t^2$ to obtain the inequality $9t^2\pi^2 < 2$, which in turn is solved by $0 < t < \sqrt{2}/3\pi$. Since $\alpha = 1/\pi^5 < \sqrt{2}/3\pi$, we have proved (c).

Verification of (d), viz., $P_r'(t) = -2\pi \sin(2\pi t) - 4\pi \sin(4\pi t) + 6\pi \sin(6\pi t) > 0$ for $t \in (0, \alpha]$. We shall prove $3\sin(6\pi t) > 2\sin(4\pi t) + \sin(2\pi t)$ for all $t \in (0, \alpha]$.

Using (A.1), we make the estimates

$$3\sin(6\pi t) \ge 3\Big(6\pi t - \frac{(6\pi t)^3}{3!}\Big),$$

$$\begin{aligned} 2\sin(4\pi t) + \sin(2\pi t) &\le 2\Big(4\pi t - \frac{(4\pi t)^3}{3!} + \frac{(4\pi t)^5}{5!}\Big) + 2\pi t - \frac{(2\pi t)^3}{3!} + \frac{(2\pi t)^5}{5!} \\ &= 10\pi t - \frac{(2\pi)^3}{3!}(t^3)(1 + 2^4) + \frac{(2\pi)^5}{5!}(t^5)(1 + 2^6). \end{aligned}$$

Hence, to prove (d), it suffices to show that, for all $t \in (0, \alpha]$,

$$10\pi t - \frac{(2\pi)^3}{3!}(t^3)(1 + 2^4) + \frac{(2\pi)^5}{5!}(t^5)(1 + 2^6) < 3\Big(6\pi t - \frac{(6\pi t)^3}{3!}\Big).$$

Rearranging the inequality, we obtain

$$10\pi t + \frac{(2\pi)^5}{5!}(t^5)(1 + 2^6) + \frac{(6\pi t)^3}{2} < 18\pi t + \frac{(2\pi)^3}{3!}(t^3)(1 + 2^4),$$

that is,

$$\frac{(2\pi)^4}{4!}13t^5+\frac{(2\pi)^2}{2}27t^3<4t+\frac{(2\pi)^2}{3!}17t^3.$$

Since $t>0$, this simplifies to

$$\frac{(2\pi)^4}{4!}13t^4+\frac{(2\pi)^2}{2!}\Big(27-\frac{17}{3}\Big)t^2<4.$$

Since we are attempting to prove that the inequality holds for $t\in(0,\alpha]$ with $\alpha<1$, we take advantage of the fact that $t^4<t^2$ when $0<t<1$ to make the estimate

$$\begin{aligned}\frac{(2\pi)^4}{4!}13t^4+\frac{(2\pi)^2}{2!}\big(27-\tfrac{17}{3}\big)t^2&<t^2\Big(\frac{(2\pi)^4(13)}{4!}+\frac{(2\pi)^2(64)}{3!}\Big)<t^2\Big(\frac{(2\pi)^4(78)}{3!}\Big)\\&=t^2(2\pi)^4(13)<t^2(2\pi)^4(2\pi)^2=t^2(2\pi)^6.\end{aligned}$$

So we obtain the inequality $t^2(2\pi)^6<4$, which is solved by $0<t<\dfrac{2}{(2\pi)^3}=\dfrac{1}{4\pi^3}$. Since $\alpha=\dfrac{1}{\pi^5}<\dfrac{1}{4\pi^3}$, we have proved (d).

Verification of (e), viz., $P_r'(t)=-2\pi\sin(2\pi t)-4\pi\sin(4\pi t)+6\pi\sin(6\pi t)<0$ for $t\in[-\alpha,0)$. We prove that $P_r(t)$ is strictly decreasing on $[-\alpha,0)$ using the fact that the sine function is odd — the same method we used to prove (b).

We know from the calculations in the previous page that $P_r'(t)=-2\pi\sin(2\pi t)-4\pi\sin(4\pi t)+6\pi\sin(6\pi t)>0$ when $t\in(0,\alpha]$. Letting $t=-s$, $s\in(0,\alpha]$, we have

$$-2\pi\sin(2\pi s)-4\pi\sin(4\pi s)+6\pi\sin(6\pi s)>0,\ \ s\in(0,\alpha],$$

which leads to

$$-2\pi\sin(2\pi t)-4\pi\sin(4\pi t)+6\pi\sin(6\pi t)<0,\ \ t\in[-\alpha,0).$$

Thus, for $t\in[-\alpha,0)$, $P_r'(t)<0$, so $P_r(t)$ is strictly decreasing on $[-\alpha,0)$.

Verification of (f), viz., $P_i'(t)=2\pi\cos(2\pi t)+4\pi\cos(4\pi t)-6\pi\cos(6\pi t)>0$ for $t\in[-\alpha,\alpha]\backslash\{0\}$. It suffices to verify the inequality for $t\in(0,\alpha]$ since the cosine function is even.

Using (A.2), we make the estimates

$$\cos(2\pi t)+2\cos(4\pi t)\ge 1-\frac{(2\pi t)^2}{2!}+2-\frac{2(4\pi t)^2}{2!}=3-\Big(\frac{(2\pi t)^2}{2}+(4\pi t)^2\Big),$$

$$3\cos(6\pi t)\le 3-\frac{3(6\pi t)^2}{2!}+\frac{3(6\pi t)^4}{4!}.$$

Hence, to prove (f), it suffices to show that for all $t\in(0,\alpha]$,

$$3-\frac{3(6\pi t)^2}{2!}+\frac{3(6\pi t)^4}{4!}<3-\Big(\frac{(2\pi t)^2}{2}+(4\pi t)^2\Big),$$

that is, $-54\pi^2t^2 + \dfrac{2^4 3^5 \pi^4 t^4}{2^3 3} < -18\pi^2 t^2$, which simplifies to

$$162\pi^4 t^4 < 36\pi^2 t^2.$$

Since $t > 0$, we divide by $6\pi^2 t^2$ to obtain the inequality

$$27\pi^2 t^2 < 6,$$

which in turn is solved by $0 < t < \dfrac{\sqrt{6}}{\pi\sqrt{27}}$. Since $\alpha = \dfrac{1}{\pi^5} < \dfrac{\sqrt{6}}{\pi\sqrt{27}}$, this proves (f).

Verification of (g), viz., $\lim_{t\to 0^+} P_i'(t)/P_r'(t)$ and $\lim_{t\to 0^-} P_i'(t)/P_r'(t)$ both exist as finite real numbers. The limits need not be equal, so we evaluate them separately.

$$\lim_{t\to 0^+} \frac{P_i'(t)}{P_r'(t)} = \lim_{t\to 0^+} \frac{2\pi(\cos(2\pi t) + 2\cos(4\pi t) - 3\cos(6\pi t))}{-2\pi(\sin(2\pi t) + 2\sin(4\pi t) - 3\sin(6\pi t))},$$

which has the form $0/0$ when plugging in $t = 0$. We use L'Hôpital's rule to get

$$\begin{aligned}\lim_{t\to 0^+} \frac{P_i'(t)}{P_r'(t)} &= \lim_{t\to 0^+} \frac{P_i''(t)}{P_r''(t)} \\ &= \lim_{t\to 0^+} \frac{-(2\pi)^2(\sin(2\pi t) + 4\sin(4\pi t) - 9\sin(6\pi t))}{-(2\pi)^2(\cos(2\pi t) + 4\cos(4\pi t) - \cos(6\pi t))} = \frac{0}{4(2\pi)^2} = 0.\end{aligned}$$

Thus, the limit exists as a finite real number.

Since $\lim_{t\to 0^-} P_i'(t)/P_r'(t)$ also has the form $0/0$, and since

$$\frac{P_i''(t)}{P_r''(t)} = \frac{-(2\pi)^2(\sin(2\pi t) + 4\sin(4\pi t) - 9\sin(6\pi t))}{-(2\pi)^2(\cos(2\pi t) + 4\cos(4\pi t) - \cos(6\pi t))}$$

is continuous at $t = 0$, we have

$$\lim_{t\to 0^-} \frac{P_i'(t)}{P_r'(t)} = \lim_{t\to 0^+} \frac{P_i'(t)}{P_r'(t)} = \lim_{t\to 0} \frac{P_i'(t)}{P_r'(t)} = \lim_{t\to 0} \frac{P_i''(t)}{P_r''(t)} = \frac{P_i''(0)}{P_r''(0)} = 0$$

as well. Hence, (g) is proved, which also shows that $P_2(t)$ admits a cusp when $t = 0$.

5. Acknowledgements

The authors acknowledge fruitful discussions several years ago with J. Donatelli, T. Dulaney, and S. Gerber. More recently, we acknowledge the invaluable scholarship of B. Saffari and the indispensable assistance of E. Au-Yeung. Benedetto gratefully acknowledges support from the AFOSR grant FA9550-05-1-0443. Sugar Moore gratefully acknowledges support from the University of Maryland, Department of Mathematics NSF VIGRE Grant, as well as the Daniel Sweet Undergraduate Research Fellowship of the Norbert Wiener Center at the University of Maryland.

References

[Benedetto 1997] J. J. Benedetto, *Harmonic analysis and applications*, CRC Press, Boca Raton, FL, 1997. MR 97m:42001

[Benke 1994] G. Benke, "Generalized Rudin–Shapiro systems", *J. Fourier Anal. Appl.* **1**:1 (1994), 87–101. MR 96d:42001 Zbl 0835.42014

[Brillhart 1973] J. Brillhart, "On the Rudin–Shapiro polynomials", *Duke Math. J.* **40** (1973), 335–353. MR 47 #3645 Zbl 0263.33012

[Brillhart and Carlitz 1970] J. Brillhart and L. Carlitz, "Note on the Shapiro polynomials", *Proc. Amer. Math. Soc.* **25** (1970), 114–118. MR 41 #5575 Zbl 0191.35101

[Brillhart and Morton 1996] J. Brillhart and P. Morton, "A case study in mathematical research: the Golay–Rudin–Shapiro sequence", *Amer. Math. Monthly* **103**:10 (1996), 854–869. MR 98g:01048 Zbl 0873.11020

[Budišin 1990] S. Z. Budišin, "New complementary pairs of sequences", *Electronics Let.* **26**:13 (1990), 881–883.

[Daubechies 1992] I. Daubechies, *Ten lectures on wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics **61**, Soc. Industrial Appl. Math., Philadelphia, 1992. MR 93e:42045 Zbl 0776.42018

[Eliahou et al. 1990] S. Eliahou, M. Kervaire, and B. Saffari, "A new restriction on the lengths of Golay complementary sequences", *J. Combin. Theory Ser. A* **55**:1 (1990), 49–59. MR 91i:11020 Zbl 0705.94012

[Eliahou et al. 1991] S. Eliahou, M. Kervaire, and B. Saffari, "On Golay polynomial pairs", *Adv. in Appl. Math.* **12**:3 (1991), 235–292. MR 93b:68066 Zbl 0767.05004

[Golay 1949] M. J. E. Golay, "Multi-slit spectrometry", *J. Opt. Soc. Amer.* **39**:6 (1949), 437–444.

[Golay 1951] M. J. E. Golay, "Static multslit spectrometry and its application to the panoramic display of infrared spectra", *J. Opt. Soc. Amer.* **41**:7 (1951), 468–472.

[Golay 1961] M. J. E. Golay, "Complementary series", *IRE Trans.* **IT-7**:2 (1961), 82–87. MR 23 #A3096

[Golay 1962] M. J. E. Golay, Note on "Complementary series", 1962. In correspondence section of *Proc. IRE* **50**:1, p. 84.

[Howard et al. 2006] S. D. Howard, A. R. Calderbank, and W. Moran, "The finite Heisenberg–Weyl groups in radar and communications", *EURASIP J. Appl. Signal Process.* (2006), Art. ID 85685. MR 2233868

[Jedwab 2005] J. Jedwab, "A survey of the merit factor problem for binary sequences", pp. 30–55 in *Sequences and their applications: Proceedings of SETA 2004*, Lecture Notes in Computer Science **3486**, Springer, Berlin, 2005.

[Jedwab and Yoshida 2006] J. Jedwab and K. Yoshida, "The peak sidelobe level of families of binary sequences", *IEEE Trans. Inform. Theory* **52**:5 (2006), 2247–2254. MR 2006m:94044

[Kahane 1970] J.-P. Kahane, *Séries de Fourier absolument convergentes*, Ergebnisse der Math. **50**, Springer, Berlin, 1970. MR 43 #801 Zbl 0195.07602

[Katznelson 1976] Y. Katznelson, *An introduction to harmonic analysis*, 2nd corr. ed., Dover Publications, New York, 1976. MR 54 #10976 Zbl 0352.43001

[Levanon and Mozeson 2004] N. Levanon and E. Mozeson, *Radar signals*, Wiley and IEEE Press, Hoboken, NJ, 2004.

[Lüke 1985] H. D. Lüke, "Sets of one and higher dimensional Welti codes and complementary codes", *IEEE Trans. Aerospace and Electronic Systems* **21** (1985), 170–179.

[Mallat 1998] S. Mallat, *A wavelet tour of signal processing*, Academic Press, San Diego, CA, 1998. MR 99m:94012 Zbl 0937.94001

[Pezeshki et al. 2008] A. Pezeshki, A. R. Calderbank, W. Moran, and S. D. Howard, "Doppler resilient Golay complementary waveforms", *IEEE Trans. Inform. Theory* **54**:9 (2008), 4254–4266.

[Rudin 1959] W. Rudin, "Some theorems on Fourier coefficients", *Proc. Amer. Math. Soc.* **10** (1959), 855–859. MR 22 #6979 Zbl 0091.05706

[Rutter 2000] J. W. Rutter, *Geometry of curves*, Chapman & Hall/CRC, Boca Raton, FL, 2000. MR 2001e:53004 Zbl 0962.53002

[Saffari 1986] B. Saffari, "Une fonction extrémale liée à la suite de Rudin–Shapiro", *C. R. Acad. Sci. Paris Sér. I Math.* **303**:4 (1986), 97–100. MR 88a:11024 Zbl 0608.10051

[Saffari 1987] B. Saffari, "Structure algébrique sur les couples de Rudin–Shapiro: Problème extrémal de Salem sur les polynômes à coefficients ± 1", *C. R. Acad. Sci. Paris Sér. I Math.* **304**:5 (1987), 127–130. MR 88d:30008 Zbl 0608.10052

[Saffari 2001] B. Saffari, "Some polynomial extremal problems which emerged in the twentieth century", pp. 201–233 in *Twentieth century harmonic analysis—a celebration* (Il Ciocco, 2000), NATO Sci. Ser. II Math. Phys. Chem. **33**, Kluwer Acad. Publ., Dordrecht, 2001. MR 2002g:26001 Zbl 0996.42001

[Searle and Howard 2007] S. Searle and S. Howard, "A novel polyphase code for sidelobe suppression", pp. 377–381 in *Proceedings of IEEE Waveform Diversity and Design* (Pisa, Italy, 2007), IEEE, Los Alamitos, CA, 2007.

[Shapiro 1951] H. S. Shapiro, *Extremal problems for polynomials*, M.S. Thesis, Massachusetts Institute of Technology, 1951.

[Tseng and Liu 1972] C. C. Tseng and C. L. Liu, "Complementary sets of sequences", *IEEE Trans. Information Theory* **IT-18** (1972), 644–652. MR 53 #2511

[Turyn 1974] R. J. Turyn, "Hadamard matrices, Baumert–Hall units, four-symbol sequences, pulse compression, and surface wave encodings", *J. Combinatorial Theory Ser. A* **16** (1974), 313–333. MR 49 #10577 Zbl 0291.05016

[Vaidyanathan 1993] P. P. Vaidyanathan, *Multirate systems and filter banks*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[Welti 1960] G. R. Welti, "Quaternary codes for pulsed radar", *IRE Trans.* **6**:3 (1960), 400–408.

Received: 2009-03-24 Accepted: 2009-08-12

jjb@math.umd.edu *Norbert Wiener Center, Department of Mathematics, University of Maryland, College Park, MD 20742-4111, United States*

sugar@math.umd.edu *Norbert Wiener Center, Department of Mathematics, University of Maryland, College Park, MD 20742-4111, United States*

Some numerical radius inequalities for Hilbert space operators

Mohsen Erfanian Omidvar,
Mohammad Sal Moslehian and Assadollah Niknam

(Communicated by Kenneth S. Berenhaut)

We present several numerical radius inequalities for Hilbert space operators. More precisely, we prove that if $A, B, C, D \in B(H)$ and $T = \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$ then $\max(w(A), w(D)) \le \frac{1}{2}(\|T\| + \|T^2\|^{1/2})$ and $\max\big((w(BC))^{1/2}, (w(CB))^{1/2}\big) \le \frac{1}{2}(\|T\| + \|T^2\|^{1/2})$. We also show that if $A \in B(H)$ is positive, then

$$w(AX - XA) \le \tfrac{1}{2}\|A\|(\|X\| + \|X^2\|^{1/2}).$$

1. Introduction and preliminaries

Let $B(H)$ denote the C^*-algebra of all bounded linear operators on a complex Hilbert space H with inner product $\langle \cdot, \cdot \rangle$. For $A \in B(H)$ let

$$\begin{aligned} w(A) &= \sup\{|\langle x, Ax\rangle| : \|x\| = 1\}, \\ \|A\| &= \sup\{\|Ax\| : \|x\| = 1\}, \\ |A| &= (A^*A)^{1/2} \end{aligned}$$

denote the numerical radius, the usual operator norm of A and the absolute value of A. It is well known that $w(\cdot)$ is a norm on $B(H)$, and that for all $A \in B(H)$,

$$\tfrac{1}{2}\|A\| \le w(A) \le \|A\|. \tag{1-1}$$

Here are some basic properties of the numerical radius:

$$w(|A|) = \|A\|, \tag{1-2}$$

$$w(A^*A) = w(AA^*), \tag{1-3}$$

$$w(UAU^*) = w(A), \tag{1-4}$$

$$w(A_1 \oplus A_2 \oplus \cdots \oplus A_n) = \max\{w(A_i) : i = 1, 2, \ldots, n\}, \tag{1-5}$$

MSC2000: primary 47A62; secondary 46C15, 47A30, 15A24.

Keywords: bounded linear operator, Hilbert space, norm inequality, numerical radius, positive operator.

for all operators $A, A_1, A_2, \dots, A_n \in B(H)$ and all unitary operators $U \in B(H)$.

Suppose $H = M_1 \oplus M_2$ and $A \in B(H)$. Then we can write A as a block matrix

$$A = \begin{bmatrix} I_1^* A I_1 & I_1^* A I_2 \\ I_2^* A I_1 & I_2^* A I_2 \end{bmatrix}, \tag{1-6}$$

where $I_i \in B(M_i, H)$ such that $I_i(x) = x$ $(i = 1, 2)$. If A and B are operators in $B(H)$ we write the direct sum $A \oplus B$ for the 2×2 operator matrix $\left[\begin{smallmatrix} A & 0 \\ 0 & B \end{smallmatrix}\right]$, regarded as an operator on $H \oplus H$. Thus

$$\|A \oplus B\| = \left\| \begin{bmatrix} 0 & A \\ B & 0 \end{bmatrix} \right\| = \max(\|A\|, \|B\|). \tag{1-7}$$

Suppose $\mathscr{A} = A_1 \oplus A_2 \oplus \cdots \oplus A_n$, where $A_i \in B(H)$ and $x_1, x_2, \dots, x_n \in H$. That is,

$$\mathscr{A} = \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_n \end{bmatrix},$$

which we also write $\mathscr{A} = \operatorname{diag}(A_1, \dots, A_n)$. Then

$$\big\langle [x_1, \dots, x_n]^T, \mathscr{A}[x_1, \dots, x_n]^T \big\rangle = \sum_{i=1}^{n} \langle x_i, A_i(x_i) \rangle,$$

$$w(\mathscr{A}) = \sup\left\{ \left| \big\langle [x_1, \dots, x_n]^T, \mathscr{A}[x_1, \dots, x_n]^T \big\rangle \right| : \sum_{i=1}^{n} \|x_i\|^2 = 1 \right\}.$$

For additional properties of the numerical radius, see [Bhatia 1997; Halmos 1982] and references therein.

Consider $A = [A_{ij}]$, where $A_{ij} \in B(H)$ and $i, j = 1, 2, \dots, n$. We define $C(A) = A_{11} \oplus A_{22} \oplus \cdots \oplus A_{nn}$, called the *n-pinching* of A. We set $z = e^{2\pi i/n}$ and $U := \operatorname{diag}(I, zI, \dots, z^{n-1}I)$, where I is the identity operator in $B(H)$. Using the identity $\sum_{k=0}^{n-1} z^k = 0$, one can see that $C(A) = (1/n) \sum_{k=0}^{n-1} U^{*k} A U^k$ (see also [Bhatia 2000; 1997]).

It is shown in [Kittaneh 2005] that if $A, B, C, D, S, T \in B(H)$, then

$$w(ATB + CSD) \le \tfrac{1}{2}\big(\|A|T^*|^{2(1-\alpha)} A^* + B^*|T|^{2\alpha} B + C|S^*|^{2(1-\alpha)} C^* + D^*|S|^{2\alpha} D\|\big),$$

for all α with $0 \le \alpha \le 1$. In particular, if $A, U, P \in B(H)$ such that U is unitary

and P is projection, we have

$$w(AU \pm U^*A) \le \tfrac{1}{2}\big\||A| + |A^*| + U^*(|A| + |A^*|)U\big\| \le \|A\| + \|A^2\|^{1/2}, \quad (1\text{-}8)$$

$$w(AP - PA) \le \tfrac{1}{2}\big\||A| + |A^*| + P(|A| + |A^*|)P\big\| \le \|A\| + \|A^2\|^{1/2}, \quad (1\text{-}9)$$

$$w(A) \le \tfrac{1}{2}\big(\|A\| + \|A^2\|^{1/2}\big). \quad (1\text{-}10)$$

The last inequality refines the second inequality in (1-1); see also [Kittaneh 2003]. In [Kittaneh 2007; Bhatia and Kittaneh 2008] it is shown that if $A, B, X \in B(H)$ such that A and B are positive, then

$$|||AX - XB||| \le \max(\|A\|, \|B\|)|||X|||,$$

where $|||\cdot|||$ is a unitarily invariant norm.

In particular,

$$\|AX - XA\| \le \|A\|\|X\|. \quad (1\text{-}11)$$

In this paper we establish some inequalities sharper than inequalities (1-9) and (1-11) to the numerical radius and we give a new proof of inequality (1-10). Some applications of these inequalities are considered as well.

2. Main results

In [Bhatia 1997] it is shown that

$$\frac{1}{2}\left|\left|\left|\begin{bmatrix} A+B & 0 \\ 0 & A+B \end{bmatrix}\right|\right| \le \left|\left|\left|\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}\right|\right| \le \left|\left|\left|\begin{bmatrix} |A|+|B| & 0 \\ 0 & 0 \end{bmatrix}\right|\right|,$$

where $|||\cdot|||$ is a unitarily invariant norm. In this paper we extend this inequality to the numerical radius. We begin by establishing an interesting property of the numerical radius.

Lemma 2.1. *Let $A \in B(H)$. Then*

$$w(C(A)) \le w(A). \quad (2\text{-}1)$$

Proof. Since $C(A) = \dfrac{1}{n}\displaystyle\sum_{k=0}^{n-1} U^{*k}AU^k$, we have

$$w(C(A)) \le \frac{1}{n}\sum_{k=0}^{n-1} w(U^{*k}AU^k) = \frac{1}{n}\sum_{k=0}^{n-1} w(A) = w(A),$$

where the inequality follows from property (1-4). □

Theorem 2.2. *Let $A_1, A_2, \ldots, A_n \in B(H)$. Then*

$$\frac{1}{n}w\Big(\operatorname{diag}\Big(\sum_{i=1}^{n} A_i, \ldots, \sum_{i=1}^{n} A_i\Big)\Big) \le w(\mathcal{A}) \le w\Big(\operatorname{diag}\Big(\sum_{i=1}^{n} |A_i|, 0, \ldots, 0\Big)\Big).$$

Proof. For the first inequality, we have, using (1-5),

$$w\Big(\operatorname{diag}\Big(\sum_{i=1}^{n}A_i,\dots,\sum_{i=1}^{n}A_i\Big)\Big)=w\Big(\sum_{i=1}^{n}A_i\Big)\le\sum_{i=1}^{n}w(A_i)$$
$$\le n\max\{w(A_i):i=1,2,\dots,n\}=nw(\mathscr{A}).$$

For the second inequality first we suppose $A_1,A_2,\dots,A_n$ to be positive, so

$$w\left(\begin{bmatrix}\sum_{i=1}^{n}A_i & 0 & \cdots & 0\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 0\end{bmatrix}\right)=w\left(\begin{bmatrix}A_1^{1/2} & A_2^{1/2} & \cdots & A_n^{1/2}\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 0\end{bmatrix}\begin{bmatrix}A_1^{1/2} & 0 & \cdots & 0\\ A_2^{1/2} & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ A_n^{1/2} & 0 & \cdots & 0\end{bmatrix}\right)$$
$$=w\left(\begin{bmatrix}A_1^{1/2} & 0 & \cdots & 0\\ A_2^{1/2} & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ A_n^{1/2} & 0 & \cdots & 0\end{bmatrix}\begin{bmatrix}A_1^{1/2} & A_2^{1/2} & \cdots & A_n^{1/2}\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 0\end{bmatrix}\right)$$
$$=w\left(\begin{bmatrix}A_1 & A_1^{1/2}A_2^{1/2} & \cdots & A_1^{1/2}A_n^{1/2}\\ A_2^{1/2}A_1^{1/2} & A_2 & \cdots & A_2^{1/2}A_n^{1/2}\\ \vdots & \vdots & \ddots & \vdots\\ A_n^{1/2}A_1^{1/2} & A_n^{1/2}A_2^{1/2} & \cdots & A_n\end{bmatrix}\right),$$

where the second equality follows from (1-3). Using the inequality (2-1), we get

$$w\left(\begin{bmatrix}A_1 & 0 & \cdots & 0\\ 0 & A_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & A_n\end{bmatrix}\right)\le w\left(\begin{bmatrix}A_1 & A_1^{1/2}A_2^{1/2} & \cdots & A_1^{1/2}A_n^{1/2}\\ A_2^{1/2}A_1^{1/2} & A_2 & \cdots & A_2^{1/2}A_n^{1/2}\\ \vdots & \vdots & \ddots & \vdots\\ A_n^{1/2}A_1^{1/2} & A_n^{1/2}A_2^{1/2} & \cdots & A_n\end{bmatrix}\right)$$
$$=w\Big(\operatorname{diag}\Big(\sum_{i=1}^{n}A_i,0,\dots,0\Big)\Big),$$

Now let $A_1,A_2,\dots,A_n$ be arbitrary. Then

$$w\left(\begin{bmatrix}|A_1| & 0 & \cdots & 0\\ 0 & |A_2| & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & |A_n|\end{bmatrix}\right)\le w\Big(\operatorname{diag}\Big(\sum_{i=1}^{n}|A_i|,0,\dots,0\Big)\Big).$$

Since

$$w\left(\begin{bmatrix} |A_1| & 0 & \cdots & 0 \\ 0 & |A_2| & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & |A_n| \end{bmatrix}\right) = w\left(|\mathscr{A}|\right) \geq w\left(\mathscr{A}\right),$$

we have $w(\mathscr{A}) \leq w\big(\mathrm{diag}\big(\sum_{i=1}^{n} |A_i|, 0, \ldots, 0\big)\big)$. □

Corollary 2.3. *Let $A \in B(H)$. Then $\frac{1}{2}w((A + A^*) \oplus (A + A^*)) \leq w(A \oplus A^*)$.*

Kittaneh [2006] showed that if $A, B, C, D \in B(H)$ and if $T = \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$, then

$$\max(r(A), r(D)) \leq \tfrac{1}{2}(\|T\| + \|T^2\|^{1/2}), \quad (r(BC))^{1/2} \leq \tfrac{1}{2}(\|T\| + \|T^2\|^{1/2}).$$

We show similar inequalities for the numerical radius. To achieve this, we need the following lemma [Kittaneh 2005].

Lemma 2.4. *If $A, B \in B(H)$ and $AB = BA$, then $w(AB) \leq 2w(A)w(B)$.*

Theorem 2.5. *If $A, B, C, D \in B(H)$ and $T = \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$, then*

$$\max(w(A), w(D)) \leq \tfrac{1}{2}(\|T\| + \|T^2\|^{1/2}), \tag{2-2}$$

and

$$\max((w(BC))^{1/2}, (w(CB))^{1/2}) \leq \tfrac{1}{2}(\|T\| + \|T^2\|^{1/2}). \tag{2-3}$$

Proof.

By (1-5), we have $\max(w(A), w(D)) = w\left(\left[\begin{smallmatrix} A & 0 \\ 0 & D \end{smallmatrix}\right]\right)$. Since D is arbitrary,

$$\max(w(A), w(D)) = w\left(\begin{bmatrix} A & 0 \\ 0 & -D \end{bmatrix}\right).$$

Consider the unitary operator $U = \left[\begin{smallmatrix} I & 0 \\ 0 & -I \end{smallmatrix}\right]$ on $H \oplus H$. Then $2\left[\begin{smallmatrix} A & 0 \\ 0 & -D \end{smallmatrix}\right] = TU + UT$. Thus

$$\max(w(A), w(D)) \leq \tfrac{1}{2}(\|T\| + \|T^2\|^{1/2}),$$

by inequality (1-8). This proves the inequality (2-2).

To prove the inequality (2-3), we note that

$$\begin{aligned} \max(w(BC), w(CB)) &= w\left(\begin{bmatrix} BC & 0 \\ 0 & CB \end{bmatrix}\right) \quad \text{(by (1-5))} \\ &= w\left(\begin{bmatrix} 0 & B \\ C & 0 \end{bmatrix}^2\right) \\ &\leq 2w\left(\begin{bmatrix} 0 & B \\ C & 0 \end{bmatrix}\right)^2 \quad \text{(by Lemma 2.4).} \end{aligned}$$

Since B is arbitrary, we have

$$\max(w(BC), w(CB)) \le 2w\left(\begin{bmatrix} 0 & -B \\ C & 0 \end{bmatrix}\right)^2.$$

We observe that $2\left[\begin{smallmatrix} 0 & -B \\ C & 0 \end{smallmatrix}\right] = TU - UT$, so

$$\max\big((w(BC))^{1/2}, (w(CB))^{1/2}\big) \le \tfrac{1}{2}(\|T\| + \|T^2\|^{1/2})$$

by inequality (1-8). □

Corollary 2.6. *If $A \in B(H)$, then*

$$w(A) \le \tfrac{1}{2}(\|A\| + \|A^2\|^{1/2}) \le \|A\|.$$

Proof. Let $T = \left[\begin{smallmatrix} A & 0 \\ 0 & 0 \end{smallmatrix}\right]$. Then

$$\begin{aligned} w(A) &\le \tfrac{1}{2}(\|T\| + \|T^2\|^{1/2}) \quad \text{(by (2-2))} \\ &= \tfrac{1}{2}(\|A\| + \|A^2\|^{1/2}) \quad \text{(by (1-7))} \\ &\le \|A\|. \end{aligned}$$

□

Corollary 2.7. *If $A \in B(H)$, then $\|A + A^*\| \le \|A\| + \|A^2\|^{1/2} \le 2\|A\|$.*

Proof. Since $A + A^*$ is self-adjoint, we have

$$\begin{aligned} \tfrac{1}{2}\|A + A^*\| &= \tfrac{1}{2}w((A + A^*) \oplus (A + A^*)) && \text{(by (1-2) and (1-5))} \\ &\le w(A \oplus A^*) && \text{(by Corollary 2.3)} \\ &\le \tfrac{1}{2}\big(\|A \oplus A^*\| + \|(A \oplus A^*)^2\|^{1/2}\big) && \text{(by Corollary 2.6)} \\ &= \tfrac{1}{2}(\|A\| + \|A^2\|^{1/2}) && \text{(by (1-7))} \\ &\le \|A\|. \end{aligned}$$

□

We use some similar strategies as in [Kittaneh 2007] to prove the next two results.

Theorem 2.8. *Let $A, P \in B(H)$ such that P is a projection. Then*

$$w(AP - PA) \le \tfrac{1}{2}(\|A\| + \|A^2\|^{1/2}). \tag{2-4}$$

Proof. Using the decomposition $H = \operatorname{ran} P \oplus \ker P$ and equality (1-6), we represent P as the form $P = \left[\begin{smallmatrix} I_1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$, where I_1 is the identity operator on $\operatorname{ran} P$. With respect to this decomposition, A can be written as $A = \left[\begin{smallmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{smallmatrix}\right]$. Then

$$PA - AP = \begin{bmatrix} 0 & A_{12} \\ -A_{21} & 0 \end{bmatrix}.$$

If I_2 is the identity operator on $\ker P$ and if $U = \left[\begin{smallmatrix} I_1 & 0 \\ 0 & -I_2 \end{smallmatrix}\right]$, then U is unitary and $\left[\begin{smallmatrix} 0 & A_{12} \\ -A_{21} & 0 \end{smallmatrix}\right] = \frac{1}{2}(UA - AU)$. Therefore

$$w(AP - PA) = w\left(\begin{bmatrix} 0 & A_{12} \\ -A_{21} & 0 \end{bmatrix}\right) = \tfrac{1}{2} w(AU - U^*A) \le \tfrac{1}{2}(\|A\| + \|A^2\|^{1/2}),$$

where the inequality follows from (1-8). □

Theorem 2.9. *Suppose that $A \in B(H)$ is positive. Then*

$$w(AX - XA) \le \tfrac{1}{2}\|A\|(\|X\| + \|X^2\|^{1/2}). \tag{2-5}$$

Proof. First we prove that if A is positive and a contraction, then

$$w(AX - XA) \le \tfrac{1}{2}(\|X\| + \|X^2\|^{1/2}).$$

If $R = \sqrt{A - A^2}$, the operator

$$P = \begin{bmatrix} A & R \\ R & I - A \end{bmatrix}$$

is a projection on $H \oplus H$, because $A\sqrt{A - A^2} = \sqrt{A - A^2}A$. If $Y = \left[\begin{smallmatrix} X & 0 \\ 0 & 0 \end{smallmatrix}\right]$, then $PY - YP = \left[\begin{smallmatrix} AX - XA & -XR \\ RX & 0 \end{smallmatrix}\right]$. By the inequality (2-4), we have

$$w(YP - PY) \le \tfrac{1}{2}(\|Y\| + \|Y^2\|^{1/2}).$$

Now if $Q = \left[\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right]$, then $\left[\begin{smallmatrix} AX - XA & 0 \\ 0 & 0 \end{smallmatrix}\right] = Q(PY - YP)Q^*$, so

$$\begin{aligned} w\left(\begin{bmatrix} AX - XA & 0 \\ 0 & 0 \end{bmatrix}\right) &= w(YP - PY) && \text{(by (1-4))} \\ &\le \tfrac{1}{2}(\|Y\| + \|Y^2\|^{1/2}) && \text{(by (2-4))} \\ &= \tfrac{1}{2}(\|X\| + \|X^2\|^{1/2}) && \text{(by (1-7)),} \end{aligned}$$

whence $w(AX - XA) \le \frac{1}{2}(\|X\| + \|X^2\|^{1/2})$. Let A be a positive operator. It follows from the inequality

$$w\Big(\frac{A}{\|A\|}X - X\frac{A}{\|A\|}\Big) \le \frac{1}{2}(\|X\| + \|X^2\|^{1/2})$$

that $w(AX - XA) \le \frac{1}{2}\|A\|(\|X\| + \|X^2\|^{1/2})$. □

Corollary 2.10. *If $A, B \in B(H)$ such that A is positive and B is self-adjoint, then*

$$\|AB - BA\| \le \|A\|\|B\|. \tag{2-6}$$

Proof. The inequality (2-6) follows from (2-5) by letting $X = B$. □

Corollary 2.11. *Suppose that $T \in B(H)$ has the cartesian decomposition $T = A + iB$ such that A is positive and B is self-adjoint. Then*

$$\|T^*T - TT^*\| \leq \|A\|^2 + \|B\|^2.$$

Proof. By (2-6) and the arithmetic–geometric mean inequality, we have

$$\|T^*T - TT^*\| = 2\|AB - BA\| \leq 2\|A\|\|B\| \leq \|A\|^2 + \|B\|^2. \qquad \square$$

References

[Bhatia 1997] R. Bhatia, *Matrix analysis*, Grad. Texts in Math. **169**, Springer, New York, 1997. MR 98i:15003 Zbl 0863.15001

[Bhatia 2000] R. Bhatia, "Pinching, trimming, truncating, and averaging of matrices", *Amer. Math. Monthly* **107**:7 (2000), 602–608. MR 2001h:15020 Zbl 0984.15024

[Bhatia and Kittaneh 2008] R. Bhatia and F. Kittaneh, "Commutators, pinchings, and spectral variation", *Oper. Matrices* **2**:1 (2008), 143–151. MR 2392772 Zbl 1147.15019

[Halmos 1982] P. R. Halmos, *A Hilbert space problem book*, 2nd ed., Grad. Texts in Math. **19**, Springer, New York, 1982. MR 84e:47001 Zbl 0496.47001

[Kittaneh 2003] F. Kittaneh, "A numerical radius inequality and an estimate for the numerical radius of the Frobenius companion matrix", *Studia Math.* **158**:1 (2003), 11–17. MR 2004i:15022 Zbl 1113.15302

[Kittaneh 2005] F. Kittaneh, "Numerical radius inequalities for Hilbert space operators", *Studia Math.* **168**:1 (2005), 73–80. MR 2005m:47009 Zbl 1072.47004

[Kittaneh 2006] F. Kittaneh, "Spectral radius inequalities for Hilbert space operators", *Proc. Amer. Math. Soc.* **134**:2 (2006), 385–390. MR 2006d:47008 Zbl 1081.47010

[Kittaneh 2007] F. Kittaneh, "Inequalities for commutators of positive operators", *J. Funct. Anal.* **250**:1 (2007), 132–143. MR 2008j:47031 Zbl 1131.47009

Received: 2009-05-05 Accepted: 2009-07-01

erfanian@mshdiau.ac.ir *Department of Mathematics, Faculty of Science, Islamic Azad University-Mashhad Branch, Mashhad 91722, Iran*

moslehian@ferdowsi.um.ac.ir *Department of Pure Mathematics, Center of Excellence in Analysis on Algebraic Structures (CEAAS), Ferdowsi University of Mashhad, P.O. Box 1159, Mashhad 91775, Iran* http://www.um.ac.ir/~moslehian

dassamankin@yahoo.co.uk *Department of Pure Mathematics, Center of Excellence in Analysis on Algebraic Structures (CEAAS), Ferdowsi University of Mashhad, P.O. Box 1159, Mashhad 91775, Iran*

On the consistency of finite difference approximations of the Black–Scholes equation on nonuniform grids

Myles D. Baker and Daniel D. Sheng

(Communicated by Johnny Henderson)

The Black–Scholes equation has been used for modeling option pricing extensively. When the volatility of financial markets creates irregularities, the model equation is difficult to solve numerically; for this reason nonuniform grids are often used for greater accuracy. This paper studies the numerical consistency of popular explicit, implicit and leapfrog finite difference schemes for solving the Black–Scholes equation when nonuniform meshes are utilized. Mathematical tools including Taylor expansions are used throughout our analysis. The consistency ensures the basic reliability of the finite difference schemes based on choices of temporal and variable spatial derivative approximations. Truncation error terms are derived and discussed, and numerical experiments using C, C++ and Matlab are given to illustrate our discussions. We show that, though orders of accuracy are lower compared with their peers on uniform grids, nonuniform algorithms are easy to implement and use for turbulent financial markets.

1. An introduction of the differential equation and nonuniform grids

Let $f = f(x,t)$ be the value of an option in the option market. In [Brandimarte 2006; Wilmott et al. 1995], for example, we find the linearized Black–Scholes equation

$$B(f)(x,t) = \frac{\partial f}{\partial t}(x,t) + \alpha \frac{\partial^2 f}{\partial x^2}(x,t) + \beta \frac{\partial f}{\partial x}(x,t) - \gamma f(x,t) = 0, \quad t \geq 0, \quad (1\text{-}1)$$

where t is time, x is the asset price (or space variable), and $B(f)$ is the so-called Black–Scholes operator, expressed in terms of partial derivatives. The constants

MSC2000: primary 65G50, 65L12, 65N06, 65N15; secondary 65L70, 65L80, 65D25.

Keywords: finite difference approximation, difference algorithm, explicity, implicity, consistency, accuracy, matrix equations.

This collaborated research was supported by a Undergraduate Research and Scholarly Achievement Grant (no. 223-08-URSA) from Baylor University.

α, β and γ are nonnegative; they represent important parameters in economic and financial calculations.

Let $T > 0$ be a sufficiently large number. We consider a rectangular space-time domain D where $0 \le x \le 1$, $0 \le t \le T$ for (1-1). We further adopt an initial option value distribution

$$f(x, 0) = \sin(a\pi x), \quad 0 \le x \le 1, \tag{1-2}$$

as well as the Dirichlet boundary conditions

$$f(0, t) = \tfrac{1}{2}\sin(b\pi t),\ f(1, t) = \sin(a\pi(t + 1)), \quad t > 0, \tag{1-3}$$

where a and b are constants.

Since conventional difference approximations are consistent to the first derivative [Jain and Sheng 2007; Urban et al. 2004], we may adopt a uniform grid in the temporal direction, while maintaining nonuniform discretization in the x-direction. Let $\tau > 0$ be the temporal step size used, and $h_0, h_1, h_2, \ldots, h_n$ the spatial step sizes, where in general

$$h_i \ne h_{i+1}, \quad i \in \{0, 1, 2, \ldots, n - 1\},$$

and

$$\sum_{k=0}^{n} h_k = 1.$$

Thus, a two-dimensional nonuniform grid

$$D_{h,\tau} = \{(x_i, t_j) : x_i = x_{i-1} + h_{i-1},\ i = 1, 2, \ldots, n + 1;\ x_0 = 0,\ x_{n+1} = 1\}$$

is a discrete set over the domain D. Any $P_{i,j} = (x_i, t_j) \in D_{h,\tau}$ is called a grid point of $D_{h,\tau}$. It is an internal grid point if $i \ne 0, n+1$ and $j \ne 0$, a boundary point if $i \in \{0, n + 1\}$ and an initial point if $j = 0$.

In Figure 1 we show a particular two-dimensional nonuniform grid. In the design of nonuniform grids, we define the *smoothness ratios*

$$r_k = \frac{h_k}{h_{k-1}}, \quad k = 1, 2, \ldots, n,$$

and we avoid extreme r_k values since they may cause nonphysical oscillations of the numerical solutions inconsistent with the assumption that (1-1) adheres to the principles of geometric Brownian motion [Sheng 2008; Wilmott et al. 1995, §3.5, pp. 41–43].

For concreteness and simplicity, we will fix the parameter values

$$\alpha = 1, \quad \beta = 2, \quad \gamma = 0 \tag{1-4}$$

throughout our investigations.

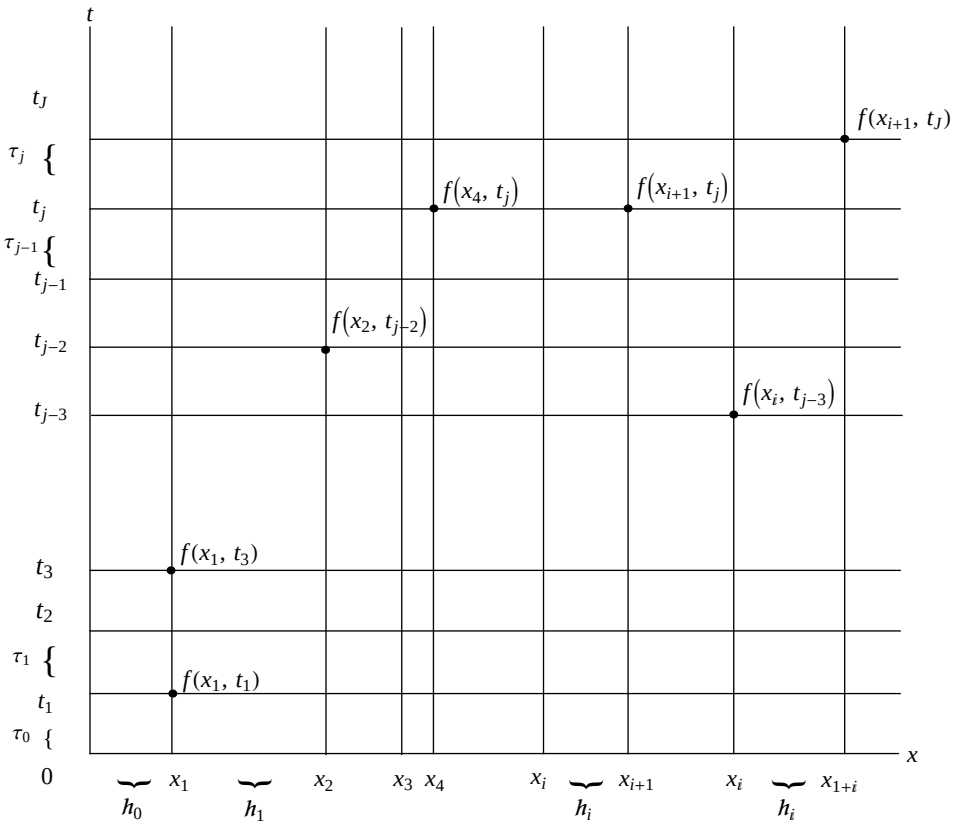


Figure 1. An illustration of the nonuniform grid $D_{h,\tau}$. A temporal step $\tau = 0.5$ is used. Particular spatial steps $h_0 = 0.15$, $h_1 = 0.08$, $h_2 = 0.07$, $h_3 = 0.06$, $h_4 = 0.02$, $h_5 = 0.02$, $h_6 = 0.01$, $h_7 = 0.02$, $h_8 = 0.08$, $h_9 = 0.09$, $h_{10} = 0.08$, $h_{11} = 0.12$, $h_{12} = 0.07$, $h_{13} = 0.1$ are used.

2. The explicit scheme

An *explicit scheme* is an algorithm which can be executed readily, or straightforwardly, without using a system of nonlinear solvers [Atkinson and Han 2004; Sheng 2008; Smith 1985]. In our case, unknowns can be evaluated by first finding the initial values (1-2) and boundary values to the left and right (1-3), and then using those values to arrive at the targeted $f_{i,j}$. We note that the scheme is being taken along a nonuniform grid rather than a uniform grid. In order to find values in the next temporal level, a recursive formula needs to be implemented. The application of computer software is very helpful in computing these f values, where else we would never have been able to see complex numerical results.

Denote the function value $f(x_i, t_j)$ by $f_{i,j}$. From the structure of $D_{h,\tau}$, we know that for any internal grid point $P_{i,j}$ we have

$$x_i = h_0 + h_1 + \cdots + h_{i-1}, \quad i \in \{1, 2, \ldots, n\}; \qquad t_j = j\tau, \quad j > 0.$$

According to [Atkinson and Han 2004, §4.1, p. 137; Jain and Sheng 2007], at $P_{i,j}$ we have

$$\left.\frac{\partial f}{\partial t}\right|_{i,j} \approx \frac{f_{i,j+1} - f_{i,j}}{\tau}, \tag{2-1}$$

$$\begin{aligned}\left.\frac{\partial^2 f}{\partial x^2}\right|_{i,j} &\approx \mathcal{D}_{2,x} f_{i,j} = \frac{D_{x,+} f_{i,j} - D_{x,-} f_{i,j}}{(h_{i-1} + h_i)/2} \\ &= \frac{2}{h_i(h_{i-1} + h_i)} f_{i+1,j} - \frac{1/h_i + 1/h_{i-1}}{(h_{i-1} + h_i)/2} f_{i,j} + \frac{2}{h_{i-1}(h_{i-1} + h_i)} f_{i-1,j} \\ &= \frac{2}{h_i(h_{i-1} + h_i)} f_{i+1,j} - \frac{2}{h_{i-1}h_i} f_{i,j} + \frac{2}{h_{i-1}(h_{i-1} + h_i)} f_{i-1,j}\end{aligned} \tag{2-2}$$

$$\left.\frac{\partial f}{\partial x}\right|_{i,j} \approx \frac{f_{i+1,j} - f_{i-1,j}}{h_{i-1} + h_i}. \tag{2-3}$$

Original concepts of the finite differences (2-1)–(2-3) can be traced back to calculations of variations in L^p spaces and beyond [Fonseca and Leoni 2007, §2.1.1].

Recall our choices $\alpha = 1$, $\beta = 2$, $\gamma = 0$. If we substitute (2-1)–(2-3) into (1-1) and remove the error terms, we acquire an explicit finite difference equation at $P_{i,j}$:

$$\begin{aligned}B_{h,\tau}(f_{i,j}) = \frac{f_{i,j+1} - f_{i,j}}{\tau} &- \frac{2}{h_{i-1} + h_i}\left(1 - \frac{1}{h_{i-1}}\right) f_{i-1,j} \\ &- \frac{2}{h_{i-1}h_i} f_{i,j} + \frac{2}{h_{i-1} + h_i}\left(1 + \frac{1}{h_i}\right) f_{i+1,j} = 0,\end{aligned}$$

where $B_{h,\tau}$ can be viewed as a discretized Black–Scholes operator. The above can be reformulated to the recursive formula

$$\begin{aligned}f_{i,j+1} &= f_{i,j} + \sigma_i\left[\left(1 - \frac{1}{h_{i-1}}\right) f_{i-1,j} + \frac{h_{i-1} + h_i}{h_{i-1}h_i} f_{i,j} - \left(1 + \frac{1}{h_i}\right) f_{i+1,j}\right], \\ \sigma_i &= \frac{2\tau}{h_{i-1} + h_i},\end{aligned}$$

or

$$\begin{aligned}f_{i,j+1} &= \sigma_i\left(1 - \frac{1}{h_{i-1}}\right) f_{i-1,j} + \left(1 + \frac{2\tau}{h_{i-1}h_i}\right) f_{i,j} - \sigma_i\left(1 + \frac{1}{h_i}\right) f_{i+1,j}, \\ \sigma_i &= \frac{2\tau}{h_{i-1} + h_i},\end{aligned} \tag{2-4}$$

which runs for the temporal level index j from 0 to J, as far as $J\tau \le T$. The numerical solution can thus be derived from the recursive relation (2-4) together

with the initial and boundary conditions (1-2), (1-3). This explicit finite difference scheme (2-4), (1-2), (1-3) is originally presented in [Sheng 2008].

To analyze (2-4) we need the following definitions:

Definition 2.1 (Order of accuracy). Consider a discretized Black–Scholes operator $B_{h,\tau}$, where $0 < h, \tau < 1$. Assume that the function $f(x,t)$ is sufficiently smooth (no cusps or discontinuities of partial derivatives up to a desired order) over D. If

$$\max_{(x_i,t_j)\in D_{h,\tau}} \left|B(f)(x_i,t_j) - B_{h,\tau}(f_{i,j})\right| = O(h^p + \tau^q), \tag{2-5}$$

where $h = \max\{h_1, h_2, \ldots, h_N\}$, we say that the difference scheme defined by $B_{h,\tau}$ is an *order*-(p,q) scheme for solving the Black–Scholes equation (1-1). We also say that $B_{h,\tau}$ is an *order-r* scheme, where $r = \min\{p,q\}$.

A difference method is practically meaningful when both p and q are positive.

Definition 2.2. A difference method $B_{h,\tau}$ is *consistent* if it has order $r > 0$.

There is always an error associated with a finite difference approximation. If we use Taylor's Theorem to expand certain known finite difference approximations along difference schemes, such as (2-4), we can empirically prove that these approximations may or may not be useful when applied to the Black–Scholes equation. For this purpose, we need the following notion:

Definition 2.3. The *truncation error function* of the difference scheme is defined as

$$\operatorname{err}(f)_{i,j} = B(f)(x_i,t_j) - B_{h,\tau}(f_{i,j}), \quad (x_i,t_j) \in D_{h,\tau}.$$

Theorem 2.4. *For the explicit finite difference scheme (2-4) we have the truncation error estimate*

$$\operatorname{err}_E(f)_{i,j} = O(h+\tau),$$

where $h = \max\{h_1, h_2, \ldots, h_N\}$. *Therefore the explicit scheme (2-4) is of first order.*

If the local spatial grid is uniform, that is, $h_i = h_{i-1} = h > 0$, *the scheme (2-4) is locally of second order in space*:

$$\operatorname{err}_E(f)_{i,j} = O(h^2+\tau).$$

Proof. We use the so-called forward, central and modified central difference operators, defined respectively by

$$\Delta_t f = \frac{f(t_{i,j+1}) - f(t_{i,j})}{\tau_{i,j+1} - \tau_{i,j}}, \quad \delta_x f = \frac{f(x_{i+1,j}) - f(x_{i-1,j})}{h_{i+1,j} - h_{i-1,j}}, \quad \mathcal{D}_{2,x} f = \frac{\Delta_x f - \nabla_x f}{(h_{i,j-1} + h_{i,j})/2},$$

for $i \in X$, $j \in T$. In the last expression Δ_x is defined like Δ_t, and the backward spatial difference operator ∇_x is similar, but with all spatial indices decreased by 1. In this notation we obtain from (2-4)

$$\mathrm{err}_E(f)_{i,j} = (f_t - \Delta_t f)_{i,j} + \alpha(f_{xx} - \mathscr{D}_{2,x} f)_{i,j} + \beta(f_x - \delta_x f)_{i,j}. \tag{2-6}$$

We have (see [Jain and Sheng 2007] for details)

$$\begin{aligned}
(f_t - \Delta_t f)_{i,j} &= -\tfrac{1}{2}\tau f_{tt}(x_i, t_j) - \tfrac{1}{6}\tau^2 f_{ttt}(x_i, t_j) + \cdots, \\
(f_{xx} - \mathscr{D}_{2,x} f)_{i,j} &= -\tfrac{1}{3}(h_i - h_{i-1}) f_{xxx}(x_i, t_j) \\
&\qquad - \tfrac{1}{12}(h_i^2 - h_i h_{i-1} + h_{i-1}^2) f_{x^4}(x_i, t_j) - \cdots, \\
(f_x - \delta_x f)_{i,j} &= -\frac{1}{2(h_i + h_{i-1})}\left(h_i^2 f_{xx}(x_i, t_j) - h_{i-1}^2 f_{xx}(x_i, t_j) + \cdots\right) \\
&= \tfrac{1}{2}(h_i - h_{i-1}) f_{xx}(x_i, t_j) + \tfrac{1}{6}(h_i^2 - h_i h_{i-1} + h_{i-1}^2) f_{xxx}(x_i, t_j) + \cdots.
\end{aligned}$$

Thus the lowest-order terms are linear in $\max\{h_i, h_{i+1}\}$ and τ_j; but when $h_i = h_{i+1} = h$, the linear contributions in h drop out. □

3. The implicit scheme

Instead of the forward finite difference (2-1), we may consider the backward difference formula

$$\left.\frac{\partial f}{\partial t}\right|_{i,j} = \frac{f_{i,j} - f_{i,j-1}}{\tau} + O(\tau). \tag{3-1}$$

Substitution of (3-1), (2-2), and (2-3) into (1-1) yields

$$\begin{aligned}
\frac{f_{i,j} - f_{i,j-1}}{\tau} &= \frac{2}{h_{i-1} + h_i}\left(1 - \frac{1}{h_{i-1}}\right) f_{i-1,j} \\
&\quad + \frac{2}{h_{i-1} h_i} f_{i,j} - \frac{2}{h_{i-1} + h_i}\left(1 + \frac{1}{h_i}\right) f_{i+1,j},
\end{aligned}$$

which is significantly different from the explicit scheme (2-4).

For later convenience we replace the index j by $j+1$ and $j-1$ by j. Our difference equation then becomes

$$\begin{aligned}
\frac{f_{i,j+1} - f_{i,j}}{\tau} &= \frac{2}{h_{i-1} + h_i}\left(1 - \frac{1}{h_{i-1}}\right) f_{i-1,j+1} \\
&\quad + \frac{2}{h_{i-1} h_i} f_{i,j+1} - \frac{2}{h_{i-1} + h_i}\left(1 + \frac{1}{h_i}\right) f_{i+1,j+1},
\end{aligned}$$

which can further be written as

$$-\sigma_i\left(1 - \frac{1}{h_{i-1}}\right) f_{i-1,j+1} - \left(\frac{2\tau}{h_{i-1} h_i} - 1\right) f_{i,j+1} + \sigma_i\left(1 + \frac{1}{h_i}\right) f_{i+1,j+1} = f_{i,j}, \tag{3-2}$$

where σ_i is the same as defined before. The implicit finite difference algorithm (3-2), together with conditions (1-2), (1-3), is studied in [Sheng 2008].

Equation (3-2) cannot be solved independently without collaboration between the rest of the equations at the temporal level $j+1$. We note that there are n internal

grid points. Thus we have n difference equations at each of the temporal levels, and we get a linear system of size n, which can be expressed in the matrix form as

$$Mf_{j+1} = g_{j+1}, \quad j \in \{1, 2, \ldots, J\}, \tag{3-3}$$

where M is a tridiagonal matrix with nontrivial elements

$$\begin{aligned} m_{i,i+1} &= -\sigma_i(1 + 1/h_i), \quad i = 1, 2, \ldots, n-1, \\ m_{i,i} &= 2\tau/(h_{i-1}h_i) - 1, \quad i = 1, 2, \ldots, n, \\ m_{i,i-1} &= \sigma_i(1 - 1/h_{i-1}), \quad i = 2, 3, \ldots, n. \end{aligned}$$

For the vectors we have

$$\begin{aligned} (f_{j+1})_i &= f_{i,j+1}, \quad i = 1, 2, \ldots, n, \\ (g_{j+1})_1 &= -f_{1,j} - \sigma_1(1 - 1/h_0) f_{0,j+1}, \\ (g_{j+1})_i &= -f_{i,j}, \quad i = 2, 3, \ldots, n-1, \\ (g_{j+1})_n &= -f_{n,j} + \sigma_n(1 + 1/h_n) f_{n+1,j+1}, \end{aligned}$$

where the values of $f_{0,j+1}$ and $f_{n+1,j+1}$ are given by the condition (1-3).

The tridiagonal system of linear equations (3-3) can be solved conveniently by using a special subroutine in C and Matlab; see [Jain and Sheng 2007; Pratap 1999; Sheng 2008].

Theorem 3.1. *For the implicit finite difference scheme (3-2) or (3-3) we have the truncation error estimate*

$$\mathrm{err}_I(f)_{i,j} = O(h + \tau),$$

where $h = \max\{h_1, h_2, \ldots, h_N\}$. *Therefore the implicit scheme is of first order.*

If the local spatial grid region is uniform, that is, $h_i = h_{i-1} = h > 0$, *the scheme is locally of second order in space*:

$$\mathrm{err}_I(f)_{i,j} = O(h^2 + \tau).$$

Proof. We have

$$\mathrm{err}_I(f)_{i,j} = (f_t - \nabla_t f)_{i,j} + \alpha(f_{xx} - \mathfrak{D}_{2,x} f)_{i,j} + \beta(f_x - \delta_x f)_{i,j}. \tag{3-4}$$

The α and β terms are as in the proof of Theorem 2.4, while for the remaining term we have (from [Jain and Sheng 2007], for instance).

$$(f_t - \nabla_t f)_{i,j} = \tfrac{1}{2}\tau f_{tt}(x_i, t_j) - \tfrac{1}{6}\tau^2 f_{ttt}(x_i, t_j) + \cdots.$$

Substitution into (3-4) and consideration of the special case $h_i = h_{i-1}$ yields the result. □

4. The leapfrog scheme

In this much more sophisticated approach, we first replace (2-1) by the central difference formula

$$\frac{\partial f}{\partial t}\Big|_{i,j} = \frac{f_{i,j+1} - f_{i,j-1}}{2\tau} + O(\tau^2). \tag{4-1}$$

We immediately notice the increase in the order of approximation. Now, instead of (2-2), (2-3) for the spatial derivatives, we consider the average formulas

$$\begin{aligned}\frac{\partial^2 f}{\partial x^2}\Big|_{i,j} &= \frac{1}{2}\Big(\frac{D_{x,+}f_{i,j-1} - D_{x,-}f_{i,j-1}}{(h_{i-1}+h_i)/2} + \frac{D_{x,+}f_{i,j+1} - D_{x,-}f_{i,j+1}}{(h_{i-1}+h_i)/2}\Big) + O(h)\\ &= \frac{1}{h_i(h_{i-1}+h_i)}(f_{i+1,j-1} + f_{i+1,j+1}) - \frac{1}{h_{i-1}h_i}(f_{i,j-1} + f_{i,j+1})\\ &\quad + \frac{1}{h_{i-1}(h_{i-1}+h_i)}(f_{i-1,j-1} + f_{i-1,j+1}) + O(h),\end{aligned} \tag{4-2}$$

$$\frac{\partial f}{\partial x}\Big|_{i,j} = \frac{1}{2}\Big(\frac{f_{i+1,j-1} - f_{i-1,j-1}}{h_{i-1}+h_i} + \frac{f_{i+1,j+1} - f_{i-1,j+1}}{h_{i-1}+h_i}\Big) + O(h), \tag{4-3}$$

where $h = \max_k\{h_k\}$. The order of approximation of the spatial derivatives is still 1 due to the nonuniform grid.

Substitution of (4-1)–(4-3) into (1-1) yields

$$\begin{aligned}\frac{f_{i,j+1} - f_{i,j-1}}{2\tau} &= \frac{1}{h_{i-1}+h_i}\Big(1 - \frac{1}{h_i}\Big)(f_{i-1,j-1} + f_{i-1,j+1}) + \frac{1}{h_{i-1}h_i}(f_{i,j-1} + f_{i,j+1})\\ &\quad - \frac{1}{h_{i-1}+h_i}\Big(1 + \frac{1}{h_i}\Big)(f_{i+1,j-1} + f_{i+1,j+1}).\end{aligned}$$

Let

$$\sigma_i = \frac{2\tau}{h_{i-1}+h_i}.$$

Then the difference equation can be rearranged as

$$\begin{aligned}f_{i,j+1} &= f_{i,j-1} + \sigma_i\Big(1 - \frac{1}{h_i}\Big)(f_{i-1,j-1} + f_{i-1,j+1})\\ &\quad + \frac{2\tau}{h_{i-1}h_i}(f_{i,j-1} + f_{i,j+1}) - \sigma_i\Big(1 + \frac{1}{h_i}\Big)(f_{i+1,j-1} + f_{i+1,j+1}),\end{aligned}$$

which leads to

$$\begin{aligned}&\sigma_i\Big(1 - \frac{1}{h_i}\Big)f_{i-1,j+1} + \Big(\frac{2\tau}{h_{i-1}h_i} - 1\Big)f_{i,j+1} - \sigma_i\Big(1 + \frac{1}{h_i}\Big)f_{i+1,j+1}\\ &\quad = -\sigma_i\Big(1 - \frac{1}{h_i}\Big)f_{i-1,j-1} - \Big(\frac{2\tau}{h_{i-1}h_i} + 1\Big)f_{i,j-1} + \sigma_i\Big(1 + \frac{1}{h_i}\Big)f_{i+1,j-1}.\end{aligned} \tag{4-4}$$

Like(3-2), the system of linear equations (4-4) can be put into matrix form:

$$Pf_{j+1} = Qf_{j-1} + s_{j+1}, \tag{4-5}$$

where P, Q are $n \times n$ tridiagonal matrices and s_{j+1} can be determined via the boundary condition (1-3). Therefore (4-5) can be readily solved by computers [Atkinson and Han 2004; Sheng 2008; Smith 1985].

The leapfrog scheme (4-4), or (4-5), is implicit since it must be solved as a linear system.

A peculiarity of the leapfrog method is that if we start from the initial value at $t = 0$, we can only obtain the numerical solution of (1-1)–(1-3) on even temporal levels. To compute numerical solutions on odd temporal levels, we need solution values on the first temporal level, which can be generated by using one step via either the explicit method or implicit method.

Theorem 4.1. *For the leapfrog implicit finite difference scheme (4-4) or (4-5) we have the truncation error estimate*:

$$\mathrm{err}_L(f)_{i,j} = O(h + \tau^2),$$

where

$$h = \max\{h_1, h_2, \ldots, h_N\}.$$

Therefore the leapfrog scheme is of first order in space and second order in time.

If the local spatial grid region is uniform, that is, $h_i = h_{i-1} = h > 0$, the leapfrog scheme becomes a second order method locally:

$$\mathrm{err}_L(f)_{i,j} = O(h^2 + \tau^2).$$

Proof. For the leapfrog scheme (4-4) or (4-5),

$$\begin{aligned}\mathrm{err}_L(f)_{i,j} = (f_t - \delta_t f)_{i,j} + \alpha\big((f_{xx})_{i,j} - \tfrac{1}{2}(\mathcal{D}_{2,x} f_{i,j-1} + \mathcal{D}_{2,x} f_{i,j+1})\big) \\ + \beta\big((f_x)_{i,j} - \tfrac{1}{2}(\delta_x f_{i,j-1} + \delta_x f_{i,j+1})\big).\end{aligned} \tag{4-6}$$

Again, according to [Jain and Sheng 2007; Sheng 2008], we have

$$(f_t - \delta_t f)_{i,j} = -\tfrac{1}{6}\tau^2 f_{ttt}(x_i, t_j) - \tfrac{1}{120}\tau^4 f_{t^5}(x_i, t_j) - \cdots. \tag{4-7}$$

Note that, since the grid distribution in space is irregular, (4-7) cannot be extended for estimating the difference $(f_x)_{i,j} - \frac{1}{2}(\delta_x f_{i,j-1} + \delta_x f_{i,j+1})$. Instead, employing the expansion

$$\delta_x f_{i,j-1} = (f_x)_{i,j-1} + \tfrac{1}{2}(h_i - h_{i-1})(f_{xx})_{i,j-1} + \tfrac{1}{6}(h_i^2 - h_i h_{i-1} + h_{i-1}^2)(f_{xxx})_{i,j-1} + \cdots$$

and performing simplifications, we obtain

$$\begin{aligned}(f_x)_{i,j} - \tfrac{1}{2}(\delta_x f_{i,j-1} + \delta_x f_{i,j+1}) \\ = -\tfrac{1}{4}(h_i - h_{i-1})\big((f_{xx})_{i,j-1} + (f_{xx})_{i,j+1}\big) \\ -\tfrac{1}{12}(h_i^2 - h_i h_{i-1} + h_{i-1}^2)\big((f_{xxx})_{i,j-1} + (f_{xxx})_{i,j+1}\big) + O(h^3). \quad (4\text{-}8)\end{aligned}$$

For the α term in (4-7) we will assume for simplicity that the grids $\{x_i\}$ remain the same on different temporal levels. Using the expansion

$$\mathscr{D}_{2,x} f_{i,j-1} = \tfrac{1}{3}(h_i - h_{i-1})3(f_{xxx})_{i,j-1} + \tfrac{1}{12}(h_i^2 - h_i h_{i-1} + h_{i-1}^2)(f_{x^4})_{i,j-1} + \cdots$$

and simplifying, we obtain

$$\begin{aligned}(f_{xx})_{i,j} - \tfrac{1}{2}(\mathscr{D}_{2,x} f_{i,j-1} + \mathscr{D}_{2,x} f_{i,j+1}) \\ = -\tfrac{1}{6}(h_i - h_{i-1})\big((f_{xxx})_{i,j-1} + 3(f_{xxx})_{i,j} + (f_{xxx})_{i,j+1}\big) \\ -\tfrac{1}{24}(h_i^2 + h_{i-1}^2)\big((f_{x^4})_{i,j-1} + 6(f_{x^4})_{i,j} + (f_{xxx})_{i,j+1}\big) \\ +\tfrac{1}{24}h_i h_{i-1}\big((f_{x^4})_{i,j-1} + (f_{x^4})_{i,j+1}\big) + O(h^3). \quad (4\text{-}9)\end{aligned}$$

Substituting (4-7)–(4-9) into (4-6) we get for $\mathrm{err}_L(f)_{i,j}$ the expression

$$\begin{aligned}-\tfrac{1}{6}\tau^2 f_{ttt}(x_i, t_j) + \alpha\Big[\tfrac{1}{6}(h_i - h_{i-1})\big((f_{xxx})_{i,j-1} + 3(f_{xxx})_{i,j} + (f_{xxx})_{i,j+1}\big) \\ +\tfrac{1}{24}(h_i^2 + h_{i-1}^2)\big((f_{x^4})_{i,j-1} + 6(f_{x^4})_{i,j} + (f_{x^4})_{i,j+1}\big) \\ +\tfrac{1}{24}h_i h_{i-1}\big((f_{x^4})_{i,j-1} + (f_{x^4})_{i,j+1}\big)\Big] \\ +\beta\Big[\tfrac{1}{4}(h_i - h_{i-1})\big((f_{xx})_{i,j-1} + (f_{xx})_{i,j+1}\big) \\ +\tfrac{1}{12}(h_i^2 - h_i h_{i-1} + h_{i-1}^2)\big((f_{xxx})_{i,j-1} + (f_{xxx})_{i,j+1}\big)\Big] \\ + O(h^3 + \tau^4),\end{aligned}$$

which is generally of first order in space and second order in time, but becomes of second order in both time and space when $h_i = h_{i-1}$. □

Since a leapfrog scheme spans three temporal levels, the computation using (4-4) or (4-5) can be started initially. One strategy is to use an implicit or an explicit scheme for calculating the numerical solution at the first temporal level, that is, when $j = 1$. Then, by using the numerical solutions at temporal levels 0 and 1, a leapfrog scheme can generate solutions at higher temporal levels.

However, this treatment may reduce the overall order of accuracy, given that an explicit or implicit method is used to generate the solution on the first temporal level. The computer club members had several fruitful discussions on the issue. At the end, we realized that such a computation is not necessary. Let τ be halved. Why not collect numerical solutions on the even number of temporal levels only?

This will guarantee our overall accuracy. We may introduce *imaginary middle temporal grid points* [Atkinson and Han 2004; Sheng 2008; Smith 1985] in the analysis, which may help to retain the overall second order accuracy in time t. We will report details elsewhere.

5. Simulation results

In our numerical experiments, we consider the following simplified Black–Scholes initial-boundary value problem [Brandimarte 2006]:

$$\frac{\partial f}{\partial t}(x,t) = -\frac{\partial^2 f}{\partial x^2}(x,t) - 2\frac{\partial f}{\partial x}(x,t), \quad 0 \le x \le 1,\ t \ge 0, \tag{5-1}$$

$$f(x,0) = \sin(2.2\pi x), \qquad 0 \le x \le 1, \tag{5-2}$$

$$\begin{aligned} f(0,t) &= 0.5\sin(7\pi t), \\ f(1,t) &= \sin(2.2\pi t), \end{aligned} \qquad t > 0. \tag{5-3}$$

For the sake of simplicity, we will only provide numerical results from the explicit scheme (2-4). Results from the other two schemes are similar.

Our nonuniform grid region is designed as follows: Let $N = 100$ and use C and Matlab programs to generate N random numbers, $x_1 < x_2 < x_3 < \cdots < x_N$, on the interval $(0, 1)$. Denote $x_0 = 0, x_{N+1} = 1$. We have the set of nonuniform spatial step sizes:

$$h_i = x_i - x_{i-1}, \quad i = 1, 2, \ldots, N+1.$$

Now, set $\tau = 1/100$. Thus a nonuniform two-dimensional grid region is completed. In Figure 2, we show the spatial step sizes across the interval $[0, 1]$. We also show the ratio distribution of the neighboring spatial step sizes, h_{i+1}/h_i, $i = 1, 2, \ldots, N$,

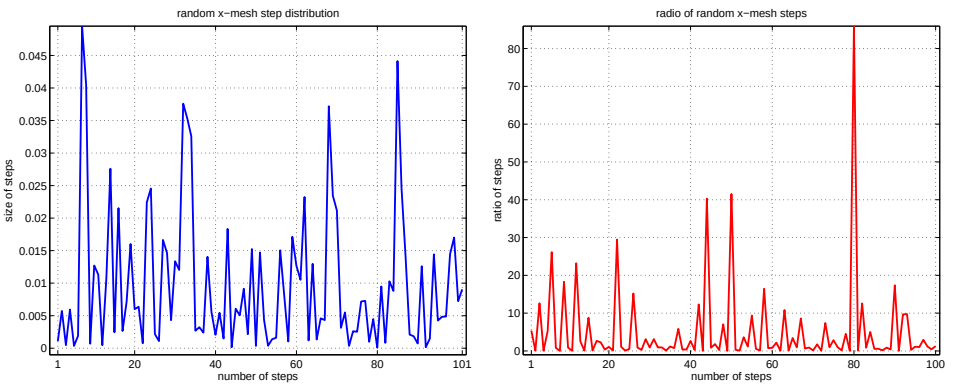


Figure 2. Left: distribution of the random spatial step sizes. Right: ratio function value distribution across the interval $[0, 1]$.

since the values are important in adaptive computations [Jain and Sheng 2007]. Note the large peak value of the ratio function near $i = 80$, exemplifying the so-called nonsmoothness phenomenon of random spatial grids [Jain and Sheng 2007; Sheng 2008]. Because the step sizes are random, each numerical experiment with the same Black–Scholes problem (5-1)–(5-3) yields a different Figure 2.

We choose the smooth test function $f(x,t) = \sin(\pi x)e^{-0.22t}$, as in [Atkinson and Han 2004; Jain and Sheng 2007]. Its partial derivatives are readily calculated. Figure 3 shows illustrative plots of the finite difference approximations for f and its derivatives. The $\mathcal{D}_{2,x}$ formula is used in approximating the second derivative since repeated use of conventional first order differences does not yield a consistent approximation of the second derivative [Fonseca and Leoni 2007; Jain and Sheng 2007]. For ease of comparison, all function surfaces are plotted from the same viewpoint.

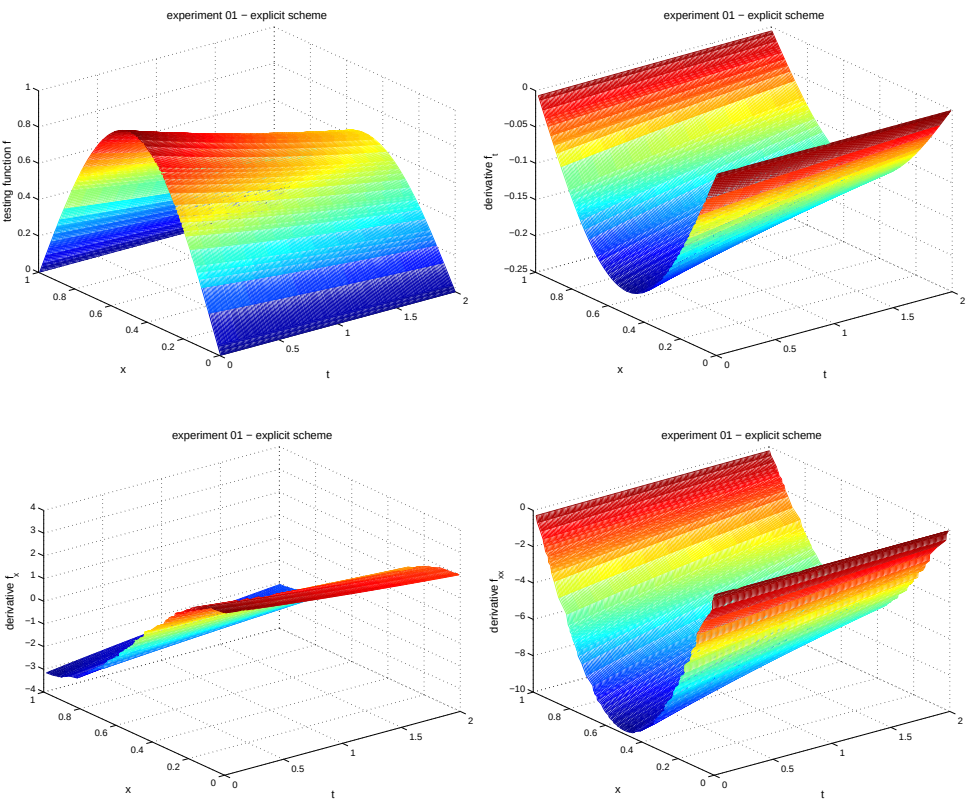


Figure 3. Clockwise from top left: graphs of $f(x,t)$, $f_t(x,t)$, $f_{xx}(x,t)$ (similar to $f_t(x,t)$) and $f_x(x,t)$. Second derivatives use the $\mathcal{D}_{2,x}$ formula.

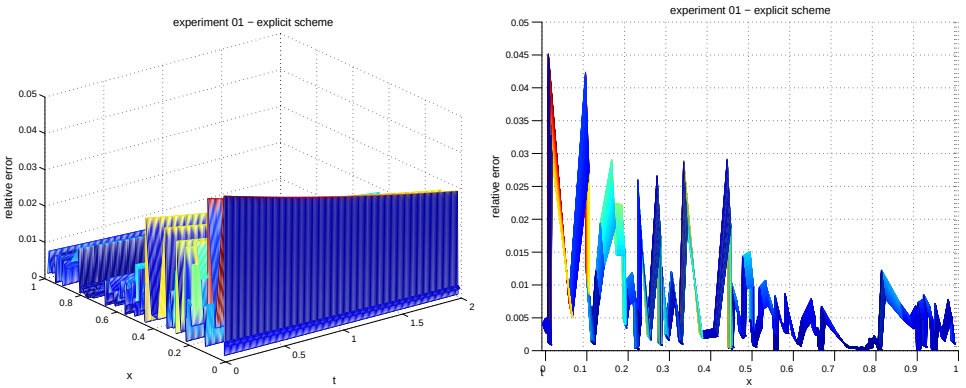


Figure 4. Left: three-dimensional distribution of the relative numerical error. Right: projection of the graph onto the XZ plane.

Consider the explicit scheme (2-4). To check the numerical truncation error of the algorithm, we submit the function $f(x,t)$ into (2-6) and evaluate the outcome on all internal nonuniform grid points. The relative error is adopted for a more reasonable evaluation of the errors [Atkinson and Han 2004].

Figure 4 gives the error distribution over the nonuniform grids of the domain $0 \le x \le 1$, $0 \le t \le 2$. In the first figure, we show a three-dimensional relative error distribution. The second figure represents the numerical error projected onto the XZ plane so that we can view more clearly the oscillatory features of the error pattern (probably due to the random spatial steps used). The overall numerical error is oscillatory but small. The maximal relative error appears to be approximately 0.045, that is, 4.5%, which is satisfactory. It is interesting that the error pattern does not seem to be similar to the mesh pattern given in Figure 2, though they must be related. A more in-depth numerical analysis may be required to reveal their internal connections.

The numerical experiments for estimating the order of convergence is more complex and tricky too, since the feature of two-dimensional problems (5-1)–(5-3) and the nonuniform grids used. Our experiments are based on the following stages:

(1) For a given testing function $f(x,t)$, let the numerical truncation error function be $(\mathrm{err}_0)_{i,j}$. Since $h_i = O(\tau)$, we may assume that

$$\left|(\mathrm{err}_0)_{i,j}\right| \approx M\tau^p, \tag{5-4}$$

where M is a positive constant, for all valid indexes i, j.

(2) Halve both the spatial and temporal step sizes. This is easy to achieve in time but relatively tricky in space. For the sake of simplicity, we may halve each of the h_i generated, although this yields pairs of identical step sizes.

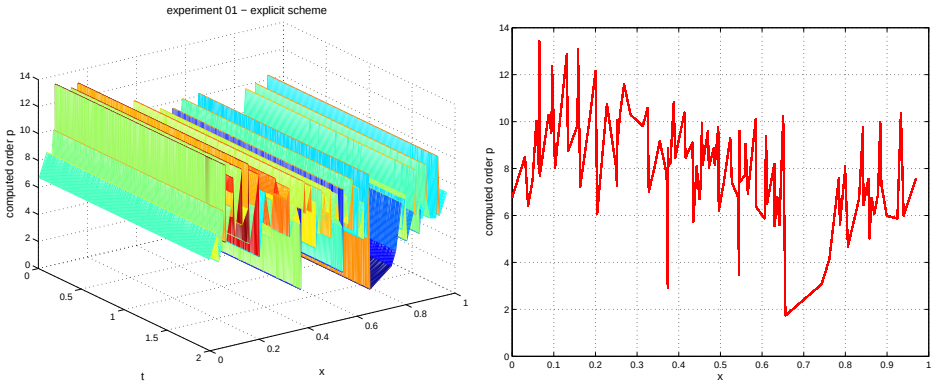


Figure 5. Left: three-dimensional distribution of the order function p. Right: contour map of the function p over the region $0 \le x \le 1,\ 0 \le t \le 2$.

(3) Repeat the computation on the refined grid region. Suppose that the new numerical truncation error function is $(\mathrm{err}_1)_{i,j}$, where the indexes are taken only on grid points where $(\mathrm{err}_0)_{i,j}$ is defined. Thus, we have

$$\left|(\mathrm{err}_1)_{i,j}\right| \approx M(\tau/2)^p \tag{5-5}$$

for all such indexes i, j.

(4) From (5-4), (5-5) we deduce that $\left|\dfrac{(\mathrm{err}_0)_{i,j}}{(\mathrm{err}_1)_{i,j}}\right| \approx 2^p$, which offers an estimate

$$p \approx \frac{1}{\ln 2} \ln \left|\frac{(\mathrm{err}_0)_{i,j}}{(\mathrm{err}_1)_{i,j}}\right|. \tag{5-6}$$

Evidently, such a p is also a function of i, j. Therefore an average value of such p values would provide a more reasonable estimate of the order for the underlying numerical method.

The process may be repeated by halving the step sizes again. However, note that (5-6) provides only an estimate which can be used as a reference.

Figure 5 demonstrates a solution surface of the p values obtained via (5-6). It is interesting to notice that

$$\max_{i,j} p = 13.4235, \quad \min_{i,j} p = 1.7357, \quad \text{average}\,(p) = 8.1013.$$

The numerical results seem to be much higher than the linear truncation error predicted by Theorem 2.4 in the situation. Most of our randomly chosen x-grids have demonstrated a similar conclusion. However, since the testing function is artificially chosen and the grid points in the x-direction are randomly chosen, we cannot conclude that the actual truncation order is much better than predicted. The

experiment results only indicate a strong possibility that our numerical scheme may behave better than anticipated.

6. Conclusion

Financial confidence is of the utmost importance when making investments, and in this paper we analyze the consistency of explicit, implicit and leapfrog finite difference schemes for solving Black–Scholes partial differential equation. We show the potential these numerical methods have for making accurate predictions of the option values when nonuniform discretizations are necessary. By using the $\mathcal{D}_{2,x}$ difference formula, we prove that all difference methods developed are consistent. While the explicit and implicit schemes provide a truncation error of the order $O(\tau+h)$, the leapfrog scheme offers a higher order $O(\tau^2+h)$. More precise error estimates for the three numerical methods are delivered in the corollaries for closer comparisons. Numerical experiments based on the explicit finite difference scheme have demonstrated a satisfactory result, indicating their good potential use in real-world implementations.

The orders of approximation can be further improved, but, the use of more sophisticated finite difference formulas may lead to complicated numerical schemes which are either difficult to use or difficult to analyze. The benefit of *order improvement* may be limited in actual financial computations, though its theory in numerical investigations is always meaningful. The exploration of better, higher-order numerical schemes for solving the Black–Scholes equation is one of the goals in our forthcoming study.

There are many interesting problems to be explored for the finite difference schemes implemented. A particularly relevant issue is numerical stability in the von Neumann sense [Atkinson and Han 2004; Sheng 2008; Smith 1985]. This concern focuses on the question: once a tiny error is introduced during computations, will it affect significantly further option values of f? If such a consequence is unavoidable, then is there a strategy to reduce the damage? We prefer to leave the answers to our forthcoming investigations. We also encourage the reader to explore any possible solutions.

Acknowledgments

The authors are grateful to Dr. Johnny Henderson and Dr. Lance Littlejohn, professors of mathematics at Baylor University, and Mr. Andrew D. Sheng, undergraduate student of computer science at Carnegie Mellon University, for reading our manuscripts, constructing and experimenting many programs via C, C++ and Matlab languages, and for sharing important research notes and thoughts. Our

special thanks go to Baylor University for its Undergraduate Research and Scholarly Achievement (URSA) grant, which enabled this collaborative research. We also thank our project advisor, Dr. Qin Sheng, professor of mathematics at Baylor University, for suggesting this line of undergraduate research and for the encouragement and numerous discussions throughout the project. Last, but not least, we appreciate our reviewers for their comments and suggestions that helped to improve the contents of this paper.

References

[Atkinson and Han 2004] K. Atkinson and W. Han, *Elementary numerical analysis*, 3rd ed., Wiley, Somerset, NJ, 2004. Zbl 0782.65002

[Brandimarte 2006] P. Brandimarte, *Numerical methods in finance and economics: A MATLAB-based introduction*, 2nd ed., Wiley, Hoboken, NJ, 2006. MR 2007d:91001 Zbl 1129.91002

[Fonseca and Leoni 2007] I. Fonseca and G. Leoni, *Modern methods in the calculus of variations: L^p spaces*, Springer, New York, 2007. MR 2008m:49001 Zbl 1153.49001

[Jain and Sheng 2007] B. Jain and A. D. Sheng, "An exploration of the approximation of derivative functions via finite differences", *Rose-Hulman Undergraduate Math Journal* **8** (2007), 172–188.

[Pratap 1999] R. Pratap, *Getting started with MatLab 5*, Oxford University Press, Oxford and New York, 1999.

[Sheng 2008] A. D. Sheng, "Optimized finite difference approximations on non-uniform grids with applications: A midterm project report to the URSA program", preprint, Baylor University, 2008, Available at http://www.baylor.edu/research/ursa/index.php?id=50690.

[Smith 1985] G. D. Smith, *Numerical solution of partial differential equations: Finite difference methods*, 3rd ed., Oxford App. Math. and Computing Science Series **13**, Oxford Univ. Press, New York, 1985. MR 87c:65002 Zbl 0576.65089

[Urban et al. 2004] P. Urban, J. Owen, D. Martin, R. Haese, S. Haese, and M. Bruce, *Mathematics for the international student : Mathematics HL (core)*, Haese & Harris Publications, Adelaide Airport, Australia, 2004.

[Wilmott et al. 1995] P. Wilmott, S. Howison, and J. Dewynne, *The mathematics of financial derivatives: A student introduction*, Cambridge University Press, Cambridge, 1995. MR 96h:90028 Zbl 0842.90008

Received: 2009-06-04 Revised: 2009-07-23 Accepted: 2009-07-27

mylesdanielbaker@gmail.com *Department of Mathematics, Baylor University, Waco, TX 76798-7328, United States*
http://www.baylor.edu/math/

danieldsheng@gmail.com *Westwood High School, Austin, TX 78750, United States*
http://schools.roundrockisd.org/Westwood/

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
2009 vol. 2 no. 4