involve

a journal of mathematics

msp

2013 vol. 6, no. 4

involve

msp.org/involve

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2013 is US $105/year for the electronic version, and $145/year (+$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to MSP.

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

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

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

INVOLVE 6:4 (2013)
dx.doi.org/10.2140/involve.2013.6.383

Embeddedness for singly periodic Scherk surfaces with higher dihedral symmetry

Valmir Bucaj, Sarah Cannon, Michael Dorff,
Jamal Lawson and Ryan Viertel

(Communicated by Frank Morgan)

The singly periodic Scherk surfaces with higher dihedral symmetry have $2n$-ends that come together based upon the value of φ. These surfaces are embedded provided that $\frac{\pi}{2} - \frac{\pi}{n} < \frac{n-1}{n}\varphi < \frac{\pi}{2}$. Previously, this inequality has been proved by turning the problem into a Plateau problem and solving, and by using the Jenkins–Serrin solution and Krust's theorem. In this paper we provide a proof of the embeddedness of these surfaces by using some results about univalent planar harmonic mappings from geometric function theory. This approach is more direct and explicit, and it may provide an alternate way to prove embeddedness for some complicated minimal surfaces.

1. Introduction

A minimal surface in $\mathbb{R}^3$ is a surface whose mean curvature vanishes at each point on the surface. One area of minimal surface theory that has seen a lot of interest and results recently is the study of complete embedded minimal surfaces. Minimal surfaces can be parametrized by the classical Weierstrass representation. However, these surfaces are not guaranteed to be complete and embedded. In this paper we will consider the family of singly periodic Scherk surfaces with higher dihedral symmetry that were first described in the seminal paper [Karcher 1988]. They belong to the larger class of embedded singly periodic minimal surfaces with Scherk ends and genus 0 in the quotient that have been completely classified in [Pérez and Traizet 2007]. The singly periodic Scherk surfaces with higher dihedral symmetry have $2n$-ends that come together based upon the value of φ. In particular, it was shown in [Weber 2005] that these surfaces are embedded provided that

$$\frac{\pi}{2} - \frac{\pi}{n} < \frac{n-1}{n}\varphi < \frac{\pi}{2}. \tag{1}$$

MSC2010: 30C45, 49Q05, 53A10.
Keywords: minimal surfaces, harmonic mappings, Scherk, univalence.
Part of this research was done during the 2010 BYU REU supported by NSF grant DMS-0755422.

Previously, this inequality has been established by turning the problem into a Plateau problem and solving, and by using the Jenkins–Serrin solution and Krust's theorem. In this paper, we will provide a proof of the embeddedness of these surfaces by using some results about univalent planar harmonic mappings from geometric function theory. This approach is more direct and explicit, and it may provide an alternate way to prove embeddedness for some complicated minimal surfaces. In the interesting paper [McDougall and Schaubroeck 2008], the authors discuss similar harmonic mappings and the corresponding minimal surfaces. They also work to prove an inequality similar to (1). While their approach is sound, there are unfortunately several small mistakes and errors, and the inequality they give is incorrect and different from the result in [Weber 2005]. In our paper, we start with planar harmonic mappings but then approach the proof of the inequality in a different way and derive the correct inequality given by (1).

This approach involves the following steps. First, we will construct a φ-variable family of planar harmonic functions that map the unit disk univalently onto a $2n$-gon region. Next, we will compute the value of φ for which these functions are convex. Then, we will use a simple convolution theorem to construct a "conjugate" family of planar harmonic functions that are also univalent. Finally, using a Weierstrass representation we will lift this last family to minimal graphs that turn out to be the singly periodic Scherk surfaces with higher dihedral symmetry. Because of the harmonic functions are univalent, the embeddedness of the Scherk surfaces is guaranteed.

2. A family of univalent planar harmonic mappings

Definition 2.1. A continuous function $f(x, y) = u(x, y) + i\,v(x, y)$ defined in a domain $G \subset \mathbb{C}$ is a *complex-valued harmonic function* in G if u and v are real harmonic functions in G.

Complex-valued harmonic functions defined on $\mathbb{D}$, the unit disk, are related to analytic functions, as the following theorem shows.

Theorem 2.2 [Clunie and Sheil-Small 1984]. *If $f = u + i\,v$ is harmonic in a simply connected domain G, then f can be written as $f = h + \bar{g}$, where h and g are analytic.*

We are interested in univalent (one-to-one) harmonic mappings. While it is often difficult to establish the univalency of a planar harmonic function, we do have the following nice result about local univalency.

Lemma 2.3 [Lewy 1936]. *The harmonic function $f = h + \bar{g}$ is locally univalent and sense-preserving in $\mathbb{D}$ if and only if $|g'(z)/h'(z)| < 1$ for all $z \in \mathbb{D}$.*

The function $\omega(z) = g'(z)/h'(z)$ is known as the dilatation and plays an important role in the theory of univalent harmonic mappings.

We will now consider a specific family of planar harmonic mappings that are related to Scherk surfaces. Let $f_n(z) = h_n(z) + \overline{g_n(z)}$ be the family of planar harmonic mappings from $\mathbb{D}$ into $\mathbb{C}$, where

$$h_n'(z) = \frac{1}{(z^n - e^{i\varphi})(z^n - e^{-i\varphi})}, \quad g_n'(z) = \frac{z^{2n-2}}{(z^n - e^{i\varphi})(z^n - e^{-i\varphi})},$$

$n \geq 2$ and $\varphi \in [0, \frac{\pi}{2}]$. Thus,

$$f_n(z) = \int_0^z \frac{d\zeta}{(\zeta^n - e^{i\varphi})(\zeta^n - e^{-i\varphi})} + \overline{\int_0^z \frac{\zeta^{2n-2}\, d\zeta}{(\zeta^n - e^{i\varphi})(\zeta^n - e^{-i\varphi})}}.$$

Note that $g_n'(z)/h_n'(z) = z^{2n-2}$. Letting ξ be the primitive n-th root of unity and using the residue theorem, we can compute that

$$\begin{aligned} h_n(z) =& \frac{1}{2n\sin\varphi} \int_0^z \left(\sum_{j=1}^n \frac{-i\, e^{-i(\frac{n-1}{n})\varphi} \xi^j}{\zeta - e^{i\frac{\varphi}{n}} \xi^j} + \sum_{j=1}^n \frac{i\, e^{i(\frac{n-1}{n})\varphi} \xi^j}{\zeta - e^{-i\frac{\varphi}{n}} \xi^j} \right) d\zeta \\ =& \frac{1}{2n\sin\varphi} \sum_{k=1}^n \Big(-i\, e^{-i(\frac{n-1}{n}\varphi + \frac{2k\pi}{n})} \log\big(z - e^{i(\frac{\varphi}{n} - \frac{2k\pi}{n})}\big) \\ & \qquad + i\, e^{i(\frac{n-1}{n}\varphi + \frac{2k\pi}{n})} \log\big(z - e^{-i(\frac{\varphi}{n} - \frac{2k\pi}{n})}\big) \Big). \end{aligned}$$

Similarly,

$$\begin{aligned} g_n(z) =& \frac{1}{2n\sin\varphi} \sum_{k=1}^n \Big(-i\, e^{i(\frac{n-1}{n}\varphi + \frac{2k\pi}{n})} \log\big(z - e^{i(\frac{\varphi}{n} - \frac{2k\pi}{n})}\big) \\ & \qquad + i\, e^{-i(\frac{n-1}{n}\varphi + \frac{2k\pi}{n})} \log\big(z - e^{-i(\frac{\varphi}{n} - \frac{2k\pi}{n})}\big) \Big). \end{aligned}$$

Since $f_n(z) = \operatorname{Re}(h_n(z) + g_n(z)) + i \operatorname{Im}(h_n(z) - g_n(z))$, after normalizing so that $f_n(0) = 0$, we get

$$\begin{aligned} f_n(z) = \frac{1}{n\sin\varphi} \sum_{k=1}^n \Bigg\{ & \cos\left(\frac{n-1}{n}\varphi + \frac{2k\pi}{n}\right)\left(\beta_1 - \beta_2 + \frac{4k\pi}{n}\right) \\ & - i \sin\left(\frac{n-1}{n}\varphi + \frac{2k\pi}{n}\right)(\beta_1 + \beta_2) \Bigg\}, \end{aligned} \tag{2}$$

where

$$\begin{aligned} \beta_1 &= \arg\big(z + e^{i(\frac{\varphi}{n} - \frac{2k\pi}{n})}\big), \\ \beta_2 &= \arg\big(z + e^{-i(\frac{\varphi}{n} - \frac{2k\pi}{n})}\big). \end{aligned}$$

Theorem 2.4. *The harmonic function f_n maps $\mathbb{D}$ onto a $2n$-gon.*

Because the dilatation $\omega_n(z)$ equals $g_n'(z)/h_n'(z) = z^{2n-2}$, we know that f_n maps arcs of $\partial\mathbb{D}$ to either concave arcs or to stationary points [Bshouty and Hengartner 1997; Bshouty et al. 2008]. Letting $z = e^{i\theta} \in \partial\mathbb{D}$, we see that the latter situation occurs. In particular, f_n maps $\partial\mathbb{D}$ to vertices, v_m ($m = 1, \ldots, 2n$), of a $2n$-gon such that

$$\arg v_m = e^{\frac{i(j-1)\pi}{n}} \quad \text{and} \quad |v_m| = \begin{cases} |v_1| & \text{if } v_m \text{ is odd,} \\ |v_2| & \text{if } v_m \text{ is even,} \end{cases}$$

where it can be computed that

$$v_1 = \frac{\pi}{n \sin\varphi}\left(\cos\frac{(n-1)\varphi}{n} + \cot\frac{\pi}{n}\sin\frac{(n-1)\varphi}{n}\right) + 0i, \tag{3}$$

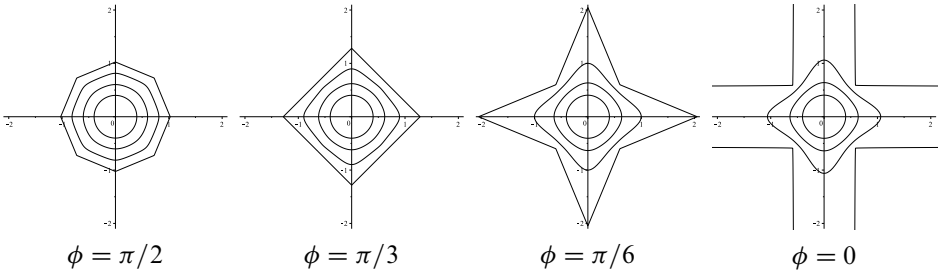
$$v_2 = \frac{\pi}{n \sin\varphi}\sin\frac{(n-1)\varphi}{n}\left(\cot\frac{\pi}{n} + i\right). \tag{4}$$

Example 2.5. For $n = 4$, we have

$$f_4(z) = \operatorname{Re}(h_4(z) + g_4(z)) + i \operatorname{Im}(h_4(z) - g_4(z)),$$

where

$$\begin{aligned}\operatorname{Re}(h_4(z)+g_4(z)) = \frac{1}{4\sin\varphi}\Bigg(&\cos\frac{3\varphi}{4}\Big[\arg\big(z-e^{i\frac{\varphi}{4}}\big) - \arg\big(z+e^{i\frac{\varphi}{4}}\big) \\ &\qquad - \arg\big(z-e^{-i\frac{\varphi}{4}}\big) + \arg\big(z+e^{-i\frac{\varphi}{4}}\big)\Big] \\ &+ \sin\frac{3\varphi}{4}\Big[\arg\big(z-e^{i(\frac{\varphi}{4}+\frac{\pi}{2})}\big) - \arg\big(z+e^{i(\frac{\varphi}{4}+\frac{\pi}{2})}\big) \\ &\qquad + \arg\big(z-e^{-i(\frac{\varphi}{4}-\frac{\pi}{2})}\big) - \arg\big(z+e^{-i(\frac{\varphi}{4}-\frac{\pi}{2})}\big)\Big]\Bigg) \\ &+ \frac{2\pi}{4\sin\varphi}\left(\cos\frac{3\varphi}{4} + \sin\frac{3\varphi}{4}\right)\end{aligned}$$



$\phi = \pi/2$ $\phi = \pi/3$ $\phi = \pi/6$ $\phi = 0$

Figure 1. Images under f_4 of concentric circles in $\mathbb{D}$ for various values of φ.

and

$$\operatorname{Im}(h_4(z)-g_4(z)) = \frac{1}{4\sin\varphi}\Bigg(\sin\frac{3\varphi}{4}\big[-\arg\big(z-e^{i\frac{\varphi}{4}}\big)+\arg\big(z+e^{i\frac{\varphi}{4}}\big) - \arg\big(z-e^{-i\frac{\varphi}{4}}\big)+\arg\big(z+e^{-i\frac{\varphi}{4}}\big)\big] + \cos\frac{3\varphi}{4}\big[\arg\big(z-e^{i(\frac{\varphi}{4}+\frac{\pi}{2})}\big)+\arg\big(z-e^{i(\frac{\varphi}{3}+\frac{2\pi}{3})}\big) - \arg\big(z-e^{-i(\frac{\varphi}{4}-\frac{\pi}{2})}\big)+\arg\big(z+e^{-i(\frac{\varphi}{4}-\frac{\pi}{2})}\big)\big]\Bigg).$$

Letting

$$M = \frac{\pi}{4\sin\varphi}\cos\frac{3\varphi}{4} \quad \text{and} \quad N = \frac{\pi}{4\sin\varphi}\sin\frac{3\varphi}{4},$$

we see that f_4 maps $\partial\mathbb{D}$ to the vertices of an octagon as follows (see Figure 1):

$$f_4(e^{i\theta}) = \begin{cases} v_1 = (M+N) & \text{if } -\frac{\varphi}{4} < \theta < \frac{\varphi}{4}, \\ v_2 = N + iN & \text{if } \frac{\varphi}{4} < \theta < \frac{\pi}{2} - \frac{\varphi}{4}, \\ v_3 = i(M+N) & \text{if } \frac{\pi}{2} - \frac{\varphi}{4} < \theta < \frac{\pi}{2} + \frac{\varphi}{4}, \\ v_4 = -N + iN & \text{if } \frac{\pi}{2} + \frac{\varphi}{4} < \theta < \pi - \frac{\varphi}{4}, \\ v_5 = -(M+N) & \text{if } \pi - \frac{\varphi}{4} < \theta < \pi + \frac{\varphi}{4}, \\ v_6 = -N - iN & \text{if } \pi + \frac{\varphi}{4} < \theta < \frac{3\pi}{2} - \frac{\varphi}{4}, \\ v_7 = -i(M+N) & \text{if } \frac{3\pi}{2} - \frac{\varphi}{4} < \theta < \frac{3\pi}{2} + \frac{\varphi}{4}, \\ v_8 = N - iN & \text{if } \frac{3\pi}{2} + \frac{\varphi}{4} < \theta < -\frac{\varphi}{4}. \end{cases}$$

Theorem 2.6. *For $n \geq 2$, f_n is univalent for all $z \in \mathbb{D}$ and $\varphi \in (0, \frac{\pi}{2}]$.*

Proof. This follows from a result by Duren, McDougall and Schaubroeck [Duren et al. 2005] that states if a harmonic function f is of the form (2) constructed with a piecewise constant boundary function and with values on the m vertices of a polygonal region Ω and with $\omega = g'(z)/h'(z)$ being a Blaschke product with at most $m-2$ factors, then

$$f(z) \text{ is univalent in } \mathbb{D} \iff \text{all the zeros of } \omega \text{ lie in } \mathbb{D}. \qquad \square$$

Remark 2.7. For $n = 3, 4$, one can simply employ the shearing technique of Clunie and Sheil-Small [1984] to prove univalency with even less background. However, for $n \geq 5$ the shearing technique cannot be applied to f_n.

Theorem 2.8. *The image $f_n(\mathbb{D})$ is convex for every $\varphi \in \left(\frac{n}{n-1}\left(\frac{\pi}{2} - \frac{\pi}{n}\right), \frac{\pi}{2}\right]$.*

Proof. Note that f_n will be convex for every φ if

$$\operatorname{Re} v_2 > \tfrac{1}{2}\operatorname{Re}(v_1 + v_3) \quad \text{and} \quad \operatorname{Im} v_2 > \tfrac{1}{2}\operatorname{Im}(v_1 + v_3).$$

From (3), it is clear that

$$\begin{aligned}
&\operatorname{Re} v_1 = v_1, && \operatorname{Im} v_1 = 0,\\
&\operatorname{Re} v_2 = v_1 - \frac{\pi \cos \frac{(n-1)\varphi}{n}}{n \sin \varphi}, && \operatorname{Im} v_2 = \frac{\pi \sin \frac{(n-1)\varphi}{n}}{n \sin \varphi},\\
&\operatorname{Re} v_3 = \operatorname{Re}(e^{i\frac{2\pi}{n}} v_1) = \cos \frac{2\pi}{n} v_1, && \operatorname{Im} v_3 = \operatorname{Im}(e^{i\frac{2\pi}{n}} v_1) = \sin \frac{2\pi}{n} v_1.
\end{aligned}$$

Setting $\operatorname{Re} v_2 = \frac{1}{2} \operatorname{Re}(v_1 + v_3)$ and solving for v_1 yields

$$v_1 = \frac{2\pi}{n} \cdot \frac{\cos \frac{(n-1)\varphi}{n}}{\sin \varphi \left(1 - \cos \frac{2\pi}{n}\right)}. \tag{5}$$

Likewise, setting $\operatorname{Im}(v_2) = \frac{1}{2} \operatorname{Im}(v_1 + v_3)$ and again solving for v_1 yields

$$v_1 = \frac{2\pi}{n} \cdot \frac{\sin \frac{(n-1)\varphi}{n}}{\sin \varphi \sin \frac{2\pi}{n}}. \tag{6}$$

Equating (5) and (6) and solving for φ we obtain

$$\varphi = \frac{n}{n-1} \arctan \frac{\sin \frac{2\pi}{n}}{1 - \cos \frac{2\pi}{n}} = \frac{n}{n-1} \left(\frac{\pi}{2} - \frac{\pi}{n}\right). \qquad \square$$

There is a convolution theorem for planar harmonic mappings that takes univalent convex maps and transforms them into new harmonic maps while preserving univalency. We will apply this convolution theorem to those functions f_n that map $\mathbb{D}$ onto a convex domain. But first, we need some background. For analytic functions

$$f(z) = \sum_{n=0}^{\infty} a_n z^n \quad \text{and} \quad F(z) = \sum_{n=0}^{\infty} A_n z^n,$$

their convolution is defined as

$$f(z) * F(z) = \sum_{n=0}^{\infty} a_n A_n z^n.$$

Note that the right half-plane mapping, $f(z) = z/(1-z)$, acts as the convolution identity; that is, if F is an analytic function, then

$$\frac{z}{1-z} * F(z) = F(z).$$

Now let's consider the case of harmonic convolutions.

Definition 2.9. Given harmonic univalent functions

$$f(z) = h(z) + \bar{g}(z) = z + \sum_{n=2}^{\infty} a_n z^n + \sum_{n=1}^{\infty} \bar{b}_n \bar{z}^n,$$

$$F(z) = H(z) + \overline{G}(z) = z + \sum_{n=2}^{\infty} A_n z^n + \sum_{n=1}^{\infty} \overline{B}_n \bar{z}^n,$$

define the *harmonic convolution* as

$$f(z) * F(z) = h(z) * H(z) + \overline{g(z) * G(z)} = z + \sum_{n=2}^{\infty} a_n A_n z^n + \sum_{n=1}^{\infty} \overline{b_n B_n} \bar{z}^n.$$

Lemma 2.10 [Clunie and Sheil-Small 1984]. *Let $f = h + \bar{g}$ be a harmonic univalent mapping from $\mathbb{D}$ onto a convex domain and normalized so that $f(0) = 0$ and $f_z(0) = 1$. Also, let ϕ be a normalized univalent analytic function from $\mathbb{D}$ onto a convex domain. Then for* ($|\alpha| \le 1$),

$$f * (\alpha\overline{\phi} + \phi) = h * \phi + \alpha\overline{g * \phi}$$

is a univalent harmonic map $\mathbb{D}$ onto a close-to-convex domain.

Theorem 2.11. *The function F_n is univalent on $\mathbb{D}$ for $\varphi \in \left(\frac{n}{n-1}\left(\frac{\pi}{2} - \frac{\pi}{n}\right), \frac{\pi}{2}\right]$.*

Proof. From Theorem 2.8, we know the f_n are convex maps for $\frac{n}{n-1}\left(\frac{\pi}{2} - \frac{\pi}{n}\right) < \varphi \le \frac{\pi}{2}$. Hence for these values of φ we can apply Lemma 2.10 with $\phi = z/(1-z)$ and $\alpha = -1$ to create the planar harmonic maps

$$F_n(z) = \operatorname{Re}\left(h_n(z) - g_n(z)\right) + i \operatorname{Im}\left(h_n(z) + g_n(z)\right)$$

which are univalent in $\mathbb{D}$. □

Example 2.12. From Theorem 2.11, we conclude that the harmonic maps $F_4(z)$ are univalent in $\mathbb{D}$ (see Figure 2).

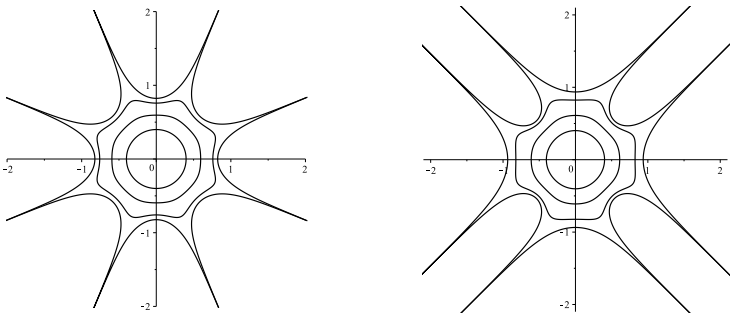
3. Singly periodic Scherk surfaces with higher dihedral symmetry

The connection between planar harmonic mappings and minimal surfaces can be seen in the following *Weierstrass representation* (see [Duren 2004], for example):

Theorem 3.1. *Let $f = h + \bar{g}$ be an orientation-preserving harmonic univalent mapping of $\mathbb{D}$ onto some domain Ω with dilatation $\omega = q^2$, where q is an analytic function in $\mathbb{D}$. Then*

$$X(z) = \left(\operatorname{Re}(h(z) + g(z)),\ \operatorname{Im}(h(z) - g(z)),\ 2\operatorname{Im}\int_0^z \sqrt{g'(\zeta)h'(\zeta)}\, d\zeta\right)$$

gives an isothermal parametrization of a minimal graph whose projection in the xy-plane is f.



$\phi = \pi/2$ $\phi = \pi/3$

Figure 2. Images under F_4 of concentric circles in $\mathbb{D}$ for various values of φ.

Thus, univalent planar harmonic mappings with a dilatation that is the square of an analytic function lift to minimal graphs in $\mathbb{R}^3$. We have shown that both families f_n and F_n of harmonic mappings satisfy the hypotheses of Theorem 3.1 for a given range of φ values and will thus lift to embedded minimal graphs. To identify these surfaces, we use the following standard Weierstrass representation.

Theorem 3.2 (Weierstrass representation (G, dh) [Weber 2005]). *Every regular minimal surface has a local isothermal parametric representation of the form*

$$X(z) = \operatorname{Re} \int_a^z \left(\frac{1}{2}\left(\frac{1}{G} - G\right),\ \frac{i}{2}\left(\frac{1}{G} + G\right),\ 1 \right) dh,$$

where G is the Gauss map, dh is the height differential, and $a \in \mathbb{D}$ is a constant.

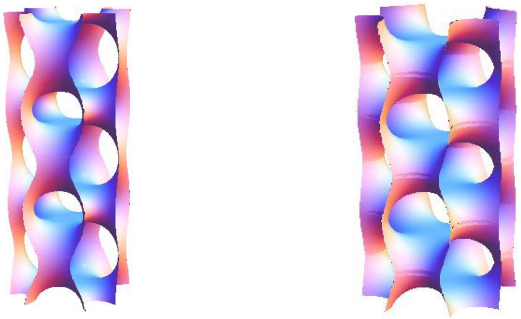
Proving the embeddedness of singly periodic Scherk surfaces with higher dihedral symmetry is not easy. However, with the material we have developed it follows naturally.

Theorem 3.3. *F_n lifts to a family of embedded singly periodic Scherk surfaces with higher dihedral symmetry for φ satisfying* (1).

Proof. Scalings and reflections across planes containing two axes do not alter the geometry of minimal surfaces. So we can use the coordinate functions from the two Weierstrass representations to get

$$h = \int_0^z \frac{1}{G}\, dh, \quad g = \int_0^z G\, dh. \tag{7}$$

In [Weber 2005] the Gauss map and height differential for a family of minimal surfaces ranging from Scherk's singly periodic surface with $2n$ ends when $\varphi = \frac{\pi}{2}$



$\phi = \pi/2$ $\phi = \pi/3$

Figure 3. Singly periodic Scherk surfaces.

to the n-noid when $\varphi = 0$ is given by

$$G = z^{n-1}, \quad dh = \frac{z^{n-1}}{(z^n - e^{i\varphi})(z^n - e^{-i\varphi})}.$$

Using the formulas in (7) we see

$$h^* = \int_0^z \frac{d\zeta}{(\zeta^n - e^{i\varphi})(\zeta^n - e^{-i\varphi})}, \quad g^* = -\int_0^z \frac{\zeta^{2n-2}\, d\zeta}{(\zeta^n - e^{i\varphi})(\zeta^n - e^{-i\varphi})}.$$

It is clear that $F_n = h^* + \overline{g^*}$. Hence, we see that F_n lifts to this family of singly periodic Scherk's surfaces for all $\varphi \in \left(\frac{n}{n-1}\left(\frac{\pi}{2} - \frac{\pi}{n}\right), \frac{\pi}{2}\right]$. □

Remark 3.4. We could have used Krust's theorem [Dierkes et al. 1992] instead of Lemma 2.10. But this convolution theorem is not well known and is a generalization of Krust's Theorem applied to planar harmonic mappings.

Remark 3.5. The harmonic maps, f_n, lift to a family of minimal surfaces that continuously transform from Scherk's first surface with $2n$-ends to a minimal surface with n-helicoidal ends. Because the harmonic maps are univalent, the resulting minimal surfaces are graphs. However, they are graphs only over the domain $\mathbb{D}$. This does not contradict the fact that the minimal surface with n helicoidal ends is not embedded since the surface is defined on a domain larger than $\mathbb{D}$.

Area for further investigation. Apply the approach used in this paper to prove the embeddedness for less symmetric Scherk-like surfaces and for the twist deformation of Scherk's singly periodic surfaces (see [Weber 2005, pp. 39–40]).

Acknowledgements

The authors would like to thank Casey Douglas for his comments and suggestions.

References

[Bshouty and Hengartner 1997] D. Bshouty and W. Hengartner, "Boundary values versus dilatations of harmonic mappings", *J. Anal. Math.* **72** (1997), 141–164. MR 99c:30061 Zbl 0908.30017

[Bshouty et al. 2008] D. Bshouty, A. Lyzzaik, and A. Weitsman, "On the boundary behaviour of univalent harmonic mappings onto convex domains", *Comput. Methods Funct. Theory* **8**:1-2 (2008), 261–275. MR 2010b:30069 Zbl 1160.30002

[Clunie and Sheil-Small 1984] J. Clunie and T. Sheil-Small, "Harmonic univalent functions", *Ann. Acad. Sci. Fenn. Ser. A I Math.* **9** (1984), 3–25. MR 85i:30014 Zbl 0506.30007

[Dierkes et al. 1992] U. Dierkes, S. Hildebrandt, A. Küster, and O. Wohlrab, *Minimal surfaces, I: Boundary value problems*, Grundl. Math. Wiss. **295**, Springer, Berlin, 1992. MR 94c:49001a Zbl 0777.53012

[Duren 2004] P. Duren, *Harmonic mappings in the plane*, Cambridge Tracts in Mathematics **156**, Cambridge University Press, 2004. MR 2005d:31001 Zbl 1055.31001

[Duren et al. 2005] P. Duren, J. McDougall, and L. Schaubroeck, "Harmonic mappings onto stars", *J. Math. Anal. Appl.* **307**:1 (2005), 312–320. MR 2006c:31002 Zbl 1112.31001

[Karcher 1988] H. Karcher, "Embedded minimal surfaces derived from Scherk's examples", *Manuscripta Math.* **62**:1 (1988), 83–114. MR 89i:53009 Zbl 0658.53006

[Lewy 1936] H. Lewy, "On the non-vanishing of the Jacobian in certain one-to-one mappings", *Bull. Amer. Math. Soc.* **42**:10 (1936), 689–692. MR 1563404 Zbl 0015.15903

[McDougall and Schaubroeck 2008] J. McDougall and L. Schaubroeck, "Minimal surfaces over stars", *J. Math. Anal. Appl.* **340**:1 (2008), 721–738. MR 2009d:31004 Zbl 1169.53007

[Pérez and Traizet 2007] J. Pérez and M. Traizet, "The classification of singly periodic minimal surfaces with genus zero and Scherk-type ends", *Trans. Amer. Math. Soc.* **359**:3 (2007), 965–990. MR 2007m:53010 Zbl 1110.53008

[Weber 2005] M. Weber, "Classical minimal surfaces in Euclidean space by examples: geometric and computational aspects of the Weierstrass representation", pp. 19–63 in *Global theory of minimal surfaces*, edited by D. Hoffman, Clay Math. Proc. **2**, Amer. Math. Soc., Providence, RI, 2005. MR 2006e:53025 Zbl 1100.53015

Received: 2012-05-23 Revised: 2012-07-24 Accepted: 2012-07-25

vbuqaj@gmail.com *Computer Science, Information Systems and Mathematics, Texas Lutheran University, Seguin, TX 78155, United States*

cannon.sarahm@gmail.com *Department of Mathematics, Tufts University, Medford, MA 02155, United States*

mdorff@math.byu.edu *Department of Mathematics, Brigham Young University, Provo, UT 84602, United States*

jelawson@loyno.edu *Mathematical Sciences, Loyola University New Orleans, New Orleans, LA 70118, United States*

rdviertel@gmail.com *Department of Mathematics, Brigham Young University, Provo, UT 84602, United States*

INVOLVE 6:4 (2013)
dx.doi.org/10.2140/involve.2013.6.393

An elementary inequality about the Mahler measure

Konstantin Stulov and Rongwei Yang

(Communicated by Andrew Granville)

Let $p(z)$ be a degree n polynomial with zeros z_j, $j = 1, 2, \ldots, n$. The total distance from the zeros of p to the unit circle is defined as $\mathrm{td}(p) = \sum_{j=1}^{n} \big||z_j| - 1\big|$. We show that up to scalar multiples, $\mathrm{td}(p)$ sits between $M(p) - 1$ and $m(p)$. This leads to an equivalent statement of Lehmer's problem in terms of $\mathrm{td}(p)$. The proof is elementary.

1. Introduction

Let $p(z) = \sum_{j=0}^{n} a_j z^j$ be a polynomial with complex coefficients of degree n. The Mahler measure $M(p)$ [Everest and Ward 1999] is defined as

$$M(p) = \exp\left(\int_0^{2\pi} \log \big| p(e^{i\theta}) \big| \, \frac{d\theta}{2\pi} \right).$$

We denote $\log M(p)$ by $m(p)$. Jensen's formula implies that

$$M(p) = |a_n| \prod_{|z_j|>1} |z_j|,$$

where throughout this paper the z_j, $j = 1, 2, \ldots, n$, are the zeros of $p(z)$, counting multiplicity. We also assume that $|a_n| = 1$. It is then clear that $M(p) \geq 1$, and

$$0 \leq m(p) = \log\big((M(p) - 1) + 1\big) \leq M(p) - 1,$$

and when $M(p)$ is close to 1, $m(p)$ is close to $M(p) - 1$. Lehmer's problem is to verify that for integer-coefficient monic polynomials, $m(p)$ is either 0 (for products of cyclotomic polynomials and possibly a factor of z^k) or is bounded away from 0 by a fixed positive constant. This is a deep and unsolved problem.

For a polynomial p of degree n, the associated polynomial $p^*(z)$ is defined as $z^n \overline{p(1/\overline{z})}$. We say p is reciprocal if $p = cp^*$ for some complex number c of modulus 1. One sees that the zeros of a reciprocal p off the unit circle appear in conjugate reciprocal pairs. Interestingly, Lehmer's problem was unsolved only for reciprocal polynomials. A key ingredient of this paper is the total distance from the

MSC2010: 11CXX.
Keywords: Mahler measure, total distance.

zeros of p to the unit circle T defined to be

$$\mathrm{td}(p) = \sum_{j=1}^{n} \big||z_j| - 1\big|.$$

Theorem. *For every complex polynomial* $p(z) = \sum_{j=0}^{n} a_j z^j$, *with* $|a_n| = |a_0| = 1$, *we have*

$$m(p) \le \mathrm{td}(p) \le 2(M(p) - 1).$$

If p *is reciprocal, then* $2m(p) \le \mathrm{td}(p)$. *Further, the equalities hold only if* $\mathrm{td}(p) = 0$.

Therefore, Lehmer's problem can be stated equivalently as follows: There is an $\epsilon > 0$ such that if p has integer coefficients with $|a_n| = |a_0| = 1$ and $\mathrm{td}(p) \neq 0$, then $\mathrm{td}(p) \ge \epsilon$.

2. Proof

Lemma 1. *If* t_j, $j = 1, 2, \ldots, k$ *are numbers in the interval* $[0, 1]$, *then*

$$\sum_{j=1}^{k} (1 - t_j) \le \frac{1}{\prod_{j=1}^{k} t_j} - 1,$$

where equality holds only if $t_j = 1$ *for each* j.

Proof. The inequality is trivial if one of the t_j is 0. Now, we assume $t_j > 0$ for each j. We prove by induction. It is easy to see that the lemma is true for $k = 1$. Assume the lemma is true for k. For s and t in $(0, 1]$, one checks that

$$\frac{1}{ts} - \left(\frac{1}{t} + 1 - s\right) = \frac{(1-s)(1-ts)}{ts} \ge 0, \tag{2-1}$$

and hence

$$\frac{1}{ts} - 1 \ge \frac{1}{t} - s.$$

Therefore

$$\begin{aligned}\sum_{j=1}^{k} (1 - t_j) + (1 - t_{k+1}) &\le \frac{1}{\prod_{j}^{k} t_j} - 1 + (1 - t_{k+1}) \\ &= \frac{1}{\prod_{j}^{k} t_j} - t_{k+1} \le \frac{1}{\prod_{j}^{k+1} t_j} - 1. \end{aligned}$$

□

If $\{\lambda_j : j = 1, 2, \ldots\}$ is a subset of the open unit disk $\mathbb{D}$, the associated Blaschke product is defined as

$$B(z) = \prod_{j=1}^{\infty} \frac{z - \lambda_j}{1 - \overline{\lambda_j} z}, \quad z \in \mathbb{D}.$$

Clearly, the product is convergent for each z if and only if $\sum_{j=0}^{\infty}(1-|\lambda_j|)<\infty$ [Garnett 2007]. In this case $B(z)$ is a bounded analytic function on $\mathbb{D}$. It follows immediately from Lemma 1 that

$$\sum_{j=1}^{\infty}(1-|\lambda_j|)\le\frac{1}{|B(0)|}-1.$$

Proof of the Theorem. For a polynomial $p(z)$, since $\prod_{j=1}^{n}|z_j|=\dfrac{|a_0|}{|a_n|}$, we have

$$\frac{M(p)}{|a_0|}-1=\frac{1}{\prod_{|z_j|\le1}|z_j|}-1\ge\sum_{|z_j|\le1}(1-|z_j|) \tag{2-2}$$

by Lemma 1. On the other hand, inductively using that $(a-1)+(b-1)<ab-1$ for $a,b>1$, we have

$$\sum_{|z_j|>1}(|z_j|-1)\le\prod_{|z_j|>1}|z_j|-1=\frac{M(p)}{|a_n|}-1.$$

Here the equality is allowed only because there may not be a z_j with $|z_j|>1$. Combining with (2-2), we have $\mathrm{td}(p)\le M(p)(1/|a_n|+1/|a_0|)-2$. In the case $|a_0|=|a_n|=1$, we have

$$\mathrm{td}(p)\le2(M(p)-1), \tag{2-3}$$

with equality occurring only if $\mathrm{td}(p)=0$. The dominance of $m(p)$ by $\mathrm{td}(p)$ is an easy consequence of the inequality $\log(1+t)\le t$. To be precise,

$$m(p)=\sum_{|z_k|>1}\log|z_k|\le\sum_{|z_k|>1}(|z_k|-1)\le\mathrm{td}(p).$$

We establish a stronger inequality for reciprocal polynomials with $|a_0|=|a_n|=1$. Let $z_1,z_2,\dots,z_k$ be the zeros of such a p that are outside of the unit circle, where $2k\le n$. Then $m(p)=\log|z_1|+\log|z_2|+\cdots+\log|z_k|$ and

$$\mathrm{td}(p)=\sum_{j=1}^{k}(|z_j|-1)+\Big(1-\frac{1}{|z_j|}\Big).$$

Let $f(t)=t-(1/t)-2\log t,\ t\ge1$. One easily checks that f is strictly increasing and $f(1)=0$. It follows that $|z_j|-1/|z_j|>2\log|z_j|$ for each $1\le j\le k$, and hence $2m(p)\le\mathrm{td}(p)$, with equality precisely when $k=0$, which occurs if and only if $\mathrm{td}(p)=0$ since $|a_0|=|a_n|=1$. □

Example. Consider Lehmer's polynomial

$$G(z)=z^{10}+z^9-z^7-z^6-z^5-z^4-z^3+z+1.$$

It is well-known that eight of its zeros lie in the unit circle and the other two are real and form a reciprocal pair. Since $M(G) \approx 1.1763$, we have

$$\begin{aligned} \mathrm{td}(G) &\approx (1.1763 - 1) + (1 - 1/1.1763) \approx 0.3262, \\ 2m(G) &\approx 2 \times 0.1624 = 0.3248, \\ 2(M(p) - 1) &\approx 0.3526. \end{aligned}$$

Our Theorem has some interesting implications. We need two more definitions to state them. Define

$$\begin{aligned} \Delta(p) &= \max\big\{\big||\alpha| - 1\big| : p(\alpha) = 0\big\}, \\ \delta(p) &= \min\big\{\big||\alpha| - 1\big| : p(\alpha) = 0\big\}. \end{aligned}$$

Then it is clear that

$$\delta(p) \le \frac{\mathrm{td}(p)}{n} \le \Delta(p). \tag{2-4}$$

When p is reciprocal and α is a zero of p, $1/\alpha$ is also a zero. Since $t - 1 \ge 1 - 1/t$ for $t \ge 1$, we have

$$\Delta(p) = \max\{|\alpha| - 1 : p(\alpha) = 0\} = \max\{|\alpha| : p(\alpha) = 0\} - 1.$$

Likewise

$$\delta(p) = 1 - \max\{|\alpha| : |\alpha| \le 1,\ p(\alpha) = 0\}.$$

For simplicity, we let

$$\lambda(p) = \max\{|\alpha| : p(\alpha) = 0\}$$

and let

$$\lambda'(p) = \max\{|\alpha| : |\alpha| \le 1,\ p(\alpha) = 0\}.$$

In [Smyth 2008], $\lambda(p)$ is called the house of the zeros of p. Geometrically, $\lambda(p)$ is the modulus of the zero that is the farthest from the unit circle, while $\lambda'(p)$ is the modulus of the zero that is the nearest to the unit circle. The next proposition then follows easily from (2-4).

Proposition. *For a reciprocal complex polynomial p of degree $n \ge 2$,*

$$\lambda(p) \ge 1 + \frac{\mathrm{td}(p)}{n} \quad \textit{and} \quad \lambda'(p) \ge 1 - \frac{\mathrm{td}(p)}{n}.$$

Regarding $\lambda(p)$, there is an unsolved conjecture by Schinzel and Zassenhaus that states that there is an absolute constant C so that if p is a monic irreducible polynomial of degree n with integer coefficients, then $\lambda(p) \ge 1 + C/n$. This inequality will follow easily from a positive answer to Lehmer's problem. Indeed, one has $\lambda(p) \ge 1 + m(p)/n$ [Smyth 2008]. But in view of Theorem, Proposition provides a better inequality for reciprocal polynomials.

Acknowledgment

The authors thank the referee for helpful and stimulating comments.

References

[Everest and Ward 1999] G. Everest and T. Ward, *Heights of polynomials and entropy in algebraic dynamics*, Springer, London, 1999. MR 2000e:11087

[Garnett 2007] J. B. Garnett, *Bounded analytic functions*, 1st ed., Graduate Texts in Mathematics **236**, Springer, New York, 2007. MR 2007e:30049

[Smyth 2008] C. Smyth, "The Mahler measure of algebraic numbers: a survey", pp. 322–349 in *Number theory and polynomials*, London Math. Soc. Lecture Note Ser. **352**, Cambridge Univ. Press, 2008. MR 2009j:11172

Received: 2012-07-09 Revised: 2013-02-12 Accepted: 2013-02-16

kstulov@stanford.edu *Institute for Computational and Mathematical Engineering, Stanford University, Stanford, NY 94305, United States*

ryang@albany.edu *Department of Mathematics and Statistics, University of Albany, State University of New York, Albany, NY 12047, United States*

INVOLVE 6:4 (2013)
dx.doi.org/10.2140/involve.2013.6.399

Ecological systems, nonlinear boundary conditions, and Σ-shaped bifurcation curves

Kathryn Ashley, Victoria Sincavage and Jerome Goddard II

(Communicated by John Baxley)

We examine a one-dimensional reaction diffusion model with a weak Allee growth rate that appears in population dynamics. We combine grazing with a certain nonlinear boundary condition that models negative density dependent dispersal on the boundary and analyze the effects on the steady states. In particular, we study the bifurcation curve of positive steady states as the grazing parameter is varied. Our results are acquired through the adaptation of a quadrature method and Mathematica computations. Specifically, we computationally ascertain the existence of Σ-shaped bifurcation curves with several positive steady states for a certain range of the grazing parameter.

1. Introduction

Within population dynamics, the most accepted exemplar for modeling a designated population is the logistic equation

$$f(u) = u(a - bu), \tag{1-1}$$

which illustrates the inference that as a population burgeons, the per capita growth rate

$$\tilde{f}(u) = a - bu \tag{1-2}$$

of that population declines linearly. Yet empirically several authors have witnessed that at lower population densities, the per capita growth rate initially increases (see [Allee 1938; Dennis 1989; Lewis and Kareiva 1993; Shi and Shivaji 2006]). This phenomenon is known in the literature as the Allee effect [1938]. Since the logistic growth model does not compensate for the initial increase, a model of the Allee effect must be implemented to account for this phenomenon.

MSC2010: 34B08, 34B18.
Keywords: nonlinear boundary conditions, weak Allee effect, positive solutions.
Research supported by National Science Foundation grant DMS 0852032.

The Allee effect can be either strong, in which the per capita growth rate is initially negative, or weak, in which the per capita growth rate is initially positive. The Allee effect is generally modeled in the literature via quadratic per capita growth rate functions of the population density. In this case, the analysis is more difficult since the per capita growth rate is not linear or even nonincreasing. As a contrast with (1-1), a weak Allee effect has been modeled as

$$f(u) = u(u+1)(b-u), \tag{1-3}$$

where $b > 1$.

By analyzing additional factors that can influence a population, such as grazing or harvesting, a better understanding can be had of the dynamics of the population. Therefore, through the inclusion of an extra term to account for these factors, specifically grazing, a more precise model can be obtained. Grazing can be considered as a category of natural predation, for example, when an owl preys upon the surrounding rodent population. The term $cu^2/(1+u^2)$ is commonly employed to model grazing of a population (see [Causey et al. 2010; Lee et al. 2011; Poole et al. 2012; van Nes and Scheffer 2005]).

Density dependent dispersal, or more specifically density dependent emigration, describes a situation in which the dispersal/emigration of individuals living within a patch is based on the population density, in our case, on the habitat border. A positive density dependent emigration characterizes a case where individuals have a tendency to leave if the population density is large and a tendency to stay if the population density is small. On the contrary, a negative density dependent emigration represents a case where individuals have a tendency to stay if the population density is large and a tendency to leave if the population density is small.

Initially and intuitively it was believed that the majority of animals exhibit positive density dependent dispersal. However, recent studies of several animals, including the bighorn sheep, roe deer, house mouse, prairie vole, European badger, and the Glanville fritillary butterfly *Melitaea cinxia*, have proven otherwise (see [Kuussaari et al. 1996; Matthysen 2005]). In the literature, several factors have been suggested as a cause of negative density dependent dispersal, including: niche breadth, increased predator abundance, and, in particular, conspecific attraction (see [Kuussaari et al. 1996; Matthysen 2005]). Conspecific attraction most simply means that there is a predisposition of individuals within a population to become enticed to areas where there are more conspecifics.

Cantrel and Cosner proposed the following nonlinear boundary condition to model conspecific attraction on the boundary of a patch (see [Cantrell and Cosner 2003; 2007; Goddard et al. 2010a; 2010b; 2011a; 2011b; 2012]):

$$d(\nabla u \cdot \eta)\alpha(x,u) + [1-\alpha(x,u)]u = 0; \quad \partial\Omega, \tag{1-4}$$

where $\alpha : \bar{\Omega} \times [0, \infty) \to [0, 1]$ is C^1 and nondecreasing, $d > 0$ is the diffusion parameter, $\nabla u \cdot \eta$ is the outward normal derivative, and $\Omega \subset \mathbb{R}^n$ $(n \geq 1)$ is a smooth bounded domain. The $\alpha(x, u)$'s of biological importance are of the form

$$\alpha(x, u) = \alpha(u) = \frac{u}{u + g(u)}, \tag{1-5}$$

where $g : [0, \infty) \to [\delta, \infty)$ is a C^1 function, $\delta > 0$, and $g(u)/u \to 0$ as $u \to \infty$. Here, $\alpha(u)$ represents the fraction of the population that stays on the boundary when reached. Notice that if $\alpha(u) \equiv 0$ then (1-4) becomes the Dirichlet boundary condition $(u = 0;\ \partial\Omega)$, and if $\alpha(u) \equiv 1$ then (1-4) becomes the Neumann boundary condition $(\nabla u \cdot \eta = 0;\ \partial\Omega)$. In terms of this paper, we consider the case when $g(u) \equiv d$, where $d > 0$ is the diffusion parameter.

Our purpose is to analyze the effects of grazing in combination with a weak Allee effect and the nonlinear boundary conditions (1-4) on the steady state solutions of a reaction diffusion model. In particular, we study the one-dimensional case when $n = 1$ and $\Omega = (0, 1)$:

$$u_t = \frac{1}{\lambda} u_{xx} + u\tilde{f}(u) - \frac{cu^2}{1 + u^2}; \quad (0, 1), \tag{1-6}$$

with nonlinear boundary conditions, namely

$$\begin{aligned} -u'' = \lambda\left[u\tilde{f}(u) - \frac{cu^2}{1 + u^2}\right] = \lambda f(u); \quad (0, 1), \\ u(0)\left[-\frac{1}{\lambda}u'(0) + \frac{1}{\lambda}\right] = 0, \\ u(1)\left[\frac{1}{\lambda}u'(1) + \frac{1}{\lambda}\right] = 0, \end{aligned} \tag{1-7}$$

where u represents the population density, $\tilde{f}(u)$ represents the per capita growth rate, $\lambda = 1/d$ and $d > 0$ represents the diffusion coefficient, and $c \geq 0$ represents the maximum grazing rate. Notice that the boundary conditions found in (1-7) can be separated into the following four cases:

$$-u'' = \lambda f(u);\ (0, 1), \quad u(0) = 0, \quad u(1) = 0, \tag{1-8}$$

$$-u'' = \lambda f(u);\ (0, 1), \quad u(0) = 0, \quad u'(1) = -1, \tag{1-9}$$

$$-u'' = \lambda f(u);\ (0, 1), \quad u'(0) = 1, \quad u(1) = 0, \tag{1-10}$$

$$-u'' = \lambda f(u);\ (0, 1), \quad u'(0) = 1, \quad u'(1) = -1. \tag{1-11}$$

Thus, the positive solutions of (1-8)–(1-11) are the positive solutions of (1-7). Further, it is clear that if $u(x)$ is a positive solution of (1-9), then $v(x) = u(1 - x)$ also satisfies (1-10). Thus, it suffices to only consider (1-8), (1-9), and (1-11).

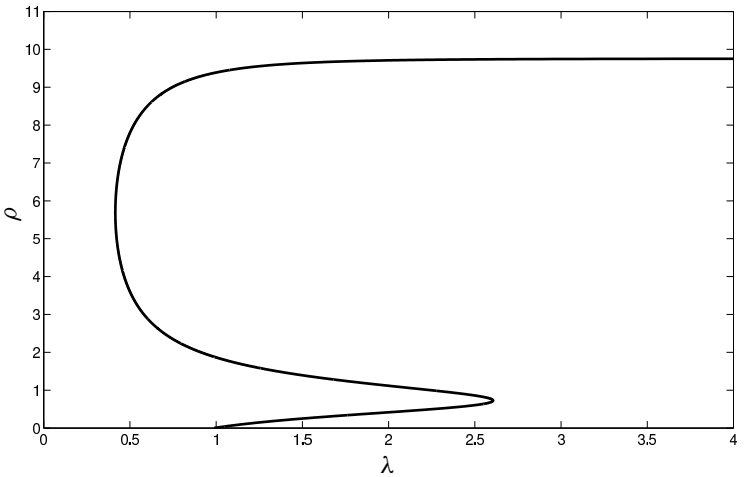


Figure 1. S-shaped bifurcation curve.

Prior studies have gathered information and analyzed the positive solutions to both strong and weak Allee problems. Additionally, the analysis of the positive solutions to the combination of grazing and the Allee effect has also been made; however, to the best of our understanding no analysis has been made in regards to the Allee effect with grazing and nonlinear boundary conditions. In the case when $\alpha(u) \equiv 0$, (1-8) has a rich history. For the logistic case with Dirichlet boundary conditions, Lee, Sasi, and Shivaji proved the existence of an S-shaped bifurcation curve in one dimension, as well as higher dimensions for a certain range of the grazing parameter [Lee et al. 2011]. Regarding the one-dimensional weak Allee effect model with Dirichlet boundary conditions, Poole, Roberson, and Stephenson showed the existence of an S-shaped bifurcation curve, resembling Figure 1, both computationally and analytically for certain parameter ranges [Poole et al. 2012]. In particular, our focus is to further examine the structure of positive solutions of (1-7) when the nonlinear boundary conditions (1-4) are satisfied for the range of the parameters where Poole et al. [2012] showed the existence of an S-shaped bifurcation curve of positive solutions. Computationally, we show the existence of Σ-shaped bifurcation curves as exemplified in Figure 2.

We employ and adapt the quadrature method first developed by Laetsch [1970] to study the structure of positive solutions of (1-7). First, some important preliminaries will be presented in Section 2, followed by a discussion of applying and adapting the quadrature method for the specific cases (1-8), (1-9), and (1-11). In Section 6, we provide the complete evolution of the bifurcation curve of positive solutions of (1-7), followed by analytical results confirming some of our observations in Section 7.

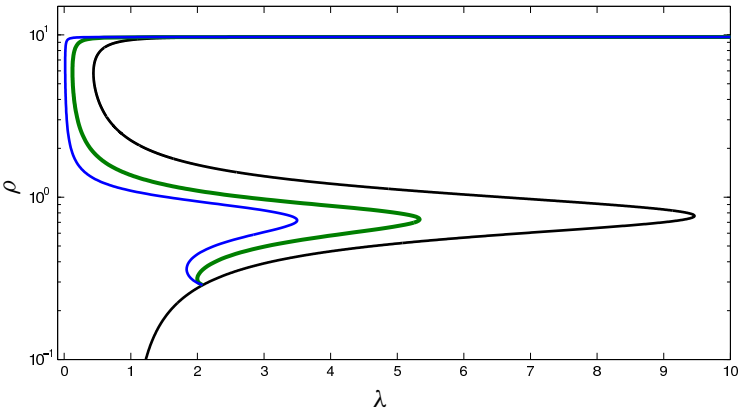


Figure 2. Σ-shaped bifurcation curves.

2. Preliminaries

We examine the combination of the weak Allee effect and grazing in the subsequent reaction term:

$$f(u) = u(u+1)(b-u) - \frac{cu^2}{1+u^2} \quad \text{for } b > 1,\ c \geq 0$$
$$= \frac{u(u+1)(b-u)(1+u^2) - cu^2}{1+u^2}.$$

Through observation it is apparent that the numerator of $f(u)$ can be written as a fifth-degree polynomial. Regardless of any specific values for b and c, by analyzing the roots of $f(u)$ the existence of three roots — a negative root, a positive root, and a root at $u = 0$ — can be determined. As c is varied the remaining three roots alternate between imaginary and real values. For the purpose of this paper, denote $\sigma = \sigma(b, c)$ as the smallest positive root of $f(u)$. Also, allow $\sigma_0 = \sigma_0(b, c)$ and $\sigma_1 = \sigma_1(b, c)$ to represent the remaining roots. Regardless of the value of $c > 0$, for certain values of b, specifically $b \in (1, b_0)$ (some $b_0 > 0$), there exists only one positive real root of $f(u)$ represented by σ.

Remark 1. Through calculation and the use of Mathematica, it is estimated that $b_0 \approx 2.852$.

Specifically, when $b \in (b_0, \infty)$, it has been determined that the shape of $f(u)$ changes when c is varied. Note when $c \in [0, c_0)$ (some $c_0 = c_0(b) > 0$), there exists exactly one positive real root denoted by $\sigma(b, c)$. Figure 3 depicts this case. The shape of $f(u)$ is modified as c becomes larger. Specifically, when $c \in [c_0, c_1)$ (some $c_1 = c_1(b) \in (c_0, \infty)$), $f(u)$ has 3 positive real roots, namely $\sigma(b, c)$, $\sigma_0(b, c)$, and $\sigma_1(b, c)$, as depicted in Figure 4. For $c > c_1$, $f(u)$ is shifted downward resulting in

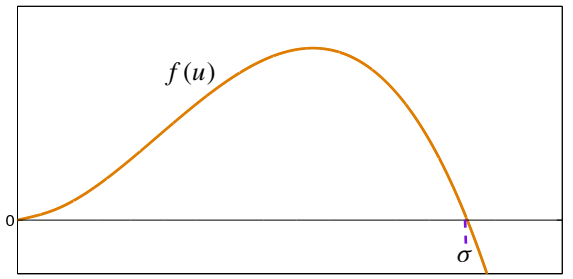


Figure 3. Graph of $f(u)$ for $b > b_0$ and $c \in [0, c_0)$.

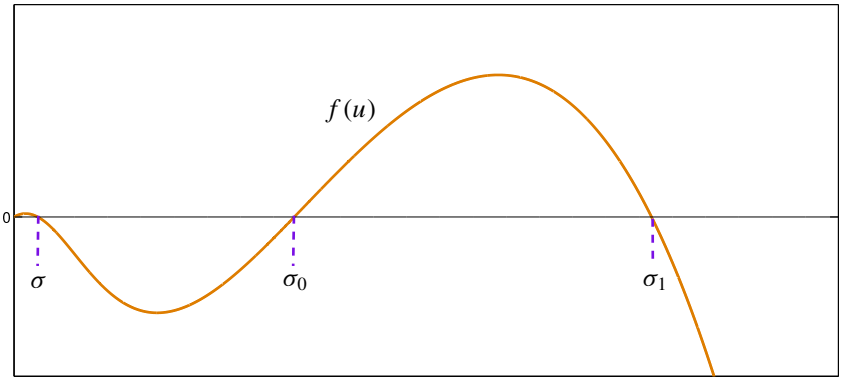


Figure 4. Graph of $f(u)$ for $b > b_0$ and $c \in [c_0, c_1)$.

exactly one positive real root $\sigma(b, c)$, meaning $\sigma_0(b, c)$ and $\sigma_1(b, c)$ are imaginary roots. This particular case is illustrated in Figure 5.

In the preceding cases the structure of the positive solutions of (1-7) varies. As our primary interest is the structure of positive solutions for the range of parameters where Poole et al. [2012] showed the existence of S-shaped bifurcation curves, we focus on the case when $c \in [0, c_0)$.

3. Quadrature method for (1-8)

For completeness, we reestablish the results obtained through the quadrature method actualized by Laetsch [1970] and Brown, Ibrahim, and Shivaji [Brown et al. 1981]. Additionally we recapitulate the subsequent boundary value problem analyzed by Poole et al. [2012] for positive solutions:

$$-u''(x) = \lambda f(u(x)); x \in (0, 1), \quad u(0) = 0, \quad u(1) = 0, \tag{3-1}$$

where $f : [0, \infty) \to (0, \infty)$ is a C^1 function. Clearly, a positive solution of (3-1) must resemble Figure 6.

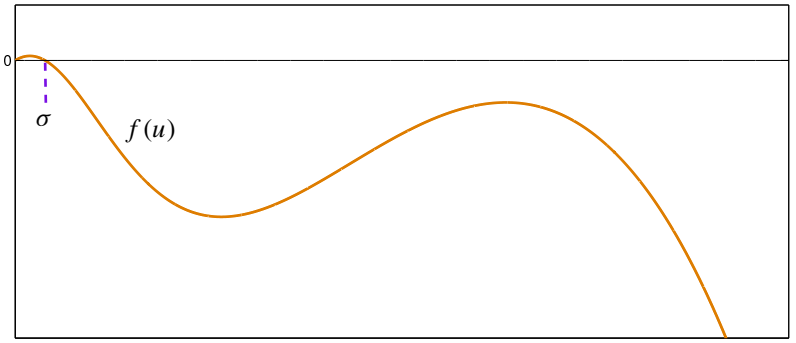


Figure 5. Graph of $f(u)$ for $b > b_0$ and $c > c_1$.

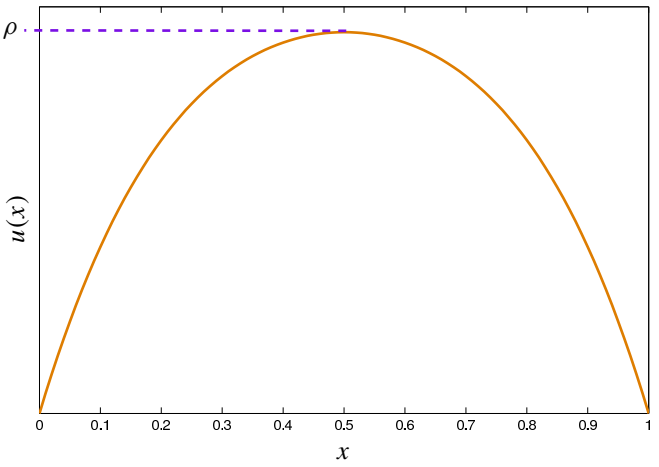


Figure 6. Graph of a typical positive solution of (3-1).

Theorem 3.1 [Brown et al. 1981; Laetsch 1970]. *Suppose $u(x)$ is a positive solution to* (3-1) *with $\|u\|_\infty = \rho = u(\frac{1}{2})$, where $\rho > 0$. Such a solution to* (3-1) *exists if and only if*

$$G_1(\rho) = \sqrt{2}\int_0^\rho \frac{ds}{\sqrt{F(\rho)-F(s)}} = \sqrt{\lambda}, \tag{3-2}$$

where $F(x) = \int_0^x f(s)\,ds$.

Proof. ($\Rightarrow$) Recognizing that (3-1) is an autonomous differential equation, we see that if u is a positive solution to (3-1) with $u'(x_0) = 0$ for a particular $x_0 \in (0, 1)$, then $m(x) = u(x_0 + x)$ and $n(x) = u(x_0 - x)$ both satiate the initial value problem

$$-k''(x) = \lambda f(k(x)), \quad k(0) = u(x_0), \quad k'(0) = 0, \tag{3-3}$$

where $x \in [0, l)$ and $l = \min\{x_0, 1 - x_0\}$. Using Picard's existence and uniqueness theorem, we have that $u(x_0 + x) \equiv u(x_0 - x)$ for all $x \in [0, l)$ and thus $u(x)$ is symmetric about $x = \frac{1}{2}$, which is notedly where $u(x)$ achieves its maximum.

By multiplying the differential equation in (3-1) by $u'(x)$, we have

$$-\left[\frac{[u'(x)]^2}{2}\right]' = \big[\lambda F(u(x))\big]'. \tag{3-4}$$

Integration of both sides of (3-4) gives

$$\frac{u'(x)}{\sqrt{F(\rho) - F(u(x))}} = \sqrt{2\lambda}; \quad x \in [0, \tfrac{1}{2}). \tag{3-5}$$

By integrating a second time and using the fact that $u(0) = 0$, we have

$$\int_0^{u(x)} \frac{dt}{\sqrt{F(\rho) - F(t)}} = \sqrt{2\lambda}x; \quad x \in [0, \tfrac{1}{2}]. \tag{3-6}$$

Substituting $x = \frac{1}{2}$ and utilizing $u(\frac{1}{2}) = \rho$, (3-6) can be written as

$$G_1(\rho) = \sqrt{2}\int_0^{\rho} \frac{dt}{\sqrt{F(\rho) - F(t)}} = \sqrt{\lambda}. \tag{3-7}$$

Therefore, if $u(x)$ is a positive solution to (3-1) where $\|u\|_\infty = \rho$, then ρ must fulfill $G_1(\rho) = \sqrt{\lambda}$.

($\Leftarrow$) Assume $G_1(\rho) = \sqrt{\lambda}$ for $\rho > 0$. Now define a function $u : [0, \frac{1}{2}] \to [0, \infty)$ by

$$\int_0^{u(x)} \frac{dt}{\sqrt{F(\rho) - F(t)}} = \sqrt{2\lambda}x; \quad x \in [0, \tfrac{1}{2}]. \tag{3-8}$$

We now show that $u(x)$ satisfies (3-1). Notice that $u(x)$ is well defined and via the implicit function theorem also twice differentiable. Hence, differentiating (3-8) yields

$$u'(x) = \sqrt{2\lambda\big[F(\rho) - F(u(x))\big]}.$$

By differentiating a second time we obtain

$$-u''(x) = \lambda f(u(x)).$$

In addition, it is clear that $u(0) = 0$. By defining $u(x)$ as a symmetric solution on $[0, 1]$, it is apparent that $u(x)$ is a positive solution to (3-1) with $\|u\|_\infty = \rho$. □

It is important to discern that $G_1(\rho)$ is well defined and the improper integral is convergent. To that end, we state an important remark.

Remark 2. The improper integral in (3-7) is both well defined and convergent for ρ-values that fulfill:

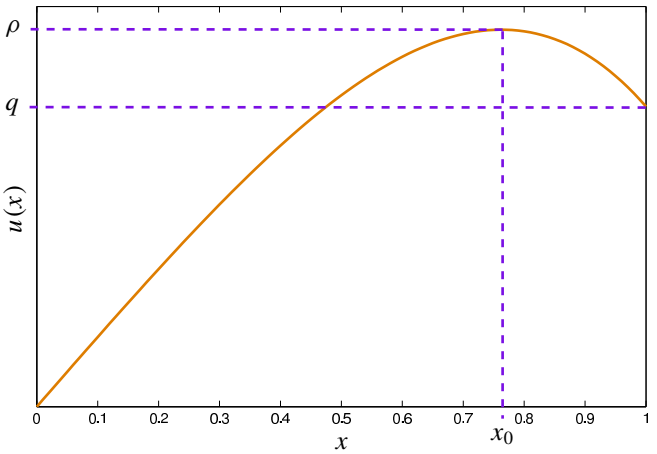


Figure 7. Graph of a typical positive solution of (4-1).

(1) $f(\rho) > 0$;

(2) $F(\rho) > F(s)$ for all $s \in [0, \rho)$.

Notice that from Figure 3 if $c \in [0, c_0)$, then both (1) and (2) will be satisfied for all $\rho \in (0, \sigma(b, c))$. We close this section by recalling an important result from Brown et al.

Theorem 3.2 [Brown et al. 1981]. *$G_1(\rho)$ is both differentiable and continuous on the defined set $T = \{\rho > 0 \mid f(\rho) > 0$ and $F(\rho) - F(s) > 0$ for all $s \in [0, \rho)\}$ where*

$$G_1'(\rho) = \sqrt{2} \int_0^1 \frac{H(\rho) - H(\rho v)}{[F(\rho) - F(\rho v)]^{3/2}}\, dv,$$

in which

$$H(s) = F(s) - \frac{s}{2} f(s).$$

4. Quadrature method for (1-9)

In this section, we adapt the quadrature method to analyze the structure of positive solutions of (1-9):

$$-u'' = \lambda \left[u(u+1)(b-u) - \frac{cu^2}{(u^2+1)} \right]; \ (0,1), \quad u(0) = 0, \quad u'(1) = -1. \ (4\text{-}1)$$

Define

$$f(u) = \left[u(u+1)(b-u) - \frac{cu^2}{(u^2+1)} \right] \quad \text{and} \quad F(x) = \int_0^x f(s)\, ds.$$

It is apparent that a positive solution of (4-1) must resemble Figure 7.

Assume $u(x)$ is a positive solution to (4-1) with $\|u\|_\infty = \rho$ and $u(1) = q$ for $q \in [0, \rho)$. By multiplying the differential equation in (4-1) by $u'(x)$ we obtain

$$-\left[\frac{[u'(x)]^2}{2}\right]' = \big[\lambda F(u(x))\big]'. \tag{4-2}$$

Integrating both sides of (4-2) yields

$$\frac{-(u'(x))^2}{2} = \lambda F(u(x)) + C. \tag{4-3}$$

Recalling that $u(x_0) = \rho$ and $u'(x_0) = 0$, (4-3) becomes

$$C = -\lambda F(\rho). \tag{4-4}$$

Similarly, using $u(1) = q$ and $u'(1) = -1$, (4-3) is utilized to determine a second value for C,

$$C = -\tfrac{1}{2} - \lambda F(q). \tag{4-5}$$

Combining (4-4) and (4-5) gives

$$\sqrt{2\lambda} = \frac{1}{\sqrt{F(\rho) - F(q)}}. \tag{4-6}$$

In utilizing the C-value from (4-4) while solving for $u'(x)$, (4-3) becomes

$$u'(x) = \sqrt{2\lambda\big[F(\rho) - F(u(x))\big]}; \qquad x \in [0, x_0], \tag{4-7}$$

$$u'(x) = -\sqrt{2\lambda\big[F(\rho) - F(u(x))\big]}; \quad x \in [x_0, 1]. \tag{4-8}$$

Rearranging (4-7) and (4-8) gives

$$\frac{u'(x)}{\sqrt{F(\rho) - F(u(x))}} = \sqrt{2\lambda}; \qquad x \in [0, x_0), \tag{4-9}$$

$$\frac{u'(x)}{\sqrt{F(\rho) - F(u(x))}} = -\sqrt{2\lambda}; \quad x \in (x_0, 1]. \tag{4-10}$$

Integration of (4-9) from 0 to x and (4-10) from x_0 to x yields

$$\int_0^x \frac{u'(x)}{\sqrt{F(\rho) - F(u(x))}} = \int_0^x \sqrt{2\lambda}; \qquad x \in [0, x_0), \tag{4-11}$$

$$\int_{x_0}^x \frac{u'(x)}{\sqrt{F(\rho) - F(u(x))}} = \int_{x_0}^x -\sqrt{2\lambda}; \quad x \in (x_0, 1]. \tag{4-12}$$

Using a change of variables and recalling $u(0) = 0$ and $u(x_0) = \rho$ we obtain

$$\int_0^{u(x)} \frac{dw}{\sqrt{F(\rho) - F(w)}} = \sqrt{2\lambda}x; \qquad x \in [0, x_0], \tag{4-13}$$

$$\int_\rho^{u(x)} \frac{dw}{\sqrt{F(\rho) - F(w)}} = -\sqrt{2\lambda}(x - x_0); \quad x \in [x_0, 1]. \tag{4-14}$$

By substituting $x = x_0$ into (4-13) and $x = 1$ into (4-14) we obtain

$$\int_0^{\rho} \frac{dw}{\sqrt{F(\rho) - F(w)}} = \sqrt{2\lambda}x_o, \tag{4-15}$$

$$\int_\rho^{q} \frac{dw}{\sqrt{F(\rho) - F(w)}} = -\sqrt{2\lambda}(1 - x_0). \tag{4-16}$$

Subtracting (4-16) from (4-15) we have

$$\sqrt{2}\int_0^{\rho} \frac{dw}{\sqrt{F(\rho) - F(w)}} - \frac{1}{\sqrt{2}}\int_0^{q} \frac{dw}{\sqrt{F(\rho) - F(w)}} = \sqrt{\lambda}. \tag{4-17}$$

By synthesizing (4-6) with (4-17) we denote

$$\tilde{G}_2(\rho, q) = \sqrt{2}\int_0^{\rho} \frac{dw}{\sqrt{F(\rho) - F(w)}} - \frac{1}{\sqrt{2}}\int_0^{q} \frac{dw}{\sqrt{F(\rho) - F(w)}} - \frac{1}{\sqrt{2}\sqrt{F(\rho) - F(q)}}. \tag{4-18}$$

By Remark 2, the improper integral in $\tilde{G}_2(\rho, q)$ exists and is convergent for ρ in $(0, \sigma(b, c))$. Also, for a given $\rho \in (0, \sigma(b, c))$ Picard's existence and uniqueness theorem guarantees that the corresponding $q = u(1) \in [0, \rho)$ must be unique. If for each $\rho \in (0, \sigma(b, c))$ there exists a unique $q(\rho) \in [0, \rho)$ where $\tilde{G}_2(\rho, q(\rho)) = 0$, then there exists a unique $\lambda \in (0, \infty)$ such that

$$\sqrt{2}\int_0^{\rho} \frac{ds}{\sqrt{F(\rho) - F(s)}} - \frac{1}{\sqrt{2}}\int_0^{q(\rho)} \frac{ds}{\sqrt{F(\rho) - F(s)}} = \frac{1}{\sqrt{2}\sqrt{F(\rho) - F(q(\rho))}} = \sqrt{\lambda} \tag{4-19}$$

will be satisfied. Therefore it is imperative to examine the existence and uniqueness of such a $q = q(\rho)$. Hence, we recall and prove Lemma 1, adapted from [Goddard et al. 2010a], which outlines necessary properties of $\tilde{G}_2(\rho, q)$.

Lemma 1 [Goddard et al. 2010a]. *If* $\rho \in (0, \sigma(b, c))$ *then*:

(1) $\tilde{G}_2(\rho, q) \to -\infty$ *as* $q \to \rho^-$ *for fixed* $\rho \in (0, \sigma(b, c))$.

(2) $[\tilde{G}_2]_q < 0$ *for every* $q \in [0, \rho)$ *and fixed* $\rho \in (0, \sigma(b, c))$.

(3) $\tilde{G}_2(\rho, 0) \to \infty$ *when* $\rho \to \sigma(b, c)^-$.

(4) $\tilde{G}_2(\rho, 0) \to -\infty$ *when* $\rho \to 0^+$.

Proof. (1) Accomplished via the mean value theorem and the fact that $F(u)$ is an increasing function on $(0, \sigma(b, c))$.

(2) Let $\rho \in (0, \sigma(b, c))$. Thus

$$[\tilde{G}_2(\rho, q)]_q = -\frac{1}{\sqrt{2}\sqrt{F(\rho) - F(q)}} - \frac{f(q)}{2\sqrt{2}[F(\rho) - F(q)]^{\frac{3}{2}}} < 0$$

for all $q \in [0, \rho)$, since $f(s) > 0$ for $s \in (0, \sigma(b, c))$.

(3) For every $\rho \in (0, \sigma(b, c))$, we have

$$\tilde{G}_2(\rho, 0) = \sqrt{2} \int_0^\rho \frac{ds}{\sqrt{F(\rho) - F(s)}} - \frac{1}{\sqrt{2}\sqrt{F(\rho)}} = G_1(\rho) - \frac{1}{\sqrt{2}\sqrt{F(\rho)}}. \quad (4\text{-}20)$$

Laetsch [1970] showed that $G_1(\rho) \to \infty$ as $\rho \to \sigma(b, c)^-$. This implies that $\tilde{G}_2(\rho, 0) \to \infty$ when $\rho \to \sigma(b, c)^-$.

(4) Ascertained via the mean value theorem and the monotonicity of $F(u)$ on $(0, \sigma(b, c))$. □

According to Lemma 1, $\tilde{G}_2(\rho, q)$ must resemble Figure 8, whereas Figures 9 and 10 illustrate $\tilde{G}_2(\rho, 0)$. Noteworthy from Lemma 1, if $\tilde{G}_2(\rho, 0) \geq 0$ then there exists a unique $q(\rho) \in [0, \rho)$ wherefore $\tilde{G}_2(\rho, q(\rho)) = 0$. We conjecture as a result of our computations that there is a unique $\rho^* = \rho^*(b, c) > 0$ wherefore if $\rho \geq \rho^*$, then $\tilde{G}_2(\rho, 0) \geq 0$. Also if $\rho < \rho^*$ then $\tilde{G}_2(\rho, 0) < 0$. So, for all $\rho \in [\rho^*, \infty)$ there exists a unique $q = q(\rho) \in [0, \rho)$ where $\tilde{G}_2(\rho, q(\rho)) = 0$. In this case, we have

$$G_2(\rho, q(\rho)) = \frac{1}{\sqrt{2}\sqrt{F(\rho) - F(q)}} = \sqrt{\lambda}. \quad (4\text{-}21)$$

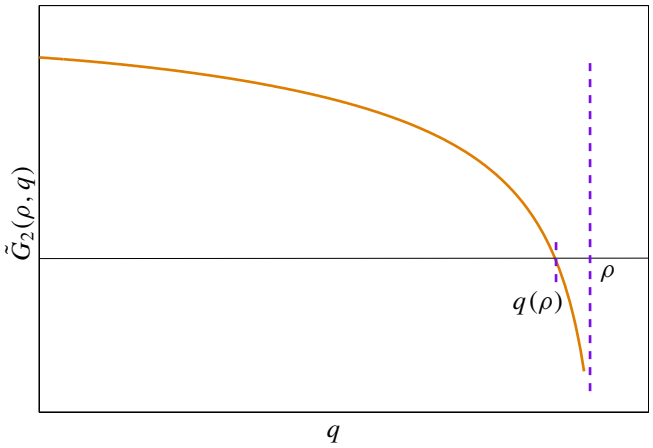


Figure 8. Graph of $\tilde{G}_2(\rho, q)$.

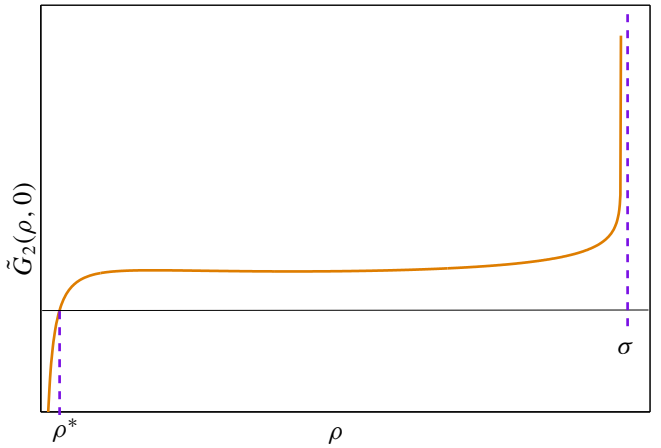


Figure 9. Graph of $\tilde{G}_2(\rho, 0)$ when $b = 10$ and $c = 0$.

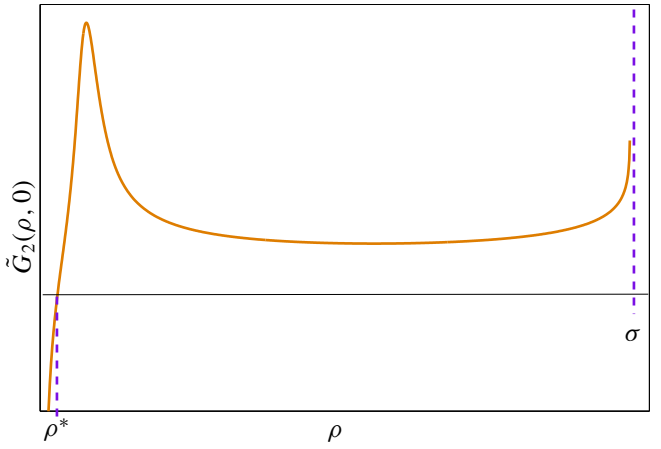


Figure 10. Graph of $\tilde{G}_2(\rho, 0)$ when $b = 10$ and $c = 33$.

We now state and prove the main theorem of the section.

Theorem 4.1. *The function $u(x)$ is a positive solution to* (4-1) *with*

$$\|u\|_\infty = \rho \in S(b, c) := \left[\rho^*(b, c), \sigma(b, c)\right)$$

if and only if

$$G_2(\rho, q(\rho)) = \frac{1}{\sqrt{2}\sqrt{F(\rho) - F(q)}} = \sqrt{\lambda}$$

for a positive λ for which $q = q(\rho) \in [0, \rho)$ is the unique solution of

$$\tilde{G}_2(\rho, q(\rho)) = \sqrt{2}\int_0^\rho \frac{dw}{\sqrt{F(\rho) - F(w)}} - \frac{1}{\sqrt{2}}\int_0^q \frac{dw}{\sqrt{F(\rho) - F(w)}} - \frac{1}{\sqrt{2}\sqrt{F(\rho) - F(q)}} = 0.$$

Proof. ($\Rightarrow$) Accomplished in the above analysis.

($\Leftarrow$) Assume that there exist $\lambda \in (0, \infty)$ and $\rho \in S(b, c)$ wherefore $G_2(\rho, q(\rho)) = \sqrt{\lambda}$ in which the unique solution of $\tilde{G}_2(\rho, q(\rho)) = 0$ is $q(\rho) \in [0, \rho)$. Define

$$u(x) : [0, 1] \to \mathbb{R}$$

via

$$\int_0^{u(x)} \frac{ds}{\sqrt{F(\rho) - F(s)}} = \sqrt{2\lambda}x; \qquad x \in [0, x_0], \tag{4-22}$$

$$\int_\rho^{u(x)} \frac{ds}{\sqrt{F(\rho) - F(s)}} = -\sqrt{2\lambda}(x - x_0); \quad x \in [x_0, 1]. \tag{4-23}$$

Now, we will exhibit $u(x)$ as a positive solution to (4-1). Note that $u(x)$ has a turning point at x_0 denoted by

$$x_0 = \frac{1}{\sqrt{2\lambda}} \int_0^{\rho} \frac{ds}{\sqrt{F(\rho) - F(s)}}. \tag{4-24}$$

For the given $\lambda > 0$, it is apparent that

$$\frac{1}{\sqrt{2\lambda}} \int_0^{u(x)} \frac{ds}{\sqrt{F(\rho) - F(s)}} \tag{4-25}$$

is both a differentiable function of u and an increasing function ranging from 0 to x_0 when u takes on the values from 0 to ρ. Therefore, for each $x \in [0, x_0]$ there is a unique $u(x)$ wherefore

$$\int_0^{u(x)} \frac{ds}{\sqrt{F(\rho) - F(s)}} = \sqrt{2\lambda}x. \tag{4-26}$$

The implicit function theorem gives that $u(x)$ is a twice-differentiable function with respect to x. Differentiating (4-26) with respect to x gives

$$u'(x) = \sqrt{2\lambda\big[F(\rho) - F(u(x))\big]}; \quad x \in [0, x_0]. \tag{4-27}$$

Through a similar argument,

$$u'(x) = -\sqrt{2\lambda\big[F(\rho) - F(u(x))\big]}; \quad x \in [x_0, 1]. \tag{4-28}$$

By utilizing (4-27) and (4-28) we obtain

$$\frac{[u'(x)]^2}{2} = \lambda\big[F(\rho) - F(u(x))\big]; \quad x \in [0, 1]. \tag{4-29}$$

Through differentiation of (4-29) we have

$$-u''u' = \lambda f(u)u'; \quad x \in (0, 1), \tag{4-30}$$

which can be rewritten as

$$-u'' = \lambda f(u); \quad x \in (0, 1). \tag{4-31}$$

Thus, we have proved that $u(x)$ satisfies the differential equation in (4-1). Now, we show that $u(x)$ satisfies the boundary value conditions in (4-1); however, it is apparent that $u(0) = 0$. Additionally, using $G_2(\rho, q(\rho)) = \sqrt{\lambda}$, we ascertain

$$\sqrt{F(\rho) - F(q)} = \frac{1}{\sqrt{2\lambda}}. \tag{4-32}$$

Substitution of $x = 1$ in (4-28) yields

$$u'(1) = -\sqrt{2\lambda}\sqrt{F(\rho) - F(q)}. \tag{4-33}$$

When (4-32) and (4-33) are synthesized we obtain

$$u'(1) = -1. \tag{4-34}$$

Therefore, the boundary conditions in (4-1) are satisfied by $u(x)$. □

5. Quadrature method for (1-11)

Further extension of the quadrature method is performed in this section to analyze the structure of positive solutions of (1-11):

$$-u'' = \lambda\left[u(u+1)(b-u) - \frac{cu^2}{(u^2+1)}\right]; (0, 1), \quad u'(0) = 1, \quad u'(1) = -1. \tag{5-1}$$

Define

$$f(u) = \left[u(u+1)(b-u) - \frac{cu^2}{(u^2+1)}\right] \quad \text{and} \quad F(x) = \int_0^x f(s)\,ds.$$

Clearly, a positive solution of (5-1) must resemble Figure 11, where $\|u\|_\infty = \rho$, $\rho \in (0, \infty)$, $q = u(0) = u(1)$, and $q \in [0, \rho)$. Through a similar argument as in Section 4, we articulate the main theorem of this section.

Theorem 5.1. *The function $u(x)$ is a positive solution of* (5-1) *with*

$$\|u\|_\infty = \rho \in S(b, c) = \left[\rho^*(b, c), \sigma(b, c)\right)$$

if and only if

$$G_3(\rho, q(\rho)) = \frac{1}{\sqrt{2}\sqrt{F(\rho) - F(q)}} = \sqrt{\lambda}, \tag{5-2}$$

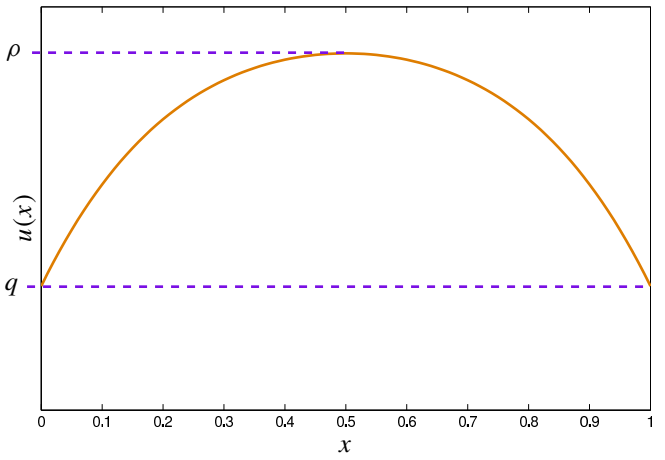


Figure 11. Graph of a typical positive solution of (5-1).

for which $q = q(\rho) \in [0, \rho)$ *is the unique solution of*

$$\tilde{G}_3(\rho, q(\rho)) = \sqrt{2}\int_0^{\rho} \frac{dw}{\sqrt{F(\rho) - F(w)}} - \sqrt{2}\int_0^{q} \frac{dw}{\sqrt{F(\rho) - F(w)}} - \frac{1}{\sqrt{2}\sqrt{F(\rho) - F(q)}} = 0.$$

6. Computational results

Within this section we exhibit the complete evolution of the bifurcation curve of positive solutions of (1-7) for $c \in [0, c_0(b))$. The results for (1-8) are reestablished via Mathematica computations and by recalling Theorem 3.1. For (1-9) and (1-11), we recall Theorems 4.1 and 5.1 and utilize a standard root-finding algorithm to find the unique $\rho^*(b, c) > 0$. Then for $\rho \in [\rho^*(b, c), \sigma(b, c))$ we employ a root-finding algorithm to find the corresponding unique $q(\rho)$, which is delineated in Theorems 4.1 and 5.1. These diagrams were acquired via Mathematica for a single b-value as c-values are varied. If $b \in (b_0, \infty)$ then there exist

$$0 < c_0^* < c_1^* < c_2^* < c_3^* < c_4^* < c_5^* < c_6^* < c_7^* < c_0(b)$$

such that we have the following cases. In the subsequent figures, (1-8) is represented in black, cases (1-9) and (1-10) in red, and (1-11) in blue.

Case 1. If $c \in [0, c_0^*)$ then there exist $\lambda_i > 0$ for $i = 1, 2, 3, 4$ such that if

- $\lambda \in [0, \lambda_2)$, then (1-7) has no positive solution;
- $\lambda = \lambda_2$, then (1-7) has a unique positive solution;
- $\lambda \in (\lambda_2, \lambda_3)$, then (1-7) has exactly 2 positive solutions;
- $\lambda = \lambda_3$ and $\lambda \in [\lambda_0, \infty)$, then (1-7) has exactly 4 positive solutions;

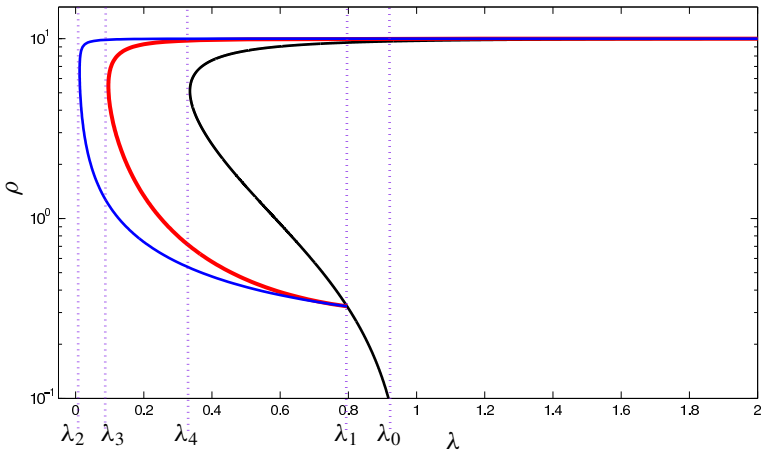


Figure 12. ρ versus λ when $b = 10$ and $c = 0$ (Case 1).

- $\lambda \in (\lambda_1, \lambda_0)$, then (1-7) has exactly 5 positive solutions;
- $\lambda \in (\lambda_3, \lambda_4)$, then (1-7) has exactly 6 positive solutions;
- $\lambda = \lambda_4$, then (1-7) has exactly 7 positive solutions;
- $\lambda \in (\lambda_4, \lambda_1]$, then (1-7) has exactly 8 positive solutions.

Figure 12 illustrates Case 1.

Case 2. If $c \in [c_0, c_1)$ (some $c_1(b) > 0$) then there exist $\lambda_i > 0$ for $i = 1, 2, \ldots, 6$ such that if

- $\lambda \in [0, \lambda_2)$, then (1-7) has no positive solution;
- $\lambda = \lambda_2$, then (1-7) has a unique positive solution;
- $\lambda \in (\lambda_2, \lambda_3)$, then (1-7) has exactly 2 positive solutions;
- $\lambda = \lambda_3$ and $\lambda \in [\lambda_0, \infty)$, then (1-7) has exactly 4 positive solutions;
- $\lambda \in (\lambda_1, \lambda_5)$ and $\lambda \in (\lambda_6, \lambda_0)$, then (1-7) has exactly 5 positive solutions;
- $\lambda \in (\lambda_3, \lambda_4)$, $\lambda = \lambda_5$, and $\lambda = \lambda_6$, then (1-7) has exactly 6 positive solutions;
- $\lambda = \lambda_4$ and $\lambda \in (\lambda_5, \lambda_6)$, then (1-7) has exactly 7 positive solutions;
- $\lambda \in (\lambda_4, \lambda_1]$, then (1-7) has exactly 8 positive solutions.

Figure 13 illustrates Case 2.

Case 3. If $c = c_1$ then there exist $\lambda_i > 0$ for $i = 1, 2, \ldots, 5$ such that if

- $\lambda \in [0, \lambda_2)$, then (1-7) has no positive solution;
- $\lambda = \lambda_2$, then (1-7) has a unique positive solution;
- $\lambda \in (\lambda_2, \lambda_3)$, then (1-7) has exactly 2 positive solutions;

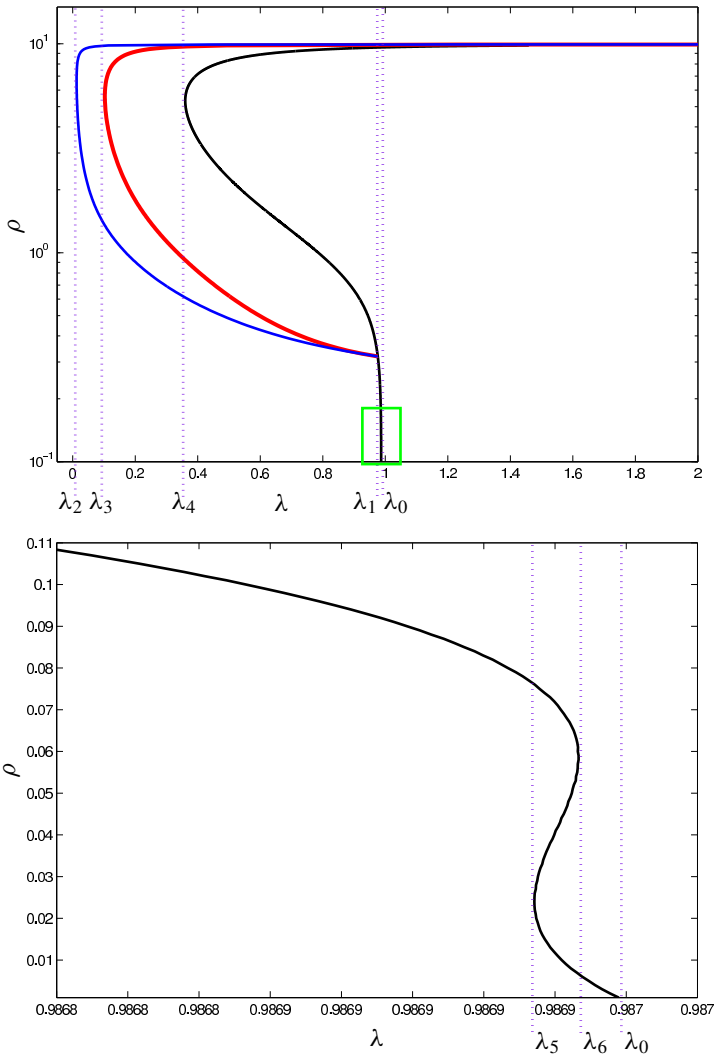


Figure 13. ρ versus λ (top) and cross-section (bottom) for $b = 10$ and $c = 8.97$ (Case 2).

- $\lambda = \lambda_3$ and $\lambda \in [\lambda_0, \infty)$, then (1-7) has exactly 4 positive solutions;
- $\lambda \in (\lambda_1, \lambda_5)$ and $\lambda = \lambda_0$, then (1-7) has exactly 5 positive solutions;
- $\lambda \in (\lambda_3, \lambda_4)$, $\lambda = \lambda_5$, then (1-7) has exactly 6 positive solutions;
- $\lambda = \lambda_4$ and $\lambda \in (\lambda_5, \lambda_0)$, then (1-7) has exactly 7 positive solutions;
- $\lambda \in (\lambda_4, \lambda_1]$, then (1-7) has exactly 8 positive solutions.

Figure 14 illustrates Case 3.

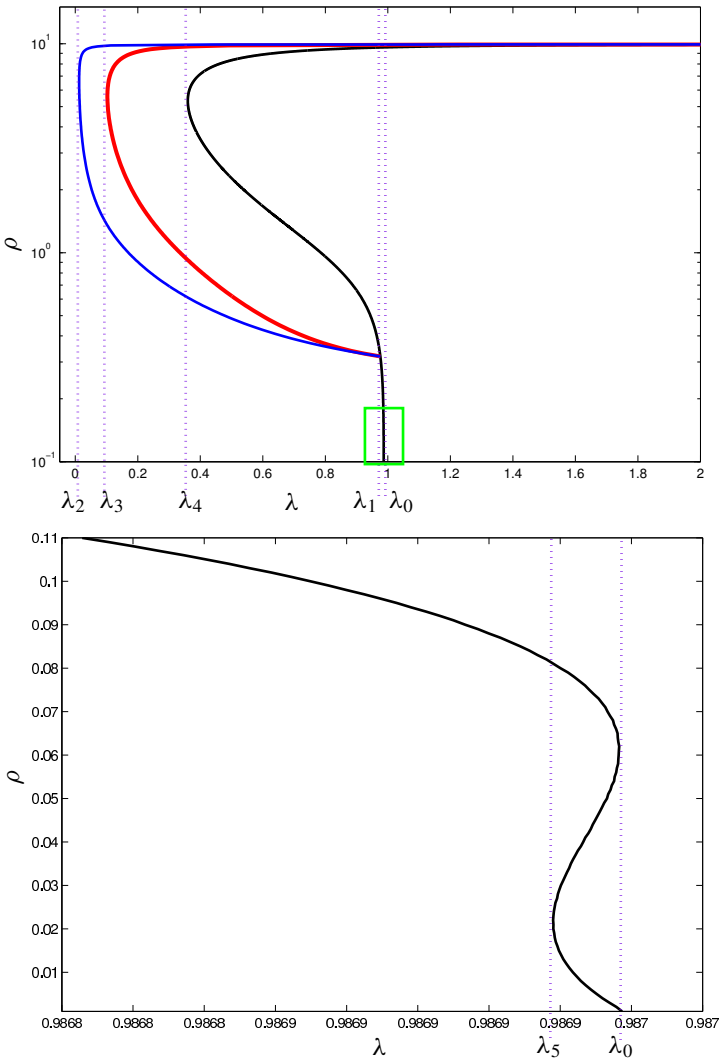


Figure 14. ρ versus λ (top) and cross-section (bottom) for $b = 10$ and $c = 8.972$ (Case 3).

Case 4. If $c \in (c_1, c_2)$ (some $c_2(b) > 0$) then there exist $\lambda_i > 0$ for $i = 1, 2, \ldots, 6$ such that if

- $\lambda \in [0, \lambda_2)$, then (1-7) has no positive solution;
- $\lambda = \lambda_2$, then (1-7) has a unique positive solution;
- $\lambda \in (\lambda_2, \lambda_3)$, then (1-7) has exactly 2 positive solutions;
- $\lambda = \lambda_3$ and $\lambda \in (\lambda_6, \infty)$, then (1-7) has exactly 4 positive solutions;
- $\lambda \in (\lambda_1, \lambda_5)$ and $\lambda = \lambda_6$, then (1-7) has exactly 5 positive solutions;

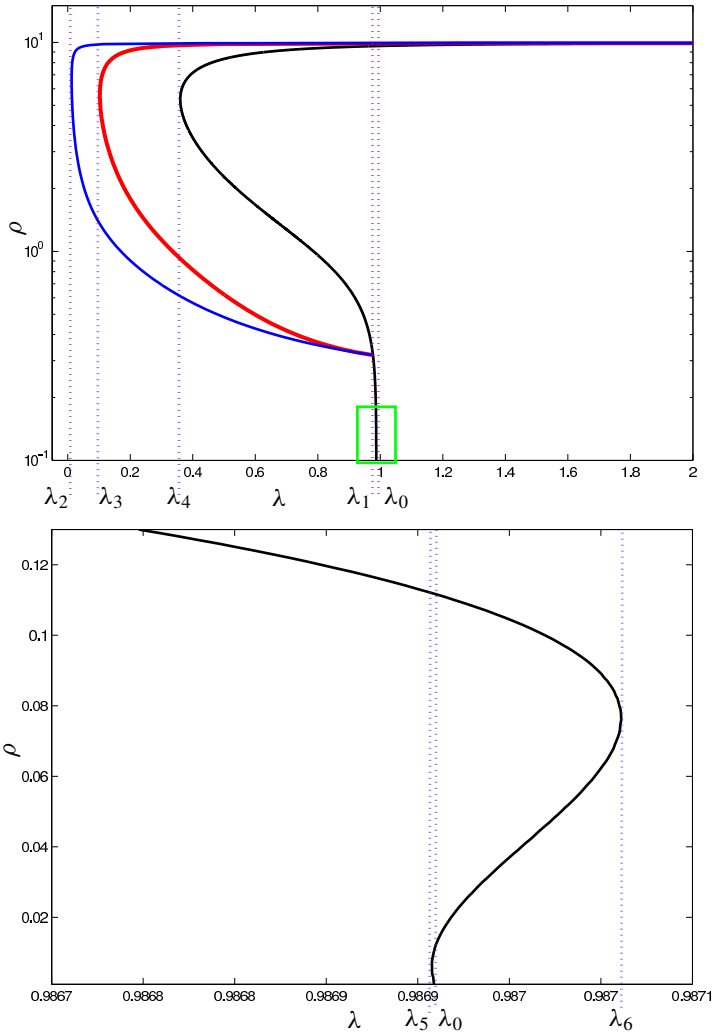


Figure 15. ρ versus λ (top) and cross-section (bottom) for $b = 10$ and $c = 8.99$ (Case 4).

- $\lambda \in (\lambda_3, \lambda_4)$, $\lambda \in [\lambda_0, \lambda_6)$, and $\lambda = \lambda_5$, then (1-7) has exactly 6 positive solutions;
- $\lambda = \lambda_4$ and $\lambda \in (\lambda_5, \lambda_0)$, then (1-7) has exactly 7 positive solutions;
- $\lambda \in (\lambda_4, \lambda_1]$, then (1-7) has exactly 8 positive solutions.

Figure 15 illustrates Case 4.

Case 5. If $c = c_2$ then there exist $\lambda_i > 0$ for $i = 1, 2, \ldots, 5$ such that if

- $\lambda \in [0, \lambda_2)$, then (1-7) has no positive solution;
- $\lambda = \lambda_2$, then (1-7) has a unique positive solution;

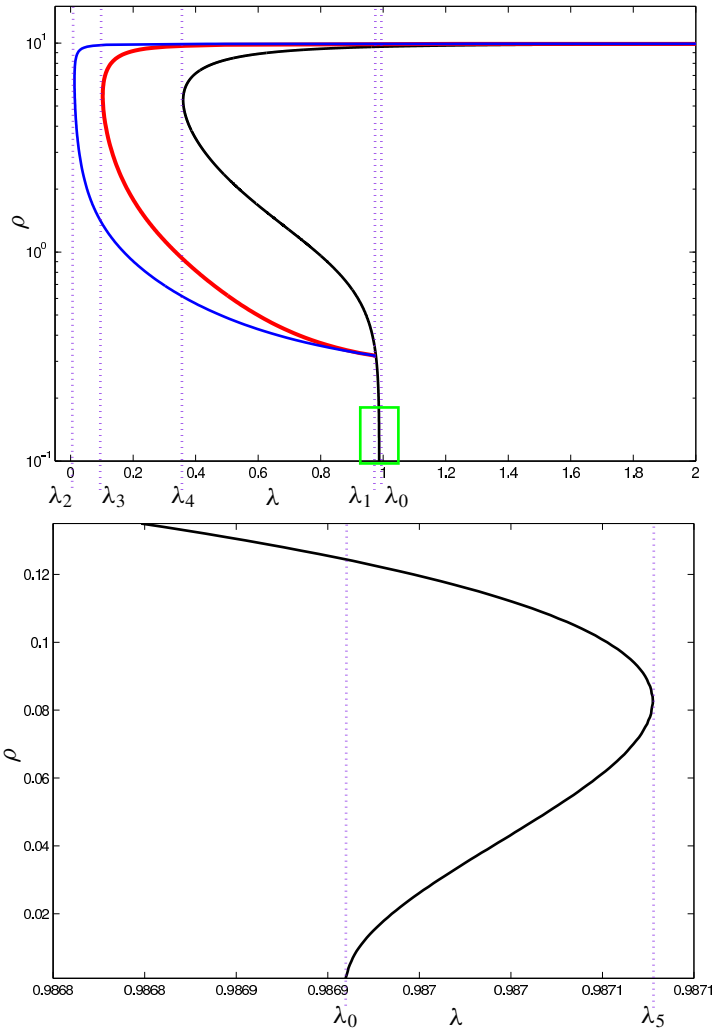


Figure 16. ρ versus λ (top) and cross-section (bottom) for $b = 10$ and $c = 9$ (Case 5).

- $\lambda \in (\lambda_2, \lambda_3)$, then (1-7) has exactly 2 positive solutions;
- $\lambda = \lambda_3$ and $\lambda \in (\lambda_5, \infty)$, then (1-7) has exactly 4 positive solutions;
- $\lambda \in (\lambda_1, \lambda_0]$ and $\lambda = \lambda_5$, then (1-7) has exactly 5 positive solutions;
- $\lambda \in (\lambda_3, \lambda_4)$ and $\lambda \in (\lambda_0, \lambda_5)$, then (1-7) has exactly 6 positive solutions;
- $\lambda = \lambda_4$, then (1-7) has exactly 7 positive solutions;
- $\lambda \in (\lambda_4, \lambda_1]$, then (1-7) has exactly 8 positive solutions.

Figure 16 illustrates Case 5.

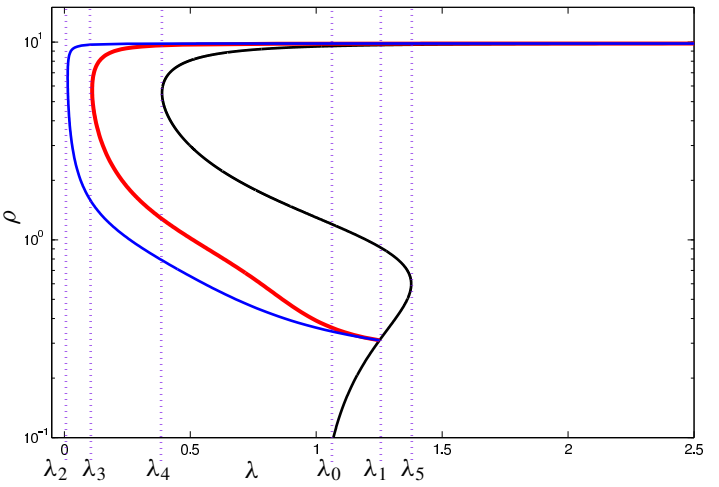


Figure 17. ρ versus λ when $b = 10$ and $c = 18$ (Case 6).

Case 6. If $c \in (c_2, c_3]$ (some $c_3(b) > 0$) then there exist $\lambda_i > 0$ for $i = 1, 2, \ldots, 5$ such that if

- $\lambda \in [0, \lambda_2)$, then (1-7) has no positive solution;
- $\lambda = \lambda_2$, then (1-7) has a unique positive solution;
- $\lambda \in (\lambda_2, \lambda_3)$, then (1-7) has exactly 2 positive solutions;
- $\lambda = \lambda_3$ and $\lambda \in (\lambda_5, \infty)$, then (1-7) has exactly 4 positive solutions;
- $\lambda = \lambda_5$, then (1-7) has exactly 5 positive solutions;
- $\lambda \in (\lambda_3, \lambda_4)$, (λ_1, λ_5), then (1-7) has exactly 6 positive solutions;
- $\lambda = \lambda_4$, then (1-7) has exactly 7 positive solutions;
- $\lambda \in (\lambda_4, \lambda_0]$, then (1-7) has exactly 8 positive solutions;
- $\lambda \in (\lambda_0, \lambda_1]$, then (1-7) has exactly 9 positive solutions.

Figure 17 illustrates Case 6.

Case 7. If $c \in (c_3, c_4)$ (some $c_4 > 0$) then there exist $\lambda_i > 0$ for $i = 1, 2, \ldots, 7$ such that if

- $\lambda \in [0, \lambda_2)$, then (1-7) has no positive solution;
- $\lambda = \lambda_2$, then (1-7) has a unique positive solution;
- $\lambda \in (\lambda_2, \lambda_3)$, then (1-7) has exactly 2 positive solutions;
- $\lambda = \lambda_3$ and $\lambda \in (\lambda_7, \infty)$, then (1-7) has exactly 4 positive solutions;
- $\lambda = \lambda_7$, then (1-7) has exactly 5 positive solutions;
- $\lambda \in (\lambda_3, \lambda_4)$, (λ_1, λ_7), then (1-7) has exactly 6 positive solutions;

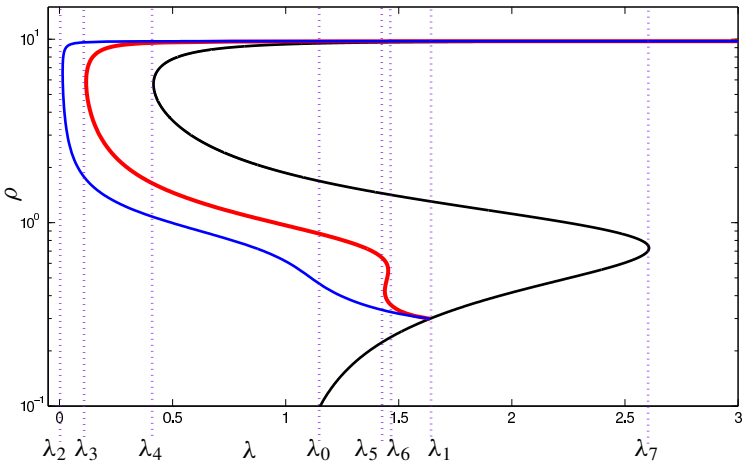


Figure 18. ρ versus λ when $b = 10$ and $c = 26$ (Case 7).

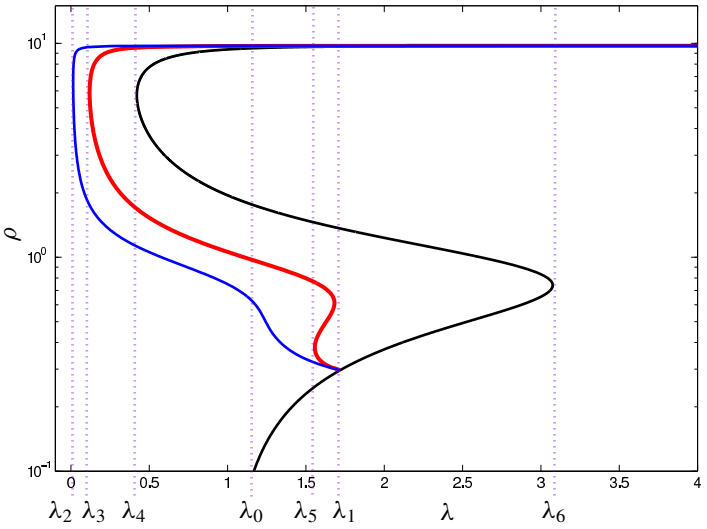


Figure 19. ρ versus λ when $b = 10$ and $c = 27.3$ (Case 8).

- $\lambda = \lambda_4$, then (1-7) has exactly 7 positive solutions;
- $\lambda \in (\lambda_4, \lambda_0]$, then (1-7) has exactly 8 positive solutions;
- $\lambda \in (\lambda_0, \lambda_5)$, $(\lambda_6, \lambda_1]$, then (1-7) has exactly 9 positive solutions;
- $\lambda = \lambda_5$, $\lambda = \lambda_6$, then (1-7) has exactly 11 positive solutions;
- $\lambda \in (\lambda_5, \lambda_6)$, then (1-7) has exactly 13 positive solutions.

Figure 18 illustrates Case 7. Notice that the red curve has become Σ-shaped and this shape persists through $c \leq c_0(b)$.

Case 8. If $c = c_4$ then there exist $\lambda_i > 0$ for $i = 1, 2, \ldots, 6$ such that if

- $\lambda \in [0, \lambda_2)$, then (1-7) has no positive solution;
- $\lambda = \lambda_2$, then (1-7) has a unique positive solution;
- $\lambda \in (\lambda_2, \lambda_3)$, then (1-7) has exactly 2 positive solutions;
- $\lambda = \lambda_3$ and $\lambda \in (\lambda_6, \infty)$, then (1-7) has exactly 4 positive solutions;
- $\lambda = \lambda_6$, then (1-7) has exactly 5 positive solutions;
- $\lambda \in (\lambda_3, \lambda_4), (\lambda_1, \lambda_6)$, then (1-7) has exactly 6 positive solutions;
- $\lambda = \lambda_4$, then (1-7) has exactly 7 positive solutions;
- $\lambda \in (\lambda_4, \lambda_0]$, then (1-7) has exactly 8 positive solutions;
- $\lambda \in (\lambda_0, \lambda_5)$, then (1-7) has exactly 9 positive solutions;
- $\lambda = \lambda_1, \lambda = \lambda_5$, then (1-7) has exactly 11 positive solutions;
- $\lambda \in (\lambda_5, \lambda_1)$, then (1-7) has exactly 13 positive solutions.

Figure 19 illustrates Case 8.

Case 9. If $c \in (c_4, c_5]$ (some $c_5(b) > 0$), then there exist $\lambda_i > 0$ for $i = 1, 2, \ldots, 7$ such that if

- $\lambda \in [0, \lambda_2)$, then (1-7) has no positive solution;
- $\lambda = \lambda_2$, then (1-7) has a unique positive solution;
- $\lambda \in (\lambda_2, \lambda_3)$, then (1-7) has exactly 2 positive solutions;
- $\lambda = \lambda_3$ and $\lambda \in (\lambda_7, \infty)$, then (1-7) has exactly 4 positive solutions;
- $\lambda = \lambda_7$, then (1-7) has exactly 5 positive solutions;
- $\lambda \in (\lambda_3, \lambda_4), (\lambda_6, \lambda_7)$, then (1-7) has exactly 6 positive solutions;
- $\lambda = \lambda_4$, then (1-7) has exactly 7 positive solutions;
- $\lambda \in (\lambda_4, \lambda_0], \lambda = \lambda_6$, then (1-7) has exactly 8 positive solutions;
- $\lambda \in (\lambda_0, \lambda_5)$, then (1-7) has exactly 9 positive solutions;
- $\lambda \in (\lambda_1, \lambda_6)$, then (1-7) has exactly 10 positive solutions;
- $\lambda = \lambda_5$, then (1-7) has exactly 11 positive solutions;
- $\lambda \in (\lambda_5, \lambda_1]$, then (1-7) has exactly 13 positive solutions.

Figure 20 illustrates Case 9.

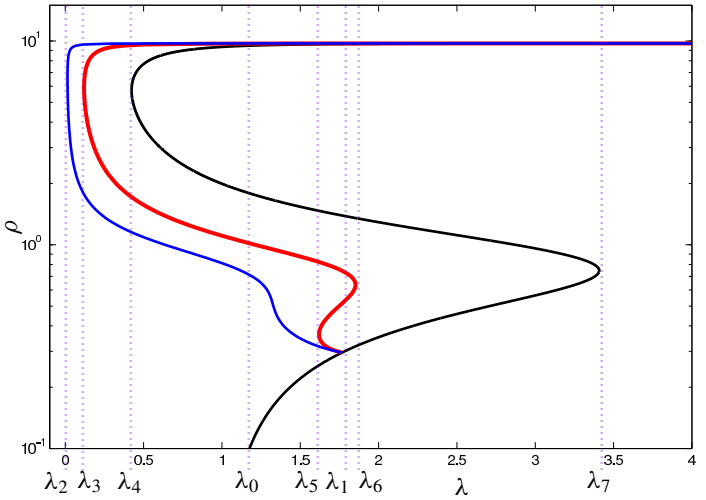


Figure 20. ρ versus λ when $b = 10$ and $c = 28$ (Case 9).

Case 10. If $c \in (c_5, c_6)$ (some $c_6(b) > 0$) then there exist $\lambda_i > 0$ for $i = 1, 2, \ldots, 9$ such that if

- $\lambda \in [0, \lambda_2)$, then (1-7) has no positive solution;
- $\lambda = \lambda_2$, then (1-7) has a unique positive solution;
- $\lambda \in (\lambda_2, \lambda_3)$, then (1-7) has exactly 2 positive solutions;
- $\lambda = \lambda_3$ and $\lambda \in (\lambda_9, \infty)$, then (1-7) has exactly 4 positive solutions;
- $\lambda = \lambda_9$, then (1-7) has exactly 5 positive solutions;

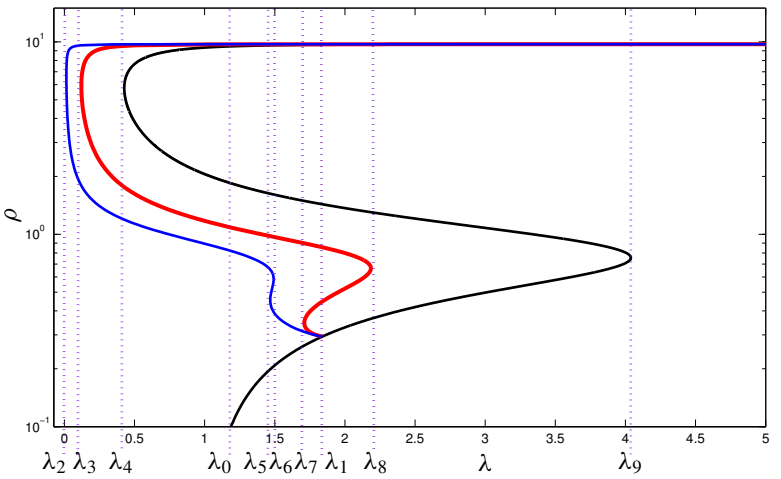


Figure 21. ρ versus λ when $b = 10$ and $c = 29$ (Case 10).

- $\lambda \in (\lambda_3, \lambda_4), (\lambda_8, \lambda_9)$, then (1-7) has exactly 6 positive solutions;
- $\lambda = \lambda_4$, then (1-7) has exactly 7 positive solutions;
- $\lambda \in (\lambda_4, \lambda_0], \lambda = \lambda_8$, then (1-7) has exactly 8 positive solutions;
- $\lambda \in (\lambda_0, \lambda_5), (\lambda_6, \lambda_7)$, then (1-7) has exactly 9 positive solutions;
- $\lambda = \lambda_5, \lambda = \lambda_6, \lambda \in (\lambda_1, \lambda_8)$, then (1-7) has exactly 10 positive solutions;
- $\lambda \in (\lambda_5, \lambda_6), \lambda = \lambda_7$, then (1-7) has exactly 11 positive solutions;
- $\lambda \in (\lambda_7, \lambda_1]$, then (1-7) has exactly 13 positive solutions.

Figure 21 illustrates Case 10. Notice that the blue curve has now also become Σ-shaped and its shape persists through $c \leq c_0(b)$.

Case 11. If $c = c_6$ then there exist $\lambda_i > 0$ for $i = 1, 2, \ldots, 8$ such that if

- $\lambda \in [0, \lambda_2)$, then (1-7) has no positive solution;
- $\lambda = \lambda_2$, then (1-7) has a unique positive solution;
- $\lambda \in (\lambda_2, \lambda_3)$, then (1-7) has exactly 2 positive solutions;
- $\lambda = \lambda_3$ and $\lambda \in (\lambda_8, \infty)$, then (1-7) has exactly 4 positive solutions;
- $\lambda = \lambda_8$, then (1-7) has exactly 5 positive solutions;
- $\lambda \in (\lambda_3, \lambda_4), (\lambda_7, \lambda_8)$, then (1-7) has exactly 6 positive solutions;
- $\lambda = \lambda_4$, then (1-7) has exactly 7 positive solutions;
- $\lambda \in (\lambda_4, \lambda_0], \lambda = \lambda_7$, then (1-7) has exactly 8 positive solutions;
- $\lambda \in (\lambda_0, \lambda_5)$, then (1-7) has exactly 9 positive solutions;

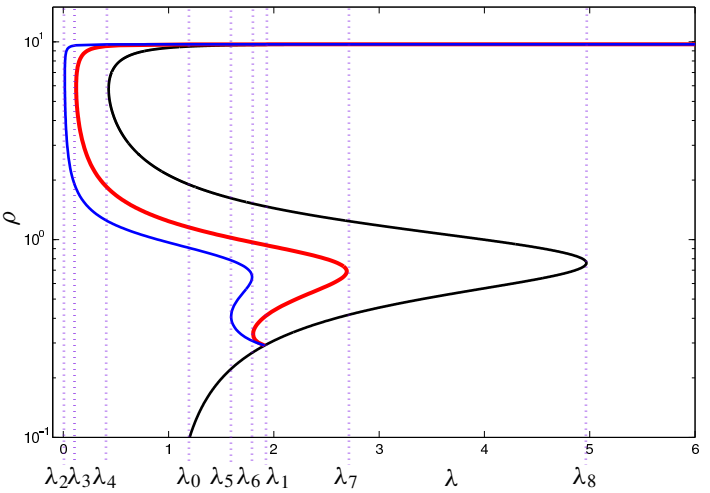


Figure 22. ρ versus λ when $b = 10$ and $c = 30$ (Case 11).

- $\lambda = \lambda_5, \lambda \in (\lambda_1, \lambda_7)$, then (1-7) has exactly 10 positive solutions;
- $\lambda \in (\lambda_5, \lambda_6)$, then (1-7) has exactly 11 positive solutions;
- $\lambda = \lambda_6$, then (1-7) has exactly 12 positive solutions;
- $\lambda \in (\lambda_6, \lambda_1]$, then (1-7) has exactly 13 positive solutions.

Figure 22 illustrates Case 11.

Case 12. If $c \in (c_6, c_7)$ (some $c_7(b) > 0$) then there exist $\lambda_i > 0$ for $i = 1, 2, \ldots, 9$ such that if

- $\lambda \in [0, \lambda_2)$, then (1-7) has no positive solution;
- $\lambda = \lambda_2$, then (1-7) has a unique positive solution;
- $\lambda \in (\lambda_2, \lambda_3)$, then (1-7) has exactly 2 positive solutions;
- $\lambda = \lambda_3$ and $\lambda \in (\lambda_9, \infty)$, then (1-7) has exactly 4 positive solutions;
- $\lambda = \lambda_9$, then (1-7) has exactly 5 positive solutions;
- $\lambda \in (\lambda_3, \lambda_4), (\lambda_8, \lambda_9)$, then (1-7) has exactly 6 positive solutions;
- $\lambda = \lambda_4$, then (1-7) has exactly 7 positive solutions;
- $\lambda \in (\lambda_4, \lambda_0], \lambda = \lambda_8$, then (1-7) has exactly 8 positive solutions;
- $\lambda \in (\lambda_0, \lambda_5)$, then (1-7) has exactly 9 positive solutions;
- $\lambda = \lambda_5, \lambda \in (\lambda_1, \lambda_8)$, then (1-7) has exactly 10 positive solutions;
- $\lambda \in (\lambda_5, \lambda_6)$, then (1-7) has exactly 11 positive solutions;
- $\lambda = \lambda_6, \lambda \in (\lambda_7, \lambda_1]$, then (1-7) has exactly 13 positive solutions;

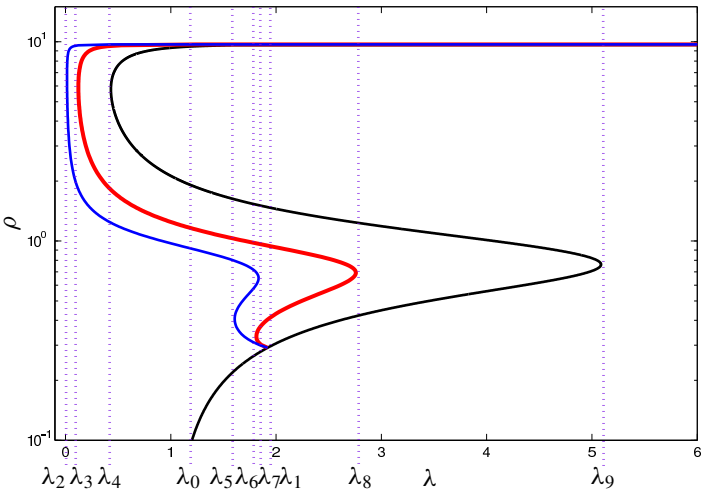


Figure 23. ρ versus λ when $b = 10$ and $c = 30.1$ (Case 12).

- $\lambda = \lambda_7$, then (1-7) has exactly 14 positive solutions;
- $\lambda \in (\lambda_6, \lambda_7)$, then (1-7) has exactly 15 positive solutions.

Figure 23 illustrates Case 12.

Case 13. If $c = c_7$ then there exist $\lambda_i > 0$ for $i = 1, 2, \ldots, 8$ such that if

- $\lambda \in [0, \lambda_2)$, then (1-7) has no positive solution;
- $\lambda = \lambda_2$, then (1-7) has a unique positive solution;
- $\lambda \in (\lambda_2, \lambda_3)$, then (1-7) has exactly 2 positive solutions;
- $\lambda = \lambda_3$ and $\lambda \in (\lambda_8, \infty)$, then (1-7) has exactly 4 positive solutions;
- $\lambda = \lambda_8$, then (1-7) has exactly 5 positive solutions;
- $\lambda \in (\lambda_3, \lambda_4), (\lambda_7, \lambda_8)$, then (1-7) has exactly 6 positive solutions;
- $\lambda = \lambda_4$, then (1-7) has exactly 7 positive solutions;
- $\lambda \in (\lambda_4, \lambda_0]$, $\lambda = \lambda_7$, then (1-7) has exactly 8 positive solutions;
- $\lambda \in (\lambda_0, \lambda_5)$, then (1-7) has exactly 9 positive solutions;
- $\lambda = \lambda_5$, $\lambda \in (\lambda_1, \lambda_7)$, then (1-7) has exactly 10 positive solutions;
- $\lambda \in (\lambda_5, \lambda_6)$, then (1-7) has exactly 11 positive solutions;
- $\lambda = \lambda_6$, then (1-7) has exactly 13 positive solutions;
- $\lambda = \lambda_1$, then (1-7) has exactly 14 positive solutions;
- $\lambda \in (\lambda_6, \lambda_1)$, then (1-7) has exactly 15 positive solutions.

Figure 24 illustrates Case 13.

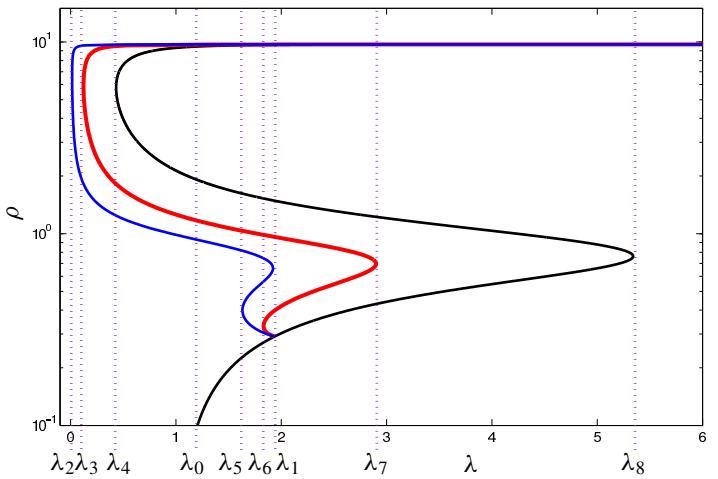


Figure 24. ρ versus λ when $b = 10$ and $c = 30.3$ (Case 13).

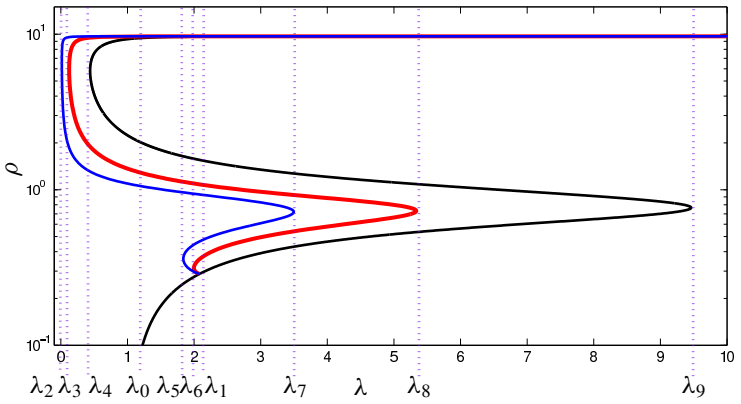


Figure 25. ρ versus λ when $b = 10$ and $c = 32$ (Case 14).

Case 14. If $c \in (c_7, c_0(b))$ then there exist $\lambda_i > 0$ for $i = 1, 2, \ldots, 9$ such that if

- $\lambda \in [0, \lambda_2)$, then (1-7) has no positive solution;
- $\lambda = \lambda_2$, then (1-7) has a unique positive solution;
- $\lambda \in (\lambda_2, \lambda_3)$, then (1-7) has exactly 2 positive solutions;
- $\lambda = \lambda_3$ and $\lambda \in (\lambda_9, \infty)$, then (1-7) has exactly 4 positive solutions;
- $\lambda = \lambda_9$, then (1-7) has exactly 5 positive solutions;
- $\lambda \in (\lambda_3, \lambda_4), (\lambda_8, \lambda_9)$, then (1-7) has exactly 6 positive solutions;
- $\lambda = \lambda_4$, then (1-7) has exactly 7 positive solutions;
- $\lambda \in (\lambda_4, \lambda_0], \lambda = \lambda_8$, then (1-7) has exactly 8 positive solutions;
- $\lambda \in (\lambda_0, \lambda_5)$, then (1-7) has exactly 9 positive solutions;
- $\lambda = \lambda_5, \lambda \in (\lambda_7, \lambda_8)$, then (1-7) has exactly 10 positive solutions;
- $\lambda \in (\lambda_5, \lambda_6), \lambda = \lambda_7$, then (1-7) has exactly 11 positive solutions;
- $\lambda \in (\lambda_1, \lambda_7)$, then (1-7) has exactly 12 positive solutions;
- $\lambda = \lambda_6$, then (1-7) has exactly 13 positive solutions;
- $\lambda \in (\lambda_6, \lambda_1]$, then (1-7) has exactly 15 positive solutions;

Figure 25 illustrates Case 14.

7. Analytical results

In order to bolster our computational results as well as elaborate on the behavior of the bifurcation curves, we procure some analytical results. First, we recall some results from [Laetsch 1970] detailing the behavior of $G_1(\rho)$ when $\rho \to \sigma(b, c)^-$

and when $\rho \to 0^+$ in the following lemmas, where $\sigma(b, c)$ represents the smallest positive root of $f(u)$.

Lemma 2 [Laetsch 1970]. $\lim_{\rho \to \sigma(b,c)^-} G_1(\rho) = \infty$.

Lemma 3 [Laetsch 1970]. $\lim_{\rho \to 0^+} G_1(\rho) = \pi/(2\sqrt{b})$.

Our main goal for this section is to establish the following analytical results for (1-9) and (1-11). Recall that $\lambda = [G_2(\rho, q)]^2$ and $\lambda = [G_3(\rho, q)]^2$ from Theorems 4.1 and 5.1, respectively. Thus, we can obtain some global behavior of the ρ versus λ bifurcation curve via study of $G_2(\rho, q)$ and $G_3(\rho, q)$.

Theorem 7.1.

(1) $$\left(\sqrt{2} - \frac{1}{\sqrt{2}}\right) \int_0^\rho \frac{dw}{\sqrt{F(\rho) - F(w)}} \le G_2(\rho, q) \le \sqrt{2} \int_0^\rho \frac{dw}{\sqrt{F(\rho) - F(w)}};$$

(2) $$G_3(\rho, q) \le \sqrt{2} \int_0^\rho \frac{dw}{\sqrt{F(\rho) - F(w)}}.$$

Proof. To prove (1), recall

$$\begin{aligned} G_2(\rho, q) &= \frac{1}{\sqrt{2}\sqrt{F(\rho) - F(q)}} \\ &= \sqrt{2} \int_0^\rho \frac{dw}{\sqrt{F(\rho) - F(w)}} - \frac{1}{\sqrt{2}} \int_0^q \frac{dw}{\sqrt{F(\rho) - F(w)}}. \end{aligned} \tag{7-1}$$

We ascertain an upper bound by substituting $q = 0$ into (7-1) yielding

$$G_2(\rho, q) \le \sqrt{2} \int_0^\rho \frac{dw}{\sqrt{F(\rho) - F(w)}}.$$

Also, recalling $q \in [0, \rho)$ and allowing $q \to \rho^-$ in (7-1) we obtain

$$G_2(\rho, q) \ge \left(\sqrt{2} - \frac{1}{\sqrt{2}}\right) \int_0^\rho \frac{dw}{\sqrt{F(\rho) - F(w)}}$$

as the lower bound. Hence,

$$\left(\sqrt{2} - \frac{1}{\sqrt{2}}\right) \int_0^\rho \frac{dw}{\sqrt{F(\rho) - F(w)}} \le G_2(\rho, q) \le \sqrt{2} \int_0^\rho \frac{dw}{\sqrt{F(\rho) - F(w)}}.$$

Now to prove (2). Recall

$$G_3(\rho, q) = \sqrt{2} \int_0^\rho \frac{dw}{\sqrt{F(\rho) - F(w)}} - \sqrt{2} \int_0^q \frac{dw}{\sqrt{F(\rho) - F(w)}}. \tag{7-2}$$

Similarly, by substituting $q = 0$ into (7-2) we obtain

$$G_3(\rho, q) \le \sqrt{2} \int_0^\rho \frac{dw}{\sqrt{F(\rho) - F(w)}}$$

as the upper bound. □

Theorem 7.2. $\lim_{\rho \to \sigma(b,c)^-} G_2(\rho, q) = \infty.$

Proof. By Theorem 7.1, we have

$$G_2(\rho, q) \geq \left(\sqrt{2} - \frac{1}{\sqrt{2}}\right) \int_0^{\rho} \frac{dw}{\sqrt{F(\rho) - F(w)}}. \tag{7-3}$$

From Lemma 2, it is clear that the right side of (7-3) approaches infinity as

$$\rho \to \sigma(b, c)^-.$$

Therefore, it is apparent that

$$\lim_{\rho \to \sigma(b,c)^-} G_2(\rho, q) = \infty. \qquad \square$$

Acknowledgement

A significant portion of this research work was completed by Kathryn Ashley and Victoria Sincavage while attending the 2011 NSF REU Program at the Center for Computational Sciences at Mississippi State University. We would like to extend a special thanks to our advisor and mentor Professor Ratnasingham Shivaji for his guidance and support throughout this research project.

References

[Allee 1938] W. C. Allee, *The social life of animals*, Norton, New York, 1938.

[Brown et al. 1981] K. J. Brown, M. M. A. Ibrahim, and R. Shivaji, "S-shaped bifurcation curves", *Nonlinear Anal.* **5**:5 (1981), 475–486. MR 82h:35007 Zbl 0458.35036

[Cantrell and Cosner 2003] R. S. Cantrell and C. Cosner, *Spatial ecology via reaction-diffusion equations*, Wiley, Chichester, 2003. MR 2007a:92069 Zbl 1059.92051

[Cantrell and Cosner 2007] R. S. Cantrell and C. Cosner, "Density dependent behavior at habitat boundaries and the Allee effect", *Bull. Math. Biol.* **69**:7 (2007), 2339–2360. MR 2341874 Zbl 05265737

[Causey et al. 2010] R. Causey, S. Sasi, and R. Shivaji, "An ecological model with grazing and constant yield harvesting", *Bull. Belg. Math. Soc. Simon Stevin* **17**:5 (2010), 833–839. MR 2012a:35103 Zbl 1208.35153

[Dennis 1989] B. Dennis, "Allee effects: population growth, critical density, and the chance of extinction", *Natur. Resource Modeling* **3**:4 (1989), 481–538. MR 91h:92032 Zbl 0850.92062

[Goddard and Shivaji 2012] J. Goddard, II and R. Shivaji, "A population model with nonlinear boundary conditions and constant yield harvesting", *Proceedings of Dynamic Systems and Applications* **6** (2012), 150–157.

[Goddard et al. 2010a] J. Goddard, II, E. K. Lee, and R. Shivaji, "A double S-shaped bifurcation curve for a reaction-diffusion model with nonlinear boundary conditions", *Bound. Value Probl.* (2010), Art. ID 357542. MR 2665818 Zbl 1211.35037

[Goddard et al. 2010b] J. Goddard, II, E. K. Lee, and R. Shivaji, "Population models with nonlinear boundary conditions", *Electron. J. Differ. Equ. Conf.* **19** (2010), 135–149. MR 2012f:35306 Zbl 1204.34030

[Goddard et al. 2011a] J. Goddard, II, E. K. Lee, and R. Shivaji, "Population models with diffusion, strong Allee effect, and nonlinear boundary conditions", *Nonlinear Anal.* **74**:17 (2011), 6202–6208. MR 2012i:35192 Zbl 1227.35172

[Goddard et al. 2011b] J. Goddard, II, R. Shivaji, and E. K. Lee, "Diffusive logistic equation with non-linear boundary conditions", *J. Math. Anal. Appl.* **375**:1 (2011), 365–370. MR 2011i:35123 Zbl 1208.35082

[Kuussaari et al. 1996] M. Kuussaari, M. Nieminen, and I. Hanski, "An experimental study of migration in the Glanville fritillary butterfly *Melitaea cinxia*", *Journal of Animal Ecology* **65**:6 (1996), 791–801.

[Laetsch 1970] T. Laetsch, "The number of solutions of a nonlinear two point boundary value problem", *Indiana Univ. Math. J.* **20** (1970), 1–13. MR 42 #4815 Zbl 0215.14602

[Lee et al. 2011] E. K. Lee, S. Sasi, and R. Shivaji, "S-shaped bifurcation curves in ecosystems", *J. Math. Anal. Appl.* **381**:2 (2011), 732–741. MR 2012e:92080 Zbl 1221.35421

[Lewis and Kareiva 1993] M. A. Lewis and P. Kareiva, "Allee dynamics and the spread of invading organisms", *Theoretical Population Biology* **43**:2 (1993), 141–158. Zbl 0769.92025

[Matthysen 2005] E. Matthysen, "Density-dependent dispersal in birds and mammals", *Ecography* **28** (2005), 403–416.

[van Nes and Scheffer 2005] E. H. van Nes and M. Scheffer, "Implications of spatial heterogeneity for catastrophic regime shifts in ecosystems", *Ecology* **86**:7 (2005), 1797–1807.

[Poole et al. 2012] E. Poole, B. Roberson, and B. Stephenson, "Weak Allee effect, grazing, and S-shaped bifurcation curves", *Involve* **5**:2 (2012), 133–158.

[Shi and Shivaji 2006] J. Shi and R. Shivaji, "Persistence in reaction diffusion models with weak Allee effect", *J. Math. Biol.* **52**:6 (2006), 807–829. MR 2007g:92070 Zbl 1110.92055

Received: 2012-07-23 Revised: 2013-04-04 Accepted: 2013-04-10

klashley01@email.wm.edu *Department of Mathematical Sciences, Clemson University, Clemson, SC 29634, United States*

vsincav@clemson.edu *Department of Mathematical Sciences, Clemson University, Clemson, SC 29634, United States*

jgoddard@aum.edu *Department of Mathematics, Auburn University Montgomery, Montgomery, AL 36124-4023, United States*

INVOLVE 6:4 (2013)
dx.doi.org/10.2140/involve.2013.6.431

The probability of randomly generating finite abelian groups

Tyler Carrico

(Communicated by Joseph Gallian)

Extending the work of Deborah L. Massari and Kimberly L. Patti, this paper makes progress toward finding the probability of k elements randomly chosen without repetition generating a finite abelian group, where k is the minimum number of elements required to generate the group. A proof of the formula for finding such probabilities of groups of the form $\mathbb{Z}_{p^m} \oplus \mathbb{Z}_{p^n}$, where $m, n \in \mathbb{N}$ and p is prime, is given, and the result is extended to groups of the form $\mathbb{Z}_{p^{n_1}} \oplus \cdots \oplus \mathbb{Z}_{p^{n_k}}$, where $n_i, k \in \mathbb{N}$ and p is prime. Examples demonstrating applications of these formulas are given, and aspects of further generalization to finding the probabilities of randomly generating any finite abelian group are investigated.

Introduction

Throughout this paper, let k be the minimum number of elements required to generate a group G, A_G be the event where k elements randomly chosen without repetition generate G, and $P(A_G)$ be the probability of A_G occurring. Massari [1979] showed that, for a finite cyclic group G of order a, $P(A_G) = \phi(a)/a$, where ϕ is the Euler phi function. Patti [2002] showed, among other things, that, for $G = \mathbb{Z}_p \oplus \cdots \oplus \mathbb{Z}_p$ (the external direct product of $\mathbb{Z}_p$ taken k times, where p is prime),

$$P(A_G) = \frac{\prod_{i=0}^{k-1}(p^k - p^i)}{\prod_{j=0}^{k-1}(p^n - j)}.$$

It is natural to ask what the probability of generating groups like G is when powers are added to the p subscripts. We now turn to this problem.

Theorem 1. *Let $G = \mathbb{Z}_{p^m} \oplus \mathbb{Z}_{p^n}$ where $m, n \in \mathbb{N}$ and p is prime. Then*

$$P(A_G) = \frac{p^{2(m+n-2)}(p^2-1)(p^2-p)}{p^{m+n}(p^{m+n}-1)}.$$

MSC2010: 20P05.
Keywords: abelian, group, generate, probability.

Proof. Partition the elements of G into p^2 subsets (these particular types of subsets will be referred to as A-subsets from this point forward):

$$A_{ij} = \{(px+i \mod p^m,\ py+j \mod p^n) : x \in \mathbb{Z}_{p^m},\ y \in \mathbb{Z}_{p^n},\ i, j \in \{0, 1, \ldots, p-1\}\}.$$

Note that $\bigcup A_{ij} = G$ and $|A_{ij}| = |G|/p^2 = p^{m+n-2}$ for $i = 0, 1, \ldots, p-1$ and $j = 0, 1, \ldots, p-1$.

From this point forward we assume without being explicit that, in any tuple $(px+i,\ py+j)$, i and j are reduced modulo p, $px+i$ is reduced modulo p^m, and $py+j$ is reduced modulo p^n.

Note that, for any $g \in G$ and any $t_1, t_2 \in \mathbb{Z}$ such that $t_1 \equiv t_2 \mod p$, $t_1 g$ and $t_2 g$ belong to the same A-subset. Therefore, any element $g \in G$ can at most generate p A-subsets since there are p possible choices for an integer t that have the potential to place tg in different A-subsets. For an element g, let F_g denote the family of A-subsets to which tg belongs for all possible values of t. Note that for each $g \notin A_{00}$ exactly p A-subsets belong to F_g (g does not necessarily generate all A-subsets belonging to F_g, but it generates at least one element belonging to each A-subset).

Let $g = (a, b) \in G$, and let (c, d) and (e, f) be two elements, each from any A-subset belonging to F_g. Then $(c, d) \equiv k_1(a, b) \mod p$ and $(e, f) \equiv k_2(a, b) \mod p$ for some $k_1, k_2 \in \mathbb{Z}$. Thus, for any $k_3, k_4 \in \mathbb{Z}$, $k_3(c, d) + k_4(e, f) \equiv k_3k_1(a, b) + k_4k_2(a, b) = [k_3k_1 + k_4k_2](a, b) \mod p$, which belongs to an A-subset in F_g. Thus, these two elements generate at most p A-subsets and thus cannot generate G. Note that, since $A_{00} \in F_g$ for all $g \in G$, it is impossible for two elements to generate G when one of them belongs to A_{00}.

Now suppose we choose elements $a, b \notin A_{00}$, say $a = (px_1 + i_1,\ py_1 + j_1)$ and $b = (px_2 + i_2,\ py_2 + j_2)$, such that a does not belong to any A-subset in F_b and b does not belong to any A-subset in F_a (thus it is not the case that $i_1 = i_2 = 0$, $j_1 = j_2 = 0$, $i_1 = j_1 = 0$, or $i_2 = j_2 = 0$). We will show that a and b together generate G.

Case 1: At least one of i_1, i_2, j_1, j_2 is zero. Without loss of generality let $i_1 = 0$. Then $i_2 \neq 0$, $j_1 \neq 0$, and because $\gcd(py_1 + j_1, p^n) = 1$ there exists $q \in \mathbb{Z}$ such that $qa = (px, 1)$ for some $x \in \mathbb{Z}_{p^m}$.

Subcase 1: $j_2 = 0$. Then by similar reasoning there exists $r \in \mathbb{Z}$ such that $rb = (1,\ py)$ for some $y \in \mathbb{Z}_{p^n}$. Now $qa - pxrb = (0, -p^2xy + 1)$, and $\gcd(-p^2xy + 1, p^n) = 1$ so there exists $s \in \mathbb{Z}$ such that $s[qa - pxrb] = (0, 1)$. Finally, $rb - pys[qa - pxrb] = (1, 0)$.

Subcase 2: $j_2 \neq 0$. Then $b - j_2qa = (p[x_2 - j_2x] + i_2,\ py_2)$, and we arrive at the same situation as Subcase 1.

Case 2: None of i_1, i_2, j_1, j_2 are zero. Let

$$e = i_2a - i_1b = (p[i_2x_1 - i_1x_2],\ p[i_2y_1 - i_1y_2] + c),$$

where $c = i_2 j_1 - i_1 j_2$. We will show that $c \neq 0$. Assume to the contrary that $c = 0$. Since $j_2 \neq 0$, $j_2 \in \{1, 2, \ldots, p-1\} \subset \mathbb{Z}_p$, and, because $\mathbb{Z}_p$ is a field, there exists $k \in \mathbb{Z}_p$ such that $kj_2 \equiv 1 \mod p$. Let $d = j_1 k$. Because $i_1 j_2 = i_2 j_1$, we now have $i_1 \equiv i_1 j_2 k = i_2 j_1 k = d i_2 \mod p$ and $j_1 \equiv j_1 j_2 k = d j_2 \mod p$ so that a and db are in the same A-subset, a contradiction. Thus, $c \neq 0$.

Now, because $\gcd(p[i_2 y_1 - i_1 y_2] + c, p^n) = 1$, there exists $q \in \mathbb{Z}$ such that $qe = (px, 1)$ for some $x \in \mathbb{Z}_{p^m}$. Further, $f = b - qe[py_2 + j_2] = (px_3 + i_2, 0)$ for some $x_3 \in \mathbb{Z}_{p^m}$ and $\gcd(px_3 + i_2, p^m) = 1$, so there exists $t \in \mathbb{Z}$ such that $tf = (1, 0)$. Finally, $qe - pxtf = (0, 1)$.

In any case, we have shown that a and b generate $(1, 0)$ and $(0, 1)$, and thus a and b together generate G.

It is left to show the value of $P(A_G)$. For the first element a, any element other than an element from A_{00} can be chosen. Thus, there are p^{m+n-2} elements from each of the $p^2 - 1$ possible A-subsets from which to choose, a total of $p^{m+n-2}(p^2 - 1)$ elements out of the possible p^{m+n}. For the second element, an element must be chosen from an A-subset not belonging to F_a. Since p A-subsets belong to F_a, there are $p^2 - p$ such A-subsets, each containing p^{m+n-2} elements. Thus, there are $p^{m+n-2}(p^2 - p)$ elements out of the remaining $p^{m+n} - 1$ possible elements from which to choose. Therefore,

$$\begin{aligned} P(A_G) &= \frac{p^{m+n-2}(p^2-1)}{p^{m+n}} \cdot \frac{p^{m+n-2}(p^2-p)}{p^{m+n}-1} \\ &= \frac{p^{2(m+n-2)}(p^2-1)(p^2-p)}{p^{m+n}(p^{m+n}-1)}. \end{aligned}$$

□

Example. Consider the group $H = \mathbb{Z}_{7^5} \oplus \mathbb{Z}_{7^{12}}$. Then

$$\begin{aligned} P(A_H) &= \frac{7^{2(5+12-2)}(7^2-1)(7^2-7)}{7^{5+12}(7^{5+12}-1)} \\ &= \frac{7^{30}(48)(42)}{7^{17}(7^{17}-1)} \\ &= 0.83965. \end{aligned}$$

This result can be extended to the external direct product of any finite number of $\mathbb{Z}_{p^{n_i}}$.

Theorem 2. *Let $G = \mathbb{Z}_{p^{n_1}} \oplus \cdots \oplus \mathbb{Z}_{p^{n_k}}$, where $n_i \in \mathbb{N}$ and p is prime. Define $n = \sum_{i=1}^{k} n_i$. Then*

$$P(A_G) = \frac{p^{k(n-k)} \prod_{i=0}^{k-1} (p^k - p^i)}{\prod_{j=0}^{k-1} (p^n - j)}.$$

Proof. Partition the elements of G into p^k A-subsets:

$$A_{i_1\cdots i_k}=\{(px_1+i_1 \bmod p^{n_1},\ldots,px_k+i_k \bmod p^{n_k}):x_j\in\mathbb{Z}_{p^{n_j}},i_j\in\{0,1,\ldots,p-1\}\}.$$

Note that $\bigcup A_{i_1\cdots i_k}=G$ and $|A_{i_1\cdots i_k}|=|G|/p^k=p^{n-k}$ for $i_j=0,1,\ldots,p-1$.

Similar to the case where $k=2$, any element $g\in G$ can at most generate p A-subsets, and for each $g\notin A_{0\cdots0}$ exactly p A-subsets belong to F_g. When k elements are chosen, if any two elements belong to A-subsets within the same family, at most p^{k-1} A-subsets can be generated. Therefore, it is impossible to generate G with such a choice of elements.

Now choose an element not in the null family, then choose another not in the family of the first element, then choose another such that it is not in any family generated by any linear combination of the first two elements, and so forth until we have chosen k elements $a_1,\ldots,a_k$. Then none of the elements can be written as a linear combination of the other $k-1$ elements. Define $A=\{a_1,\ldots,a_k\}$. Assume that none of the elements of A are part of an A-subset with zero in its subscript. Then for any a_{m_1} and a_{m_2} there exist integers c_1 and c_2 so that $c_1a_{m_1}+c_2a_{m_2}\in A_{0i_2\cdots i_k}$, where not all of the i_j are zero. Therefore, we can generate $k-1$ elements $a'_1,\ldots,a'_{k-1}$ where $a'_m=c_1a_m+c_2a_{m+1}$ and c_1 and c_2 are such that $a'_m\in A_{0i_2\cdots i_k}$, where not all of the i_j are zero. Assume that $i_j\neq0$ for $j=2,\ldots,k$. Define $A'=\{a'_1,\ldots,a'_{k-1}\}$. Note that none of $a'_1,\ldots,a'_{k-1}$ can be written as linear combinations of the other $k-2$ elements, for, if this were possible, some a_j could be written as a linear combination of the elements in A other than a_j, which contradicts our choice of the elements of A. We can now generate $k-2$ elements $a''_1,\ldots,a''_{k-2}$ in a similar manner so that $a''_m\in A_{00i_3\cdots i_k}$, and similar conditions and assumptions hold. Continuing in this manner, we generate an element $a^{(k)}\in A_{0\cdots0i_k}$, where $i_k\neq0$. Now, because $\gcd(px_k+i_k,p^{n_k})=1$, there exists c such that $ca^{(k)}=(py_1 \mod p^{n_1},\ldots,py_{k-1} \mod p^{n_{k-1}},1)$ for some $y_i\in\mathbb{Z}_{p^{n_i}}$.

Following a procedure similar to the one previously described, only changing the order by which the linear combinations of the elements are taken, we can generate $k-1$ other elements so that we have a total of k elements $b_1,\ldots,b_k$ such that the j-th coordinate of b_j is 1 and the remaining coordinates are multiples of p. Now linear combinations of these elements can be taken so that k elements $c_1,\ldots,c_k$ are generated, where the j-th coordinate of c_j is not a multiple of p and the remaining coordinates are 0. Thus, the greatest common divisor of the j-th coordinate of each c_j and p^{n_j} is 1, and thus there exists t_j for each c_j such that t_jc_j has 1 for the j-th coordinate and zero for the remaining coordinates.

If, unlike our earlier assumptions, it happens at any point that some i_j is zero, notice that this is a subcase of our original case, where we already possess elements

which otherwise we would have had to generate as we did in our original case. Thus, our assumption that each i_j be nonzero at each step is unnecessary and, in any case, $a_1, \ldots, a_k$ together generate G.

It is left to show the value of $P(A_G)$. For the first element a_1, any element other than an element from $A_{0\ldots0}$ can be chosen. Thus, there are p^{n-k} elements from each of the $p^k - 1$ possible A-subsets from which to choose, a total of $p^{n-k}(p^k - 1)$ elements out of the possible p^n. For the second element a_2, an element must be chosen from an A-subset not belonging to F_{a_1}. Since p A-subsets belong to F_{a_1}, there are $p^k - p$ such A-subsets, each containing p^{n-k} elements. Thus, there are $p^{n-k}(p^k - p)$ elements out of the remaining $p^n - 1$ possible elements from which to choose. Continuing in this manner and multiplying the resulting probabilities, we have

$$\begin{aligned} P(A_G) &= \frac{p^{n-k}(p^k-1)}{p^n} \cdot \frac{p^{n-k}(p^k-p)}{p^n-1} \cdots \frac{p^{n-k}(p^k-p^{k-1})}{p^n-k} \\ &= \frac{p^{k(n-k)} \prod_{i=0}^{k-1}(p^k-p^i)}{\prod_{j=0}^{k-1}(p^n-j)}. \end{aligned}$$

□

Example. Consider the group $I = \mathbb{Z}_{29} \oplus \mathbb{Z}_{29^2} \oplus \mathbb{Z}_{29^3} \oplus \mathbb{Z}_{29^4}$. Then $1+2+3+4=10$, so

$$\begin{aligned} P(A_I) &= \frac{29^{4(10-4)}(29^4-1)(29^4-29)(29^4-29^2)(29^4-29^3)}{29^{10}(29^{10}-1)(29^{10}-2)(29^{10}-3)} \\ &= 0.964. \end{aligned}$$

Extension. The fundamental theorem of finite abelian groups states that every finite abelian group is isomorphic to a direct product of cyclic groups of prime-power order, that is, groups of the form $\mathbb{Z}_{p_1{}^{n_1}} \oplus \cdots \oplus \mathbb{Z}_{p_k{}^{n_k}}$, where $n_i \in \mathbb{N}$ and p_i are prime [Gallian 2006]. We would thus hope that extending the previous theorem by varying the primes would be simple. This is not the case, however. Consider the following three groups and the probabilities of generating them:

(1) Let $G_1 = \mathbb{Z}_2 \oplus \mathbb{Z}_2$. Then $P(A_{G_1}) = 1/2$.

(2) Let $G_2 = \mathbb{Z}_3 \oplus \mathbb{Z}_3$. Then $P(A_{G_2}) = 2/3$.

(3) Let $G_3 = \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_3$. Then $P(A_{G_3}) = 8/35$.

Notice that, although we have a formula for finding the probabilities of generating (1) and (2) and G_3 is isomorphic to $G_1 \oplus G_2$, the relationship between the probabilities of generating each of the three groups (1/2, 2/3, and 8/35) is not obvious. The following is a conjecture for the probability of generating groups of form similar to G_3.

Conjecture. Let $G = \mathbb{Z}_{p^{n_1}} \oplus \mathbb{Z}_{p^{n_2}} \oplus \mathbb{Z}_{q^{n_3}} \oplus \mathbb{Z}_{q^{n_4}}$, where $n_i \in \mathbb{N}$ and p and q are prime. Then

$$P(A_G) = \frac{p^{2(n_1+n_2-2)}q^{2(n_3+n_4-2)}(p^2q^2-p^2-q^2+1)(p^2q^2-p^2q-q^2p+pq)}{p^{n_1+n_2}q^{n_3+n_4}(p^{n_1+n_2}q^{n_3+n_4}-1)}.$$

This equation is similar in form to our previous theorem, yet it differs significantly in the number of A-subsets from which elements can be chosen that successfully generate G; the first element can be chosen from $p^2q^2-p^2-q^2+1$ A-subsets and the second element can be chosen from $p^2q^2-p^2q-q^2p+pq$ A-subsets. The following conjecture shows how similar complexities arise in a group form similar to the previous case:

Conjecture. Let $G = \mathbb{Z}_{p^{n_1}} \oplus \mathbb{Z}_{p^{n_2}} \oplus \mathbb{Z}_{p^{n_3}} \oplus \mathbb{Z}_{q^{n_4}}$ where $n_i \in \mathbb{N}$ and p and q are prime. Then

$$P(A_G) = \frac{p^{3(n_1+n_2+n_3-3)}q^{n_4-1}(p^3-1)(p^3-p)(p^3-p^2)(q^3-1)}{p^{n_1+n_2+n_3}q^{n_4}(p^{n_1+n_2+n_3}q^{n_4}-1)(p^{n_1+n_2+n_3}q^{n_4}-2)}.$$

Finally, since a set of elements from a group will either generate the whole group or a proper subgroup, if we let B_G be the event where k elements randomly chosen without repetition generate a proper subgroup of G, then $P(B_G)$, the probability of B_G occurring, is

$$P(B_G) = 1 - P(A_G).$$

Acknowledgements

I would like to express my gratitude to Dr. Daniel Kiteck for dedicating time to oversee the research, continually checking my proofs, and giving guidance, advice, and encouragement. I would also like to thank Dr. Bob Mallison for invaluable hints which led to the completion of the proofs of the theorems.

References

[Gallian 2006] J. A. Gallian, *Contemporary abstract algebra*, 6th ed., Houghton Mifflin, Boston and New York, 2006.

[Massari 1979] D. L. Massari, "The probability of generating a cyclic group", *Pi Mu Epsilon Journal* **7**:1 (1979), 3–6. Zbl 0435.20055

[Patti 2002] K. L. Patti, "The probability of randomly generating a finite group", *Pi Mu Epsilon Journal* **11**:6 (2002), 313–316.

Received: 2012-07-26 Revised: 2012-10-26 Accepted: 2012-11-13

supernaturalgospel@gmail.com *202-0004 Tokyo, Nishitokyo-shi, Shimohoya 3-11-23, Japan*

INVOLVE 6:4 (2013)
dx.doi.org/10.2140/involve.2013.6.437

Free and very free morphisms into a Fermat hypersurface

Tabes Bridges, Rankeya Datta, Joseph Eddy,
Michael Newman and John Yu

(Communicated by Ravi Vakil)

This paper studies the existence of free and very free curves on the degree 5 Fermat hypersurface in $\mathbb{P}^5$ over an algebraically closed field of characteristic 2. We explicitly compute a free curve in degree 8, and a very free curve in degree 9. We also prove that free and very free curves cannot exist in lower degrees.

1. Introduction

Any smooth projective Fano variety in characteristic zero is rationally connected and hence contains a very free rational curve. In positive characteristic a smooth projective Fano variety is rationally chain-connected. However, it is not known whether such varieties are separably rationally connected, or equivalently, whether they have a very free rational curve. This is an open question even for nonsingular Fano hypersurfaces. See [Kollár 1996], as well as [Debarre 2001].

Following [Shen 2012], we consider the degree 5 Fermat hypersurface

$$X: \quad X_0^5 + X_1^5 + X_2^5 + X_3^5 + X_4^5 + X_5^5 = 0$$

in $\mathbb{P}^5$ over an algebraically closed field k of characteristic 2. This is a nonsingular projective Fano variety.

Theorem 1.1. *Any free rational curve $\varphi : \mathbb{P}^1 \to X$ has degree ≥ 8, and there exists a free rational curve of degree 8. Any very free rational curve $\varphi : \mathbb{P}^1 \to X$ has degree ≥ 9, and there exists a very free rational curve of degree 9.*

This result, although perhaps expected, is interesting for several reasons. First, it is known that X is unirational; see [Debarre 2001, p. 52] (the corresponding rational map $\mathbb{P}^4 \dashrightarrow X$ is inseparable). Second, in [Beauville 1990], it is shown that

MSC2010: primary 14-02; secondary 14M22.

Keywords: free morphisms, very free morphisms, Fermat hypersurface, Fermat hypersurface over a field of characteristic 2.

Results in this paper were obtained during an REU at Columbia University led by A. J. de Jong in the summer of 2012.

every nonsingular hyperplane section of X is isomorphic to a Fermat hypersurface of dimension 3, and this property characterizes Fermat hypersurfaces among all hypersurfaces of degree 5 in characteristic 2. We believe that these facts single out the Fermat as a likely candidate for a counterexample to the conjecture below; instead, our theorem shows that they are evidence for it.

Conjecture 1.2. Nonsingular Fano hypersurfaces have very free rational curves.

Zhu [2011] discusses this question more broadly. Let us discuss a little bit about the method of proof. In Section 2, we translate the geometric question into an algebraic question which is computationally more accessible. In Sections 3, 4, and 5, we exclude low-degree solutions by theoretical methods. Finally, in Sections 6 and 7, we explicitly describe some curves which are free and very free in degrees 8 and 9, respectively.

2. The overall setup

In the rest of this paper, k will be an algebraically closed field of characteristic 2 and X will be the Fermat hypersurface of degree 5 over k. Let $\varphi : \mathbb{P}^1 \to X$ be a nonconstant morphism. We will repeatedly use that every vector bundle on $\mathbb{P}^1$ is a direct sum of line bundles; see [Grothendieck 1957]. Thus we can choose a splitting

$$\varphi^* T_X = \mathcal{O}_{\mathbb{P}^1}(a_1) \oplus \mathcal{O}_{\mathbb{P}^1}(a_2) \oplus \mathcal{O}_{\mathbb{P}^1}(a_3) \oplus \mathcal{O}_{\mathbb{P}^1}(a_4).$$

Recall that φ is said to be a *free curve* on X if $a_i \geq 0$, and φ is said to be *very free* if $a_i > 0$. Consider the commutative diagram

$$\begin{array}{ccccccccc}
 & & 0 & & 0 & & & & \\
 & & \downarrow & & \downarrow & & & & \\
 & & \mathcal{O}_X & = & \mathcal{O}_X & & & & \\
 & & \downarrow & & \downarrow & & & & \\
0 & \longrightarrow & E_X & \longrightarrow & \mathcal{O}_X(1)^{\oplus 6} & \longrightarrow & \mathcal{O}_X(5) & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & \| & & \\
0 & \longrightarrow & T_X & \longrightarrow & T_{\mathbb{P}^5}|_X & \longrightarrow & N_{X/\mathbb{P}^5} & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & & & \\
 & & 0 & & 0 & & & &
\end{array} \tag{2-1}$$

with exact rows and columns as indicated. We will call E_X the *extended tangent bundle of* X. The left vertical exact sequence determines a short exact sequence

$$0 \to \mathcal{O}_{\mathbb{P}^1} \to \varphi^* E_X \to \varphi^* T_X \to 0.$$

The splitting type of $\varphi^* E_X$ will consistently be denoted $(f_1, f_2, f_3, f_4, f_5)$ in this paper. Since $\mathrm{Hom}_{\mathbb{P}^1}(\mathcal{O}_{\mathbb{P}^1}(f), \mathcal{O}_{\mathbb{P}^1}(a)) = 0$ if $f > a$, we conclude:

(1) If $f_i \geq 0$ for all i, then φ is free.

(2) If $f_i > 0$ for all i, then φ is very free.

For the converse, note that the map $\mathcal{O}_{\mathbb{P}^1} \to \varphi^* E_X$ has image contained in the direct sum of the summands with $f_i \geq 0$. Hence, if $f_i < 0$ for some i, then φ is not free. Finally, suppose that $f_i \geq 0$ for all i. If there are at least two f_i equal to 0, then we see that φ is free but not very free. We conclude:

(3) If φ is free, then $f_i \geq 0$ for all i.

(4) If φ is very free, then either

  (a) $f_i > 0$ for all i, or

  (b) exactly one f_i vanishes and all others are positive.

We do not know if (4b) occurs.

Translation into algebra. Here we work over the graded k-algebra $R = k[S, T]$. As usual, we let $R(e)$ be the graded free R-module whose underlying module is R with grading given by $R(e)_n = R_{e+n}$. A *graded free R-module* will be any graded R-module isomorphic to a finite direct sum of $R(e)$'s. Such a module M has a *splitting type* which is uniquely defined up to reordering, namely, the sequence of integers $u_1, \ldots, u_r$ such that $M \cong R(u_1) \oplus \cdots \oplus R(u_r)$.

We will think of a degree d morphism $\varphi : \mathbb{P}^1 \to \mathbb{P}^5$ as a 6-tuple $(G_0, \ldots, G_5)$ of homogeneous elements in R of degree d with no common factors. Then φ is a morphism into X if and only if $G_0^5 + \cdots + G_5^5 = 0$. In this situation we define two graded R-modules. The first is called the *pullback of the cotangent bundle*

$$\Omega_X(\varphi) = \mathrm{Ker}(\tilde{\varphi} : R^{\oplus 6}(-d) \to R),$$

where the map $\tilde{\varphi}$ is given by $(A_0, \ldots, A_5) \mapsto \sum A_i G_i$. The second is called the *the pullback of the extended tangent bundle*

$$E_X(\varphi) = \mathrm{Ker}(R^{\oplus 6}(d) \to R(5d)),$$

where the map is given by $(A_0, \ldots, A_5) \mapsto \sum A_i G_i^4$. Since the kernel of a map of graded free R-modules is a graded free R-module, both $\Omega_X(\varphi)$ and $E_X(\varphi)$ are themselves graded free R-modules of rank 5.

Lemma 2.1. *The splitting type of $\varphi^* E_X$ is equal to the splitting type of the R-module $E_X(\varphi)$.*

Proof. Recall that $\mathbb{P}^1 = \mathrm{Proj}(R)$. Thus, a finitely generated graded R-module corresponds to a coherent sheaf on $\mathbb{P}^1$; see [Hartshorne 1977, Proposition 5.11]. Under this correspondence, the module $R(e)$ corresponds to $\mathcal{O}_{\mathbb{P}^1}(e)$. The lemma follows

if we show that $\varphi^* E_X$ is the coherent sheaf associated to $E_X(\varphi)$. Diagram (2-1) shows that $\varphi^* E_X$ is the kernel of a map $\mathbb{O}_{\mathbb{P}^1}(d)^{\oplus 6} \to \mathbb{O}_{\mathbb{P}^1}$ given by substituting $(G_0, \ldots, G_5)$ into the partial derivatives of the polynomial defining X. Since the equation is $X_0^5 + \cdots + X_5^5$, the derivatives are X_i^4, and substituting we obtain G_i^4 as desired. □

3. Relating the splitting types

Observe that $\Omega_X(\varphi)$ is also a graded free module of rank 5 and so has a splitting type, which we denote using $e_1, \ldots, e_5$. In this section, we relate the splitting type of $\Omega_X(\varphi)$ to the splitting type of $E_X(\varphi)$.

If $(A_0, \ldots, A_5) \in \Omega_X(\varphi)$, then $A_0G_0 + \cdots + A_5G_5 = 0$ so that

$$A_0^4 G_0^4 + \cdots + A_5^4 G_5^4 = 0$$

by the Frobenius endomorphism in characteristic 2. Let

$$\mathcal{T} = \{(A_0^4, \ldots, A_5^4) \mid (A_0, \ldots, A_5) \in \Omega_X(\varphi)\}$$

in $E_X(\varphi)$. We denote the R-module generated by $\mathcal{T}$ as $R\langle\mathcal{T}\rangle$.

Lemma 3.1. *In the notation above,* $E_X(\varphi) = R\langle\mathcal{T}\rangle$.

Proof. Let $(B_0, \ldots, B_5)$ be an element of $E_X(\varphi)$, where B_i is a homogeneous polynomial of degree b. We consider the case $b \equiv 0 \bmod 4$.

Observe that we can rewrite each monomial term of B_i as $(c^{1/4} S^\ell T^k)^4 S^i T^{4-i}$ or $(c^{1/4} S^\ell T^k)^4$ for some integers ℓ, k, where $c \in k$ and $0 < i < 4$. After collecting terms and applying the Frobenius endomorphism, we obtain

$$B_i = a_{i1}^4 + a_{i2}^4 S^3 T + a_{i3}^4 S^2 T^2 + a_{i4}^4 S T^3,$$

where each a_{ij} is an element of R. Then, since $B_0G_0^4 + \cdots + B_5G_5^4 = 0$, substituting our expression for the B_i's and applying Frobenius, we obtain

$$\left(\sum_{i=0}^{5} a_{i1} G_i\right)^4 + \left(\sum_{i=0}^{5} a_{i2} G_i\right)^4 S^3 T + \left(\sum_{i=0}^{5} a_{i3} G_i\right)^4 S^2 T^2 + \left(\sum_{i=0}^{5} a_{i4} G_i\right)^4 S T^3 = 0.$$

The sums $\sum_{i=0}^{5} a_{ij} G_i$ are each themselves homogeneous polynomials. But since the degree of T in each term above is distinct modulo 4, the equation $\sum_{i=0}^{5} a_{ij} G_i = 0$ implies that $(a_{0j}, \ldots, a_{5j}) \in \Omega_X(\varphi)$ so that $(a_{0j}^4, \ldots, a_{5j}^4) \in \mathcal{T}$ for $1 \le j \le 4$.

Hence, every homogeneous element of $E_X(\varphi)$ is contained in the submodule generated by $\mathcal{T}$. Since the reverse containment is trivial, it follows that $E_X(\varphi) = R\langle\mathcal{T}\rangle$. The cases for $b \equiv 1, 2, 3 \bmod 4$ follow similarly. □

Proposition 3.2. *If* $x_i = (x_{i0}, \ldots, x_{i5})$, *for* $1 \le i \le 5$, *form a basis for* $\Omega_X(\varphi)$, *then* $y_i = (x_{i0}^4, \ldots, x_{i5}^4)$, *for* $1 \le i \le 5$, *form a basis for* $E_X(\varphi)$.

Proof. If $x_i \in \Omega_X(\varphi)$, then $y_i \in \mathcal{T}$, and every element of $\mathcal{T}$ is an R-linear combination of the y_i's. Since $E_X(\varphi) = R\langle\mathcal{T}\rangle$, every element of $E_X(\varphi)$ is also an R-linear combination of the y_i's so that the y_i's generate $E_X(\varphi)$. Moreover, $E_X(\varphi)$ is a free module of rank 5 over a domain, so the generators y_i for $E_X(\varphi)$ must also be linearly independent and hence form a basis. □

Accounting for twist, a simple computation using the results above gives us the following.

Corollary 3.3. *Let φ be a degree d morphism, and $e_1, \ldots, e_5$ be the splitting type of $\Omega_X(\varphi)$. If $f_1 = 4e_1 + 5d$, $f_2 = 4e_2 + 5d, \ldots, f_5 = 4e_5 + 5d$, then $f_1, \ldots, f_5$ is the splitting type of $E_X(\varphi)$.*

4. Numerology

We now utilize some facts about graded free modules in order to give constraints on potential splitting types. Given a graded free module

$$M = R(u_1) \oplus \cdots \oplus R(u_r),$$

one can observe that the Hilbert polynomial H_M is given by

$$H_M(m) = rm + u_1 + \cdots + u_r + r.$$

Let φ denote a free morphism of degree d into X. Noting that the map

$$\tilde{\varphi} : R(-d)_m^{\oplus n+1} \to R_m$$

is surjective for $m \gg 0$, we obtain

$$\begin{aligned} H_{\Omega(\varphi)}(m) &= \dim_k\big(\ker(R(-d)_m^{\oplus n+1} \to R_m)\big) \\ &= (n+1)(-d+m+1) - (m+1) \\ &= nm + -d(n+1) + n. \end{aligned}$$

A similar calculation shows that

$$H_{E_X(\varphi)}(m) = nm + d(n+1-5) + n.$$

We continue to refer to the splitting type components of $\Omega(\varphi)$ and $E_X(\varphi)$ as e_i and f_i, respectively. In both cases $n = r = 5$, so combining these two equations with the general form for the Hilbert polynomial of a graded free module, we obtain our first constraints:

$$\begin{aligned} e_1 + e_2 + e_3 + e_4 + e_5 &= -6d, \\ f_1 + f_2 + f_3 + f_4 + f_5 &= d. \end{aligned}$$

Recall from Section 2 that a curve is free or very free if $f_i \geq 0$ or $f_i > 0$, respectively, for each i. Since $f_i = 4e_i + 5d$, it follows that

$$e_i \geq -\frac{5d}{4},$$

where strict inequality implies the curve is very free. With these two bounds, we can quickly observe a few facts about curves of different degrees.

Remarks. (1) There exist no free curves in degrees 1, 2, 3, 6, and 7.

(2) Any free curve of degree not divisible by 4 must be very free.

(3) There are no very free curves in degrees 4 or 8.

(4) The splitting type of $\Omega(\varphi)$ of a free curve of degree 4 must be

$$(-5, -5, -5, -5, -4).$$

(5) The splitting type of $\Omega(\varphi)$ of a very free curve of degree 5 must be

$$(-6, -6, -6, -6, -6).$$

All of these observation follow directly from the two constraints. For example, in degree 6, $e_1 + e_2 + e_3 + e_4 + e_5 = -6d = -36$. However, each $e_i \geq -30/4 = -7.5$. So even if each e_i is at best -7, the e_i cannot sum to -36.

The rest of the remarks follow in a similar manner. Note that one can glean even more information about these curves from the constraints, but the remarks listed above are sufficient for our purposes.

5. Degree 4 and 5 morphisms into X

We will now show that there are no free morphisms of degrees 4 or 5 into X. A morphism $\varphi = (G_0, \ldots, G_5)$, where each $G_i = \sum_{j=0}^{d} a_{ij} S^{d-j} T^j$ is a homogeneous polynomials of degree d, gives us a $6 \times (d+1)$ matrix (a_{ij}). We will denote this matrix as M_φ.

Lemma 5.1. *If φ is a degree 4 or 5 free morphism into X, then M_φ has maximal rank.*

Proof. This follows from Remarks(4) and (5) by observing that for a degree d morphism into X, the transpose of M_φ is the matrix of the k-linear map

$$\tilde{\varphi}_d : (R(-d)^{\oplus 6})_d \to R_d. \qquad \square$$

Lemma 5.2.

(a) *There are no degree 4 free morphisms into X.*

(b) *There are no degree 5 free morphisms into X.*

Proof. (a) Assume a degree 4 free morphism $\varphi = (G_0, \ldots, G_5)$ exists. By the previous lemma, the 6×5 matrix $M_\varphi = (a_{ij})$ has maximal rank. Since permuting the G_i's does not affect the splitting type of $E_X(\varphi)$, we can assume that the first 5 rows of M_φ are linearly independent over k. Then $\det\big((a_{ij})_{i\le 4}\big) \neq 0$. Now consider the matrix $\overline{M}_\varphi = (a_{ij}^4)$. By the Frobenius endomorphism on k,

$$\det\big((a_{ij}^4)_{i\le 4}\big) = \det\big((a_{ij})_{i\le 4}\big)^4 \neq 0,$$

proving that $\overline{M}_\varphi$ has maximal rank as well.

Since $G_0^5 + \cdots + G_5^5 = 0$, computing the coefficients of $G_0^5 + \cdots + G_5^5$, we obtain for $0 \le j \le 4$

$$\sum_{i=0}^{5} a_{ij}^4 a_{i1} = 0 \quad \text{and} \quad \sum_{i=0}^{5} a_{ij}^4 a_{i3} = 0. \tag{5-1}$$

The kernel of the map $k^6 \to k^5$ given by right multiplication by the matrix $\overline{M}_\varphi$ has dimension 1 because $\operatorname{rank}(\overline{M}_\varphi) = 5$. By (5-1),

$$(a_{01}, a_{11}, \ldots, a_{51}), (a_{03}, a_{13}, \ldots, a_{53}) \in \ker(k^6 \to k^5),$$

and since these 6-tuples are columns of M_φ, they are linearly independent over k. Then $\dim_k\big(\ker(k^6 \to k^5)\big) \ge 2$, a contradiction.

(b) Assume $\varphi = (G_0, \ldots, G_5)$ is a degree 5 free morphism. By the previous lemma, the matrix $M_\varphi = (a_{ij})$ has maximal rank and is invertible. Thus $\overline{M}_\varphi = (a_{ij}^4)$ is invertible by the same argument above. Since $G_0^5 + \cdots + G_5^5 = 0$, computing the coefficients of the polynomial $G_0^5 + \cdots + G_5^5$, we get

$$\sum_{i=0}^{5} a_{ij}^4 a_{i2} = 0 \quad \text{for } 0 \le j \le 5.$$

Thus, the product of the row matrix $(a_{02}, a_{12}, \ldots, a_{52})$ and the matrix $\overline{M}_\varphi$ is 0, which is impossible because $(a_{02}, a_{12}, \ldots, a_{52}) \neq 0$ and $\overline{M}_\varphi$ is invertible. □

6. Computations for the degree 8 free curve

Let $\varphi : \mathbb{P}^1 \to \mathbb{P}^5$ be a morphism given by the 6-tuple

$$\begin{aligned}
&G_0 = S^7T, && G_1 = S^4T^4 + S^3T^5,\\
&G_2 = S^4T^4 + S^3T^5 + T^8, && G_3 = S^7T + S^6T^2 + S^5T^3 + S^4T^4 + S^3T^5,\\
&G_4 = S^8 + S^7T + S^6T^2 + S^5T^3 + S^4T^4 + S^3T^5 + T^8,\\
&G_5 = S^8 + S^7T + S^6T^2 + S^5T^3 + S^4T^4 + S^3T^5 + S^2T^6 + ST^7.
\end{aligned}$$

One can check by computer or by hand that this curve lies on the Fermat hypersurface $X \subset \mathbb{P}^5$. Due to twisting, the domain of the map $\tilde{\varphi} : R(-8)^{\oplus 6} \to R$ has its first non-trivial graded piece in dimension 8. The G_i are linearly independent over k, hence the kernel is trivial in dimension 8. The matrix for the map $\tilde{\varphi}_9 : R(-8)_9^{\oplus 6} \to R_9$ is

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

where each direct summand of the domain has a basis $\{(S, 0), (0, T)\}$, of which we take six copies (for total dimension 12), and the range has basis given by the degree 9 monomials in S and T, ordered by increasing T-degree (for total dimension 10). This matrix has rank 10, which means that the map in degree 9 is surjective. By rank-nullity, two dimensions of the kernel live in degree 9; denote the generators by x_1, x_2. Surjectivity of $\tilde{\varphi}$ in degree 9 implies surjectivity in all higher degrees. A second application of rank-nullity gives $\dim_k \Omega(\varphi)_{10} = 7$. Four of the generators are inherited from the previous degree, taking the forms

$$x_1 S, x_2 S, x_1 T, x_2 T.$$

We conclude that there are three additional generators in degree 10. Therefore, the splitting type of $\Omega_X(\varphi)$ is $(e_1, \ldots, e_5) = (-10, -10, -10, -9, -9)$, which corresponds to a splitting type for $E_X(\varphi)$ of $(f_1, \ldots, f_5) = (0, 0, 0, 4, 4)$, hence the curve is free.

7. A very free rational curve of degree 9

We conclude by giving an example of a degree 9 very free curve lying on X. Let $\varphi : \mathbb{P}^1 \to \mathbb{P}^5$ be a morphism into the Fermat hypersurface given by the 6-tuple

$$\begin{aligned}
&G_0 = S^4T^5, && G_1 = S^9 + S^8T + S^5T^4, \\
&G_2 = S^9 + S^4T^5 + ST^8, && G_3 = S^9 + S^8T + S^4T^5 + S^3T^6 + S^2T^7 + ST^8, \\
&G_4 = S^9 + S^5T^4 + S^3T^6 + S^2T^7 + ST^8 + T^9, \\
&G_5 = S^7T^2 + S^6T^3 + S^5T^4 + S^3T^6 + S^2T^7 + ST^8 + T^9.
\end{aligned}$$

Let $e_1, \ldots, e_5$ again denote the splitting type of $\Omega_X(\varphi)$. As in Section 6, we know that $e_i \leq -9$. Since the G_i are linearly independent over k, $\dim_k(\Omega_X(\varphi)_9) = 0$. Next we claim that $\tilde{\varphi_{10}} : R_1^{\oplus 6} \to R_{10}$ is surjective. In fact, it can be checked that the $\tilde{\varphi}(b_i)$ span R_{10}, where the b_i are distinct basis elements of $R_1^{\oplus 6}$. It follows that $\tilde{\varphi_n} : R(-9)_n^{\oplus 6} \to R_n$ is surjective for $n \geq 10$. Hence,

$$\dim_k(\Omega_X(\varphi)_{10}) = \dim_k(R_1^{\oplus 6}) - \dim_k(R_{10}) = 1,$$
$$\dim_k(\Omega_X(\varphi)_{11}) = \dim_k(R_2^{\oplus 6}) - \dim_k(R_{11}) = 6.$$

After reordering, this yields $(e_1, \ldots, e_5) = (-11, -11, -11, -11, -10)$, which corresponds to the splitting type $(1, 1, 1, 1, 5)$ of $E_X(\varphi)$, showing that φ is very free. This completes the proof of Theorem 1.1.

References

[Beauville 1990] A. Beauville, "Sur les hypersurfaces dont les sections hyperplanes sont à module constant", pp. 121–133 in *The Grothendieck Festschrift, I*, edited by P. Cartier et al., Progr. Math. **86**, Birkhäuser, Boston, 1990. MR 91m:14070 Zbl 0723.14031

[Debarre 2001] O. Debarre, *Higher-dimensional algebraic geometry*, Springer, New York, 2001. MR 2002g:14001 Zbl 0978.14001

[Grothendieck 1957] A. Grothendieck, "Sur la classification des fibrés holomorphes sur la sphère de Riemann", *Amer. J. Math.* **79** (1957), 121–138. MR 19,315b Zbl 0079.17001

[Hartshorne 1977] R. Hartshorne, *Algebraic geometry*, Graduate Texts in Mathematics **52**, Springer, New York, 1977. MR 57 #3116 Zbl 0367.14001

[Kollár 1996] J. Kollár, *Rational curves on algebraic varieties*, Ergebnisse der Math. (3) **32**, Springer, Berlin, 1996. MR 98c:14001 Zbl 0877.14012

[Shen 2012] M. Shen, "Rational curves on Fermat hypersurfaces", *C. R. Math. Acad. Sci. Paris* **350**:15-16 (2012), 781–784. MR 2981353 Zbl 06100626

[Zhu 2011] Y. Zhu, "Fano hypersurfaces in positive characteristic", preprint, 2011. arXiv 1111.2964

Received: 2012-08-05 Revised: 2012-11-06 Accepted: 2012-11-08

mbridg5@uic.edu *Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, IL 60607-7045, United States*

rankeya@umich.edu *Department of Mathematics, University of Michigan, Ann Arbor, MI 48109, United States*

jle2129@columbia.edu *Department of Mathematics, Columbia University, New York, NY 10027, United States*

mgnewman@umich.edu *Department of Mathematics, University of Michigan, Ann Arbor, MI 48109, United States*

jy2433@columbia.edu *Department of Mathematics, Columbia University, New York, NY 10027, United States*

INVOLVE 6:4 (2013)
dx.doi.org/10.2140/involve.2013.6.447

Irreducible divisor simplicial complexes

Nicholas R. Baeth and John J. Hobson

(Communicated by Scott Chapman)

For an integral domain D, the *irreducible divisor graph* $G_D(x)$ of a nonunit $x \in D$ gives a visual representation of the factorizations of x. Here we consider a higher-dimensional generalization of this notion, the *irreducible divisor simplicial complex* $S_D(x)$. We show how this new structure is a true generalization of $G_D(x)$, and show that it often carries more information about the element x and the domain D than its two-dimensional counterpart.

1. Introduction and preliminaries

The concept of an irreducible divisor graph of an element x in an integral domain D was introduced in [Coykendall and Maney 2007]. The vertices of this graph are a prechosen set of irreducible divisors of x, and any pair of vertices are connected by an edge if and only if the corresponding irreducible divisors appear in the same factorization of x. The relevance of the irreducible divisor graph was illustrated in the same paper and in [Axtell et al. 2011]: an integral domain D is a unique factorization domain if and only if each irreducible divisor graph is complete if and only if each irreducible divisor graph is connected.

Since their introduction, irreducible divisor graphs have been studied in the context of integral domains [Axtell et al. 2011; Maney 2008] and in more general contexts [Axtell and Stickles 2008; Bachman et al. 2012; Smallwood and Swartz 2009; 2008]. Despite the appealing result mentioned above, it is difficult to pick out the factorizations of an element given its irreducible divisor graph. In short, irreducible divisor graphs fail to give us all the information we might wish to glean about an element's factorizations.

Our main goal is to introduce the concept of an *irreducible divisor simplicial complex*, effectively a generalization of the irreducible divisor graph to higher dimensions. As we shall see, irreducible divisor simplicial complexes often convey more information about the factorization of an element than its two-dimensional

MSC2010: primary 13A05; secondary 55U10.
Keywords: factorization, simplicial complex.
The research reported here is part of Hobson's Masters thesis at the University of Central Missouri.

counterpart. Maney [2008] uses homologies to study irreducible divisor graphs, linking irreducible divisor graphs to certain zeroth and first homologies. Moreover, higher homologies are considered which, although not explicitly mentioned in [Maney 2008], are related to irreducible divisor simplicial complexes. This gives yet another motivation for studying this new construct.

We now provide a brief overview of the ring- and graph-theoretic terminology that will be required in the sequel. Throughout, D will denote an integral domain, D^* the nonzero elements of D, and $U(D)$ the units of D. It will often be convenient for us to speak of the set of nonzero nonunits of D which will be denoted by $D^* \setminus U(D)$. An element $x \in D^* \setminus U(D)$ is *irreducible* if whenever $x = yz$ with $y, z \in D$, then either $y \in U(D)$ or $z \in U(D)$. We say $x \in D^* \setminus U(D)$ is *prime* if whenever $x \mid yz$ with $y, z \in D$, then either $x \mid y$ or $x \mid z$. An element $x \in D$ is *square-free* if it is not divisible by any perfect square $z^2 \in D^* \setminus U(D)$, that is, if z^2 divides x for $z \in D$, then $z \in U(D)$. Two elements a and b of D are called *associates* if $a = ub$ where $u \in U(D)$. The relation $a \sim b$ on elements of D is an equivalence relation that partitions D into *associate classes*. We denote the set of irreducibles in D as $\mathrm{Irr}(D) = \{x : x \text{ is irreducible}\}$ and define $\overline{\mathrm{Irr}}(D)$ to be a (prechosen) set of associate class representatives, one from each class of nonzero associates. We denote the irreducible divisors of a particular element $x \in D$ as $\mathrm{Irr}(x)$ and set $\overline{\mathrm{Irr}}(x) = \mathrm{Irr}(x) \cap \overline{\mathrm{Irr}}(D)$. Note that by considering only $\overline{\mathrm{Irr}}(D)$, we do not distinguish elements in D from their associates and we are implicitly working in the *reduced multiplicative monoid* $D^\bullet_{\mathrm{red}}$ (see [Geroldinger and Halter-Koch 2006, Chapter 1]), which is the multiplicative monoid whose elements are associate classes and whose identity is the set of units.

As we will be studying the factorization of elements of D as products of irreducibles, it will be useful to restrict our study to only *atomic domains* where each element $x \in D^* \setminus U(D)$ can be factored into a finite product of irreducible elements. Clearly every prime in an integral domain is irreducible. If D is an atomic domain, then D is a *unique factorization domain* (UFD) if and only if all irreducibles in D are prime [Geroldinger and Halter-Koch 2006, Theorem 1.1.10.2]. We now give a brief introduction of some special types of atomic domains and related terminology. We say D is a *finite factorization domain* (FFD) if every nonzero nonunit in D has only finitely many distinct nonassociate irreducible divisors. If D is a finite factorization domain, the *set of lengths* (of factorizations) of $x \in D^* \setminus U(D)$ is $L(x) = \{t : x = a_1 a_2 \cdots a_t \text{ where each } a_i \text{ is irreducible}\}$. A FFD D is a *bounded factorization domain* (BFD) if there is a bound on the length of factorization into products of irreducible elements for each nonzero nonunit element in D. If $|L(x)| = 1$ for all $x \in D^* \setminus U(D)$, we say that D is a *half-factorial domain* (HFD). The *elasticity* $\rho(D)$ of D gives a measure of how far D is from being a HFD; it is defined as the supremum of the *elasticity* $\rho(x) := \max L(x)/\min L(x)$ of each element $x \in D^* \setminus U(D)$.

A *graph* is an ordered pair of sets (V, E), where V is called the *vertex set*, and E is the *edge set*, whose elements are subsets of V of cardinality 2. We denote an edge between vertices a and b as $\{a, b\}$ and note that the edge $\{a, b\}$ is the same as the edge $\{b, a\}$. The edge $\{a, b\}$ is said to be *incident* with both vertices a and b. We denote the set of vertices of a graph G as $V(G)$ and the set of edges of G as $E(G)$. In addition, we define a *loop* to be an edge between a and itself. We now define a higher-dimensional analog of graphs. A *simplicial complex* S is an ordered pair (V, F) where V is a set of vertices and the *set of faces* F is a collection of subsets of V satisfying: (1) $\{v\} \in F$ for all $v \in V$ (vertices are faces) and (2) if $\sigma \in F$ and $\tau \subseteq \sigma$, then $\tau \in F$ (subsets of faces are faces). As with graphs, we denote the set of vertices of S as $V(S)$, and the set of faces of S as $F(S)$. The *dimension* of a face β of finite cardinality in a simplicial complex S is one less than its cardinality and is denoted as $\dim(\beta) = |\beta| - 1$. Faces with maximal cardinality (with respect to inclusion) are referred to as *facets*. For a nonnegative integer k, the *k-skeleton* $K_k(S)$ of a simplicial complex S is the subsimplex of S consisting of all the faces of S whose dimension is at most k. We note that $K_1(S)$ is a graph.

2. Irreducible divisor graphs

In this section, we introduce the irreducible divisor graph of an element in an atomic domain and summarize results from [Axtell et al. 2011; Coykendall and Maney 2007].

Definition 2.1. Let D be an atomic domain and let $x \in D^* \setminus U(D)$. The *irreducible divisor graph* of x, denoted $G_D(x)$, is given by (V, E) where the vertex set $V = \{\overline{\text{Irr}}(x) : x \in D\}$, and given $y_1, y_2 \in V$, there is an edge $\{y_1, y_2\} \in E$ between vertices y_1 and y_2 if and only if $y_1 y_2 \mid x$.

When it is clear from context, we will drop the subscript D from $G_D(x)$ and write $G(x)$. If the same element $a \in \overline{\text{Irr}}(D)$ appears multiple times in a particular factorization of $x \in D^* \setminus U(D)$, then we add one or more loops to the vertex a in $G(x)$. We place n loops on vertex a provided $a^{n+1} \mid x$ and $a^{n+2} \nmid x$. When a vertex has more than one loop, we will denote the number of loops in the graph with a superscript over the loop.

Example 2.2. Let $D = \mathbb{Z}[\sqrt{-5}]$ and consider the irreducible divisor graph $G(18)$. Recall that 18 factors as

$$18 = 2 \cdot 3^2 = 3(1 + \sqrt{-5})(1 - \sqrt{-5}) = 2(2 + \sqrt{-5})(2 - \sqrt{-5}).$$

To simplify notation, we set $\alpha = (1 + \sqrt{-5})$ and $\beta = (2 + \sqrt{-5})$, with $\overline{\alpha}$ and $\overline{\beta}$ denoting their complex conjugates. Using the rules provided in Definition 2.1, we construct the irreducible divisor graph shown in Figure 1. For example, $\{2, \beta\}$ is an

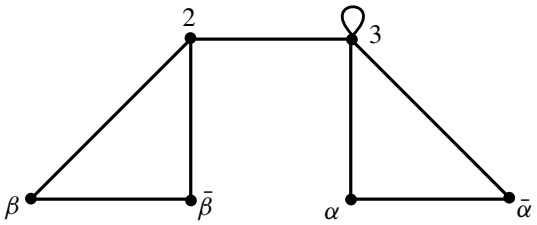


Figure 1. $G(18)$ in $\mathbb{Z}[\sqrt{-5}]$.

edge in $G(18)$ since $2\beta \mid 18$. Since $3^2 \mid 18$ but $3^3 \nmid 18$, we place a single loop on vertex 3. We note $G(18)$ is connected but not complete.

One of the goals in studying the irreducible divisor graph of an element x in an integral domain D is to be able to draw conclusions about the factorization of the element in question and of other elements of D. When we look closely at $G(18)$ and briefly try to forget what the factorizations of 18 look like, we can see that there will be some factorization that will include β, $\bar{\beta}$, and 2. Since β and $\bar{\beta}$ are connected by an edge in $G(18)$, $18 = \beta\bar{\beta}x$ for some $x \in D^* \setminus U(D)$. Similarly, $18 = 2\beta y$ and $18 = 2\bar{\beta}z$ for some $y, z \in D^* \setminus U(D)$. Since all irreducible factors of x appear together with β and $\bar{\beta}$ in a factorization of 18 and since none of 2, β, or $\bar{\beta}$ are looped in $G(18)$, it must be the case that $x = 2$. Similarly, $y = \bar{\beta}$ and $z = \beta$. Thus 18 factors as $18 = 2\beta\bar{\beta}$, and this factorization corresponds to the complete subgraph with vertex set $\{2, \beta, \bar{\beta}\}$. Note that the maximal complete subgraphs $\{2, 3\}$ and $\{3, \alpha, \bar{\alpha}\}$ also correspond to factorizations of 18. However, this correspondence requires a priori knowledge of the factorizations of 18 in $\mathbb{Z}[\sqrt{-5}]$ and we cannot see simply by looking at the graph $G(18)$ what the remaining factorizations of 18 are. This problem occurs because of the loop on the vertex 3. When we look at the graph, we really have no way of assigning the element 3^2 to any one factorization. As irreducible divisor graphs get more complicated with more irreducible divisors, we will have much difficulty in deciphering what the factorization of a particular element is by simply looking at its irreducible divisor graph.

In most situations, factorizations do not correspond to complete subgraphs. Conversely, complete subgraphs need not correspond to factorizations. We now consider another example where this is certainly the case.

Example 2.3. Let $D = \mathbb{Z}[\sqrt{-5}]$ and consider $G(108)$. By considering norms, we see that 108 factors only as

$$\begin{aligned} 108 &= 2^2 3^3 = 2 \cdot 3^2 (1 + \sqrt{-5})(1 - \sqrt{-5}) \\ &= 2^2 \cdot 3(2 + \sqrt{-5})(2 - \sqrt{-5}) = 3(1 + \sqrt{-5})^2 (1 - \sqrt{-5})^2 \\ &= 2(1 + \sqrt{-5})(1 - \sqrt{-5})(2 + \sqrt{-5})(2 - \sqrt{-5}). \end{aligned}$$

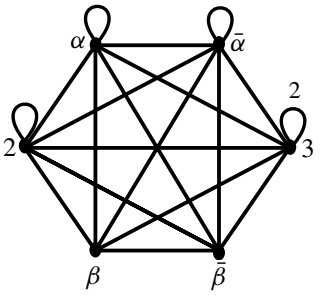


Figure 2. $G(108)$ in $D = \mathbb{Z}[\sqrt{-5}]$.

As before, let $\alpha = (1 + \sqrt{-5})$ and $\beta = (2 + \sqrt{-5})$ with $\overline{\alpha}$ and $\overline{\beta}$ denoting their complex conjugates. The irreducible divisor graph is given in Figure 2. This graph is complete, even though D is not a UFD. Certainly not all complete subgraphs correspond to factorizations of 108, thus making it hard to glean factorization-theoretic information from the irreducible divisor graph. We will return to this example later in Example 3.4.

For variety, we now give an example of the irreducible divisor graph of an element in a nonhalf-factorial domain.

Example 2.4. Let k be a field and let $D = k[x^{10}, x^{12}, x^{18}, x^{33}]$ denote the subring of the polynomial ring $k[x]$. Then $x^{66} \in D$ and the only irreducible divisors of x^{66} in D are x^{10}, x^{12}, x^{18}, and x^{33}. Moreover, x^{66} factors only as

$$x^{66} = (x^{12})(x^{18})^3 = (x^{12})^4(x^{18}) = (x^{10})^3(x^{18})^2 = (x^{10})^3(x^{12})^3 = (x^{33})^2.$$

Therefore, the irreducible divisor graph $G_D(x^{66})$, shown in Figure 3, consists of a complete graph on three vertices (x^{10}, x^{12}, and x^{18}) with 2, 3, and 2 loops on these respective vertices, along with a single vertex (x^{33}) having a single loop.

We now turn to several important results that can be found in [Axtell et al. 2011; Coykendall and Maney 2007]. The first result gives necessary and sufficient

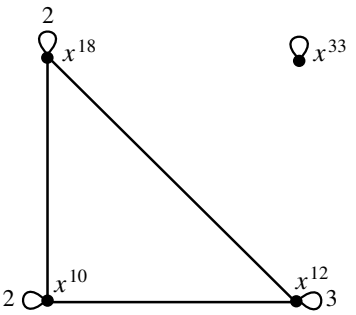


Figure 3. $G(x^{66})$ in $D = k[x^{10}, x^{12}, x^{18}, x^{33}]$.

conditions for an atomic domain to be a UFD. The second result gives a bound on the elasticity of an element given by its irreducible divisor graph. We will prove generalizations of these results in Section 3.

Theorem 2.5 [Axtell et al. 2011, Theorem 2.1]. *Let D be an atomic domain. The following statements are equivalent.*

(1) *D is a UFD.*

(2) *$G(x)$ is complete for all $x \in D^* \setminus U(D)$.*

(3) *$G(x)$ is connected for all $x \in D^* \setminus U(D)$.*

Proposition 2.6 [Axtell et al. 2011, Proposition 4.1]. *Let x be an element which is not irreducible of a BFD D. Then $\rho(x)$ does not exceed*

$$\tfrac{1}{2}\max\{t + l : G(x) \text{ contains a complete subgraph with } t \text{ vertices and } l \text{ loops}\}.$$

The proof of this result makes note of the fact that if $x = a_1^{m_1} a_2^{m_2} \cdots a_n^{m_n}$, where $a_1, a_2, \ldots, a_n \in \overline{\text{Irr}}(x)$, then $G(x)$ contains a complete subgraph with a vertex corresponding to each a_i. In the special case that x is square-free we produce a more accurate result.

Corollary 2.7 [Axtell et al. 2011, Corollary 4.4]. *Let x be a square-free nonirreducible element of a domain D. Then*

$$\rho(x) \leq \tfrac{1}{2}\max\{t : G(x) \text{ contains a complete subgraph with } t \text{ vertices}\}.$$

We note that the bounds given in Proposition 2.6 and Corollary 2.7 are, in general, not tight. There are three reasons: First, it is often the case that not all vertices belonging to a complete subgraph of $G_D(x)$ are involved in a single factorization of x. Second, the minimal length of a factorization of x is often larger than 2. Finally, when counting loops, it is impossible to know how many come from a given factorization of x. As was done in Corollary 2.7, assuming that x is square-free eliminates the third problem. We will consider these other two issues in Section 3.

3. Irreducible divisor simplicial complexes

We now extend the definition of irreducible divisor graphs given in Section 2 to higher dimensions. We do this in the hopes that this extension will yield more information about the factorization of elements in an atomic domain. After giving a couple of examples, we generalize the results given in Section 2, but in terms of irreducible divisor simplicial complexes.

Definition 3.1. Let D be an atomic domain and let $x \in D^* \setminus U(D)$. The *irreducible divisor simplicial complex* of x, denoted $S_D(x)$, is given by (V, F) with vertex set V given by $V = \{\overline{\text{Irr}}(x) : x \in D\}$ and with $\{y_1, y_2, \ldots, y_n\} \in F$ a face if and only if $y_1 y_2 \cdots y_n \mid x$. In addition, to satisfy convention, we also put $\varnothing \in F$.

Whenever the context is clear we will drop the subscript D from $S_D(x)$ giving $S(x)$.

Remark 3.2. Let $S(x) = (V, F)$ be an irreducible simplicial complex. Clearly F is a collection of subsets of V. If $y \in V = \overline{\mathrm{Irr}}(x)$, then $\{y\} \in F$ since $y \mid x$ and hence vertices are faces. Second, suppose that $\sigma \in F$ and $\tau \subseteq \sigma$. Since $\sigma \in F$, we know $\sigma = \{y_1, \ldots, y_n\}$ where $y_1 \cdots y_n \mid x$. Hence $\tau = \{y_{i_1}, \ldots, y_{i_j}\}$, some subcollection of the y_i, and clearly $y_{i_1} \cdots y_{i_j} \mid x$. Thus $\tau \in F$, and hence subsets of faces are faces. Therefore irreducible simplicial complexes are indeed simplicial complexes.

We graphically represent irreducible divisor simplicial complexes and irreducible divisor graphs in similar ways. Points represent vertices and edges represent faces of dimension 1. If we have some element x which factors into irreducibles as $x_1^{m_1} \cdots x_n^{m_n}$ with distinct irreducible x_i and $m_i \geq 1$ for all i, then the vertex representing x_i will be drawn with $m_i - 1$ loops. Graphically, we illustrate two-dimensional faces by shaded triangles and three-dimensional faces by solid tetrahedra. We have no effective way to graphically depict higher-dimensional faces, so readers are on their own.

Example 3.3. Recall the irreducible divisor graph $G(18)$ in Figure 1. We now show the corresponding irreducible divisor simplicial complex $S(18)$:

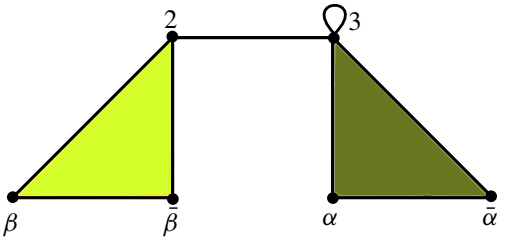


Figure 4. $S(18)$ in $\mathbb{Z}[\sqrt{-5}]$.

Here we have the same general structure as $G(18)$, but we now have two-dimensional facets $\{2, \beta, \bar{\beta}\}$ and $\{3, \alpha, \bar{\alpha}\}$ which are represented graphically as shaded faces. In this higher-dimensional structure, we avoid the difficulty in determining factorizations as in Example 2.2. Indeed, the facets $\{2, 3\}$, $\{2, \beta, \bar{\beta}\}$, and $\{3, \alpha, \bar{\alpha}\}$ correspond directly to the factorizations of 18. We will make this idea more precise in Propositions 3.7 and 3.8.

Example 3.4. We now consider the irreducible divisor simplicial complex $S(108)$ in $D = \mathbb{Z}[\sqrt{-5}]$. Recall that 108 factors as

$$108 = 2^2 3^3 = 2 \cdot 3^2 \alpha\bar{\alpha} = 2^2 3 \beta\bar{\beta} = 3\alpha^2\bar{\alpha}^2 = 2\alpha\bar{\alpha}\beta\bar{\beta}.$$

If we investigate Figure 2, we can see the difficulty in extracting a particular factorization by simply analyzing the graph. However, this becomes much easier if we consider the irreducible divisor simplicial complex $S(108)$. We have that

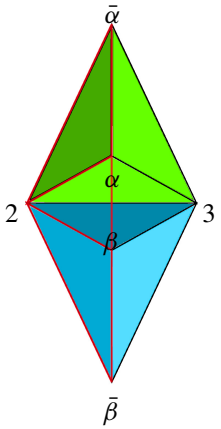


Figure 5. $S(108)$ in $\mathbb{Z}[\sqrt{-5}]$.

$S(108) = (V, F)$, with $V = \{2, 3, \alpha, \bar{\alpha}, \beta, \bar{\beta}\}$ and $F = \{\varnothing\} \cup F_0 \cup F_1 \cup F_2 \cup F_3 \cup F_4$, where F_i denotes the set of faces of $S(108)$ with dimension i:

$$F_0 = \{\{v\} : v \in V\},$$
$$F_1 = \{S \subseteq V : |S| = 2\},$$
$$F_2 = \{S \subseteq V : |S| = 3\} - \{\{3, \alpha, \beta\}, \{3, \alpha, \bar{\beta}\}, \{3, \bar{\alpha}, \beta\}, \{3, \bar{\alpha}, \bar{\beta}\}\},$$
$$F_3 = \big\{\{2, 3, \beta, \bar{\beta}\}, \{2, 3, \alpha, \bar{\alpha}\}, \{\alpha, \bar{\alpha}, \beta, \bar{\beta}\}, \{2, \bar{\alpha}, \beta, \bar{\beta}\},$$
$$\{2, \alpha, \beta, \bar{\beta}\}, \{2, \alpha, \bar{\alpha}, \bar{\beta}\}, \{2, \alpha, \bar{\alpha}, \beta\}\big\},$$
$$F_4 = \{2, \alpha, \bar{\alpha}, \beta, \bar{\beta}\}.$$

The maximal faces (facets) of $S(x)$ are

$$\{2, \alpha, \bar{\alpha}, \beta, \bar{\beta}\}, \quad \{2, 3, \beta, \bar{\beta}\}, \quad \{2, 3, \alpha, \bar{\alpha}\}.$$

Note that in Figure 5, the red-colored outline illustrates the 4-dimensional facet $\{2, \alpha, \bar{\alpha}, \beta, \bar{\beta}\}$. Unlike in $G(108)$, we can actually see that there are factorizations of 108 that contain $2, \alpha, \bar{\alpha}, \beta$, and $\bar{\beta}$, since they form a face of $S(108)$. We can also conclude that there is a factorization of x that only contains $2, 3, \alpha$, and $\bar{\alpha}$, a fact that is not immediately apparent when examining $G(108)$. If we consider only $G(108)$ and consider the set $A = \{2, 3, \alpha, \bar{\alpha}\}$, we see no clear way of proving that a factorization of 108 given by $2 \cdot 3 \cdot \alpha\bar{\alpha}$ will not include β or $\bar{\beta}$. After all, there are edges connecting β or $\bar{\beta}$ to each element of A. In other words, the graph $G(108)$ does not seem to provide enough information to support the conclusion that $108 = 2^i 3^j \alpha^k \bar{\alpha}^l$ for $i, j, k, l \geq 1$. In contrast, $S(108)$ contains far more information, as we will see in the results that follow.

Example 3.5. Recall the element $x^{66} \in D = k[x^{10}, x^{12}, x^{18}, x^{33}]$ from Example 2.4. Since no three distinct irreducible divisors of x^{66} occur together in a factorization

of x^{66}, the irreducible divisor simplicial complex contains no faces of dimension higher than 1 and $S_D(x^{66}) = G_D(x^{66})$ as shown in Figure 3. Even though these two constructions give identical objects in this case, the simplicial complex carries more information. In particular, only by looking at $S_D(x^{66})$ can we see that there are no factorizations involving more than two distinct irreducible factors.

We now generalize and extend the results from Section 2. First we note that the irreducible divisor simplicial complex $S_D(x)$ properly contains as a subsimplex the irreducible divisor graph $G_D(x)$.

Proposition 3.6. *For D an atomic domain and $x \in D^* \setminus U(D)$, we have*

$$K_1(S(x)) = G(x).$$

Proof. Let $G(x) = (V, E)$ denote the irreducible divisor graph of x, and let $S(x) = (V', F)$ denote the irreducible divisor simplicial complex of x. By definition,

$$V' = V = \overline{\mathrm{Irr}}(x).$$

Furthermore, $E \subseteq F$ since if $\{a, b\} \in E$, then $ab \mid x$ and hence $\{a, b\} \in F$. Moreover, if $\{a, b\}$ is a one-dimensional face of F, then $ab \mid x$ and hence $\{a, b\} \in E$. That is, the one-dimensional faces of $S(x)$ are precisely the edges of $G(x)$. □

The following results give a means for finding factorizations of an element x by considering $S_D(x)$.

Proposition 3.7. *For D an atomic domain and $x \in D^* \setminus U(D)$, let $A = \{a_1, \ldots, a_n\}$ be a facet of the irreducible divisor simplicial complex $S(x)$. Then there exists a factorization of x given by $x = a_1^{m_1} \cdots a_n^{m_n}$, where $m_i \geq 1$ for each i.*

Proof. Since A is a face of $S(x)$, we know that $a_1 \cdots a_n \mid x$. In fact, since $x/(a_1 a_2 \cdots a_n)$ also has a factorization, there is a factorization of x that involves each a_i. Suppose, by way of contradiction, that there exists some factorization of x of the form $a_1^{m_1} \cdots a_n^{m_n} b_1 \cdots b_k$, where each b_j is irreducible and b_j is not an associate of a_i for all i, j. Then by the definition of $S(x)$, $\{a_1, \ldots, a_n, b_1\}$ is a face of $S(x)$ properly containing A, contradicting the fact that A is a facet of $S(x)$. □

The converse to Proposition 3.7 does not hold in general as seen in Example 3.4. Indeed, $108 = 2^2 3^3 = 3\alpha^2\overline{\alpha}^2$ and yet neither $\{2, 3\}$ nor $\{3, \alpha, \overline{\alpha}\}$ is a facet since they are properly contained in the facet $\{2, 3, \alpha, \overline{\alpha}\}$. However, if we apply an additional restriction we find a partial converse.

Proposition 3.8. *Let D be an atomic domain and suppose $x \in D^* \setminus U(D)$ is square-free. Then every factorization of x corresponds to a facet of $S(x)$.*

Proof. By way of contradiction, suppose there exists a factorization $x = a_1 a_2 \cdots a_n$, with each a_i irreducible, corresponding to the face $A = \{a_1, a_2, \ldots, a_n\}$ of $S(x)$ that is not a facet. That is, $A \subsetneq B$ for some facet $B = \{a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_m\}$ of $S(x)$, where no b_i is associate to any a_i. Applying Proposition 3.7 and the fact that x is square-free, the facet B corresponds to the factorization $x = a_1 a_2 \cdots a_n b_1 b_2 \cdots a_m$. Setting these two factorizations equal, we have

$$x = a_1 a_2 \cdots a_n = a_1 a_2 \cdots a_n b_1 b_2 \cdots b_m.$$

As D is an integral domain, we may repeatedly apply left-cancellation to find that $1 = b_1 \cdots b_m$. This is a contradiction, since each of the b_i is a nonunit irreducible of D. Hence, $A = B$ is a facet of $S(x)$. □

We now produce a result analogous to Theorem 2.5 providing another necessary and sufficient condition for an integral domain D to be a UFD. Recall that if X is a set, then $\mathcal{P}(X)$ denotes the power set of X consisting of all subsets of X. We abuse notation and write $\mathcal{P}(X)$ to denote the simplicial complex $(X, \mathcal{P}(X))$ with vertex set X and face set $\mathcal{P}(X)$. Recall that $V(\mathcal{P}(X)) = X$ and $F(\mathcal{P}(X)) = \mathcal{P}(X)$. Recall from Definition 3.1 that if the singleton $\{y\}$ is a face of $S(x)$, then y is an irreducible divisor of x. In the next theorem, we may safely ignore all loops in both $S(x)$ and $G(x)$.

Theorem 3.9. *Let D be an atomic domain. The following are equivalent.*

(1) *For every $x \in D^* \setminus U(D)$, $S(x) = \mathcal{P}(A)$ for some $A \subseteq \overline{\mathrm{Irr}}(x)$.*

(2) *D is a UFD.*

Proof. Assume (1) and let $x \in D^* \setminus U(D)$. Then $S(x) = \mathcal{P}(A)$ for some $A \subseteq \overline{\mathrm{Irr}}(x)$. Since $G(x) = K_1(S(x))$ by Proposition 3.6, and since $K_1(\mathcal{P}(A))$ is a complete graph, $G(x)$ is complete. Since this holds for all $x \in D^* \setminus U(D)$, D is a UFD by Theorem 2.5.

If D is a UFD, any x factors uniquely as $x = a_1^{m_1} \cdots a_n^{m_n}$, $m_i \geq 1$. Then $a_{i_1} \cdots a_{i_t} \mid x$ for any subset $\{a_{i_1}, \ldots, a_{i_t}\} \subseteq \{a_1, \ldots, a_n\}$, and hence $F(S(x)) = \mathcal{P}(\{a_i, \ldots, a_n\})$. That is, $S(x) = \mathcal{P}(\overline{\mathrm{Irr}}(x))$. □

We now examine another necessary and sufficient condition for an integral domain D to be a UFD. First, we require a definition and two lemmas. Recall that for two simplicial complexes $S = (V, F)$ and $T = (W, G)$, their *join* $S * T$ is the simplicial complex with vertex set $V \cup W$ and with face set $\{A \cup B : A \in F, B \in G\}$.

Lemma 3.10. *Let A and B be two sets. As simplicial complexes, $\mathcal{P}(A \cup B) = \mathcal{P}(A) * \mathcal{P}(B)$.*

Proof. First we show that the vertex sets are equal. Suppose $a \in V(\mathcal{P}(A \cup B))$. Then $a \in A \cup B$, which by definition means $a \in V(\mathcal{P}(A) * \mathcal{P}(B))$. For the other containment, suppose $b \in V(\mathcal{P}(A) * \mathcal{P}(B))$. By definition, $b \in A \cup B$ and hence $b \in V(\mathcal{P}(A \cup B))$.

Now we show that $\mathcal{P}(A \cup B)$ and $\mathcal{P}(A) * \mathcal{P}(B)$ have the same face set. Let $\alpha \in F(\mathcal{P}(A \cup B))$, that is, $\alpha \subseteq A \cup B$. Set $\alpha_A := \alpha \cap A \subseteq A$ and $\alpha_B := \alpha \setminus \alpha_A \subseteq B$. Clearly $\alpha = \alpha_A \cup \alpha_B$, and hence $\alpha \in F(\mathcal{P}(A) * \mathcal{P}(B))$. To show the other containment, select $\alpha \in F(\mathcal{P}(A) * \mathcal{P}(B))$ and write $\alpha = \alpha_A \cup \alpha_B$ for some $\alpha_A \subseteq A$, $\alpha_B \subseteq B$. Then $\alpha \subseteq A \cup B$, and thus $\alpha \in F(\mathcal{P}(A \cup B))$. Since $\mathcal{P}(A \cup B)$ and $\mathcal{P}(A) * \mathcal{P}(B)$ have the same vertex and face sets, they are equal as simplicial complexes. □

Lemma 3.11. *Let $a, b \in D^* \setminus U(D)$. Then $V(S(b)) \cup V(S(a)) \subseteq V(S(ab))$. Moreover, if D is a UFD, then equality holds.*

Proof. Suppose $x \in V(S(a)) \cup V(S(b))$. If $x \in V(S(a))$, then $x \mid a$. If $x \in V(S(b))$, then $x \mid b$. In either case, $x \mid ab$ and hence $x \in V(S(ab))$.

Now suppose that D is a UFD and let $x \in V(S(ab))$. Then $x \mid ab$, with x irreducible and hence prime. If $x \mid a$, then $x \in V(S(a))$. If $x \nmid a$, then $x \mid b$ and hence $x \in V(S(b))$. Thus $x \in V(S(a)) \cup V(S(b))$. □

Theorem 3.12. *Let D be an atomic domain. The following are equivalent.*

(1) *$S(a) * S(b) = S(ab)$ for all $a, b \in D^* \setminus U(D)$.*

(2) *D is a UFD.*

Proof. Suppose D is not a UFD. Then there exists an irreducible $z \in D$ that is not prime. That is, there exists $a, b \in D$ where $z \mid ab$, but $z \nmid a$ and $z \nmid b$. Since $z \mid ab$, we have $z \in V(S(ab))$. We now consider $S(a) * S(b)$. By definition, $z \notin V(S(a))$ and $z \notin V(S(b))$, and hence $z \notin V(S(a)) \cup V(S(b))$. But then $z \notin V(S(a) * S(b))$, since $V(S(a) * S(b)) = V(S(a)) \cup V(S(b))$. Therefore $S(a) * S(b) \neq S(ab)$.

Now let D be a UFD and let $a, b \in D^* \setminus U(D)$. We want to show that $S(a) * S(b) = S(ab)$. Since D is a UFD, we know from Theorem 3.9 that $S(x) = \mathcal{P}(V(S(x)))$ for any $x \in D^* \setminus U(D)$. From Lemma 3.11, we have $V(S(ab)) = V(S(a)) \cup V(S(b))$. Also, using Lemma 3.10, we have

$$\begin{aligned} S(ab) &= \mathcal{P}\big(V(S(ab))\big) = \mathcal{P}\big(V(S(a)) \cup V(S(b))\big) \\ &= \mathcal{P}\big(V(S(a))\big) * \mathcal{P}\big(V(S(b))\big) = S(a) * S(b). \end{aligned}$$

Thus $S(ab) = S(a) * S(b)$ for all $a, b \in D^* \setminus U(D)$. □

We now provide improvements to the elasticity results of Section 2.

Theorem 3.13. *Let D be a BFD. For $x \in D^* \setminus U(D)$ a non irreducible element, let $A(x)$ and $B(x)$ be sets of positive integers defined as:*

$$A(x) = \{v + l : S(x) \text{ contains a facet with } v \text{ vertices and } l \text{ loops}\},$$

$$B(x) = \{v + l : G(x) \text{ contains a complete subgraph with } v \text{ vertices and } l \text{ loops}\}.$$

Then

$$\max L(x) \leq \max A(x) \leq \max B(x).$$

Moreover,

$$\rho(x) \leq \tfrac{1}{2} \max A(x) \leq \tfrac{1}{2} \max B(x).$$

Note that $\frac{1}{2}B(x)$ is precisely the bound given in Proposition 2.6.

Proof. Let $\{a_1, \ldots, a_v\}$ be a facet of $S(x)$ with a total of l loops on these vertices. Then $a_1 \cdots a_v \mid x$, and thus $\{a_1, \ldots, a_v\}$ is the vertex set of a complete subgraph of $G(x)$. Loops are preserved when moving from $S(x)$ to $G(x)$. Therefore if $n \in A(x)$, then $n \in B(x)$. Thus $\max A(x) \leq \max B(x)$. If $M = \max L(x)$, then we can write $x = a_1^{n_1} a_2^{n_2} \cdots a_t^{n_t}$, where the a_i are distinct irreducibles and $\sum_{i=1}^{t} n_i = M$. The set $\{a_1, a_2, \ldots, a_t\}$ is a face in $S(x)$ which is contained in some facet of $S(x)$. Also, for each i with $1 \leq i \leq t$, there are $n_i - 1$ loops drawn on the vertex a_i. Thus for any factorization of x of length M we can find a facet of $S(x)$ that contains at least M vertices/loops, and hence $\max L(x) \leq \max X(x)$. Finally, since x is not irreducible, $\min(L(x)) \geq 2$ and thus $\rho(x) \leq \frac{1}{2} \max(A(x)) \leq \frac{1}{2} \max(B(x))$. □

We now consider the sharpness of these bounds by looking at two examples.

Example 3.14. Consider $G(108)$ in Figure 2. The graph $G(108)$ is complete and thus to find the bound on elasticity using Corollary 2.7 we count all vertices and all loops giving us $\rho(108) \leq \frac{1}{2}(6+5) = \frac{11}{2}$. Though not explicitly mentioned in Corollary 2.7, we also see that $\max L(x) \leq 11$. Now consider $S(108)$ in Example 3.4. In order to maximize the total of vertices of and loops in a facet of $S(108)$, we select the facet $\{2, \alpha, \overline{\alpha}, \beta, \overline{\beta}\}$. By Theorem 3.13, $\max L(x) \leq 5$ and $\rho(108) \leq \frac{1}{2}(5+3) = 4$. Here we see that the bound on $\max L(x)$ achieved by Theorem 3.13 is sharp, while the bound on $\max L(x)$ from Corollary 2.7 is not. Since $\mathbb{Z}[\sqrt{-5}]$ is half-factorial, $\rho(108) = 1$ and neither of the bounds on elasticity are sharp.

Example 3.15. Consider $x^{66} \in D = k[x^{10}, x^{12}, x^{18}, x^{33}]$ from Examples 2.4 and 3.5. Since we know precisely the factorizations of x^{66}, we see that $\max L(x^{66}) = 6$, $\min L(x^{66}) = 2$, and $\rho(x^{66}) = 3$. The bounds given by Corollary 2.7 are $\max L(x^{66}) \leq 13$ and $\rho(x^{66}) \leq \frac{13}{2}$. The bounds from Theorem 3.13 are much sharper, with $\max L(x^{66}) \leq 7$ and $\rho(x^{66}) \leq \frac{7}{2}$.

In the special case where x is square-free, we determine in Theorem 3.16 both the minimum and maximum of $L(x)$ as well as the elasticity precisely when using irreducible divisor simplicial complexes, which is a vast improvement over the bound given in Corollary 2.7.

Theorem 3.16. *Let D be a BFD and let $x \in D^* \setminus U(D)$ be square-free. Choose facets β and α such that β has maximal cardinality and α has minimal cardinality among the set of all facets of $S(x)$. Then*

$$\max L(x) = \dim(\beta) + 1, \quad \min L(x) = \dim(\alpha) + 1, \quad \rho(x) = \frac{\dim(\beta) + 1}{\dim(\alpha) + 1}.$$

Proof. By Proposition 3.8, each factorization of x corresponds to a facet of $S(x)$. Therefore

$$\max L(x) = \max\{|\beta| : \beta \in F(S(x))\},$$
$$\min L(x) = \min\{|\alpha| : \alpha \in F(S(x))\}.$$

By definition,

$$\rho(x) = \frac{\max L(x)}{\min L(x)},$$

and thus

$$\rho(x) = \frac{\dim(\beta)+1}{\dim(\alpha)+1}.$$ □

Example 3.17. Let k be a field and let

$$D = k\big[xy^2w, xz, y^2, z^3w, x^2y^2, y^2z^2, z^2w^2\big]$$

be a subring of the polynomial ring $k[x, y, z, w]$. Then the element $x^2y^4z^4w^2$ factors in D only as

$$x^2y^4z^4w^2 = (xy^2w)(xz)(y^2)(z^3w) = (x^2y^2)(y^2z^2)(z^2w^2).$$

Thus $L(x^2y^4z^4w^2) = \{3, 4\}$ and $\rho(x^2y^4z^4w^2) = \frac{4}{3}$. The irreducible divisor graph $G_D(x^2y^4z^4w^2)$, shown at the top of Figure 6, consists of two disjoint components, a 4-clique and a 3-clique, with no looped vertices. The irreducible divisor simplicial complex, shown at the bottom of Figure 6, consists of two disjoint facets, one of dimension 3, the other of dimension 2. Again, no vertices are looped. The bounds

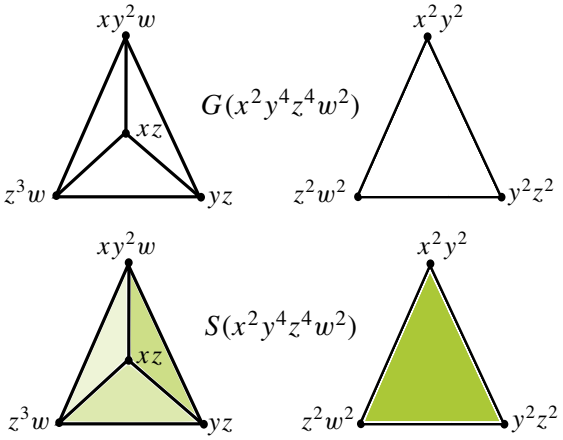


Figure 6. $G(x^2y^4z^4w^2)$ (top) and $S(x^2y^4z^4w^2)$ (bottom) in $D = k[xy^2w, xz, y^2, z^3w, x^2y^2, y^2z^2, z^2w^2]$.

from Corollary 2.7 are

$$\max L(x^2y^4z^4w^2) \leq 4 \quad \text{and} \quad \rho(x^2y^4z^4w^2) \leq 2.$$

The values from Theorem 3.13 are precise, with

$$\max L(x^2y^4z^4w^2) = 4, \quad \min L(x^2y^4z^4w^2) = 3, \quad \rho(x^2y^4z^4w^2) = \tfrac{4}{3}.$$

Acknowledgements

We would like to thank the referee for a careful reading of this manuscript and for several comments which helped to improve this paper.

References

[Axtell and Stickles 2008] M. Axtell and J. Stickles, "Irreducible divisor graphs in commutative rings with zero divisors", *Comm. Algebra* **36**:5 (2008), 1883–1893. MR 2010c:13004 Zbl 1142.13003

[Axtell et al. 2011] M. Axtell, N. Baeth, and J. Stickles, "Irreducible divisor graphs and factorization properties of domains", *Comm. Algebra* **39**:11 (2011), 4148–4162. MR 2012k:13002 Zbl 06067789

[Bachman et al. 2012] D. Bachman, N. Baeth, and C. Edwards, "Irreducible divisor graphs for numerical monoids", *Involve* **5**:4 (2012), 449–462. MR 3069047

[Coykendall and Maney 2007] J. Coykendall and J. Maney, "Irreducible divisor graphs", *Comm. Algebra* **35**:3 (2007), 885–895. MR 2008a:13001 Zbl 1114.13001

[Geroldinger and Halter-Koch 2006] A. Geroldinger and F. Halter-Koch, *Non-unique factorizations: Algebraic, combinatorial and analytic theory*, Pure and Applied Mathematics **278**, Chapman & Hall/CRC, Boca Raton, FL, 2006. MR 2006k:20001 Zbl 1113.11002

[Maney 2008] J. Maney, "Irreducible divisor graphs. II", *Comm. Algebra* **36**:9 (2008), 3496–3513. MR 2009h:13001 Zbl 1153.13300

[Smallwood and Swartz 2008] H. Smallwood and D. Swartz, "Properties of the diameter and girth of the hybrid irreducible divisor graph", Technical Report, Wabash Summer Institute in Mathematics, 2008.

[Smallwood and Swartz 2009] H. Smallwood and D. Swartz, "An investigation of the structure of underlying irreducible divisors", *Amer. J. Undergrad. Res.* **36**:2/3 (2009), 5–12.

Received: 2012-08-06 Revised: 2012-10-12 Accepted: 2012-10-15

baeth@ucmo.edu *Mathematics and Computer Science, University of Central Missouri, W. C. Morris 213, Warrensburg, MO 64093, United States*

jake.hobson13@gmail.com *University of Central Missouri, Warrensburg, MO 64093, United States*

INVOLVE 6:4 (2013)
dx.doi.org/10.2140/involve.2013.6.461

Smallest numbers beginning sequences of 14 and 15 consecutive happy numbers

Daniel E. Lyons

(Communicated by Nigel Boston)

It is well known that there exist arbitrarily long sequences of consecutive happy numbers. In this paper we find the smallest numbers beginning sequences of fourteen and fifteen consecutive happy numbers.

1. Introduction

Guy [1994, Problem E34] defines a happy number in the following way: "If you iterate the process of summing the squares of the decimal digits of a number, then it is easy to see that you either reach the cycle $4 \to 16 \to 37 \to 58 \to 89 \to 145 \to 42 \to 20 \to 4$ or arrive at 1. In the latter case you started from a happy number." Written another way, a happy number N is one for which some iteration of the function $S(N) = \sum_{j=0}^{k} a_j^2$ returns a value of 1, where $\sum_{j=0}^{k} a_j 10^j$ is the decimal expansion of N. According to Guy, the problem was first brought to the attention of the Western mathematical world when Reginald Allenby's daughter returned with it from school in Britain. It is thought to have originated in Russia.

The first pair of consecutive happy numbers is 31, 32. The first example of three consecutive happy numbers is 1880, 1881, 1882. The smallest N beginning a sequence of four and five consecutive happy numbers are 7839 and 44488, respectively. El-Sedy and Siksek [2000] were the first to publish a proof that there exist arbitrarily long sequences of happy numbers, although Lenstra is known to have had an unpublished proof before them. Styer [2010] found the smallest examples of sequences of j consecutive happy numbers, for j from 6 to 13.

In this paper, we will use a period (.) to denote the concatenation operator to group sets of digits together within a large number. For convenience and clarity,

MSC2010: 11A63.

Keywords: happy numbers, consecutive happy numbers, strings of happy numbers, in a row, fourteen consecutive, fifteen consecutive.

we will also write large strings of 9 by their quantity in parenthesis. For example, $615 \cdot 10^{157} + (10^{155} - 1) \cdot 10^2 + 71$ will be written as 615.(155 nines).71.

Define the function $S\left(\sum_{j=0}^{k} a_j 10^j\right) = \sum_{j=0}^{k} a_j^2$ and

$$N_0 = 7888.(1604938271577 \text{ nines}).1.(345696 \text{ nines}).3.$$

2. Fourteen consecutive happy numbers

Theorem 1. $N_0 = 7888.(1604938271577 \text{ nines}).1.(345696 \text{ nines}).3$ *is the smallest N that begins a sequence of fourteen consecutive happy numbers. Note*: N_0 *has* 1604938617279 *digits.*

Because the S function simply sums the squares of the digits of a number, and because addition is commutative, the ordering of the digits has no effect on the function's output. In other words,

Lemma 1. *For every choice of positive integers A, B, and C,*

$$S(A.B.C) = S(B.A.C) = S(A.C.B) = S(A) + S(B) + S(C).$$

Lemma 2. N_0 *begins a sequence of fourteen consecutive happy numbers.*

Proof. Before the carry:

$$\begin{aligned}
N_0 &= 7888.(1604938271577 \text{ nines}).1.(345696 \text{ nines}).3,\\
S(N_0) &= 130000027999364,\\
N_0 + 1 &= 7888.(1604938271577 \text{ nines}).1.(345696 \text{ nines}).4,\\
S(N_0 + 1) &= 130000027999371,\\
N_0 + 2 &= 7888.(1604938271577 \text{ nines}).1.(345696 \text{ nines}).5,\\
S(N_0 + 2) &= 130000027999380,\\
N_0 + 3 &= 7888.(1604938271577 \text{ nines}).1.(345696 \text{ nines}).6,\\
S(N_0 + 3) &= 130000027999391,\\
N_0 + 4 &= 7888.(1604938271577 \text{ nines}).1.(345696 \text{ nines}).7,\\
S(N_0 + 4) &= 130000027999404,\\
N_0 + 5 &= 7888.(1604938271577 \text{ nines}).1.(345696 \text{ nines}).8,\\
S(N_0 + 5) &= 130000027999419,\\
N_0 + 6 &= 7888.(1604938271577 \text{ nines}).1.(345696 \text{ nines}).9,\\
S(N_0 + 6) &= 130000027999436.
\end{aligned}$$

After the carry:

$$\begin{aligned}
N_0 + 7 &= 7888.(1604938271577 \text{ nines}).2.(345696 \text{ zeros}).0,\\
S(N_0 + 7) &= 129999999997982,\\
N_0 + 8 &= 7888.(1604938271577 \text{ nines}).2.(345696 \text{ zeros}).1,\\
S(N_0 + 8) &= 129999999997983,\\
N_0 + 9 &= 7888.(1604938271577 \text{ nines}).2.(345696 \text{ zeros}).2,\\
S(N_0 + 9) &= 129999999997986,\\
N_0 + 10 &= 7888.(1604938271577 \text{ nines}).2.(345696 \text{ zeros}).3,\\
S(N_0 + 10) &= 129999999997991,\\
N_0 + 11 &= 7888.(1604938271577 \text{ nines}).2.(345696 \text{ zeros}).4,\\
S(N_0 + 11) &= 129999999997998,\\
N_0 + 12 &= 7888.(1604938271577 \text{ nines}).2.(345696 \text{ zeros}).5,\\
S(N_0 + 12) &= 129999999998007,\\
N_0 + 13 &= 7888.(1604938271577 \text{ nines}).2.(345696 \text{ zeros}).6,\\
S(N_0 + 13) &= 129999999998018.
\end{aligned}$$

It is not difficult to see that each of these numbers is happy. The iterations of the S function get small rather quickly, and, after at most nine steps, reach 1. □

Lemma 3. *If $N_a < N_0$ is another example of a number beginning a sequence of fourteen consecutive happy numbers, then $S(N_a) < 9^2 \cdot 1604938617279 = 130000027999599$.*

Proof. In order for N_a to be smaller than N_0, it must not contain more digits than N_0. N_0 contains 1604938617279 digits. The largest number containing no more than 1604938617279 digits is $10^{1604938617279} - 1$, or 1604938617279 digits 9, which has an S value of $9^2 \cdot 1604938617279 = 130000027999599$. Therefore, if there were a number $N_a < N_0$ beginning a sequence of fourteen consecutive happy numbers, it would necessarily have $S(N_a) < 130000027999599$. □

We will let N_1 denote any candidate less its final digit. Thus we write $N_a = N_1.x$, where x is the final digit. So, in our case, $N_0 = N_1.3$. Let d be the first (rightmost) non-nine digit of N_1, and let N_2 be the remaining digits of N_1, to the left of d. Thus we have

$$N_1 = N_2.d.(k \text{ nines})$$

for an integer $k \geq 0$.

Lemma 4. $S(N_1 + 1) \leq S(N_1) + 17$.

Proof.
$$N_1 = N_2.d.(k \text{ nines}),$$
$$N_1 + 1 = N_2.(d+1).(k \text{ zeros}),$$
$$S(N_1) = S(N_2) + d^2 + 9^2 k,$$
$$S(N_1 + 1) = S(N_2) + (d+1)^2,$$
$$S(N_1 + 1) - S(N_1) = (d+1)^2 - d^2 - 81k \le 9^2 - 8^2 = 17. \qquad \square$$

Lemma 5. *Let M have four or more digits and let m, f, g, h be integers. Define*

$$M = M_2.f.(m \text{ nines}).g.h,$$

where $m \ge 0$, $0 \le f \le 8$, $0 \le e, g, h \le 9$, and M_2 either is a positive integer or else is possibly vacuous (in which case we define $S(M_2) = 0$). Then

$$S(M + e^2) = S(M_2) + S(f.(m \text{ nines}).g.h + e^2).$$

Proof. Since $e^2 \le 81$, then $g.h + e^2 \le 180$. Now $g.h + e^2 = i.j$ or $g.h + e^2 = 1.i.j$ for some digits i and j. Then we have $M + e^2 = M_2.f.(m \text{ nines}).i.j$ or $M + e^2 = M_2.(f+1).(m \text{ zeros}).i.j$. Now Lemma 1 completes the argument. $\square$

Note that $130000027999599 + 17 = 130000027999616$.

Lemma 6. *If each member of the set $\{M + e^2 \mid e = 2, 3, 4, 5, 6, 7, 8, 9\}$ is happy, then $M > 130000027999616$.*

Proof. Styer [2010], when dealing with fewer than fourteen consecutive happy numbers, did an exhaustive search on all values of M up to the needed bounds for his purposes. In order to reach a bound as high as 130000027999599, we order the digits of M. This makes the search approximately seven million times more efficient.

Write $M = M_2.f.(m \text{ nines}).g.h$ as in Lemma 5. Assume the digits of M_2 are ordered in nondecreasing order. For each m from 0 to 12, we have a separate Maple script that checks every possible M with the digits of M_2 ordered to see if each member of $\{M + e^2 \mid e = 2, 3, 4, 5, 6, 7, 8, 9\}$ is happy. A Maple program shows there are none. (For the relevant Maple worksheets, see [Lyons 2012].) $\square$

Lemma 7. *The final digit x of N_a satisfies $x \ge 3$.*

Proof. We assumed the existence of $N_a < N_0$ that begins a sequence of 14 consecutive happy numbers and we have written $N_a = N_1.x$ where x is a single digit. Suppose $x = 0$, 1, or 2. Then $N_1.e$ is happy with $e = 2, \ldots, 9$. Thus $S(N_1) + e^2$ is happy with $e = 2, \ldots, 9$. By the previous lemma, we have $S(N_1) > 13000002799916$. But $S(N_a) < 13000002799599$ by Lemma 3. Moreover,

$$S(N_1) = S(N_a) - x^2 \le S(N_a) - 4 < 13000002799595.$$

The upper and lower bounds we have for $S(N_1)$ contradict each other, so $x \ge 3$. $\square$

A set of Maple calculations similar to those in Lemma 6 yields the following lemma:

Lemma 8. *If each member of the set* $\{M + e^2 \mid e = 0, 1, 2, 3, 4, 5, 6, 7\}$ *is happy, then* $M > 130000027999616$.

Lemma 9. *The final digit* x *of* N_a *is* $x = 3$.

Proof. We know that $x \geq 3$ by Lemma 8. Suppose $x \geq 4$. Now the numbers $N_a + u = N_1.x + u$ are happy for $u = 0, 1, \ldots, 14$. If $x \geq 4$ these numbers include $(N_1+1).e$ with $e = 0, 1, \ldots, 7$. Therefore $S(N_1+1) > 13000002777616$. However, by Lemmas 3 and 4,

$$S(N_1 + 1) < 13000002799599 + 17 = 13000002799616,$$

giving a contradiction. Therefore $x = 3$. □

Lemma 10. *The value* $M_3 = 129999999997982$ *is the only* $M < 130000027999616$ *such that every member of* $\{M + e^2 \mid e = 0, 1, 2, 3, 4, 5, 6\}$ *is a happy number.*

Proof. Maple calculations similar to Lemma 5 give this single example with digits in nondecreasing order. While any other permutation of the leading 11 digits (the M_2 portion of M_3) will also result in every member of $\{M + e^2 \mid e = 0, 1, 2, 3, 4, 5, 6\}$ being a happy number, these permutations will give us an M value which exceeds our bound. □

Lemma 11. *The value of* $S(N_1)$ *must satisfy*

$$129999999997982 - 17 < S(N_1) < 130000027999599.$$

Lemma 12. *The only* M *with* $129999999997982 - 17 < M < 130000027999599$ *such that every member of* $\{M + e^2 \mid e = 3, 4, 5, 6, 7, 8, 9\}$ *is a happy number is* $M = 130000027999355$.

A Maple search over all the numbers within the bounds listed above returned this single result. Call this value M_1.

We now have the following relationships:

$$S(N_1) = S(N_2) + d^2 + 81k = 1300000027999355 = M_1,$$
$$S(N_0 + 7) = S(N_2) + (d + 1)^2 = 129999999997982 = M_3,$$
$$M_1 - M_3 = 81k - 2d - 1 = 28001373.$$

We look for integers k and d that satisfy this last relationship and find the sole solution $k = 345696$ and $d = 1$.

Now all that is left is to find the smallest N_2 that will satisfy these three equations. With $d = 1$, it reduces to $S(N_2) = 129999999997978$. Using the methods elaborated by Styer [2010], we easily find that the minimal N_2 with $S(N_2) = 129999999997978$

is $N_2 = 7888.(1604938271577 \text{ nines})$. Putting all this together we see that the smallest N beginning a sequence of fourteen consecutive happy numbers is indeed $N_0 = 7888.(1604938271577 \text{ nines}).1.(345696 \text{ nines}).3$.

3. Fifteen consecutive happy numbers

Using the same methods as outlined above, we have confirmed Styer's previous conjecture that the smallest number beginning a sequence of fifteen consecutive happy numbers is $N = 77.(2222222222222220 \text{ nines}).3.(97388).3$.

References

[El-Sedy and Siksek 2000] E. El-Sedy and S. Siksek, "On happy numbers", *Rocky Mountain J. Math.* **30**:2 (2000), 565–570. MR 2002c:11011 Zbl 1052.11008

[Guy 1994] R. K. Guy, *Unsolved problems in number theory*, 2nd ed., Springer, New York, 1994. MR 96e:11002 Zbl 0805.11001

[Lyons 2012] D. Lyons, Maple programs, 2012, http://homepage.villanova.edu/robert.styer/HappyNumbers/happy_numbers.htm.

[Styer 2010] R. Styer, "Smallest examples of strings of consecutive happy numbers", *J. Integer Seq.* **13**:6 (2010), Article 10.6.3, 10. MR 2011f:11007 Zbl 1238.11007

Received: 2012-08-30 Revised: 2012-10-21 Accepted: 2012-10-30

danlyons811@gmail.com *Villanova University, 800 Lancaster Avenue, Villanova, PA 19085, United States*

INVOLVE 6:4 (2013)
dx.doi.org/10.2140/involve.2013.6.467

An orbit Cartan type decomposition of the inertia space of SO($2m$) acting on $\mathbb{R}^{2m}$

Christopher Seaton and John Wells

(Communicated by Michael Dorff)

We study the inertia space of $\mathbb{R}^{2m}$ with the standard action of the special orthogonal group SO($2m$). In particular, we indicate a decomposition of the inertia space that induces the *orbit Cartan type stratification* of the inertia space recently defined by C. Farsi, M. Pflaum, and the first author for an arbitrary smooth G-manifold where G is a compact Lie group.

1. Introduction

Let G be a compact Lie group, let M be a smooth, left G-manifold, and let $X = G\backslash M$ denote the orbit space of M. The *inertia space* ΛX is a topological space given by a subquotient of $G \times M$ under the diagonal G-action, where G acts by conjugation on the first factor. In [Farsi et al. 2012], an explicit Whitney stratification of the inertia space is presented, called the *orbit Cartan type stratification*, giving the inertia space the structure of a differentiable stratified space. This structure coincides with the notion of a stratified space with smooth structure — see [Pflaum 2001] — and simultaneously a differentiable space in the sense of [Navarro González and Sancho de Salas 2003]. In the case that G acts locally freely, so that $G\backslash M$ is an orbifold, the inertia space has played a major role in the study of the geometry of orbifolds; see [Adem et al. 2007], for instance. In general, the inertia space has appeared in connection with equivariant homology theories in noncommutative geometry [Brylinski 1987].

Recall that a *decomposition* of a topological space X is a locally finite partition of X into locally closed, smooth manifolds, called *pieces*, such that the frontier condition is satisfied: if $R \cap \bar{S} \neq \varnothing$ for pieces R and S, then $R \subseteq \bar{S}$. A *stratification* of X is an equivalence class of essentially identical decompositions, defined by assigning to each point $x \in X$ the germ at x of the piece containing x in a decomposition of a neighborhood of x. A decomposition of X *induces* a stratification

MSC2010: 57S15, 58A35.
Keywords: inertia space, stratification, special orthogonal group, Lie group.

if the germ assigned to x by the stratification coincides with the germ at x of the piece of the decomposition containing x. See [Pflaum 2001] for background on decomposed and stratified spaces.

In this note, we determine a decomposition of the inertia space for the standard action of the even special orthogonal group $\mathrm{SO}(2m)$ on $\mathbb{R}^{2m}$ that induces the orbit Cartan type stratification. Our goal is to illustrate the computability of the stratification and to develop a large class of examples through which to better understand its properties.

The outline of this paper is as follows. In Section 2, we recall the definition of the inertia space and the orbit Cartan type stratification, and discuss facts about $\mathrm{SO}(2m)$ that we will need. In Section 3, we define the decomposition and prove that it has the required properties, recalling necessary information about the centralizers of elements of the standard maximal torus in $\mathrm{SO}(2m)$. We prove Theorem 3.2 by verifying the decomposition of the inertia space as well as its relationship to the stratification.

2. Background

In this section, we recall the orbit Cartan type stratification of the inertia space and collect the results we will need in the sequel. We use R_θ to indicate the 2×2 matrix

$$R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

that acts (on the left) on $\mathbb{R}^2$ as a rotation through the angle θ. We say a value of θ is *generic* if θ is not congruent to $0 \bmod 2\pi$ or $\pi \bmod 2\pi$. Additionally, we use $\mathrm{diag}(A_1, \ldots, A_\ell)$ to indicate the matrix in block form with diagonal blocks $A_1, \ldots, A_\ell$ and 0 elsewhere. We let I_n denote the $n\times n$ identity matrix, or simply I when the dimensions are clear from the context, and we let $\langle x \rangle$ denote the span of an element $x \in \mathbb{R}^{2m}$.

The inertia space of a G-manifold and its stratification. We recall the following from [Farsi et al. 2012]. Note that $\mathrm{SO}(2m)$ is connected, and with respect to the standard action of $\mathrm{SO}(2m)$ on $\mathbb{R}^{2m}$, the isotropy group of each point $x \in \mathrm{SO}(2m)$ is connected. As this is our intended application, we specialize to this case for simplicity.

Let G be a compact, connected Lie group, let M be a smooth, left G-manifold, and let $X = G\backslash M$ denote the quotient space. The *loop space* ΛM of the G-manifold M is the set

$$\Lambda M := \{(h, x) \in G \times M \mid hx = x\}.$$

The loop space ΛM is clearly invariant under the action of G on $G \times M$ given by

$$g(h, x) = (ghg^{-1}, gx),$$

and the *inertia space* ΛX is defined to be the quotient of the loop space under this action.

Now, assume the isotropy group of each $x \in M$ is connected, and let $(h, x) \in \Lambda M$. Let $H = G_{(h,x)}$ denote the isotropy group of (h, x) in G, which is given by the centralizer $Z_{G_x}(h)$ of h in the isotropy group G_x of x, and choose a linear slice $V_{(h,x)}$ at (h, x) for the G-action on $G \times M$. By a *slice*, we mean a submanifold $V_{(h,x)}$ of $G \times M$ transversal to the orbit of (h, x) and satisfying these properties:

- $V_{(h,x)}$ is closed in its orbit $GV_{(h,x)}$, which is an open neighborhood of (h, x) in $G \times M$.
- $HV_{(h,x)} = V_{(h,x)}$.
- $gV_{(h,x)} \cap V_{(h,x)} \neq \varnothing$ implies $g \in H$.

A *linear slice* is H-equivariantly diffeomorphic to an H-invariant neighborhood of the origin in the normal space $T_{(h,x)}(G \times M)/T_{(h,x)}G(h, x)$ to the orbit at (h, x), on which H acts linearly. See [Bredon 1972, II, Theorem 4.4] and [Koszul 1953].

As G_x is connected by hypothesis, we have, by [Duistermaat and Kolk 2000, Theorem 3.3.1(i)], that h is contained in the connected component of the identity H° of H. Therefore, we may choose a maximal torus $\mathsf{T}_{(h,x)}$ of H° containing h. We define an equivalence relation $\sim$ on $\mathsf{T}_{(h,x)}$ by declaring that $t_1 \sim t_2$ for $t_1, t_2 \in \mathsf{T}_{(h,x)}$ if there is an open G-invariant neighborhood U of (h, x) such that $U^{t_1} = U^{t_2}$. This is the case if and only if $(GV_{(h,x)})^{t_1} = (GV_{(h,x)})^{t_2}$. We let $\mathsf{T}^*_{(h,x)}$ denote the $\sim$ class of h in $\mathsf{T}_{(h,x)}$.

With this, the stratification of ΛM is given by assigning to (h, x) the germ of the set

$$G\left(V^H_{(h,x)} \cap (\mathsf{T}^*_{(h,x)} \times M)\right), \tag{2-1}$$

and the stratification of ΛX is given by assigning to the orbit $G(h, x)$ the germ of this G-invariant set. It is demonstrated in [Farsi et al. 2012] that ΛM equipped with this stratification has the structure of a differentiable Whitney stratified space, and moreover that ΛX inherits from this G-invariant stratification the structure of a differentiable Whitney stratified space. In particular, the germ at (h, x) of the set defined in (2-1) does not depend on the choice of slice nor on the choice of maximal torus $\mathsf{T}_{(h,x)}$, and the germ at $G(h, x)$ of the corresponding stratification of ΛX does not depend on the choice of representative (h, x) from the orbit $G(h, x)$.

Example 2.1. Consider the case $G = \mathrm{SO}(2)$ with its standard action on $M = \mathbb{R}^2$. It is easy to see that

$$\Lambda\mathbb{R}^2 = \{(I, x) : x \in \mathbb{R}^2 \smallsetminus \{0\}\} \cup \{(h, 0) : h \in \mathrm{SO}(2)\} \subseteq \mathrm{SO}(2) \times \mathbb{R}^2,$$

where $I \in \mathrm{SO}(2)$ denotes the identity matrix. That is, $\Lambda\mathbb{R}^2$ is homeomorphic to $\mathbb{R}^2$ with a circle attached at the origin. The SO(2)-isotropy group of points of the form $(h, 0)$ is SO(2), while all other points have trivial isotropy. In particular, note that the partition of $\Lambda\mathbb{R}^2$ into isotropy types is not a decomposition, as the frontier condition fails at the point $(I, 0)$.

Any invariant neighborhood of a point $(h, 0)$ contains points with nonzero $\mathbb{R}^2$-coordinate. Hence, the maximal torus $\mathsf{T}_{(h,0)} = \mathrm{SO}(2)$ consists of two $\sim$ classes: the identity fixing each point in any SO(2)-invariant neighborhood, and $\mathrm{SO}(2) \smallsetminus \{I\}$, whose elements fix points of the form $(h, 0)$. Clearly, $\mathsf{T}_{(I,x)}$ is trivial for $x \neq 0$. It follows that a decomposition of $\Lambda\mathbb{R}^2$ inducing the orbit Cartan type stratification consists of three pieces:

$$\begin{aligned}\mathcal{P}_1 &= \{(I, x) : x \in \mathbb{R}^2 \smallsetminus \{0\}\},\\ \mathcal{P}_2 &= \{(h, 0) : h \in \mathrm{SO}(2) \smallsetminus \{I\}\},\\ \mathcal{P}_3 &= \{(I, 0)\}.\end{aligned}$$

The SO(2)-action on $\mathcal{P}_1$ is identified with the standard action on $\mathbb{R}^2 \smallsetminus \{0\}$, while the action is trivial on $\mathcal{P}_2$ and $\mathcal{P}_3$. Hence, the quotient space ΛX is homeomorphic to a ray with a circle attached to its endpoint.

***The special orthogonal group* SO(n).** The material in this section is well-known, and can be found in [Tapp 2005, Chapter 9]. See also [Bröcker and tom Dieck 1995, IV Section 3; Humphreys 1978, pages 64–5] for a description of the Weil group of $\mathrm{SO}(n)$.

The *special orthogonal group* $\mathrm{SO}(n)$ is the group of $n \times n$ orthogonal matrices with determinant 1. It is a compact, connected Lie group of dimension $n(n-1)/2$. For an element $k \in \mathrm{SO}(n)$, we let (k) denote the $\mathrm{SO}(n)$-conjugacy class of k.

If $n = 2m$ is even, then the *standard maximal torus* $\mathbb{T}^{\mathrm{st}}_{2m}$ in $\mathrm{SO}(n)$ is an m-dimensional torus given by the set of matrices of the form

$$\mathbb{T}^{\mathrm{st}}_{2m} := \{\mathrm{diag}(R_{\theta_1}, \ldots, R_{\theta_m}) \mid \theta_i \in [0, 2\pi)\}.$$

The center of $\mathrm{SO}(2m)$ is $\{I, -I\}$. The Weil group $N_{\mathrm{SO}(2m)}(\mathbb{T}^{\mathrm{st}}_{2m})/\mathbb{T}^{\mathrm{st}}_{2m}$ is generated by all permutations of the angles $\theta_1, \ldots, \theta_m$ as well as all transformations multiplying two angles by $-1 \bmod 2\pi$.

If $n = 2m + 1$ is odd, then the standard maximal torus $\mathbb{T}^{\mathrm{st}}_{2m+1}$ of $\mathrm{SO}(2m + 1)$ is m-dimensional of the form

$$\mathbb{T}^{\mathrm{st}}_{2m+1} := \{\mathrm{diag}(1, R_{\theta_1}, \ldots, R_{\theta_m}) \mid \theta_i \in [0, 2\pi)\},$$

and the center of $\mathrm{SO}(2m + 1)$ is trivial. The Weil group of $\mathrm{SO}(2m + 1)$ is generated by all permutations of $\theta_1, \ldots, \theta_m$ and all transformations multiplying any angle by $-1 \bmod 2\pi$.

3. The decomposition of $\Lambda\mathbb{R}^{2m}$

Statement of the decomposition. Let $k \in \mathrm{SO}(2m)$. As every element of $\mathrm{SO}(2m)$ is conjugate to an element of $\mathbb{T}^{\mathrm{st}}_{2m}$, we may choose an element $h = \mathrm{diag}(R_{\theta_1}, \ldots, R_{\theta_m})$ of the $\mathrm{SO}(2m)$-conjugacy class of k contained in the standard maximal torus. Using the action of the Weil group, we may choose h with the θ_i listed in the following order. We first list all $\theta_i = 0$, followed by all $\theta_i = \pi$. Then, we list the remaining $\theta_i \neq 0$ in such a way that any angles that agree up to a sign $\bmod 2\pi$ are listed consecutively.

Given such a choice of h, define $(a_0(h), a_\pi(h), \rho(h), s(h))$ as follows. Let $a_0(h)$, with $0 \leq a_0(h) \leq m$, denote the multiplicity of the angle 0; let $a_\pi(h)$, with $0 \leq a_\pi(h) \leq m - a_0(h)$, denote the multiplicity of π; let ρ denote the (possibly empty) partition of $m - a_0(h) - a_\pi(h)$ indicating the number of generic angles that coincide up to a sign for each angle that occurs. Finally, if it is possible by the action of the Weil group to list all angles that coincide up to a sign with the same sign, we let $s(h) = +$; otherwise, we let $s(h) = -$. As elements of $\mathbb{T}^{\mathrm{st}}_{2m}$ are conjugate in $\mathrm{SO}(2m)$ if and only they are conjugate via an element of $N_{\mathrm{SO}(2m)}(\mathbb{T}^{\mathrm{st}}_{2m})$, it is easy to see that $(a_0(h), a_\pi(h), \rho(h), s(h))$ does not depend on the choice of h, and hence is constant on the conjugacy class of k. Hence, we define

$$(a_0(k), a_\pi(k), \rho(k), s(k)) = (a_0(h), a_\pi(h), \rho(h), s(h)).$$

We refer to $T(k) = (a_0(k), a_\pi(k), \rho(k), s(k))$ as the *type* of k, denoted simply $T = (a_0, a_\pi, \rho, s)$ when k is clear from the context.

Example 3.1. We now illustrate the types of elements of $\mathbb{T}^{\mathrm{st}}_{2m}$.

(1) The identity element I has type $(m, 0, \varnothing, +)$, while $-I$ has type $(0, m, \varnothing, +)$.

(2) The element $h = \mathrm{diag}(R_\theta, R_\theta, R_{-\theta}) \in \mathrm{SO}(6)$ with θ generic has type $(0, 0, \{3\}, -)$. Note that any permutation of angles or multiplication of an even number of angles by $-1 \bmod 2\pi$ will result in angles with different signs.

(3) The element $h = \mathrm{diag}(R_\theta, R_{-\theta}, R_\phi, R_\phi, R_{-\phi}) \in \mathrm{SO}(10)$ with θ, ϕ generic has type $(0, 0, \{2, 3\}, +)$ because it is conjugate to $\mathrm{diag}(R_\theta, R_\theta, R_\phi, R_\phi, R_\phi)$. On the other hand, $\mathrm{diag}(R_\theta, R_{-\theta}, R_\phi, R_\phi, R_\phi)$ has type $(0, 0, \{2, 3\}, -)$.

(4) An element $h = \mathrm{diag}(R_0, R_\theta, R_{-\theta}) \in \mathrm{SO}(6)$ with θ generic has type $(1, 0, \{2\}, +)$, because it is conjugate to $\mathrm{diag}(R_0, R_\theta, R_\theta)$ via multiplication of the first and third angles by $-1 \bmod 2\pi$.

For any k of type (a_0, a_π, ρ, s), we have $0 \leq a_0 \leq m$ and $0 \leq a_\pi \leq m - a_0$. If $a_0 > 0$ or $a_\pi > 0$, then $s = +$; this follows from the fact that multiplication by $-1 \bmod 2\pi$ fixes angles 0 and π, as in Example 3.1(4) above. We specify a specific partition $\rho(k)$ by a set with multiplicity, such as $\{1, 1, 2\}$, and adapt ordinary set operations in the obvious way: $\{1\} \cup \{1, 2\} = \{1, 1, 2\}$.

Given $k \in \mathrm{SO}(2m)$ of type (a_0, a_π, ρ, s) with $\rho = \{\rho_1, \dots, \rho_\ell\}$ listed in non-decreasing order, by the above observations, there is an element $h \in (k) \cap \mathbb{T}^{\mathrm{st}}_{2m}$ such that

$$h = \mathrm{diag}(I_{2a_0}, -I_{2a_\pi}, \underbrace{R_{\theta_1}, \dots, R_{\theta_1}}_{\rho_1}, \dots, \underbrace{R_{\theta_\ell}, \dots, R_{\pm\theta_\ell}}_{\rho_\ell}),$$

with each θ_i generic and $\theta_i \neq \pm\theta_j$ for $i \neq j$. In other words, the order and signs of the angles are chosen as above according to the ordering of ρ with at most one sign change, which is required to occur in the last position. We then say that h is in *standard form*. Note that h is unique if and only if $\rho_i \neq \rho_{i+1}$ for each i. We say that $h' \in \mathbb{T}^{\mathrm{st}}_{2m}$ of the same type as h is in the *same standard form as* h if

$$h' = \mathrm{diag}(I_{2a_0}, -I_{2a_\pi}, \underbrace{R_{\phi_1}, \dots, R_{\phi_1}}_{\rho_1}, \dots, \underbrace{R_{\phi_\ell}, \dots, R_{\pm\phi_\ell}}_{\rho_\ell}),$$

with each ϕ_i generic and $\phi_i \neq \pm\phi_j$ for $i \neq j$, so that the repeated angles and the single sign discrepancy, if it occurs, occur in the same positions. Given h and h' in the same standard form, for any $g \in N_{\mathrm{SO}(2m)}(\mathbb{T}^{\mathrm{st}}_{2m})$, we say that ghg^{-1} and $gh'g^{-1}$ are in the *same form*. That is, elements of $\mathbb{T}^{\mathrm{st}}_{2m}$ are in the same form if they are of the same type and can be put into the same standard form by the same element of the Weil group.

With this, we are ready to state our main result, which describes a decomposition of $\Lambda\mathbb{R}^{2m}$ that induces the orbit Cartan type stratification given by (2-1). We state the decomposition for the loop space $\Lambda\mathbb{R}^{2m}$, though a direct consequence is that the quotients of the pieces of the decomposition, which are $\mathrm{SO}(2m)$-invariant and consist of points of the same isotropy type, define a decomposition of the inertia space $\Lambda(\mathrm{SO}(2m)\backslash\mathbb{R}^{2m})$ that induces the orbit Cartan type stratification.

Theorem 3.2. *For each type $T = (a_0, a_\pi, \rho, s)$, let*

$$P_{T,0} = \{(h, 0) \in \mathrm{SO}(2m) \times \mathbb{R}^{2m} : h \text{ has type } T\},$$

and let

$$P_{T,1} = \{(h, x) \in \mathrm{SO}(2m) \times (\mathbb{R}^{2m} \smallsetminus \{0\}) : hx = x, h \text{ has type } T\}.$$

Then a decomposition of $\Lambda\mathbb{R}^{2m}$ inducing the orbit Cartan type stratification described by (2-1) *is given by the following pieces*:

I. $P_{T,0}$ *for each type $T = (a_0, a_\pi, \rho, s)$ such that $a_0 > 0$, $a_\pi > 1$, $s = -$, or $a_\pi = 0$ and $1 \notin \rho$,*

II. $P_{(0,1,\rho,+),0} \cup P_{(0,0,\{1\}\cup\rho,+),0}$ *for each partition ρ of $m-1$,*

III. $P_{T,1}$ *for each type $T = (a_0, a_\pi, \rho, +)$ such that $a_0 > 0$.*

Note that an element h fixes a nonzero element of $\mathbb{R}^{2m}$ if and only if $a_0(h) > 0$, and recall that $s(h) = +$ whenever $a_0(h) > 0$ or $a_\pi(h) > 0$.

***Centralizers in* SO(2m).** Let $h \in \mathbb{T}^{\mathrm{st}}_{2m} \leq \mathrm{SO}(2m)$ be in standard form. Then the centralizer of h in $\mathrm{SO}(2m)$ is determined by the form of h. Specifically, let

$$h = \mathrm{diag}(I_{2a_0}, -I_{2a_\pi}, \underbrace{R_{\theta_1}, \ldots, R_{\theta_1}}_{\rho_1}, \ldots, \underbrace{R_{\theta_\ell}, \ldots, R_{\pm\theta_\ell}}_{\rho_\ell}),$$

with each θ_i generic and $\theta_i \neq \pm\theta_j$ for $i \neq j$, where we may have $a_0 = 0$ or $a_\pi = 0$. The centralizer of h in $\mathrm{SO}(2m)$ is a set of matrices in blocks given by

$$\mathrm{diag}(A, B, C_1, \ldots, C_\ell),$$

where $A \in \mathrm{O}(2a_0)$, $B \in \mathrm{O}(2a_\pi)$, and each $C_i \in \mathrm{O}(2\rho_i)$. Note that in general, $Z_{\mathrm{SO}(2m)}(h)$ contains $\mathbb{T}^{\mathrm{st}}_{2m}$. We first discuss the matrices C_i.

By direct computation, it is easy to see that the only 2×2 matrices that commute with R_θ for θ generic are given by

$$\begin{bmatrix} c_1 & -c_2 \\ c_2 & c_1 \end{bmatrix},$$

i.e., a scalar multiple of a rotation matrix. The only $2\rho_i \times 2\rho_i$ matrices C_i that commute with $\mathrm{diag}(R_\theta, \ldots, R_\theta)$, where θ is generic and R_θ occurs ρ_i times, are matrices whose 2×2 blocks are scalar multiples of rotation matrices as above. Similarly, the only $2\rho_i \times 2\rho_i$ matrices that commute with $\mathrm{diag}(R_\theta, \ldots, R_\theta, R_{-\theta})$, where θ is generic and R_θ occurs $\rho_i - 1$ times, are matrices whose 2×2 block are as above except for the blocks in the last two rows and columns, excluding the lower-right 2×2 block, which are given by

$$\begin{bmatrix} c_1 & c_2 \\ c_2 & -c_1 \end{bmatrix}.$$

In particular, as these computations require only that θ is generic, all elements of the same form have the same centralizer.

If $a_\pi = 0$, then the set of elements of the same form as h is an open, dense subset of a torus of $\mathrm{SO}(2m)$ of dimension ℓ, and $Z_{\mathrm{SO}(2m)}(h)$ coincides with the centralizer of this torus. In particular, $Z_{\mathrm{SO}(2m)}(h)$ is connected by [Duistermaat and Kolk 2000, Theorem 3.3.1]. Then as the determinant of each block is a continuous function from $Z_{\mathrm{SO}(2m)}(h)$ to $\{\pm1\}$, it must be that each C_i has determinant 1. It follows that A must have determinant 1, and hence A can be any element of $\mathrm{SO}(2a_0)$. Note that we may also conclude for arbitrary h that each $C_i \in \mathrm{SO}(2\rho_i)$.

If $a_\pi \neq 0$ and $a_0 \neq 0$, then $A \in \mathrm{O}(2a_0)$ and $B \in \mathrm{O}(2a_\pi)$ can be any elements with the same determinant ±1, and the centralizer of h has two connected components.

If $a_\pi \neq 0$ and $a_0 = 0$, then as the determinant of each C_i is 1, B must also have determinant 1 and can be any element of $\mathrm{SO}(2a_\pi)$.

The reader is cautioned that it is possible for elements of different types to have identical centralizers. For instance, for θ_1 and θ_2 generic, $\theta_1 \neq \pm\theta_2$, the centralizers of the elements $\mathrm{diag}(R_0, R_{\theta_2})$, $\mathrm{diag}(R_\pi, R_{\theta_2})$, and $\mathrm{diag}(R_{\theta_1}, R_{\theta_2})$ coincide and are equal to the standard maximal torus $\mathrm{SO}(2) \times \mathrm{SO}(2) \leq \mathrm{SO}(4)$, though these elements are in standard from of type $(1, 0, \{1\}, +)$, $(0, 1, \{1\}, +)$, and $(0, 0, \{1, 1\}, +)$, respectively. More generally, if ρ is any partition of $m-1$, then elements of type $(1, 0, \rho, +)$, $(0, 1, \rho, +)$, and $(0, 0, \{1\} \cup \rho, +)$ in standard form have the same centralizer, as in either case, the first 2×2-block is forced to be an element of $\mathrm{SO}(2)$. If h is in standard form and of either of these types, then any element with the same centralizer as h is also in standard form.

However, if $a_0(h) = a_0(h') > 0$ for elements $h, h' \in \mathbb{T}^{\mathrm{st}}_{2m}$ such that h is in standard form and both h and h' have centralizer H, then h and h' are in the same standard form. In particular, if H is connected, then $a_\pi(h) = a_\pi(h') = 0$, and if H is not connected, then the size of the second block in elements of H determines that $a_\pi(h) = a_\pi(h') > 0$. The size and structure of the later blocks in elements of H determine the values of ρ and $s = +$ for both h and h' as well as their form. Similarly, if $a_0(h) = a_0(h') = 0$ and $a_\pi(h) > 1$ with h in standard form, then the size and structure of the blocks in the centralizer again determine the form of h and hence h'.

Finally, suppose $a_0(h) = a_\pi(h) = 0$ with h in standard form. If $1 \notin \rho(h)$, then the size and structure of the blocks of H determine the form of h and hence h' so that h' is in the same standard form as h. Otherwise, h has type $(0, 0, \{1\} \cup \rho, s)$ for a partition ρ of $m-1$ and is in standard form

$$h = \mathrm{diag}(R_{\theta_1}, \underbrace{R_{\theta_2}, \ldots, R_{\theta_2}}_{\rho_1}, \ldots, \underbrace{R_{\theta_\ell}, \ldots, R_{\pm\theta_\ell}}_{\rho_{\ell-1}})$$

so that $\theta_1, \ldots, \theta_\ell$ are generic. If $a_0(h') = 0$, then either h' is in the same standard form as h or

$$h' = \mathrm{diag}(R_\pi, \underbrace{R_{\phi_1}, \ldots, R_{\phi_1}}_{\rho_1}, \ldots, \underbrace{R_{\phi_\ell}, \ldots, R_{\pm\phi_\ell}}_{\rho_{\ell-1}}),$$

which has type $(0, 1, \rho, +)$ and is not in standard form if $s(h) = -$.

We now summarize these observations.

Proposition 3.3. *Elements of $\mathbb{T}^{\mathrm{st}}_{2m}$ in the same standard form have the same centralizer, and elements of $\mathrm{SO}(2m)$ of the same type have conjugate centralizers. Conversely, if $h, h' \in \mathbb{T}^{\mathrm{st}}_{2m}$ with h in standard form and $Z_{\mathrm{SO}(2m)}(h) = Z_{\mathrm{SO}(2m)}(h')$, then*:

- *If $a_0(h) = a_0(h') > 0$, then h and h' are in the same standard form.*

- *If* $a_0(h) = a_0(h') = 0$ *and* $a_\pi(h) > 1$, *then* h *and* h' *are in the same standard form.*
- *If* $a_0(h) = a_\pi(h) = 0$ *and* $1 \notin \rho(h)$, *then* h *and* h' *are in the same standard form.*
- *If* $a_0(h) = a_0(h') = 0$ *and* h *has type* $(0, 1, \rho, +)$ *or* $(0, 0, \{1\} \cup \rho, +)$, *then* h' *is in standard form and has type* $(0, 1, \rho, +)$ *or* $(0, 0, \{1\} \cup \rho, +)$.
- *If* $a_0(h) = a_0(h') = 0$ *and* h *has type* $(0, 0, \{1\} \cup \rho, -)$, *then either* h' *is in standard form of type* $(0, 0, \{1\} \cup \rho, -)$ *or* h' *is of type* $(0, 1, \rho, +)$ *and is not in standard form.*

Note that Proposition 3.3 does not exhaust all cases but considers those that we will need below.

Proof of Theorem 3.2. In this section, we demonstrate that the partition defined in Theorem 3.2 is indeed a decomposition that induces the orbit Cartan type stratification. First, we establish the following.

Lemma 3.4. *Let* $h \in \mathbb{T}^{\mathrm{st}}_{2m}$ *be in standard form. Then there is a neighborhood* U *of* h *in* $\mathbb{T}^{\mathrm{st}}_{2m}$ *small enough so that every* $h' \in U$ *of the same type as* h *is in the same standard form as* h. *If* h *has type* $(0, 0, \{1\} \cup \rho, -)$, *then we may choose* U *so that it contains no elements* h' *such that* $a_\pi(h') > 0$.

Proof. Let $h = \mathrm{diag}(R_{\theta_1}, \ldots, R_{\theta_m})$, where angles need not be distinct or generic. Choose $\epsilon > 0$ such that $(\theta_i - \epsilon, \theta_i + \epsilon)$ contains 0 (respectively π) if and only if $\theta_i = 0$ (respectively π), and, for $i \neq j$, the intersection $(\theta_i - \epsilon, \theta_i + \epsilon) \cap (\pm\theta_j - \epsilon, \pm\theta_j + \epsilon)$ is nonempty if and only if $\theta_i = \pm\theta_j$. Then for any $h' = \mathrm{diag}(\phi_i, \ldots, \phi_m)$ such that $|\phi_i - \theta_i| < \epsilon$ for each i, h' is of the same type as h if and only if it is in the same standard form. Moreover, if h has type $(0, 0, \{1\} \cup \rho, -)$, then as $\theta_i \neq \pi$ for each i, U contains no elements of type $(0, a_\pi, \sigma, +)$ for $a_\pi > 0$ and any partition σ. □

Lemma 3.5. *Let* $h \in \mathbb{T}^{\mathrm{st}}_{2m}$ *be an element of the maximal torus of* $\mathrm{SO}(2m)$.

(i) *A linear slice* $V_{(h,0)}$ *for the diagonal* $\mathrm{SO}(2m)$*-action on* $\mathrm{SO}(2m) \times \mathbb{R}^{2m}$ *at* $(h, 0)$ *can be chosen such that* $V_{(h,0)}$ *contains* $U_h \times U_0$, *where* U_h *is a neighborhood of* h *in* $\mathbb{T}^{\mathrm{st}}_{2m}$ *and* U_0 *is a neighborhood of* 0 *in* $\mathbb{R}^{2m}$.

(ii) *If* $0 \neq x \in \mathbb{R}^{2m}$ *such that* $hx = x$, *then a linear slice* $V_{(h,x)}$ *for the diagonal* $\mathrm{SO}(2m)$*-action on* $\mathrm{SO}(2m) \times \mathbb{R}^{2m}$ *at* (h, x) *can be chosen such that* $V_{(h,x)}$ *contains* $U_h \times U_x$ *where* U_h *is a neighborhood of* h *in* $\mathbb{T}^{\mathrm{st}}_{2m}$ *and* U_x *is a connected neighborhood of* x *in the span* $\langle x \rangle$ *of* x *in* $\mathbb{R}^{2m}$.

Proof. Fix the standard ($\mathrm{SO}(2m)$-invariant) Riemannian metric on $\mathbb{R}^{2m}$, choose a bi-invariant metric on $\mathrm{SO}(2m)$, and let $\mathrm{SO}(2m) \times \mathbb{R}^{2m}$ carry the product metric. Recall that (h) denotes the $\mathrm{SO}(2m)$-conjugacy class of h. By [Duistermaat and Kolk

2000, Proposition 3.1.1], the only slice at h for the $SO(2m)$-action on $SO(2m)$ by conjugation is given by a neighborhood S_h of h in the centralizer $Z_{SO(2m)}(h)$, where the linear structure is inherited from the Lie algebra $\mathfrak{z}_h$ of $Z_{SO(2m)}(h)$ via a logarithmic chart. Because the orthogonal complement of $T_h(h)$ in $T_h\mathrm{SO}(2m)$ with respect to the metric is mapped to a slice by the exponential map (see [Duistermaat and Kolk 2000, Theorem 2.3.3]), it follows that $T_h S_h = T_h Z_{SO(2m)}(h)$ is the orthogonal complement of $T_h(h)$ in $T_h\mathrm{SO}(2m)$.

As $SO(2m)(h,0) = (h) \times \{0\} \subset \mathrm{SO}(2m) \times \mathbb{R}^{2m}$, using the isometry

$$T_{(h,0)}(\mathrm{SO}(2m) \times \mathbb{R}^{2m}) \to T_h\mathrm{SO}(2m) \oplus T_0\mathbb{R}^{2m},$$

we have that

$$\begin{aligned} T_{(h,0)}(Z_{SO(2m)}(h) \times \mathbb{R}^{2m}) &\cong T_h Z_{SO(2m)}(h) \oplus T_0\mathbb{R}^{2m} \\ &\cong \big(T_{(h,0)}\mathrm{SO}(2m)(h,0)\big)^{\perp}. \end{aligned}$$

Hence, a slice for the $\mathrm{SO}(2m)$-action on $\mathrm{SO}(2m) \times \mathbb{R}^{2m}$ may be chosen to be a suitably small neighborhood of $(h,0)$ in $Z_{SO(2m)}(h) \times \mathbb{R}^{2m}$. Clearly $\mathbb{T}^{\mathrm{st}}_{2m} \leq Z_{SO(2m)}(h)$, proving (i).

To prove (ii), note that the orbit $\mathrm{SO}(2m)x$ of x is given by the sphere of radius $\|x\|$, so that in $T_x\mathbb{R}^{2m}$, $(T_x\mathrm{SO}(2m)x)^{\perp} = T_x\langle x\rangle$. Then as

$$T_{(h,x)}\mathrm{SO}(2m)(h,x) \subseteq T_h(h) \oplus T_x\mathrm{SO}(2m)x,$$

we have

$$\begin{aligned} T_h Z_{SO(2m)}(h) \oplus T_x\langle x\rangle &= (T_h(h))^{\perp} \oplus (T_x\mathrm{SO}(2m)x)^{\perp} \\ &\subseteq (T_h(h) \times T_x\mathrm{SO}(2m)x)^{\perp} \\ &\subseteq (T_{(h,x)}\mathrm{SO}(2m)(h,x))^{\perp}. \end{aligned}$$

It follows that we may choose a slice $V_{(h,x)}$ at (h,x) such that

$$T_{(h,x)}V_{(h,x)} = (T_{(h,x)}\mathrm{SO}(2m)(h,x))^{\perp},$$

and hence an open neighborhood of (h,x) in $Z_{SO(2m)}(h) \times \langle x\rangle$ is contained in $V_{(h,x)}$. □

Proof of Theorem 3.2. Given an arbitrary element $(k,x) \in \Lambda\mathbb{R}^{2m}$, as k is conjugate to an element of $\mathbb{T}^{\mathrm{st}}_{2m}$, the type of k is defined. Moreover, as the type is conjugation invariant, it is well defined, so that the pieces defined in I, II, and III clearly form a partition of $\Lambda\mathbb{R}^{2m}$. Moreover, as the number of types is finite, the partition is finite and hence trivially locally finite.

For each element (k,y) of a piece P, we now demonstrate that for some (h,x) in the orbit of (k,y) and appropriate choices of slice and maximal torus,

there is an open, $\mathrm{SO}(2m)$-invariant neighborhood of (h, x) within which the set $P \cap \mathrm{SO}(2m)V_{(h,x)}$ coincides with the set defined in (2-1). This implies that the decomposition induces the orbit Cartan type stratification. Moreover, as the germs defining the stratification are germs of locally closed, smooth manifolds, it follows that each piece P is a locally closed, smooth submanifold of $\mathrm{SO}(2m) \times \mathbb{R}^{2m}$. With this, we will need only show that the pieces satisfy the frontier condition.

I. Suppose $(k, 0)$ is of type $T = (a_0, a_\pi, \rho, s)$ with $a_0 > 0$, $a_\pi > 1$, $s = -$, or $a_\pi = 0$ and $1 \notin \rho$. Choose an element $h \in (k) \cap \mathbb{T}^{\mathrm{st}}_{2m}$ in standard form and a slice $V_{(h,0)}$ at $(h, 0)$ for the $\mathrm{SO}(2m)$-action on $\mathrm{SO}(2m) \times \mathbb{R}^{2m}$ with $U_h \times U_0 \subseteq V_{(h,0)}$ as in Lemma 3.5. Applying Lemma 3.4 and shrinking $V_{(h,0)}$ if necessary, we assume that if $(h', x) \in V_{(h,0)}$ with $h' \in \mathbb{T}^{\mathrm{st}}_{2m}$ of the same type as h, then h' is in the same standard form as h. Moreover, if h has type $(0, 0, \{1\} \cup \rho, -)$, we assume that $V_{(h,0)}$ contains no elements of the form (h', x) such that $a_\pi(h') > 0$. Let $H = \mathrm{SO}(2m)_{(h,0)} = Z_{\mathrm{SO}(2m)}(h)$, and define the set

$$Q_{(h,0)} := V^H_{(h,0)} \cap \left((\mathbb{T}^{\mathrm{st}}_{2m})^*_{(h,0)} \times \mathbb{R}^{2m}\right).$$

That is, the $\mathrm{SO}(2m)$-saturation $\mathrm{SO}(2m)Q_{(h,0)}$ is the set that defines the germ of the stratum containing $(h, 0)$ in (2-1). Note that as H contains $\mathbb{T}^{\mathrm{st}}_{2m}$, which only fixes the origin in $\mathbb{R}^{2m}$, any element of $V^H_{(h,0)}$ is of the form $(h', 0)$ for $h' \in \mathrm{SO}(2m)$. Moreover, as $h \in H$, it must be that for any $(h', 0) \in V^H_{(h,0)}$, the element h' commutes with h.

Let $(h', 0) \in Q_{(h,0)}$ be arbitrary. Then $h' \in (\mathbb{T}^{\mathrm{st}}_{2m})^*_{(h,0)}$, implying that the h and h' fix the same subset of $\mathrm{SO}(2m)V_{(h,0)}$. In particular, as $\{h\} \times U_0 \subseteq V_{(h,0)}$, with U_0 a neighborhood of the origin in $\mathbb{R}^{2m}$, and as h' commutes with h, it follows that $(\mathbb{R}^{2m})^h = (\mathbb{R}^{2m})^{h'}$, so $a_0(h) = a_0(h')$. Additionally, by the definition of slice, every point in $V_{(h,0)}$ has isotropy group contained in H, so $V^H_{(h,0)}$ consists only of points with isotropy group equal to H. Hence $Z_{\mathrm{SO}(2m)}(h) = Z_{\mathrm{SO}(2m)}(h')$, so by Proposition 3.3 and the choice of slice, h and h' are in the same standard form. It follows that the orbit of any element of $Q_{(h,0)}$ is contained in $P_{T,0}$ and hence $\mathrm{SO}(2m)Q_{(h,0)} \subseteq P_{T,0}$.

Conversely, if $(k', 0) \in P_{T,0} \cap \mathrm{SO}(2m)V_{(h,0)}$ so that k' is of the same type as h, then by the choice of $V_{(h,0)}$, there is an $(h', 0) \in V_{(h,0)} \cap \mathrm{SO}(2m)(k', 0)$ such that h' is in the same standard form as h. Then h and h' have the same centralizer by Proposition 3.3 so that $(h', 0) \in V^H_{(h,0)}$. Moreover, because $Z_{SO(2m)}(h) = Z_{SO(2m)}(h')$ and the angle 0 occurs in the same positions in both, h and h' fix the same elements of $\mathrm{SO}(2m) \times \mathbb{R}^{2m}$ so that clearly $(\mathrm{SO}(2m)V_{(h,0)})^h = (\mathrm{SO}(2m)V_{(h,0)})^{h'}$. Hence $(h', 0) \in Q_{(h,0)}$. Therefore, we have that $\mathrm{SO}(2m)Q_{(h,0)} = P_{T,0} \cap \mathrm{SO}(2m)V_{(h,0)}$, so that $\mathrm{SO}(2m)Q_{(h,0)}$ and $P_{T,0}$ define the same germ at $(h, 0)$.

II. The argument in this case is similar to I above. Choosing a representative $(h, 0)$ of the orbit of an arbitrary point with $h \in \mathbb{T}^{\mathrm{st}}_{2m}$ in standard form, for any $h' \in (\mathbb{T}^{\mathrm{st}}_{2m})^*_{(h,0)}$, as h and h' have the same fixed point set in $\mathbb{R}^{2m}$, $a_0(h) = 0$ implies $a_0(h') = 0$. In this case, however, while elements $h, h' \in \mathbb{T}^{\mathrm{st}}_{2m}$ of the same type have the same centralizer, the centralizers do not distinguish between group elements in standard form of type $(0, 1, \rho, +)$ and $(0, 0, \{1\} \cup \rho, +)$ by Proposition 3.3. Moreover, any neighborhood of an element in standard form of type $(0, 1, \rho, +)$ clearly contains elements in standard form of type $(0, 0, \{1\} \cup \rho, +)$. As the fixed-point sets of such elements in $\mathrm{SO}(2m) \times \mathbb{R}^{2m}$ coincide, the argument is identical to that of I combining these two types.

III. Let $(k, x) \in \Lambda\mathbb{R}^{2m}$ and let T be the type of k. As the $\mathrm{SO}(2m)$-action on $\mathbb{R}^{2m}$ is transitive on spheres about the origin, we may assume that x has coordinates $(\|x\|, 0, \ldots, 0)$, and hence $\mathrm{SO}(2m)_x = \{\mathrm{diag}(1, A) : A \in \mathrm{SO}(2m-1)\} \cong \mathrm{SO}(2m-1)$. As any element of $\mathrm{SO}(2m)_x$ is conjugate to an element of the standard maximal torus $\mathbb{T}^{\mathrm{st}}_{2m-1}$ via an element of $\mathrm{SO}(2m)_x$, we may choose an element (h, x) in the orbit $\mathrm{SO}(2m)_x(k, x)$ such that $h \in \mathbb{T}^{\mathrm{st}}_{2m-1}$ is in standard form. Note that as h fixes x, we have $a_0(h) > 0$.

Choose a slice $V_{(h,x)}$ at (h, x) that contains $U_h \times U_x$ as in Lemma 3.5, and shrink $V_{(h,x)}$ if necessary so that $V_{(h,x)} \cap (\mathrm{SO}(2m) \times -U_x) = \varnothing$. We again assume by Lemma 3.4 and shrinking $V_{(h,x)}$ that for any $(h', y) \in V_{(h,x)}$ such that $h' \in \mathbb{T}^{\mathrm{st}}_{2m}$ has the same type as h, h' must also have the same form.

It will be convenient to restrict to a smaller open neighborhood of (h, x) in $\mathrm{SO}(2m) \times \mathbb{R}^{2m}$. To do so, recall that the Weil group $N_{\mathrm{SO}(2m)}(\mathbb{T}^{\mathrm{st}}_{2m})/\mathbb{T}^{\mathrm{st}}_{2m}$ is finite. Hence, by [tom Dieck 1987, Proposition 3.23], we may shrink U_h to assume that for $g \in N_{\mathrm{SO}(2m)}(\mathbb{T}^{\mathrm{st}}_{2m})$, we have $gU_h = U_h$ if $g \in Z_{SO(2m)}(h)$ and $U_h \cap gU_h = \varnothing$ otherwise. Moreover, letting $\mathrm{SO}(2m)_*$ denote the set of conjugacy classes in $\mathrm{SO}(2m)$ equipped with its natural quotient topology, we may assume that the quotient of U_h by $N_{\mathrm{SO}(2m)}(\mathbb{T}^{\mathrm{st}}_{2m})/Z_{SO(2m)}(h)$ is homeomorphic to an open subset of $\mathrm{SO}(2m)_*$ containing (h). In particular, as the quotient map $\mathrm{SO}(2m) \to \mathrm{SO}(2m)_*$ is continuous, $\mathrm{SO}(2m)U_h$ is open in $\mathrm{SO}(2m)$. Let $W = (\mathrm{SO}(2m)U_h) \times (\mathrm{SO}(2m)U_x)$, and then as $\mathrm{SO}(2m)U_x = \{z \in \mathbb{R}^{2m} : \epsilon_1 < \|z\| < \epsilon_2\}$ for some $0 < \epsilon_1 < \epsilon_2$, $\mathrm{SO}(2m)U_x$ is open in $\mathbb{R}^{2m}$. Hence W is an open, $\mathrm{SO}(2m)$-invariant neighborhood of (h, x) in $\mathrm{SO}(2m) \times \mathbb{R}^{2m}$. Finally, we further shrink $V_{(h,x)}$ if necessary to assume that it does not intersect $gU_h \times \mathbb{R}^{2m}$ for any of the finite translates of U_h by $g \in N_{\mathrm{SO}(2m)}(\mathbb{T}^{\mathrm{st}}_{2m})$ such that $g \notin Z_{\mathrm{SO}(2m)}(h)$. We will show that the piece $P_{T,1}$ coincides with the set given in (2-1) when intersected with W.

Let $H = \mathrm{SO}(2m)_{(h,x)} = Z_{\mathrm{SO}(2m-1)}(h)$ so that H consists of those elements of $Z_{\mathrm{SO}(2m)}(h)$ whose first row and column are that of the identity. Define the set

$$Q_{(h,x)} := V^H_{(h,x)} \cap \left((\mathbb{T}^{\mathrm{st}}_{2m-1})^*_{(h,x)} \times \mathbb{R}^{2m}\right) \cap W.$$

Fix $(h', y) \in Q_{(h,x)}$ so that $h' \in (\mathbb{T}^{\mathrm{st}}_{2m-1})^*_{(h,x)}$. Therefore, as any neighborhood of (h, x) contains points (h, y') where any coordinate of y' except the first may be chosen to be zero or nonzero, and as $h \in H$ so that h and h' commute, we have that h and h' must have 0 occur as an angle with the same multiplicity in the same positions. Therefore, $a_0(h') = a_0(h) > 0$. Note that $(h', y) \in V^H_{(h,x)}$ so that $\mathrm{SO}(2m)_{(h',y)} = Z_{\mathrm{SO}(2m)_y}(h') = H$. In particular, as $(h', y) \in W \cap V_{(h,x)}$ and $h' \in \mathbb{T}^{\mathrm{st}}_{2m-1} \leq \mathbb{T}^{\mathrm{st}}_{2m}$, we may conclude h' is in U_h. We consider two cases:

If $a_0(h) > 1$ or $a_\pi(h) > 0$, then H contains $\mathbb{T}^{\mathrm{st}}_{2m-1}$ as well as the element $g = \mathrm{diag}(1, -1, -1, 1, I_{2m-4})$. The fixed point set in $\mathbb{R}^{2m}$ of the group generated by g and $\mathbb{T}^{\mathrm{st}}_{2m-1}$ is $\langle x \rangle$, so that $y \in \langle x \rangle$. Then as $a_0(h') = a_0(h)$, connectedness of H determines whether $a_\pi(h)$, and hence $a_\pi(h')$, vanish. If not, the second block of elements of H indicates that $a_\pi(h') = a_\pi(h)$, and the following blocks further indicate that h and h' have the same type. Therefore, $(h', y) \in P_{T,1}$.

If $a_0(h) = 1$ and $a_\pi(h) = 0$, then every element of H, and in particular h', is given by $\mathrm{diag}(I_2, D)$ for a $(2m-2)\times(2m-2)$ matrix D. As H contains $\mathbb{T}^{\mathrm{st}}_{2m-1}$ which then must fix y, it follows that $y = (a, b, 0, \ldots, 0)$ for some $a, b \in \mathbb{R}$. Then there is a $\bar{g} = \mathrm{diag}(R_\theta, I_{2m-2})$ such that $\bar{g}y = (\|y\|, 0, \ldots, 0)$. Moreover, as $h' = \mathrm{diag}(I_2, D)$ for some D, we have $\bar{g}h'\bar{g}^{-1} = h'$, and $\bar{g}(h', y) = (h', (\|y\|, 0, \ldots, 0))$. However, as $y \in \mathrm{SO}(2m)U_x$, and $\bar{g}y \in \langle x \rangle$ has positive first coordinate, it follows that $\bar{g}y \in U_x$. Moreover, as $h' \in U_h$, we have $\bar{g}(h', y) \in U_h \times U_x \subseteq V_{(h,x)}$, so that as $(h', y) \in V_{(h,x)}$, it follows from the definition of slice that $\bar{g} \in H$. Then as elements of H fix y, we have that $y = (\|y\|, 0, \ldots, 0)$ to begin with.

With this, the element $g = \mathrm{diag}(1, -1, -1, 1, I_{2m-4})$ fixes y and hence, as it is not an element of H, cannot commute with h'. It follows that $a_\pi(h') = 0$, and then the structure of blocks of elements of H imply that h and h' have the same type. We again have $(h', y) \in P_{T,1}$, and hence $\mathrm{SO}(2m)Q_{(h,x)} \subseteq P_{T,1}$, since $P_{T,1}$ is $\mathrm{SO}(2m)$-invariant.

Conversely, if $(k, y) \in P_{(T,1)} \cap \mathrm{SO}(2m)V_{(h,x)} \cap W$, then (k, y) is in the orbit of an element $(h', y') \in V_{(h,x)}$. Then as h' has the same type as h, it must have the same standard form as h. This implies that h' and h have the same centralizer, and moreover that $a_0(h') = a_0(h) > 0$. Noting that h' fixes y', and hence that y' has nonzero coordinates only in the first $2a_0(h)$ positions, there is an element $\bar{g} = \mathrm{diag}(D, I_{2(m-a_0(h))}) \leq \mathrm{SO}(2m)$ for some $D \in \mathrm{SO}(2a_0(h))$ such that $\bar{g}y' = (\|y'\|, 0, \ldots, 0)$. As $(h', y') \in W \cap V_{(h,x)}$ and $h' \in \mathbb{T}^{\mathrm{st}}_{2m}$, $h' \in U_h$. Hence, as $\bar{g}$ commutes with h', $\bar{g}(h', y') = (h', (\|y'\|, 0, \ldots, 0)) \in U_h \times U_x \subseteq V_{(h,x)}$. That h and h' have the same centralizer and $\bar{g}y' \in \langle x \rangle$ implies $\bar{g}(h', y') \in V^H_{(h,x)}$. In addition, that h and h' have the same type implies $h' \in (\mathbb{T}^{\mathrm{st}}_{2m-1})^*_{(h,x)}$. It follows that $\bar{g}(h', y') \in Q_{(h,x)} \cap W$ so that $(k, y) \in \mathrm{SO}(2m)Q_{(h,x)} \cap W$, completing the proof that $\mathrm{SO}(2m)Q_{(h,x)} \cap W = P_{(T,1)} \cap \mathrm{SO}(2m)V_{(h,x)} \cap W$.

The frontier condition. To show that the pieces defined in Theorem 3.2 satisfy the frontier condition, we first claim that $k \in \mathrm{SO}(2m)$ is in the closure in $\mathrm{SO}(2m)$ of the set of elements of type T if and only if some conjugate gkg^{-1} of k is in the closure in $\mathbb{T}^{\mathrm{st}}_{2m}$ of the set of elements of type T in standard form. Note that gkg^{-1} itself need not be in standard form.

Let $\{k_i\}_{i\in\mathbb{N}}$ be a convergent sequence of elements of $\mathrm{SO}(2m)$ that are all of the same type $T = (a_0, a_\pi, \rho, s)$, and let $k = \lim_{i\to\infty} k_i \in \mathrm{SO}(2m)$. Then for each i, there is a g_i such that $g_i k_i g_i^{-1} \in \mathbb{T}^{\mathrm{st}}_{2m}$ is of standard form. By compactness of $\mathrm{SO}(2m)$, we may assume by passing to a subsequence that the g_i converge to some $g \in \mathrm{SO}(2m)$. Then by continuity of the action by conjugation and as $\mathbb{T}^{\mathrm{st}}_{2m}$ is closed, we have

$$gkg^{-1} = \lim_{i\to\infty} g_i k_i g_i^{-1} \in \mathbb{T}^{\mathrm{st}}_{2m}.$$

Conversely, if k is conjugate to some $gkg^{-1} \in \mathbb{T}^{\mathrm{st}}_{2m}$, where gkg^{-1} is the limit of a sequence $\{h_i\}_{i\in\mathbb{N}}$ of elements in $\mathbb{T}^{\mathrm{st}}_{2m}$ of the same type T in standard form, then $g^{-1}h_i g$ is a sequence of elements of type T that converges to k.

Now, for a type $T = \{a_0, a_\pi, \rho, s\}$ with $\rho = \{\rho_1, \ldots, \rho_\ell\}$, let $\mathbb{T}^{\mathrm{st}}_{2m}(T)$ denote the set of elements in $\mathbb{T}^{\mathrm{st}}_{2m}$ in standard form of type T. Suppose $h \in \overline{\mathbb{T}^{\mathrm{st}}_{2m}(T)}$ so that there is a sequence $\{h_i\}_{i\in\mathbb{N}} \subseteq \mathbb{T}^{\mathrm{st}}_{2m}(T)$ such that $h_i \to h$. Recall that if $s = -$, then the sign discrepancy in the angles of the h_i is taken to be in the final position, corresponding to ρ_ℓ. As each h_i has I and $-I$ in the first a_0 and a_π positions, respectively, it follows that h must as well. Similarly, letting $\theta_{j,i}$ denote the angle in the ρ_j position of h_i for $j = 1, \ldots, \ell$, we have that $\lim_{i\to\infty} \theta_{j,i}$ exists and is given by θ_j, the angle in the corresponding position of h, which can have any value. Let $J = \{j \in \{1, \ldots, \ell\} : \theta_j = 0\}$, and let $J' = \{j \in \{1, \ldots, \ell\} : \theta_j = \pi\}$. As it may be the case that the θ_j are not distinct, let σ denote the partition formed from $\rho \setminus \{\rho_j : j \in I \cup J\}$ by summing elements ρ_j and $\rho_{j'}$ when $\theta_j = \theta_{j'}$. Then if $s = +$ or θ_ℓ is generic, h has type

$$\left(a_0 + \sum_{j\in J} \rho_j,\ a_\pi + \sum_{j\in J'} \rho_j,\ \sigma,\ s\right),$$

while if $s = -$ and $\theta_\ell \in \{0, \pi\}$, h has type

$$\left(a_0 + \sum_{j\in I} \rho_j,\ a_\pi + \sum_{j\in J} \rho_j,\ \sigma,\ +\right).$$

Given an arbitrary element h' of $\mathbb{T}^{\mathrm{st}}_{2m}$ of the same form as h, it is easy to see that one can define a sequence $\{h'_i\}_{i\in\mathbb{N}}$ of elements of type T such that $h'_i \to h'$ simply by redefining the angles in the h_i corresponding to $j \notin J \cup J'$ to converge to those of h', choosing distinct sequences when $\theta_j = \theta_{j'}$ for $j \neq j'$ as above. It follows that

if $h \in \overline{\mathbb{T}^{\mathrm{st}}_{2m}(T)}$, then every element of $\mathbb{T}^{\mathrm{st}}_{2m}$ of the same form of h is contained in $\overline{\mathbb{T}^{\mathrm{st}}_{2m}(T)}$. However, by applying the Weil group to this sequence, it then follows that every element of $\mathbb{T}^{\mathrm{st}}_{2m}$ of the same type as h is contained in the closure of elements of type T in $\mathbb{T}^{\mathrm{st}}_{2m}$. This claim extends by conjugation to all of $\mathrm{SO}(2m)$ as above, so we conclude that the partition of $\mathrm{SO}(2m)$ into types satisfies the frontier condition.

Finally, note that this partition still satisfies frontier if we combine types of the form $(0, 1, \rho, s)$ and $(0, 0, \{1\} \cup \rho, s)$. If the set of elements of type T contains points of type $(0, 1, \sigma, s)$ in its closure, then T must itself be of the form either $(0, 1, \rho, s)$ or $(0, 0, \{1\} \cup \rho, s)$, where σ is formed from ρ or $\{1\} \cup \rho$ by summing elements as above. As these types are also combined, the resulting set must contain all elements of type $(0, 1, \sigma, s)$ and $(0, 0, \{1\} \cup \sigma, s)$ in its closure.

With this, we need only note that as the closure of $\mathbb{R}^{2m} \smallsetminus \{0\}$ is clearly $\mathbb{R}^{2m}$, by inspection, the pieces of type I, II, and III satisfy the frontier condition. Hence, by $\mathrm{SO}(2m)$-invariance of these pieces, frontier is satisfied in the quotient as well. □

It is of interest to note that the sets of type III form a decomposition of the loop space of the $\mathrm{SO}(2m)$-space $\mathbb{R}^{2m} \smallsetminus \{0\}$. Because each point in $\Lambda(\mathbb{R}^{2m} \smallsetminus \{0\})$ is contained in an $\mathrm{SO}(2m)$-invariant neighborhood in $\Lambda\mathbb{R}^{2m}$ that does not intersect $\mathrm{SO}(2m) \times \{0\}$, it follows that this decomposition induces the orbit Cartan type stratification of the inertia space $\Lambda(\mathrm{SO}(2m)\backslash(\mathbb{R}^{2m} \smallsetminus \{0\}))$.

The loop space $\Lambda(\mathbb{R}^{2m} \smallsetminus \{0\})$ is the loop space of a $\mathrm{SO}(2m)$-manifold with a single isotropy type and hence is a smooth manifold by [Farsi et al. 2012, Proposition 4.4]. Given an element $(h, x) \in \Lambda(\mathbb{R}^{2m} \smallsetminus \{0\})$, where we may assume up to conjugation that $x = (\|x\|, 0, \ldots, 0)$ and $h \in \mathbb{T}^{\mathrm{st}}_{2m-1}$ is in standard form as above, it must be that $a_0(h) > 0$. Hence, as the types of such elements are determined by their centralizers by Proposition 3.3, the decomposition of $\Lambda(\mathbb{R}^{2m} \smallsetminus \{0\})$, and hence the associated inertia space, corresponds to the decomposition into isotropy types, demonstrating that the orbit Cartan type stratification of this $\mathrm{SO}(2m)$-manifold coincides with its stratification by isotropy types. This is not generally true for the odd case, as it fails in the case of $\mathrm{SO}(3)$ acting on $\mathbb{R}^3 \smallsetminus \{0\}$ described in [Farsi et al. 2012, Section 4.2.6].

Acknowledgements

This paper is the result of the second author's Senior Seminar project in the Department of Mathematics and Computer Science at Rhodes College, and both authors express appreciation to the department and college for support during the preparation of this manuscript. The first author would like to thank Carla Farsi and Markus Pflaum for helpful conversations and support. The first author was partially supported by a Rhodes College Faculty Development Endowment Grant. The second author was partially supported by a Rhodes College Fellowship.

References

[Adem et al. 2007] A. Adem, J. Leida, and Y. Ruan, *Orbifolds and stringy topology*, Cambridge Tracts in Mathematics **171**, Cambridge University Press, Cambridge, 2007. MR 2009a:57044 Zbl 1157.57001

[Bredon 1972] G. E. Bredon, *Introduction to compact transformation groups*, Pure and Applied Mathematics **46**, Academic Press, New York, 1972. MR 54 #1265 Zbl 0246.57017

[Bröcker and tom Dieck 1995] T. Bröcker and T. tom Dieck, *Representations of compact Lie groups*, Graduate Texts in Mathematics **98**, Springer, New York, 1995. MR 97i:22005 Zbl 0874.22001

[Brylinski 1987] J.-L. Brylinski, "Cyclic homology and equivariant theories", *Ann. Inst. Fourier* (*Grenoble*) **37**:4 (1987), 15–28. MR 89j:55008 Zbl 0625.55003

[tom Dieck 1987] T. tom Dieck, *Transformation groups*, de Gruyter Studies in Mathematics **8**, De Gruyter, Berlin, 1987. MR 89c:57048 Zbl 0611.57002

[Duistermaat and Kolk 2000] J. J. Duistermaat and J. A. C. Kolk, *Lie groups*, Springer, Berlin, 2000. MR 2001j:22008 Zbl 0955.22001

[Farsi et al. 2012] C. Farsi, M. Pflaum, and C. Seaton, "Inertia spaces of proper Lie group actions and their topological properties", preprint, 2012. arXiv 1207.0595v1

[Humphreys 1978] J. E. Humphreys, *Introduction to Lie algebras and representation theory*, Graduate Texts in Mathematics **9**, Springer, New York, 1978. MR 81b:17007 Zbl 0447.17001

[Koszul 1953] J. L. Koszul, "Sur certains groupes de transformations de Lie", *Colloques Int. Centre Nat. Rech. Sci.* **52** (1953), 137–141. MR 15,600g Zbl 0101.16201

[Navarro González and Sancho de Salas 2003] J. A. Navarro González and J. B. Sancho de Salas, C^∞-*differentiable spaces*, Lecture Notes in Mathematics **1824**, Springer, Berlin, 2003. MR 2005c:58008 Zbl 1039.58001

[Pflaum 2001] M. J. Pflaum, *Analytic and geometric study of stratified spaces*, Lecture Notes in Mathematics **1768**, Springer, Berlin, 2001. MR 2002m:58007 Zbl 0988.58003

[Tapp 2005] K. Tapp, *Matrix groups for undergraduates*, Student Mathematical Library **29**, American Mathematical Society, Providence, RI, 2005. MR 2006d:20001 Zbl 1089.20001

Received: 2012-08-30 Revised: 2013-05-17 Accepted: 2013-05-19

seatonc@rhodes.edu *Department of Mathematics and Computer Science, Rhodes College, 2000 North Parkway, Memphis, TN 38112, United States*

jwwells@math.hawaii.edu *Department of Mathematics, University of Hawaii at Manoa, 2565 McCarthy Mall, Honolulu, HI 96822, United States*

INVOLVE 6:4 (2013)
dx.doi.org/10.2140/involve.2013.6.483

Optional unrelated-question randomized response models

Sat Gupta, Anna Tuck, Tracy Spears Gill and Mary Crowe

(Communicated by Kenneth S. Berenhaut)

We propose a generalization of Greenberg's unrelated-question randomized response model allowing subjects the option of giving a correct response if they find the survey question nonsensitive, and to give a scrambled response if they find the question sensitive. Models are provided for both the binary response and the quantitative response situations. Mathematical properties of the proposed models are examined and validated with computer simulations.

1. Introduction

Obtaining accurate information is essential in all surveys, particularly in public health research where respondents often face sensitive and personal questions. Examples include surveys of sexual behavior, drug use, or illegal activities. Despite assurances of anonymity, subjects often give untruthful responses leading to problematic response bias.

One method of reducing this bias is the randomized response technique (RRT), originally introduced in [Warner 1965], and subsequently developed and generalized by many researchers [Greenberg et al. 1969; Gupta et al. 2002; 2010; Mehta et al. 2012; Sousa et al. 2010]. We will focus on the unrelated-question RRT method, developed in [Greenberg et al. 1969]. Compared to direct questioning methods, all RRT methods lead to more accurate estimates of sensitive behaviors, because of increased anonymity of the subject's response. In the unrelated-question model, a predetermined proportion of subjects are randomized to answer an innocuous unrelated question with known prevalence level. The researcher is unaware of which question (actual or innocuous) any particular respondent answered, although the mean of the research question can be estimated at the aggregate level. Unrelated-question RRT has been used extensively over the past fifty years to estimate

MSC2010: 62D05.

Keywords: parameter estimation, randomized response technique, unrelated-question model, optional scrambling.

Research supported by NSF grants DBI 0926288 and DMS 0850465.

prevalence of behaviors ranging from induced abortion [Chow et al. 1979] to software piracy [Kwan et al. 2010] and livestock disease prevalence [Cross et al. 2010]. This technique avoids the ethical issues associated with the bogus pipeline technique [Jones and Sigall 1971] and is not as lengthy as the Marlowe–Crowne social desirability scale method [Crowne and Marlowe 1960]. Here the increase in anonymity offered by the technique lessens respondent anxiety during the survey, resulting in more truthful responses [Stem and Bozman 1988].

The original unrelated-question RRT model makes no differentiation as to whether an individual actually considers the topic sensitive; every subject is assumed to find the research question sensitive, so all subjects utilize the randomization device to produce a scrambled response. However, a topic or question may be sensitive for one person, but not sensitive for another. Optional RRT models, introduced in [Gupta et al. 2002], take this into account by allowing subjects who do not find the question sensitive to answer it without utilizing the randomization step. Subjects who find the research question sensitive still use the randomization device prior to providing a response. In this optional model, the researcher remains unaware as to whether or not the subject used the scrambling device or provided a truthful response.

We propose a generalization of the unrelated-question RRT, which takes this difference into account by allowing the randomization step to be optional for the subjects. We deal with both the binary response and the quantitative response situations and estimate the prevalence (π) of the sensitive behavior and the mean response (μ) of the quantitative sensitive question. In addition, the model also estimates the sensitivity level (W) of the underlying question, which is the proportion of subjects who consider the question to be sensitive, and hence choose to provide a scrambled response. We provide the theoretical framework for the two models and examine their mathematical properties, which are also validated by computer simulations.

2. Proposed quantitative response model

We begin first with the quantitative response case, where the researcher is interested in estimating population mean. A randomization device provided to the respondent by the researcher determines whether the subject receives the sensitive research question or the innocuous, unrelated question.

Let X be the true sensitive variable of interest with unknown mean μ_X and unknown variance σ_X^2, and Y be a nonsensitive variable with known mean μ_Y and known variance σ_Y^2. Let p represent the probability of receiving the sensitive question from the randomization device.

The reported response Z is given by

$$Z = \begin{cases} X & \text{with probability } p, \\ Y & \text{with probability } 1 - p. \end{cases}$$

Let W be the sensitivity level of the question. That is, a proportion W of the respondents considers the question sensitive and will choose to provide a scrambled response. Others will provide a direct response with probability $1 - W$. Then

$$Z = \begin{cases} X & \text{with probability } (1-W) + Wp, \\ Y & \text{with probability } W(1-p), \end{cases}$$

with

$$E(Z) = (1-W)E(X) + W(pE(X) + (1-p)E(Y)), \tag{2-1}$$
$$\mathrm{Var}(Z) = [(1-W) + Wp]E(X^2) + W(1-p)E(Y^2) - [E(Z)]^2.$$

Here, both μ_X and W are unknown parameters. To solve the above equation for two unknowns, we use a split-sample approach where the total sample size may be split into two subsamples, each receiving a randomization device with a different probability (p_i, $i = 1, 2$) of receiving the sensitive question. The expected response in the i-th ($i = 1, 2$) subsample then is given by

$$E(Z_i) = (1-W)E(X) + W(p_i E(X) + (1-p_i)E(Y)), \quad \text{where } i = 1, 2. \tag{2-2}$$

2.1. ***Estimation of population mean.*** Solving the system of two equations (2-2) for the parameters of interest, we get

$$\frac{E(Z_1) - E(X)}{E(Z_2) - E(X)} = \frac{1-p_1}{1-p_2}.$$

Solving for $E(X)$, we get

$$E(X) = \frac{E(Z_1) - \lambda E(Z_2)}{1-\lambda}, \quad \text{where } \lambda = \frac{1-p_1}{1-p_2}.$$

This suggests estimating μ_X by

$$\hat{\mu}_X = \frac{\bar{Z}_1 - \lambda \bar{Z}_2}{1-\lambda}, \tag{2-3}$$

where $\bar{Z}_i$ is the sample mean of reported responses in the i-th subsample. The variance of this estimator is given by

$$\mathrm{Var}(\hat{\mu}_X) = \frac{\mathrm{Var}(\bar{Z}_1) + \lambda^2 \mathrm{Var}(\bar{Z}_2)}{(1-\lambda)^2}, \tag{2-4}$$

where

$$\mathrm{Var}(\bar{Z}_1) = \frac{[(1-W) + Wp_1]E(X^2) + W(1-p_1)E(Y^2) - [E(Z_1)]^2}{n_1},$$
$$\mathrm{Var}(\bar{Z}_2) = \frac{[(1-W) + Wp_2]E(X^2) + W(1-p_2)E(Y^2) - [E(Z_2)]^2}{n_2}.$$

It is easy to see that $E(\hat{\mu}_X) = \mu_X$, so the estimator $\hat{\mu}_X$ is unbiased. Also, $\hat{\mu}_X$ is a linear combination of independent sample means; hence it has an asymptotic normal distribution. More formally, we have the following asymptotic result:

Theorem 1. *The estimator $\hat{\mu}_X$ is distributed as $AN(\mu_X, V)$, where*

$$V = \frac{\text{Var}(\bar{Z}_1) + \lambda^2 \,\text{Var}(\bar{Z}_2)}{(1-\lambda)^2}$$

with

$$\text{Var}(\bar{Z}_i) = \frac{[(1-W) + Wp_i]E(X^2) + W(1-p_i)E(Y^2) - [E(Z_2)]^2}{n_i}, \quad i = 1, 2$$

and

$$E(Z_i) = (1-W)E(X) + W(p_i E(X) + (1-p_i)E(Y)).$$

2.2. ***Optimal allocation of sample size.*** For the optimal sample split (n_1, n_2), we look at the first derivative of $\text{Var}(\hat{\mu}_X)$ from (2-3), given by

$$\frac{\partial \text{Var}(\hat{\mu}_X)}{\partial n_1} = \frac{1}{(1-\lambda)^2}\left\{\frac{-\sigma_1^2}{n_1^2} + \lambda^2 \frac{\sigma_2^2}{(n-n_1)^2}\right\}.$$

Setting this equal to zero, we get

$$\begin{aligned} 0 &= \frac{1}{(1-\lambda)^2}\left(\frac{-\sigma_1^2}{n_1^2} + \lambda^2\left(\frac{\sigma_2}{n-n_1}\right)^2\right), \\ \frac{\sigma_1^2}{n_1^2} &= \lambda^2 \frac{\sigma_2^2}{(n-n_1)^2}, \\ \frac{n-n_1}{n_1} &= \sqrt{\lambda^2 \frac{\sigma_2^2}{\sigma_1^2}} = \left|\lambda \frac{\sigma_2}{\sigma_1}\right|. \end{aligned}$$

Therefore,

$$\frac{n_2}{n_1} = \lambda \frac{\sigma_2}{\sigma_1} \tag{2-5}$$

gives the optimal ratio of subjects split in the two subsamples. This will result in the minimum variance of the estimator $\hat{\mu}_X$ since the second derivative of $\text{Var}\,(\hat{\mu}_X)$ is positive. Equation (2-5) assumes rough preliminary estimates of σ_1 and σ_2 are available. These may be obtained through a pilot study.

2.3. ***Estimation of sensitivity level.*** In addition to estimating the mean $(\hat{\mu}_X)$, the proportion of subjects who scramble their response (W) is also estimated. We can easily solve (2-2) for W, which will lead to the possible estimator

$$\hat{W} = \frac{\bar{Z}_1 - \bar{Z}_2}{(p_2 - p_1)(\mu_Y - \hat{\mu}_X)}. \tag{2-6}$$

This representation of $\hat{W}$ as a ratio of two random variables presents difficulties in deriving its properties. We can, however, rewrite $\hat{W}$ in terms of $\bar{Z}_1$ and $\bar{Z}_2$ to get

$$\hat{W} = \frac{\bar{Z}_1 - \bar{Z}_2}{\mu_Y(p_2 - p_1) + (1 - p_2)\bar{Z}_1 - (1 - p_1)\bar{Z}_2}. \tag{2-7}$$

Using the first-order bivariate Taylor approximation, with $A = E(\bar{Z}_1)$ and $B = E(\bar{Z}_2)$, we get

$$\begin{aligned}
\hat{W} &\approx \hat{W}(A, B) + \frac{\partial \hat{W}(\bar{Z}_1, \bar{Z}_2)}{\partial \bar{Z}_1}\bigg|_{A,B} (\bar{Z}_1 - A) + \frac{\partial \hat{W}(\bar{Z}_1, \bar{Z}_2)}{\partial \bar{Z}_2}\bigg|_{A,B} (\bar{Z}_2 - B) \\
&= \frac{A - B}{\mu_Y(p_2 - p_1) + (1 - p_2)A - (1 - p_1)B} \\
&\quad + \frac{(p_2 - p_1)(\mu_Y - B)(\bar{Z}_1 - A)}{[\mu_Y(p_2 - p_1) + (1 - p_2)A - (1 - p_1)B]^2} \\
&\quad + \frac{(p_2 - p_1)(A - \mu_Y)(\bar{Z}_2 - B)}{[\mu_Y(p_2 - p_1) + (1 - p_2)A - (1 - p_1)B]^2} =: \hat{W}_1.
\end{aligned}$$

Taking the expected value, we get $(Z_1 - \mu_\gamma) \to (\Lambda - \mu_\gamma)$:

$$\begin{aligned}
E(\hat{W}_1) &= \frac{A - B}{\mu_Y(p_2 - p_1) + (1 - p_2)A - (1 - p_1)B} \\
&\quad + \frac{(p_2 - p_1)(\mu_Y - B)(E(\bar{Z}_1) - A)}{[\mu_Y(p_2 - p_1) + (1 - p_2)A - (1 - p_1)B]^2} \\
&\quad + \frac{(p_2 - p_1)(\bar{Z}_1 - \mu_Y)(E(\bar{Z}_2) - B)}{[\mu_Y(p_2 - p_1) + (1 - p_2)A - (1 - p_1)B]^2} \\
&= \frac{A - B}{\mu_Y(p_2 - p_1) + (1 - p_2)A - (1 - p_1)B} = W.
\end{aligned}$$

Thus $\hat{W}_1$, the first-order approximation of $\hat{W}$, is an unbiased estimator of W with variance given by

$$\begin{aligned}
\mathrm{Var}(\hat{W}_1) &= \left(\frac{(p_2 - p_1)(\mu_Y - B)}{[\mu_Y(p_2 - p_1) + (1 - p_2)A - (1 - p_1)B]}\right)^2 \frac{\sigma_1^2}{n_1} \\
&\quad + \left(\frac{(p_2 - p_1)(\mu_Y - B)(A - \mu_Y)}{[\mu_Y(p_2 - p_1) + (1 - p_2)A - (1 - p_1)B]}\right)^2 \frac{\sigma_2^2}{n_2},
\end{aligned} \tag{2-8}$$

where

$$\sigma_i^2 = [1 - W + Wp_i]E(X^2) + W(1 - p_i)E(Y^2) - [E(Z_i)]^2, \quad i = 1, 2.$$

Also, $\hat{W}_1$ is asymptotically normal since it is a linear combination of independent sample means $\bar{Z}_1$ and $\bar{Z}_2$. This property is later confirmed by simulation. This result is summarized in the following theorem.

Theorem 2. $\hat{W}_1 \sim AN(W, V_w)$, *where*

$$\mathrm{Var}(\hat{W}_1) = \left(\frac{(p_2 - p_1)(\mu_Y - B)}{[\mu_Y(p_2 - p_1) + (1 - p_2)A - (1 - p_1)B]}\right)^2 \frac{\sigma_1^2}{n_1} + \left(\frac{(p_2 - p_1)(A - \mu_Y)}{[\mu_Y(p_2 - p_1) + (1 - p_2)A - (1 - p_1)B]}\right)^2 \frac{\sigma_2^2}{n_2},$$

$$\sigma_i^2 = [1 - W + Wp_i]E(X^2) + W(1 - p_i)E(Y^2) - [E(Z_i)]^2, \quad i = 1, 2.$$

3. Proposed binary response model

The estimator proposed in the preceding section is used when an estimate of the population mean is needed. In many cases the main research interest is in the prevalence of a particular sensitive behavior or characteristic. In this case the research question demands a binary response, such as "yes" or "no". We modify the preceding estimator to be used in these cases.

3.1. ***Proposed model.*** Let X be a sensitive binary variable of interest with unknown mean π_X, and Y be a nonsensitive binary variable with known mean π_Y. Let p represent probability of receiving the sensitive question from the randomization device. Here the probability of a "yes" response (P_Y) is given by

$$P_Y = (1 - W)\pi_X + W[p\pi_X + (1 - p)\pi_Y].$$

Again, the sample is split into two subsamples to solve for both π_X and W. The probability of a "yes" response in the i-th ($i = 1, 2$) subsample is given by

$$P_{Y_i} = (1 - W)\pi_X + W[p_i\pi_X + (1 - p_i)\pi_Y], \quad i = 1, 2.$$

Solving this system of two equations for π_X gives

$$\pi_X = \frac{P_{Y_1} - \lambda P_{Y_2}}{1 - \lambda}, \quad \text{where } \lambda = \frac{1 - p_1}{1 - p_2}. \tag{3-1}$$

3.2. ***Estimation of population proportion.*** Using (3-1), we obtain the estimate for the population proportion (π_X) of the sensitive characteristic as

$$\hat{\pi}_X = \frac{\hat{P}_{Y_1} - \lambda \hat{P}_{Y_2}}{1 - \lambda}, \tag{3-2}$$

with variance given by

$$\mathrm{Var}(\hat{\pi}_X) = \frac{\mathrm{Var}(\hat{P}_{Y_1}) + \lambda^2 \mathrm{Var}(\hat{P}_{Y_2})}{(1 - \lambda)^2}, \tag{3-3}$$

where

$$\mathrm{Var}(\hat{P}_{Y_1}) = \frac{P_{Y_1}(1-P_{Y_1})}{n_1} \quad \text{and} \quad \mathrm{Var}(\hat{P}_{Y_2}) = \frac{P_{Y_2}(1-P_{Y_2})}{n_2}.$$

Again, it can easily be seen that $E(\hat{\pi}_X) = \pi_X$, so the estimator $\hat{\pi}_X$ is unbiased. Also $\hat{\pi}_X$ is a linear combination of independent sample means, and hence has an asymptotic normal distribution.

3.3. ***Optimal allocation of sample size.*** Just as in the quantitative response case, the optimal sample split is given by

$$\frac{n_2}{n_1} = \lambda\sqrt{\frac{P_{Y_2}(1-P_{Y_2})}{P_{Y_1}(1-P_{Y_1})}}. \tag{3-4}$$

3.4. ***Estimation of sensitivity level.*** From (3-1), an estimator for the sensitivity level (W) in the binary case can be represented as

$$\hat{W}_\pi = \frac{\hat{P}_{Y_1} - \hat{P}_{Y_2}}{(p_2-p_1)(\pi_Y - \hat{\pi}_X)} = \frac{\hat{P}_{Y_1} - \hat{P}_{Y_2}}{\pi_Y(p_2-p_1) + (1-p_2)\hat{P}_{Y_1} - (1-p_1)\hat{P}_{Y_2}}. \tag{3-5}$$

Applying the first-order Taylor approximation expansion for a bivariate function, and assuming $A = P_{Y_1}$, $B = P_{Y_2}$, this can be approximated by

$$\begin{aligned}\hat{W}_\pi \approx{}& \frac{A-B}{\pi_Y(p_2-p_1) + (1-p_2)A - (1-p_1)B} \\ &+ \frac{(p_2-p_1)(\pi_Y - B)(\hat{P}_{Y_1} - A)}{[\pi_Y(p_2-p_1) + (1-p_2)A - (1-p_1)B]^2} \\ &+ \frac{(p_2-p_1)(A-\pi_Y)(\hat{P}_{Y_2} - B)}{[\pi_Y(p_2-p_1) + (1-p_2)A - (1-p_1)B]^2} =: \hat{W}_{\pi_1}.\end{aligned}$$

It can be verified that

$$E(\hat{W}_{\pi_1}) = \frac{A-B}{\mu_Y(p_2-p_1) + (1-p_2)A - (1-p_1)B} = W_\pi.$$

Thus, $\hat{W}_{\pi_1}$ is an unbiased estimator of W with variance given by

$$\begin{aligned}\mathrm{Var}(\hat{W}_{\pi_1}) = {}& \left(\frac{(p_2-p_1)(\pi_Y - B)}{[\pi_Y(p_2-p_1) + (1-p_2)A - (1-p_1)B]}\right)^2 \frac{\sigma_1^2}{n_1} \\ &+ \left(\frac{(p_2-p_1)(A-\mu_Y)}{[\pi_Y(p_2-p_1) + (1-p_2)A - (1-p_1)B]}\right)^2 \frac{\sigma_2^2}{n_2},\end{aligned} \tag{3-6}$$

where

$$\sigma_1^2 = \frac{P_{Y_1}(1-P_{Y_1})}{n_1} \quad \text{and} \quad \sigma_2^2 = \frac{P_{Y_2}(1-P_{Y_2})}{n_2}.$$

W	n_1	$\hat{\mu}_X$	$\mathrm{Var}(\hat{\mu}_X)$	$\widehat{\mathrm{Var}}(\hat{\mu}_X)$	$\hat{W}_1$	$\mathrm{Var}(\hat{W}_1)$	$\widehat{\mathrm{Var}}(\hat{W}_1)$
0.0	698	1.9988	0.0058	0.0058	−0.0001	0.0065	0.0067
0.1	674	2.0012	0.0066	0.0065	0.0982	0.0069	0.0069
0.2	680	2.0013	0.0068	0.0068	0.1982	0.0073	0.0073
0.3	690	2.0008	0.0072	0.0072	0.2984	0.0077	0.0077
0.4	699	2.0004	0.0075	0.0075	0.3989	0.0080	0.0080
0.5	710	2.0005	0.0079	0.0079	0.4991	0.0082	0.0082
0.6	722	2.0006	0.0081	0.0082	0.5996	0.0083	0.0084
0.7	737	1.9998	0.0084	0.0084	0.7004	0.0085	0.0085
0.8	753	1.9996	0.0086	0.0088	0.8005	0.0087	0.0087
0.9	774	1.9999	0.0089	0.0090	0.9002	0.0090	0.0090
1.0	800	1.9991	0.0091	0.0092	1.0001	0.0095	0.0097

Table 1. Estimates of μ_X and W with optimized subsamples. X and Y have Poisson distributions with $\mu_X = 2.0$, $\mu_Y = 4.0$. Total sample size is 1000, $p_1 = 0.8$, $p_2 = -0.2$.

Also, $\hat{W}_{\pi_1}$ clearly has an asymptotic normal distribution being a linear combination of independent sample means.

4. Simulation study

The preceding theoretical formulas are tested empirically through computer simulations. Poisson distribution is assumed for both X and Y. The subsample split (n_1, n_2) is obtained by the optimal split method described above. Table 1 and Table 2 present simulation results obtained with SAS.

The simulation results provide strong support for the theoretical results that $\hat{\mu}_X$ and $\hat{\pi}_X$ are unbiased. The theoretical and simulated variances of $\hat{\mu}_X$ and $\hat{\pi}_X$ can also be seen to be very close. The simulations also support that $\hat{W}_1$ and $\hat{W}_{\pi_1}$ are good estimators of W for the quantitative case and the binary case, respectively.

We also note that $\hat{W}_1$ and $\hat{W}_{\pi_1}$ may occasionally give estimates that are outside of the normal range [0, 1]. This happens when the true value of W is close to zero or 1. As in [Warner 1965], this is because our estimators are unconstrained. In such cases we recommend using an estimate of zero if $\hat{W}_1 < 0$, and 1 if $\hat{W}_1 > 1$.

The Kolmogorov–Smirnov normality test is used in SAS to check the sampling distributions of $\hat{\mu}_X$, $\hat{\pi}_X$, $\hat{W}_1$, and $\hat{W}_{\pi_1}$ against the normal distribution. The p-values for $\hat{\mu}_X$, $\hat{\pi}_X$, $\hat{W}_1$, and $\hat{W}_{\pi_1}$ are all greater than 0.15, indicating that their distributions are not significantly different from the normal distribution.

W	n_1	$\hat{\pi}_X$	$\mathrm{Var}(\hat{\pi}_X)$	$\widehat{\mathrm{Var}}(\hat{\pi}_X)$	$\hat{W}_{\pi_1}$	$\mathrm{Var}(\hat{W}_{\pi_1})$	$\widehat{\mathrm{Var}}(\hat{W}_{\pi_1})$
0.0	800	0.1508	0.0003	0.0003	−0.0037	0.0045	0.0047
0.1	786	0.1500	0.0004	0.0004	0.1005	0.0050	0.0050
0.2	777	0.1500	0.0004	0.0004	0.2007	0.0053	0.0052
0.3	772	0.1499	0.0005	0.0005	0.3008	0.0054	0.0055
0.4	770	0.1498	0.0005	0.0005	0.4009	0.0054	0.0055
0.5	770	0.1501	0.0005	0.0005	0.5008	0.0053	0.0055
0.6	772	0.1502	0.0005	0.0005	0.6005	0.0052	0.0052
0.7	776	0.1502	0.0006	0.0006	0.7007	0.0049	0.0050
0.8	782	0.1501	0.0006	0.0006	0.8006	0.0046	0.0046
0.9	790	0.1500	0.0006	0.0006	0.9008	0.0042	0.0042
1.0	800	0.1500	0.0006	0.0006	0.9996	0.0038	0.0038

Table 2. Estimates of π_X and W with optimized subsamples. The true values are $\pi_X = 0.15$, $\pi_Y = 0.85$. Total sample size is 1000, $p_1 = 0.8$, $p_2 = -0.2$.

5. Concluding remarks

The optional unrelated-question RRT proposed above provides models for simultaneously estimating both the mean and sensitivity level of a sensitive behavior. This is distinct from previous unrelated-question RRT models, which estimate only the mean. Estimators are derived for both the quantitative and binary response cases. In both cases, estimators of the mean ($\hat{\mu}_X$, $\hat{\pi}_X$) and first-order Taylor approximations of the sensitivity level ($\hat{W}_1$, $\hat{W}_{\pi_1}$) are shown to be asymptotically normal and unbiased.

Of note in Table 1, the variances of both $\hat{\mu}_X$ and $\hat{W}_1$ increase as W increases (when more subjects choose to provide scrambled responses). In Table 2 the variance of $\hat{\pi}_X$ increases slightly as W increases. When optionality is incorporated into this model, when even a small proportion of subjects do not find the question sensitive (and thus answer directly) the variance of the estimator is smaller than in a comparable model where all subjects must provide a scrambled response ($W = 1.0$).

References

[Chow et al. 1979] L. P. Chow, W. Gruhn, and W. P. Chang, "Feasibility of the randomized response technique in rural Ethiopia", *Amer. J. Public Health* **69**:3 (1979), 273–276.

[Cross et al. 2010] P. Cross, G. Edwards-Jones, H. Omed, and A. Williams, "Use of a randomized response technique to obtain sensitive information on animal disease prevalence", *Prev. Vet. Med.* **96**:3–4 (2010), 252–262.

[Crowne and Marlowe 1960] D. P. Crowne and D. Marlowe, "A new scale of social desirability independent of psychopathology", *J. Consult. Psychol.* **24**:4 (1960), 349–54.

[Greenberg et al. 1969] B. G. Greenberg, A.-L. A. Abul-Ela, W. R. Simmons, and D. G. Horvitz, "The unrelated question randomized response model: Theoretical framework", *J. Amer. Statist. Assoc.* **64** (1969), 520–539. MR 40 #982

[Gupta et al. 2002] S. Gupta, B. Gupta, and S. Singh, "Estimation of sensitivity level of personal interview survey questions", *J. Statist. Plann. Inference* **100**:2 (2002), 239–247. MR 1877192 Zbl 0985.62010

[Gupta et al. 2010] S. Gupta, J. Shabbir, and S. Sehra, "Mean and sensitivity estimation in optional randomized response models", *J. Statist. Plann. Inference* **140**:10 (2010), 2870–2874. MR 2011h:62027 Zbl 1191.62009

[Jones and Sigall 1971] E. E. Jones and H. Sigall, "The bogus pipeline: A new paradigm for measuring affect and attitude", *Psychological Bulletin* **76**:5 (1971), 349–364.

[Kwan et al. 2010] S. S. K. Kwan, M. K. P. So, and K. Y. Tam, "Applying the randomized response technique to elicit truthful responses to sensitive questions in IS research: The case of software piracy behavior", *Info. Sys. Research* **21**:4 (2010), 941–959.

[Mehta et al. 2012] S. Mehta, B. K. Dass, J. Shabbir, and S. Gupta, "A three stage optional randomized response model", *J. Stat. Theory Pract.* **6**:3 (2012), 417–427.

[Sousa et al. 2010] R. Sousa, J. Shabbir, P. C. Real, and S. Gupta, "Ratio estimation of the mean of a sensitive variable in the presence of auxiliary information", *J. Stat. Theory Pract.* **4**:3 (2010), 495–507. MR 2758690 Zbl 05902629

[Stem and Bozman 1988] D. E. Stem and C. S. Bozman, "Respondent anxiety reduction with the randomized response technique", *Adv. Consum. Res.* **15**:1 (1988), 595–599.

[Warner 1965] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias", *J. Am. Stat. Assoc.* **60**:309 (1965), 63–69.

Received: 2012-10-01 Revised: 2012-11-05 Accepted: 2012-11-08

sngupta@uncg.edu *Department of Mathematics and Statistics, University of North Carolina at Greensboro, 317 College Avenue, Greensboro, NC 27412, United States*

avmikh@uw.edu *Department of Biostatistics, University of Washington, 1705 NE Pacific Street, Seattle, Washington 98195, United States*

tgspears@uncg.edu *School of Nursing, University of North Carolina at Greensboro, P.O. Box 26170, Greensboro, NC 27402, United States*

tgspears@unc.edu *Department of Biostatistics, University of North Carolina at Chapel Hill, 135 Dauer Drive, Chapel Hill, NC 27599, United States*

INVOLVE 6:4 (2013)
dx.doi.org/10.2140/involve.2013.6.493

On the difference between an integer and the sum of its proper divisors

Nichole Davis, Dominic Klyve and Nicole Kraght

(Communicated by Kenneth S. Berenhaut)

Let $\sigma(n)$ be the sum of the divisors of n. Although much attention has been paid to the possible values of $\sigma(n)-n$ (the sum of proper divisors), comparatively little work has been done on the possible values of $e(n) := \sigma(n)-2n$. Here we present some theoretical and computational results on these values. In particular, we exhibit some infinite and possibly infinite families of integers that appear in the image of $e(n)$. We also find computationally all values of $n < 10^{20}$ for which $e(n)$ is odd, and we present some data from our computations. At the end of this paper, we present some conjectures suggested by our computational work.

1. Introduction and background

Let $s(n)$ be the sum of the proper divisors of n, so that $s(n) = \sigma(n)-n$, where $\sigma(n)$ represents the standard sum-of-divisors function. We shall refer to the value by which the sum of the proper divisors of an integer n exceeds n as the *excedent* of n, which we denote by $e(n)$, so that $e(n) := s(n)-n$, or $e(n) := \sigma(n)-2n$. In a sense, values of $e(n)$ have been studied since antiquity. The Pythagoreans, for example, were especially interested in finding those n for which $e(n) = 0$. These are the *perfect* numbers. Today we also use the ancient Greek descriptors *deficient* and *abundant* to refer to those integers n for which $e(n) < 0$ and $e(n) > 0$, respectively.

More recently, some particular values of the excedent of n have been studied in the literature. Most noteworthy is the case where $e(n) = 1$. An integer for which $e(n) = 1$ is said to be *quasiperfect*. Quasiperfect numbers were first studied by Cattaneo [1951], who referred to $e(n)$ as the *eccedenza* of n, partly inspiring our choice of the English word *excedent*. Cattaneo showed that a quasiperfect number must be an odd square, and that if it is relatively prime to 3, it must have at least seven distinct prime factors. These results have since been improved. We now know, for example, that if n is a quasiperfect number:

MSC2010: 11A25, 11Y70.
Keywords: sigma function, sum of divisors, excedents, computational mathematics.

(1) $n = k^2$, where k is odd [Cattaneo 1951];

(2) if m is a proper divisor of n, then $\sigma(m) < 2n$ [Cattaneo 1951];

(3) if $r \mid \sigma(n)$ then $r \equiv 1$ or $3 \pmod 8$ [Cattaneo 1951];

(4) n has at least seven prime factors [Hagis and Cohen 1982];

(5) $n > 10^{35}$ [Hagis and Cohen 1982].

Despite this impressive list, however, the biggest question concerning quasiperfect numbers, namely, *do quasiperfect numbers exist?*, remains unanswered. In the language of this paper, we could say that we still don't know whether there are integers n for which $e(n) = 1$.

There seems to have been only one attempt to pursue more general questions of this sort. In his Ph.D. thesis (see [Cohen 1982] for a summary), Cohen considered a generalization of quasiperfect numbers. According to his definition, a *k-quasiperfect number* is an integer n for which $s(n) - n = k^2$ for a positive integer k relatively prime to n. He proved, among other things, that if such numbers exist, they must be larger than 10^{20} and have at least four distinct prime factors.

In this work, we wish to broaden Cohen's definition of a k-quasiperfect number to allow for any integer value of k. Then a 0-*quasiperfect number* is just a perfect number, a 1-*quasiperfect number* is the integer normally defined as a quasiperfect number, and we could denote the integers which Cohen considered simply as k-*quasiperfect numbers for square* k.

Our primary goal is to classify those integers m that are in the image of the excedent function. We call these integers *excedents*. Integers not in the image of $e(n)$ we call *nonexcedents*. The general problem of how to determine whether a given integer is an excedent seems very hard, however, and we are far from a complete classification. We do, however, give a few results concerning infinite and two potentially infinite families of excedents. We also give a conjecture, based on extensive computational evidence, about which small values of m are excedents, and which are nonexcedents.

2. Related work

It is worth noting that although references to values of $s(n) - n$ in the literature are fairly rare, some work has been done on values of $\sigma(n) - n$. Erdős [1973] showed that there are infinitely many numbers m for which $\sigma(n) - n = m$ has no solution, and furthermore that these m have positive lower density. Chen and Zhao [2011] have recently improved this to show that the density of these m is at least 0.06. Pomerance [1975] has considered a more general case, considering the set

$$S(a) = \{n : \sigma(n) \equiv a \pmod n\}.$$

He showed that for all a, the set $S(a)$ has at least two elements.

More recently, there has been an increase in interest in topics related to the values of $s(n)-n$. Anavi, Pollack and Pomerance [Anavi et al. 2013] show that the number of elements not greater than x in $S(a)$ (not counting those in a certain "obvious" set involving multiples of perfect and multiply perfect numbers) is bounded by $x^{1/2+o(1)}$ for each $|a| \le x^{1/4}$. Since $\sigma(n) \equiv e(n) \pmod{n}$, this immediately gives an upper bound on the number of n up to x for which $e(n) \equiv a \pmod{n}$ as well. One conclusion is that there can be no more than $x^{1/2+o(1)}$ k-quasiperfect integers up to x (outside of the obvious set) for any $k \le x^{1/4}$.

is studied in [Pollack and Shevelev 2012]. These are integers whose excedent is equal to one of the divisors. Finally, it is shown in [Pollack and Pomerance 2013] that for odd k, the number of k-quasiperfect numbers that are $\le x$ is at most $x^{1/4+o(1)}$ as $x \to \infty$.

Somewhat disappointingly, a close study of the references in this paper, including several suggested by the referee, show that the first three theorems in this paper have already appeared in some form in the literature. We shall still give our statements (and in one case, our proof) of these theorems in the hope that they may offer two things. First, we present and prove our theorems in an elementary manner. Second, our independent discovery of these results play an important role in our story, and help to motivate much of the computational work in the latter parts of the paper.

3. Computational experiments

Most computations for this work were conducted using PARI/GP. Initially, we computed $e(n)$ for all n in the range $[1, 10^{10}]$. We then recorded the number of times an integer m occurred as a value of $e(n)$ in this range. Let $N_m(x)$ be the number of integers $n \le x$ for which $e(n) = m$. Values for some small m from our computation are given in Table 1. It is worth noting that there are several methods which can speed up the computation of many values of $e(n)$. We worked primarily by isolating the values in which we were especially interested. A clever method for finding all numbers not in the image of $s(n)$ up to a given bound has recently been described in [Pomerance and Yang 2012].

In looking at this data, a few things immediately stand out. The most obvious is that there are many integers whose excedent is 12. Slightly less obvious, perhaps, is what seems to be a bias toward even values of the excedent function. These observations would guide our initial work.

It is clear that 12 is in the image of $e(n)$ quite often, leading us to ask immediately if there are other values which appear very often. Extending our search, we found a few other values of m for which there are a large number of integers n with $e(n) = m$, namely $m = 56$ and $m = 992$. A bit of consideration reveals that the

m	$N_m(10^{10})$	m	$N_m(10^{10})$
1	0	−1	32
2	9	−2	4
3	1	−3	0
4	10	−4	14
5	0	−5	1
6	3	−6	8
7	1	−7	1
8	25	−8	15
9	0	−9	0
10	3	−10	9
11	0	−11	1
12	78505339	−12	7
13	0	−13	0
14	6	−14	4
15	0	−15	0
16	20	−16	35
17	1	−17	0
18	10	−18	5
19	1	−19	2
20	20	−20	15

Table 1. Number of integers $n \leq 10^{10}$ for which $e(n) = m$ for small m.

numbers $12, 56, 992$ are precisely double the first three perfect numbers, $6, 28, 496$, leading us to suspect that integers that are exactly double perfect numbers may come up unusually often. We proved that each of these numbers in fact occurs infinitely often (see Theorem 2), and we note that these numbers are a special case of the set described in [Anavi et al. 2013]. The "obvious" set mentioned above contains multiples of both perfect and multiply perfect numbers — it's clear that these values of the excedent function behave quite differently than do other values. Anavi et al. [2013] refer to these as *regular* solutions of $\sigma(n) \equiv a \pmod n$, as opposed to the other, *sporadic* solutions.

Similarly, the observation that odd numbers occur in the image of $e(n)$ infrequently led us to seek a classification for those n with $e(n)$ odd. We succeeded in completely classifying these values; see Theorem 1.

A final observation we made was that, among the odd values of $s(n)$, many integers that are one less than a power of 2 seemed to appear. An inquiry into these numbers led us to the discovery that every Mersenne prime is the excedent of at least one positive integer. This is proven in Theorem 3.

m	rank of apparition of m	m	rank of apparition of m
−2	3	2	20
−4	5	4	12
−6	7	6	8925
−8	22	8	56
−10	11	10	40
−12	13	12	24
−14	27	14	272
−16	17	16	550
−18	19	18	208
−20	46	20	176

Table 2. The smallest n for which $e(n) = m$ for small even m.

4. Ranks of apparition

If we wish to decide whether an integer is an excedent, it would be helpful to know how far we ought to search via brute force before believing that an integer which has not yet appeared as an excedent will never appear. We ask then, for a given excedent m, what is the smallest integer n for which $e(n) = m$? We shall refer to this n as the *rank of apparition* of m. If all m that are excedents have small rank of apparition, we may trust that for all m, either m is the excedent of a small integer, or it is never the excedent of any.

Table 2 gives the rank of apparition of all even integers m with $|m| \leq 20$. It suggest that if m is an even excedent of any integer, it is likely the excedent of a rather small integer. Indeed the rank of apparition of all even m in the range $-20 \leq m \leq 20$ is under 10,000.

For odd excedents, the situation is quite different. Recall that for some odd values m (including 1), we do not know whether m is ever an excedent. Table 3 lists every small odd integer which is the excedent of some $n < 10^{20}$, together with its rank of apparition.

The fact that some values, say $m = -11$, don't appear until over 200,000 makes us hesitate to claim that values which don't come up early will never appear. In fact, the situation is even worse than this. The smallest integer we found (in absolute value) whose rank of apparition is more than 10^6 is 127, for which $g(127) = 1032256$. Similarly, $g(1529) = 66324736$ is the smallest known m for which the rank of apparition is greater than 10^7. If we want any hope of putting together a list of excedents and nonexcedents, then, we shall clearly have to extend our search beyond these small values.

m	rank of apparition of m	m	rank of apparition of m
-1	1	3	18
-5	9	7	196
-7	50	17	100
-11	244036	19	36
-19	25	31	15376
-25	98	39	162
-47	484		

Table 3. The smallest n for which $e(n) = m$ for small odd m.

5. Results

As described above, most of our results were motivated by a careful observation of a large amount of data. We here state and prove the three theorems briefly mentioned above, which constitute the primary theoretical results of our research.

Before proceeding, we wish to remind the reader of some basic facts about the function $\sigma(n)$. There are two properties of $\sigma(n)$ which we shall need. First, for a prime power p^k, we have

$$\sigma(p^k) = \frac{p^{k+1} - 1}{p - 1}.$$

Second, $\sigma(n)$ is multiplicative. That is, if a and b are relatively prime, then $\sigma(ab) = \sigma(a)\sigma(b)$. It seems likely that this property, which is so useful in a large number of applications, is the primary reason that so much more attention has been paid to $\sigma(n)$ than to $s(n)$. From these two facts, it is fairly straightforward to show the following theorem.

Theorem 1. *The excedent of n, $e(n)$, is odd if and only if $n = k^2$ or $n = 2k^2$ for some integer k.*

This theorem is enormously useful. Since we have already determined (computationally) that odd excedents are rare, we wish to extend our search for these numbers. Thanks to Theorem 1, if we wish to look for odd excedents, we now know that we need to consider only squares and numbers that are double a square. We will use this to great effect for our computations in Section 7. While Theorem 1 is crucial in our work below, it is not original. It is similar to one often encountered in number theory courses; see [Burton 1976, Chapter 6, Exercise 7], for example.

Our second theorem concerns one family of numbers (probably infinite), all of which appear in the image of the excedent function. Although this theorem was

new to us, we have learned that this is not the first time it has appeared in the literature. The referee pointed out that this theorem appears in more general form in [Pomerance 1975], from which we learned that the first appearance of Theorem 2 was in a note by Mąkowski [1960].

Theorem 2. *If N is a perfect number, then $2N$ will be the excedent of infinitely many integers m. In particular, if p is a prime not dividing $2N$, then $2N$ is the excedent of $2pN$.*

Although the proof of this theorem can be found with a literature search, the reader is encouraged to try to prove it herself. It takes only straightforward calculation. Indeed, this result also can now be found in some elementary number theory texts; Theorem 2 appears, for example, as Exercise 21 of [Robbins 2006].

Somewhat disappointingly, despite the fact that our third theorem was new when we proved it, a proof appeared in a paper by Pollack and Shevelev [2012] after our work was submitted to *Involve*. We discovered this work while reading references recommended by the referee during revisions.

Theorem 3. *Let $M_p = 2^p - 1$ be a Mersenne prime. Then $2^p - 1$ will be in the image of the excedent function. In particular,*

$$e(2^{p-1}M_p^2) = M_p.$$

Proof. Let $M_p = 2^p - 1$ be a Mersenne prime, and let $n = 2^{p-1}M_p^2$. We wish to show that $e(n) = \sigma(n) - 2n = M_p$. Because n is already written as a power of 2 multiplied by an odd prime, we can use the multiplicativity of $\sigma(n)$ to write

$$\begin{aligned}
\sigma(2^{p-1}M_p^2) &= \sigma(2^{p-1})\sigma(M_p^2) \\
&= (2^p - 1)(M_p^2 + M_p + 1) \\
&= (2^p - 1)(M_p^2 + 2^p) \\
&= 2^p M_p^2 + 2^{2p} - M_p^2 - 2^p \\
&= 2^p M_p^2 + 2^p M_p - M_p^2 \\
&= 2^p M_p^2 + M_p(2^p - M_p) \\
&= 2^p M_p^2 + M_p.
\end{aligned}$$

Then, since $n = 2^{p-1}M_p^2$ and $\sigma(n) = 2^p M_p^2 + M_p$, we have that the excedent of n, $\sigma(n) - 2n$, is

$$e(n) = (2^p M_p^2 + M_p) - 2(2^{p-1}M_p^2) = M_p,$$

as desired. □

6. Arithmetic progressions

As we noted above, the set of Mersenne primes is a (probably) infinite family of values of the excedent function. We might then ask: are there any provably infinite families of excedents? A bit of thought reveals the answer to be in the affirmative. For example, $e(p) = -(p-1)$ for any prime p, so any integer of the form $-(p-1)$ is certainly an excedent. Indeed, we could find several other infinite families of excedents in terms of their prime factorization as well. Rather than pursue this avenue of study, however, we would like to turn our attention to one more idea — looking for excedents in arithmetic progressions.

To this end, we present one more theorem, and the result of one intriguing computation. The demonstration of the following theorem relies on the Goldbach conjecture. The Goldbach conjecture, as it is usually stated, is that every even integer greater than 2 is the sum of two primes. Although the problem remains open, van der Corput [1936; 1938], Estermann [1938] and Chudakov [1937] each proved independently that *almost* every even number is the sum of two primes — that is, every even number is the sum of two primes, except possibly for a set of density zero.

We should note that this implies a related fact which will prove useful to us. Since the density of integers of the form $2p$ for prime p has density zero, we can also say that almost every even number is the sum of two *distinct* primes. This fact will allow us to prove the following.

Theorem 4. *Every integer* $n \equiv 12 \pmod{24}$ *is contained in the image of the excedent function, except perhaps for a set of density* 0.

Proof. Let $n = pq$, with p and q both prime. Then $s(n) = p + q + 1$. Since, by the discussion above, we know that almost all even integers can be written in the form $p + q$ for distinct p and q, it follows that almost odd integers can be written in the form $p + q + 1$. Thus, we have that almost all odd integers are in the image of $s(n)$.

Now let m be any integer relatively prime to 6, so that $m \equiv 1$ or $5 \pmod{6}$. For such an m, $e(12m)$ has an interesting form. We see this by writing

$$\begin{aligned} e(6m) &= \sigma(6m) - 2(6m) = \sigma(6)\sigma(m) - 12m \\ &= 12(m + s(m)) - 12m = 12s(m). \end{aligned}$$

Since numbers relatively prime to 6 are odd, they can almost all be written as $s(m)$ for some m, and therefore almost all numbers of the form $12(2k + 1)$ lie in the image of $e(n)$, from which the theorem follows. □

Finding this arithmetic progression of excedents raises the obvious question of whether there are other arithmetic progressions that are (almost) all contained in the image of the excedent function. Preliminary computations show that this may be a

m	residue class $k \pmod m$
8	4
12	4, 8
16	4, 8, 12
18	6
20	4, 8, 12, 16
24	4, 8, 12, 16, 20
26	2
28	4, 8, 12, 14, 16, 20, 24
30	12, 14, 18, 26
32	4, 8, 12, 16, 20, 24, 28
34	2, 10, 12, 22, 26
36	4, 6, 8, 12, 16, 18, 20, 24, 28, 32
38	8, 12, 20, 22, 30
40	4, 8, 12, 16, 20, 24, 28, 32, 34, 36
42	2, 6, 12, 14, 18, 24, 28, 32, 34, 36, 38
44	4, 8, 12, 16, 20, 24, 28, 32, 36, 40
46	2, 10, 14, 18, 22, 26, 30, 40
48	4, 8, 12, 14, 16, 20, 24, 28, 30, 32, 36, 40, 42, 44
50	4, 6, 12, 16, 20, 22, 24, 32, 38, 46
52	2, 4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44, 48
54	2, 6, 8, 12, 14, 18, 20, 24, 32, 36, 42, 52
56	4, 8, 10, 12, 14, 16, 20, 24, 28, 32, 34, 36, 40 42, 44, 48, 52, 54
58	26, 34, 38, 40, 42, 44, 46, 50, 52, 54
60	4, 8, 12, 14, 16, 18, 20, 24, 26, 28, 30, 32, 36, 40, 42, 44, 48, 52

Table 4. Every integer up to 10,000 lying in one of the residue classes listed here is contained in the image of the excedent function.

promising line of inquiry. We searched for arithmetic progressions all of whose members up to 10,000 are contained in the excedents we have found. They are listed in Table 4.

Of all the residue classes in Table 4, we have succeeded in explaining only the class 12 (mod 24). We encourage others to use the ideas above to see if more of these classes can be proven to lie entirely (or almost entirely) in the image of $e(n)$.

7. Computational results (redux)

By Theorem 1, we know that all odd excedents are the image under $e(n)$ of an integer of the form k^2 or $2k^2$. Therefore, if we wish to search just for odd excedents, we need only look at numbers of this specialized form. We therefore revised our earlier search to consider only squares and double squares, and were able to extend

Bound on n	$e(n)$ even $0 < e(n)$	$e(n)$ even $e(n) < 0$	$e(n)$ odd $0 < e(n)$	$e(n)$ odd $e(n) < 0$	Total $-10^4 \le e(n) \le 10^4$
10^4	0.6126	0.2202	0.0166	0.0134	0.2157
10^5	0.9378	0.5888	0.0320	0.0240	0.3956
10^6	0.9722	0.6922	0.0370	0.0310	0.4330
10^7	0.9832	0.7618	0.0400	0.0328	0.4544
10^8	0.9894	0.8390	0.0408	0.0334	0.4756
10^9	0.9894	0.8390	0.0408	0.0334	0.4756
10^{10}	0.9894	0.8390	0.0408	0.0334	0.4756

Table 5. The proportion of integers m with $|e(m)| \le 10^4$ in various classes that are excedents of a number less than the given bound.

our preliminary computation by several orders of magnitude.

In the end, we computed the value of $e(n)$ for $n = k^2$ and $n = 2k^2$ for all n up to 10^{20}. Despite searching to this large value, we find that of the fifty odd values of m with $-50 < m < 50$, thirty-two of them are never in the image of the excedent function. The values that never occur are

$$-49, -45, -43, -39, -35, -33, -31, -29, -27, -23, -21, -17, -15, -13, -9, -3,$$
$$1, 5, 9, 11, 13, 15, 21, 23, 25, 27, 29, 33, 35, 37, 43, 45. \quad (1)$$

Among the positive nonexcedents are those studied by Cohen — all the odd squares appear on the list. There are, however, many other odd values that never appear. We cannot explain these values or find any way to classify them, nor do they appear in Sloane's *Online encyclopedia of integer sequences*.

We can, however, use our data to speculate about the density of integers that are excedents. We recorded all excedents of integers up to 10^{10} with absolute value less than 10,000, and we shall use these to get an idea about the density of integers that are excedents. In the table below, we give the proportion of integers that are excedents. Because even excedents behave differently than odd excedents, and because the sign of an integer also seems to affect its probability of being an excedent, we first break integers into four groups (by parity and sign) and consider these proportions separately.

Based on this (admittedly limited) data, it seems reasonable to conjecture that most positive integers are excedents.

8. Conjectures and future work

The theorems above represent observations we made based on our data, and which we have been able to prove. We have also made other observations which we have been unable to prove. Among these are:

Conjecture 5. Every even number is the excedent of at least one positive integer.

Up to 10^8, our computational data show that every even integer n satisfying

$$-480 < n < 130$$

is the excedent of some integer, and we see no reason to expect that any even number will not appear on the list of excedents at some point. We saw in Table 2 that it seems to be the case that if m is an even excedent of any integer, it is likely the excedent of a rather small integer, but when we extend to m beyond the range of Table 2, we actually do find some even numbers k appear in the image of $e(n)$ only for fairly large n. For example, the smallest n such that $e(n) = -384$ is $n = 99413968$.

Conjecture 6. The values given in (1), giving integers that are not in the image of the excedent function for any $n \leq 10^{20}$, are in fact nonexcedents, and will never be in the image of this function.

This conjecture seems less certain. Since we know that there are some k which appear in the image of $e(n)$ only for large n, it is certainly possible that one (or more!) of the values in (1) may yet appear. However, we find no new excedents with absolute value less than 100 appear for any integers greater than 10^9 — we believe that these exceptional values are unlikely to appear after 10^{20}.

There are several other open questions. Can one find an infinite family of integers all of which are nonexcedents? Possibly easier: do the excedents have a density in the integers? If so, what is it? It is striking that more than 2500 years after the concept was first considered by the Pythagoreans, questions about the excedent of an integer continue to beguile and challenge us. It is our hope that these preliminary investigations may serve as a catalyst for further research on the excedent function.

References

[Anavi et al. 2013] A. Anavi, P. Pollack, and C. Pomerance, "On congruences of the form $\sigma(n) \equiv a \pmod{n}$", *Int. J. Number Theory* **9**:1 (2013), 115–124. MR 2997493 Zbl 06132096

[Burton 1976] D. M. Burton, *Elementary number theory*, Allyn and Bacon, Boston, 1976. Reprinted McGraw-Hill, 1998. MR 81c:10001a Zbl 0314.10001

[Cattaneo 1951] P. Cattaneo, "Sui numeri quasiperfetti", *Boll. Un. Mat. Ital* (3) **6** (1951), 59–62. Zbl 0042.26804

[Chen and Zhao 2011] Y.-G. Chen and Q.-Q. Zhao, "Nonaliquot numbers", *Publ. Math. Debrecen* **78**:2 (2011), 439–442. MR 2011m:11014 Zbl 1240.11007

[Chudakov 1937] N. G. Chudakov, "О проблеме Гольдбаха", *Dokl. Akad. Nauk SSSR* (*N.S.*) **17** (1937), 335–338. Zbl 0018.00603

[Cohen 1982] G. L. Cohen, *Generalised quasiperfect numbers*, Ph.D. thesis, Univ. New South Wales, Sydney, 1982. Abstracted in *Bull. Australian Math. Soc.* **27**:1 (1983), 153–155. Zbl 0494.10003

[van der Corput 1936] J. van der Corput, "Sur l'hypothèse de Goldbach pour presque tous les nombres pairs", *Acta Arith.* **2**:2 (1936), 266–290.

[van der Corput 1938] J. van der Corput, "Sur l'hypothèse de Goldbach", *Proc. Akad. Wet. Amsterdam* **41** (1938), 76–80. Zbl 0018.24408 JFM 64.0127.01

[Erdős 1973] P. Erdős, "Über die Zahlen der Form $\sigma(n)-n$ und $n-\phi(n)$", *Elem. Math.* **28** (1973), 83–86. MR 49 #2502 Zbl 0272.10003

[Estermann 1938] T. Estermann, "On Goldbach's problem: proof that almost all even positive integers are sums of two primes", *Proc. Lond. Math. Soc.* (2) **2**:1 (1938), 307–314. Zbl 0020.10503 JFM 64.0126.05

[Hagis and Cohen 1982] P. Hagis, Jr. and G. L. Cohen, "Some results concerning quasiperfect numbers", *J. Austral. Math. Soc. Ser. A* **33**:2 (1982), 275–286. MR 84f:10008 Zbl 0494.10002

[Mąkowski 1960] A. Mąkowski, "Remarques sur les fonctions $\theta(n)$, $\varphi(n)$ et $\sigma(n)$", *Mathesis* **69** (1960), 302–303. MR 23 #A834

[Pollack and Pomerance 2013] P. Pollack and C. Pomerance, "On the distribution of some integers related to perfect and amicable numbers", *Colloq. Math.* **130**:2 (2013), 169–182. MR 3049062 Zbl 06156689

[Pollack and Shevelev 2012] P. Pollack and V. Shevelev, "On perfect and near-perfect numbers", *J. Number Theory* **132**:12 (2012), 3037–3046. MR 2965207 Zbl 06097278

[Pomerance 1975] C. Pomerance, "On the congruences $\sigma(n) \equiv a \pmod{n}$ and $n \equiv a \pmod{\varphi(n)}$", *Acta Arith.* **26**:3 (1975), 265–272. MR 52 #5535 Zbl 0266.10005

[Pomerance and Yang 2012] C. Pomerance and H.-S. Yang, "Variant of a theorem of Erdős on the sum-of-proper-divisors function", (2012).

[Robbins 2006] N. Robbins, *Beginning number theory*, 2nd ed., Jones & Bartlett, Sudbury, MA, 2006.

Received: 2012-12-07 Revised: 2013-05-18 Accepted: 2013-05-20

davisni@gwmail.cwu.edu *13561 Macadam Road, South, Tukwila, WA 98168, United States*

klyved@cwu.edu *2400 North Ellington Street, Ellenburg, WA 98926, United States*

kraghtn@gwmail.cwu.edu *14235 SE 224th Street, Kent, WA 98042, United States*

INVOLVE 6:4 (2013)
dx.doi.org/10.2140/involve.2013.6.505

A Pexider difference associated to a Pexider quartic functional equation in topological vector spaces

Saeid Ostadbashi, Abbas Najati,
Mahsa Solaimaninia and Themistocles M. Rassias

(Communicated by Martin Bohner)

Let $(G, +)$ be an Abelian group and X be a sequentially complete Hausdorff topological vector space over the field $\mathbb{Q}$ of rational numbers. We deal with a Pexider difference

$$2f(2x+y)+2f(2x-y)-2g(x+y)-2g(x-y)-12g(x)+3g(y),$$

where f and g are mappings defined on G and taking values in X. We investigate the Hyers–Ulam stability of the Pexiderized quartic functional equation

$$2f(2x+y)+2f(2x-y)=2g(x+y)+2g(x-y)+12g(x)-3g(y)$$

in topological vector spaces.

1. Introduction and preliminaries

The stability problem concerning the stability of group homomorphisms originated from a question of Ulam [1964] and was answered affirmatively by Hyers [1941] for Banach spaces. This result was generalized by Aoki [1950] for additive mappings and by Rassias [1978] for linear mappings by considering an unbounded Cauchy difference. The question of stability can be raised not only concerning the Cauchy functional equation but also in connection with other functional equations. For more concerning the stability results of functional equations, see [Czerwik 2002; 2003; Hyers et al. 1998; Jung 2001; Forti 1995; Hyers and Rassias 1992]. The stability of the quartic functional equation has been investigated in [Cădariu and Radu 2004; Chung and Sahoo 2003; Lee et al. 2005; Najati 2008].

Adam and Czerwik [2007] investigated the problem of the Hyers–Ulam stability of a generalized quadratic functional equation in linear topological spaces. In this

MSC2010: primary 39B82; secondary 34K20, 54A20.
Keywords: Hyers–Ulam stability, quartic mapping, topological vector space.

paper, we prove that the Pexiderized quartic functional equation

$$2f(2x+y)+2f(2x-y)=2g(x+y)+2g(x-y)+12g(x)-3g(y)$$

is stable for functions f, g defined on an Abelian group and taking values in a topological vector space.

Let G be an Abelian group and throughout this paper let X be a sequentially complete Hausdorff topological vector space over the field $\mathbb{Q}$ of rational numbers. A mapping $f : G \to X$ is *quartic* if it satisfies the functional equation

$$f(2x+y)+f(2x-y)=4f(x+y)+4f(x-y)+24f(x)-6f(y)$$

for all $x, y \in G$. This equation is called the *quartic functional equation*. For a given $f : G \to X$, we will use the notation

$$Df(x, y) := f(2x+y)+f(2x-y)-4f(x+y)-4f(x-y)-24f(x)+6f(y).$$

For given sets $A, B \subseteq X$ and a number $k \in \mathbb{R}$, we define the well-known operations

$$A+B := \{a+b : a \in A, b \in B\}, \quad kA := \{ka : a \in A\}.$$

We denote the convex hull of a set $U \subseteq X$ by $\operatorname{conv}(U)$ and the sequential closure of U by $\overline{U}$. Moreover it is well-known that:

(i) If $A, B \subseteq X$ are bounded sets, then $A+B$ $\operatorname{conv}(A)$ and $\bar{A}$ are bounded subsets of X.

(ii) If $A, B \subseteq X$ and $\alpha, \beta \in \mathbb{R}$, then $\alpha \operatorname{conv}(A) + \beta \operatorname{conv}(B) = \operatorname{conv}(\alpha A + \beta B)$.

(iii) Let X_1 and X_2 be linear spaces over $\mathbb{R}$. If $f : X_1 \to X_2$ is a quartic function, then $f(rx) = r^4 f(x)$ for all $x \in X_1$ and all $r \in \mathbb{Q}$.

2. Main results

We start with the following lemma.

Lemma 2.1. *Let G be an Abelian group and $B \subseteq X$ be a nonempty set. If the even functions $f, g : G \to X$ satisfy*

$$f(2x+y)+f(2x-y)-g(x+y)-g(x-y)-6g(x)+\tfrac{3}{2}g(y) \in B \qquad (2\text{-}1)$$

for all $x, y \in G$, then

$$Df(x, y)+24f(0) \in 16\operatorname{conv}(B-B), \qquad (2\text{-}2)$$

$$Dg(x, y)+24g(0) \in 4\operatorname{conv}(B-B) \qquad (2\text{-}3)$$

for all $x, y \in G$.

Proof. Putting $x = 0$ in (2-1), we get

$$4f(y) - g(y) - 12g(0) \in 2B \tag{2-4}$$

for all $y \in G$. If we put $x = y = 0$ in (2-1), then we have

$$4f(0) - 13g(0) \in 2B. \tag{2-5}$$

It follows from (2-4) and (2-5) that, for all $x, y \in G$,

$$\begin{aligned}
&Df(x, y) + 24f(0)\\
&= [f(2x + y) + f(2x - y) - g(x + y) - g(x - y) - 6g(x) + \tfrac{3}{2}g(y)]\\
&\quad - [4f(x + y) - g(x + y) - 12g(0)] - [4f(x - y) - g(x - y) - 12g(0)]\\
&\quad - [24f(x) - 6g(x) - 72g(0)] + [6f(y) - \tfrac{3}{2}g(y) - 18g(0)] + [24f(0) - 78g(0)],
\end{aligned}$$

which lies in $12\,\mathrm{conv}(B) + 12\,\mathrm{conv}(-B) = 16\,\mathrm{conv}(B - B)$. This proves (2-2). Moreover, we have, for all $x, y \in G$,

$$\begin{aligned}
&Dg(x, y) + 24g(0)\\
&= [4f(2x + y) + 4f(2x - y) - 4g(x + y) - 4g(x - y) - 24g(x) + 6g(y)]\\
&\quad - [4f(2x + y) - g(2x + y) - 12g(0)] - [4f(2x - y) - g(2x - y) - 12g(0)]
\end{aligned}$$

which lies in $4\,\mathrm{conv}(B) + 4\,\mathrm{conv}(-B) = 4\,\mathrm{conv}(B - B)$. Hence we get (2-3). □

Theorem 2.2. *Let G be an Abelian group and $B \subseteq X$ be a nonempty bounded set. Suppose that the even functions $f, g : G \to X$ satisfy* (2-1) *for all $x, y \in G$. Then there exists exactly one quartic function $\mathcal{Q} : G \to X$ such that*

$$\begin{aligned}
\mathcal{Q}(x) - f(x) + f(0) &\in \tfrac{8}{15}\,\overline{\mathrm{conv}(B - B)},\\
4\mathcal{Q}(x) - g(x) + g(0) &\in \tfrac{2}{15}\,\overline{\mathrm{conv}(B - B)}
\end{aligned}$$

for all $x \in G$. Moreover, the function $\mathcal{Q}$ is given by

$$\mathcal{Q}(x) = \lim_{n\to\infty} \frac{1}{2^{4n}} f(2^n x) = \frac{1}{4} \lim_{n\to\infty} \frac{1}{2^{4n}} g(2^n x) \quad \text{for } x \in G,$$

and the convergence of the sequences are uniform on G.

Proof. By Lemma 2.1, we have

$$Df(x, y) \in -24f(0) + 16\,\mathrm{conv}(B - B) \tag{2-6}$$

for all $x, y \in G$. Setting $y = 0$ in (2-6), we get

$$2f(2x) - 32f(x) \in -30f(0) + 16\,\mathrm{conv}(B - B)$$

for all $x \in G$. Therefore

$$\frac{1}{2^4} f(2x) - f(x) \in \frac{1}{2^4} \tilde{B} \tag{2-7}$$

for all $x \in G$, where $\tilde{B} := -15f(0) + 8\,\mathrm{conv}(B - B)$. It is clear that $\tilde{B}$ is convex. Replacing x by $2^n x$ in (2-7), we infer that

$$\frac{1}{2^{4(n+1)}} f(2^{n+1}x) - \frac{1}{2^{4n}} f(2^n x) \in \frac{1}{2^{4(n+1)}} \tilde{B}$$

for all $x \in G$ and all integers $n \geqslant 0$. Therefore

$$\begin{aligned} \frac{1}{2^{4n}} f(2^n x) - \frac{1}{2^{4m}} f(2^m x) &= \sum_{k=m}^{n-1} \Big[\frac{1}{2^{4(k+1)}} f(2^{k+1}x) - \frac{1}{2^{4k}} f(2^k x) \Big] \\ &\in \sum_{k=m}^{n-1} \frac{1}{2^{4(k+1)}} \tilde{B} \subseteq \frac{1}{15 \times 2^{4m}} \tilde{B} \end{aligned} \tag{2-8}$$

for all $x \in G$ and all integers $n > m \geqslant 0$. Since B is bounded, we conclude that $\tilde{B}$ is bounded. It follows from (2-8) and boundedness of the set $\tilde{B}$ that the sequence $\{(1/2^{4n}) f(2^n x)\}$ is (uniformly) Cauchy in X for all $x \in G$. Since X is a sequential complete topological vector space, the sequence $\{(1/2^{4n}) f(2^n x)\}$ is convergent for all $x \in G$, and the convergence is uniform on G. Define

$$\mathcal{Q}_1 : G \to X, \quad \mathcal{Q}_1(x) := \lim_{n\to\infty} \frac{1}{2^{4n}} f(2^n x).$$

Since $-24f(0) + 16\,\mathrm{conv}(B - B)$ is bounded, it follows from (2-6) that

$$D\mathcal{Q}_1(x, y) = \lim_{n\to\infty} \frac{1}{2^{4n}} Df\big(2^n x, 2^n y\big) = 0$$

for all $x, y \in G$. So $\mathcal{Q}_1$ is quartic. Letting $m = 0$ and $n \to \infty$ in (2-8), we get

$$\mathcal{Q}_1(x) - f(x) + f(0) \in \tfrac{8}{15}\,\overline{\mathrm{conv}(B - B)} \tag{2-9}$$

for all $x \in G$. Applying (2-3) as before, we have

$$\frac{1}{2^{4n}} g(2^n x) - \frac{1}{2^{4m}} g(2^m x) \in \sum_{k=m}^{n-1} \frac{1}{2^{4(k+1)}} \tilde{C} \subseteq \frac{1}{15 \times 2^{4m}} \tilde{C} \tag{2-10}$$

for all $x \in G$, where $\tilde{C} := -15g(0) + 2\,\mathrm{conv}(B - B)$. Then $\{(1/2^{4n}) g(2^n x)\}$ is a (uniformly) Cauchy sequence in X for all $x \in G$. Define

$$\mathcal{Q}_2 : G \to X, \quad \mathcal{Q}_2(x) := \lim_{n\to\infty} \frac{1}{2^{4n}} g(2^n x).$$

As before, we can check that $\mathcal{Q}_2$ is a quartic function satisfying

$$\mathcal{Q}_2(x) - g(x) + g(0) \in \tfrac{2}{15}\,\overline{\mathrm{conv}(B - B)} \tag{2-11}$$

for all $x \in G$. To prove the equality $4\mathcal{Q}_1 = \mathcal{Q}_2$, we have

$$4\mathcal{Q}_1(x) - \mathcal{Q}_2(x) = [4\mathcal{Q}_1(x) - 4f(x)] - [\mathcal{Q}_2(x) - g(x)] + [4f(x) - g(x)]$$

for all $x \in G$. Applying (2-4), (2-5), (2-9) and (2-11) in the above equation, we get

$$4\mathcal{Q}_1(x) - \mathcal{Q}_2(x) \in M := 2\,\overline{\mathrm{conv}(B - B)} + 2(B - B) \tag{2-12}$$

for all $x \in G$. Replacing x by $2^n x$ in (2-12), we get

$$4\mathcal{Q}_1(2^n x) - \mathcal{Q}_2(2^n x) \in M$$

for all $x \in G$ and all integers n. Since $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are quartic, we obtain

$$4\mathcal{Q}_1(x) - \mathcal{Q}_2(x) \in \frac{1}{2^{4n}} M \tag{2-13}$$

for all $x \in G$. Since M is bounded, letting $n \to \infty$ in (2-13) we obtain $4\mathcal{Q}_1 = \mathcal{Q}_2$. Assuming $\mathcal{Q} := \mathcal{Q}_1$, we can see that the conditions of theorem are satisfied.

To prove uniqueness, suppose that there exists another quartic function $\mathcal{Q}' : G \to X$ satisfying

$$\mathcal{Q}'(x) - f(x) + f(0) \in \tfrac{8}{15}\,\overline{\mathrm{conv}(B - B)}$$

for all $x \in G$. Then we have

$$\mathcal{Q}'(x) - \mathcal{Q}(x) = [\mathcal{Q}'(x) - f(x) + f(0)] - [\mathcal{Q}(x) - f(x) + f(0)] \in \tfrac{16}{15}\,\overline{\mathrm{conv}(B - B)}$$

for all $x \in G$. Applying the same method as before, we get $\mathcal{Q}' = \mathcal{Q}$. This completes the proof. □

Acknowledgements

The authors would like to thank the referee for useful comments.

References

[Adam and Czerwik 2007] M. Adam and S. Czerwik, "On the stability of the quadratic functional equation in topological spaces", *Banach J. Math. Anal.* **1**:2 (2007), 245–251. MR 2366108 Zbl 1130.39021

[Aoki 1950] T. Aoki, "On the stability of the linear transformation in Banach spaces", *J. Math. Soc. Japan* **2** (1950), 64–66. MR 12,717a Zbl 0040.35501

[Cădariu and Radu 2004] L. Cădariu and V. Radu, "A Hyers–Ulam–Rassias stability theorem for a quartic functional equation", *Automat. Comput. Appl. Math.* **13**:1 (2004), 31–39. MR 2009f:39050

[Chung and Sahoo 2003] J. K. Chung and P. K. Sahoo, "On the general solution of a quartic functional equation", *Bull. Korean Math. Soc.* **40**:4 (2003), 565–576. MR 2004h:39059 Zbl 1048.39017

[Czerwik 2002] S. Czerwik, *Functional equations and inequalities in several variables*, World Scientific, River Edge, NJ, 2002. MR 2003b:39027 Zbl 1011.39019

[Czerwik 2003] S. Czerwik, *Stability of functional equations of Ulam–Hyers–Rassias type*, Hadronic Press, Palm Harbor, FL, 2003.

[Forti 1995] G. L. Forti, "Hyers–Ulam stability of functional equations in several variables", *Aequationes Math.* **50**:1-2 (1995), 143–190. MR 96i:39033 Zbl 0836.39007

[Hyers 1941] D. H. Hyers, "On the stability of the linear functional equation", *Proc. Nat. Acad. Sci. U. S. A.* **27** (1941), 222–224. MR 2,315a Zbl 0061.26403

[Hyers and Rassias 1992] D. H. Hyers and T. M. Rassias, "Approximate homomorphisms", *Aequationes Math.* **44**:2-3 (1992), 125–153. MR 93i:39007 Zbl 0806.47056

[Hyers et al. 1998] D. H. Hyers, G. Isac, and T. M. Rassias, *Stability of functional equations in several variables*, Progress in Nonlinear Differential Equations and their Applications **34**, Birkhäuser, Boston, MA, 1998. MR 99i:39035 Zbl 0907.39025

[Jung 2001] S.-M. Jung, *Hyers–Ulam–Rassias stability of functional equations in mathematical analysis*, Hadronic Press, Palm Harbor, FL, 2001. MR 2003b:39035 Zbl 0980.39024

[Lee et al. 2005] S. H. Lee, S. M. Im, and I. S. Hwang, "Quartic functional equations", *J. Math. Anal. Appl.* **307**:2 (2005), 387–394. MR 2006d:39047 Zbl 1072.39024

[Najati 2008] A. Najati, "On the stability of a quartic functional equation", *J. Math. Anal. Appl.* **340**:1 (2008), 569–574. MR 2376178 Zbl 1133.39030

[Rassias 1978] T. M. Rassias, "On the stability of the linear mapping in Banach spaces", *Proc. Amer. Math. Soc.* **72**:2 (1978), 297–300. MR 80d:47094 Zbl 0398.47040

[Ulam 1964] S. M. Ulam, *Problems in modern mathematics*, Wiley, New York, 1964. MR 43 #6031 Zbl 0137.24201

Received: 2013-01-23 Accepted: 2013-01-28

s.ostadbashi@urmia.ac.ir *Department of Mathematics, Urmia University, Urmia 57561-51818, Iran*

a.najati@uma.ac.ir *Department of Mathematics, Faculty of Mathematical Sciences, University of Mohaghegh Ardabili, Ardabil 56199-11367, Iran*

solaimaninia@urmia.ac.ir *Department of Mathematics, Urmia University, Urmia 57561-51818, Iran*

trassias@math.ntua.gr *Department of Mathematics, Zografou Campus, National Technical University of Athens, 15780 Athens, Greece*

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@msp.org with details about how your graphics were generated.

White space. Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Proofs. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

involve
2013 vol. 6 no. 4

1944-4176(2013)6:4;1-7